*DISS. ETH NO...24808…*

# *Investigation into the cause of fertility restoration in cytoplasmic male sterile perennial ryegrass (Lolium perenne* L.*)*

*A thesis submitted to attain the degree of*
*DOCTOR OF SCIENCES of ETH ZURICH*
*(Dr. sc. ETH Zurich)*

*presented by*

*TIMOTHY ANDREW SYKES*
*MSc. Biotechnology (Distinction), University of Essex, UK*

*born on 10.08.1984*

*citizen of*
*Australia*

*accepted on the recommendation of*

*Prof. Dr. Bruno Studer, examiner*
*Prof. Dr. Torben Asp, co-examiner*
*Prof. Dr. Ian Max Møller, co-examiner*

*2017*

# TABLE OF CONTENTS

**SUMMARY** 5

**ZUSAMMENFASSUNG** 7

**CHAPTER 1 - GENERAL INTRODUCTION** 9

**THE IMPORTANCE OF PLANT BREEDING** 9
**CYTOPLASMIC MALE STERILITY** 10
**FORAGE GRASSES** 12
**CONTROVERSIES** 13
**THESIS OUTLINE** 13
**REFERENCES** 15

**CHAPTER 2 - IN SILICO IDENTIFICATION OF CANDIDATE GENES FOR FERTILITY RESTORATION IN CYTOPLASMIC MALE STERILE PERENNIAL RYEGRASS (*LOLIUM PERENNE* L.)** 19

**ABSTRACT** 20
**INTRODUCTION** 21
**RESULTS** 22
PENTATRICOPEPTIDE REPEAT (PPR) AND RESTORER OF FERTILITY-LIKE PPR GENE IDENTIFICATION AND CLASSIFICATION 22
RESTORER OF FERTILITY-LIKE PPR GENE COMPARISON IN MULTIPLE SPECIES 22
PHYLOGENETIC ANALYSIS OF THE RFL CLUSTER 25
SYNTENY ANALYSIS 29
**DISCUSSION** 30
PPR AND RFL GENES IN PERENNIAL RYEGRASS 30
ORTHOLOGY-BASED STRATEGIES FOR RFL IDENTIFICATION 30
GENOME REGIONS OF ACTIVE RFL GENE GENERATION 31
ACCURACY AND USEFULNESS OF THIS APPROACH 32
THE VALUE OF RF GENES FOR CMS-BASED POLLINATION CONTROL IN FORAGE GRASSES 32
CONCLUSION 33
**METHODS** 34
IDENTIFICATION OF PENTATRICOPEPTIDE REPEAT (PPR) PROTEINS 34
CLASSIFICATION OF PENTATRICOPEPTIDE REPEAT (PPR) PROTEINS 34
IDENTIFICATION OF RESTORER OF FERTILITY-LIKE PPR (RFL) PROTEINS 34
DATABASES 34
ORTHOLOGOUS CLUSTERING OF SPECIES 35
PHYLOGENETIC RECONSTRUCTION AND ANALYSIS 35
**REFERENCES** 37
**SUPPLEMENTARY DATA** 41

**CHAPTER 3 - GENOTYPING BY SEQUENCING OF A POPULATION OF PERENNIAL RYEGRASS *(LOLIUM PERENNE* L.) SEGREGATING FOR CYTOPLASMIC MALE STERILITY RESTORATION** 45

**ABSTRACT** 46
**INTRODUCTION** 47
**METHODS** 48
PLANT MATERIAL 48

SEQUENCING LIBRARY PREPARATION AND DNA SEQUENCING 49
GBS DATA PROCESSING 49
MARKER MAPPING 50
SYNTENY-BASED MARKER MAPPING USING THE PERENNIAL RYEGRASS GENOMEZIPPER 50
QUANTITATIVE TRAIT LOCI IDENTIFICATION 50
**RESULTS** **51**
PHENOTYPING OF FERTILITY RESTORATION 51
GENOTYPING BY SEQUENCING, IDENTIFICATION OF SNP MARKERS SIGNIFICANTLY ASSOCIATED WITH FERTILITY RESTORATION AND MAKER LOCALISATION ON THE PERENNIAL RYEGRASS GENOME 51
VALIDATION OF THE SNP DISCOVERY, ASSOCIATION ANALYSIS AND GENOME POSITIONING APPROACH 52
POSITIONING OF SIGNIFICANT MARKERS USING THE PERENNIAL RYEGRASS GENOMEZIPPER 52
CONSOLIDATION OF THE MARKER LOCATIONS VIA ITALIAN RYEGRASS AND RICE 54
SNP DISCOVERY USING A HIGH-QUALITY GENOME ASSEMBLY OF ITALIAN RYEGRASS 54
**DISCUSSION** **57**
PHENOTYPING RESULTS 57
EFFECTIVENESS OF GBS AS A TOOL TO IDENTIFY QTL 57
MARKER IDENTIFICATION VS QTL IDENTIFICATION 58
**REFERENCES** **59**
**SUPPLEMENTARY DATA** **61**

**CHAPTER 4 - BULK SERGEANT ANALYSIS IN A POPULATION OF PERENNIAL RYEGRASS (*LOLIUM PERENNE* L.) SEGREGATING FOR RESTORATION OF CYTOPLASMIC MALE STERILITY** **67**

**ABSTRACT** **68**
**INTRODUCTION** **69**
**METHODS** **71**
DNA LIBRARY PREPARATION 71
SNP DISCOVERY AND ALLELE FREQUENCY CALCULATION USING A HIGH-QUALITY GENOME ASSEMBLY OF ITALIAN RYEGRASS 71
**RESULTS** **72**
SEQUENCING AND SNP IDENTIFICATION 72
GENETIC CONTROL OF FERTILITY RESTORATION 74
GENES AND PROTEIN DOMAINS LOCATED WITHIN THE SIGNIFICANT SCAFFOLDS AND THEIR QTL 77
REPRESENTATION OF THE MITOCHONDRIAL GENOME WITHIN THE NUCLEAR GENOMES OF PERENNIAL RYEGRASS AND ITALIAN RYEGRASS 77
**DISCUSSION** **79**
THE GENETICS OF CMS FERTILITY RESTORATION IN PERENNIAL RYEGRASS 79
MITOCHONDRIAL ORIGINS OF THE RESTORATION LINKED QTL 79
COMPARISON TO QTL PREVIOUSLY IDENTIFIED USING GENOTYPING BY SEQUENCING 80
CANDIDATE GENE FOR FERTILITY RESTORATION 81
**REFERENCES** **82**
**SUPPLEMENTARY DATA** **84**

**CHAPTER 5 - EXPRESSION ANALYSIS OF GENES INVOLVED IN THE RESTORATION OF CYTOPLASMIC MALE STERILITY IN PERENNIAL RYEGRASS (*LOLIUM PERENNE* L.)** **87**

**ABSTRACT** **88**
**INTRODUCTION** **89**
**METHODS** **91**

TISSUE COLLECTION, RNA ISOLATION AND SEQUENCING 91
RNA SEQUENCING DATA ANALYSIS 91
IDENTIFICATION AND ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES 92
CHLOROPHYLL CONTENT MEASUREMENTS 92
**RESULTS 93**
RNA SEQUENCING AND READ MAPPING TO THE HIGH-QUALITY GENOME ASSEMBLY OF ITALIAN RYEGRASS 93
ANALYSIS OF GENERAL DIFFERENTIAL EXPRESSION DATA BETWEEN GENOTYPE AND TISSUE TYPE SAMPLES 93
CHLOROPHYLL CONTENT ANALYSIS OF PLANTS FROM POPULATIONS SEGREGATING FOR FERTILITY RESTORATION 97
ANALYSIS OF GENE EXPRESSION ACROSS PREVIOUSLY IDENTIFIED QUALITATIVE TRAIT LOCI (QTL) FOR FERTILITY RESTORATION 98
CANDIDATE GENE IDENTIFICATION 103
**DISCUSSION 104**
SYSTEMIC CHANGES IN GENE EXPRESSION 104
CANDIDATE GENES FOR FERTILITY RESTORATION 106
**REFERENCES 107**
**SUPPLEMENTARY DATA 110**

**CHAPTER 6 – GENERAL DISCUSSION 113**

**NUCLEAR INTEGRATED MITOCHONDRIAL GENES AS RESTORERS OF FERTILITY 113**
**ADVANTAGES OF A MULTI-TECHNIQUE APPROACH TO CANDIDATE GENE IDENTIFICATION 115**
**MECHANISMS OF FERTILITY RESTORATION FOR OTHER CYTOPLASMIC MALE STERILITY SYSTEMS 115**
**CYTOPLASMIC MALE STERILITY IN HYBRID BREEDING SCHEMES 117**
**THE GLOBAL POTENTIAL OF HYBRID VARIETIES 118**
**REFERENCES 119**

**ACKNOWLEDGEMENTS 121**

**APPENDICES 123**

**APPENDIX I - GENETIC LOCI GOVERNING ANDROGENIC CAPACITY IN PERENNIAL RYEGRASS (*LOLIUM PERENNE* L.) 123**
ABSTRACT 124
INTRODUCTION 125
MATERIALS AND METHODS 127
RESULTS 130
DISCUSSION 137
REFERENCES 142
SUPPLEMENTARY DATA 146
**APPENDIX II - GENOME EDITING: SCIENTIFIC OPPORTUNITIES, PUBLIC INTERESTS AND POLICY OPTIONS IN THE EUROPEAN UNION 151**
SUMMARY 152
INTRODUCTION 155
PLANTS 160
REFERENCES 163
**APPENDIX III - THE BENEFITS OF NEW PLANT BREEDING TECHNIQUES 165**
**APPENDIX IV - EUCARPIA - PLENARY DISCUSSION: INNOVATION VS REGULATION 167**

**CURRICULUM VITAE 169**

# Summary

Understanding the genetic architecture underlying cytoplasmic male sterility (CMS) is vital if this trait is to be utilised in even more species for the production of high-yielding hybrid varieties. CMS has already been applied to great effect in several agricultural species with the largest yield gains seen in maize, where hybrids now make up to 65% of the global maize production. As global food demand increases, new hybrid varieties, especially in crops where the potential of hybrid breeding is yet to be fully exploited, will play a key role in ensuring that agricultural production can continue to maintain global food security.

The CMS trait is charactorised as an inability of affected individuals to produce viable pollen. This phenotype allows plant breeders to use CMS as a pollination control mechanism during the production of hybrid seed. In order to maintain the CMS trait within breeding programs, molecular markers are needed that allow breeders to track genes that can restore male fertility. Although the cause of CMS is inherited through the cytoplasm, the genes that can restore fertility (*Rf* genes) are nuclear encoded. This makes the CMS trait not just agriculturally important for the production of hybrid seed, but also of fundamental interest in the study of mitochondrial-nuclear genome interactions.

This thesis documents an investigation into the cause of CMS restoration in perennial ryegrass (*Lolium perenne* L). Although the importance of forage grasses such as perennial ryegrass is often over-looked, they account for 80% of milk and 70% of meat production in Europe. A CMS system has been identified in perennial ryegrass although it has yet to be fully charactorised, making it a challenge to utilise for hybrid seed production. The goal of this research was to identify the genes or genomic loci responsible for fertility restoration and in doing so, to reveal the genetic architecture of CMS fertility-restoration in perennial ryegrass. This was achieved through the use of four contrasting approaches, reported in chapters 2-5, which in concert were able to reveal the genetic control of this important trait.

Firstly, a study of the relevant literature revealed that in almost all cases, identified *Rf* genes belong to a class of RNA binding proteins called the pentatricopeptide repeat (PPR) proteins. Although being particularly numerous in plant genomes, with hundreds of members identified, a sub-class of CMS restoring PPRs, the restorer of fertility-like PPRs (RFLs), has been identified. In **chapter 2**, a bioinformatics pipeline is outlined that allows the rapid identification of these *RFL* genes from genomic or transcriptomic sequence data, exploiting sequence similarities between *RFLs*. This pipeline not only allows the simultaneous identification of *RFLs* from multiple species but also permits the identification of active sites of *RFL* generation within a genome. The application of which, to 14 plant species, revealed that 50-90% of *RFL* genes can be found within two to three distinct genomic loci per species.

A population of over 1,200 perennial ryegrass plants segregating for CMS fertility restoration was identified and formed the basis for the investigation into fertility restoration presented here. This population showed no intermediate phenotypes and a fertility restoration rate of 27%, suggesting that a two loci co-dominant control may be responsible for fertility restoration. Three sequencing-based approaches were employed to interrogate this population, the first of which, genotyping by sequencing (GBS), is presented in **chapter 3**. The application of GBS to this population identified four quantitative trait loci (QTL) associated to the restoration of the CMS phenotype, spanning a genome region of 87.3 Mb.

Following this, a bulk sergeant analysis (BSA) was performed on pools of DNA from sterile or fertile plants. As the phenotype of interest is binary, sterile or fertile, this proved to be a powerful way to further resolve the previously identified QTL. The results of the BSA are given in **chapter 4**, where two loci were identified, covering 74 kb of the reference genome. Most strikingly, these two loci were both identified as containing DNA sequence of mitochondrial origin, with one nuclear encoded mitochondrial gene shown to be mutated in the sterile pool. This suggested that functional copies of deficient CMS-causal mitochondrial genes may be responsible for fertility restoration.

Finally, total RNA was collected across three tissue types from the four genotypes present in a CMS breeding program, and a gene expression analysis performed. The results of which, presented in **chapter 5**, revealed several expressed mitochondrial genes within the two QTL identified in chapter 4. Once fully assessed, three strong candidate genes for fertility restoration were identified, two of which were subunits of the mitochondrial respiratory complex IV and have been previously implicated in CMS causation. Mutations in the coding sequence of one of these genes was associated to the sterile phenotype, with both containing predicted mitochondrial transit peptides. In addition to these results, the gene expression analysis uncovered pleiotropic effects associated to the presence of *Rf* genes and highlighted large expression shifts in restored hybrids which suggest a cause for the observations of practical plant breeders that restored hybrids are ‘more vigorous’ than unrestored hybrids.

The results presented here identify a previously undescribed mechanism for CMS fertility restoration and strong targets for marker development within the population studied. Moreover, they add to the body of knowledge concerning nuclear-mitochondrial genome interactions and will be of special interest to researchers in the field of evolutionary gene transfer from mitochondrial to nuclear genomes. The most important outcome of these results is the possibility for the identification and engineering of this novel CMS system in new crops. This will give plant breeders a new approach for integrating CMS into hybrid breeding schemes, and may lead to the development of hybrids in species where the potential of hybrid breeding is yet to be fully realised.

# Zusammenfassung

Das Verständnis der genetischen Architektur, die der cytoplasmatisch männlichen Sterilität (CMS) zugrunde liegt, ist von entscheidender Bedeutung, wenn dieses Merkmal in weiteren Kulturarten für die Produktion von Hochertrags-Hybridsorten genutzt werden soll. CMS wurde bereits bei einer Vielzahl landwirtschaftlicher Arten mit großem Erfolg eingesetzt, wobei der größte Ertragszuwachs bei Mais zu verzeichnen ist, dessen Anteil an Hybriden, gemessen an der weltweiten Maisproduktion 65% ausmacht. Da die weltweite Nachfrage nach Nahrungsmitteln zunimmt, werden neue Hybridsorten, insbesondere in Kulturarten, in denen das Potenzial der Hybridzüchtung noch nicht vollständig ausgeschöpft ist, eine Schlüsselrolle spielen, um sicherzustellen, dass die landwirtschaftliche Produktion weiterhin die globale Ernährungssicherheit aufrechterhalten kann.

Die CMS-Eigenschaft wird durch eine Unfähigkeit der betroffenen Pflanzen charakterisiert, lebensfähigen Pollen zu bilden. Dieser Phänotyp gibt Pflanzenzüchtern die Möglichkeit, CMS als Bestäubungskontrollmechanismus während der Produktion von Hybridsaatgut zu verwenden. Um das CMS-Merkmal innerhalb von Zuchtprogrammen zu verwenden, werden molekulare Marker benötigt, die es Züchtern erlauben, Pflanzenmaterial zu identifizieren, die die männliche Fruchtbarkeit wiederherstellen können. Obwohl die Ursache von CMS durch das Cytoplasma vererbt wird, sind Gene, die die Fertilität wiederherstellen können (*Rf*-Gene), nuklear codiert. Dadurch ist das CMS-Merkmal nicht nur für die Produktion von Hybridsaatgut von agronomischer Bedeutung, sondern auch von grundlegendem Interesse für die Untersuchung von Mitochondrien-Zellkern-Wechselwirkungen.

In dieser Dissertation wird die Untersuchung zur Ursache der CMS-Restauration bei Deutschem Weidelgras (*Lolium perenne* L.) dokumentiert. Obwohl die Bedeutung von Futtergräsern wie Weidelgras häufig vernachlässigt wird, sind sie in Europa für 80% der Milch und 70% der Fleischproduktion verantwortlich. Ein CMS-System konnte in Deutschem Weidelgras identifiziert werden. Da es jedoch noch nicht vollständig charakterisiert wurde, ist dessen Einsatz in der Produktion von Hybrid-Saatgut nur bedingt praktikabel. Das Ziel dieser Untersuchungen war es, die für die Fertilitätsrestauration verantwortlichen Gene oder Genom-Abschnitte zu identifizieren und damit die genetische Architektur der CMS-Fertilitätsrestauration bei Deutschem Weidelgras aufzuklären. Durch die Verwendung von vier verschiedenen Ansätzen, über die in den Kapiteln 2-5 berichtet wird, sowie deren Zusammenspiel, konnte die genetische Kontrolle dieses wichtigen Merkmals aufgezeigt werden.

Zu Beginn der Dissertation zeigte eine Recherche der relevanten Literatur, dass identifizierte *Rf*-Gene in fast allen Fällen zu einer Klasse von RNA-bindenden Proteinen gehören, die als Pentatricopeptid-Repeat (PPR) Proteine bezeichnet werden. Obwohl sie in Pflanzengenomen, mit Hunderten von identifizierten Mitgliedern, besonders zahlreich sind, konnte eine Unterklasse von PPRs identifiziert werden, die CMS wiederherstellt - die restorer of fertility-like PPRs (RFLs). In **Kapitel 2** wird eine Bioinformatik-Pipeline vorgestellt, die die schnelle Identifizierung dieser *RFL*-Gene in genomischen- oder transkriptomischen Sequenzdaten ermöglicht, wobei Sequenzähnlichkeiten zwischen *RFL*s ausgenutzt werden. Diese Pipeline ermöglicht nicht nur die gleichzeitige Identifizierung von *RFL*s mehrerer Spezies, sondern auch die Identifizierung von aktiven Abschnitten im Genom, die zur Neubildung von *RFL*s innerhalb eines Genoms verantwortlich sind. Die Analyse von 14 Pflanzenarten ergab, dass 50-90% der *RFL*-Gene innerhalb von zwei bis drei verschiedenen genomischen Loci pro Art zu finden sind.

Eine in CMS-Fertilitätsrestauration spaltende Population von über 1.200 Weidelgraspflanzen, wurde identifiziert und bildete die Grundlage für die hier vorgestellte Untersuchung zur Wiederherstellung der Fertilität. Diese Population zeigte keine intermediären Phänotypen und eine Wiederherstellungsrate der Fertilität von 27%, was darauf hindeutet, dass eine co-dominante Kontrolle mit zwei Loci für die Wiederherstellung der Fruchtbarkeit verantwortlich sein könnte. Drei sequenzierungsbasierte Ansätze wurden verwendet, um diese Population zu untersuchen, der erste davon, Genotyping by sequencing (GBS), wird in **Kapitel 3** vorgestellt. Die Anwendung von GBS in dieser Population identifizierte vier quantitative trait loci (QTL) im Zusammenhang mit der Wiederherstellung des CMS-Phänotyps mit einer Grösse von insgesamt 87.3 Mb.

Im Anschluss daran wurde eine Bulk Sergeant Analysis (BSA) an DNA-Pools aus sterilen oder fertilen Pflanzen durchgeführt. Da der untersuche Phänotyp entweder steril oder fruchtbar ist, erwies sich dies als ein effektiver Ansatz, um die zuvor identifizierte QTL weiter aufzulösen. Die Ergebnisse der BSA sind in **Kapitel 4** angegeben, in denen zwei Loci identifiziert wurden, die 74 Kb des Referenzgenoms abdecken. Interessanterweise konnte festgestellt werden, dass diese zwei Loci DNA-Sequenz mitochondrialer Herkunft enthielten, wovon bei einem nuklear codierten mitochondrialen Gen im sterilen Pool gezeigt werden konnte, dass es Mutationen aufweist. Dies legt die Vermutung nahe, dass funktionelle Kopien von defizienten CMS-kausalen mitochondrialen Genen für die Wiederherstellung der Fruchtbarkeit verantwortlich sein könnten.

Letztendlich wurde die Gesamt-RNA von vier in einem CMS-Züchtungsprogramm verwendeter Genotypen, einschliesslich drei verschiedener Gewebetypen in einer Genexpressionsanalyse untersucht. Die Ergebnisse, die in **Kapitel 5** vorgestellt werden, zeigten mehrere exprimierte mitochondriale Gene innerhalb der beiden in Kapitel 4 identifizierten QTL. Nach vollständiger Auswertung wurden drei mögliche Kandidaten-Gene für die Fertilitätswiederherstellung identifiziert, von denen zwei für Untereinheiten des mitochondrialen Atmungskomplexes IV codieren und die schon zuvor in Zusammenhang mit der Ursache von CMS gebracht werden konnten. Mutationen in der kodierenden Sequenz eines dieser Gene sind mit dem sterilen Phänotyp assoziiert, wobei beide Gene errechnete mitochondriale Transitpeptide enthalten. Zusätzlich zu diesen Ergebnissen konnte die Genexpressionsanalyse pleiotrope Effekte aufzeigen, die das Vorhandensein von *Rf*-Genen mit enormen Expressionsverschiebungen in restaurierten Hybriden assoziiert, was die Beobachtung von praktischen Pflanzenzüchtern untermauert, die restaurierte Hybriden im Gegensatz zu nicht-restaurierten Hybriden als "wüchsiger" beschreiben.

Die hier präsentierten Ergebnisse zeigen einen noch nie zuvor beschriebenen Mechanismus für die Wiederherstellung der CMS-Fertilität und identifizieren ideale Ansatzpunkte für die Markerentwicklung innerhalb der untersuchten Population. Darüber hinaus ergänzen sie das Wissen über nuklear-mitochondriale Genom-Interaktionen und sind von besonderem Interesse für Forscher auf dem Gebiet des evolutionären Gentransfers von mitochondrialen zu nuklearen Genomen. Die bedeutendste Erkenntnis dieser Dissertation ist die Möglichkeit zur Identifizierung und Entwicklung dieses neuartigen CMS-Systems in neuen Kulturen. Dies wird Pflanzenzüchtern eine neue Möglichkeit eröffnen, CMS in Hybrid-Züchtungsschemata zu integrieren und dadurch zur Entwicklung von Hybriden in Arten beitragen, bei denen das Potenzial der Hybridzüchtung noch nicht vollständig verwirklicht ist.

# Chapter 1

# General Introduction

### The Importance of Plant Breeding

Food security is rapidly becoming one of the greatest challenges of our time, driven by an ever-growing population that is predicted to increase by 34% by 2050 [1]. Along with this population increase, food demand is predicted to rise by 46% by 2050 [2]. Malnutrition, both general (starvation) and specific (micronutrient deficiencies), is already a problem and disproportionately affects people in developing or underdeveloped countries [3]. In order to meet these increasing demands, agriculture will need to continue to innovate and modernise both on the farm and in the lab.

Ever since the 'Green Revolution' of the 1950s-1980s, agriculture has managed to keep pace with food demand through the discovery and use of large scale irrigation systems, mineral fertilizers, pesticides, mechanization and monocultures [4] with global crop yields increasing by 56% between 1965 and 1985. Although food supply has been outpacing demand since the Green Revolution, yearly yield increases are now falling below yearly demand increases and innovations are required to maintain food security [5]. This problem is exacerbated by the decreasing amount of arable land available per person. This figure has halved in the last 50 years, due to both land loss (mainly through desertification and urbanisation) and population growth [6]. Compounding this even further are the challenges arising from climate change that also face agricultural production, such as increasing average temperatures and destructive weather events such as storms and droughts [7].

Although there are several approaches to alleviating food security concerns within the realm of politics, legal regulations and development assistance, this thesis is focussed on a solution offered by molecular plant breeding. Plant breeding has been employed to improve crops both qualitatively and quantitatively for over 12,000 years, beginning with the earliest wheat (*Triticum aestivium* L.) varieties all the way through to modern genome editing techniques [8]. Targeted or systematic breeding has only been performed for the last 200 years and 'scientific' breeding has been employed since the confirmation of Mendels' laws in 1900 [8]. All plant breeding consists of two main phases: the first, in which (genetic) variation is identified or created, and the second, in which desirable traits are selected for and fixed within genotypes. This basic process has made use of several emerging techniques and technologies over the years including artificial crossing, induced mutation, polyploidisation, cell fusion, tissue culture, inter-specific hybridistion [9], and more recently, transgenic technologies to introduce foreign genes, targeted mutation and gene editing.

One of the most effective techniques employed by plant breeding since the 1930s has been hybrid breeding. Hybrid breeding exploits the phenomenon of heterosis, whereby offspring from genetically distinct inbred parental lines show significantly increased vigour. This manifests most appreciably as an increase in yield for the first generation of hybrids and has been most effectively utilised in maize (*Zea mays* L.) breeding programs, where yield has increased from 1.8 to 7.8 t/ha in the United States through the cultivation of hybrid varieties. Hybrid varieties now make up 65% of globally produced maize, 60% of sunflower (*Helianthus annuus* L.), 48% of sorghum (*Sorghum bicolor* L.), and 12% of rice (*Oryza sative* L.), with the majority of radish (*Raphanus raphanistrum* L.), pepper (*Capsicum* spp.) and cabbage (*Brassica oleracea* L.) varieties also being hybrids [8]. There are also current efforts to introduce hybrids

into breeding programs for wheat [10], barley (*Hordeum vulgare* L.) [11] and other important crops. There are three main hurdles that must be overcome for a successful hybrid breeding scheme. Firstly, inbred lines need to be established which can be confounded by the self-incompatibility mechanisms seen in many crop species [12]. Secondly, inbred lines with good general combining ability need to be identified followed by parental crosses with good specific combining ability. Finally, a pollination control mechanism to allow large-scale hybrid seed production is required such as emasculation, cytoplasmic or genic male sterility, self-incompatibility or gametocides [13]. Exactly what drives heterosis towards increased yields is still to be elucidated, although mechanisms that include epistatic interactions, dominance or overdominance and epigenetics have been hypothesised [14].

**Cytoplasmic Male Sterility**

Cytoplasmic male sterility (CMS) was first described in the 1920s [15] and is one of the most widely applied pollination control mechanisms utilised in hybrid breeding schemes. CMS has been characterised in over 140 self-pollinating and cross-pollinating plant species [16] and is an example of extra-nuclear heredity, where the male sterilising factor is inherited from the maternal mitochondria. In nature, CMS is observed in gynodioecious species where hermaphrodite and male sterile (female) individuals are found within the same population [17]. Why CMS is found in nature is puzzling, as it confers on an individual a seeming fitness disadvantage, allowing propagation of genes through only the maternal organs. Some hypothesise that this could be evolutionarily maintained as it allows natural populations to obtain some of the advantages conferred by heterosis, or that it allows dense populations to become more energy and nitrogen efficient by reducing the overall pollen production [18]. Another, more plausible, explanation lies in the origins of the mitochondrial genome, where the 'invading' mitochondrial genome and 'native' nuclear genome are still undergoing a genetic arms race [17]. This neatly explains why the CMS source is mitochondrial, as the mitochondrial genome is maternally inherited and thus garners no evolutionary advantage from pollen production. This theory is further bolstered by the fact that genes that restore pollen production, and thus male fertility, are nuclear in origin [19].

The mitochondrial origin of CMS is almost always a recombination event within the mitochondrial genome that creates a novel open reading frame (ORF), containing a portion of a functional mitochondrial gene [20]. The exact mechanism whereby this ORF causes CMS is yet to be elucidated, with the leading theory postulating that the translated product of the ORF may interfere with the respiratory electron transport chain making it 'leaky'. This leads to the production of reactive oxygen species and a decrease in the proton potential across the inner mitochondrial membrane, thus lowering energy output to below what is required for pollen formation as well as triggering programmed cell death [21]. Plant breeders have also been able to generate CMS systems through the use of mutagenesis treatments, such as ethyl methanesulfonate (EMS) [22], by disrupting mitochondrial/nuclear communication through the use of forced wide crosses [23], and through the use of other techniques such as protoplast fusion [24].

The CMS phenotype can be rescued through the action of nuclear genes called restorer of fertility (*Rf*) genes. These genes are usually from the large pentatricopeptide repeat (PPR) gene family that code for RNA-binding proteins [24]. PPRs are very prevalent in higher plant genomes with around 450 *PPR* genes found per genome [24] as opposed to the six *PPR*s in the human genome. This expansion in plants is thought to have occurred after the separation of the land plant lineage from green algae, as the green algae *Chlamydomonas reinhardtii* has only 12 *PPR* genes [25], whereas the moss *Physcomitrella patens* has 103 [26]. Due to the lack of

introns in *PPR* genes, this expansion is thought to have been driven by a retrotransposon-like process [27]. Although PPR proteins are encoded by the nuclear genome, they most often function within organelles to mediate gene expression, facilitating the processing and translation of RNAs as well as being involved in transcript editing [28]. The majority of transcript editing events, mediated by PPRs, involve C-to-U transitions [29] in both the chloroplasts and mitochondria. This pattern is consistent with the theory that *PPR* expansion in land plants was in response to the increased amounts of UV light penetrating an ozone-poor early atmosphere when land plants first evolved around 700 million years ago. As UV light causes C-to-T mutations in DNA [30], PPR-mediated RNA editing can rescue these mutations before the transcript is translated. This expansion appears to have occurred quickly during land plant evolution as PPRs are very well conserved between mono- and dicots, suggesting that the period of expansion was relatively short [31].

*Rf* genes from within the *PPR* gene family come from a subset of *PPRs*, called the restorer of fertility-like PPRs (*RFLs*) [32]. The *RFLs* can be identified from within the *PPR* gene family by their relative homology and similarities to *RFLs* in other species [33]. The RFL proteins are also notable for not containing the domains associated with transcript editing and are – as with nearly all *PPR* genes – intron-less [29]. Perhaps the most unique feature of the RFLs from within the PPR protein family is that they are under a diversifying selection pressure [34]. This signifies that mutations that lead to amino acid changes are selected for, with this process being driven by homologous recombination events leading to clusters of *RFL*s within the genome [35]. The exact mechanism whereby RFLs restore male fertility to CMS-affected plants appears to be diverse with transcript degradation, cleavage, sequestration and translational blocking all being implicated [36]. Additionally, some environmentally sensitive genic male sterility restoration mechanisms have been identified with this being most effectively utilised in the rice hybrid breeding systems to create a two, rather than three, line CMS hybrid breeding scheme [37].

There are few exceptions to *PPR-Rf* genes as restorers of CMS, with restoration of fertility genes including other mechanisms. These mechanisms often act at a metabolic level with an aldehyde dehydrogenase [38], acyl-carrier protein [39] and peptidase [40] all being identified. The exception to non-RFL mediated restoration occurring through metabolic pathways is the glycine-rich proteins that work in concert with a PPR to achieve fertility restoration [41].

As mentioned above, CMS is widely applied in hybrid breeding programs as a pollination control mechanism [42]. To utilise CMS, breeders ensure that one of their parental lines is affected by CMS; this line is known as the mother-line as it is now functionally female. When crossing the mother-line to the (unaffected by CMS) father-line, the breeder can be sure that all seed collected from the mother-line is from a hybridisation between the two parental lines. Thus, the use of a pollination control mechanism in hybrid breeding is a very effective way to ensure purity of hybrid seed, especially when generating hybrid seed at a commercial level [42]. CMS and other non-manual pollination control mechanisms have allowed breeders to get away from very time- and money-consuming approaches such as emasculation, where the male flower organs are removed by hand. Incidentally, this is another reason that maize was one of the first crop plants in which hybrids were developed, as they are a monoecious species, where male and female flowers are physically separate on an individual plant [45]. This makes manual emasculation much easier as compared to gynodioecious species, where individual flowers contain both male and female organs.

One issue with CMS-assisted hybrid breeding schemes is the need to maintain the CMS phenotype in one of the parental lines, as this line is no longer creating pollen and thus cannot

propagate itself. To overcome this, breeders retain a CMS-free population of the mother-line, called the maintainer-line, that when crossed to the CMS-affected population can maintain that CMS-affected population through to the next generation. This is called the 'three line' system with mother-maintainer, mother-CMS and father lines. The other aspect of this system is the need for the father-line to be carrying an *Rf* gene. This is to ensure that the F1 hybrids have a restored phenotype and are able to produce pollen in the field, and is especially important for grain crops such as rice and maize. There are some indications that restored F1 hybrids might also have an intrinsic yield advantage over unrestored hybrids [44]. To ensure that the father-line does carry the *Rf* gene, breeders often utilise molecular markers that denote the restoring genotype [45]. This is important, as unlike the CMS-affected plants, the presence of a restorer gene shows no phenotype in the absence of CMS. Molecular markers for *Rf* genes are also necessary to ensure that maintainer populations for the mother lines are *Rf*-free; this ensures no unwanted restoration within the CMS-affected mother line that would lead to impurities in the produced hybrid seed.

Research into the molecular underpinning of CMS-*Rf* systems has important practical applications in hybrid breeding, but is also an important topic for basic research. This is due to the fact that CMS-*Rf* systems are models for eukaryotic nuclear-mitochondrial communication that are relatively easy to study, since the breakdown of communication that can cause CMS has an easily observable phenotype. Greater understanding of how the nuclear and mitochondrial genomes communicate allows researchers to better understand a range of topics from ageing and cell death to transcription coordination [46].

**Forage Grasses**

The importance of forage grasses is often over-looked as they are not directly consumed by the human population but rather by our livestock. However, grasslands are vital agro-ecosystems that cover 39% of agricultural land in Europe [47] accounting for 80% of milk production and 70% of meat production [48]. Perennial ryegrass (*Lolium perenne* L.) is a major component of temperate grassland systems with an annual seed production output of 90 thousand metric tonnes. It accounts for almost 50% of total grass production (forage and turf grasses), making it the most important grass species both in Europe and worldwide [49].

In forage grasses, biomass is the primary yield target. Despite intensive breeding efforts over the last decades, increases in biomass yield are below that of other major crop species [50]. Currently there are attempts to introduce hybrid breeding in several forage crops including perennial ryegrass. Introducing hybrid breeding schemes in perennial ryegrass presents several problems. The first of which is that perennial ryegrass is self-incompatible (SI), making the creation of inbred-lines difficult, although recent studies focussing on SI [51], self-fertility [52] and doubled haploids [53] all hold promise to overcome this. Secondly, although there is a CMS system described in perennial ryegrass [54], it has yet to be fully elucidated, allowing it to be easily utilised in hybrid breeding schemes. The CMS causing cytoplasm was induced using EMS treatment and the CMS causing mitochondrial genome has been sequenced [55]. What is currently lacking from the understanding of this CMS system is the identification and characterisation of the restoration mechanism to both provide breeders with molecular markers to track *Rf* genes through their populations and to further our general understanding of CMS-*Rf* interactions.

## Controversies

Although CMS is a plant trait utilised in hybrid breeding, it is often mislabelled as a 'new plant breeding technique' (NPBT) in itself. The catalogue of NPBTs includes relatively new innovations such as cisgenesis, intragenesis, targeted mutagenesis, transient introduction of recombinant DNA, RNAi induced gene silencing and genome editing. Some older techniques such as protoplast fusion, grafting to GMO rootstocks and – erroneously – CMS are also often included in this group [56], although rarely by plant scientists. It is notable that GMOs, at least transgenic GMOs, are not considered NPBTs [57]. This has led some people and organisations, often linked to organic production ideology, to call for a ban on the use of CMS-affected plants in the generation of hybrid seed as they believe them to be 'engineered'. Most significantly this has occurred in Switzerland, where the former President of the National Council of Switzerland presented an inquiry to the Swiss Parliament entitled 'CMS hybrids and other potentially problematic plant breeding techniques' requesting that the Swiss Confederation consider CMS as a 'genetic engineering' technique and thus ban it [58]. Somewhat ironically, CMS hybrids are the only source of commercially viable production in several vegetable species and have been used extensively in both conventional and organic farming since the 1950s [59]. It is possible that CMS is grouped with these other more modern techniques due to public perception driven by its seemingly ominous name – a prime example of this being seen in the television show 'Helix', in which the villain is attempting to sterilise the human race using 'cytoplasmic male sterile apples' [60]. Currently the European Union (EU) is considering policy options for the use of NPBT within the EU, although this is focussing more on genome editing than other techniques [61] and CMS is likely to be ignored at the policy level considering its agronomical and economical importance as well as its long history of use.

## Thesis Outline

This thesis contributes to the body of knowledge that confirms CMS as a naturally occurring trait, further supporting its continued use in plant breeding. The body of the thesis is divided into six chapters. The first is this general introduction. In the second chapter, a bioinformatics pipeline for the rapid identification of *RFL* genes from genomic or transcriptomic sequencing data is described. Several previous studies have identified the unique nature of RFLs from within the PPR family, but none have utilised this uniqueness for rapid identification. A bioinformatics pipeline utilising an orthologous clustering approach using protein sequences for the rapid identification of RFLs was developed. This pipeline was tested on a genomic draft of the perennial ryegrass genome, identifying 373 PPRs and 25 RFLs. The orthologous clustering method was also able to identify known *PPR-Rf* genes in other species and revealed that on average, 50-90% of candidate *RFL* genes are found in two or three genomic clusters within 0.01-0.1% of the genome.

In chapter three, a genotyping by sequencing (GBS) approach to identify quality trait loci (QTL) in a population of perennial ryegrass segregating for fertility restoration is described. This segregating population was provided by the breeding company Norddeutsche Pflanzenzucht Hans-Georg Lembke KG (NPZ) and consisted of several sub-populations totalling over 1,200 individual plants. Over 3,000 markers were generated across the perennial ryegrass genome by GBS, with 44 of these showing a significant association to the restoration phenotype. This enabled the identification of four QTL and has provided valuable marker data for further investigation and possible use in ongoing hybrid plant breeding efforts.

Chapter four presents a bulk segregate analysis (BSA) of sterile vs. fertile samples from the above chapter. In order to further define the QTL identified, all sterile and fertile samples were

separately pooled and shotgun sequenced. The generated sequence data was aligned to the high quality genome sequence of the Italian ryegrass (*Lolium multiflorum* Lam.) variety 'Rabiosa' (Molecular Plant Breeding, ETH Zurich, Switzerland, unpublished), with subsequent allele frequency analysis revealing two QTL that matched with the QTL identified in chapter three. This analysis further defined these QTL as being of mitochondrial origin with a strong candidate gene identified at one locus. These QTL represent the best opportunity to develop molecular markers for *Rf* in perennial ryegrass.

Chapter five consists of a gene expression investigation into the transcriptional changes observed between four different genotypes (maintainer, sterile, restored and restorer) and three different tissue types (leaf, flower and anther) of perennial ryegrass. This analysis revealed three genes within the previously identified QTL that are very strong candidates for fertility restoration as they are both expressed in restored anther tissue and contain mutations that affect the coding sequence. This analysis also shows wholesale changes in gene expression in restored leaf tissue, raising questions about a possible fitness advantage of restored hybrids.

Finally, in chapter six, the outcomes of this thesis are contextualised, their practical uses in plant breeding explored and further experimental designs to complement the results presented here are explored.

## References


1. Fao, U., 2009. How to Feed the World in 2050. In *Rome: High-Level Expert Forum*.
2. Gouel, C. and Guimbard, H., 2017. Nutrition transition and the structure of global food demand. (April 7). IFPRI Discussion Paper 1631. Available at SSRN: https://ssrn.com/abstract=2950524
3. Pinstrup-Andersen, P., and Cohen, M., 2000. Modern biotechnology for food and agriculture: Risks and opportunities for the poor. Agricultural biotechnology and the poor. Consultative Group on International Agricultural Research, Washington DC, USA, pp.159-172.
4. Myers, N., 1999. The next green revolution: its environmental underpinnings. Current Science, 76(4), pp.507-513.
5. FAO. 2002. World Agriculture: towards 2015/2030. Food and Agriculture Organization of the United Nations, Rome.
6. Miller, J.K., Herman, E.M., Jahn, M. and Bradford, K.J., 2010. Strategic research, education and policy goals for seed science and crop improvement. Plant science, 179(6), pp.645-652.
7. Nelson, G.C. and Shively, G.E., 2014. Modelling climate change and agriculture: an introduction to the special issue. *Agricultural Economics*, *45*(1), pp.1-2.
8. Lee, J., Chin, J.H., Ahn, S.N. and Koh, H.J., 2015. Brief History and Perspectives on Plant Breeding. In Current Technologies in Plant Molecular Breeding (pp. 1-14). Springer Netherlands.
9. Lusser, M., Parisi, C., Plan, D. and Rodríguez-Cerezo, E., 2011. New plant breeding techniques. Stateof-the-art and prospects for commercial development. JRC Scientific and Technical Reports/EUR 24760 EN.
10. Whitford, R., Fleury, D., Reif, J.C., Garcia, M., Okada, T., Korzun, V. and Langridge, P., 2013. Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. Journal of experimental botany, 64(18), pp.5411-5428.
11. Mühleisen, J., Maurer, H.P., Stiewe, G., Bury, P. and Reif, J.C., 2013. Hybrid breeding in barley. *Crop Science*, *53*(3), pp.819-824.
12. Takayama, S. and Isogai, A., 2005. Self-incompatibility in plants. *Annu. Rev. Plant Biol.*, *56*, pp.467-489.
13. Fu, D., Xiao, M., Hayward, A., Fu, Y., Liu, G., Jiang, G. and Zhang, H., 2014. Utilization of crop heterosis: a review. Euphytica, 197(2), pp.161-173.
14. Chen, Z.J., 2013. Genomic and epigenetic insights into the molecular bases of heterosis. Nature reviews. Genetics, 14(7), p.471.
15. Bateson, W. and Gairdner, A.E., 1921. Male-sterility in flax, subject to two types of segregation. *Journal of Genetics*, *11*(3), pp.269-275.
16. Laser, K.D. and Lersten, N.R., 1972. Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms. The Botanical Review, 38(3), pp.425-454.
17. Touzet, P. and Budar, F., 2004. Unveiling the molecular arms race between two conflicting genomes in cytoplasmic male sterility?. Trends in plant science, 9(12), pp.568-570.
18. Frank, S.A., 1989. The evolutionary dynamics of cytoplasmic male sterility. *The American Naturalist*, *133*(3), pp.345-376.
19. Hanson, M.R. and Bentolila, S., 2004. Interactions of mitochondrial and nuclear genes that affect male gametophyte development. *The Plant Cell*, *16*(suppl 1), pp.S154-S169.
20. Schnable, P.S. and Wise, R.P., 1998. The molecular basis of cytoplasmic male sterility and fertility restoration. Trends in plant science, 3(5), pp.175-180.Balk, J. and Leaver, C.J., 2001. The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release. *The Plant Cell*, *13*(8), pp.1803-1818.
21. Sasakuma, T., Maan, S.S. and Williams, N.D., 1978. EMS-induced male-sterile mutants in euplasmic and alloplasmic common wheat. *Crop Science*, *18*(5), pp.850-853.
22. Malik, M., Vyas, P., Rangaswamy, N.S. and Shivanna, K.R., 1999. Development of two new cytoplasmic male-sterile lines in Brassica juncea through wide hybridization. *Plant Breeding*, *118*(1), pp.75-78.
23. Zelcer, A., Aviv, D. and Galun, E., 1978. Interspecific transfer of cytoplasmic male sterility by fusion between protoplasts of normal Nicotiana sylvestris and X-ray irradiated protoplasts of male-sterile N. tabacum. *Zeitschrift für Pflanzenphysiologie*, *90*(5), pp.397-407.

24. Barkan, A. and Small, I., 2014. Pentatricopeptide repeat proteins in plants. *Annual review of plant biology*, *65*, pp.415-442.
25. Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L. and Marshall, W.F., 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science, 318(5848), pp.245-250.
26. Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y. and Tanahashi, T., 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science, 319(5859), pp.64-69.
27. Fujii, S. and Small, I., 2011. The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytologist*, *191*(1), pp.37-47.
28. Small, I.D., Rackham, O. and Filipovska, A., 2013. Organelle transcriptomes: products of a deconstructed genome. Current opinion in microbiology, 16(5), pp.652-658.
29. Jobson, R.W. and Qiu, Y.L., 2008. Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift?. *Biology Direct*, *3*(1), p.43.
30. Harris, R.S., 2013. Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications. *Genome medicine*, *5*(9), p.87.
31. O'toole, N., Hattori, M., Andres, C., Iida, K., Lurin, C., Schmitz-Linneweber, C., Sugita, M. and Small, I., 2008. On the expansion of the pentatricopeptide repeat gene family in plants. *Molecular Biology and Evolution*, *25*(6), pp.1120-1128.
32. Andrés, C., Lurin, C. and Small, I.D., 2007. The multifarious roles of PPR proteins in plant mitochondrial gene expression. Physiologia plantarum, 129(1), pp.14-22.
33. Sykes, T., Yates, S., Nagy, I., Asp, T., Small, I. and Studer, B., 2017. In silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (Lolium perenne L.). *Genome biology and evolution*, *9*(2), pp.351-362.
34. Geddy, R. and Brown, G.G., 2007. Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. *BMC genomics*, *8*(1), p.130.
35. Melonek, J., Stone, J.D. and Small, I., 2016. Evolutionary plasticity of restorer-of-fertility-like proteins in rice. *Scientific reports*, *6*, p.35152.
36. Chen, L. and Liu, Y.G., 2014. Male sterility and fertility restoration in crops. *Annual review of plant biology*, *65*, pp.579-606.
37. Li, Q., Zhang, D., Chen, M., Liang, W., Wei, J., Qi, Y. and Yuan, Z., 2016. Development of japonica photo-sensitive genic male sterile rice lines by editing carbon starved anther using CRISPR/Cas9. *Journal of Genetics and Genomics*, *43*(6), pp.415-419.
38. Liu, F., Cui, X., Horner, H.T., Weiner, H. and Schnable, P.S., 2001. Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. *The Plant Cell*, *13*(5), pp.1063-1078.
39. Fujii, S. and Toriyama, K., 2009. Suppressed expression of RETROGRADE-REGULATED MALE STERILITY restores pollen fertility in cytoplasmic male sterile rice plants. *Proceedings of the National Academy of Sciences*, *106*(23), pp.9513-9518.
40. Kitazaki, K., Arakawa, T., Matsunaga, M., Yui-Kurino, R., Matsuhira, H., Mikami, T. and Kubo, T., 2015. Post-translational mechanisms are associated with fertility restoration of cytoplasmic male sterility in sugar beet (Beta vulgaris). *The Plant Journal*, *83*(2), pp.290-299.
41. Hu, J., Wang, K., Huang, W., Liu, G., Gao, Y., Wang, J., Huang, Q., Ji, Y., Qin, X., Wan, L. and Zhu, R., 2012. The rice pentatricopeptide repeat protein RF5 restores fertility in Hong-Lian cytoplasmic male-sterile lines via a complex with the glycine-rich protein GRP162. *The Plant Cell*, *24*(1), pp.109-122.
42. Bohra, A., Jha, U.C., Adhimoolam, P., Bisht, D. and Singh, N.P., 2016. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. *Plant cell reports*, *35*(5), pp.967-993.
43. Irish, E.E. and Nelson, T., 1989. Sex determination in monoecious and dioecious plants. *The plant cell*, *1*(8), p.737.
44. Weingartner, U., Kaeser, O., Long, M. and Stamp, P., 2002. Combining cytoplasmic male sterility and xenia increases grain yield of maize hybrids. *Crop Science*, *42*(6), pp.1848-1856.

45. Akagi, H., Yokozeki, Y., Inagaki, A., Nakamura, A. and Fujimura, T., 1996. A codominant DNA marker closely linked to the rice nuclear restorer gene, Rf-1, identified with inter-SSR fingerprinting. *Genome*, *39*(6), pp.1205-1209.
46. Kotiadis, V.N., Duchen, M.R. and Osellame, L.D., 2014. Mitochondrial quality control and communications with the nucleus are important in maintaining mitochondrial function and cell health. *Biochimica Et Biophysica Acta (BBA)-General Subjects*, *1840*(4), pp.1254-1265.
47. Huyghe C., De Vliegher A., van Gils B and Peeters A. (2014) Grassland and herbivore production in Europe and effect of common policies; Ed Quae, 287 pp.
48. Wilkins PW, Humphreys MO. Progress in breeding perennial forage grasses for temperate agriculture. J Agric Sci 2003, 140:129–150.
49. Breeding Targets for Ryegrass in Europe. Europeanseed. Vol. 3 Issue 2 (2016).
50. van der Heijden SAG, Roulund N. Genetic gain in agronomic value of forage crops and turf: a review. Sustainable use of genetic diversity in forage and turf breeding (2010), pp. 247–260.
51. Do Canto, J., Studer, B. and Lubberstedt, T., 2016. Overcoming self-incompatibility in grasses: a pathway to hybrid breeding. *Theoretical and applied genetics*, *129*(10), pp.1815-1829.
52. Aguirre, A.A., Studer, B., Do Canto, J., Frei, U. and Lübberstedt, T., 2013. Mapping a New Source of Self-fertility in Perennial Ryegrass (Lolium perenne L.). *Plant Breeding and Biotechnology*, *1*(4), pp.385-395.
53. Begheyn, R.F., Vangsgaard, K., Roulund, N. and Studer, B., 2016. Efficient doubled haploid production in perennial ryegrass (Lolium perenne L.). In *Breeding in a World of Scarcity* (pp. 151-155). Springer International Publishing.
54. Nitzsche, W., 1971. Cytoplasmic male sterility in ryegrass (Lolium Spp). *Zeitschrift fur Pflanzenzuchtung*, *65*(3), p.206.
55. Nagy, I., Islam, M.S., Møller, I.M., Byrne, S. and Asp, T., 2014. Sequencing and analysis of mitochondrial genomes of fertile and male-sterile lines in perennial ryegrass (Lolium perenne L.). In *Plant Biotech Denmark Annual Meeting 2014*.
56. Schaart, J.G., van de Wiel, C.C., Lotz, L.A. and Smulders, M.J., 2016. Opportunities for products of new plant breeding techniques. *Trends in plant science*, *21*(5), pp.438-449.
57. Madre, Y., and Agostino, V.D., 2017. New plant-breeding techniques: What are we talking about? *Farm Europe,* Growth, 18 May. (http://www.farm-europe.eu/travaux/new-plant-breeding-techniques-what-are-we-talking-about/)
58. INTERPELLATION 14.3935, 2014. Hybrides CMS et autres techniques de sélection végétale potentiellement problématiques. *L'Assemblée fédérale — Le Parlement suisse.* (https://www.parlament.ch/fr/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20143935)
59. Dhall, R.K., 2010. Status of male sterility in vegetables for hybrid development. A Review. *Advances in Horticultural Science*, pp.263-279.
60. “Cross Polination”. *Helix.* Syfy, Comcast, Philadelphia, Pennsylvania.
61. European Academies Science Advisory Council. “Genome editing: scientific opportunities, public interests and policy options in the European Union.” 2017. (http://www.easac.eu/home/press-releases/detail-view/article/new-easac-re.html)

# Chapter 2

# *In silico* identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.)

**Timothy Sykes[1], Steven Yates[1], Istvan Nagy[2], Torben Asp[2], Ian Small[3] and Bruno Studer[1]***

[1] Institute of Agricultural Sciences, Forage Crop Genetics, ETH Zurich, 8092 Zurich, Switzerland.

[2] Department of Molecular Biology and Genetics, Research Centre Flakkebjerg, Aarhus University, Forsøgsvej 1, 4200 Slagelse, Denmark.

[3] Plant Energy Biology, ARC Centre of Excellence, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia.

* Author for Correspondence: Bruno Studer, Institute of Agricultural Sciences, Forage Crop Genetics, ETH Zurich, 8092 Zurich, Switzerland, +41446320157, bruno.studer@usys.ethz.ch

## Abstract

Perennial ryegrass (*Lolium perenne* L.) is widely used for forage production in both permanent and temporary grassland systems. To increase yields in perennial ryegrass, recent breeding efforts have been focused on strategies to more efficiently exploit heterosis by hybrid breeding. Cytoplasmic male sterility (CMS) is a widely applied mechanism to control pollination for commercial hybrid seed production and although CMS systems have been identified in perennial ryegrass, they are yet to be fully characterised.

Here, we present a bioinformatics pipeline for efficient identification of candidate restorer of fertility (*Rf*) genes for CMS. From a high-quality draft of the perennial ryegrass genome, 373 *pentatricopeptide repeat (PPR)* genes were identified and classified, further identifying 25 *restorer of fertility-like PPR* (*RFL*) genes through a combination of DNA sequence clustering and comparison to known *Rf* genes. This extensive gene family was targeted as the majority of *Rf* genes in higher plants are *RFL* genes. These *RFL* genes were further investigated by phylogenetic analyses, identifying three groups of perennial ryegrass *RFL*s. These three groups likely represent genomic regions of active *RFL* generation and identify the probable location of perennial ryegrass *PPR-Rf* genes.

This pipeline allows for the identification of candidate *PPR-Rf* genes from genomic sequence data and can be used in any plant species. Functional markers for *PPR-Rf* genes will facilitate map-based cloning of *Rf* genes and enable the use of CMS as an efficient tool to control pollination for hybrid crop production.

**Key words:** Cytoplasmic male sterility (CMS), Hybrid breeding, Pentatricopeptide repeat (PPR) proteins, Perennial ryegrass (*Lolium perenne* L.), Restoration of fertility, Restorer of fertility-like PPR (RFL)

## Introduction

The agronomical value of perennial ryegrass (*Lolium perenne* L.) comes from its ability to produce high forage yield of good feed quality in both permanent and temporary grassland systems [1]. Due to the increasing global demand for animal products, improved varieties of forage grasses are becoming an important aspect of global food security. Thus, perennial ryegrass has been the subject of intensive breeding efforts over recent decades. However, these breeding efforts are mainly focused on the improvement of population and synthetic varieties and show limited increases in biomass yield [2, 3], which is one of the most important traits in forage grasses.

Hybrid breeding, by efficiently exploiting the phenomenon of heterosis, has been successfully used in breeding programs to increase yield in several important crop species including rice (*Oryza sativa* L.), maize (*Zea mays* L.) and rapeseed (*Brassica napus* L.) [4, 5]. Due to its significant impact, there are currently considerable efforts to establish hybrid breeding schemes for other crops including wheat (*Triticum aestivium* L.) [6]. The development and application of hybrid breeding in forage crops has the potential to result in similar yield increases [2]. To employ hybrid breeding in perennial ryegrass, one of the major challenges is the absence of a pollination control strategy that would allow the efficient production of hybrid seed on a commercial level. In several plant species including maize, onion (*Allium cepa* L.), sorghum (*Sorghum bicolor* L.), sugar beet (*Beta vulgaris* L.), sunflower (*Helianthus annuus* L.), rapeseed, common beans (*Phaseolus vulgaris* L.) and rice, cytoplasmic male sterility (CMS) has been successfully applied to control pollination for hybrid seed production [7-14]. Although CMS systems have been identified in perennial ryegrass [15-17], they are yet to be fully characterised [18-21].

CMS in flowering plants is characterised by a maternally inherited inability to produce functional pollen [22]. This functional defect is often attributed to aberrant transcripts originating from the mitochondrial genome, with these CMS causing transcripts usually coding for novel chimeric open reading frames (ORFs) containing part of a functional mitochondrial gene [23, 22]. The translated products of these chimeric transcripts disrupt normal mitochondrial function such that the energy requirements for pollen formation cannot be met, rendering the pollen unviable [12].

The CMS phenotype is often restored through the action of nuclear-derived RNA binding proteins that are generally members of the large family of pentatricopeptide repeat (PPR) proteins [24]. Exceptions are the CMS-T restoration in maize [25], the *Rf* gene *bvORF20* in sugar beet [26] as well as other RNA-binding proteins that have been implicated in fertility restoration [27, 28]. PPR proteins are particularly numerous in land plants, with 450 PPRs identified in Arabidopsis (*Arabidopsis thaliana* L.) and 477 in rice [29-33]. Although PPR proteins are encoded by the nuclear genome, they most often function within organelles to mediate gene expression, facilitating the processing and translation of RNAs [34]. PPR proteins contain tandem arrays of a degenerate 35 amino acid motif that bind to RNA in a sequence-specific manner [32]. PPR proteins appear to be functional only in organelles and as such have been described as the chaperones of organelle gene expression [35]. PPR proteins have previously been divided into subclasses based on PPR motif variations and a series of conserved C-terminal domains [36, 37]. The two main subclasses of PPR proteins, the P and PLS subclasses, are defined by the organisation of the individual PPR motifs within a PPR gene. The P-type PPRs are comprised almost entirely from the canonical 35-amino-acid P motif. In contrast, the PLS subclass of PPRs is comprised of triplet repeats containing one P motif, one L motif ('long', usually 36 amino acids) and one S ('short', usually 31 amino acids). This PLS

subclass is also characterised by three distinctive C-terminal motifs: E (extended), E+ (slightly longer version of the E-domain) and DYW (named for terminating with a conserved Asp-Tyr-Trp triplet). All PPRs that have been shown to be involved in RNA editing, in both mitochondria and chloroplasts, are members of these three sub-groups [38]. The E/E+ domains are believed to provide an essential recognition site for an (as yet unidentified) editing complex. The DYW domain, which usually includes an E domain, shows similarity to deaminases and is possibly directly involved in RNA editing [39-43].

A subgroup of the P-type PPRs is specifically linked to fertility restoration of CMS: the restorer of fertility-like PPR (RFL) proteins. This group is identified by their relative homology from within the PPR family, their identity with other known CMS restorer PPRs from related plant species and their tendency to be present in several homologous copies clustered within the genome. These RFLs comprise around 10 to 30 members per plant genome from the full set of PPRs [44, 45]. It has been shown previously that *RFL* genes appear to be under different selection pressures when compared to the rest of the *PPR* gene family members. Within the *RFL* subgroup, high ratio of non-synonymous versus synonymous nucleotide substitutions indicates diversifying selection [45, 46]. This suggests, in conjunction with gene duplication, that the generation of new *RFL* genes and subsequent loss of non-functional *RFL*s is relatively rapid, keeping pace with the generation of novel CMS sources. CMS is also used as a model system for studying nuclear/mitochondrial genome interactions, as its easy detection allows researchers to rapidly identify individuals with a breakdown in nuclear/mitochondrial signalling [47].

In order to provide plant breeders with a molecular tool for candidate *Rf* gene identification and thus facilitate the implementation of hybrid breeding schemes in perennial ryegrass, this study aimed to locate, *in silico*, regions of active *RFL* generation in the perennial ryegrass genome by i) the development and validation of a bioinformatics pipeline for the identification of *PPR*s and *RFL*s from genomic sequence, ii) utilizing this pipeline for identification of *PPR* genes within the perennial ryegrass genome, iii) classifying these *PPR* genes in order to isolate the *RFL*s as potential candidate *Rf* genes, iv) phylogenetically analyzing the *RFL* genes from several grass species to identify groups of rapidly diverging *RFL* genes within the perennial ryegrass genome, and v) using this analysis to locate genomic regions of novel *RFL* generation.

## Results

### Pentatricopeptide repeat (PPR) and restorer of fertility-like PPR gene identification and classification

A draft of the perennial ryegrass genome sequence [48] was scanned to identify *PPR* genes using a Hidden Markov Model (HMM) profile matrix [49]. From a total of 71,009 genes, obtained from *ab initio* and evidence-based gene predictions in the perennial ryegrass genome, 373 *PPR* genes were identified. These 373 *PPR* genes were classified into two subfamilies, P and PLS, based on the arrangements of the repeated PPR motifs. Each of these subfamilies contained roughly half of the identified *PPR* genes with the P subfamily being slightly larger with 207 members, representing 55% of the total *PPR*s. The PLS subfamily was further grouped based on the presence or absence of the C-terminal domains implicated in RNA editing. From a total of 166 PLS subfamily genes, 40 were missing RNA editing-specific C-terminal motifs (PLS subclass), while the remaining 126 were organized into the E class (72), the E+ class (23) and the DYW class (31) (Figure 1). Analysis of the 25 RFLs, identified by homology to known restorers from other grass species, revealed that they all belonged to the P subfamily. Further analysis identified five pseudogenes that were truncated and lacking start/stop codons. These identified *RFLs* have an average of 16 PPR domains as compared to 9.7 PPR domains for the remainder of the *PPR* genes.

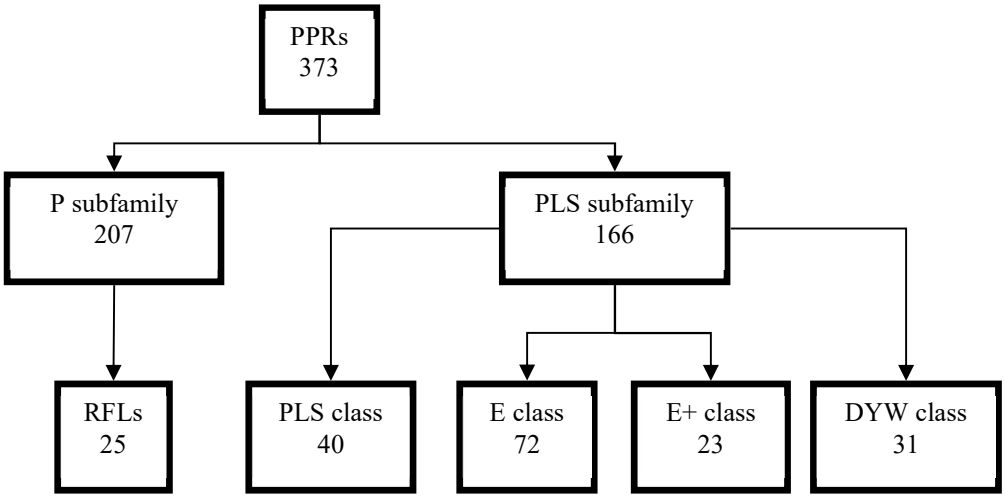



**Figure 1.** Classification of identified *pentatricopeptide repeat* (*PPR*) genes in perennial ryegrass (*Lolium perenne* L.). Sequences were classified into P and PLS subfamilies, based on the architecture of the repeated PPR motifs. The PLS subfamily was further classified by the presence of several non-PPR C-terminal domains. All identified *restorer of fertility-like PPR* (*RFL*) genes were part of the P subfamily.

### Restorer of fertility-like PPR gene comparison in multiple species

Orthologous clustering of protein sequences from 14 species was performed to ascertain whether the identified perennial ryegrass *RFL* genes are similar to *RFL* genes from other plant species. For this clustering, the canonical coding sequences (CDS) [50] of 14 species were used, comprising a total of 561,090 protein sequences. Of these, 554,468 passed the quality checking by OrthoMCL [51], of which 403,713 proteins were grouped into 44,672 clusters (Figure 2A). A subset of 5,054 clusters contained proteins from all species, representing 30.6% of the

403,713 clustered proteins. In contrast, 17.3% of the sequences were species-specific and contained in 39.7% of clusters (Figure 2A).

Further analysis identified 287 clusters that contain at least one of the 373 perennial ryegrass *PPR*s found previously. Plotting the number of species represented in these 287 clusters against the number of proteins present revealed a linear relationship with one clear outlier. This outlying cluster contained 154 proteins originating from 13 species and is more than three times bigger than the second largest cluster. This cluster was entirely comprised of PPR proteins and contained 16 of the previously identified 25 RFLs from perennial ryegrass. The nine RFLs not present in this cluster were found to be either pseudogenes or poorly annotated genes leading to them not clustering with the remainder of the RFLs. The following species were dropped: Italian ryegrass (*Lolium multiflorum* L.*)* and meadow fescue (*Festuca pratensis* L.*)* as their sequences originated from transcriptome sequencing and a low number of RFLs were identified; bamboo (*Phyllostachys heterocla* L.) and teff (*Eragrostis tef* L.*)* as, although their genomes have been sequenced into scaffolds, these were not organised into contiguous sequences and thus did not provide precise information about genome positions. No RFLs were identified from banana (*Musa acuminate* L.*).* This approach not only showed that *RFL* genes form a distinct orthologous group, but also validated the approach used for *RFL* identification within the perennial ryegrass genome.

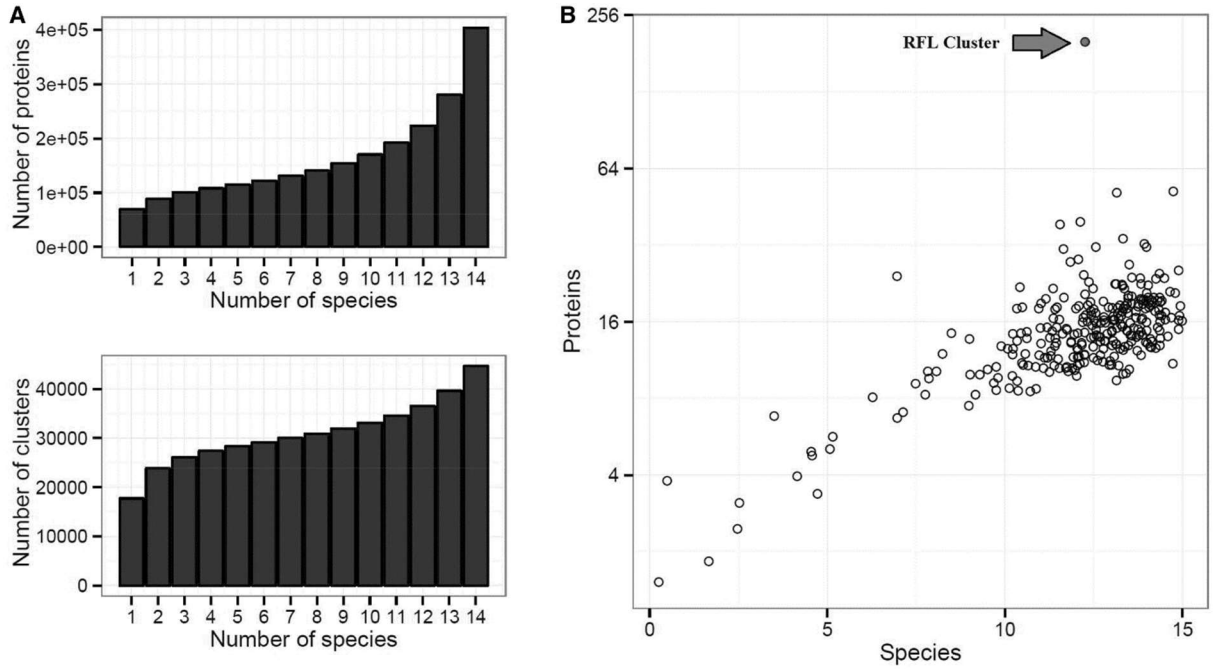



**Figure 2. A.** Histogram showing the number of proteins and clusters in relation to the number of species per cluster from the OrthoMCL protein sequence clustering of 14 species. **B**. Scatterplot showing the number of species (x-axis) and the number of proteins (y-axis, log2 scale) present in the 287 clusters containing at least one perennial ryegrass *PPR* gene. The outlying cluster containing 16 out of 25 identified RFLs is indicated with an arrow.

## Phylogenetic analysis of the RFL cluster

Having identified a set of *RFL* genes from multiple species (supplementary Table 1), a phylogenetic analysis was performed in order to understand the evolutionary ancestry underpinning the *RFL* genes. Protein sequences from the OrthoMCL-generated RFL cluster (indicated with an arrow in Figure 2B) were phylogenetically analysed, revealing four major clades of RFLs (Figure 3, Table 1). The only dicot included, Arabidopsis, was represented within a clade of its own (clade 3). The other three clades encompassed all the monocot sequences, with perennial ryegrass and Brachypodium (*Brachypodium distachyon* L.) being the only species represented in only one clade, and wild einkorn wheat (*Triticum urartu* L.) being the only species represented in all three monocot clades. All species, with the exceptions of wild einkorn wheat and foxtail millet (*Setaria italica* L.), had a majority of sequences present in only one clade.

**Table 1.** The number of *restorer of fertility-like PPR* (*RFL*) genes in each clade as well as the total number of *RFL*s identified are given for each species.

| Species | Number of sequences | | | | Totals |
|---|---|---|---|---|---|
| | Clade 1 | Clade 2 | Clade 3 | Clade 4 | |
| ***perennial ryegrass*** | - | 25 | - | - | 25 |
| ***wild einkorn wheat*** | 4 | 9 | - | 7 | 20 |
| ***barley*** | - | 6 | - | 1 | 7 |
| ***foxtail millet*** | 8 | - | - | 5 | 13 |
| ***rice*** | - | 2 | - | 10 | 12 |
| ***sorghum*** | 1 | - | - | 17 | 18 |
| ***maize*** | 7 | - | - | 2 | 9 |
| ***Brachypodium*** | - | 9 | - | - | 9 |
| ***Arabidopsis*** | - | - | 28 | - | 28 |
| **Totals** | 20 | 51 | 32 | 42 | 145 |

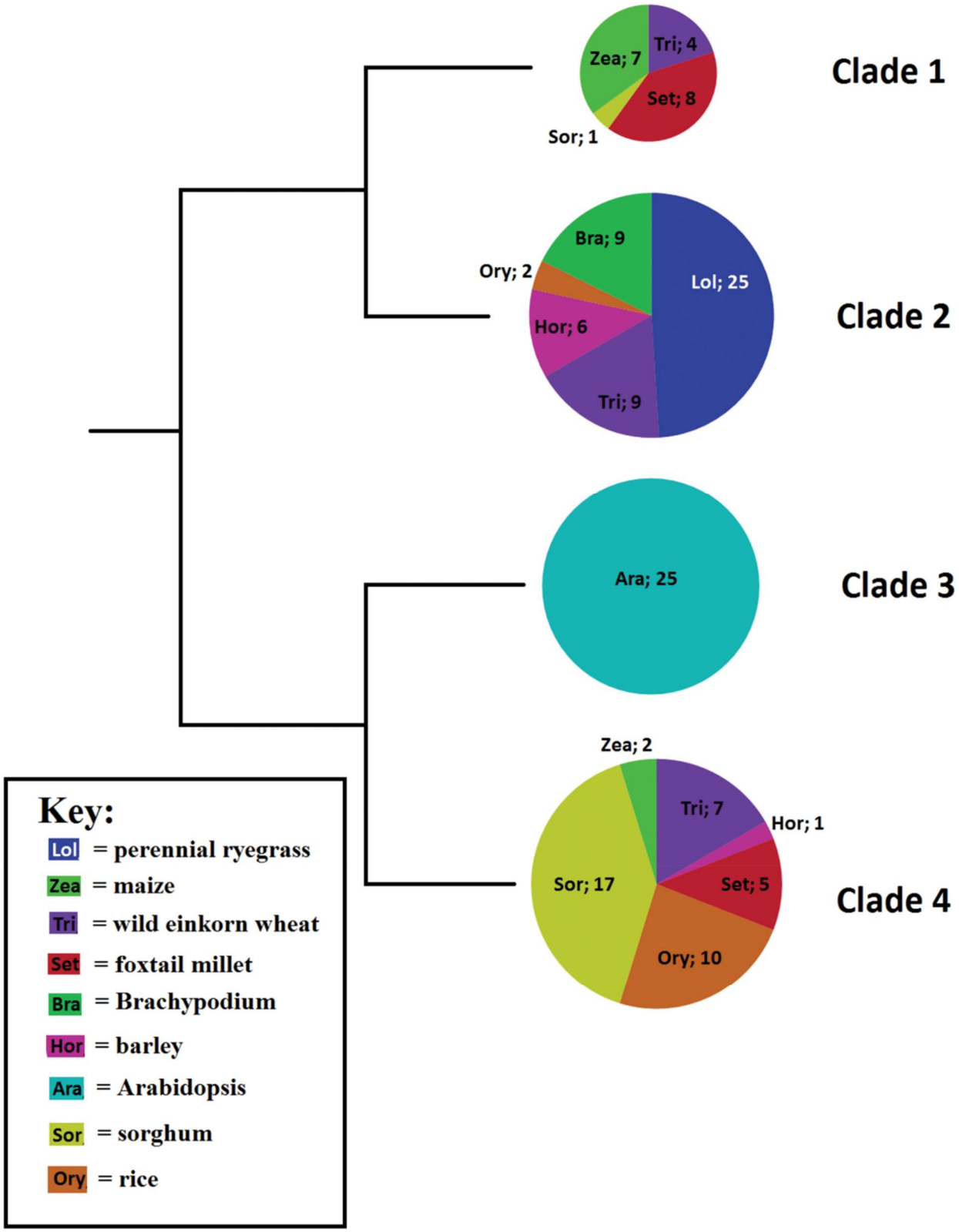



**Figure 3**. Phylogenetic tree showing the four identified clades within the RFL cluster containing a total of 154 proteins originating from 14 species. Colours are species-specific with abbreviated species names (see Table 1) shown along with the number of RFLs present.

To identify the *RFL* genes from each species that most recently evolved, detailed phylogenetic trees of each clade were coupled with genome location data available from Ensemble Plants (http://plants.ensembl.org/index.html). This revealed that within each clade, *RFLs* from the same species tend to cluster together, with the tightest clusters containing *RFLs* from the same genomic region of a single species (Figure 4, Table 2). Clades 1, 2, 3 and 4 had 65%, 23%, 75% and 52% of the *RFLs* represented in these species-specific clusters respectively. Considering only those species with whole genome sequence information available, 68% of their *RFL* genes were present in 13 clusters comprising 0.13% of their combined genomes. For example, in rice, 50% of the identified *RFL*s were found within 320kb of chromosome 10 (Table 2).

**Table 2.** Details of the species-specific genomic regions having a high density of *RFL* genes.

| | | Region of high *RFL* density | | |
|---|---|---|---|---|
| **Clade** | **Species** | **Genome location (bp)** | **Size (kb)** | **No. of genes present in cluster** |
| 1 | *Setaria italica* | Ch8:29882484-31204264 | 1322 | 5 |
| 1 | *Zea mays* | Ch2:227716868-228633247 | 917 | 6 |
| 1 | *Setaria italica* | Ch7:15683154-15692828 | 9 | 2 |
| 2 | *Oryza sativa* | Ch4:16684906-16757223 | 9 | 2 |
| 2 | *Hordeum vulgare* | Ch1:47176692-50263441 | 3087 | 3 |
| 2 | *Brachypodium distachyon* | Ch2:38479458-39012768 | 533 | 7 |
| 3 | *Arabidopsis thalians* | Ch1:4183066-4355929 | 172 | 4 |
| 3 | *Arabidopsis thalians* | Ch1:23176930-23988740 | 812 | 17 |
| 4 | *Sorghum bicolor* | Ch2:5169697-5744703 | 575 | 3 |
| 4 | *Zea mays* | Ch8:76606724-76690742 | 84 | 2 |
| 4 | *Sorghum bicolor* | Ch5:2222303-2776884 | 554 | 9 |
| 4 | *Oryza sativa* | Ch10:18823675-19143586 | 320 | 6 |
| 4 | *Oryza sativa* | Ch8:374091-383986 | 10 | 2 |

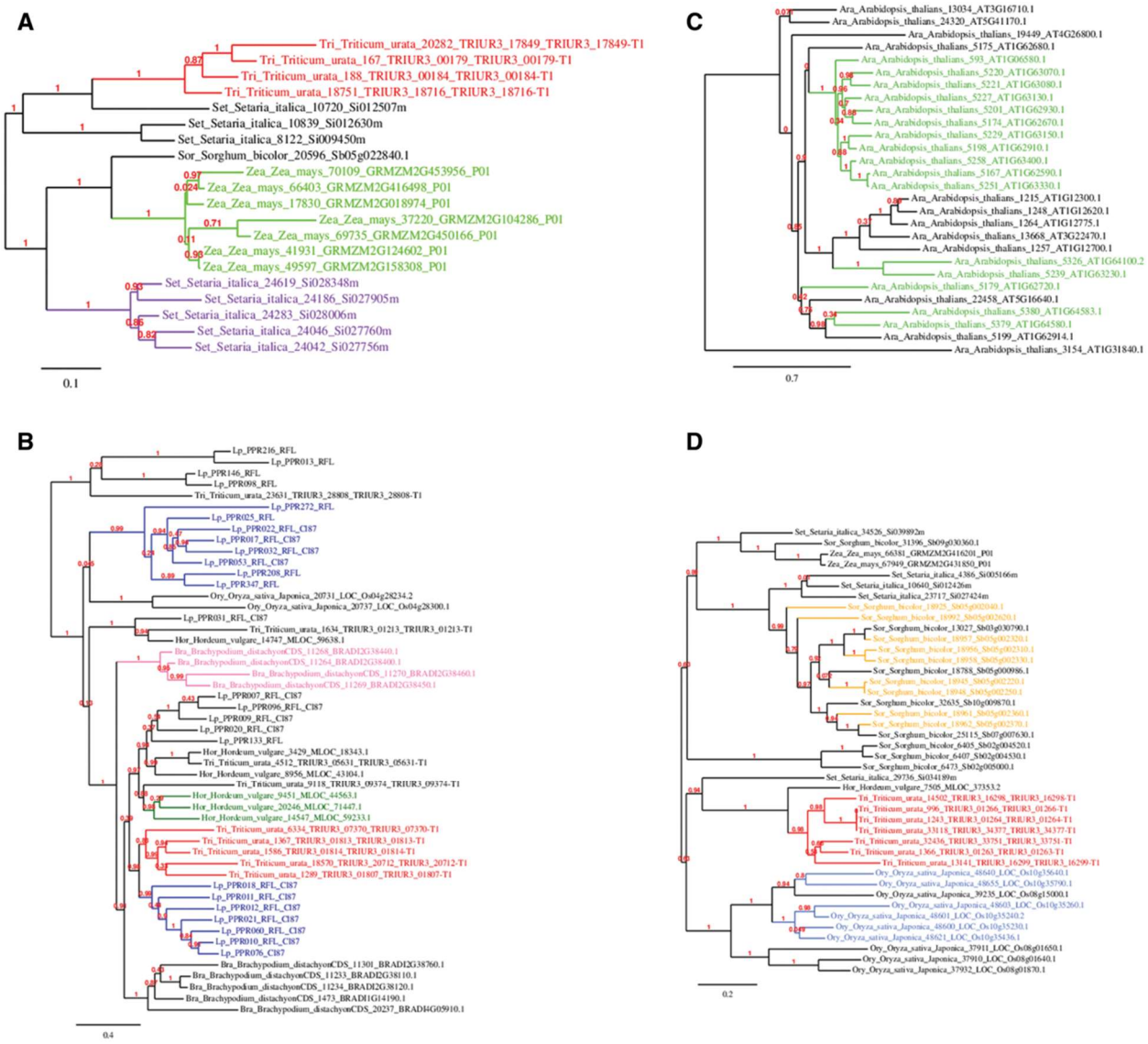



**Figure 4.** Phylogenetic trees generated using protein sequences from the OrthoMCL generated RFL cluster. **A**, Clade 1. **B**, Clade 2. **C**, Clade 3. **D**, Clade 4. Colours represent clusters of sequences originating from the same genomic region of that species or showing a similar arrangement when genome information is not available.

Given the abundance of *RFL*s within these relatively small genome regions, these sites can be considered hotspots for *RFL* recombination that exhibit elevated rates of recombination relative to a neutral expectation. *RFL* genes within these clusters, at the same genome region, will be the youngest as they are still present within this *RFL* recombination hotspot. This implies that any list of candidate *PPR-Rf* genes can be further narrowed to *RFL*s present within these zones. These regions of active *RFL* generation contained known *Rf* genes, with the rice *Rf1* [52] and *Rf4* [53] genes being present in the *RFL*-rich region of rice chromosome 10. This allowed us to further refine the list of possible *Rf* genes in perennial ryegrass by looking for groups of tightly clustering sequences that show a similar pattern to other species. From Clade 2, three groups of perennial ryegrass *RFL*s meeting these criteria were identified, comprising seven, eight and five sequences respectively (given in blue, Figure 4B). The first of these groups contained only sequences present in the OrthoMCL RFL cluster, the second group four sequences from this cluster and four from the original RFL genome scan, and the third cluster four from the RFL cluster and one from the genome scan.

## Synteny analysis

In order to identify the genome position of *RFL* generation in perennial ryegrass, a comparative genomics approach based on the Genome Zipper [54] was applied. The *RFL*-rich zones from species with available genomic information were searched for conserved synteny with the genomes of other related plant species. The comparative genomics tool available at Ensemble Plants (http://plants.ensembl.org/index.html) was used to discern if any synteny exists between the *RFL*-rich genomic regions, from different species, within a single clade. This revealed no synteny between any *RFL*-rich regions neither within each clade nor between clades.

## Discussion

A bioinformatics pipeline targeting candidate *Rf* genes for CMS was successfully established and identified three clusters of possible *RFL* generation in perennial ryegrass. This pipeline, consisting of three complementing steps (supplementary Fig.1), is based on genomic sequence data and thus can be used in any plant species for which such data is available. Validation of the pipeline in fully sequenced grass species such as rice, Brachypodium and sorghum revealed that 50 to 90% of candidate *RFL* genes are found within no more than three genomic regions consisting of 0.01 to 0.1% of the genome. A similar approach could now be applied to cereals, where efficient access to *Rf* genes is an integral part of CMS-based hybrid production [55].

The first step of this pipeline utilises protein domain profile matrices and sequence comparisons to identify PPR and RFL proteins from translated CDS. The second step involves orthologous clustering of multiple species to identify *RFL* genes. This second step does not undermine the first step of this pipeline but is complementary, as the first step identifies a more complete set of *RFL*s including pseudogenes and poorly annotated genes, both of which are important in identifying *RFL* recombination hotspots. The second step is also integral as it provides the data to complete the third and final step, which employs phylogenetic analysis to recognise areas of *RFL* diversification within the genome. This method not only identifies candidate *PPR-Rf* genes from restoring genotypes, but also enables efficient identification of dynamic *RFL* clusters from non-restoring phenotypes [56, 57, 36, 37].

### PPR and *RFL* genes in perennial ryegrass

In the draft genome sequence of perennial ryegrass [48], 373 *PPR* genes were identified and classified, revealing 25 *RFL*s. The number of *RFL*s identified here is consistent with other studies that have reported 10-30 *RFL*s per genome [45], for example in Arabidopsis [36]. These *RFL*s have, on average, six more PPR domains than non-RFL PPR proteins. This possibly indicates that in perennial ryegrass, RFLs have a higher RNA sequence specificity than other PPR proteins. This was expected, as known PPR-Rf proteins bind to a specific mRNA sequence whereas other PPRs have been shown to bind to multiple mRNAs [33]. Further evidence for multiple binding specificities comes from the number of transcript editing sites being present in mitochondrial genomes compared to PPRs with editing domains. The Arabidopsis mitochondrial genome encompasses 441 cytosine to uracil editing sites, although only 193 *PPR* genes, containing the E domain required for transcript editing, can be found in the nuclear genome [58, 36]. It appears that RFL proteins, unlike some other PPR proteins, are highly specialised, targeting a single transcript within the mitochondria [24].

### Orthology-based strategies for RFL identification

By using orthologous clustering, *RFL* genes from nine species were identified, showing that *RFL*s are distinct enough to be identified directly from whole genome sequence data without first identifying the *PPR* gene family [59, 46, 37]. This was exemplified in Figure 2B where the only cluster containing more than 50 *PPR*s was the *RFL* cluster. Strikingly, all known *Rf* genes that were present in the original genomes used for clustering were found in the *RFL* cluster. This also validates the sequence alignment and comparison approach used to identify *RFL*s from the whole set of *PPR* genes in perennial ryegrass. Non-*RFL PPR*s also clustered together with their orthologs from different species, but in contrast to *RFLs*, most of these clusters contained only one orthologous *PPR* gene per species.

Although the orthologous clustering and phylogenetic approach is an effective method to identify regions of active *RFL* generation, it was unable to identify all PPRs and was also

less successful at identifying *RFL* genes from perennial ryegrass than the genome scanning approach. The effectiveness of the orthologous clustering and phylogenetic approach is dependent upon the type and quality of the input data. The type of data used is important as genomic sequence information may contain a more complete set of *RFL* genes than transcriptome data because of the tissue- and time-specific expression of *RFL* genes [60]. This is highlighted by the Italian ryegrass and meadow fescue transcriptomes, comprising relatively few *RFLs.* On the other hand, due to this tissue- and time-specific expression, transcriptome data could also be used to enrich for *Rf* genes by sampling from tissues known to be expressing *Rf* proteins, such as anthers [61]. Although the use of genomic sequence data is preferable, individual *RFLs* can still be overlooked by orthologous clustering if they are poorly annotated or pseudogenes. Moreover, using incomplete genome assemblies as input data may not reveal all *RFL* clusters as they can be difficult to assemble, due to the repetitive features of *RFL*-rich genomic regions [62]. This was observed in barley (*Hordeum vulgare* L.), where an *RFL* was identified on an unordered contig from the same chromosome 6HS containing a recently mapped *Rf* locus that could not be associated with an *RFL* cluster [63]. In cereals, further functional restorer loci have been described in wheat [64] and rye (*Secale cereal* L.) [65]. The use of sequence data from restoring individuals in conjunction with the pipeline described here could help to identify candidate *Rf-PPR* genes within the identified regions.

**Genome regions of active RFL gene generation**

To identify the likely location of any *PPR-Rf* genes, a phylogenetic approach was applied to find clusters of highly similar *RFL* genes within single species, allowing the genomic regions of *RFL* generation to be distinguished. By comparing species with genome location information to perennial ryegrass, three regions of possible *RFL* generation were identified. Through further phylogenetic analysis of *RFL*s from several species, the fine structure of *RFL* organisation in grasses was resolved and regions of novel *RFL* generation in species with positional genome information identified. This understanding of the architecture of *RFL* genes within other grass species led to the identification of similar groups of *RFL* genes in perennial ryegrass. Given the phylogenetic similarities between these groups, we can confidently assume that each of these groups of *RFL* genes in perennial ryegrass will be represented at single loci within the genome. These loci could be elucidated with more detailed genomic information or the use of a mapping population for genetic linkage mapping. Wild einkorn wheat, another species without genome location information, also showed a similar pattern, with three tight clusters indicating the likelihood of three *RFL* generation loci.

The rate of recombination within the mitochondrial genome, which is the source of novel CMS mechanisms, is high [66, 10], requiring a relatively rapid generation of new *RFL* genes through recombination driven diversifying selection [45]. The likelihood of functional *PPR-Rf* genes being present in these zones of active *RFL* generation is a function of how long it takes for fertility restoration to become fixed within a population (the time it takes for an *Rf* gene to restore CMS in an entire population) and the rate at which *RFL* genes are shuffled throughout the genome (how long a newly functional *Rf* gene is likely to stay within the genome region of active *RFL* generation). This suggests that if the rate of fixation is faster than the rate of shuffling, *Rf* genes will always be found within these *RFL* clusters. This is further borne out by the genome synteny results, showing a breakdown of synteny in the region of *RFL* generation zones, indicating that novel *RFL* generation occurs faster than speciation, unlike other *PPR* genes that are highly conserved between species. Similar findings were reported for barley and rye, where *Rf* containing regions showed synteny to regions from rice, Brachypodium and sorghum that contained no *RFLs* [65, 63]. These results indicate not only that *RFL*s are being shuffled around the genome at a rate faster than that of speciation but also that they are being rapidly lost when non-functional [56].

In the four clades identified within the *RFL* cluster, all the dicot *RFL* genes fell within a single clade, representing the split between monocots and dicots. Although the dicot sequences were in a separate clade, the fact that *RFL*s from both monocot and dicot species were identified within a single cluster based on orthologous clustering is consistent with the hypothesis that monocot and dicot *RFL* genes share a common ancestor. This also suggests that this common ancestor is distinct from all other *PPR* genes and predates the monocot/dicot split, meaning that *RFL* genes evolved before this split [37].

**Accuracy and usefulness of this approach**

The approach presented here allows efficient targeting of *RFL* containing genomic region(s) in multiple species. These regions have previously been shown to contain *Rf-PPR* genes [61, 67-73]. In grasses, examples can be found in maize with the *Rf8* locus mapping to an *RFL* cluster on chromosome 2 [74], and in rice with the *Rf1* [52] and *Rf4* [53] genes being present within the *RFL* cluster of rice chromosome 10. The most recent example is the *Rf6* restorer in rice [75]. *Rf6* was mapped to a 200 kb region on rice chromosome 8, which contains three *RFL*s identified in this study, with one of these genes (Os08g01870) being located within 15kb of the marker shown to be cosegregating with the restorer gene [76]. The only identified *PPR-Rf* gene that is located outside of the *RFL*-rich regions is *Rf1* from sorghum. The *Rf1* locus, most likely encoded by *PPR13*, is located as a single *PPR-Rf* gene on chromosome 8 although *PPR13* was not cloned from a restoring genotype [77]. *PPR13* is different in its structure from all other identified *RFL-Rf* genes as it is of the PLS subtype and contains domains linked with RNA editing, indicating that the mechanism for restoration of the CMS phenotype may also be unique [56, 32]. *PPR13* also exemplifies the complementarity of protein domain profile matrix scans and orthologous clustering, the latter of which would have been unable to detect a gene like *PPR13.*

The clustering approach assumes that newly functional *PPR-Rf* genes are the result of recombination events within an *RFL* genomic cluster and not an existing *RFL* that has gained a restoring function through the serendipitous recognition of a novel CMS causing transcript within the mitochondria. This balance will most likely differ between species and between populations of the same species under differing environmental conditions. It is important to note here that this approach will be most successful in identifying *PPR-Rf* genes in naturally occurring CMS systems (where the rapid evolution of *RFL*s has had time to overcome the damage in the mitochondria), but will also find traction in induced CMS systems where the CMS phenotype still has a mitochondrial ORF as its source and as such a possible *PPR-Rf* gene as a restorer.

**The value of Rf genes for CMS-based pollination control in forage grasses**

This pipeline provides an efficient first approach for *Rf* gene identification as it permits researchers to target the most likely genomic regions to contain *Rf* genes. Rapid identification of *Rf* candidate *RFL* genes will facilitate the development of functional markers for restoration of fertility, enabling efficient exploitation of CMS as a tool to control pollination for hybrid breeding in forage grasses. However, fertile hybrid seed is not necessarily needed for temporary forage production as biomass and not seed is the primary yield target [21]. Indeed, it is often unwelcome, as any partial or full restoration of male fertility during hybrid seed production would decrease the purity and value of that seed. Nevertheless, *Rf* gene identification is important to ensure that markers can be designed and populations screened to prevent unwanted fertility restoration. This will help to overcome the main challenge in outbreeding forage grasses with highly heterozygous genomes, which is the maintenance of the CMS trait. The ability to rapidly identify individuals carrying an *Rf* gene within a breeding population would assist

breeders in maintaining the commercially important CMS phenotype as well as ensuring hybrid seed purity. For breeding purposes, the exact position of the *Rf* gene does not need to be identified as genetic markers tightly linked to the functional *Rf* gene might be sufficient to identify restoring phenotypes. The approach used in this study can provide this by identifying *RFL* clusters within the genome, allowing the relatively rapid identification of useful markers. Further dissection of *RFL* clusters, possible through BAC library screen and subsequent BAC clone sequencing, would allow the identification and cloning of the responsible *Rf* gene.

## Conclusion

Here, we have designed and implemented an *in silico* pipeline to identify candidate *Rf-PPR* genes and demonstrated its effectiveness by pinpointing known *Rf* genes. This study focused on perennial ryegrass and identified three regions of active *RFL* generation, providing excellent targets for marker development and future mapping approaches. Information is also provided for other species such as wild einkorn wheat, showing the wider applications of this method. As demonstrated, this pipeline can also be used to characterise *RFLs* in both monocots and dicots to provide new insights into their evolution. The predictive power of this approach will improve as more genome sequence data becomes available. Knowledge of *RFL*-rich genomic regions within a genome might also be used for targeted sequencing of such regions in restorer plants and facilitate the expedient determination of *Rf* genes, the knowledge of which would not only be useful for breeding programs but also for fundamental research into nuclear/mitochondrial interactions.

## Methods

### Identification of pentatricopeptide repeat (PPR) proteins

To identify, *in silico*, members of the PPR protein family in the genome assembly of perennial ryegrass (http://185.45.23.197:5080/ryegrassgenome), all available PPR domain sequences from the Pfam database (http://pfam.xfam.org) were collected and used for the development of a Hidden Markov Model (HMM) profile matrix using the *hmmbuild* program of the HMMER package (v3.1b1, http://hmmer.org). This HMM profile matrix was used to identify members of the PPR family in a total of 71,009 translated DNA transcript sequences obtained from *ab initio* and evidence-based predictions from a high-quality genomic draft of the perennial ryegrass genome sequence [48].

### Classification of pentatricopeptide repeat (PPR) proteins

PPR-containing transcript sequences were analyzed on a standalone PfamScan pipeline to ascertain the exact co-ordinates of each PPR domain within a scaffold sequence as well as information on the frequencies and distribution of the PPR domains. Predictive information on protein functions and conserved sequence elements was obtained by sending all PPR containing sequences through a standalone InterProScan (version 5) [78] pipeline by scanning the PANTHER, PROSITE profiles, Pfam and SUPERFAMILY databases. Sequences were identified as belonging to the P or PLS subfamilies through analysis of PPR motif lengths, with the PLS subfamily having longer (L) and shorter (S) subdomains [36]. The identified members of the PLS family were processed using the online domain elicitation tool MEME [79] and conserved blocks representing the E, E+ and DYW C-terminal domains identified. To ensure all possible C-terminal domains have been identified, the PPR domains were masked out using the *maskfeat* program of the EMBOSS package (Rice). The masked sequences were aligned and clustered to identify any conserved regions outside of the PPR domains. All sequences were also searched using HMM profiles for the E, E+ and DYW domains.

### Identification of restorer of fertility-like PPR (RFL) proteins

All identified PPR genomic sequences were clustered using CD-hit [80] at 90%, 80%, 70%, 60% and 50% identity. Clustering at 90% to 70% revealed no clusters of more than three members. Clustering at 60% revealed 3 clusters containing 9, 6 and 4 PPRs, respectively. All PPR sequences were then aligned, using the NCBI BLAST platform (http://blast.ncbi.nlm.nih.gov/), to known or predicted restorer genes from; brachypodium (gi|357139997), rice (gi|33859441) and maize (gi|662249846). Hits with at least 50% identity and 50% query cover were collected. PPRs that were present on at least three of these four lists were considered candidate RFLs.

### Databases

The CDS of the following species were downloaded from Ensembl Plants (http://plants.ensembl.org/index.html, on 10/04/2014) [81] using the Perl API tool [50]; *Arabidopsis thaliana* (TAIR10), *Brachypodium distachyon* (V1.0), *Hordeum vulgare* (European Nucleotide assembly (ENA): GCA_000326085.1), *Musa acuminata* (ENA: GCA_000313855.1), *Oryza sativa Japonica* (ENA: GCA_000005425.2), *Setaria italica* (ENA: GCA_000263155.1), *Sorghum bicolor* (ENA: GCA_000003195.1), *Triticum urartu* (ENA: GCA_000347455.1), *Zea mays* (ENA: GCA_000005005.5). The following CDS of *Phyllostachys heterocla* da (v1.0) was downloaded from http://www.bamboogdb.org/ [82]. The

CDS of *Lolium perenne* was received from Ruttink et al. (2013) [83] and the respective CDS of *Lolium multiflorum* and *Festuca pratense* were kindly provided by Stoces et al. (in preparation). The *Eragrostis tef* cDNA was downloaded from http://www.tef-research.org/genome.html (Extended.gte200.cDNA.fa) [84] and its CDS determined using OrfPrdictor [85]. The cDNA was then searched against a protein BLAST database comprising of *Arabidopsis thaliana*, *Glycine max*, *Oryza sativa Japonica*, *Populus trichocarpa*, and *Manihot esculenta*, using BLASTP [86] with minimum e-value $1e^{-5}$. The BLASTP results were used to infer coding frame, all other parameters and methods used were as described by Min et al. (2005) [85].

**Orthologous clustering of species**

To cluster the protein sequences into orthologous clusters, the offline version of OrthoMCL [51] was used. Briefly, the protein names within a fasta file (per species) were first changed for consistency (also for simplicity) and to ameliorate any problems arising later from special characters and similarities between names. This was done using an in house Perl script. The resulting fasta file was then formatted to make it compliant with the OrthoMCL algorithm (a short species-specific prefix was added to each name for subsequent species identification). The sequences were then filtered for low quality, based on sequence length (>30 aa, retained) and percentage of stop codons (>10%, discarded). From these high quality proteins, an all-vs-all BLASTP was run where all proteins were searched against all proteins (minimum E-value $1e^{-5}$); the database was not split into subgroups when doing this so no corrections for E-score where necessary. The results of the BLASTP were collated and then parsed before loading into a local MySQL orthoMCL database. In the next stage, pairs of proteins that are potentially orthologs, in-paralogs or co-orthologs were identified using the OrthoMCL algorithm [51], where protein pairwise connections were normalised for ortholog pairs between and within species. The resulting potential pairs were then organised in clusters using the MCL alogirthm [87]. The results were output and the names were changed back to their original for subsequent work.

**Phylogenetic reconstruction and analysis**

The phylogenetic relationships between the protein sequences from the OrthoMCL generated cluster containing RFLs, including the 9 putative ryegrass RFLs not present, were reconstructed and analysed using web tools made available by The Montpellier Laboratory of Informatics, Robotics and Microelectronics LIRMM (http://www.phylogeny.fr/) [88]. Sequence alignments were completed using MUSCLE [89], phylogenetic analysis using PhyML [90, 91] and the resulting tree viewed using TreeDyn [92].

## Acknowledgment

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no PITN-GA-2013-608422 – IDP BRIDGES. The project was also supported by the Swiss National Science Foundation (SNSF Professorship grant no: PP00P2 138988). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We sincerely thank Prof. Dr. Achim Walter for hosting the Forage Crop Genetics group at ETH Zurich.

# References

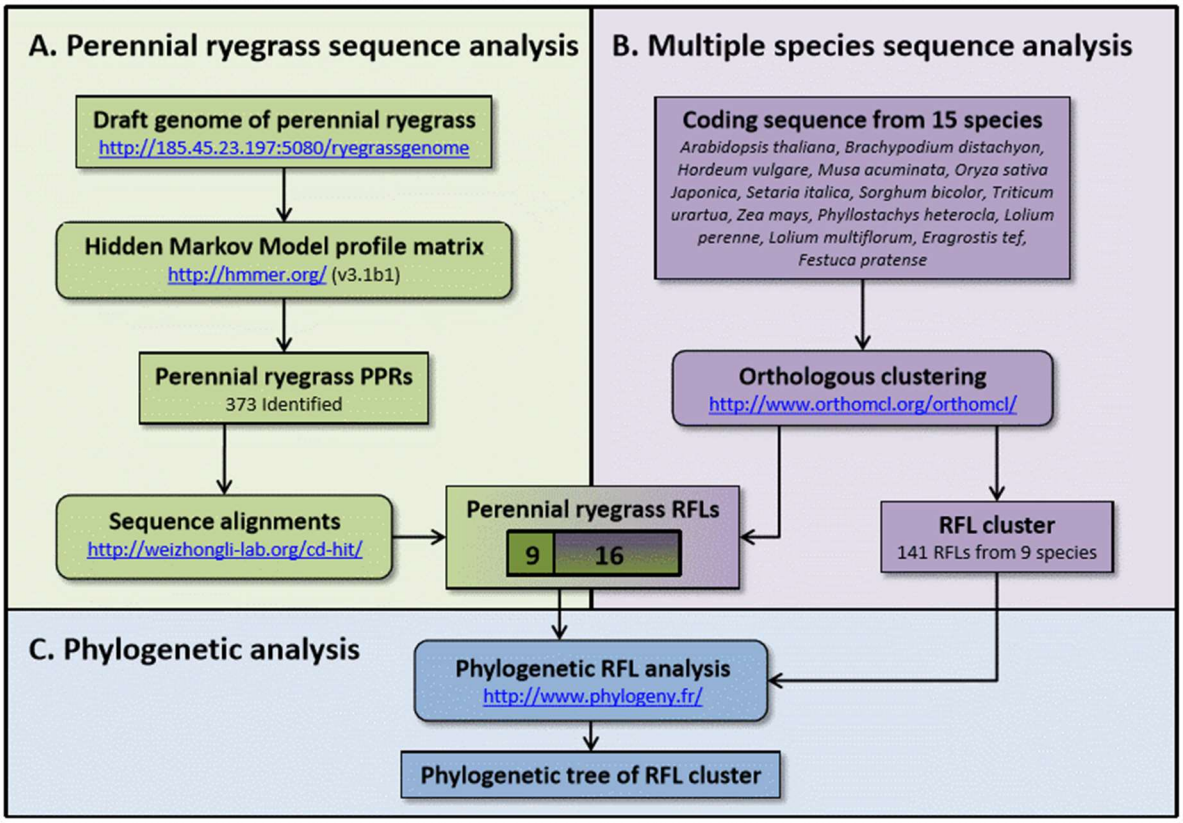

1. Wilkins PW. 1991. Breeding perennial ryegrass for agriculture. *Euphytica.* 52(3):201-214.
2. Pembleton LW, et al. 2015. Design of an F1 hybrid breeding strategy for ryegrasses based on selection of self-incompatibility locus-specific alleles. Frontiers in plant science. 6:764-64.
3. van der Heijden SAG, and Roulund N, editors. 2010. Genetic Gain in Agronomic Value of Forage Crops and Turf: A Review. *Sustainable Use of Genetic Diversity in Forage and Turf Breeding*. pp.247-60.
4. Duvick DN. 2001. Biotechnology in the 1930s: the development of hybrid maize. Nature Reviews Genetics. 2(1):69-74.
5. Melchinger AE. 2010. The International Conference on 'Heterosis in Plants'. Theor Appl Genet. 120(2):201-03.
6. Longin CFH, et al. 2012. Hybrid breeding in autogamous cereals. Theor Appl Genet. 125(6):1087-96.
7. Virmani SS, editor. 1994. Hybrid rice technology: new developments and future prospects. *Hybrid rice technology: new developments and future prospects.* pp.296.
8. Ahokas H. 1983. Cytoplasmic male sterility in barley. XV. PI 296897 as a restorer of fertility in *msml* and *msm2* cytoplasms. Hereditas. 99(1):157-59.
9. Havey MJ. 2004. The use of cytoplasmic male sterility for hybrid seed production. *Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria*. pp. 623-34. Springer Netherlands.
10. Kubo T, et al. 2011. Male Sterility-Inducing Mitochondrial Genomes: How Do They Differ? Critical Reviews in Plant Sciences. 30(4):378-400.
11. Martin AC. et al. 2009. Chromosome engineering in wheat to restore male fertility in the msH1 CMS system. Mol Breed. 24(4):397-408.
12. Schnable PS, and Wise RP. 1998. The molecular basis of cytoplasmic male sterility and fertility restoration. Trends in Plant Science. 3(5):175-80.
13. Singh SK, Chatrath R, and Mishra B. 2010. Perspective of hybrid wheat research A review. Indian Journal of Agricultural Sciences. 80(12):1013-27.
14. Yuan LP, Virmani SS. 1988. Status of hybrid rice research and development. Hybrid rice. Proceedings of an international symposium, Changsha, China, 6-10 October 1986. 7-24.
15. Wit F. 1974. Cytoplasmic male sterility in ryegrass (*Lolium* spp.) detected after intergenic hybridization. Euphytica. 23:31-38.
16. Connoly V, Wright-Turner R. 1984. Induction of cytoplasmic male-sterility into ryegrass (*Lolium perenne*). Theor Appl Genet. 68(5):449-53.
17. Creemersmolenaar J, Hall RD, Krens FA. 1992. Asymmetric protoplast fusion aimed at intraspecific transfer of cytoplasmic male sterility (CMS) in *Lolium perenne* L. Theor Appl Genet. 84(5-6):763-70.
18. Kiang AS, et al. 1993. Cytoplasmic male sterility (CMS) in *Lolium perenne* L.: 1. Development of a diagnostic probe for the male-sterile cytoplasm. Theor Appl Genet. 86(6):781-87.
19. Kiang AS, Kavanagh TA. 1996. Cytoplasmic male sterility (CMS) in *Lolium perenne* L .2. The mitochondrial genome of a CMS line is rearranged and contains a chimaeric *atp9* gene. Theor Appl Genet. 92(3-4):308-15.
20. McDermott P, Connolly V, Kavanagh T. 2008. The mitochondrial genome of a cytoplasmic male sterile line of perennial ryegrass (*Lolium perenne* L.) contains an integrated linear plasmid-like element. Theor Appl Genet. 117(3):459-70.
21. Islam MS, et al. 2014. Genetics and biology of cytoplasmic male sterility and its applications in forage and turf grass breeding. Plant Breeding. 133(3):299-312.
22. Hanson MR, Bentolila S. 2004. Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell. 16:154-S69.
23. Chase CD, Babay-Laughnan S. 2004. Cytoplasmic male sterility and fertility restoration by nuclear genes. *Molecular Biology and Biotechnology of Plant Organelles*. Springer Netherlands, pp. 593-622.
24. Barkan A, Small I. 2014. Pentatricopeptide Repeat Proteins in Plants. S. S. Merchant (ed.), *Annual Review of Plant Biology.* 65:415-442.
25. Cui XQ, Wise RP, Schnable PS. 1996. The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. Science. 272(5266):1334-36.

26. Kitazaki K, et al. 2015. Post-translational mechanisms are associated with fertility restoration of cytoplasmic male sterility in sugar beet (*Beta vulgaris*). Plant J. 83(2):290-99.
27. Hu J, et al. 2012. The Rice Pentatricopeptide Repeat Protein RF5 Restores Fertility in Hong-Lian Cytoplasmic Male-Sterile Lines via a Complex with the Glycine-Rich Protein GRP162. Plant Cell. 24(1):109-22.
28. Itabashi E, et al. 2011. The fertility restorer gene, *Rf2*, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein. Plant Journal. 65(3):359-67.
29. Castandet B, Araya A. 2011. RNA Editing in Plant Organelles. Why Make It Easy? Biochemistry-Moscow. 76(8):924-31.
30. Chateigner-Boutin AL, Small I. 2010. Plant RNA editing. RNA Biology. 7(2):213-19.
31. Fujii S, Small I. 2011. The evolution of RNA editing and pentatricopeptide repeat genes. New Phytologist. 191(1):37-47.
32. Schmitz-Linneweber C, Small I. 2008. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends in Plant Science. 13(12):663-70.
33. Zehrmann A, et al. 2009. A DYW Domain-Containing Pentatricopeptide Repeat Protein Is Required for RNA Editing at Multiple Sites in Mitochondria of *Arabidopsis thaliana*. Plant Cell. 21(2):558-67.
34. Small I, Rackham O, Filipovska A. 2013. Organelle transcriptomes: products of a deconstructed genome. Current Opinion in Microbiology. 16(5):652-58.
35. Colcombet J, et al. 2013. Systematic study of subcellular localization of Arabidopsis PPR proteins confirms a massive targeting to organelles. RNA Biology. 10(9):1557-75.
36. Lurin C, et al. 2004. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell. 16(8):2089-103.
37. O'Toole N, et al. 2008. On the expansion of the pentatricopeptide repeat gene family in plants. Mol Biol Evol. 25(6):1120-28.
38. Schallenberg-Ruedinger M, et al. 2013. A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. RNA Biology. 10(9):1549-56.
39. Hammani K, et al. 2009. A Study of New Arabidopsis Chloroplast RNA Editing Mutants Reveals General Features of Editing Factors and Their Target Sites. Plant Cell. 21(11):3686-99.
40. Okuda K, et al. 2009. Pentatricopeptide Repeat Proteins with the DYW Motif Have Distinct Molecular Functions in RNA Editing and RNA Cleavage in Arabidopsis Chloroplasts. Plant Cell. 21(1):146-56.
41. Okuda K, Shikanai T. 2012. A pentatricopeptide repeat protein acts as a site-specificity factor at multiple RNA editing sites with unrelated cis-acting elements in plastids. Nucleic Acids Research. 40(11):5052-64.
42. Tasaki E, Hattori M, Sugita M. 2010. The moss pentatricopeptide repeat protein with a DYW domain is responsible for RNA editing of mitochondrial *ccmFc* transcript. Plant J. 62(4):560-70.
43. Toda T, et al. 2012. Rice *MPR25* encodes a pentatricopeptide repeat protein and is essential for RNA editing of *nad5* transcripts in mitochondria. Plant J. 72(3):450-60.
44. Andrés C, Lurin C, Small I. 2007. The multifarious roles of PPR proteins in plant mitochondrial gene expression. Physiol Plant. 129(1):14-22.
45. Fujii S, Bond CS, Small I. 2011. Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. Proc Natl Acad Sci. 108(4):1723-28.
46. Geddy R, Brown GG. 2007. Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. BMC Genomics. 8:13.
47. Chase CD. 2007. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. Trends in Genetics. 23(2):81-90.
48. Byrne S, et al. 2015. A synteny-based draft genome sequence of the forage grass *Lolium perenne*. The Plant J. 84(4):816-826
49. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:29-37.
50. McLaren W, et al. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 26(16):2069-70.
51. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178-89.
52. Wang Z, Zon Y, Li X. 2006. Cytoplasmic male sterility of rice with BoroII cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. *Plant Cell.* 18:676 - 87.
53. Luo D, et al. 2013. A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. Nature Genetics. 45(5):573-U157.

54. Pfeifer M, et al. 2013. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. Plant physiology. 161(2):571-582.
55. Whitford R, et al. 2013. Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. J Exp Bot. 64(18):5411-28.
56. Dahan J, Mireau H. 2013. The Rf and Rf-like PPR in higher plants, a fast-evolving subclass of PPR genes. RNA Biology. 10(9):1469-76.
57. Li SB, et al. 2012. Phylogenetic Genomewide Comparisons of the Pentatricopeptide Repeat Gene Family in indica and japonica Rice. Biochemical Genetics. 50(11-12):978-89.
58. Giege P, Brennicke A. 1999. RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci. 96(26):15324-29.
59. Desloire S, et al. 2003. Identification of the fertility restoration locus, Rfo, in radish, as a member of the pentatricopeptide-repeat protein family. Embo Reports. 4(6):588-94.
60. Prasad K, Kushalappa K, Vijayraghavan U. 2003. Mechanism underlying regulated expression of RFL, a conserved transcription factor, in the developing rice inflorescence. Mech Dev. 120(4):491-502.
61. Kazama T, Toriyama K. 2014. A fertility restorer gene, *Rf4*, widely used for hybrid rice breeding encodes a pentatricopeptide repeat protein. Rice. 7(1):1-5.
62. Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology. 11(4):R41.
63. Ui H, et al. 2015. High-resolution genetic mapping and physical mapconstruction for the fertility restorer *Rfm1* locus in barley. Theor Appl Genet. 128:283-290.
64. Ma ZQ, Zhao YH, Sorrells ME. 1995. Inheretence and chromosomal locations of male fertility restoring gene transferred from *Aegilops umbellulata* Zhuk. to *Triticum aestivum* L. Mol Gen Genet. 247:351-357.
65. Hackauf B, Korzun V, Wortmann H, Wilde P, Whehling P. 2012. Development of conserved ortholog set markers linked to the restorer gene *Rfp1* in rye. Mol Breed. 30:1507-1518.
66. Sloan DB, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS-Biology. 10(1):53.
67. Bentolila S, Alfonso AA, Hanson MR. 2002. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci. 99:10887–10892.
68. Kazama T, Nakamura T, Watanabe M, Sugita M, Toriyama K. 2008. Suppression mechanism of mitochondrial ORF79 accumulation by Rf1 protein in BT-type cytoplasmic male sterile rice. Plant J. 55:619–628.
69. Uyttewaal M, et al. 2008. Characterization of Raphanus sativus pentatricopeptide repeat proteins encoded by the fertility restorer locus for Ogura cytoplasmic male sterility. Plant Cell. 20:3331–3345.
70. Barr CM, Fishman L. 2010. The nuclear component of a cytonuclear hybrid incompatibility in Mimulus maps to a cluster of pentatricopeptide repeat genes. Genetics. 184(2):455-465.
71. Jo YD, Kim YM, Park MN, Yoo JH, Park M, Kim BD, Kang BC. 2010. Development and evaluation of broadly applicable markers for Restorer-of-fertility in pepper. Mol breed. 25(2):187-201.
72. Jordan DR, Klein RR, Sakrewski KG, Henzell RG, Klein PE, Mace ES. 2011. Mapping and characterization of Rf 5: a new gene conditioning pollen fertility restoration in A1 and A2 cytoplasm in sorghum (*Sorghum bicolor* (L.) Moench). Theor Appl Genet. 123(3):383-396.
73. Bisht DS, Chamola R, Nath M, Bhat SR. 2015. Molecular mapping of fertility restorer gene of an alloplasmic CMS system in Brassica juncea containing Moricandia arvensis cytoplasm. Mol Breed. 35(1):1-11.
74. Meyer J, Pei D, Wise RP. 2011. Rf8-Mediated T-Transcript Accumulation Coincides with a Pentatricopeptide Repeat Cluster on Maize Chromosome 2L. The Plant Genome. 4(3): 283-299.
75. Huang W, Yu C, Hu J, Wang L, Dan Z, Zhou W, He C, Zeng Y, Yao G, Qi J, Zhang Z. Pentatricopeptide-repeat family protein RF6 functions with hexokinase 6 to rescue rice cytoplasmic male sterility. Proceedings of the National Academy of Sciences. 2015 Dec 1;112(48):14984-9.
76. Huang W, Hu J, Yu C, Huang Q, Wan L, Wang L, Qin X, Ji Y, Zhu R, Li S, Zhu Y. Two non-allelic nuclear genes restore fertility in a gametophytic pattern and enhance abiotic stress tolerance in the hybrid rice plant. Theoretical and applied genetics. 2012 Mar 1;124(5):799-807.
77. Klein RR, et al. 2006. 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. 112(2):388-88.
78. Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236-40.

79. Bailey TL, Gribskov M. 1998. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 14(1):48-54.
80. Li WZ, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658-59.
81. Flicek P, et al. 2012. Ensembl 2012. Nucleic Acids Res. 40(1):84-90.
82. Peng Z, et al. 2013. The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). Nature Genetics. 45(4):456-61.
83. Ruttink T, et al. 2013. Orthology Guided Assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in *Lolium perenne*. *Plant* Biotechnology J. 11(5):605-17.
84. Cannarozzi G, et al. 2014. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). BMC Genomics. 15:581.
85. Min XJ, et al. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 33:677-80.
86. Altschul SF, et al. 1990. Basic Local Alignment Search Tool. J Mol Biol. 215(3):403-10.
87. Enright AJ, van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575-84.
88. Dereeper A, et al. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36:465-69.
89. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-97.
90. Anisimova M, et al. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 55(4):539-52.
91. Guindon S, and Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52(5):696-704.
92. Chevenet F, et al. 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics. 7(1):439.

## Supplementary Data




**Supplementary Figure 1.** Work flow diagram illustrating the three-step bioinformatics pipeline used to isolate *RFL* genes in perennial ryegrass (*Lolium perenne* L.). Boxes with rounded corners contain bioinformatics processes, boxed with square corners contain input/output data.

**Supplementary Table 1.** Gene identifiers for all sequence present in Figure 3 and Figure 4.

**Clade 1:** Tri_Triticum_urartu_20282_TRIUR3_17849_TRIUR3_17849-T1,
Tri_Triticum_urartu_167_TRIUR3_00179_TRIUR3_00179-T1,
Tri_Triticum_urartu_188_TRIUR3_00184_TRIUR3_00184-T1,
Tri_Triticum_urartu_18751_TRIUR3_18716_TRIUR3_18716-T1, Set_Setaria_italica_10720_Si012507m,
Set_Setaria_italica_24042_Si027756m, Set_Setaria_italica_24619_Si028348m,
Set_Setaria_italica_24186_Si027905m, Set_Setaria_italica_24283_Si028006m,
Set_Setaria_italica_24046_Si027760m, Set_Setaria_italica_10839_Si012630m,
Set_Setaria_italica_8122_Si009450m, Sor_Sorghum_bicolor_20596_Sb05g022840.1,
Zea_Zea_mays_17830_GRMZM2G018974_P01, Zea_Zea_mays_70109_GRMZM2G453956_P01,
Zea_Zea_mays_66403_GRMZM2G416498_P01, Zea_Zea_mays_37220_GRMZM2G104286_P01,
Zea_Zea_mays_69735_GRMZM2G450166_P01, Zea_Zea_mays_41931_GRMZM2G124602_P01,
Zea_Zea_mays_49597_GRMZM2G158308_P01

**Clade 2:** Lp_PPR216_RFL (scaffold_6346_ref0010855), Lp_PPR013_RFL (scaffold_3083_ref0029026),
Lp_PPR146_RFL (scaffold_16939_ref0006639), Lp_PPR098_RFL (scaffold_37184_ref0046281),
Lp_PPR032_RFL_Cl87 (scaffold_11058_ref0036534), Lp_PPR053_RFL_Cl87 (scaffold_1938_ref0031963),
Lp_PPR025_RFL (scaffold_18885_ref0023357), Lp_PPR272_RFL (scaffold_6077_ref0030816),
Lp_PPR208_RFL (scaffold_909_ref0007600), Lp_PPR022_RFL_Cl87 (scaffold_9752_ref0003803),
Lp_PPR017_RFL_Cl87 (scaffold_12_ref0042665), Lp_PPR347_RFL (scaffold_11058_ref0036534),
Lp_PPR031_RFL_Cl87 (scaffold_6777_ref0044368), Lp_PPR133_RFL (scaffold_29153_ref0009698),
Lp_PPR007_RFL_Cl87 (scaffold_3597_ref0021217), Lp_PPR009_RFL_Cl87 (scaffold_13850_ref0032247),
Lp_PPR020_RFL_Cl87 (scaffold_5048_ref0037858), Lp_PPR096_RFL_Cl87 (scaffold_13850_ref0032247),
Lp_PPR012_RFL_Cl87 (scaffold_5310_ref0017452), Lp_PPR011_RFL_Cl87 (scaffold_5948_ref0038870),
Lp_PPR018_RFL_Cl87 (scaffold_8666_ref0039153), Lp_PPR021_RFL_Cl87 (scaffold_10202_ref0031571),
Lp_PPR060_RFL_Cl87 (scaffold_2792_ref0011578), Lp_PPR010_RFL_Cl87 (scaffold_15350_ref0024553),
Lp_PPR076_Cl87 (scaffold_14826_ref0021051), Tri_Triticum_urata_23631_TRIUR3_28808_TRIUR3_28808-
T1, Tri_Triticum_urata_1634_TRIUR3_01213_TRIUR3_01213-T1,
Tri_Triticum_urata_9118_TRIUR3_09374_TRIUR3_09374-T1,
Tri_Triticum_urata_4512_TRIUR3_05631_TRIUR3_05631-T1,
Tri_Triticum_urata_6334_TRIUR3_07370_TRIUR3_07370-T1,
Tri_Triticum_urata_1367_TRIUR3_01813_TRIUR3_01813-T1,
Tri_Triticum_urata_1586_TRIUR3_01814_TRIUR3_01814-T1,
Tri_Triticum_urata_18570_TRIUR3_20712_TRIUR3_20712-T1,
Tri_Triticum_urata_1289_TRIUR3_01807_TRIUR3_01807-T1,
Ory_Oryza_sativa_Japonica_20731_LOC_Os04g28234.2,
Ory_Oryza_sativa_Japonica_20737_LOC_Os04g28300.1, Hor_Hordeum_vulgare_14747_MLOC_59638.1,
Hor_Hordeum_vulgare_8956_MLOC_43104.1, Hor_Hordeum_vulgare_3429_MLOC_18343.1,
Hor_Hordeum_vulgare_9451_MLOC_44563.1, Hor_Hordeum_vulgare_20246_MLOC_71447.1,
Hor_Hordeum_vulgare_14547_MLOC_59233.1,
Bra_Brachypodium_distachyonCDS_11264_BRADI2G38400.1,
Bra_Brachypodium_distachyonCDS_11270_BRADI2G38460.1,
Bra_Brachypodium_distachyonCDS_11269_BRADI2G38450.1,
Bra_Brachypodium_distachyonCDS_11268_BRADI2G38440.1,
Bra_Brachypodium_distachyonCDS_11233_BRADI2G38110.1,
Bra_Brachypodium_distachyonCDS_11234_BRADI2G38120.1,
Bra_Brachypodium_distachyonCDS_11301_BRADI2G38760.1,
Bra_Brachypodium_distachyonCDS_20237_BRADI4G05910.1,
Bra_Brachypodium_distachyonCDS_1473_BRADI1G14190.1

**Clade 3:** Ara_Arabidopsis_thalians_3154_AT1G31840.1, Ara_Arabidopsis_thalians_5239_AT1G63230.1,
Ara_Arabidopsis_thalians_5326_AT1G64100.2, Ara_Arabidopsis_thalians_13668_AT3G22470.1,
Ara_Arabidopsis_thalians_1257_AT1G12700.1, Ara_Arabidopsis_thalians_1264_AT1G12775.1,
Ara_Arabidopsis_thalians_1215_AT1G12300.1, Ara_Arabidopsis_thalians_1248_AT1G12620.1,
Ara_Arabidopsis_thalians_13034_AT3G16710.1, Ara_Arabidopsis_thalians_24320_AT5G41170.1,
Ara_Arabidopsis_thalians_19449_AT4G26800.1, Ara_Arabidopsis_thalians_5174_AT1G62670.1,
Ara_Arabidopsis_thalians_5198_AT1G62910.1, Ara_Arabidopsis_thalians_5221_AT1G63080.1,
Ara_Arabidopsis_thalians_5199_AT1G62914.1, Ara_Arabidopsis_thalians_5258_AT1G63400.1,
Ara_Arabidopsis_thalians_5201_AT1G62930.1, Ara_Arabidopsis_thalians_5227_AT1G63130.1,
Ara_Arabidopsis_thalians_5229_AT1G63150.1, Ara_Arabidopsis_thalians_5167_AT1G62590.1,

Ara_Arabidopsis_thalians_5251_AT1G63330.1, Ara_Arabidopsis_thalians_5220_AT1G63070.1, Ara_Arabidopsis_thalians_5175_AT1G62680.1, Ara_Arabidopsis_thalians_5380_AT1G64583.1, Ara_Arabidopsis_thalians_5379_AT1G64580.1, Ara_Arabidopsis_thalians_593_AT1G06580.1, Ara_Arabidopsis_thalians_22458_AT5G16640.1, Ara_Arabidopsis_thalians_5179_AT1G62720.1

**Clade 4:** Set_Setaria_italica_29736_Si034189m, Set_Setaria_italica_34526_Si039892m, Set_Setaria_italica_4386_Si005166m, Set_Setaria_italica_23717_Si027424m, Set_Setaria_italica_10640_Si012426m, Sor_Sorghum_bicolor_6473_Sb02g005000.1, Sor_Sorghum_bicolor_6405_Sb02g004520.1, Sor_Sorghum_bicolor_6407_Sb02g004530.1, Sor_Sorghum_bicolor_31396_Sb09g030360.1, Sor_Sorghum_bicolor_13027_Sb03g030790.1, Sor_Sorghum_bicolor_32635_Sb10g009870.1, Sor_Sorghum_bicolor_25115_Sb07g007630.1, Sor_Sorghum_bicolor_18788_Sb05g000986.1, Sor_Sorghum_bicolor_18957_Sb05g002320.1, Sor_Sorghum_bicolor_18925_Sb05g002040.1, Sor_Sorghum_bicolor_18992_Sb05g002620.1, Sor_Sorghum_bicolor_18961_Sb05g002360.1, Sor_Sorghum_bicolor_18962_Sb05g002370.1, Sor_Sorghum_bicolor_18956_Sb05g002310.1, Sor_Sorghum_bicolor_18958_Sb05g002330.1, Sor_Sorghum_bicolor_18945_Sb05g002220.1, Sor_Sorghum_bicolor_18948_Sb05g002250.1, Hor_Hordeum_vulgare_7505_MLOC_37353.2, Tri_Triticum_urata_13141_TRIUR3_16299_TRIUR3_16299-T1, Tri_Triticum_urata_14502_TRIUR3_16298_TRIUR3_16298-T1, Tri_Triticum_urata_996_TRIUR3_01266_TRIUR3_01266-T1,, Tri_Triticum_urata_1243_TRIUR3_01264_TRIUR3_01264-T1, Tri_Triticum_urata_33118_TRIUR3_34377_TR, UR3_34377-T1, Tri_Triticum_urata_1366_TRIUR3_01263_TRI, R3_01263-T1, Tri_Triticum_urata_32436_TRIUR3_33751_TR, UR3_33751-T1, Zea_Zea_mays_66381_GRMZM2G416201_P0, Zea_Zea_mays_67949_GRMZM2G431850_P01, Ory_Oryza_sativa_Japonica_48640_LOC_Os10g35640.1, Ory_Oryza_sativa_Japonica_48603_LOC_Os10g35260.1, Ory_Oryza_sativa_Japonica_48601_LOC_Os10g35240.2, Ory_Oryza_sativa_Japonica_48621_LOC_Os10g35436.1, Ory_Oryza_sativa_Japonica_48600_LOC_Os10g35230.1, Ory_Oryza_sativa_Japonica_48655_LOC_Os10g35790.1, Ory_Oryza_sativa_Japonica_39235_LOC_Os08g15000.1, Ory_Oryza_sativa_Japonica_37932_LOC_Os08g01870.1, Ory_Oryza_sativa_Japonica_37911_LOC_Os08g01650.1, Ory_Oryza_sativa_Japonica_37910_LOC_Os08g01640.1

# Chapter 3

# Genotyping by sequencing of a perennial ryegrass (*Lolium perenne* L.) population segregating for cytoplasmic male sterility restoration

**Timothy Sykes[1], Steven Yates[1] and Bruno Studer[1]***

[1] Institute of Agricultural Sciences, Molecular Plant Breeding, ETH Zurich, 8092 Zurich, Switzerland.

*Corresponding author: bruno.studer@usys.ethz.ch

## Abstract

Perennial ryegrass (*Lolium perenne* L.) is the world's most important forage crop, accounting for 50% of all forage and turf grass production. With current efforts to introduce hybrid breeding schemes in perennial ryegrass, a pollination control mechanism that can ensure hybrid seed purity is required. Cytoplasmic male sterility (CMS) is such a mechanism and has been described in perennial ryegrass with both sterile and fertility-restoring plants being identified. However, the loci underpinning these phenotypes have yet to be elucidated. What is required by plant breeders is a set of molecular markers that can accurately predict the restoring capabilities of individual plants being introduced into hybrid breeding schemes.

To identify molecular markers for fertility restoration genes, a genotyping by sequencing (GBS) approach was applied to a population of plants segregating for fertility restoration. By genotyping a total of 1,103 plants, 44 polymorphic markers denoting two strong and two possible quantitative trait loci (QTL) for fertility restoration were identified. These data suggested that two loci containing sterility-maintaining genes are present, with the possible confounding effects of additional loci. We were also able to demonstrate the effectiveness of using several genome reference assemblies for the analysis of GBS data, to capture as many informative polymorphic sites as possible.

The polymorphic sites presented here can now be used for further genotyping to generate higher resolution QTL, to both identify the likely genes responsible for fertility restoration and provide breeders with markers to track restoration genes in their breeding populations.

**Key words:** Cytoplasmic male sterility (CMS), Hybrid breeding, Pentatricopeptide repeat (PPR) proteins, Perennial ryegrass (*Lolium perenne* L.), Restoration of fertility, Genotyping by sequencing (GBS)

## Introduction

Grasslands are important agro-ecosystems; worldwide, they account for 80% of milk production and 70% of meat production [1]. Perennial ryegrass (*Lolium perenne* L.) is a major component of temperate grassland systems, with an annual seed production output of 90,000 metric tonnes. It accounts for almost 50% of the total forage and turf grass production, making it the most important grass species both in Europe and worldwide [2]. In forage grasses, biomass is the primary yield target, but despite intensive breeding efforts over the last decades, increases in biomass yield are below that of other major crop species [3].

One of the most promising options to improve this below-average yield increase is the implementation of hybrid breeding. By exploiting the phenomenon of heterosis, hybrid breeding has significantly contributed to major yield increases in several important crop species including rice (*Oryza sativa* L.), maize (*Zea mays* L.) and rapeseed (*Brassica napus* L.) [4, 5]. One of the major challenges faced by plant breeders trying to employ hybrid breeding in perennial ryegrass is the need for a pollination control strategy that would allow efficient production of hybrid seed on a commercial level. One mechanism to control pollination, successfully applied for hybrid seed production in other plant species, is cytoplasmic male sterility (CMS) [6-11].

CMS in flowering plants is characterised by a maternally-inherited inability to produce functional pollen [12]. This trait allows breeders to ensure complete outcrossing between inbred lines, assuming that one line is affected by CMS. This functional defect is often attributed to aberrant transcripts originating from the mitochondria, with these CMS causing transcripts usually consisting of novel chimeric open reading frames containing part of a functional mitochondrial gene [12].

The CMS phenotype is often restored through the action of nuclear-derived RNA binding proteins [13]. These RNA binding proteins are members of the huge family of pentatricopeptide repeat (PPR) proteins that are particularly numerous in land plants, with 450 PPRs identified in Arabidopsis (*Arabidopsis thaliana* L.) and 477 in rice [14-18]. A subgroup of PPRs is specifically linked to fertility restoration of CMS: the Restorer of Fertility like PPR (RFL) proteins. This group is identified by their relative homology from within the PPR family [19]. Around ten to 30 members of RFL encoding genes are found in plant genomes [20, 21]. Identifying genetic markers for CMS restorer (*Rf*) genes is important as it allows plant breeders to easily track *Rf* genes through their breeding material, thus saving time and speeding up the breeding process.

One way to generate genome-wide marker data is genotyping by sequencing (GBS), a low cost, high throughput genotyping method that has been used for marker discovery in multiple crop species [22]. GBS utilises genome complexity reduction and high throughput sequencing and hence does not require previous genomic sequence information to generate markers that can be used for marker-assisted breeding [23]. GBS has become popular due to the relative simplicity of generating barcoded sequencing libraries that can then be combined into a single library for sequencing. Here, we utilise the power of GBS to identify quantitative trait loci (QTL) for CMS restorer genes. Specifically, we aimed at i) applying GBS to a perennial ryegrass population segregating for fertility restoration, ii) establishing an efficient pipeline for marker discovery on the basis of multiple reference sequences, iii) associating the marker genotypes with the sterility and fertility phenotypes for QTL identification and iv) locating the QTL for fertility restoration in the perennial ryegrass genome.

## Methods

### Plant Material

To identify QTL for fertility restoration by GBS, a population of plants segregating for fertility restoration was used. This population was developed and established at Norddeutsche Pflanzenzucht Hans-Georg Lembke KG (NPZ) on the island of Poel, Germany. The plant material originated from a hybrid breeding program utilising a previously identified CMS system [24]. A cross between a CMS mother plant and a maintainer (non-CMS) father plant yielded a population segregating for fertility restoration. This father plant was then self-fertilised yielding 211 offspring, 74 of which were crossed to a population of CMS mother plants yielding seven further segregating populations (Figure 1). The initial mother plant and subsequent mother plants used were all from the same population of plants. Phenotyping was conducted in the field on Poel, Germany over two years (2014, 2015). Male sterility/fertility was visually scored as the presence or absence of anthers in mature flowers.

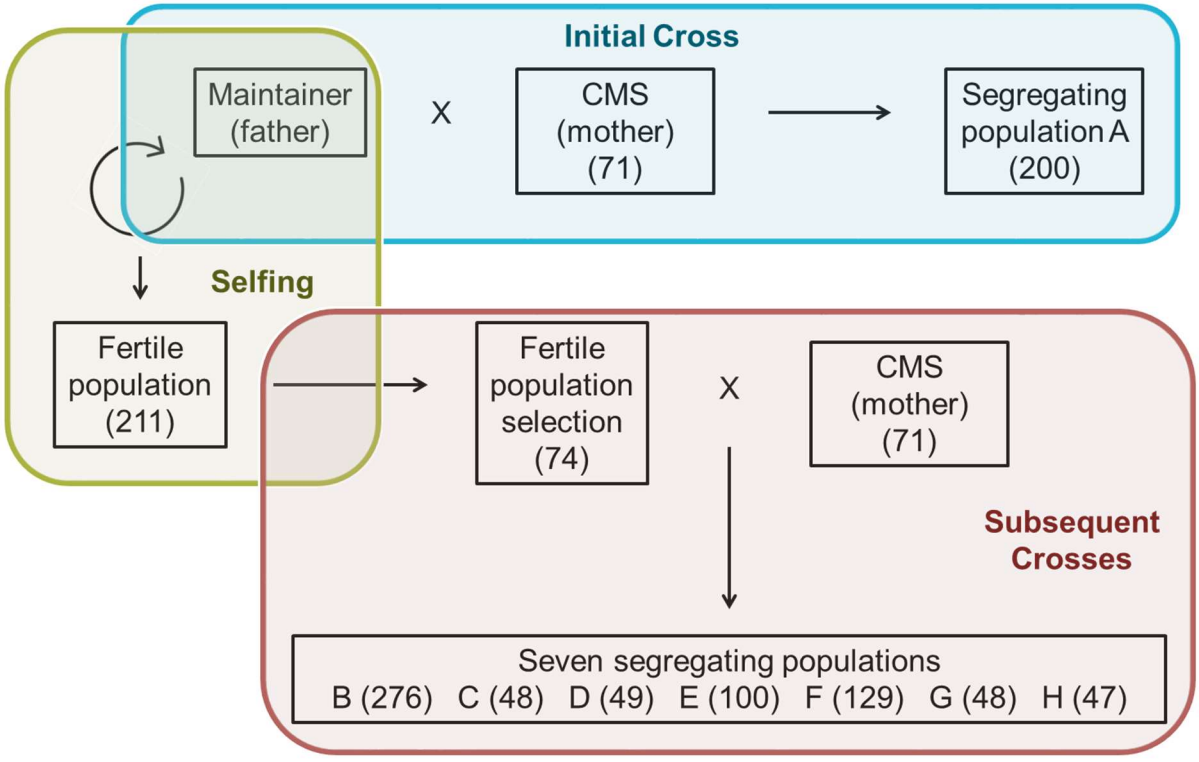



**Figure 1. Schematic showing the design of the plant populations segregating for fertility restoration.** The initial cross is shown in blue, the selfing of the father plant in green and subsequent crosses to generate further segregating populations in red. Numbers in brackets indicate the number of individual plants within each population.

**Sequencing library preparation and DNA sequencing**

Genomic DNA from all individuals was isolated from powered freeze-dried leaf tissue using either the Omega Mag-Bind® Plant DNA DS 96 Kit (Omega Bio-tek, Inc., USA) on the Kingfisher MagMAX™ Magnetic Particle Processor (Thermo Fisher Scientific Inc., USA) or by a manual isopropanol precipitation method. DNA quality and quantity were assessed using a QIAxpert (QIAGEN Inc., USA) if the samples were extracted with the Omega kit or using qPCR if manually extracted. The GBS method of Elshire et al. [25] was followed to generate PstI-digested DNA fragments for sequencing on the Illumina HiSeq® 2500 platform (Illumina Inc., USA). A set of 192 barcodes were generated using the Deena Bioinformatics online tool for the Pst1 enzyme (Deena Bioinformatics, Netherlands) and synthesized by Microsynth (Microsynth AG, Switzerland). Each adaptor contained a three base overhang for ligation with PstI digested genomic DNA. Six sequencing libraries were generated containing 192 pooled barcoded samples per library. Each library was sequenced on a single Illumina HiSeq® 2500 lane with three libraries at 100bp single end sequencing and the other three at 125bp paired end sequencing.

**GBS data processing**

Sequencing reads were demultiplexed using Sabre (v1.000, https://github.com/najoshi/sabre), allowing no mismatches. The data were then analysed using an in-house GBS pipeline. Briefly, the reads were trimmed to 100bp and the frequency (counts) of unique sequences (tags) was obtained using *Bash* commands. The count data were then summarised for each population and the unique sequences were back-transformed to a Fastq file using *Perl* scripts. The resulting Fastq file was then mapped to the perennial ryegrass draft genome sequence [26] using bowtie (v0.12.7) [28] with "—best –strata" and a maximum of 2 alignments (-m 2). The resulting Sam file was filtered for alignments to the genome and headers omitted using "*grep*" commands. The prior generated population count file was then filtered, using the Sam file, for tags which were aligned to the genome (to reduce memory requirements in *R*) using *Perl*.

The resulting trimmed count and Sam files were processed in *R*. Numerical factors were set to constrain genotyping based upon the population design: the basal ploidy level of the genotypes used (ploidy, 2), the maximum number of alleles (4), minor allele frequency (minAF, 100) and the minimum allele count (minAC, 8). To identify all unique positions, the direction (Flag), location (Ref) and position (Pos) data, from the Sam file, were concatenated together to produce a unique position identifier (Upos). To eliminate low coverage sites, only Upos with at least one tag greater than the minor allele frequency (minAF, 100) were used. Upon iteration through the potentially informative loci, first all tags with the corresponding Upos are retrieved. From the resulting tags, tags occurring at a frequency greater than 5% were retained. An Upos was considered polymorphic if the number of Itags was greater than one. If any genotype had more unique Itags than its ploidy (2) or if the number of observed alleles (Itags) exceeded the maximum number of alleles possible (in the population), then the Upos was discarded.

For genotype calling, all polymorphic nucleotide sites across Itags, termed informative sites (Isites), were identified. Once Isites were determined, each unique tag was allocated/changed to a corresponding Itag, based on all Isite positions. If a tag did not fit perfectly to any Itag, its tags count data were omitted. For genotyping, if a genotype had two alleles observed from the Itags then it was considered heterozygous (for those alleles). Homozygotes were genotyped when an individual had a single Itag with greater than 8 (minAC)

frequency (>99% confidence). A chi-squared test (bonferroni correction) was performed to identify any heterozyogous tags with a significant correlation to the sterile/fertile phenotype.

**Marker mapping**

All 100 bp tags containing the markers that passed filtering were aligned against the scaffolds incorporated into the ultra-high density genetic linkage map of perennial ryegrass generated by Velmurugan et al. [26] using BLASTn analysis [27]. Top hits with an E-value of less than 1.00E-05 were extracted, providing linkage group (LG) and genetic distance (in centimorgan, cM) information for all markers with a significant BLAST hit.

**Synteny-based marker mapping using the perennial ryegrass GenomeZipper**

Markers with a significant association to the sterility/fertility phenotype that did not have a significant BLAST hit to the genetic map were positioned through the use of the perennial ryegrass GenomeZipper [29]. The significant tag containing scaffolds were compared against the rice (IRGSP-1.0), sorghum (*Sorghum bicolor* L., v2) and Brachypodium (*Brachypodium distachyon* L., v1.0) genomes using the online BLAST function provided by EnsemblPlants (http://plants.ensembl.org). BLAST results for the significant scaffolds against the three genomes used to develop the perennial ryegrass GenomeZipper were consolidated to identify regions of conserved gene order within the scaffolds, and single representative genes were used to locate that scaffold on the zipper. Scaffolds were only considered as being "placed" on the zipper if all hits from the three genomes used to make up the zipper (rice, Brachypodium and sorghum) were consistent (or there were at least two hits in agreement if there was no BLAST result for one species).

**Quantitative Trait Loci identification**

As the overlap of generated marker position data (LG, cM), established on the basis of the ultra-high density genetic linkage map and the perennial ryegrass GenomeZipper, was not absolute, an alternative method for QTL localisation was sought. Scaffolds containing significant tags were compared to the high-quality genome assembly of the Italian ryegrass (*Lolium multiflorum* Lam.) genotype 'Rabiosa' (Molecular Plant Breeding, ETH Zurich, Switzerland, unpublished), using BLAST analysis. Top hits were extracted to identify the corresponding Rabiosa scaffolds. As the Rabiosa scaffolds were much longer than the scaffolds present in the other perennial ryegrass genome assemblies (N50=3.3Mb), the identified scaffolds contained numerous genes, allowing areas of conserved gene order in the rice genome [30] to be identified.

## Results

### Phenotyping of fertility restoration

All of the 1,465 plants evaluated for fertility restoration were either male sterile or fertile and no intermediate phenotypes were observed. Only seven plants were not scored identically over the two years and therefore discarded from further analysis. A set of 1,152 (6 x 192) samples were selected for GBS with 299 showing a fertile and 804 a sterile phenotype, giving an overall restoration rate (percentage of fertile plants) of 27.1%. Restoration rates for the segregating sub populations shown in figure 1 were; A: 23.1%, B: 6.5%, C: 25.0%, D: 10.4%, E: 39.1%, F: 23.1%, G: 25.0% and H: 21.7%.

### Genotyping by sequencing, identification of SNP markers significantly associated with fertility restoration and maker localisation in the perennial ryegrass genome

Illumina sequencing reads from the 1,152 barcoded samples were mapped to the draft assembly of the perennial ryegrass genome reported by Velmurugan et al. [26], hereafter referred to as the "Teagasc" genome assembly. After filtering, 1,620 100bp informative tags containing one or more SNP markers were identified across the perennial ryegrass genome. Of these, 1,211 contained a single SNP, 389 two SNPs and 20 three or more SNPs. Tags with several SNPs were treated as haplotypes. A chi-square test of the SNP and the phenotypic data identified 13 loci to be significantly ($P \leq 0.05$) associated with fertility restoration using a Bonferonni corrected logarithm of the odds (LOD) threshold of 4.51 (Figure 2).

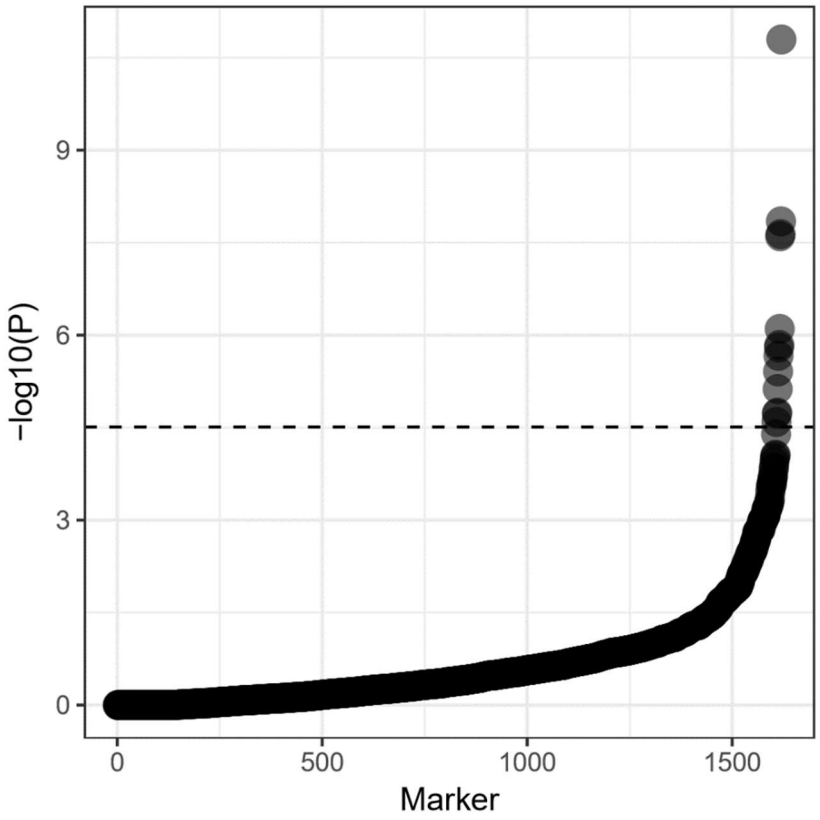



**Figure 2. SNP markers significantly associated with fertility restoration.** The y-axis shows the logarithm of the odds (LOD, -log10(P)) scores for all 1,620 markers (x-axis) identified using genotyping by sequencing. Markers are ordered by LOD scores, from smallest to highest. The dotted line at LOD=4.51 indicates the Bonferroni corrected significance threshold of the chi-square test.

The genotyping results of the13 significant tags were further analysed to ascertain which SNP variant was associated with the restored fertile phenotype using a chi-square test (Supplementary Table 1). For the majority of the 13 tags, one homozygous SNP variant was associated with the sterility phenotype, the heterozygous and the homozygous complementary SNP variant being significantly ($P \leq 0.05$) linked to the restored fertile phenotype.

To position the GBS markers on the perennial ryegrass genome, LG and map position of the 1,620 Teagasc scaffolds containing GBS makers were inferred using to the ultra-high density genetic linkage map reported by Velmurugan et al. [26]. From the total of 1,620 scaffolds, 959 were assigned to genome positions. The position of the markers significantly associated with fertility restoration indicated two potential QTL on LG3 and LG6 (Figure 3A).

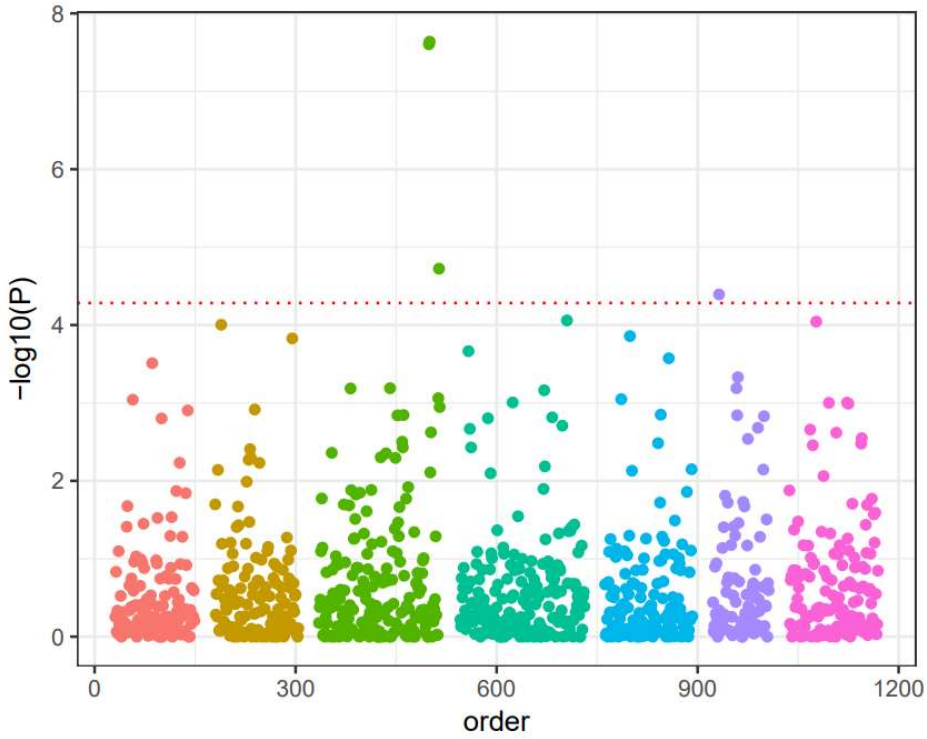
**Validation of SNP discovery, association analysis and genome positioning**

In order to assess the reliability of SNP discovery, the GBS data analysis was repeated using the genome assembly from Bryne et al, 2015 [31], hereafter referred to as the "Aarhus" genome assembly. This identified 935 informative tags with 691 containing a single SNP, 245 two SNPs and nine three or more SNPs. In total, 14 informative tags were significantly associated (LOD value above 4.52 after Bonforonni correction) with fertility restoration, eight of which were also found within the original data set generated using the Teagasc assembly (Figure 3B, Table 1). The scaffolds from the Aarhus genome assembly containing the 14 significant tags were extracted and their corresponding Teagasc scaffolds identified using BLAST. Of the six unique significant tags, four were anchored to the ultra-high density map (Table 1).

**Positioning of significant markers using the perennial ryegrass GenomeZipper**

Scaffolds from both the Teagasc and Aarhus genome assemblies that contained significant tags were used for genome synteny analysis using the GenomeZipper approach [29]. Based on genome synteny to barley (*Hordeum vulgare* L.) and the model grass genome Brachypodium as well as rice and sorghum, a total of eleven out of the 19 significant tags could be assigned to a perennial ryegrass LG and a location (in cM) within that LG. Of these eleven tags, eight were also positioned on the Teagasc map with the remaining three uniquely positioned via the GenomeZipper approach. Only one marker (RfMkr12), present in both data sets, was mapped differently with the remainder showing similar or identical results (Table 1). To further examine the compatibility of LG and cM information generated using these two approaches, the Teagasc scaffolds containing the 495 markers used to create the perennial ryegrass GenomeZipper [29] were identified using BLAST. Of these scaffolds, 163 were anchored into the ultra-high density Teagasc map (Supplementary Table 2). Comparison of the location (LG and cM) of these markers in the ultra-high density map with their corresponding position inferred by the perennial ryegrass GenomeZipper revealed that 87% of the markers were assigned to the same LG. This comparison also revealed that the order of the markers within each LG was not well conserved (Table 1).

**A**




**B**

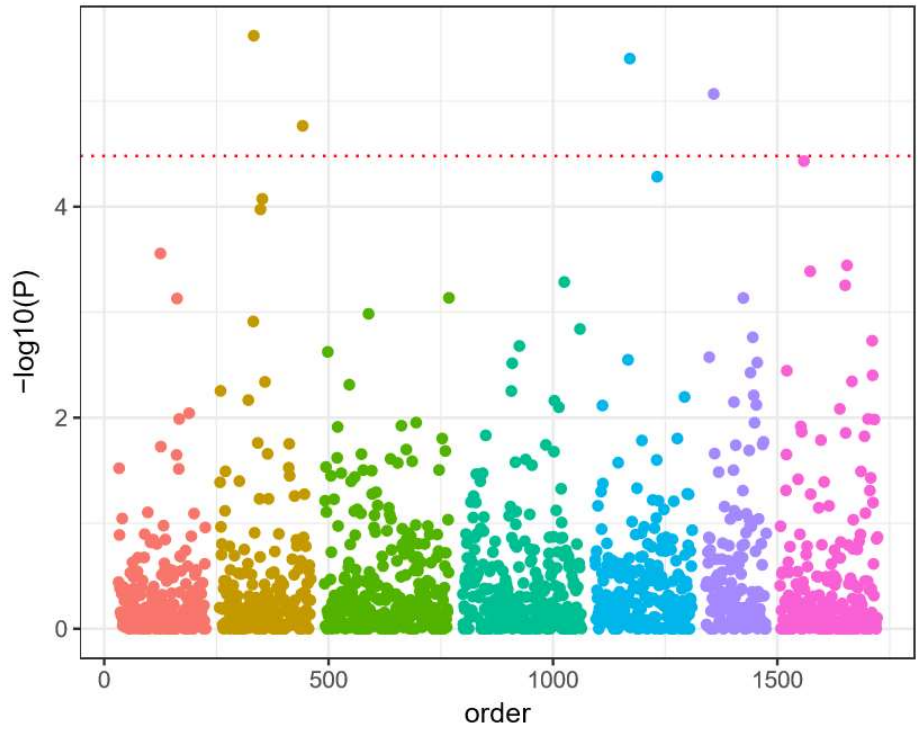



**Figure 3. Identification of QTL for fertility restoration along the seven linkage groups (LGs) representing the perennial ryegrass genome.** The Manhatten plots illustrate the marker significance (-log10(P), y-axis) and the genome location (genome position, x-axis). Markers from the seven LGs of perennial ryegrass are depicted with different colours. The dotted red line indicates the significance threshold. **A.** Data was generated using the Teagasc genome assembly. **B.** Data was generated using the Aarhus genome assembly.

**Consolidation of the marker locations via Italian ryegrass and rice**

To further consolidate the genome locations of the identified significant markers, the scaffolds containing these markers (from both assemblies) were used to identify the corresponding Rabiosa scaffolds using BLAST (Table 1). The high quality draft assembly of Rabiosa was chosen as it is more complete than either the Teagasc or Aarhus genome assembly as well as having much larger N50 value (Table 2). This allowed the identification of a sufficient number of genes per Rabiosa scaffold, to identify homologs in rice [30] and consequently areas of conserved gene order. This facilitated the consolidation of several markers revealing possible QTL on rice chromosomes 1, 2, 3 and 11 (Table 1). These correspond, via synteny, to perennial ryegrass LG3, LG6, LG4 for rice chromosomes 1, 2 and 3, respectively, with chromosome 11 having no significant synteny with the perennial ryegrass genome.

**SNP discovery using a high-quality genome assembly of Italian ryegrass**

Given the success of using Rabiosa scaffolds to identify regions of synteny within the rice genome, the original GBS sequencing reads were aligned to the Rabiosa genome and the GBS pipeline run against this data set. This identified 3,349 informative tags of which 1,432 contained a single SNP, 652 two SNPs, 939 three of more SNPs and 327 indels (DNA insertions or deletions). After statistical analysis, 32 tags with a significant association (LOD value above 4.7 after Bonforonni correction) to the phenotype were identified with six of these also present in the original data set given in table 3. Again conserved gene order was used to identify areas of synteny with the rice genome, revealing two likely QTL on rice chromosome one and six (corresponding to perennial ryegrass LG3 and LG7) and two further possible QTL on rice chromosome three and four (corresponding to perennial ryegrass LG4 and LG2) (Table 3).

**Table 1.** Genome position of the 19 significant tags identified using both the Aarhus and the Teagasc draft genome assembly as a reference for SNP discovery. Genome positions were inferred according to the high density linkage map by Velmurugan et al. [26] and compared to the perennial ryegrass GenomeZipper. Corresponding scaffolds from the Rabiosa assembly and zones of synteny with the rice genome are also shown.

| Marker | Marker Name Aarhus | Marker Name Teagasc | LOD | Map | | Zipper | | Rabiosa scaffold | Rice | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | LG | cM | LG | cM | | Chr | Mb |
| RfMrk10 | X16_scaffold_18576.ref0034088_7680_59 | X16_scaffold24527.size9309_6847_59 | 5.1 | - | - | - | - | scaffold11329 | 1 | 22.8-23.9 |
| RfMrk1 | X16_scaffold_26970.ref0002691_437_90 | X16_scaffold5846.size41115_23349_90 | 10.8 | - | - | 2 | 79 | scaffold1368 | 1 | 40.2-40.6 |
| RfMrk15 | - | X16_scaffold7199.size36343_28357_32 | 7.6 | 3 | 131 | 3 | 60 | scaffold2431 | 1 | 41.2-42.3 |
| RfMrk16 | - | X0_scaffold7199.size36343_28329_97 | | | | | | | | |
| RfMrk13 | X0_scaffold_100.ref0001926_3476_96 | X0_scaffold24681.size9221_3278_96 | 4.6 | - | - | - | - | scaffold6483 | 2 | 0.3-0.6 |
| RfMrk12 | X0_scaffold_12379.ref0036453_1034_37 | X16_scaffold10951.size24828_22867_37 | 4.7 | 3 | 154 | 6 | 50 | scaffold194 | 2 | 1.2-1.4 |
| RfMrk8 | X16_scaffold_1306.ref0035694_89626_53 | - | 5.2 | 7 | 42 | 7 | 63 | scaffold5171 | 2 | 3.3-3.6 |
| RfMrk9 | X0_scaffold_1306.ref0035694_89589_85 | - | | | | | | | | |
| RfMrk17 | - | X16_scaffold1060.size83129_74273_100 | 6.1 | - | - | 6 | 42 | scaffold8015 | 2 | 3.7-4.0 |
| RfMrk19 | - | X16_scaffold28713.size6917_5787_75 | 4.4 | 6 | 10 | 6 | 91 | scaffold3413 | 2 | 34.9-35.1 |
| RfMrk18 | - | X16_scaffold6840.size37579_20207_36_55 | 5.8 | - | - | - | - | scaffold4840 | 3 | 16.3-17.4 |
| RfMrk4 | X0_scaffold_6663.ref0003691_16698_91 | - | 6.0 | 1 | 74 | 1 | 31 | scaffold4016 | 3 | 35.0-35.2 |
| RfMrk5 | X16_scaffold_3876.ref0034878_69365_33_100 | - | 5.8 | - | - | - | - | scaffold3487 | 3 | 35.1-35.3 |
| RfMrk7 | X16_scaffold_10382.ref0044772_9923_83 | X0_scaffold14243.size18503_8618_83 | 5.4 | - | - | - | - | scaffold2279 | 6 | 9.3-11.2 |
| RfMrk14 | X16_scaffold_11858.ref0027935_13246_66 | - | 4.4 | 5 | 64 | - | - | scaffold4584 | 11 | 1.1-1.4 |
| RfMrk6 | X0_scaffold_108.ref0001929_81555_48 | X16_scaffold11408.size23758_9444_48 | 5.7 | - | - | - | - | scaffold1369 | 11 | 3.0-3.2 |
| RfMrk3 | X16_scaffold_17592.ref0036612_6574_56 | X16_scaffold27354.size7554_3084_56 | 7.8 | - | - | 2 | 92 | scaffold1253 | 11 | 3.3-4.7 |
| RfMrk11 | X0_scaffold_17592.ref0036612_6556_63 | X0_scaffold27354.size7554_3066_63 | 4.7 | | | | | | | |
| RfMrk2 | X0_scaffold_7519.ref0016284_33214_93 | - | 8.7 | - | - | - | - | scaffold2008 | 12 | 3.3-3.8 |

**Table 2.** Comparison of the three genome assemblies used for GBS analysis. Lp refers to perennial ryegrass, Lm to Italian ryegrass. (* contains both copies of the diploid genome).

| | Genome Assembly | | |
|---|---|---|---|
| | **Teagasc (Lp)** | **Aarhus (Lp)** | **Rabiosa (Lm)** |
| **Genome size (Gb)** | 1.1 | 1.2 | 4.5* |
| **N50 (Mb)** | 0.025 | 0.07 | 3.3 |
| **Markers identified** | 1620 | 935 | 3349 |
| **Significant markers** | 13 | 14 | 32 |

**Table 3.** Significant markers identified within the Rabiosa genome assembly showing LOD score and location of the area of synteny within the rice genome (in mega base pair, Mb).

| **Marker** | **Marker Name Rabiosa** | **LOD** | **Rabiosa Scaffold** | **Rice** | |
|---|---|---|---|---|---|
| | | | | **Chromsome** | **Mb** |
| RfMkr20 | X16_scaffold1422_302531_59_67_84 | 5.0 | scaffold1422 | 1 | 8.2 |
| RfMkr21 | X16_scaffold156_18745192_4_64_73 | 5.2 | scaffold156 | 1 | 8.6 |
| RfMkr22 | X16_scaffold1334_4266632_29_93 | 7.7 | scaffold1334 | 1 | 24.3 |
| RfMkr23 | X0_scaffold4067_2010133_7 | 5.0 | scaffold4067 | 1 | 33.4 |
| RfMkr24 | X0_scaffold4067_2306689_5 | 5.5 | | | |
| RfMkr25 | X0_scaffold1368_1248705_14 | 9.2 | scaffold1368 | 1 | 40.5 |
| RfMkr26 | X0_scaffold2431_7110718_10 | 5.9 | scaffold2431 | 1 | 41.1 |
| RfMkr27 | X16_scaffold2431_7110704_77 | 7.1 | | | |
| RfMkr28 | X0_scaffold11432_3235398_5_39 | 5.2 | scaffold11432 | 1 | 42.6 |
| RfMkr29 | X0_scaffold5137_93087_19_37_38 | 5.7 | scaffold5137 | 2 | 33.7 |
| RfMrk6 | X0_scaffold9245_42380_48 | 5.7 | scaffold9245 | 3 | 6.0 |
| RfMrk18 | X16_scaffold715_501723_36_55 | 5.8 | scaffold715 | 3 | 17.2 |
| RfMkr30 | X16_scaffold3321_100994_51_57_78 | 5.9 | scaffold3321 | 3 | 29.3 |
| RfMrk5 | X16_scaffold3487_623356_33_100 | 5.8 | scaffold3487 | 3 | 35.2 |
| RfMkr31 | X0_scaffold2769_852601_95 | 5.2 | scaffold2769 | 4 | 25.7 |
| RfMkr32 | X0_scaffold1758_4582452_73 | 5.0 | scaffold1758 | 4 | 28.2 |
| RfMkr33 | X0_scaffold4842_870704_21_24_26_31 | 5.6 | scaffold4842 | 4 | 33.9 |
| RfMkr34 | X0_scaffold7768_85789_62_79_indel | 6.7 | scaffold7768 | 6 | 1.4 |
| RfMrk7 | X16_scaffold2279_1867119_83 | 5.4 | scaffold2279 | 6 | 10.7 |
| RfMkr35 | X0_scaffold2279_1867121_16 | 5.7 | | | |
| RfMkr36 | X0_scaffold1777_314768_45_indel | 5.4 | scaffold1777 | 6 | 24.1 |
| RfMrk9 | X16_scaffold1913_2717258_85 | 5.1 | scaffold1913 | 6 | 28.0 |
| RfMrk8 | X16_scaffold5171_1706723_53 | 5.2 | scaffold5171 | 6 | 28.3 |
| RfMkr37 | X0_scaffold5268_317079_26_37_indel | 5.2 | scaffold5268 | 7 | 24.5 |
| RfMkr38 | X0_scaffold1773_3358245_4_38 | 6.1 | scaffold1773 | 7 | 25.8 |
| RfMkr39 | X0_scaffold1773_4407066_15_54 | 6.7 | | | |
| RfMkr40 | X16_scaffold866_5429328_7_32_66 | 6.0 | scaffold866 | 10 | 20.9 |
| RfMrk3 | X16_scaffold1253_8100548_56 | 7.8 | scaffold1253 | 11 | 4.6 |
| RfMkr41 | X0_scaffold8735_215237_28 | 4.8 | scaffold8735 | 12 | 15.8 |
| RfMkr42 | X0_scaffold54136_6644_69_indel | 5.4 | scaffold54136 | - | - |
| RfMkr43 | X0_scaffold84607_21672_98 | 6.9 | scaffold84607 | - | - |
| RfMkr44 | X16_scaffold131866_3354_97 | 7.6 | scaffold131866 | - | - |

## Discussion

A GBS analysis was successfully used to generate marker data and to identify polymorphic sites linked to the restoration of fertility in CMS-affected perennial ryegrass plants. This approach utilised a large population of 1,103 individual plants, segregating for fertility restoration that were DNA sequenced in barcoded pools. By analysing the generated sequencing data using three different ryegrass genome assemblies (two from perennial and one from Italian ryegrass), we were able to identify more informative markers than by using any single genome assembly publicly available for perennial ryegrass. Through synteny analysis to the rice genome, two strong and two possible QTL linked to the restored fertile phenotype were identified.

### Phenotyping results

From the pattern of roughly 25% restoration frequency observed within the populations segregating for fertility restoration, it can be inferred that two co-dominant loci are most likely responsible for this CMS-restoration system. Analysis of the individual markers linked to the phenotype of interest suggest that they are linked to non-restoring alleles, indicating that sterility may be actively maintained. This would be unique to CMS restoration since in other described CMS systems, restoration is an active process and not the maintenance of sterility. In order to validate these results further, the significant markers need to be genotyped across the whole population to allow pairwise analyses of markers from different QTL to ascertain whether they can, in concert, explain a large portion of observed fertility restoration. As this pattern is not perfectly explained by this hypothesis, with restoration rates varying from 6.5-39.1% within sub-populations, it is possible that other loci may be affecting fertility restoration of CMS affected plants. Also, given that no intermediate phenotypes were observed, it seems likely that a single mechanism is responsible for restoration. These assumptions are corroborated by the results presented here, as we identified two strong and two possible QTL for fertility restoration. However, as attempts to test this hypothesis by looking at the predictive power of these loci in combination were hindered by the high missing value rates of GBS, further genotyping data is required to confirm this hypothesis. All segregating populations present in this study were treated as one large population, as they all originate from a single father plant and a small population of mother plants, restricting the number of possible alleles. Also, attempts to treat sub-populations separately were again hindered by the missing value rate which hampered statistical analysis on smaller populations.

### Effectiveness of GBS as a tool to identify QTL

By using three different genome assemblies for the analysis of the GBS data, the effect of the completeness of a reference DNA sequence used for marker calling on the effectiveness of SNP discovery could be assessed. As expected, the number of identified polymorphic sites increased with the size of the assembly, although this relationship may have an upper limit dictated by the actual genome size. This increase unsurprisingly manifested itself in the discovery of an increased number of sites that can be statistically linked to the phenotype of interest, which in turn informed a better identification of QTL. Interestingly, nearly twice as many polymorphic sites were identified using the Teagasc genome assembly as compared to the Aarhus genome assembly, even though the overall genome size is very similar. Also interesting was that the number of informative sites was roughly the same. This may indicate that although the genomes are similar in size they are not similar in content, which is possible given that they both represent about half of the entire perennial ryegrass genome of 2.3 Gbp. A

comparison to the Rabiosa genome, which covers almost the entire genome, revealed around three times as many markers, despite being the genome of a different, although highly similar, species. Perhaps the greatest benefit of the Rabiosa genome was its high N50 value, which allowed multiple genes to be identified on each scaffold containing a significant tag. This allowed synteny analysis to be achieved via rice, giving a clearer picture of the QTL responsible for fertility restoration. This indicates the importance of the reference genome used for GBS data analysis and also shows the value of using several different genomes in order to identify as many informative molecular markers as possible.

**Marker identification vs. QTL identification**

Although the use of GBS has managed to identify several QTL for fertility restoration, they are reasonably broad QTL and do not allow resolution to the gene level. This is mainly due to the high missing value rates seen when analysing GBS data. The resolution of QTL can be directly linked to the number of recombination events being captured by any given data set [32], in that the more events captured, the greater the resolution. Although GBS is a very powerful tool for identifying polymorphic sites, it is not a particularly effective specific genotyping tool. In other words, what one gains in the total number of sites identified within the tested population, one loses in the representation of each of these sites in any one individual. For example, a missing value rate of 80% indicates that the chance of any individual having a particular marker identified is 20%, so that the chance of any two markers being identified within one single individual is 4%. Each time two markers are represented within one individual some information regarding the frequency of recombination between these two markers is recorded, but with such a low percentage of marker pairs being identified within a single individual, the actual number of these events recorded is relatively low, which is reflected in the broad nature of any identified QTL. What is needed now, within the data set presented here, is to genotype the markers identified as significantly linked to the restored fertile phenotype across the whole population, which will greatly increase the resolution of identified QTL.

Here, we demonstrated the applicability of GBS for the development of marker data useful for trait dissection and QTL analysis. A total of 44 polymorphic sites significantly associated to fertility restoration indicated the presence of two strong and two possible QTL for this trait. The QTL and the underpinning molecular markers reported here can now be further investigated to ascertain their value for marker-assisted breeding in CMS-based hybrid breeding schemes of perennial ryegrass.

# References


1. Wilkins PW, Humphreys MO. Progress in breeding perennial forage grasses for temperate agriculture*. J Agric Sci* 2003, 140:129–150.
2. Breeding Targets for Ryegrass in Europe. *Europeanseed.* Vol. 3 Issue 2 (2016).
3. van der Heijden SAG, Roulund N. Genetic gain in agronomic value of forage crops and turf: a review. *Sustainable use of genetic diversity in forage and turf breeding* (2010), pp. 247–260.
4. Duvick DN. Biotechnology in the 1930s: the development of hybrid maize. *Nature Reviews Genetics* 2001, 2:69-74.
5. Melchinger AE. The International Conference on 'Heterosis in Plants'. *Theoretical and Applied Genetics* 2010, 120:201-203.
6. Havey MJ. The use of cytoplasmic male sterility for hybrid seed production. *Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria* 2004:623-634.
7. Singh SK, Chatrath R, Mishra B: Perspective of hybrid wheat research A review. *Indian Journal of Agricultural Sciences* 2010, 80:1013-1027.
8. Martin AC, Atienza SG, Ramirez MC, Barro F, Martin A: Chromosome engineering in wheat to restore male fertility in the msH1 CMS system. *Molecular Breeding* 2009, 24:397-408.
9. Yuan LP, Virmani SS: Status of hybrid rice research and development. *Hybrid rice Proceedings of an international symposium*, Changsha, China, 6-10 October 1986 1988:7-24.
10. Hybrid rice technology: new developments and future prospects. *Hybrid rice technology: new developments and future prospects* 1994:viii + 296 pp.-viii + 296 pp.
11. Ahokas H. Cytoplasmic male sterility in barley. XV. PI 296897 as a restorer of fertility in msml and msm2 cytoplasms. *Hereditas* 1983, 99(1):157-9.
12. Hanson MR, Bentolila S. Interactions of mitochondrial and nuclear genes that affect male gametophyte development. *Plant Cell* 2004, 16:S154-S169.
13. Schnable PS, Wise RP. The molecular basis of cytoplasmic male sterility and fertility restoration. *Trends Plant Sci* 1998, 3:175 - 180.
14. Castandet B, Araya A. RNA Editing in Plant Organelles. Why Make It Easy? *Biochemistry-Moscow* 2011, 76:924-931.
15. Schmitz-Linneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science* 2008, 13:663-670.
16. Chateigner-Boutin A-L, Small I: Plant RNA editing. *Rna Biology* 2010, 7:213-219.
17. Fujii S, Small I: The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytologist* 2011, 191:37-47.
18. Zehrmann A, Verbitskiy D, van der Merwe JA, Brennicke A, Takenaka M: A DYW Domain-Containing Pentatricopeptide Repeat Protein Is Required for RNA Editing at Multiple Sites in Mitochondria of Arabidopsis thaliana. *Plant Cell* 2009, 21:558-567.
19. Sykes T, Yates S, Nagy I, Asp T, Small I, Studer B. In-silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (Lolium perenne L.). *Genome biology and evolution* 2016, Mar 6:evw047.
20. Fujii S, Bond CS, Small ID. Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proceedings of the National Academy of Sciences of the United States of America* 2011, 108:1723-1728.
21. Andres C, Lurin C, Small ID: The multifarious roles of PPR proteins in plant mitochondrial gene expression. *Physiologia Plantarum* 2007, 129:14-22.
22. Malmberg, M.M., Pembleton, L.W., Baillie, R.C., Drayton, M.C., Sudheesh, S., Kaur, S., Shinozuka, H., Verma, P., Spangenberg, G.C., Daetwyler, H.D. and Forster, J.W., 2017. Genotyping-by-sequencing through transcriptomics: Implementation in a range of crop species with varying reproductive habits and ploidy levels. Plant Biotechnology Journal.
23. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J.M., Pérez-Elizalde, S., Beyene, Y. and Dreisigacker, S., 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science.
24. Islam MS, et al. 2014. Genetics and biology of cytoplasmic male sterility and its applications in forage and turf grass breeding. Plant Breeding. 133(3):299-312.

25. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6(5), p.e19379.
26. Velmurugan J, Mollison E, Barth S, Marshall D, Milne L, Creevey CJ, Lynch B, Meally H, McCabe M, Milbourne D. An ultra-high density genetic linkage map of perennial ryegrass (Lolium perenne) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly. Annals of botany. 2016 Jun 6;118(1):71-87.
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology, 215(3), pp.403-410.
28. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology, 10(3), p.R25.
29. Pfeifer M, et al. 2013. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. Plant physiology. 161(2):571-582.
30. Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C., Iwamoto, M., Abe, T. and Yamada, Y., 2013. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant and Cell Physiology, 54(2), pp.e6-e6.
31. Byrne, S.L., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., Mayer, K., Campbell, J.D., Czaban, A., Hentrup, S. and Panitz, F., 2015. A synteny-based draft genome sequence of the forage grass Lolium perenne. The Plant Journal, 84(4), pp.816-826.
32. Complex Trait Consortium, 2003. The nature and identification of quantitative trait loci: a community's view. Nature reviews. Genetics, 4(11), p.911.

## Supplementary Data

**Supplementary Table 1.** Results of chi-square test for all significant tags from the Teagasc assembly showing all results for each genotyping call against both phenotypes. Significant *P*-values above 0.05 are highlighted in yellow.

| **RfMkr1** | C/C | | | C/G | | | G/G | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 14.3% | 16.9% | 0.5842 | 9.3% | 16.9% | 0.0067 | 34.5% | 16.9% | 1.15E-05 |
| Sterile | 85.7% | 83.1% | 0.8222 | 90.7% | 83.1% | 0.2654 | 65.5% | 83.1% | 0.0715 |

| **RfMkr12** | C/C | | | C/T | | | T/T | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 26.9% | 15.9% | 7.82E-05 | 8.1% | 15.9% | 0.0011 | 10.0% | 15.9% | 0.6117 |
| Sterile | 73.1% | 84.1% | 0.1157 | 91.9% | 84.1% | 0.1935 | 90.0% | 84.1% | 0.8398 |

| **RfMkr3** | C/C | | | C/T | | | T/T | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 30.2% | 15.3% | 1.16E-04 | 11.2% | 15.3% | 0.0585 | 0.0% | 15.3% | 0.1793 |
| Sterile | 69.8% | 84.7% | 0.1320 | 88.8% | 84.7% | 0.4595 | 100.0% | 84.7% | 0.5996 |

| **RfMkr15** | G/G | | | G/T | | | T/T | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 13.2% | 15.7% | 0.4028 | 11.7% | 15.7% | 0.0784 | 35.7% | 15.7% | 4.33E-06 |
| Sterile | 86.8% | 84.3% | 0.7401 | 88.3% | 84.3% | 0.4852 | 64.3% | 84.3% | 0.0685 |

| **RfMkr16** | A/A | | | A/C | | | C/C | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 38.2% | 16.6% | 1.62E-06 | 12.7% | 16.6% | 0.0673 | 14.0% | 16.6% | 0.4621 |
| Sterile | 61.8% | 83.4% | 0.0507 | 87.3% | 83.4% | 0.4560 | 86.0% | 83.4% | 0.7645 |

| RfMkr17 | G/G | | | G/T | | | T/T | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 46.7% | 14.3% | 4.01E-07 | 11.3% | 14.3% | 0.1247 | - | - | - |
| Sterile | 53.3% | 85.7% | 0.0554 | 88.7% | 85.7% | 0.5618 | - | - | - |

| RfMkr18 | AA/AA | | | AA/CG | | | CG/CG | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 24.1% | 12.8% | 0.0129 | 12.3% | 12.8% | 0.8203 | 8.9% | 12.8% | 0.1928 |
| Sterile | 75.9% | 87.2% | 0.3740 | 87.7% | 87.2% | 0.9353 | 91.1% | 87.2% | 0.6416 |

| RfMkr6 | A/A | | | A/G | | | G/G | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 0.0% | 14.7% | 0.2719 | 10.8% | 14.7% | 0.0686 | 35.7% | 14.7% | 9.04E-06 |
| Sterile | 100.0% | 85.3% | 0.6735 | 89.2% | 85.3% | 0.4849 | 64.3% | 85.3% | 0.0887 |

| RfMkr7 | A/A | | | A/G | | | G/G | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 38.5% | 15.5% | 7.62E-05 | 12.5% | 15.5% | 0.1557 | 13.3% | 15.5% | 0.8149 |
| Sterile | 61.5% | 84.5% | 0.1191 | 87.5% | 84.5% | 0.5759 | 86.7% | 84.5% | 0.9265 |

| RfMkr10 | A/A | | | A/C | | | C/C | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | - | - | - | 10.1% | 17.9% | 0.0122 | 23.8% | 17.9% | 0.0313 |
| Sterile | - | - | - | 89.9% | 82.1% | 0.2882 | 76.2% | 82.1% | 0.3617 |

| RfMkr11 | A/A | A/G | G/G |
|---|---|---|---|

| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
|---|---|---|---|---|---|---|---|---|---|
| Fertile | 8.3% | 13.7% | 0.5895 | 11.2% | 13.7% | 0.2265 | 23.4% | 13.7% | 0.0134 |
| Sterile | 91.7% | 86.3% | 0.8418 | 88.8% | 86.3% | 0.6545 | 76.6% | 86.3% | 0.3601 |

| **RfMkr13** | G/G | | | G/T | | | T/T | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 29.5% | 15.1% | 3.63E-04 | 11.5% | 15.1% | 0.0755 | - | - | - |
| Sterile | 70.5% | 84.9% | 0.1666 | 88.5% | 84.9% | 0.4905 | - | - | - |

| **RfMkr19** | C/C | | | C/T | | | T/T | | |
|---|---|---|---|---|---|---|---|---|---|
| | Observed | Expected | p-value | Observed | Expected | p-value | Observed | Expected | p-value |
| Fertile | 12.2% | 14.9% | 0.3358 | 28.9% | 14.9% | 0.0153 | 0.0% | 14.9% | 0.2678 |
| Sterile | 87.8% | 85.1% | 0.7100 | 71.1% | 85.1% | 0.3487 | 100.0% | 85.1% | 0.6686 |

**Supplementary Table 2.** Comparison of the location (Linkage group and centimorgan) of 163 markers from the lolium genome zipper to the teagasc ultra-high density perennial ryegrass map. Markers are ordered by zipper location.

| Zipper Marker | LG | cM | Teagasc scaffold | LG | cM |
|---|---|---|---|---|---|
| ve_003c_f04 | 1 | 29.0 | scaffold6354 | 1 | 74.0 |
| PTA.1969.C1 | 1 | 30.1 | scaffold7669 | 1 | 47.9 |
| PTA.604.C1 | 1 | 31.5 | scaffold2954 | 3 | 116.9 |
| PTA.1095.C1 | 1 | 31.8 | scaffold72 | 1 | 13.0 |
| PTA.1450.C1 | 1 | 32.0 | scaffold5782 | 1 | 16.1 |
| PTA.2153.C1 | 1 | 32.1 | scaffold4771 | 1 | 45.2 |
| r_008a_f02 | 1 | 33.0 | scaffold1506 | 1 | 74.0 |
| P5G07 | 1 | 33.0 | scaffold6580 | 1 | 28.2 |
| PTA.1025.C1 | 1 | 33.3 | scaffold3528 | 1 | 94.1 |
| ve_001a_e04 | 1 | 34.5 | scaffold462 | 1 | 45.5 |
| PTA.1742.C1 | 1 | 34.9 | scaffold8303 | 1 | 40.2 |
| PTA.2113.C1 | 1 | 35.1 | scaffold2539 | 1 | 29.8 |
| PTA.978.C1 | 1 | 38.2 | scaffold1983 | 1 | 62.7 |
| r_010c_f02 | 1 | 43.2 | scaffold6001 | 1 | 94.2 |
| ve_006a_f09 | 1 | 43.5 | scaffold9680 | 1 | 102.7 |
| PTA.2760.C1 | 1 | 49.3 | scaffold1163 | 1 | 98.2 |
| PTA.1007.C1 | 1 | 62.0 | scaffold1651 | 1 | 118.1 |
| PTA.1007.C2 | 1 | 63.0 | scaffold1651 | 1 | 118.1 |
| PTA.279.C1 | 1 | 67.5 | scaffold2308 | 1 | 24.5 |
| r_006c_h02 | 1 | 93.3 | scaffold6065 | 1 | 0.0 |
| PTA.403.C1 | 1 | 94.1 | scaffold1104 | 1 | 7.8 |
| PTA.1053.C1 | 2 | 12.6 | scaffold683 | 2 | 123.5 |
| PTA.1395.C1 | 2 | 35.7 | scaffold1589 | 2 | 99.6 |
| PTA.486.C1 | 2 | 66.8 | scaffold321 | 2 | 55.3 |
| PTA.664.C1 | 2 | 67.6 | scaffold10427 | 2 | 61.4 |
| PTA.615.C1 | 2 | 72.8 | scaffold2021 | 2 | 46.7 |
| PTA.1958.C1 | 2 | 79.9 | scaffold2708 | 2 | 87.1 |
| PTA.538.C1 | 2 | 80.1 | scaffold3296 | 2 | 35.0 |
| PTA.669.C1 | 2 | 80.2 | scaffold1599 | 2 | 35.0 |
| PTA.1430.C1 | 2 | 80.2 | scaffold4680 | 2 | 38.1 |
| vr_002b_h03 | 2 | 80.4 | scaffold2009 | 2 | 72.9 |
| PTA.830.C1 | 2 | 80.4 | scaffold158 | 3 | 14.7 |
| PTA.262.C1 | 2 | 80.4 | scaffold3077 | 2 | 46.7 |
| rg1_012d_f09 | 2 | 80.5 | scaffold6541 | 2 | 39.9 |
| PTA.1755.C1 | 2 | 80.8 | scaffold458 | 2 | 41.6 |
| PTA.663.C1 | 2 | 80.9 | scaffold3280 | 2 | 39.9 |
| PTA.1487.C1 | 2 | 81.3 | scaffold1571 | 2 | 46.7 |
| PTA.26.C1 | 2 | 82.0 | scaffold2009 | 2 | 72.9 |
| PTA.236.C1 | 2 | 82.5 | scaffold3060 | 2 | 46.7 |
| PTA.2280.C1 | 2 | 83.9 | scaffold7425 | 2 | 64.1 |
| PTA.1473.C1 | 2 | 86.3 | scaffold2635 | 2 | 29.7 |
| PTA.1036.C1 | 2 | 90.7 | scaffold3948 | 2 | 35.0 |
| PTA.1161.C1 | 2 | 140.3 | scaffold29 | 2 | 0.0 |
| PTA.2970.C1 | 3 | 0.3 | scaffold109 | 2 | 85.6 |
| PTA.1219.C2 | 3 | 29.4 | scaffold2475 | 3 | 41.2 |
| PTA.2225.C1 | 3 | 29.5 | scaffold1189 | 3 | 41.1 |
| PTA.429.C2 | 3 | 31.3 | scaffold8879 | 3 | 76.0 |
| gsa_002a_a07 | 3 | 33.0 | scaffold4410 | 3 | 46.2 |
| PTA.204.C2 | 3 | 37.1 | scaffold3967 | 3 | 89.7 |
| PTA.1045.C1 | 3 | 37.2 | scaffold5875 | 3 | 74.4 |
| ve_004c_b06 | 3 | 37.2 | scaffold1706 | 3 | 74.5 |
| rg6_011d_b05 | 3 | 39.5 | scaffold8629 | 7 | 84.0 |
| PTA.2492.C1 | 3 | 40.4 | scaffold3424 | 3 | 86.2 |
| r_013d_g11 | 3 | 41.3 | scaffold11368 | 3 | 74.6 |
| PTA.918.C1 | 3 | 44.8 | scaffold2601 | 3 | 70.2 |
| P7G05 | 3 | 46.8 | scaffold11338 | 3 | 88.5 |
| PTA.1113.C1 | 3 | 48.5 | scaffold1075 | 3 | 108.1 |
| PTA.2768.C1 | 3 | 49.0 | scaffold331 | 3 | 76.7 |
| RGC2 | 3 | 49.7 | scaffold1980 | 3 | 107.9 |
| PTA.1162.C1 | 3 | 50.3 | scaffold359 | 3 | 107.3 |
| PTA.396.C2 | 3 | 51.0 | scaffold331 | 3 | 76.7 |
| r_010b_g07 | 3 | 51.5 | scaffold8392 | 3 | 91.6 |
| PTA.813.C1 | 3 | 56.7 | scaffold384 | 3 | 0.0 |
| PTA.1637.C1 | 3 | 58.7 | scaffold275 | 3 | 135.9 |
| PTA.1900.C1 | 3 | 60.5 | scaffold3448 | 3 | 131.5 |
| PTA.1842.C1 | 3 | 63.3 | scaffold9106 | 3 | 118.7 |
| PTA.1792.C1 | 4 | 17.6 | scaffold2161 | 4 | 146.2 |
| PTA.770.C1 | 4 | 20.6 | scaffold6357 | 7 | 0.3 |
| r_011c_f09 | 4 | 21.6 | scaffold8599 | 4 | 136.9 |
| PTA.749.C2 | 4 | 25.1 | scaffold7428 | 4 | 148.1 |
| PTA.2126.C1 | 4 | 29.3 | scaffold4659 | 4 | 130.8 |
| LpMADS01 | 4 | 32.6 | scaffold2255 | 3 | 75.1 |
| PTA.623.C1 | 4 | 35.0 | scaffold3597 | 4 | 139.3 |
| PTA.2931.C1 | 4 | 35.2 | scaffold3597 | 4 | 139.3 |
| PTA.72.C4 | 4 | 35.8 | scaffold4754 | 4 | 132.7 |
| PTA.967.C1 | 4 | 36.3 | scaffold10467 | 4 | 132.3 |
| PHYA | 4 | 38.5 | scaffold193 | 4 | 136.1 |
| PTA.1.C1 | 4 | 41.5 | scaffold30788 | 3 | 109.8 |

| Zipper Marker | LG | cM | Teagasc scaffold | LG | cM |
|---|---|---|---|---|---|
| PTA.3.C1 | 4 | 47.6 | scaffold462 | 1 | 45.5 |
| ve_003b_h12 | 4 | 48.0 | scaffold3361 | 4 | 87.8 |
| PTA.1031.C1 | 4 | 49.4 | scaffold403 | 2 | 11.4 |
| PTA.133.C2 | 4 | 49.8 | scaffold858 | 4 | 47.6 |
| PTA.2064.C1 | 4 | 50.0 | scaffold7973 | 4 | 112.2 |
| PTA.2469.C1 | 4 | 50.5 | scaffold6834 | 4 | 81.6 |
| PTA.2787.C1 | 4 | 50.5 | scaffold6799 | 4 | 43.7 |
| PTA.1451.C1 | 4 | 50.6 | scaffold6176 | 4 | 54.2 |
| PTA.1565.C1 | 4 | 50.8 | scaffold3627 | 4 | 37.4 |
| PTA.7.C3 | 4 | 50.9 | scaffold1159 | 4 | 130.8 |
| PHYB | 4 | 51.0 | scaffold11323 | 4 | 64.8 |
| ve_004a_f04 | 4 | 51.0 | scaffold5145 | 4 | 83.5 |
| PTA.1130.C1 | 4 | 51.0 | scaffold8374 | 4 | 101.9 |
| ve_004b_d01 | 4 | 51.0 | scaffold71 | 4 | 47.6 |
| PTA.2550.C1 | 4 | 51.0 | scaffold7246 | 4 | 50.2 |
| PTA.2015.C1 | 4 | 51.0 | scaffold9955 | 4 | 77.2 |
| PTA.1026.C1 | 4 | 51.0 | scaffold272 | 5 | 36.2 |
| PTA.946.C1 | 4 | 51.5 | scaffold2377 | 4 | 76.5 |
| PTA.1021.C1 | 4 | 51.8 | scaffold6797 | 4 | 131.1 |
| PTA.1758.C1 | 4 | 53.8 | scaffold6663 | 7 | 112.5 |
| PTA.609.C3 | 4 | 54.3 | scaffold279 | 4 | 91.2 |
| PTA.32.C6 | 4 | 55.5 | scaffold5246 | 4 | 64.8 |
| r_005d_h01 | 4 | 55.6 | scaffold40 | 1 | 93.6 |
| PTA.600.C1 | 4 | 63.7 | scaffold15350 | 4 | 32.4 |
| PTA.161.C2 | 4 | 63.9 | scaffold647 | 3 | 93.9 |
| PTA.1455.C1 | 4 | 71.2 | scaffold8956 | 4 | 10.8 |
| PTA.475.C1 | 4 | 72.0 | scaffold621 | 4 | 10.8 |
| ve_003b_h42 | 4 | 84.8 | scaffold7810 | 4 | 9.1 |
| r_004d_b05 | 4 | 101.7 | scaffold2831 | 4 | 0.0 |
| PTA.359.C1 | 4 | 109.4 | scaffold484 | 4 | 0.0 |
| PTA.684.C2 | 4 | 119.2 | scaffold1390 | 4 | 58.5 |
| PTA.53.C1 | 5 | 13.5 | scaffold3708 | 5 | 30.5 |
| PTA.2547.C1 | 5 | 24.2 | scaffold14027 | 5 | 55.9 |
| PTA.1462.C1 | 5 | 26.9 | scaffold1702 | 5 | 46.9 |
| ve_005b_h05 | 5 | 27.0 | scaffold22728 | 5 | 52.3 |
| PTA.2438.C1 | 5 | 27.2 | scaffold12508 | 5 | 60.0 |
| PTA.1727.C1 | 5 | 27.4 | scaffold6517 | 5 | 42.9 |
| PTA.1433.C1 | 5 | 27.5 | scaffold939 | 5 | 77.2 |
| PTA.394.C2 | 5 | 27.8 | scaffold5422 | 5 | 74.0 |
| PTA.1451.C2 | 5 | 28.7 | scaffold5993 | 5 | 75.0 |
| PTA.3.C2 | 5 | 28.9 | scaffold596 | 4 | 76.5 |
| r_003c_h12 | 5 | 34.3 | scaffold2160 | 5 | 91.8 |
| rg1_009d_g08 | 5 | 38.1 | scaffold909 | 5 | 70.8 |
| PTA.505.C2 | 5 | 40.3 | scaffold46 | 5 | 61.3 |
| PTA.1197.C1 | 5 | 44.9 | scaffold1258 | 5 | 111.4 |
| PTA.1577.C1 | 5 | 55.2 | scaffold15288 | 5 | 97.8 |
| PTA.1750.C1 | 6 | 30.8 | scaffold415 | 3 | 51.9 |
| PTA.414.C2 | 6 | 31.6 | scaffold415 | 3 | 51.9 |
| PTA.27.C1 | 6 | 40.9 | scaffold9116 | 6 | 40.7 |
| PTA.216.C1 | 6 | 41.8 | scaffold4634 | 6 | 42.2 |
| PTA.1012.C2 | 6 | 43.6 | scaffold1957 | 6 | 24.9 |
| PTA.1671.C1 | 6 | 47.3 | scaffold11647 | 6 | 59.8 |
| PTA.2339.C1 | 6 | 51.5 | scaffold7522 | 2 | 67.1 |
| PTA.796.C2 | 6 | 52.2 | scaffold2111 | 6 | 33.4 |
| CRY2 | 6 | 52.6 | scaffold2971 | 6 | 33.4 |
| PTA.1374.C1 | 6 | 54.3 | scaffold7834 | 1 | 32.5 |
| PTA.2659.C1 | 6 | 57.3 | scaffold6237 | 6 | 41.6 |
| PTA.2129.C1 | 6 | 58.2 | scaffold3432 | 6 | 21.4 |
| PTA.386.C1 | 6 | 60.9 | scaffold5689 | 3 | 71.8 |
| PTA.3133.C1 | 6 | 65.9 | scaffold3184 | 2 | 38.1 |
| PTA.1421.C1 | 7 | 28.9 | scaffold84 | 7 | 14.7 |
| PTA.2484.C1 | 7 | 35.4 | scaffold817 | 7 | 19.9 |
| PTA.1769.C1 | 7 | 35.9 | scaffold537 | 7 | 18.9 |
| PTA.419.C1 | 7 | 38.6 | scaffold65 | 7 | 41.4 |
| PTA.9.C1 | 7 | 41.4 | scaffold11275 | 7 | 7.9 |
| ve_001a_a11 | 7 | 44.8 | scaffold1248 | 7 | 40.5 |
| PTA.2014.C1 | 7 | 47.9 | scaffold4586 | 7 | 51.7 |
| PTA.1499.C1 | 7 | 48.5 | scaffold3424 | 3 | 86.2 |
| PTA.2552.C1 | 7 | 48.8 | scaffold10288 | 7 | 28.4 |
| PTA.644.C1 | 7 | 49.2 | scaffold483 | 3 | 86.2 |
| PTA.160.C1 | 7 | 49.4 | scaffold9453 | 7 | 40.5 |
| PTA.2190.C1 | 7 | 50.5 | scaffold1099 | 7 | 75.3 |
| PTA.1997.C1 | 7 | 50.5 | scaffold8025 | 7 | 32.2 |
| PTA.438.C1 | 7 | 50.6 | scaffold7508 | 7 | 61.7 |
| PTA.936.C2 | 7 | 50.9 | scaffold8025 | 7 | 32.2 |
| PTA.198.C2 | 7 | 51.2 | scaffold2057 | 7 | 30.9 |
| PTA.529.C2 | 7 | 51.5 | scaffold15625 | 7 | 35.7 |
| PTA.1252.C1 | 7 | 53.4 | scaffold4945 | 7 | 32.2 |
| PTA.762.C1 | 7 | 55.3 | scaffold15204 | 7 | 28.4 |
| PTA.2013.C1 | 7 | 55.5 | scaffold1251 | 7 | 40.5 |
| P5G15 | 7 | 56.7 | scaffold3685 | 7 | 33.9 |
| PTA.315.C1 | 7 | 62.3 | scaffold11530 | 7 | 41.7 |
| PTA.550.C1 | 7 | 62.8 | scaffold4378 | 7 | 40.5 |
| PTA.586.C2 | 7 | 98.2 | scaffold14058 | 7 | 52.6 |
| PTA.265.C1 | 7 | 106.0 | scaffold8602 | 7 | 31.2 |

# Chapter 4

# Bulk sergeant analysis in a population of perennial ryegrass (*Lolium perenne* L.) segregating for restoration of cytoplasmic male sterility

**Timothy Sykes[1], Steven Yates[1] and Bruno Studer[1]***

[1] Institute of Agricultural Sciences, Molecular Plant Breeding, ETH Zurich, 8092 Zurich, Switzerland.

*Corresponding author: bruno.studer@usys.ethz.ch

## Abstract

The ability to link observed phenotypes to their underlying genes is essential for the progress of molecular plant breeding. One of the most powerful approaches used to achieve this is bulk sergeant analysis (BSA), which allows the relatively quick and cheap identification of qualitative trait loci (QTL) for a given phenotype.

Here, we applied BSA to a population of perennial ryegrass (*Lolium perenne* L.) segregating for cytoplasmic male sterility (CMS) restoration and identified two major QTL for this trait. These restoration QTL are linked to regions of nuclear-integrated mitochondrial genome segments, implying that fertility restoration is mediated by functional nuclear copies of mutated CMS-causal mitochondrial genes. This is further supported by the observation of a mutation in the mitochondrial 18S ribosomal RNA subunit, suggesting that disrupted protein synthesis may be the cause of CMS. Although nuclear copies of mitochondrial genes have been observed in many plant species, this is the first time they have been implicated in CMS restoration.

The results presented here uncover a new level of understanding in the complex interplay between mitochondrial and nuclear genomes. Moreover, the two QTL represent the best targets for CMS-restoration genes and markers that could be utilised in hybrid breeding programs of perennial ryegrass, to efficiently identify CMS-maintaining and fertility-restoring plants.

**Key words:** Pooled sequencing, bulk sergeant analysis (BSA), perennial ryegrass (*Lolium perenne* L.), cytoplasmic male sterility (CMS), restoration of fertility.

## Introduction

Linking observed phenotypes with their underlying genetic variation is a fundamental concept of molecular plant breeding. Identifying the key genetic loci involved in agronomically important traits is a vital step towards crop improvement and can help to isolate the genetic pathways underlying these traits, providing further targets for plant breeding. On a broader scale, evolutionary researchers are interested in the genetic interactions underpinning phenotypic adaptive change. One approach, often applied in species that can be experimentally crossed, is quantitative trait locus (QTL) mapping, where genomic regions containing causal genetic variants for phenotypic traits of interest can be identified. This approach was first described in the 1920s [1] but was not widely adopted until DNA markers became available in the 1980s and 1990s, allowing researchers to determine that QTL could explain a significant proportion of observed phenotypic difference in maize and tomatoes [2]. Subsequently, many studies were initiated to map QTL for quantitative traits such as quality and yield as well as resistance to biotic and abiotic stresses [3].

QTL mapping, in its most basic form, involves the genotyping of F2 offspring from a cross between parents with contrasting phenotypes. This allows the identification of genomic regions that are inherited non-randomly with respect to the phenotype of interest. To achieve this on a resolution sufficient for gene discovery, such forward genetic studies often require the genotyping of many individuals. Although this is a very powerful approach, it is extremely time-consuming and expensive, especially when populations consist of hundreds or thousands of individuals. With advances in technology allowing the sequencing and assembly of whole genomes an alternative method has become available.

Pooling DNA samples by phenotype, followed by bulk shotgun sequencing and subsequent SNP allele frequency analysis, provides a more efficient tool to screen the large number of plants required to map genes at a fine scale [4]. This approach, referred to as bulk segregant analysis (BSA), is especially effective when studying absolute segregating populations where the phenotype of interest is binary [5], and has the added benefit of requiring less DNA from each individual. Using pooled sequencing approaches can accurately estimate allele frequencies across the genome identifying causal loci [6]. One difficulty of this approach is differentiating between sequencing errors and rare variants. Although increasing sequencing depth can account for this, recently developed filtering methods have also been developed that can assist in this differentiation [8].

BSA approaches have been successfully employed to identify candidate genes for a variety of traits seen in model species such as Drosophila (*Drosophila melanogaster* L.) [9], Arabidopsis (*Arabidopsis lyrata* L.) [10], yeast (*Saccharomyces cerevisiae* L.) [11] and even humans as early as 1991 [12]. BSA studies have also been carried out in several crop species including in maize (*Zea mays* l.) to identify QTL for drought resistance [13], in rice (*Oryza sativa* L.) to identify QTL for increased yield [14], in barley (*Hordeum vulgare* L) for leaf rust resistance QTL [15], in wheat (*Triticum aestivum* L.) for Fusarium head blight resistance QTL [16] and many other important crop species.

In this study, the power of pooled sequencing was harnessed by applying it to a large population of perennial ryegrass (*Lolium perenne L.)* plants segregating for fertility restoration of cytoplasmic male sterility (CMS). This population was chosen since molecular markers for CMS fertility restoration would be of great benefit to plant breeders attempting to create commercial hybrid varieties. CMS is a maternally inherited trait, resulting in the inability to

produce viable pollen [17] and is utilised in hybrid breeding schemes as a pollination control mechanism [18]. Restoration of the CMS caused pollen deficiency can be mediated by nuclear restoration of fertility (*Rf*) genes. Several of these genes have been identified in other plant species [18] and they are usually members of the restorer of fertility-like pentatricopeptide repeat (*RFL*) gene sub-family [19]. The CMS/restoration system used in this study was chosen as it presents a discrete binary phenotype, fertile or sterile (chapter 2), making this trait ideal to be studied with pooled sequencing and BSA.

## Methods

### DNA library preparation

DNA samples from the previous genotyping by sequencing (GBS) study (Chapter 3) were used for this study. All 1,103 samples were pooled by phenotype with pools of 299 and 804 fertile and sterile plants, respectively. All DNA samples were normalised to 10 ng/ul prior to pooling and equal amounts of each sample added to the pools, as unequal contributions of different individuals to the DNA pool can affect the standard error of allele frequency estimates [6]. The sequencing libraries were prepared using covaris shearing for fragmentation, standard Illumina library construction using the Ovation Rapid DR library system (NuGEN Technologies Inc., USA), followed by normalisation using the Trimmer kit (Evrogen JSC, Russia). The hybridization step during normalization was extended to 16h (instead of 3-7h for cDNA normalization) and reamplification was done with standard Illumina library primers (3PE and 5PE) by LGC Genomics (LGC Limited, UK). Each library was sequenced on a single Illumina HiSeq® 2500 lane.

### SNP discovery and allele frequency calculation using a high-quality genome assembly of Italian ryegrass

Quality-checked and trimmed sequencing reads of both pools were mapped to the high-quality genome assembly of the Italian ryegrass (*Lolium multiflorum* Lam.) genotype Rabiosa, hereafter referred to as the Rabiosa genome assembly (Molecular Plant Breeding, ETH Zurich, Switzerland, unpublished), using Bowtie2 (V.2.2.3, default settings) [20]. To determine the allele frequencies within each pool, SAMtools (V1.2) [21] was used to convert the alignment (SAM files) into sorted binary alignments (BAM). Next SAMtools (V1.6) was used to generate a 'mpileup' file with SNP counts at each position. These data were then passed to bcftools (Samtools V1.6) to call variants and report the depth at each locus. Subsequently, the Perl script vcfutils.pl was used to remove SNPs with a depth less than 50. These data were then imported into a Python Environment (2.7.13), using Conda (V4.3.21, Anaconda Inc., USA) package manager. The depth at each allele (reference and alternative) was extracted using regular expressions for both fertile and sterile pools. The allele frequencies were then determined per pool ((100/Total count) x Alternative count). To remove SNPs which differ from the reference sequence (Rabiosa) but are conserved in the sequenced material, SNPs with allele frequencies equal to 100 in both pools were removed. Furthermore, all SNPs with less than 50 calls were excluded. Finally, to reduce the data size, only SNPs which segregated in one or both pools were retained. This was done by removing SNPs where the allele frequencies differed by more than 10%. The filtered data was then exported as a text file.

To identify loci with segregation distortions, the above expression data were loaded into R statistical environment (V.3.4.1). Prior to analysis, SNPs were further filtered based on abundance in each pool, by selecting those with at least 20 counts in each pool. Additionally, upon manual inspection, it was apparent that chloroplast sequences were also prevalent. These were (mostly) omitted based on coverage, where only SNPs with less than 1,000 calls were retained. Next, the allele frequencies per SNP were converted to a contingency table and then assessed for differences by use of a Chi-square test (core function). From these data, only SNPs with a *P*-value less than 1E-50 were retained. To identify loci (groups of SNPs) and remove spurious results from single SNPs, only scaffolds with at least five highly significant (*P*-value < 1e-50) SNPs were retained. These SNPs and corresponding scaffolds were then manually inspected and subject to further analysis.

## Results

### Sequencing and SNP identification

Illumina sequencing of the sterile and fertile pool produced 331,949,038 and 302,367,854 paired-end reads, respectively. These were mapped to the Rabiosa genome assembly and led to the identification of 58,579 SNPs, of which 51,751 passed filtering. *P*-values from the subsequent chi-square test for independence from the phenotype were converted into LOD scores (-log10(p-value)) and visualised across the Rabiosa genome (Figure 1).

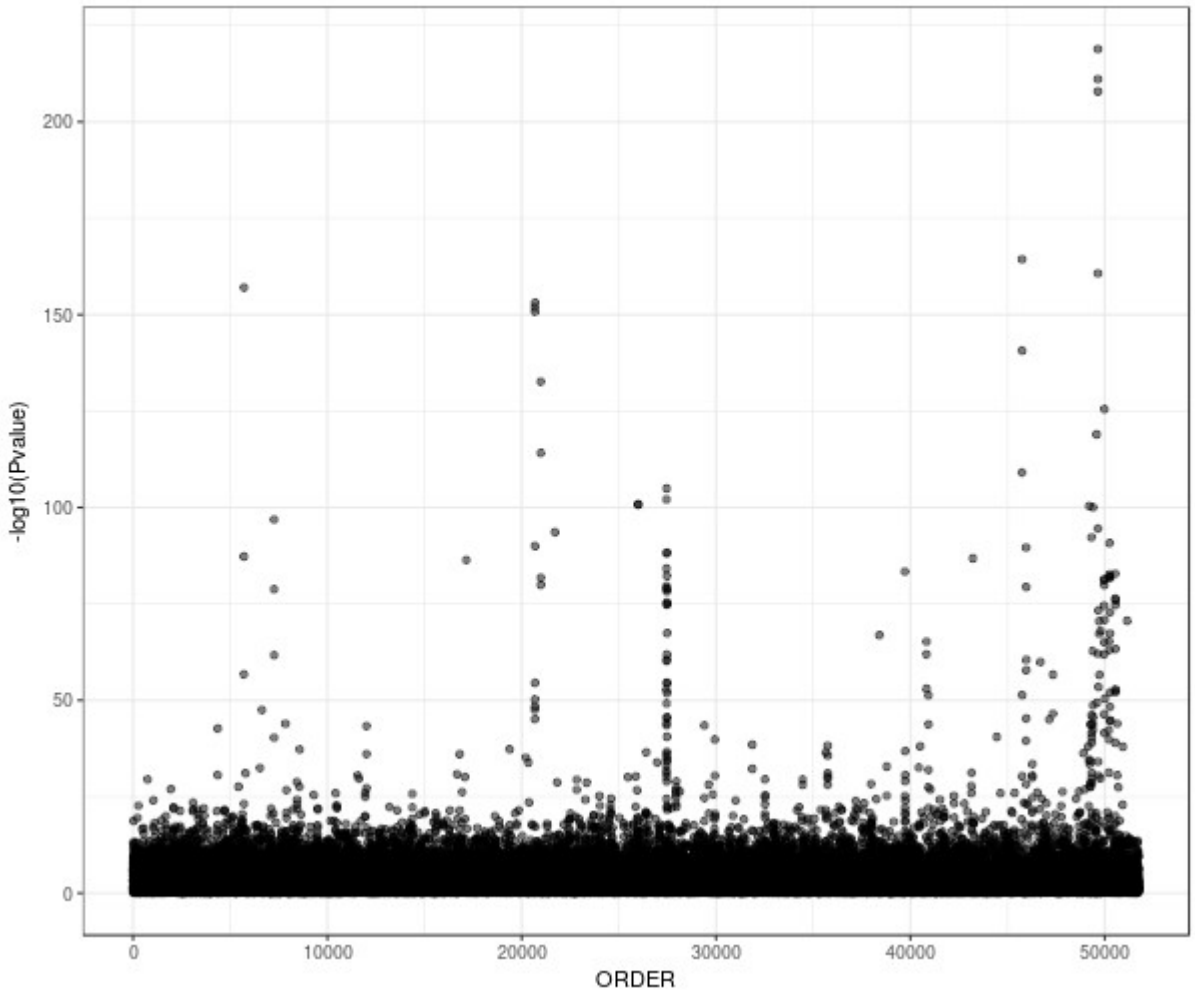


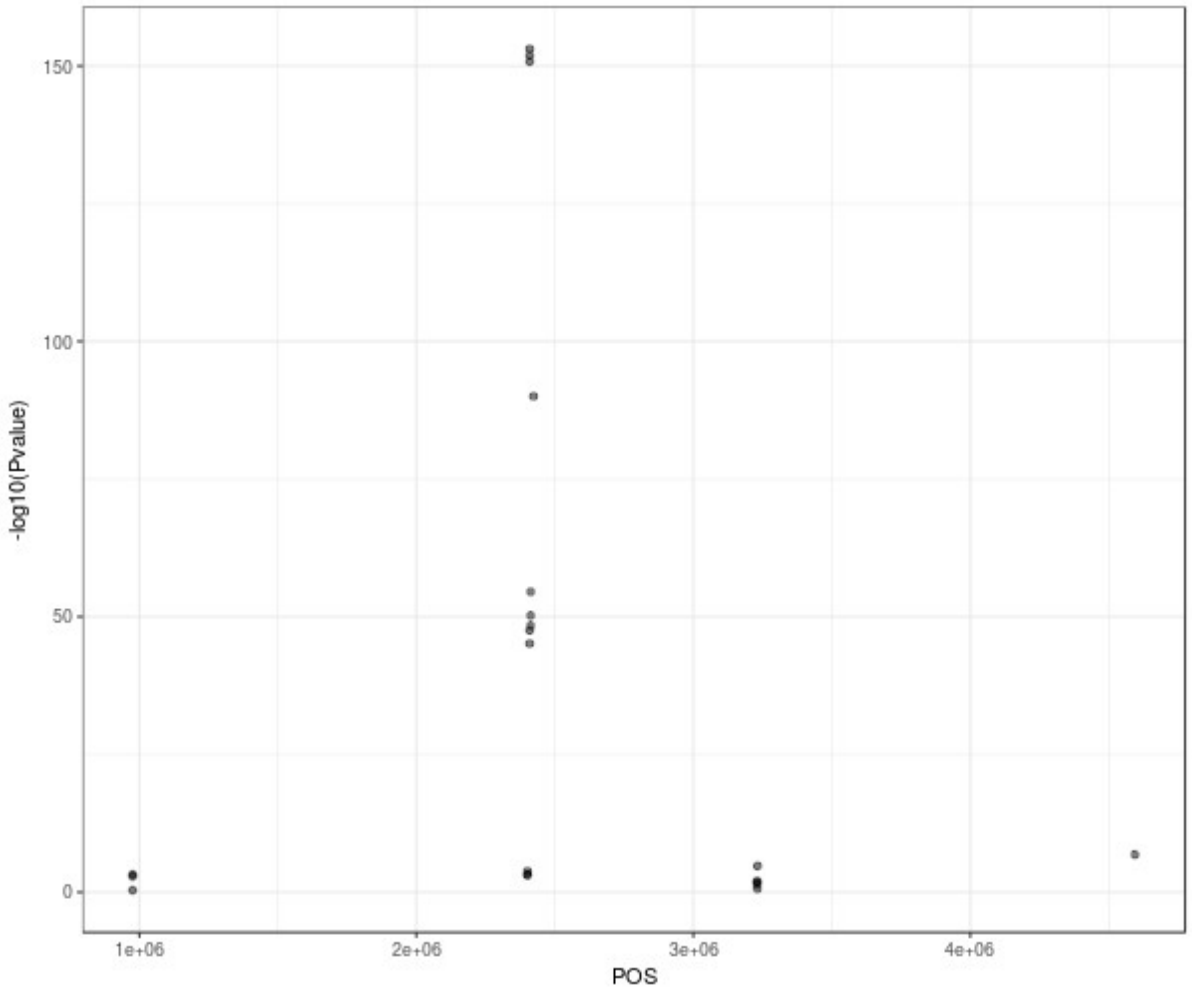

**Figure 1. Allele frequency shifts across the Rabiosa genome**. The graph shows the LOD score values (-log10 (*P* value), y-axis) for all SNPs that passed filtering. Rabiosa scaffolds are ordered by scaffold number (ORDER, x-axis).

Only five Rabiosa scaffolds contained five or more SNPs with a LOD score of above 50, referred to as "significant scaffolds" containing QTL (Table 1). Two of these scaffolds, Scaffold2257 and Scaffold3325, had a length of 2,740,423 and 975,733 bp, respectively, and contained several annotated genes. The positions of the SNPs with a LOD score above 50 sharply peaked on those two scaffolds, as illustrated in figure 2. The other three scaffolds, Scaffold142754, Scaffold153099 and Scaffold172632, were 2,572, 1,016 and 2,185 bp long, respectively, and contained significantly associated SNPs evenly distributed across the whole scaffold. They were considered as being QTL in their entirety.

**Table 1.** Rabiosa scaffolds containing at least five SNPs with a LOD score of above 50, their lengths, number of SNPs (LOD>50) and the location of the identified QTL.

| Significant Scaffold | Length (bp) | No. Of SNPs (LOD>50) | Location of QTL (kb) | Size of QTL (kb) |
|---|---|---|---|---|
| Scaffold2257 | 2,740,423 | 21 | 2,409-2,423 | 14 |
| Scaffold3325 | 975,733 | 62 | 557-617 | 60 |
| Scaffold142754 | 2,572 | 7 | 1-2.5 | 2.5 |
| Scaffold153099 | 1,016 | 7 | 1-1.0 | 1 |
| Scaffold172632 | 2,185 | 11 | 1-2.2 | 2.2 |

**A.**



**B.**

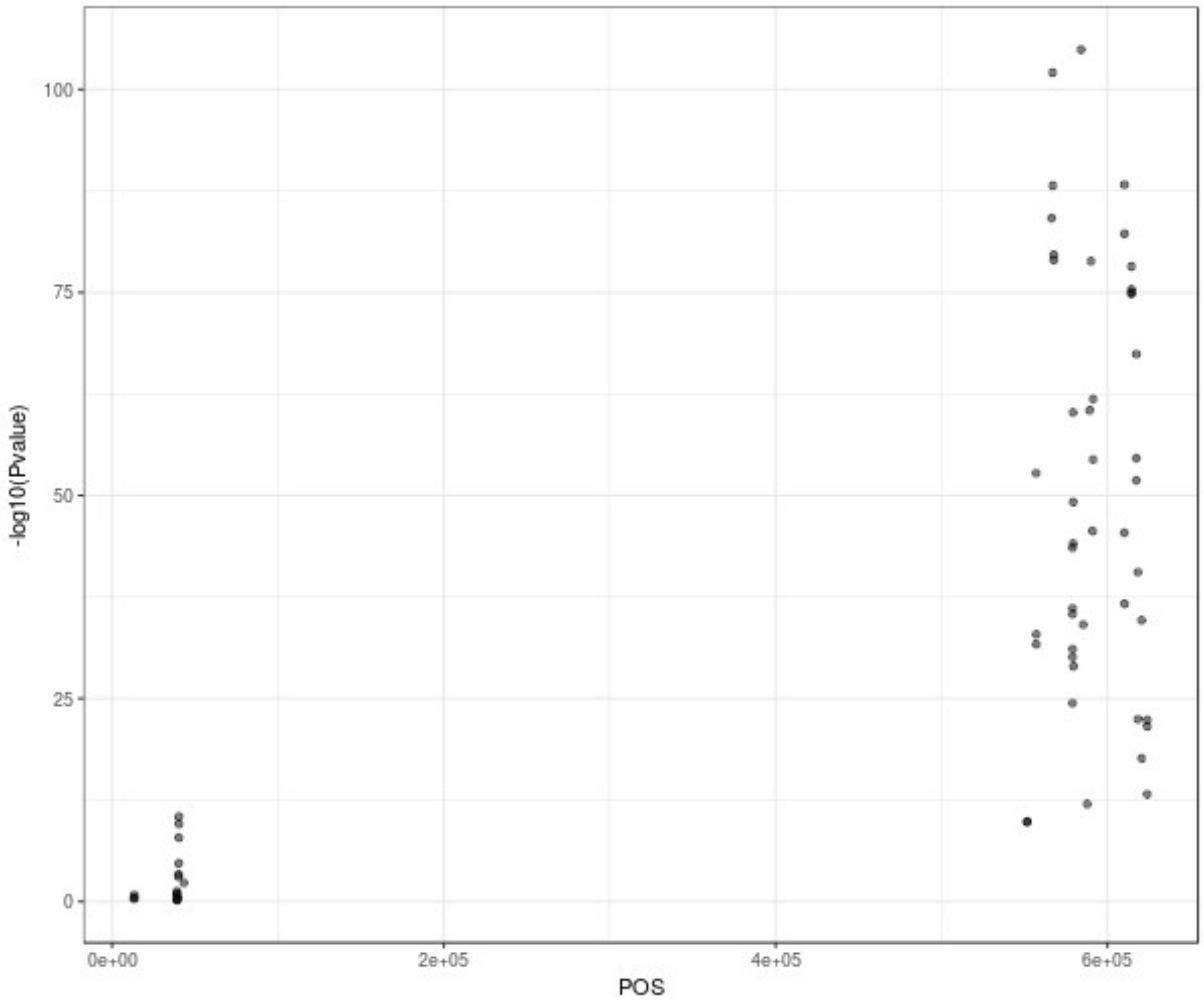



**Figure 2. Position of the QTL for fertility restoration on scaffold2257 (A) and scaffold3325 (B).** SNP positions in base pairs (POS, x-axis) are given with the corresponding LOD score values (-log10 (*P* value), y-axis).

## Genetic control of fertility restoration

Analysis of the SNP calls within the significant scaffolds revealed that within the two QTL, the loci were homozygous for the sterile and heterozygous for the fertile pools. Within the QTL on Scaffold2257, the common call, i.e. the base call at any one SNP that was seen more often, was recorded with a frequency of 0.9408 in the sterile pool, whereas the frequency was 0.5947 in the fertile pool. The same pattern was observed within Scaffold3325, with the common allele having a frequency of 0.9651 in the sterile and 0.5971 in the fertile pool (Table 2). When taking all the 58,579 SNPs recorded across the Rabiosa genome into account, the frequency that the reference SNP call was recorded in the sterile pool is 0.5528 and 0.5279 in the fertile pool. From this, it can be concluded that in the regions of the identified QTL, there is a highly significant shift towards homozygosity in the sterile, but not in the fertile pool.

**Table 2.** Location, genotyping calls and LOD scores of each SNP in Scaffold2257 and scaffold3325.

| | | | Scaffold3325 | | | | |
|---|---|---|---|---|---|---|---|
| **Position (bp)** | **Reference base** | **Alternate base** | **Sterile reference base count** | **Sterile alternate base counts** | **Fertile reference base count** | **Fertile alternate base counts** | **LOD** |
| 13264 | T | A | 15 | 16 | 12 | 21 | 0.343783 |
| 13295 | T | G | 20 | 16 | 15 | 25 | 0.452467 |
| 13431 | T | C | 33 | 32 | 31 | 18 | 0.807325 |
| 39034 | G | C | 17 | 11 | 11 | 12 | 0.233063 |
| 39053 | C | A | 23 | 9 | 12 | 14 | 1.243234 |
| 39070 | C | G | 22 | 9 | 15 | 13 | 0.955505 |
| 39117 | C | T | 18 | 16 | 12 | 18 | 0.166044 |
| 39248 | A | G | 11 | 21 | 14 | 10 | 0.818065 |
| 39662 | T | C | 12 | 14 | 18 | 10 | 0.338386 |
| 40082 | C | T | 27 | 7 | 16 | 9 | 3.067131 |
| 40085 | T | G | 28 | 7 | 17 | 9 | 3.333032 |
| 40119 | G | T | 30 | 6 | 14 | 8 | 4.70E+00 |
| 40262 | T | C | 40 | 5 | 21 | 9 | 7.87E+00 |
| 40269 | G | A | 48 | 6 | 22 | 9 | 1.05E+01 |
| 40306 | T | C | 48 | 8 | 20 | 10 | 9.54E+00 |
| 43319 | C | T | 28 | 23 | 29 | 8 | 2.296987 |
| 551600 | A | C | 3 | 39 | 46 | 14 | 9.81E+00 |
| 551601 | G | A | 3 | 39 | 46 | 14 | 9.81E+00 |
| 556973 | T | G | 179 | 3 | 50 | 47 | 5.28E+01 |
| 557107 | G | T | 97 | 2 | 28 | 13 | 3.29E+01 |
| 557113 | A | C | 94 | 1 | 29 | 13 | 3.17E+01 |
| 566364 | C | A | 268 | 2 | 64 | 72 | 8.42E+01 |
| 566871 | C | G | 9 | 369 | 117 | 103 | 1.02E+02 |
| 567182 | T | G | 4 | 322 | 108 | 95 | 8.82E+01 |
| 567687 | A | T | 1 | 287 | 94 | 88 | 7.96E+01 |
| 567689 | A | T | 1 | 285 | 94 | 87 | 7.90E+01 |
| 578940 | G | C | 123 | 8 | 48 | 55 | 2.44E+01 |
| 578964 | C | A | 138 | 7 | 50 | 56 | 3.01E+01 |
| 578975 | A | T | 143 | 8 | 50 | 59 | 3.11E+01 |
| 578980 | A | T | 153 | 9 | 50 | 55 | 3.54E+01 |
| 578981 | A | C | 154 | 8 | 51 | 55 | 3.61E+01 |
| 579000 | C | T | 175 | 8 | 49 | 65 | 4.36E+01 |
| 579329 | A | T | 260 | 18 | 76 | 101 | 6.02E+01 |
| 579366 | T | G | 18 | 232 | 97 | 77 | 4.92E+01 |
| 579402 | G | T | 17 | 208 | 87 | 67 | 4.41E+01 |
| 579586 | A | T | 9 | 126 | 50 | 35 | 2.90E+01 |
| 584107 | A | G | 3 | 339 | 98 | 79 | 1.05E+02 |
| 585474 | A | C | 2 | 117 | 24 | 42 | 3.41E+01 |
| 587822 | A | C | 9 | 66 | 71 | 31 | 1.20E+01 |
| 589381 | T | A | 4 | 201 | 59 | 46 | 6.05E+01 |
| 590079 | G | C | 4 | 277 | 78 | 84 | 7.89E+01 |
| 591111 | C | A | 157 | 1 | 47 | 43 | 4.56E+01 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 591374 | A | C | 214 | 2 | 68 | 56 | 6.19E+01 |
| 591387 | G | A | 5 | 200 | 55 | 66 | 5.44E+01 |
| 610227 | T | A | 147 | 1 | 42 | 34 | 4.54E+01 |
| 610276 | C | A | 122 | 0 | 33 | 34 | 3.67E+01 |
| 610304 | G | C | 245 | 2 | 61 | 49 | 8.22E+01 |
| 610326 | A | G | 255 | 2 | 59 | 49 | 8.83E+01 |
| 614526 | T | C | 25 | 325 | 267 | 85 | 7.50E+01 |
| 614527 | C | T | 27 | 327 | 272 | 86 | 7.48E+01 |
| 614529 | T | C | 26 | 327 | 270 | 85 | 7.54E+01 |
| 614531 | G | T | 19 | 326 | 267 | 85 | 7.82E+01 |
| 617397 | T | G | 202 | 6 | 77 | 48 | 5.46E+01 |
| 617398 | C | A | 198 | 7 | 79 | 48 | 5.19E+01 |
| 617484 | A | C | 8 | 246 | 66 | 78 | 6.74E+01 |
| 618433 | A | C | 8 | 151 | 29 | 61 | 4.06E+01 |
| 618440 | C | A | 23 | 127 | 40 | 48 | 2.24E+01 |
| 620610 | C | T | 121 | 17 | 34 | 14 | 3.46E+01 |
| 620643 | A | T | 87 | 21 | 26 | 18 | 1.76E+01 |
| 623976 | G | T | 10 | 101 | 60 | 23 | 2.16E+01 |
| 624000 | A | G | 10 | 99 | 19 | 57 | 2.23E+01 |
| 624028 | G | A | 16 | 81 | 21 | 53 | 1.32E+01 |

**Scaffold2257**

| Position (bp) | Reference base | Alternate base | Sterile reference base count | Sterile alternate base counts | Fertile reference base count | Fertile alternate base counts | LOD |
|---|---|---|---|---|---|---|---|
| 975855 | T | G | 9 | 15 | 17 | 13 | 0.338386 |
| 975949 | G | A | 28 | 8 | 12 | 10 | 3.214594 |
| 975955 | G | A | 26 | 7 | 10 | 12 | 2.837389 |
| 2401175 | A | C | 6 | 19 | 25 | 8 | 3.129467 |
| 2401176 | A | T | 6 | 19 | 25 | 8 | 3.129467 |
| 2401179 | A | C | 25 | 1 | 18 | 17 | 3.821184 |
| 2409094 | G | T | 65 | 11 | 13 | 155 | 4.75E+01 |
| 2409100 | T | A | 48 | 29 | 159 | 12 | 4.51E+01 |
| 2409291 | T | G | 485 | 21 | 137 | 87 | 1.52E+02 |
| 2409292 | C | A | 488 | 20 | 140 | 87 | 1.53E+02 |
| 2409293 | G | T | 482 | 21 | 133 | 89 | 1.51E+02 |
| 2412506 | G | T | 30 | 256 | 133 | 79 | 4.84E+01 |
| 2412867 | T | G | 31 | 261 | 108 | 89 | 5.02E+01 |
| 2412893 | C | A | 30 | 268 | 117 | 79 | 5.45E+01 |
| 2422932 | G | T | 301 | 7 | 86 | 72 | 9.00E+01 |
| 3230241 | A | G | 24 | 9 | 9 | 12 | 1.997685 |
| 3230318 | T | G | 9 | 26 | 12 | 16 | 1.822968 |
| 3230387 | A | G | 18 | 14 | 18 | 5 | 1.376191 |
| 3230404 | T | C | 11 | 19 | 11 | 10 | 0.607269 |
| 3231715 | T | A | 2 | 21 | 6 | 22 | 4.75E+00 |
| 4594236 | G | A | 38 | 6 | 31 | 10 | 6.82E+00 |

**Genes and protein domains located within the significant scaffolds and their QTL**

The annotation of the Rabiosa genome sequence (Molecular Plant Breeding, ETH Zurich, unpublished) was used to extract the genes contained within the two significant scaffolds (Scaffold2257 and Scaffold3325, Supplementary Table 1). In total, 71 genes were present with 59 on Scaffold2257 and 12 on Scafold3325. Only one gene was located within the defined QTL on Scaffold2257 (MSTRG.21923.1). This gene encodes the mitochondrial 18S ribosomal RNA subunit and contains a three base pair polymorphism (TCG/GAT) 26 bp from the 5' end of the resultant rRNA. One pentatricopeptide repeat (*PPR*) gene was identified in proximity to the QTL identified on scaffold3325 and although *PPR* genes have been implicated in the restoration of other CMS systems [19], this *PPR* gene is not a member of the *RFL* sub-family of *PPRs* known to act as *Rf* genes.

As there was only one gene present within the QTL from scaffold2257 and scaffold3325, these QTL as well as the entire Scaffold142753, Scaffold153099 and Scaffold172632 were translated into protein sequence and searched for protein domains using pfam [22]. This identified 16 protein domains in total of which three were from scaffold153099, two from scaffold172632, two from scaffold2257, nine from scaffold3325 and none from scaffold142754 (Table 3). Of these 16 identified protein domains, twelve are known to be encoded by the mitochondrial genome. The DNA sequence of the five QTL were compared to the entire nucleotide collection NCBI (National Center for Biotechnology Information, USA) using the BLAST search tool [23]. The top hit for all five QTL was the '*Lolium perenne* mitochondrion, complete genome' sequence (GenBank: JX999996.1).

**Representation of the mitochondrial genome within the nuclear genomes of perennial ryegrass and Italian ryegrass**

To confirm whether finding a large compliment of the mitochondrial genome within the nuclear genome was common for perennial ryegrass, two perennial ryegrass draft genome assemblies [24, 25] along with the Rabiosa genome were compared to the perennial ryegrass mitochondrial genome [26] using a BLAST search. This revealed that all three nuclear genome assemblies contained significant amounts of the mitochondrial genome with over 90% of the mitochondrial genome present on four scaffolds of the genome presented by Velmurugan *et al* (2016) [24], 20% present on five scaffolds of the genome presented by Bryne *et al* (2015) [25] and 45% present on 16 Rabiosa scaffolds including the five significant scaffolds previously identified.

**Table 3.** Results of the pfam protein domain search within the QTL identified on all five significant scaffolds.

| Sequence | Frame | Domain | Start | End | Accession | E-value | Description |
|---|---|---|---|---|---|---|---|
| scaffold153099 | (+2) | YMF19 | 6 | 88 | PF02326.12 | 6.50E-20 | Plant ATP synthase F0 |
| scaffold153099 | (+2) | DUF1082 | 97 | 144 | PF06449.8 | 5.20E-31 | Mitochondrial domain of unknown function (DUF1082) |
| scaffold153099 | (+3) | Cytochrome_B | 268 | 342 | PF00033.16 | 3.60E-09 | Cytochrome b/b6/petB |
| scaffold172632 | (+1) | ELF | 393 | 639 | PF03317.10 | 3.30E-86 | ELF protein |
| scaffold172632 | (+3) | COX1 | 630 | 727 | PF00115.17 | 1.30E-30 | Cytochrome C and Quinol oxidase polypeptide I |
| scaffold2257 | (-3) | Retrotrans_gag | 567 | 666 | PF03732.14 | 4.40E-13 | Retrotransposon gag protein |
| scaffold2257 | (-2) | zf-CCHC | 749 | 766 | PF00098.20 | 3.90E-06 | Zinc knuckle |
| scaffold3325 | (-2) | DNA_pol_B_2 | 3559 | 3696 | PF03175.10 | 7.60E-18 | DNA polymerase type B, organellar and viral |
| scaffold3325 | (+3) | DUF4283 | 16934 | 17082 | PF14111.3 | 9.30E-24 | Domain of unknown function (DUF4283) |
| scaffold3325 | (+3) | Oxidored_q3 | 19188 | 19329 | PF00499.17 | 3.70E-15 | NADH-ubiquinone/plastoquinone oxidoreductase chain 6 |
| scaffold3325 | (-2) | COX3 | 20120 | 20357 | PF00510.15 | 5.80E-80 | Cytochrome c oxidase subunit III |
| scaffold3325 | (-1) | COX1 | 22593 | 22921 | PF00115.17 | 1.30E-111 | Cytochrome C and Quinol oxidase polypeptide I |
| scaffold3325 | (-3) | Cytochrom_B559 | 23928 | 23956 | PF00283.16 | 6.50E-12 | Cytochrome b559 |
| scaffold3325 | (-1) | Cytochrom_B559a | 24362 | 24397 | PF00284.17 | 9.70E-10 | Lumenal portion of Cytochrome b559 |
| scaffold3325 | (+3) | Oxidored_q3 | 25169 | 25342 | PF00499.17 | 2.70E-07 | NADH-ubiquinone/plastoquinone oxidoreductase chain 6 |
| scaffold3325 | (-2) | COX1 | 29419 | 29793 | PF00115.17 | 4.00E-131 | Cytochrome C and Quinol oxidase polypeptide I |

## Discussion

A BSA approach was applied to a population of perennial ryegrass segregating for CMS fertility restoration and successfully identified QTL for this trait. Using the high quality genome assembly of Italian ryegrass, over 58,000 SNPs were identified, with 103 of these showing a significant link to the restored fertility phenotype. This revealed two strong QTL for fertility restoration that align to the previously identified QTL detected by GBS (chapter 3). Sequence analysis of these two loci revealed that they are of mitochondrial origin and likely represent functional copies of the mitochondrial genes involved in the cause of CMS.

### The genetics of CMS fertility restoration in perennial ryegrass

Previously, it has been shown that the populations segregating for CMS fertility restoration used in this study had an average restoration rate of 27%, with this rate varying from 6.5 to 39.1% within individual populations (chapter 3). It has also been hypothesised that a co-dominant two-locus system might control fertility restoration (chapter 3). The data presented here support this hypothesis, as the SNP calls at the two identified QTL revealed sterility to be associated with the homozygous and fertility with a heterozygous genotype. The same result, i.e. the sterile phenotype being associated with homozygous SNP calls, was also found in the previous GBS study. Thus, the allele or haplotype present in only the fertile pools is dominantly responsible for restoration, with the segregation pattern observed suggesting that this is a co-dominant system where the restoration allele needs to be present at both loci for fertility restoration to occur.

### Mitochondrial origins of the restoration linked QTL

Analysis of the genes present at the identified QTL (Supplementary Table 1), as well as protein domain search (Table 3) and BLAST search results showed that the DNA sequence at these loci is of a mitochondrial origin. With only one gene being present within the identified QTL, a protein domain search was performed to identify any possible protein domains that were not identified by gene annotation. By this approach, any non-functional, un-annotated or pseudogenes within the target regions could be identified. This search did reveal several mitochondrial protein domains, giving more evidence for a mitochondrial origin of the fertility-restoration linked QTL. To further validate that these QTL are of mitochondrial origin, a BLAST search was performed, identifying all five QTL as being of perennial ryegrass mitochondrial genome origin.

Mitochondrial genome sequences integrated into the nuclear genome, known as nuclear mitochondrial DNA (NUMT) [27], have been reported in many species [28], including plants [29-33]. For the first time, we established the link between these NUMTs and CMS *Rf* genes. By far the most common *Rf* genes are members of a sub-set of the *PPR* gene family, known as *RFL* genes. The major difference between the *RFL* restored CMS systems and the system under study here are their origins, with *RFL* restored systems evolving naturally and the system under study here being induced through the use of a mutagenic [34]. This could explain why this CMS system is not restored by an *RFL*, as these naturally occurring systems have evolved into their current form and are observed to be tightly regulated, with CMS causal and *Rf* genes often being expressed only in reproductive tissues [35, 36]. An induced system could logically lack this kind of fine control and the CMS causal mutation could take the form of any mitochondrial interfering mutation as opposed to the more subtle open reading frames (ORFs) that can be regulated by PPR proteins, implemented in natural CMS systems [19]. If a CMS causal

mutation, from an induced system, were to disrupt the normal functioning of a mitochondria, having a back-up nuclear encoded version of this gene could restore mitochondrial function and thus fertility. This mode of action would explain the results presented here and could be verified through genomic and transcriptomic data comparison from restoring and non-restoring genomes and sterility- and non-sterility-inducing mitochondrial genomes.

The fact that in the fertile pools the QTL are heterozygous shows that the data presented here are not just the record of mitochondrial genes that have been miss-assembled into the nuclear genome. As mitochondrial inheritance is maternal [37], it is only possible for a single mitochondrial genome to be present within the whole data set as the individuals used for this analysis were germinated from seed collected from the CMS-mother line. As we see fragments of a second mitochondrial genome present in the fertile pools, the origin of these fragments must be nuclear. Also, the strong segregation of these two mitochondrial genomes indicates that they are involved in restoration. This explains why in the sterile pools the SNP calls are homozygous, as this is recording the sequence of the mitochondrial genome as compared to the NUMT fragments found within the reference sequence. In the fertile pools the heterozygous SNP calls are recording both the mitochondrial genome (which is the same as the sterile mitochondrial genome) and the NUMT fragments within the fertile pool, allowing the differentiation of the mitochondrial and NUMT genomes. This raises the question of why there are *Rf* QTL present in the Rabiosa genome. Previously it has been shown that *Rf* genes are tightly linked to their causal CMS sources [35] but for this CMS system there appears to be *Rf* genes present in a different species. This is explained by attempts to introgress the perennial ryegrass CMS system into Italian ryegrass that have been unsuccessful as no maintaining genotypes can be identified (Wilbert Lüsink, personal communication). Every cross between a male sterile perennial ryegrass plant and a fertile Italian ryegrass plant produces fertile offspring, suggesting that most, if not all, Italian ryegrass varieties contain *Rf* genes for this particular CMS causing cytoplasm.

**Comparison to QTL previously identified using genotyping by sequencing (GBS)**

Previous investigation into the genetic underpinning of the CMS fertility restoration seen in the populations used in this study revealed four QTL (chapter 3, Table 3). To identify these QTL, the Rabiosa scaffolds containing significantly associated SNPs, generated by GBS, were used for a synteny analysis to the rice genome. The scaffolds containing the five QTL identified in this study were used to identify areas of conserved gene order (for the two scaffolds containing genes) or regions of synteny to the rice genome. This revealed that all five scaffolds associated to the previously identified QTL on rice chromosome (ch) ch1 (8.2-42.6 Mb), ch3 (17.2-35.2 Mb), ch4 (25.7-33.9 Mb) and ch6 (1.4-28.1 Mb). Scaffold2257 showed synteny to both the rice QTL locations on ch3 and ch4 and scaffold3325 showed synteny to the rice QTL on ch6. As the other three significant scaffolds are less than 3kb in length and contain no genes, finding zones of synteny with the rice genome is less precise, although all three generated results matched to rice QTL on ch1 (scaffold142754 and scaffold172632) and rice QTL on ch3 (scaffold153099). The fact that the QTL from these two different approaches concur validates the accuracy of both these sets of results and narrows the size of the identified QTL from the mega-base to the kilo-base range.

**Candidate gene for fertility restoration**

From the analysis of the genes present at the QTL, it is possible to identify candidate genes for both the CMS-cause and CMS-restoration, as this should be the same genes, with dysfunctional mitochondrial genes causing CMS and functional NUMT genes restoring fertility. The best candidate from this set of genes (Supplementary Table 1) encodes the mitochondrial 18s rRNA subunit. This gene was the only identified gene within the defined QTL and contains a highly significant three bp polymorphism 26bp from its beginning. As the 18S rRNA functions in the small ribosomal subunit at the active centre of protein synthesis [38], an alteration in this gene could affect the efficiency of mitochondrial protein synthesis.

These results reveal a previously unknown CMS/restoration mechanism involving nuclear copies of mitochondrial genes, with the 18S rRNA subunit being directly implicated as both the source (in the mitochondrial genome) and the restorer (in the nuclear genome) of CMS. The QTL identified here provide an excellent target for marker development, where these markers could be used in practical hybrid breeding programs to identify restoring genotypes.

# References


1. Sax, K., 1923. The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics, 8(6), p.552.
2. Tanksley, S.D., 1993. Mapping polygenes. Annual review of genetics, 27(1), pp.205-233.
3. Lee, M., 1995. DNA markers and plant breeding programs. *Advances in agronomy*, *55*, pp.265-344.
4. Schneeberger, K., 2014. Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nature Reviews Genetics, 15(10), pp.662-676.
5. Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C. and Estoup, A., 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Molecular Ecology, 22(14), pp.3766-3779.
6. Rode, N.O., Holtz, Y., Loridon, K., Santoni, S., Ronfort, J. and Gay, L., 2017. How to optimize the precision of allele and haplotype frequency estimates using pooled-sequencing data. Molecular Ecology Resources.
7. Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy, A.A. and Kruglyak, L., 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature, 464(7291), pp.1039-1042.
8. Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Boneschi, F.M., D'Alfonso, S. and De Bellis, G., 2016. Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. Scientific reports, 6.
9. Wei, K.H.C., Reddy, H.M., Rathnam, C., Lee, J., Lin, D., Ji, S., Mason, J.M., Clark, A.G. and Barbash, D.A., 2017. A pooled sequencing approach identifies a candidate meiotic driver in Drosophila. Genetics, 206(1), pp.451-465.
10. Fracassetti, M., Griffin, P.C. and Willi, Y., 2015. Validation of pooled whole-genome re-sequencing in Arabidopsis lyrata. PloS one, 10(10), p.e0140462.
11. Wenger, J.W., Schwartz, K. and Sherlock, G., 2010. Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. PLoS genetics, 6(5), p.e1000942.
12. Michelmore, R.W., Paran, I. and Kesseli, R.V., 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. Proceedings of the national academy of sciences, 88(21), pp.9828-9832.
13. Quarrie, S.A., Lazić-Jančić, V., Kovačević, D., Steed, A. and Pekić, S., 1999. Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. Journal of experimental botany, 50(337), pp.1299-1306.
14. Venuprasad, R., Dalid, C.O., Del Valle, M., Zhao, D., Espiritu, M., Cruz, M.S., Amante, M., Kumar, A. and Atlin, G.N., 2009. Identification and characterization of large-effect quantitative trait loci for grain yield under lowland drought stress in rice using bulk-segregant analysis. Theoretical and Applied Genetics, 120(1), pp.177-190.
15. Poulsen, D.M.E., Henry, R.J., Johnston, R.P., Irwin, J.A.G. and Rees, R.G., 1995. The use of bulk segregant analysis to identify a RAPD marker linked to leaf rust resistance in barley. Theoretical and applied genetics, 91(2), pp.270-273.
16. Shen, X., Zhou, M., Lu, W. and Ohm, H., 2003. Detection of Fusarium head blight resistance QTL in a wheat population using bulked segregant analysis. TAG Theoretical and Applied Genetics, 106(6), pp.1041-1047.
17. Hanson MR, Bentolila S. 2004. Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell. 16:154-S69.
18. Schnable PS, and Wise RP. 1998. The molecular basis of cytoplasmic male sterility and fertility restoration. Trends in Plant Science. 3(5):175-80.
19. Barkan A, Small I. 2014. Pentatricopeptide Repeat Proteins in Plants. S. S. Merchant (ed.), Annual Review of Plant Biology. 65:415-442.
20. Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), pp.357-359.

21. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), pp.2078-2079.
22. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. and Salazar, G.A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic acids research, 44(D1), pp.D279-D285.
23. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology, 215(3), pp.403-410.
24. Velmurugan J, Mollison E, Barth S, Marshall D, Milne L, Creevey CJ, Lynch B, Meally H, McCabe M, Milbourne D. An ultra-high density genetic linkage map of perennial ryegrass (Lolium perenne) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly. Annals of botany. 2016 Jun 6;118(1):71-87.
25. Byrne, S.L., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., Mayer, K., Campbell, J.D., Czaban, A., Hentrup, S. and Panitz, F., 2015. A synteny-based draft genome sequence of the forage grass Lolium perenne. The Plant Journal, 84(4), pp.816-826.
26. Islam, M.S., Studer, B., Byrne, S.L., Farrell, J.D., Panitz, F., Bendixen, C., Møller, I.M. and Asp, T., 2013. The genome and transcriptome of perennial ryegrass mitochondria. BMC genomics, 14(1), p.202.
27. Lopez, J.V., Yuhki, N., Masuda, R., Modi, W. and O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. Journal of Molecular Evolution, 39(2), pp.174-190.
28. Hazkani-Covo, E., Zeller, R.M. and Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS genetics*, *6*(2), p.e1000834.
29. Noutsos, C., Richly, E. and Leister, D., 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Research, 15(5), pp.616-628.
30. Michalovova, M., Vyskot, B. and Kejnovsky, E., 2013. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. Heredity, 111(4), pp.314-320.
31. Huang, C.Y., Grünheit, N., Ahmadinejad, N., Timmis, J.N. and Martin, W., 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiology, 138(3), pp.1723-1733.
32. Lough, A.N., Roark, L.M., Kato, A., Ream, T.S., Lamb, J.C., Birchler, J.A. and Newton, K.J., 2008. Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics, 178(1), pp.47-55.
33. Zhang, J., Jia, J., Breen, J. and Kong, X., 2011. Recent insertion of a 52-kb mitochondrial DNA segment in the wheat lineage. Functional & integrative genomics, 11(4), pp.599-609.
34. Gaue, I. and Baudis, H., 2006. *Male sterility in grasses of the genus Lolium*. U.S. Patent Application 11/520,186.
35. Tang, H., Xie, Y., Liu, Y.G. and Chen, L., 2017. Advances in understanding the molecular mechanisms of cytoplasmic male sterility and restoration in rice. Plant reproduction, pp.1-6.
36. Luo, D., Xu, H., Liu, Z., Guo, J., Li, H., Chen, L., Fang, C., Zhang, Q., Bai, M., Yao, N. and Wu, H., 2013. A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. Nature genetics, 45(5), pp.573-577.
37. Reboud, X. and Zeyl, C., 1994. Organelle inheritance in plants. Heredity, 72(2), pp.132-140.
38. Decatur, W.A. and Fournier, M.J., 2002. rRNA modifications and ribosome function. Trends in biochemical sciences, 27(7), pp.344-351.

# Supplementary Data

**Supplementary Table 1.** The genes identified in scaffold2257 and scaffold3325 and their GO annotation.

**Scaffold2257**

| Rabiosa Gene | Arabidopsis homologue | Description |
|---|---|---|
| MSTRG.21888.1 | AT5G51630.1 | Disease resistance protein (TIR-NBS-LRR class) family |
| MSTRG.21889.1 | AT3G03710.1 | RIF10, PNP polyribonucleotide nucleotidyltransferase, putative |
| MSTRG.21890.1 | AT3G03710.1 | RIF10, PNP polyribonucleotide nucleotidyltransferase, putative |
| MSTRG.21891.1 | AT3G04010.1 | O-Glycosyl hydrolases family 17 protein |
| MSTRG.21892.1 | AT3G24330.1 | O-Glycosyl hydrolases family 17 protein |
| MSTRG.21893.1 | AT5G40190.1 | RNA ligase/cyclic nucleotide phosphodiesterase family protein |
| MSTRG.21894.1 | AT3G12630.1 | A20/AN1-like zinc finger family protein |
| MSTRG.21895.1 | AT3G12630.1 | A20/AN1-like zinc finger family protein |
| MSTRG.21896.1 | AT5G07480.1 | KUOX1 KAR-UP oxidoreductase 1 |
| MSTRG.21897.1 | AT5G07480.1 | KUOX1 KAR-UP oxidoreductase 1 |
| MSTRG.21898.1 | AT5G42540.2 | XRN2 exoribonuclease 2 |
| MSTRG.21899.1 | AT4G10710.1 | SPT16 global transcription factor C |
| MSTRG.21900.1 | AT4G10710.1 | SPT16 global transcription factor C |
| MSTRG.21901.1 | AT4G04530.1 | transposable element gene |
| MSTRG.21902.1 | AT1G55510.1 | BCDH BETA1 branched-chain alpha-keto acid decarboxylase |
| MSTRG.21903.1 | AT1G55510.1 | BCDH BETA1 branched-chain alpha-keto acid decarboxylase |
| MSTRG.21904.1 | AT2G23840.1 | HNH endonuclease |
| MSTRG.21905.1 | AT2G23840.1 | HNH endonuclease |
| MSTRG.21906.1 | AT1G55480.1 | ZKT protein containing PDZ domain, a K-box domain, and a TPR region |
| MSTRG.21907.1 | AT1G55480.1 | ZKT protein containing PDZ domain, a K-box domain, and a TPR region |
| MSTRG.21908.1 | AT5G18200.1 | UTP: galactose-1-phosphate_uridylyltransferases |
| MSTRG.21909.1 | AT5G18200.1 | UTP: galactose-1-phosphate_uridylyltransferases |
| MSTRG.21910.1 | AT3G03940.1 | Protein kinase family protein |
| MSTRG.21911.1 | AT3G13670.1 | Protein kinase family protein |
| MSTRG.21912.1 | AT5G50580.2 | SAE1B, AT-SAE1-2 SUMO-activating enzyme 1B |
| MSTRG.21913.1 | AT5G50580.2 | SAE1B, AT-SAE1-2 SUMO-activating enzyme 1B |
| MSTRG.21914.1 | AT5G19090.1 | Heavy metal transport/detoxification superfamily protein |
| MSTRG.21915.1 | AT3G26540.1 | Tetratricopeptide repeat (TPR)-like superfamily protein |
| MSTRG.21916.1 | AT4G14830.1 | HSP1 unknown protein |
| MSTRG.21917.1 | AT3G22440.1 | FRIGIDA-like protein |
| MSTRG.21918.1 | AT3G22440.1 | FRIGIDA-like protein |
| MSTRG.21919.1 | AT5G36740.1 | Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger protein |
| MSTRG.21920.1 | AT5G36740.1 | Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger protein |
| MSTRG.21921.1 | AT1G34030.1 | Ribosomal protein S13/S18 family |
| MSTRG.21922.1 | AT1G34030.1 | Ribosomal protein S13/S18 family |
| MSTRG.21923.1 | ATMG01390 | RRN18 Mitochondrial 18S ribosomal RNA |
| MSTRG.21924.1 | AT5G42290.1 | transcription activator-related |
| MSTRG.21925.1 | AT1G56140.1 | Leucine-rich repeat transmembrane protein kinase |
| MSTRG.21926.1 | AT3G63270.1 | PIF / Ping-Pong family of plant |

| | | |
|---|---|---|
| MSTRG.21927.1 | AT3G63270.1 | PIF / Ping-Pong family of plant |
| MSTRG.21928.1 | AT4G37680.1 | HHP4 heptahelical protein 4 |
| MSTRG.21929.1 | AT4G37680.1 | HHP4 heptahelical protein 4 |
| MSTRG.21930.1 | AT5G46870.1 | RNA-binding (RRM/RBD/RNP motifs) family protein |
| MSTRG.21931.1 | AT5G46870.1 | RNA-binding (RRM/RBD/RNP motifs) family protein |
| MSTRG.21932.1 | AT5G14250.1 | COP13, CSN3, FUS11 Proteasome component (PCI) domain protein |
| MSTRG.21933.1 | AT3G49810.1 | ARM repeat superfamily protein |
| MSTRG.21934.1 | AT2G22870.1 | EMB2001 P-loop containing nucleoside triphosphate hydrolases |
| MSTRG.21935.1 | AT2G22870.1 | EMB2001 P-loop containing nucleoside triphosphate hydrolases |
| MSTRG.21936.1 | AT2G22860.1 | ATPSK2, PSK2 phytosulfokine 2 precursor |
| MSTRG.21937.1 | AT2G22860.1 | ATPSK2, PSK2 phytosulfokine 2 precursor |
| MSTRG.21938.1 | AT4G13340.1 | Leucine-rich repeat (LRR) family protein |
| MSTRG.21939.1 | AT3G08850.1 | RAPTOR1B, ATRAPTOR1B, RAPTOR1 HEAT repeat |
| MSTRG.21940.1 | AT3G08850.1 | RAPTOR1B, ATRAPTOR1B, RAPTOR1 HEAT repeat |
| MSTRG.21941.1 | AT1G58440.1 | XF1, SQE1 FAD/NAD(P)-binding oxidoreductase family protein |
| MSTRG.21942.1 | AT1G58440.1 | XF1, SQE1 FAD/NAD(P)-binding oxidoreductase family protein |
| MSTRG.21943.1 | AT1G10070.1 | ATBCAT-2, BCAT-2 branched-chain amino acid transaminase 2 |
| MSTRG.21944.1 | AT1G10070.1 | ATBCAT-2, BCAT-2 branched-chain amino acid transaminase 2 |
| MSTRG.21945.1 | AT5G06710.1 | HAT14 homeobox from Arabidopsis thaliana |
| MSTRG.21946.1 | AT5G06710.1 | HAT14 homeobox from Arabidopsis thaliana |

**Scaffold3325**

| Rabiosa Gene | Arabidopsis homologue | Description |
|---|---|---|
| MSTRG.30872.1 | AT4G39540.3 | SK2 shikimate kinase 2 |
| MSTRG.30873.1 | AT4G39540.3 | SK2 shikimate kinase 2 |
| MSTRG.30874.1 | AT1G50140.1 | P-loop containing nucleoside triphosphate hydrolases |
| MSTRG.30875.1 | AT1G50140.1 | P-loop containing nucleoside triphosphate hydrolases |
| MSTRG.30876.1 | AT3G28690.2 | Protein kinase superfamily protein |
| MSTRG.30877.1 | AT5G15280.1 | Pentatricopeptide repeat (PPR) superfamily protein |
| MSTRG.30878.1 | AT2G43140.2 | basic helix-loop-helix (bHLH) DNA-binding superfamily protein |
| MSTRG.30879.1 | AT2G43140.2 | basic helix-loop-helix (bHLH) DNA-binding superfamily protein |
| MSTRG.30880.1 | AT3G61930.1 | Unknown protein |
| MSTRG.30881.1 | AT1G74520.1 | ATHVA22A, HVA22A HVA22 homologue A |
| MSTRG.30882.1 | AT1G74520.1 | ATHVA22A, HVA22A HVA22 homologue A |
| MSTRG.30883.1 | AT3G47620.1 | AtTCP14, TCP14 TEOSINTE BRANCHED, cycloidea and PCF (TCP) |

# Chapter 5

# Expression analysis of genes involved in the restoration of cytoplasmic male sterility in perennial ryegrass (*Lolium perenne* L.)

**Timothy Sykes[1], Steven Yates[1], Martin Schuler[1], Maximilian Vogt[1] and Bruno Studer[1]***

[1] Institute of Agricultural Sciences, Molecular Plant Breeding, ETH Zurich, 8092 Zurich, Switzerland.

*Corresponding author: bruno.studer@usys.ethz.ch

## Abstract

Gene expression studies are vital to our understanding of living organisms, allowing us to not only identify what genes are present in a genome but also how an organism uses its genes. Among different technologies for gene expression profiling, RNA sequencing has evolved as the most widely used and precise tool. To understand what genes are involved in cytoplasmic male sterility (CMS) and its restoration in perennial ryegrass (*Lolium perenne* L.), this study aimed at precise profiling of the genes expressed in different tissue types of cytoplasmic male sterile and fertile plants.

Total RNA of leaf, flower and anther tissue in a cytoplasmic male sterile, a fertile maintainer, a fertile restorer and a genotype where CMS had been restored was extracted and used for next generation sequencing. Genome-wide transcriptome profiling and the expression of genes present in previously identified quantitate trait loci (QTL) revealed three nuclear encoded mitochondrial genes (18S 5S rRNA, COX1 and COX3) that showed increased expression in the anthers of plants known to contain a restorer gene. These results support the theory that this CMS system is restored by two co-dominant loci as well as confirming the mitochondrial origins of these nuclear genes. Finally, this data also suggests that restored perennial ryegrass hybrids have a fitness advantage over unrestored hybrids, although further study is required to fully elucidate this phenomenon.

The results presented here suggest a possible mechanism for both the cytoplasmic cause of male sterility and the nuclear encoded fertility restoration as well as providing gene targets for further study and marker development.

**Keywords:** RNA sequencing (RNASeq), gene expression, perennial ryegrass (*Lolium perenne* L.), cytoplasmic male sterility (CMS), restoration of fertility

## Introduction

RNA molecules are essential for all living cells as they act as the intermediary between genes and their encoded proteins [1-3]. Information about the abundance of different RNAs allows us to move beyond the presence/absence understanding of the genome towards a more nuanced picture: that of gene expression. Gene expression data can tell us not only if a particular gene is present within an organism but when, where and how much that particular gene is transcribed [4]. This enables the study of how changes in the levels of gene expression can change between tissue types and over time, and how this can be affected by an organism's environment.

The first high-throughput techniques designed to study RNA became available in the 1990s with the advent of the expressed sequence tag (EST) method [5]. This method partially sequenced complementary DNA (cDNA) clones of RNA, uncovering both their sequence and abundance. Despite the EST method being only semi-quantitative, it was able to identify new genes in previously un-sequenced genomes. However, the high cost of EST analysis per gene led researchers to develop methods, such as the serial analysis of gene expression (SAGE) method. SAGE, by sequencing a short portion of a cDNA molecule, managed to reduce the overall cost per gene [6]. EST and SAGE were quickly superseded in the late 1990s by DNA microarray technology, which proved more affordable for large-scale experiments [7]. DNA microarrays depend on the hybridisation of cDNA derived fluorescently labelled targets to probes that have been affixed or printed on a surface [8]. Although this method presents a relatively cheap way to capture gene expression on a large scale, it does require knowledge of the sequence of the transcripts to be studied when designing probes and thus cannot be used for gene discovery. Another concern for DNA microarrays is cross-hybridization, where transcripts other than the ones intended hybridize to a probe leading to false positive results or an overestimation of gene expression [9].

In the early 2000s, the advent of next generation sequencing (NGS) allowed researchers to gain access to previously unthinkable amounts of DNA sequence data [10-15]. This shift in sequencing technology was rapidly applied to gene expression studies and the technique of RNA Sequencing (RNASeq) was born in its many forms [16-20]. As compared to microarrays, RNASeq allows the capture of unknown RNA sequences, enabling its use for both gene detection and expression studies in previously unsequenced species [21]. Crucially, RNASeq also offers a greater dynamic range of detection and lower background noise, permitting the discovery of relatively rare RNA molecules [22].

Armed with this new technique, gene expression profiling was widely applied to understand the genetic architecture of certain traits and diseases in multiple species [23]. Some of the most successful and interesting applications of RNASeq range from fundamental understandings, with the revelation that although only 3% of the human genome encodes genes, up to 85% is in fact transcribed [24, 25], to specific discoveries such as the native RNA molecule required for the functioning of the CRISPER/Cas9 system [26]. In the field of agricultural science, RNASeq has been used to study agronomically important traits in several species and continues to be utilised today. Recent studies have revealed the effects of drought on gene expression in maize [27], the changes of gene expression upon fire-blight infection in apples (the so called *reactome*) [28] and the validation of seed defence biopriming as an alternative to chemical crop protection in cucumber and pepper [29]. RNASeq continues to be the workhorse of gene expression studies, generating knowledge that can be translated into real-world applications designed to improve the quality and sustainability of agriculture.

CMS is one such important trait used as a pollination control mechanism during the production of large quantities of hybrid seed [40]. For this use to be successful, plant breeders need to track restoration genes throughout their breeding material to ensure that no unwanted fertility restoration occurs that could negatively affect the quality of hybrid seed produced. In order to identify causative genes and the molecular pathways involved in both CMS and restoration of fertility, gene expression studies have been performed in several species including chili pepper (*Capsicum annuum* L.), cotton (*Gossypium arboreum* L.), rapeseed (*Brassica napus* L.), cabbage (*Brassica oleracea* L.), soybean (*Glycine max* L.), radish (*Raphanus sativus* L.) and watermelon (*Citrullus lanatus* L.) [30-39]. These studies showed large changes in gene expression, involving thousands of genes, within genetic pathways including starch and sucrose metabolism [30], oxidation-reduction processes [31], signalling [32], metabolic pathways [33], ion binding [34], biosynthesis of secondary metabolites [35], glycolysis [36] and organelle genes [38]. However, these studies were unable to accurately define the CMS restoration causes.

The main goal of this study was the analysis of the differences in gene expression between four different genotypes (maintainer, restorer, sterile and restored) in three tissue types (leaf, flower and anther) in a perennial ryegrass CMS system. This design was chosen as previously described CMS-restoration systems have shown that the expression of restorer genes can be tightly linked to the development of pollen precursor tissues such as tapetum [41]. This will allow the identification of genes expressed uniquely during pollen formation in restored or restoring genotypes, or conversely, genes uniquely expressed in the sterile genotype.

## Methods

### Tissue collection, RNA isolation and sequencing

Three tissue types were sampled (leaf, flower and anther) into 1.5 ml Eppendorf tubes and immediately snap frozen in liquid nitrogen. To ensure as much consistency as possible, all tissue samples were sampled between 11am and 1pm during a ten day window that encompassed flowering period in 2016.

RNA isolation was performed using the TRIzol reagent (Thermo Fisher Scientific Inc., USA) as described by Rio *et al.* [42] with samples being homogenised using liquid nitrogen and a precooled mortar and pestle. Quality control (QC) was performed on an Agilent 2200 TapeStation using the High Sensitivity RNA ScreenTape Kit (Agilent Technologies, USA). Any samples with an RNA integrity number (RIN) of less than 6 were re-extracted from new tissue samples. All sampling, RNA isolation and QC was performed on three biological replicates per genotype and tissue type, resulting in a total of 36 samples. NGS and library construction was delivered via the BBSRC National Capability in Genomics (BB/CCG1720/1) at Earlham Institute by members of the Genomics Pipelines Group (Earlham Institute, UK).

### RNA sequencing data analysis

RNASeq data was aligned to the Rabiosa reference genome using STAR (V.020201) [43]. To identify transcripts and gene models from the data, the resulting alignments were analysed using cufflinks (V.2.2.1), separately for each replicate. The resulting output (transcript.gtf files) were merged using the cuffmerge program. This generated a consensus file of transcripts which was then used as the basis to test for differential expression using the cuffdiff program. The following modifications were used; "-max-bundle-frags 5000000" and "--library-type fr-firststrand", the former of which reduced over-expressed genes and the latter accounted for the STAR alignment software used. Tests for all gene models (XLOC) were made for all pairwise combinations of treatments (Restored, Restorer, Maintainer, Sterile & Flowers, Anthers, Leaves).

To identify coding sequences (CDS), the spliced exons for each GFF transcript were retrieved using gffread (part of the Cufflinks tool suite). To identify the correct open reading frames (ORF) for protein sequences, the program ORrfPredictor (version: 3.0) [44] was used. For frame selection, the transcripts were first BLASTX [45] searched against a protein database consisting of the proteomes from Arabidopsis (*Arabidopsis thaliana* L) TAIR (version: 10, [46]), rice (*Oryza sativa* L) (downloaded from Ensembl [47]), soybean (Ensembl), poplar (*Populus trichocarpa* L) (Ensembl) and cassava (*Manihot esculenta* L) [48], downloaded from Phytozome [49]. This database, although not exhaustive, provided a broad basis of existing plant proteins. OrfPredictor was then used to identify CDS by use of the best BLAST hits frame selection. In the absence of a homologous BLAST hit, OrfPredictor selected the longest ORF. These results were then used to annotate the GFF file created by Cufflinks for CDS using scripts kindly provided by Palmieri et al. [50]. For functional annotation, the resulting proteins were searched against the Arabidopsis TAIR10 proteome using BLASTP. The functional annotation of the best BLAST hit (based on E-value, minimum 1e-15) protein was used to assign annotations for functional description and gene ontology (GO). Candidate genes were further interrogated with the use of BLAST searches NCBI (National Center for Biotechnology Information, USA) using the BLAST search tool [51] and 'localizer' [52] for mitochondrial transit peptides prediction.

## Identification and analysis of differentially expressed genes

To identify differentially expressed genes, the output from cuffdiff was loaded into R statistical environment [53]. Genes were selected as differentially expressed when satisfying the following criteria: i) the gene must be expressed above 2 FPKM (fragments per kilo base of transcript per million mapped reads), to remove genes with low abundance where expression estimates may be stochastic, ii) the fold change must be greater than 2 fold between a pairwise comparison, and iii) the corrected *P*-value must be less than 5% for significance. The subset of genes that met these criteria were used for further analysis.

To identify global changes in gene expression between tissues and genotypes, a holistic overview was taken of all differentially expressed genes. To analyse the results the gene expression values were normalized into Z-score (Z-score is the observed expression minus the mean expression divided by the standard deviation per gene). The primary purpose of this was to mitigate against genes with large abundance and changes, dominating the output. The resulting Z-scores were then subject to principal component analysis (PCA) and the first two components were extracted and used for graphical illustration. This was done for all replicates and the average values per treatment. From these analyses, it was apparent that the treatments could be divided into three clusters. To validate this, the Euclidean distance between genes was made form the first two principle components. However, for computational reasons, genes were further filtered by excluding genes with a mean FPKM less than 2 across all treatments. The relationships between genes was then calculated from the distance matrix using hierarchical clustering (R, “complete method”) and the function “cutree” was used to partition the genes into three clusters. To assess clusters for changes in biological processes, gene ontology (GO) analysis was implemented using TopGO from Bioconductor in R. A one-sided Fisher`s exact test was used to identify enriched GO terms with a minimum *P*-value of $< 0.05$. The GO terms were visualised using REVIGO (Ruđer Bošković Institute, Croatia) which groups terms using word space.

## Chlorophyll content measurements

Chlorophyll content was estimated using a Soil-Plant Analyses Development (SPAD) unit (Minolta Camera Co., Japan) as described by Rodrigues *et al* (2000) [54]. Four populations of plants, from the NPZ hybrid breeding program, were chosen that were segregating for fertility restoration with six leaves sampled from each plant and five measurements taken per leaf.

## Results

### RNA sequencing and read mapping to the high-quality genome assembly of Italian ryegrass

Illumina sequencing of the 36 samples returned an average of 30,008,998 reads per sample. The details of the assembled transcriptome can be seen in table 1, where the longest transcript at each gene was considered canonical.

**Table 1.** Details of the transcriptome assembly.

| | |
|---|---|
| **Number of genes** | **105,860** |
| **Total length** | 157 Mb |
| **Mean gene length** | 1,487 bp |
| **Median gene length** | 1240 bp |
| **N50** | 1887 bp |
| **Smallest gene** | 126 bp |
| **Largest gene** | 14,569 bp |

### Analysis of general differential expression data between genotype and tissue type samples

After filtering, 56,523 genes were identified that were considered differentially expressed between at least two samples. In order to both ascertain if there were any broad gene expression differences between genotype and tissue type combinations, and to check the consistency of the biological replicates, a PCA was performed. The accumulated variance explained was calculated with the first principal component explaining 38% of the variance, the top two explaining 67%, top three 78%, top four 85%, top five 90%, top six 93%, top seven 95%, top eight 97%, top nine 98%, top ten 99% and top twelve explaining almost 100% of the variance (Figure 1). The first two factors, explaining two thirds of the variance, were plotted for all samples (Figure 2).

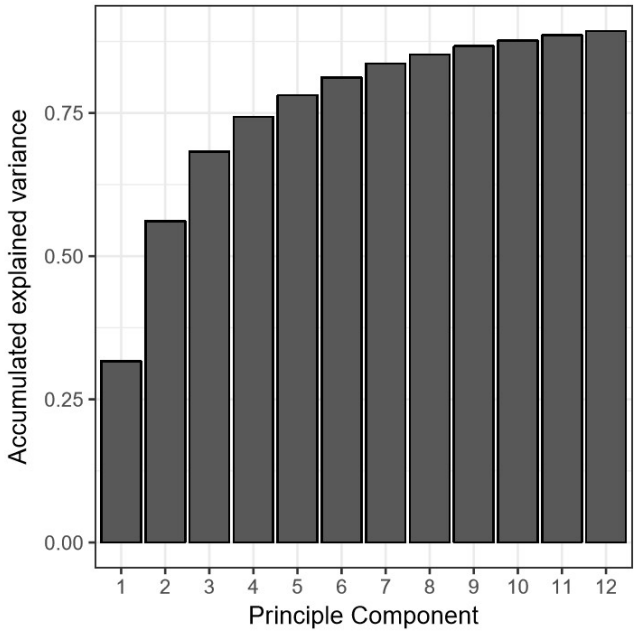
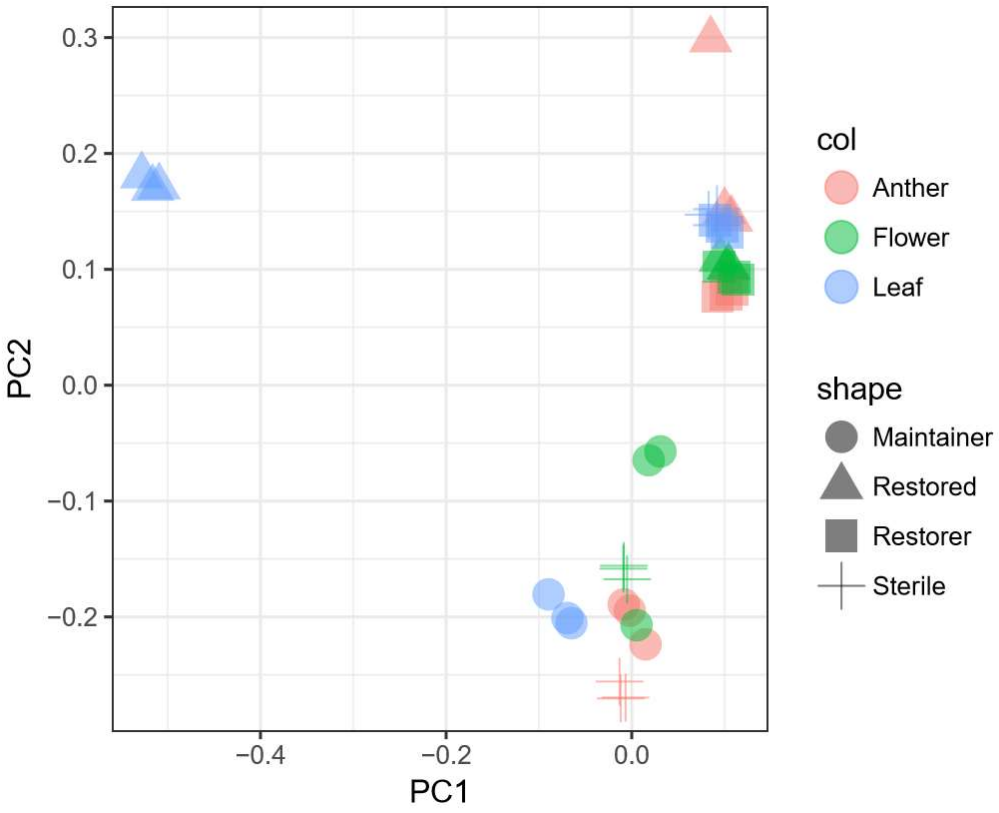



**Figure 1.** Results of the principal component analysis showing accumulated explained variance.

**A**




**B**

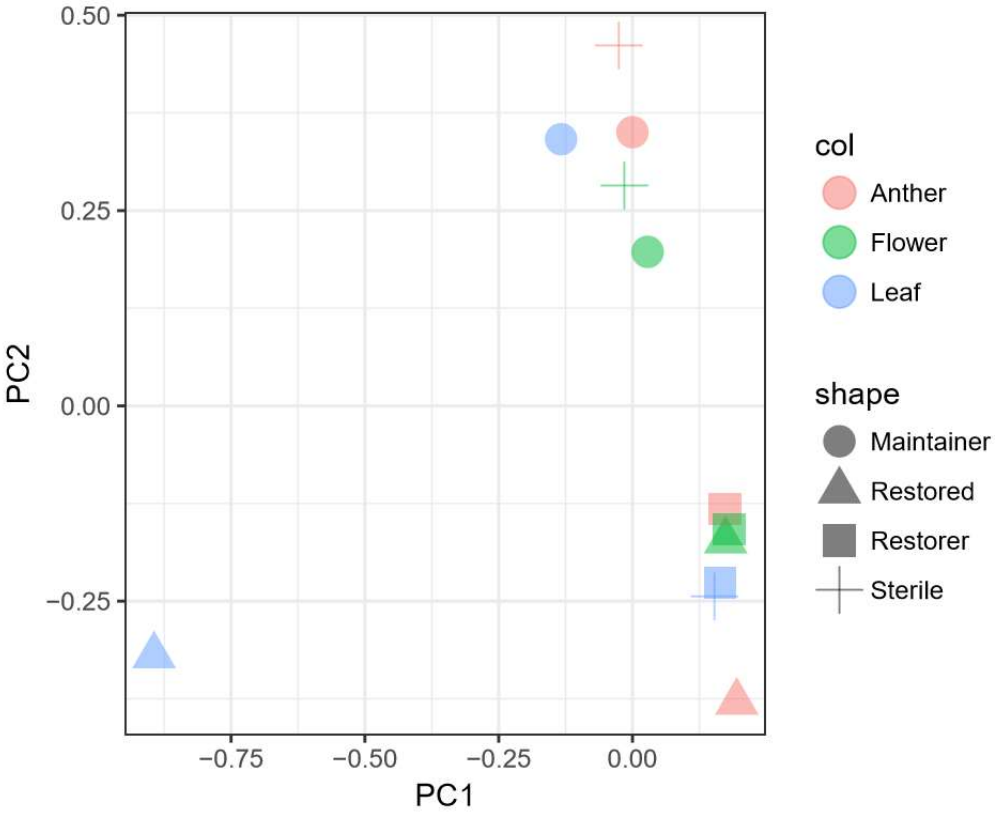



**Figure 2. A** Top two factors from the principal component analysis (PCA) for all 36 samples. **B.** Top two factors from the principal component analysis for each of the twelve treatments.

This analysis revealed that the biological replicates were consistent and three clusters could be identified. To validate these three groups, the Euclidean distance between genes was made from the first two principal components (Figure 3), revealing that restored-leaf samples made up one group, all maintainer (leaf, flower and anther) as well as sterile anther and sterile flower samples made up another with all restorer, restored flower and anther and sterile leaf samples making up the last group. Setting aside the grouping of restored leaf samples, this leaves two groups made up of either maintainer and sterile or restorer and restored samples (with the exception of sterile-leaf samples).

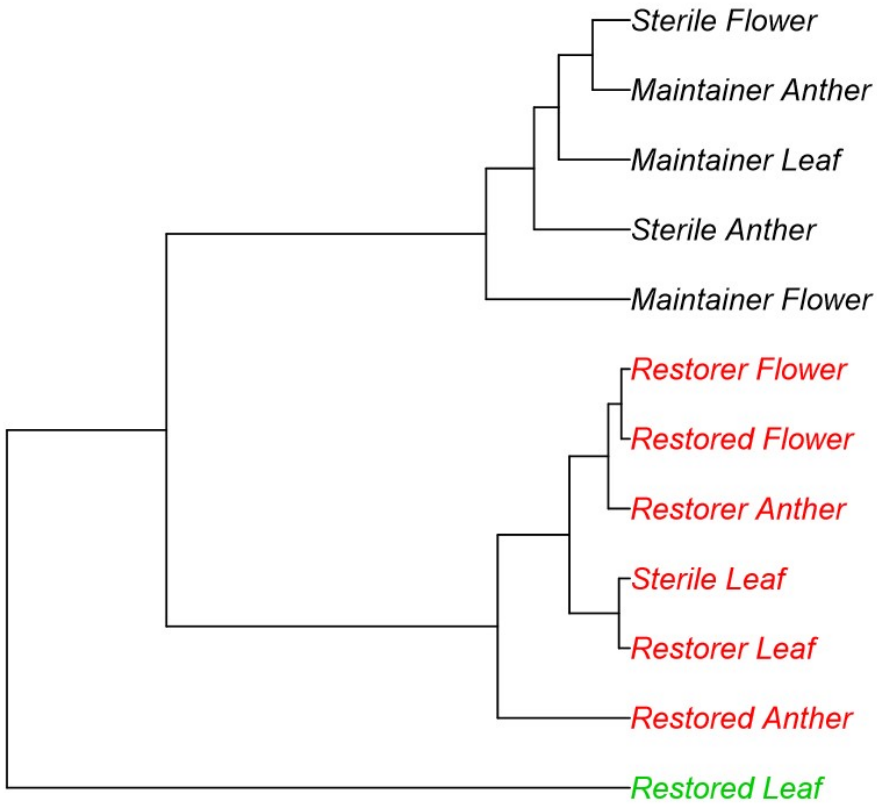



**Figure 3.** Dendrogram of all twelve samples identifying the three clusters.

To further investigate what might be driving the differential expression between these three groups, all genes with a mean FPKM of less than 2 across all treatments were excluded, leaving 29’997 genes. These, remaining, genes were then analysed for overall changes in expression values (Z-score) between all genotype and tissue type combinations (Figure 4).

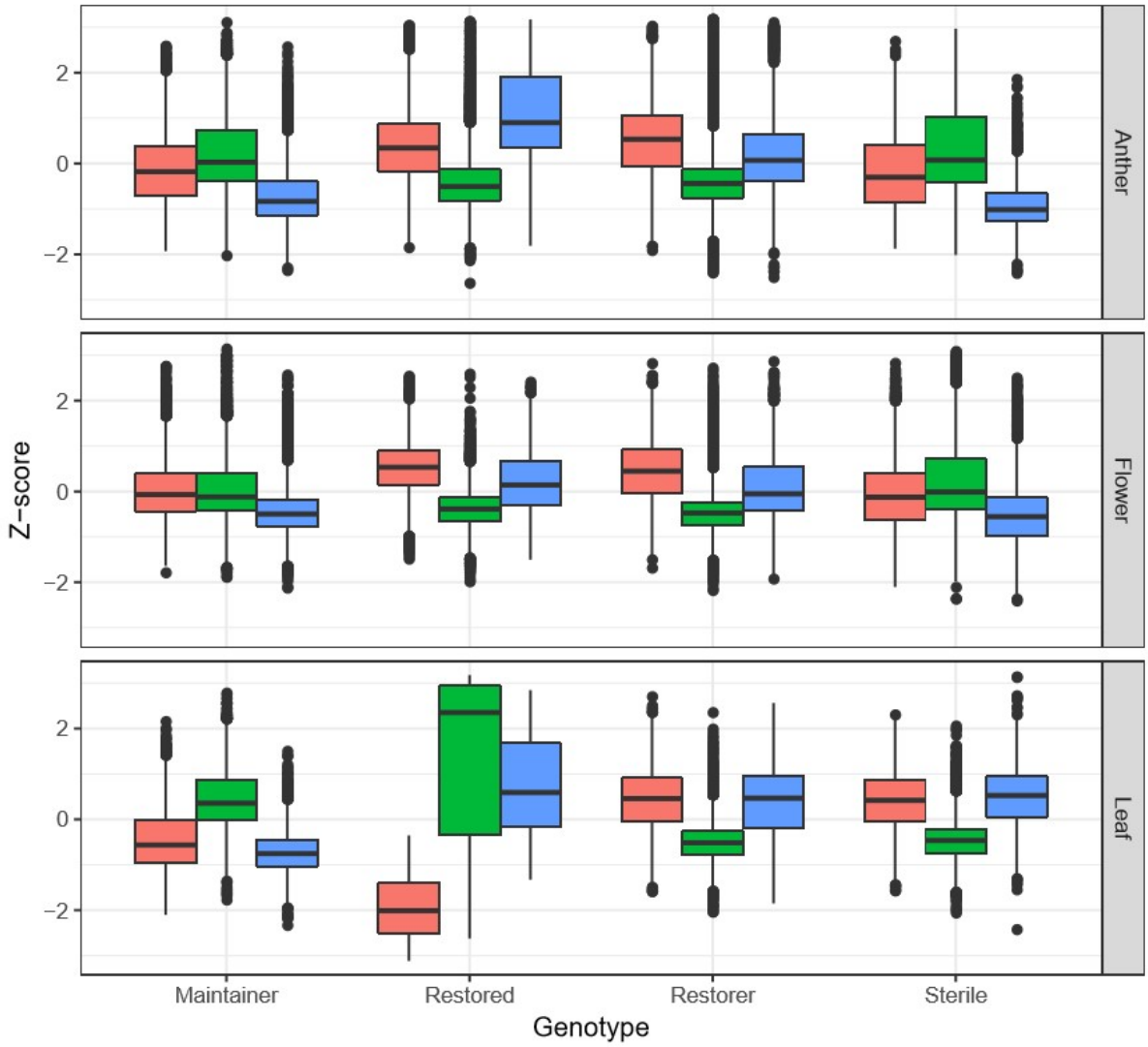



**Figure 4.** Box plots showing the changes in general gene expression (Z-score) of the genes in the three gene clusters (given in different colours). These expression scores are shown against both the genotype and tissue types.

Analysis of the GO annotation for the genes present in these clusters using REVIGO (Supplementary Figure 2) revealed that cluster one (Figure 4, red) included genes that are involved in cell cycle processes, with terms including; 'methylation' (GO:0032259), 'metabolic process' (GO:0008152), 'cell cycle phase' (GO:0022403) and nucleotide biosynthetic process (GO:0009165) being enriched. Cluster two (Figure 4, green) contained genes with photosynthesis related annotation including; 'photosynthesis, light harvesting' (GO:0009765), 'photosynthetic electron transport chain' (GO:0009767), and 'response to blue light' (GO:0009637). Cluster three (Figure 4, blue) also contains genes involved in cell cycle processes but has more enriched terms for plastid and chromosome organisation including; 'chromatin modification' (GO:0016568), 'plastid organization' (GO:0009657) and 'cytoskeleton organization' (GO:0007010). The most notable shifts in gene expression within any single genotype and tissue type combination were seen in restored leaf samples with an increase in gene expression in cluster two and a decrease of gene expression in cluster one.

The analysis also revealed pleiotropic effects of the CMS and restoration system, manifested in clusters one and three, which showed increased expression in all samples

containing *Rf* genes (restored and restorer), as well as in sterile-leaf samples (with the exception of restored-leaf which shows a large decrease in expression of cluster one). Within anther and flower tissue samples, the presence/absence of *Rf* genes is accounted for by these shifts in the expression profiles of clusters one and three, although within leaf samples these observations do not hold true.

**Chlorophyll content analysis of plants from populations segregating for fertility restoration**

Given that the analysis of cluster 2 (Figure 4) indicated an increase in photosynthetic activity in restored leaf tissue, chlorophyll measurements using a SPAD were taken to assess whether restored leaf tissue contained more chlorophyll. Although populations including the plants used for RNA sampling were not available, other populations of plants made up of restored (fertile) and unrestored (sterile) plants were investigated (Figure 5).

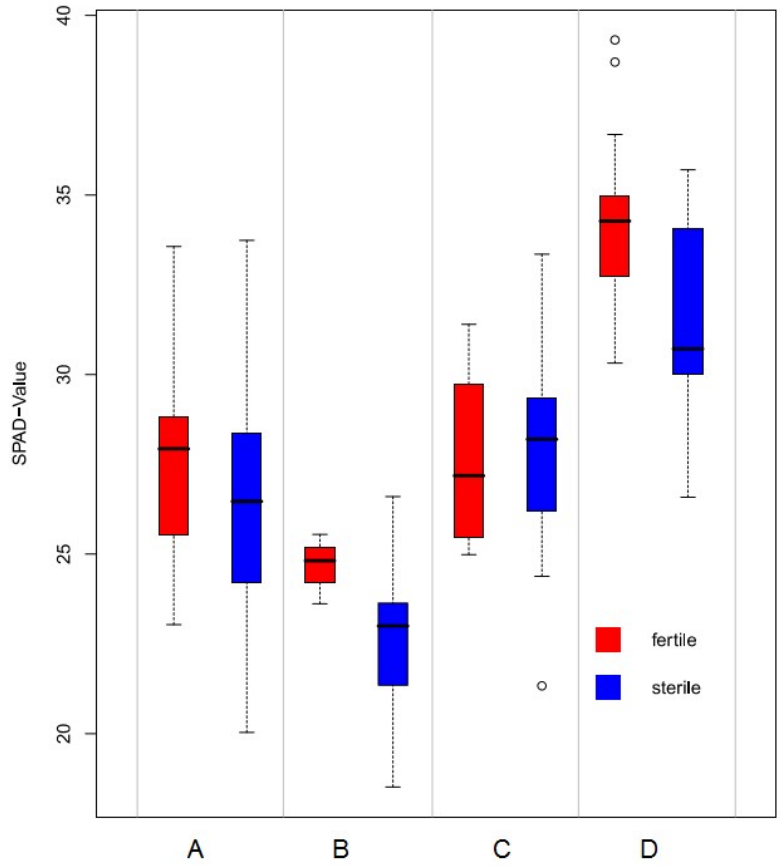


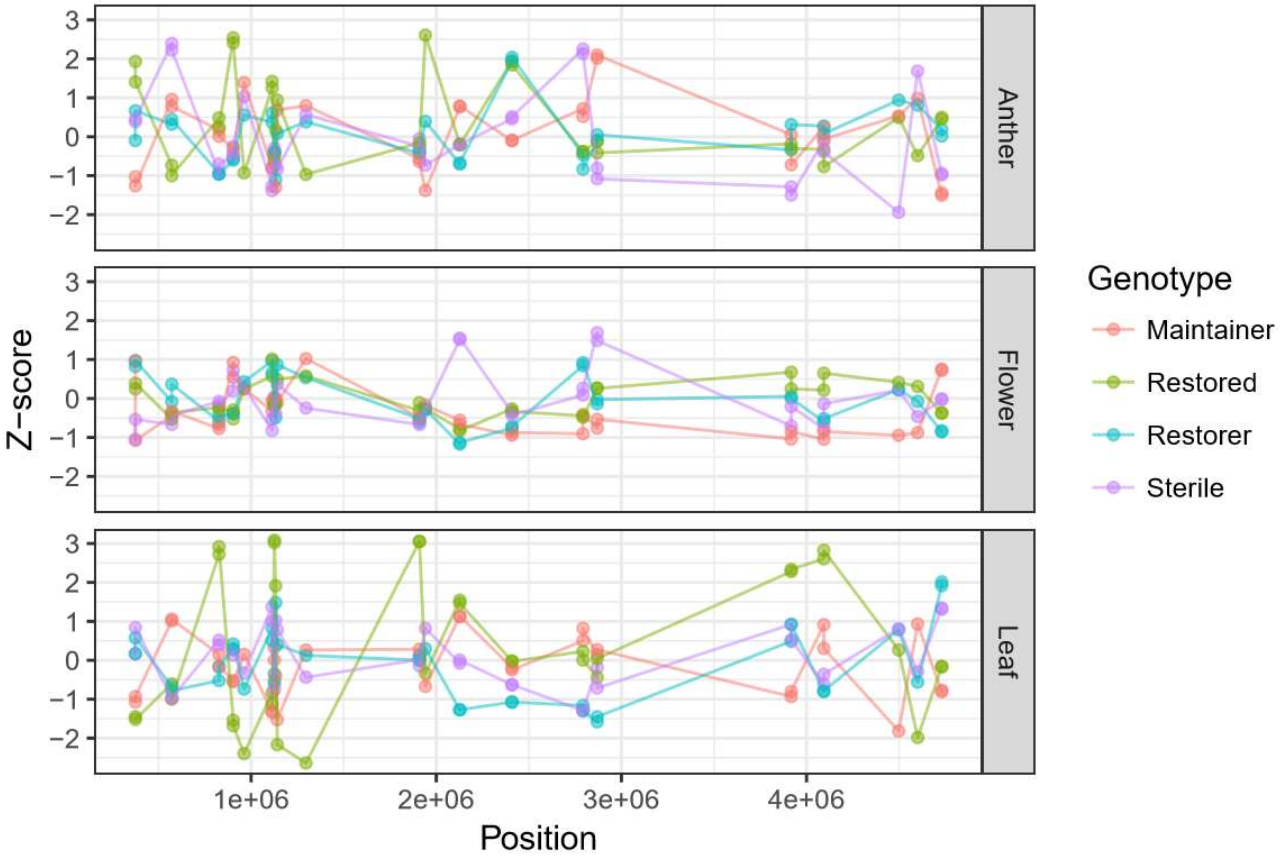

**Figure 5.** Box plot showing the SPAD results of four populations (A, B, C and D) of perennial ryegrass segregating for fertility restoration.

These results showed that the only population where fertile plants had a significantly higher chlorophyll content was population D with a *P*-value of 0.028. The other three segregating populations showed no significant difference between fertile and sterile individuals (*P*-values of A; 0.051, B; 0.059 and C; 0.669). An ANOVA test revealed that fertile plants from population D were unique from populations A, B and C and that sterile plants from all four populations were unique from each other (Supplementary Figure 1).

**Analysis of gene expression across previously identified qualitative trait loci (QTL) for fertility restoration**

Previously, QTL for fertility restoration were identified with the use of genotyping by sequencing (chapter 3) and bulk sergeant analysis (BSA, chapter 4). To assess if any genes were differentially expressed at these loci, the expression patterns of differentially expressed genes across the two Rabiosa scaffolds containing the two QTL (scaffold2257 and scaffold3325) for all genotype and tissue type combinations were visualised (Figure 6).

**A**



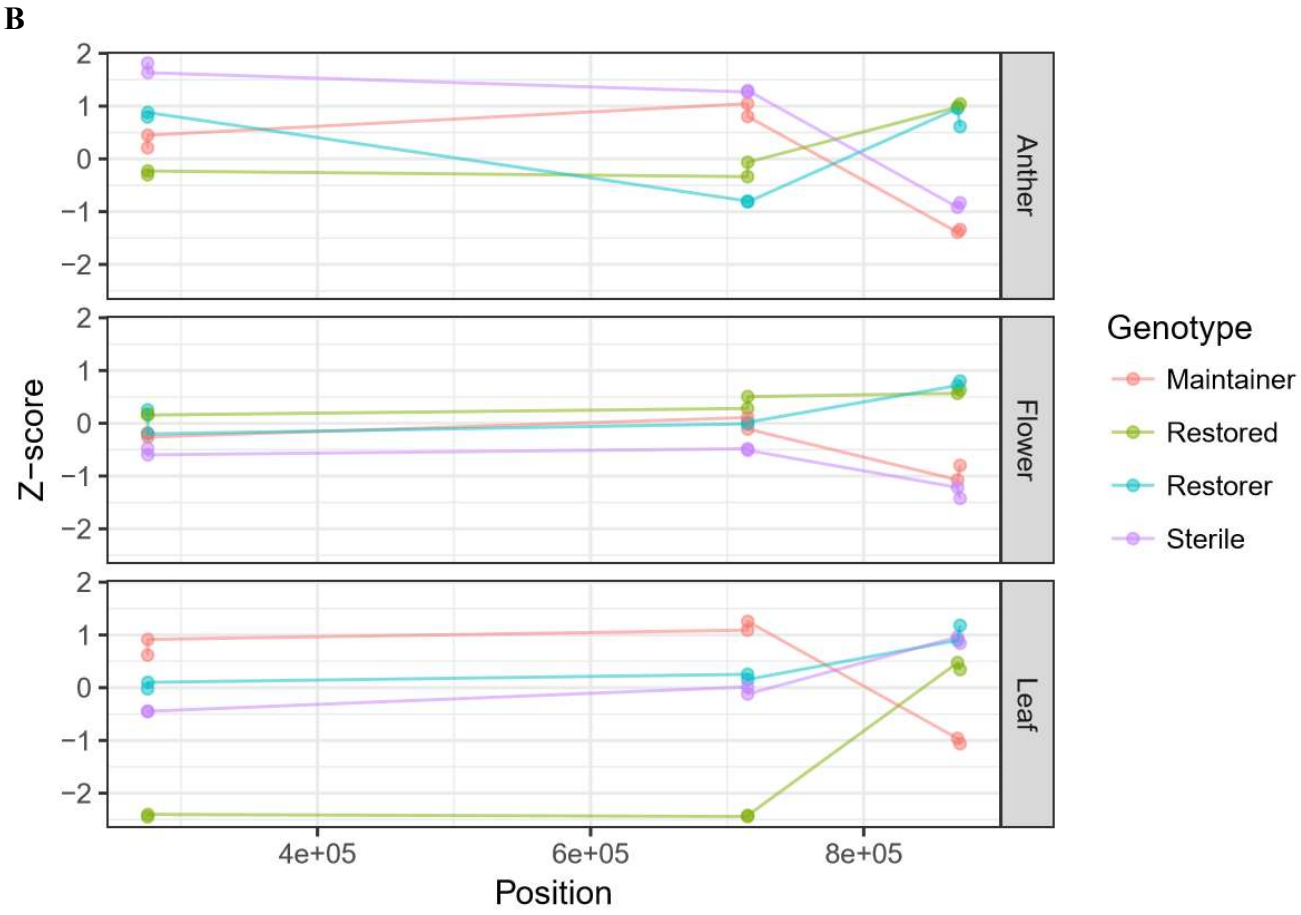



**Figure 6.** Differential gene expression (Z-score, y-axis) against scaffold location (Position, x-axis) for all differentially expressed genes identified on **A:** scaffold2257 and **B:** scaffold3325.

On scaffold2257, 92 genes were identified of which 26 were differentially expressed between at least two genotype and tissue type combinations. For scaffold3325, 34 genes were identified of which six were differentially expressed. To ascertain if any genes were being expressed within the QTL, the un-filtered gene expression data for the genes identified was analysed. Included in this analysis were 15 ORFs identified near or within the defined QTL regions, which were not identified in the original gene annotation. Of these 15 ORFs, five were within the QTL on scaffold2257 and ten within the QTL on scaffold3325 (Table 2). The location, expression data and annotation for these ORFs and genes, as well as the locations of the SNPs significantly linked to restoration of fertility (chapter 4, table 2), can be seen in figure 7 and table 2.

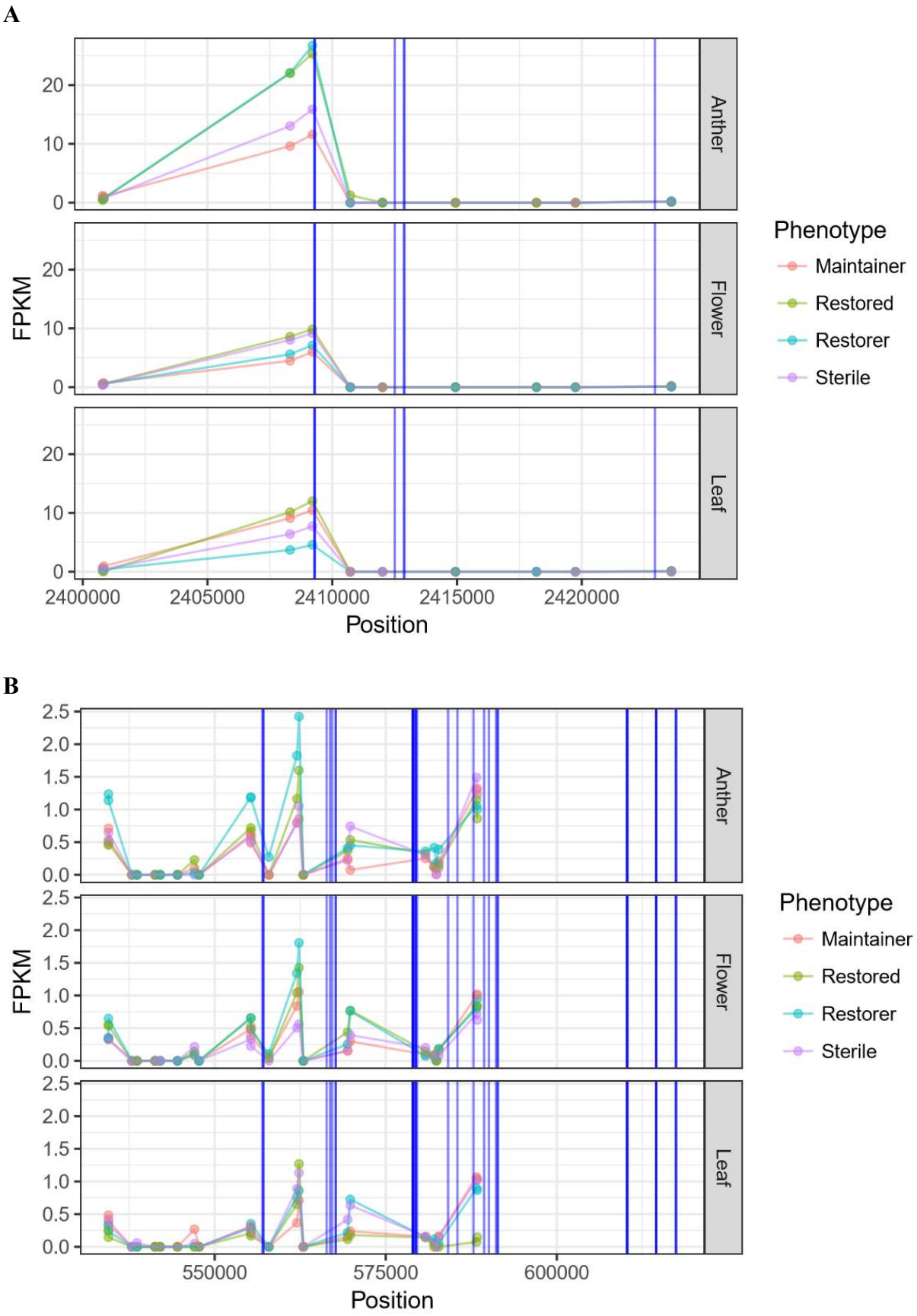



**Figure 7.** Gene expression values (FPKM, y-axis) against QTL location (Position, x-axis) for the two QTL on **A:** scaffold2257 and **B:** scaffold3325. Blue vertical lines represent the locations of SNPs singnificantly linked (LOD>50) to fertility restoration.

**Table 2.** Identification, location and identity of all annotated genes, open reading frames and SNPs within the two fertility restoration QTL (* indicates genes with a predicted mitochondrial signal peptide).

**Scaffold2257**

| Gene/SNP | Location (bp) | | Arabidopsis Homologue | Gene description |
|---|---|---|---|---|
| | Start | End | | |
| XLOC_036337 | 2400838 | 2402919 | AT3G03600 | ribosomal protein S2 |
| XLOC_036279 | 2408308 | 2411748 | - | mitochondrial 18S 5S ribosomal RNA |
| XLOC_036338 | 2409204 | 2410892 | - | mitochondrial 18S 5S ribosomal RNA |
| SNP6 | 2409291 | - | - | - |
| SNP7 | 2409292 | - | - | - |
| SNP8 | 2409293 | - | - | - |
| XLOC_orf1 | 2410720 | 2411256 | ATMG01390 | mitochondrial 18S ribosomal RNA |
| XLOC_orf2 | 2412014 | 2412346 | - | mitochondrial origin* |
| SNP9 | 2412506 | - | - | - |
| SNP10 | 2412867 | - | - | - |
| SNP11 | 2412893 | - | - | - |
| XLOC_orf3 | 2414938 | 2415243 | - | mitochondrial origin |
| XLOC_orf4 | 2418184 | 2418495 | - | mitochondrial origin |
| XLOC_orf5 | 2419746 | 2420567 | AT1G01930 | zinc finger protein-like protein |
| SNP12 | 2422932 | - | - | - |
| XLOC_036339 | 2423588 | 2424959 | ATMG00900 | cytochrome C biogenesis 256 |
| XLOC_036280 | 2423589 | 2424752 | ATMG00900 | cytochrome C biogenesis 256 |

**Scaffold3325**

| Gene/SNP | Location (bp) | | Arabidopsis Homologue | Gene description |
|---|---|---|---|---|
| | Start | End | | |
| XLOC_051294 | 534468 | 535761 | ATMG01360 | cytochrome oxidase |
| XLOC_051278 | 534470 | 535750 | ATMG01360 | cytochrome oxidase |
| XLOC_orf8 | 537897 | 538358 | - | mitochondrial origin |
| XLOC_orf9 | 538640 | 539095 | - | mitochondrial origin |
| XLOC_orf10 | 541286 | 541747 | - | mitochondrial origin* |
| XLOC_orf11 | 542029 | 542484 | - | mitochondrial origin |
| XLOC_orf12 | 544558 | 545727 | - | mitochondrial origin |
| XLOC_orf13 | 547042 | 547665 | - | mitochondrial origin |
| XLOC_orf14 | 547711 | 548163 | - | mitochondrial origin |
| XLOC_051295 | 555279 | 557170 | ATMG01360 | cytochrome c oxidase subunit 1 |
| XLOC_051279 | 555281 | 557449 | ATMG01360 | cytochrome c oxidase subunit 1* |
| SNP1 | 556973 | - | - | - |
| SNP2 | 557107 | - | - | - |
| SNP3 | 557113 | - | - | - |
| XLOC_orf15 | 557905 | 558483 | - | mitochondrial origin |
| XLOC_051280 | 562015 | 563541 | ATMG00730 | cytochrome c oxidase subunit 3* |
| XLOC_051296 | 562315 | 563580 | ATMG00730 | cytochrome c oxidase subunit 3* |
| XLOC_orf16 | 562952 | 564037 | ATMG00730 | cytochrome c oxidase subunit 3 |
| SNP4 | 566364 | - | - | - |

| | | | | |
|---|---|---|---|---|
| SNP5 | 566871 | - | - | - |
| SNP6 | 567182 | - | - | - |
| SNP7 | 567687 | - | - | - |
| SNP8 | 567689 | - | - | - |
| XLOC_051297 | 569427 | 570360 | ATMG01360 | cytochrome c oxidase subunit 1* |
| XLOC_orf17 | 569807 | 570296 | AT3G27300 | glucose-6-phosphate dehydrogenase 5* |
| SNP9 | 578940 | - | - | - |
| SNP10 | 578964 | - | - | - |
| SNP11 | 578975 | - | - | - |
| SNP12 | 578980 | - | - | - |
| SNP13 | 578981 | - | - | - |
| SNP14 | 579000 | - | - | - |
| SNP15 | 579329 | - | - | - |
| SNP16 | 579366 | - | - | - |
| SNP17 | 579402 | - | - | - |
| SNP18 | 579586 | - | - | - |
| XLOC_051298 | 582070 | 583687 | ATMG00180 | cytochrome C biogenesis 452 |
| XLOC_051281 | 582748 | 583816 | ATMG00180 | cytochrome C biogenesis 452 |
| SNP19 | 584107 | - | - | - |
| SNP20 | 585474 | - | - | - |
| SNP21 | 587822 | - | - | - |
| XLOC_051299 | 588257 | 591906 | ATMG00270 | NADH dehydrogenase 6 |
| XLOC_051282 | 588363 | 591906 | ATMG00270 | NADH dehydrogenase 6 |
| SNP22 | 589381 | - | - | - |
| SNP23 | 590079 | - | - | - |
| SNP24 | 591111 | - | - | - |
| SNP25 | 591374 | - | - | - |
| SNP26 | 591387 | - | - | - |
| SNP27 | 610227 | - | - | - |
| SNP28 | 610276 | - | - | - |
| SNP29 | 610304 | - | - | - |
| SNP30 | 610326 | - | - | - |
| SNP31 | 614526 | - | - | - |
| SNP32 | 614527 | - | - | - |
| SNP33 | 614529 | - | - | - |
| SNP34 | 614531 | - | - | - |
| SNP35 | 617397 | - | - | - |
| SNP36 | 617398 | - | - | - |
| SNP37 | 617484 | - | - | - |

**Candidate gene identification**

In total, ten genes or ORFs and seven SNPs were located within the QTL on Scaffold2257. Of these ten genes, three have mitochondrial homologues in Arabidopsis, two have nuclear homologues in Arabidopsis and remaining five have been identified as mitochondrial in origin (Table 2). Although most do not show any expression, there are two genes that show increased expression in restored and restorer anther tissue (Figure 7). Both these genes are mitochondrial 18S 5S ribosomal RNA genes and slightly overlap (XLOC_036279 and XLOC_036338). Within the coding region of both these genes are three consecutive SNPs (SNP6, SNP7 and SNP8).

On Scaffold3325, 21 genes or ORFs and 37 SNPs were located within the QTL for fertility restoration. Of these 21 genes, twelve have mitochondrial homologues in Arabidopsis, one has a nuclear homologue in Arabidopsis and the remaining eight have a mitochondrial origin (Table 2). There is varied expression of these genes across the QTL with the most relevant difference being an increase in expression in restorer and restored anther and flower tissue of two copies of the mitochondrial cytochrome c oxidase subunit 3 (COX3) (XLOC_051280 and XLOC_051296). Again these two annotated genes overlap although they do not contain any SNPs within their coding sequence. The previous gene, cytochrome c oxidase subunit 1 (COX1) (XLOC_051279), which does show an increase of expression in restorer anther tissue, contains three SNPs within its coding sequence. The first of these SNPs is a synonymous change with the other two causing a Gly276Val and Val278Gly amino acid changes. The positions of these mutations correspond to the beginning of a helical transmembrane domain in the folded COX1 protein [55] (UniProtKB - P00395). Both the COX1 and COX3 gene sequences were predicted to contain mitochondrial signal peptides with of 0.97 and 0.98 certainty, respectively [52].

## Discussion

Gene expression analysis of four CMS associated genotypes, across three tissue types, was successfully applied to identify candidate genes for fertility restoration in perennial ryegrass. The four essential genotypes for a CMS-based hybrid-breeding program were chosen for this study: the 'maintainer', 'sterile', 'restorer' and 'restored' genotypes. Between these four genotypes, all possible combinations of sterilising cytoplasm and restoring nuclear genes could be investigated.

### Systemic changes in gene expression

This study revealed pleiotropic changes in gene expression that appear to be linked to the CMS and its fertility restoration process. This was unexpected as it would seem reasonable that a CMS fertility restoration system should operate without greatly affecting general gene expression, especially in leaf tissue which was included as a control. Previous gene expression studies for other CMS systems have focused on either flower developmental stages or flowering/pollen formation tissue types and have not included non-reproductive tissues. Many genes and molecular pathways have been implemented in the cellular responses to CMS and fertility restoration and large scale changes in gene expression have been observed. Interestingly, the systemic gene expression changes between tissue types from CMS-sterile and restored-fertile plants suggest that for this particular CMS system, large changes in gene expression may not be contained to just flowering/pollen production tissues but may be systemic throughout the plant. A reason as to why large changes in gene expression are observed in leaf tissue may lie in the fact that this CMS system was induced through the use of a mutagen, thus it is possible that the CMS causal mutation(s) may be active in all tissues. This is on contrast to a natural system where such large scale changes would be detrimental upon fitness and thus eliminated by selection.

By the identification of three distinct groups of genotype and tissue type sample combinations, it could be revealed that increased chlorophyll/photosynthesis activity and cell cycle genes were responsible for the bulk of these differences. When these results are reviewed by comparing the three identified clusters of gene expression (Figure 4) against both the genotype and tissue type data, the observed pattern can be explained by three assumptions: firstly, there appears to be a difference between samples without *Rf* genes and samples with *Rf* genes, as seen in the shifts in expression from clusters one and three in Figure 4 (with the exception of sterile and restored leaf tissue). This suggests that there is a tissue-independent response in reaction to the presence of *Rf* genes while maintainer and sterile genotypes show no response. Secondly, there is another similar stress response seen in the presence of the sterilising cytoplasm that is only present in leaf tissue. This explains why sterile-leaf samples show a similar expression pattern to other samples that contain *Rf* genes. The last assumption is that these two responses interact when both present to produce an increase in photosynthesis/chlorophyll as seen in restored-leaf tissue. However, this assumption is confounded by the fact that the restored genotype is a hybrid variety and what may be being measured here is heterosis [56]. These three hypothesis are summarised in the Venn diagram below (Figure 8).

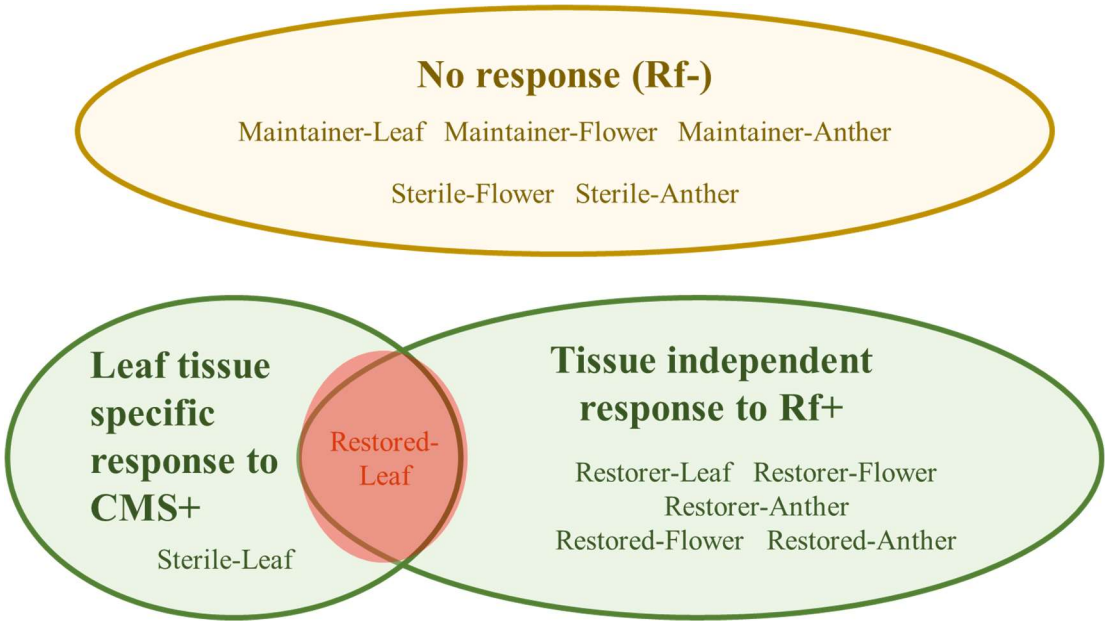



**Figure 8.** Venn diagram showing the three clusters different expression patterns seen in figure 4 and the presence (+) and absence (-) of *RF* genes and the CMS causal cytoplasm where appropriate.

The leaf specific response to the presence of CMS cytoplasm, which appears similar to the response seen in the presence of restorer genes, suggests that there is a secondary CMS recovery system active in leaf tissue. As CMS causal factors have been implemented in a reduction of energy output from affected mitochondria [57], a leaf specific CMS mitigating response would allow leaf tissue to function more normally. Given that the only easily observable phenotypic difference between restored-fertile and sterile plants is the lack of pollen formation, it is possible that this system is active in all somatic tissues. Mechanisms that tightly control the effects of CMS cytoplasm to reproductive tissues have been described in rice [58], although this is observed in a naturally occurring CMS systems that has evolved into this regulation. What is implied here, as this CMS system was induced, is that leaf tissue had a recovery/response pathway already active or awaiting activation by the CMS causal factor. This would explain why an induced CMS system, which may lack the nuances of an evolved system, would not cause significant negative effects throughout the whole plant.

What is most remarkable within these systemic shifts in gene expression is the profile seen in restored-leaf tissue (Figure 4). As the restored genotype is the only genotype included in this study that is a hybrid variety, it is possible that these results are the outcome of heterosis as several studies have observed 'increased photosynthesis' in hybrid plants (as reviewed by Offermann and Peterhansel (2014)) [59]. The other possible explanation is that the tissue independent response seen in the presence of *Rf* genes and the leaf-specific response to the sterilising cytoplasm combine uniquely to cause the observed gene expression changes. What lends credence to this hypothesis are observations made that suggest that restored hybrids are more vigorous than unrestored hybrids (Wilbert Lüsink, personal communication). Although dry matter yield data comparisons between restored and unrestored hybrids within the same populations were inconclusive (Wilbert Lüsink, personal communication), the gene expression results presented here suggest an increase in photosynthesis related activity. Chlorophyll content measurements using a SPAD chlorophyll meter showed that one of the four tested populations, segregating for restored and unrestored hybrids, had a significant difference in measured chlorophyll content, with restored-fertile plants containing more chlorophyll. These results are tantalising and suggest that further investigations into this phenomenon are warranted.

**Candidate genes for fertility restoration**

From the previous GBS and BSA studies (chapter 3 and chapter 4, respectively), two major QTL were identified for fertility restoration. Genes at these loci were investigated to identify any differential expression. This revealed three genes: a mitochondrial 18S 5S ribosomal RNA (rRNA) gene from the QTL on scaffold2257, and a *COX1* and *COX3* gene from the QTL on scaffold3325. Only the rRNA gene had been previously identified in the BSA study and the same three SNPs were identified within the coding sequence affecting the conserved 5' end of the resultant rRNA molecule. Although this gene has not been previously implicated in CMS sterility/fertility systems, its location within the QTL, combined with its increased transcription within restorer and restored anther tissue, and the presence of SNPs that correlate to fertility restoration, make it a strong candidate for an *Rf* gene.

Within the *COX3* gene, no SNPs were recorded, although the *COX1* gene did have three SNPs that cause two amino acid changes within the COX1 protein. These two changes are two residues apart and mirror each other with a glycine to valine change at position 276 and a valine to glycine change at position 278. As this region of the folded COX1 protein is the initiation point for a transmembrane helix it is likely that the glycine residue is involved in giving the protein the required flexibility to begin a helical region and that the valine is involved in hydrophobic interactions with the membrane itself. It is plausible that a swap of these two amino acids could affect the formation of the helical transmembrane. As COX1 is involved in pumping protons across the inner mitochondrial membrane, this may affect its efficiency, which is supported by the suggestion that proton leakage is the primary cause of CMS in sunflower [60]. Several other CMS systems have been linked to COX subunit dysfunction including in: Rice (*COX1* and *COX3*), Sorghum (*COX1)*, wheat (*COX1*), maize (*COX1*), radish (*COX1*), pepper (*COX2*) and sugar beet (*COX2*) (as reviewed by Chen and Liu (2014)) [61]. More recently *COX11* has also been implemented in Rice [62] and *COX1* and *COX3* in cabbage [63].

These three genes represent very strong candidates for fertility restoration within the two previously identified QTL. Not only do these results strengthen the theory that nuclear encoded mitochondrial genes are responsible for fertility restoration in perennial ryegrass, but they also provide excellent targets for further study as well as marker development for use in hybrid breeding schemes.

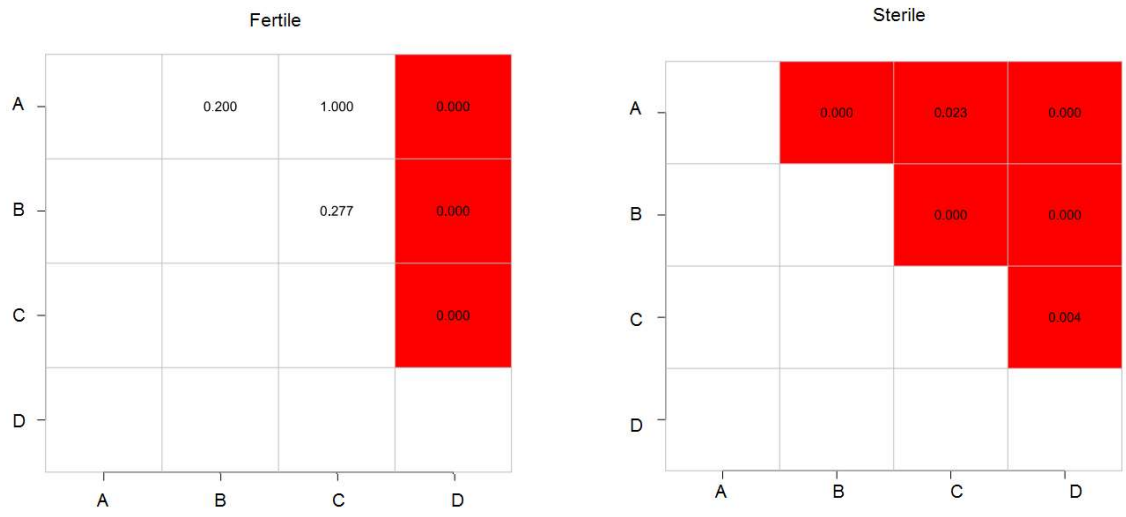
# References


1. Brenner, S., Jacob, F. and Meselson, M., 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature, 190(4776), pp.576-581.
2. Gros, F., Hiatt, H., Gilbert, W., Kurland, C.G., Risebrough, R.W. and Watson, J.D., 1961. Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. Nature, 190(4776), pp.581-585.
3. Jacob, F. and Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. Journal of molecular biology, 3(3), pp.318-356.
4. Lockhart, D.J. and Winzeler, E.A., 2000. Genomics, gene expression and DNA arrays. Nature, 405(6788), pp.827-836.
5. Nagle, J.W., Fields, C. and Venter, J.C., 1992. Sequence identification of 2,375 human brain genes. Nature, 355, p.13.
6. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W., 1995. Serial analysis of gene expression. Science, 270(5235), p.484.
7. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H. and Brown, E.L., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature biotechnology, 14(13), pp.1675-1680.
8. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. SCIENCE-NEW YORK THEN WASHINGTON-, pp.467-467.
9. Wu, C., Carta, R. and Zhang, L., 2005. Sequence dependence of cross-hybridization on short oligo microarrays. Nucleic acids research, 33(9), pp.e84-e84.
10. Voelkerding, K.V., Dames, S.A. and Durtschi, J.D., 2009. Next-generation sequencing: from basic research to diagnostics. Clinical chemistry, 55(4), pp.641-658.
11. Pettersson, E., Lundeberg, J. and Ahmadian, A., 2009. Generations of sequencing technologies. Genomics, 93(2), pp.105-111.
12. Tucker, T., Marra, M. and Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. The American Journal of Human Genetics, 85(2), pp.142-154.
13. John, R. and Grody, W.W., 2008. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. The Journal of Molecular Diagnostics, 10(6), pp.484-492.
14. Morozova, O. and Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics, 92(5), pp.255-264.
15. Fullwood, M.J., Wei, C.L., Liu, E.T. and Ruan, Y., 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome research, 19(4), pp.521-532.
16. Wang, Z., Gerstein, M. and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), pp.57-63.
17. Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlén, M. and Nielsen, J., 2012. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic acids research, 40(20), pp.10084-10097.
18. Li, S., Tighe, S.W., Nicolet, C.M., Grove, D., Levy, S., Farmerie, W., Viale, A., Wright, C., Schweitzer, P.A., Gao, Y. and Kim, D., 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nature biotechnology, 32(9), pp.915-925.
19. Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., Fennell, T. and Gnirke, A., 2013. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature methods, 10(7), pp.623-629.
20. van Dijk, E.L., Jaszczyszyn, Y. and Thermes, C., 2014. Library preparation methods for next-generation sequencing: tone down the bias. Experimental cell research, 322(1), pp.12-20.
21. Han, Y., Gao, S., Muegge, K., Zhang, W. and Zhou, B., 2015. Advanced applications of RNA sequencing and challenges. Bioinformatics and biology insights, 9(Suppl 1), p.29.
22. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A., 2011. Differential expression in RNA-seq: a matter of depth. Genome research, 21(12), pp.2213-2223.
23. Alvarez, M., Schrey, A.W. and Richards, C.L., 2015. Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution?. *Molecular Ecology*, *24*(4), pp.710-725.

24. Hangauer, M.J., Vaughn, I.W. and McManus, M.T., 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS genetics, 9(6), p.e1003569.
25. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. and Xue, C., 2012. Landscape of transcription in human cells. Nature, 489(7414), pp.101-108.
26. Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J. and Charpentier, E., 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 471(7340), pp.602-607.
27. Kakumanu, A., Ambavaram, M.M., Klumas, C., Krishnan, A., Batlang, U., Myers, E., Grene, R. and Pereira, A., 2012. Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq. Plant Physiology, 160(2), pp.846-867.
28. Kamber, T., Buchmann, J.P., Pothier, J.F., Smits, T.H., Wicker, T. and Duffy, B., 2016. Fire blight disease reactome: RNA-seq transcriptional profile of apple host plant defense responses to Erwinia amylovora pathogen infection. Scientific reports, 6, p.21600.
29. Song, G.C., Choi, H.K., Kim, Y.S., Choi, J.S. and Ryu, C.M., 2017. Seed defense biopriming with bacterial cyclodipeptides triggers immunity in cucumber and pepper. Scientific Reports, 7, p.14209.
30. Liu, C., Ma, N., Wang, P.Y., Fu, N. and Shen, H.L., 2013. Transcriptome sequencing and de novo analysis of a cytoplasmic male sterile line and its near-isogenic restorer line in chili pepper (Capsicum annuum L.). PloS one, 8(6), p.e65209.
31. Yang, P., Han, J. and Huang, J., 2014. Transcriptome sequencing and de novo analysis of cytoplasmic male sterility and maintenance in JA-CMS cotton. *PloS one*, *9*(11), p.e112320.
32. Suzuki, H., Rodriguez-Uribe, L., Xu, J. and Zhang, J., 2013. Transcriptome analysis of cytoplasmic male sterility and restoration in CMS-D8 cotton. *Plant cell reports*, *32*(10), pp.1531-1542.
33. An, H., Yang, Z., Yi, B., Wen, J., Shen, J., Tu, J., Ma, C. and Fu, T., 2014. Comparative transcript profiling of the fertile and sterile flower buds of pol CMS in B. napus. *BMC genomics*, *15*(1), p.258.
34. Zheng, B.B., Wu, X.M., Ge, X.X., Deng, X.X., Grosser, J.W. and Guo, W.W., 2012. Comparative transcript profiling of a male sterile cybrid pummelo and its fertile type revealed altered gene expression related to flower development. PLoS One, 7(8), p.e43758.
35. Wang, S., Wang, C., Zhang, X.X., Chen, X., Liu, J.J., Jia, X.F. and Jia, S.Q., 2016. Transcriptome de novo assembly and analysis of differentially expressed genes related to cytoplasmic male sterility in cabbage. *Plant Physiology and Biochemistry*, *105*, pp.224-232.
36. Li, J., Han, S., Ding, X., He, T., Dai, J., Yang, S. and Gai, J., 2015. Comparative transcriptome analysis between the cytoplasmic male sterile line NJCMS1A and its maintainer NJCMS1B in soybean (Glycine max (L.) Merr.). *PloS one*, *10*(5), p.e0126771.
37. . Lee, Y.P., Cho, Y. and Kim, S., 2014. A high-resolution linkage map of the Rfd1, a restorer-of-fertility locus for cytoplasmic male sterility in radish (Raphanus sativus L.) produced by a combination of bulked segregant analysis and RNA-Seq. *Theoretical and applied genetics*, *127*(10), pp.2243-2252.
38. Mei, S., Liu, T. and Wang, Z., 2016. Comparative transcriptome profile of the cytoplasmic male sterile and fertile floral buds of radish (Raphanus sativus L.). *International journal of molecular sciences*, *17*(1), p.42.
39. Rhee, S.J., Seo, M., Jang, Y.J., Cho, S. and Lee, G.P., 2015. Transcriptome profiling of differentially expressed genes in floral buds and flowers of male sterile and fertile lines in watermelon. *BMC genomics*, *16*(1), p.914.
40. Havey MJ. The use of cytoplasmic male sterility for hybrid seed production. Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria 2004:623-634.
41. Tang, H., Xie, Y., Liu, Y.G. and Chen, L., 2017. Advances in understanding the molecular mechanisms of cytoplasmic male sterility and restoration in rice. Plant reproduction, pp.1-6.
42. Rio, D.C., Ares, M., Hannon, G.J. and Nilsen, T.W., 2010. Purification of RNA using TRIzol (TRI reagent). Cold Spring Harbor Protocols, 2010(6), pp.pdb-prot5439.
43. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), pp.15-21.
44. Min, X.J., Butler, G., Storms, R. and Tsang, A., 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic acids research, 33(suppl_2), pp.W677-W680.

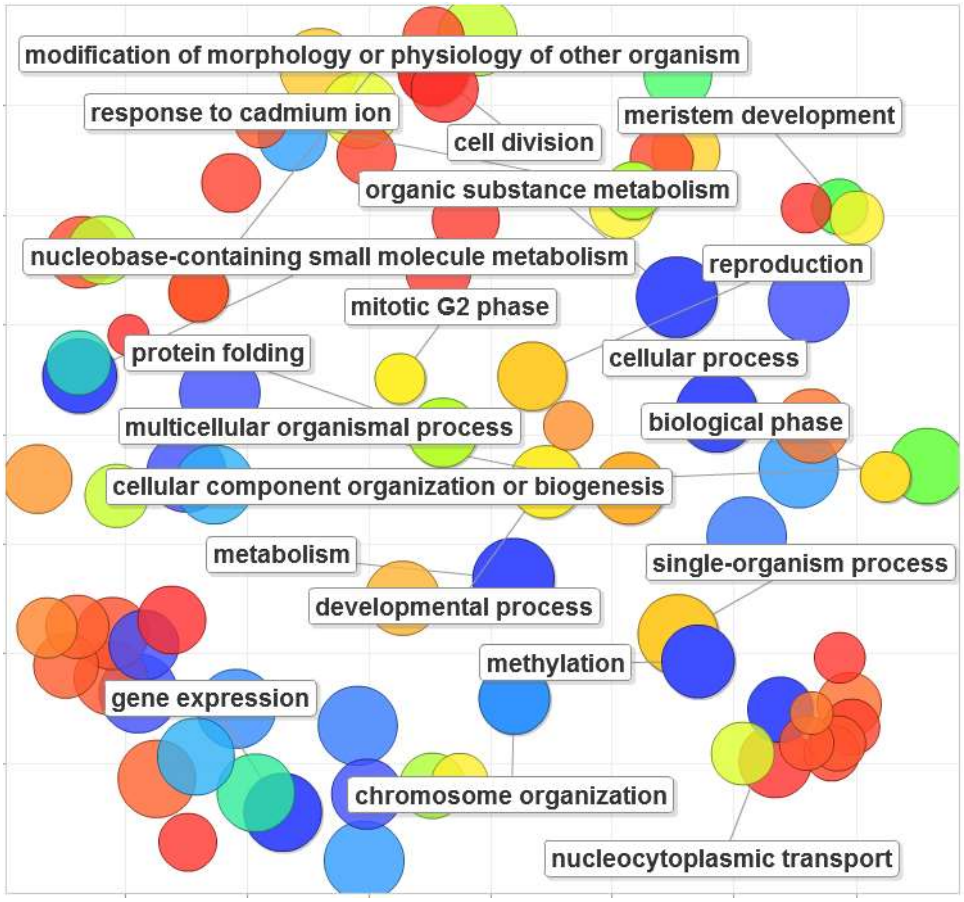
45. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), pp.3389-3402.
46. Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L. and Radenbaugh, A., 2007. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic acids research, 36(suppl_1), pp.D1009-D1014.
47. Kersey, P.J., Lawson, D., Birney, E., Derwent, P.S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kähäri, A. and Kinsella, R.J., 2009. Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic acids research*, *38*(suppl_1), pp.D563-D569.
48. Prochnik, S., Marri, P.R., Desany, B., Rabinowicz, P.D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T. and Rokhsar, D.S., 2012. The cassava genome: current progress, future directions. Tropical plant biology, 5(1), pp.88-94.
49. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S., 2011. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, *40*(D1), pp.D1178-D1186.
50. Palmieri, N., Nolte, V., Suvorov, A., Kosiol, C. and Schlötterer, C., 2012. Evaluation of different reference based annotation strategies using RNA-Seq–a case study in Drososphila pseudoobscura. PloS one, 7(10), p.e46415.
51. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology, 215(3), pp.403-410.
52. Sperschneider, J., Catanzariti, A.M., DeBoer, K., Petre, B., Gardiner, D.M., Singh, K.B., Dodds, P.N. and Taylor, J.M., 2017. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Scientific Reports, 7.
53. Team, R.C., 2000. R language definition. Vienna, Austria: R foundation for statistical computing.
54. Rodriguez, I.R. and Miller, G.L., 2000. Using a chlorophyll meter to determine the chlorophyll concentration, nitrogen concentration, and visual quality of St. Augustinegrass. HortScience, 35(4), pp.751-754.
55. Tsukihara, T., Aoyama, H., Yamashita, E. and Tomizaki, T., 1996. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 angstrom. Science, 272(5265), p.1136.
56. Baranwal, V.K., Mikkilineni, V., Zehr, U.B., Tyagi, A.K. and Kapoor, S., 2012. Heterosis: emerging ideas about hybrid vigour. *Journal of experimental botany*, *63*(18), pp.6309-6314.
57. Bohra, A., Jha, U.C., Adhimoolam, P., Bisht, D. and Singh, N.P., 2016. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. *Plant cell reports*, *35*(5), pp.967-993.
58. Luo, D., Xu, H., Liu, Z., Guo, J., Li, H., Chen, L., Fang, C., Zhang, Q., Bai, M., Yao, N. and Wu, H., 2013. A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. *Nature genetics*, *45*(5), pp.573-577.
59. Offermann, S. and Peterhansel, C., 2014. Can we learn from heterosis and epigenetics to improve photosynthesis?. *Current opinion in plant biology*, *19*, pp.105-110.
60. Balk, J. and Leaver, C.J., 2001. The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release. *The Plant Cell*, *13*(8), pp.1803-1818.
61. Chen, L. and Liu, Y.G., 2014. Male sterility and fertility restoration in crops. *Annual review of plant biology*, *65*, pp.579-606.
62. Luo, D., Xu, H., Liu, Z., Guo, J., Li, H., Chen, L., Fang, C., Zhang, Q., Bai, M., Yao, N. and Wu, H., 2013. A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. *Nature genetics*, *45*(5), pp.573-577.
63. Wang, S., Wang, C., Zhang, X.X., Chen, X., Liu, J.J., Jia, X.F. and Jia, S.Q., 2016. Transcriptome de novo assembly and analysis of differentially expressed genes related to cytoplasmic male sterility in cabbage. *Plant Physiology and Biochemistry*, *105*, pp.224-232.

# Supplementary Data

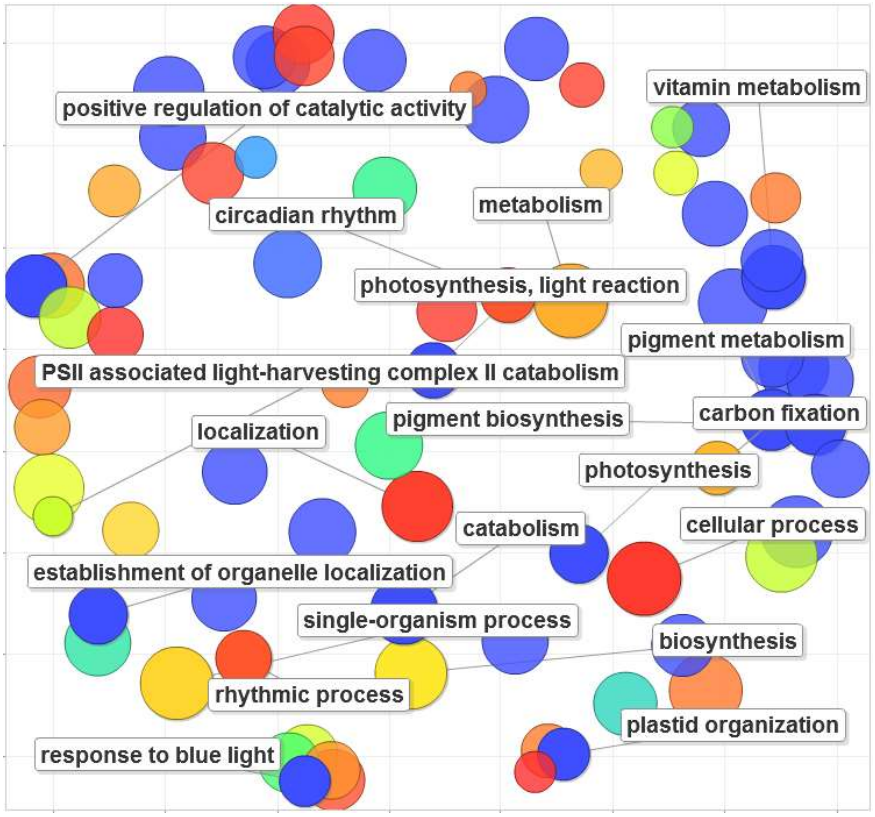



**Supplimental figure 1.** One factor ANOVA analysis of the SPAD results for the four populations segregating for fertility restoration.

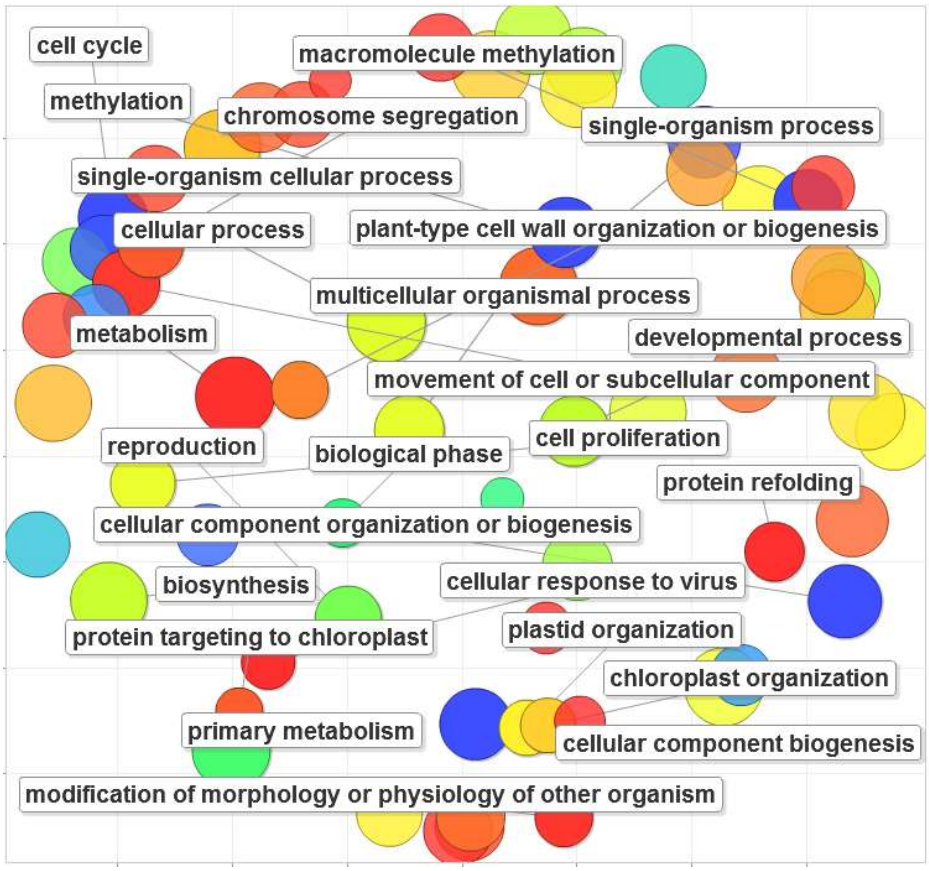
**A.**


modification of morphology or physiology of other organism
response to cadmium ion
meristem development
cell division
organic substance metabolism
nucleobase-containing small molecule metabolism
reproduction
mitotic G2 phase
protein folding
cellular process
multicellular organismal process
biological phase
cellular component organization or biogenesis
metabolism
single-organism process
developmental process
methylation
gene expression
chromosome organization
nucleocytoplasmic transport

**B.**


positive regulation of catalytic activity
vitamin metabolism
circadian rhythm
metabolism
photosynthesis, light reaction
pigment metabolism
PSII associated light-harvesting complex II catabolism
localization
pigment biosynthesis
carbon fixation
photosynthesis
catabolism
cellular process
establishment of organelle localization
single-organism process
biosynthesis
rhythmic process
plastid organization
response to blue light

**C.**




**Supplimental figure 2.** Graphical output of GO terms from REVIGO for the three differentially expressed gene clusters. **A:** Cluster one. **B:** cluster two. **C:** cluster three. Circle colours refer to p-values with blue being more significant, circle sizes correspond to the number of genes containing that annotation. The GO terms presented represent the most important to the identity of that particular cluster.

# Chapter 6

# General Discussion

### Nuclear-integrated mitochondrial genes as restorers of fertility

The application of a combination of sequencing-based techniques to interrogate the genetic cause of fertility restoration in CMS affected perennial ryegrass identified three strong candidate genes at two genetic loci. These two loci contain stretches of nuclear-integrated mitochondrial genes (NUMTs) that appear to be responsible for fertility restoration. They likely encode functional copies of their defective mitochondrial counterparts, allowing them to act as restorers. This hypothesis is strengthened by the observation that two of these genes have mutations in the sterile phenotype, with the fertile plants carrying an un-mutated copy. Further evidence for their restorative powers comes from their gene expression profiles (chapter 5) that show increased expression in the anthers of plants that carry them. This suggests that these genes are under some form of transcriptional control that is targeting them to the time and place (anther development) that they are needed to ensure pollen formation.

The bulk sergeant analysis (BSA) of restored-fertile and sterile plants revealed that within the two NUMT loci, the sterile plants are homozygous and the fertile plants heterozygous. Apart from the obvious link between the haplotype only present in the fertile pool and the fertile phenotype, this data also reveals more about the possible nature of this restoration system. Across both NUMT QTL, increased numbers of sequences were identified suggesting that both mitochondrial and nuclear copies of these genes were being mapped to these regions. As only one haplotype was observed in the sterile pool, it is likely that this is the mitochondrial copy of this region and therefore there is no NUMT DNA that matches this region (unless both the nuclear and mitochondrial copies of this region have almost exactly the same sequence). In the fertile pool, the same haplotype as in the sterile pool with the additional restoring haplotype was observed. This suggests that the NUMT regions associated with fertility are unique to the restored-fertile plants. Given this assumption, it would be expected that these regions would not be present in any gene expression analysis across the QTL in the genotypes without *Rf* genes. This is not the case, with these genes being present within the gene expression data for non *Rf* containing genotypes. This could be, once again, the mitochondrial transcripts being mapped to the nuclear genome reference, although the library preparation for the RNA sequencing was poly-A enriched, making this unlikely. This puts these two data sets slightly in conflict, although this could be resolved by long-read sequencing of the two QTL, which would definitively identify the haplotypes present at these loci.

As NUMTs have not previously been identified as restoring genes, the CMS/restoration system described here is unique within the literature. The fertility-linked NUMT presented here show the same conserved gene order as seen on the perennial ryegrass mitochondrial genome [1], confirming that they have been integrated as whole sections of mitochondrial genome and not as separate smaller events. One aspect of having NUMTs as restorers is that identifying the restoration genes also results in the identification of the CMS causal genes. This is also true for any mutations observed within the genes, with changes in the sterile versions of these genes likely being causal.

The candidate genes identified here – two subunits of the mitochondrial respiratory complex IV and a ribosomal RNA (rRNA) gene – are excellent candidates for CMS causation. The three

subunits of cytochrome c oxidase (COX1, COX2 and COX3) have all been implicated in CMS causation in several plant species (chapter 5). The other set of genes often identified as causal for CMS are subunits of ATP synthase [2]. This suggests that nuclear copies of genes encoding ATP synthase subunits could also function as CMS fertility-restoration genes in the presence of defective mitochondrial copies. The primary cause of CMS in these systems is often a chimeric open reading frame (ORF) that contains part or all of one of these genes, with the ultimate cause of CMS being increased production of reactive oxygen species and programmed cell death (PCD) [3]. These ORFs appear to interfere with the respiratory chain proteins leading to PCD and pollen abortion. What is interesting here is that this appears to only affect pollen formation, suggesting that a fully functional respiratory chain is only needed during pollen formation [4]. This concept is supported by the identification of an aldehyde dehydrogenase protein as being responsible for fertility restoration of the Texas-cytoplasm in maize (*Zea mays* L), as this protein is responsible for the removal of malondialdehyde, a toxic product of reactive oxygen species overproduction [4, 5]. Although rRNA genes have not previously been implicated as CMS causal factors, it is possible that a disruption in mitochondrial protein synthesis could produce similar effects that the CMS-causal ORFs are implicated in. This possible mechanism for CMS causation warrants further study, possibly beginning with investigation of the cellular location of this nuclear encoded rRNA to confirm its mitochondrial destination.

Gene transfer from mitochondrial to nuclear genomes is a constant feature of eukaryotic evolution, with this ongoing process relocating large numbers of mitochondrial genes into the nucleus, especially during the early phase of organellar evolution [6]. Plants are no exception to this, with Arabidopsis (*Arabidopsis thaliana* L) having 305 kb, rice (*Orzya sativa* L) 824 kb, sorghum (*Sorghum bicolor* L) 539 kb and maize 71 kb of nuclear encoded regions of their current mitochondrial genomes present in the nucleus [7]. The evolutionary drivers of this transfer continue to be hotly debated, with neutral gene loss, adaptive processes, mitochondrial competition, sexual reproductive strategies and epistatic models all being leading theories (as reviewed by Brandvain and Wade [8]). The epistatic model, stating that gene transfer is a co-evolutionary process where fitness is a function of gene combinations rather than individual genes, is perhaps the most relevant to this research [9]. Recent studies have even identified horizontal gene transfer of mitochondrial genes across species borders, adding another level of complexity to this evolutionary conundrum (reviewed by Bock [10]). Several studies have posited the evolutionary reasons for the natural maintenance of CMS, including speciation and the generation of distinct sexes in flowering plants [11, 12]. These factors may also be affecting the evolutionary maintenance of NUMT CMS restorers. The fact that the CMS system under investigation here was induced complicates this reasoning, as the system has not evolved into its current state.

What is interesting about this system is that it hints at a possible mechanism for gene transfer from mitochondrial to nuclear genomes. Mitochondrial genes can be transferred into the nucleus and, over an evolutionary period of time, acquire the necessary machinery to be expressed, and the protein product active in the mitochondria (transfer peptide, codon usage, promoter, poly-A signal, etc.). At this point, the two copies of a gene are perfectly redundant, with a loss-of-function causing no selectable phenotype and the eventual loss of that gene. If a mutation occurs in the mitochondrial copy of this gene that causes CMS, this changes that neutral situation. Now, a loss-of-function in the nuclear gene will cause or reveal CMS due to the defective mitochondrial copy. Once the individual carrying these ‘defective mitochondrial’

and ‘functional nuclear’ gene copies mates with other individuals, these two genes will be separated. If the defective mitochondrial gene copy lands in a ‘functional nuclear copy’ free environment, it will cause CMS which likely gives it a slight fitness advantage. The functional copy of the nuclear gene is now free of the CMS causal mitochondrial copy and is once again neutral, unless, through further crossing, it comes back into an individual with the CMS cytoplasm, where it will act as a CMS fertility restorer. Given this, this functional nuclear copy now confers a significant fitness advantage and will be selected for within an outcrossing population. Due to the inheritance dynamics (bi-parental vs. maternal) and the scale of conferred fitness advantage, the functional nuclear copy will likely become fixed within a population, leaving the defective copy to be eventually lost from the mitochondrial genome, thus completing the transfer of the gene from the mitochondrial to the nuclear genome.

**Advantages of a multi-technique approach to candidate gene identification**

The multi-technique approach presented here has allowed the initial hypothesis of a two loci co-dominant system being responsible for fertility restoration not only to be confirmed, but the likely genes involved to be identified. The genotyping by sequencing (GBS) study (chapter 3) allowed the identification of four QTL, totalling 87.3 Mb in length, which was further narrowed down by the use of BSA to two QTL totalling 74 kb in length, which was again narrowed down to three genes with the use of RNA sequencing (RNASeq). Taking this approach was exceedingly effective as the three techniques complement each other. GBS is a very cost-effective way to generate marker data across a genome and is particularly useful as it does not require a reference sequence to generate this marker data. As it allows the individual genotyping of hundreds of samples simultaneously, it is useful for unpicking the genetics underlying complex traits. BSA, on the other hand, is a much more powerful approach for identifying the genetics underlying binary phenotypes. This permits pools of absolute phenotypes, not groups of similar phenotypes, to be interrogated, allowing, in this case, the resolution of QTL to the kb range from the Mb range derived from GBS. Finally, the use of RNASeq allowed the identification of the individual genes being actively transcribed within these QTL. RNASeq, as it was implemented here, is a holistic approach that identifies all genes showing differential expression between sample treatments. Although this does allow broad, pleiotropic, effects to be identified, the sheer number of differentially expressed genes can make identifying causal genes impossible. Thus, RNASeq is not an appropriate technique for identifying QTL but rather a powerful way to interrogate the expression of genes within previously identified QTL. This was shown in chapter 5, where RNASeq data allowed the 31 genes identified within the two QTL to be narrowed down to just three candidate genes.

**Mechanisms of fertility restoration for other cytoplasmic male sterility systems**

There are two broad classes of CMS restorer genes. The first and easily most numerous class is made up of RNA binding proteins, of which the vast majority are pentatricopeptide repeat (PPR) genes. The second is a small group of genes encoding mitochondrial proteins of which the origins are likely mitochondrial. This second group includes an aldehyde dehydrogenase in maize [13], an acyl-carrier protein in rice [14] and a peptidase in sugar beet (*Beta vulgaris* L) [15]. Only the aldehyde dehydrogenase presents a different mode of restoration from the RNA binding proteins in this class as it can mitigate the overproduction of reactive oxygen species during pollen formation in CMS affected plants. It has been pointed out that this broader mechanism of restoration, and the fact that this gene is present in most fertile maize varieties makes this gene a fertility factor and not a restorer gene [16, 17]. Both

the other two examples, from rice and sugar beet, appear to be acting on the cause of CMS rather than the symptoms, with the interesting identification of retrograde signalling regulation in the rice system [14]. Although the genes encoding these proteins are likely of mitochondrial origin, there are no longer copies of these genes within the mitochondrial genome as they have completed their transfer from mitochondrial to nuclear genome, possibly through the mechanism described above.

The other class of restorers, containing almost exclusively *PPR* genes, are the most frequently identified and examined restorers of male fertility to CMS affected plants. The *PPR* gene family is especially numerous in land plants, with around 400 members found in most higher plant genomes [18]. These PPRs function as chaperones of mitochondrial transcription and translation and represent a mechanism whereby the nuclear genome controls mitochondrial gene expression. A small subset of PPRs has been identified that contain the CMS restorer of fertility genes. These genes are denoted as restorer of fertility-like *PPR* genes (*RFL*s) and are usually present in 10-30 copies per genome. Although the exact mechanism of action can vary for the *RFL*s that have been characterised, they all appear to interact with the CMS-causal ORF to prevent it from disrupting mitochondrial function [19].

Given that the mode of fertility restoration appears to be functionally simpler in the system uncovered by the research presented here (direct gene duplication as opposed to varied transcriptional/translational control), this raises the question as to why evolution prefers *RFL* mediated CMS-restoration. The answer here may lie in the evolutionary dynamics of the *RFL* genes. *RFLs* are under a diversifying selection pressure [20] and are maintained within the nuclear genome as clusters of genes that are undergoing recombination to constantly create new *RFL* genes and alleles [21]. In this respect, the dynamics of *RFL*s closely mirrors that of disease resistance genes, which is unsurprising as mitochondria were originally invading foreign bacteria [22]. This allows a stock of encoded proteins to be maintained that can restore varied causes of CMS. On the other hand, NUMTs as restorers are necessarily redundant in the absence of CMS and will be lost as they need to contain genes that are still present and functional within the mitochondrial genome. This makes their maintenance in the absence of a causal CMS cytoplasm impossible, as their function will be lost due to random mutation.

The approach presented in chapter 2 allows the rapid identification of *RFL* genes from large volumes of both transcriptomic and genomic sequencing data, the desired outcome being the identification of zones of active *RFL* generation that present good targets for the identification of *RFL-Rf* genes. This approach has already been successfully applied to genomic sequence data from barley (*Hordeum vulgare* L.) to clarify the source of CMS restoration previously identified (unpublished). As mentioned, *RFLs* are under a positive selective pressure and as such present an interesting target for evolutionary study. The RFL identification pipeline from chapter 2 has already been applied to sequence data from several rice sub species to identify *RFLs* for an evolutionary study revealing that DNA recombination is the driver of *RFL* diversification [21].

## Cytoplasmic male sterility in hybrid breeding schemes

The CMS trait is not just an interesting model for nuclear/mitochondrial genome interactions but also serves a practical purpose in plant breeding. CMS is the most popular of a host of techniques for controlling pollination during the production of commercial hybrid seed in many species [23]. The identification of a new class of restorer genes that indicate a novel mechanism of restoration will be of great interest to plant breeders, especially in crops without functional CMS systems such as wheat [24]. The possibility of not only identifying this CMS/*Rf* system in other crop species but also of engineering it will also be explored extensively by plant breeding companies. Perhaps the first example of this will come in Italian ryegrass (*Lolium multiflorum* L) where current breeding efforts to introgress the perennial ryegrass CMS cytoplasm into breeding material are being hampered by the inability of breeders to identify maintainer lines. On a genetic level, this means that all tested Italian ryegrass varieties are restorers and by implication, contain either the NUMTs identified here or other restorer genes. This hypothesis can be easily tested by mutating the restoring NUMTs in Italian ryegrass which should result in a non-restoring or maintainer genotype. This could either be achieved through a broad mutagenic approaches such as TILLING [25] or a more targeted approach such as CRISPER/Cas gene editing [26].

Practically, molecular markers for fertility restoration genes will assist breeders in integrating new material into their breeding programs. Currently when new plants are to be introduced into the perennial ryegrass hybrid-breeding program they need to be screened for the presence of restorer genes. This means crossing them in the field or greenhouse to CMS affected plants, collecting the seed, germinating that seed and phenotyping the offspring to make sure there are no fertile plants. This is a time consuming and costly process, often creating a two-year delay in the development of new hybrid varieties. It is vital to know if new plant material to be integrated into a hybrid-breeding scheme contains any restorer genes, as any unwanted restoration originating from the maintainer populations will cause impurities in the eventual hybrid seed to be sold to farmers. Maintainer lines are the fertile counterparts to CMS affected lines and are otherwise genetically identical to the sterile-CMS line but are male fertile. The maintainer plants do not carry any restorer genes and are free from the sterility inducing cytoplasm. This allows the sterile CMS line to be maintained – hence the name 'maintainer-line'. Crosses between the maintainer and sterile counterpart lines will result in more CMS sterile plants, while within the population, fertilisation in the maintainer line will perpetuate the male-fertile counterpart line. Therefore, the practical value of molecular markers for restorer genes is that they allow plant breeders to rapidly screen any new breeding material for the presence of restorer genes, speeding up the breeding process by up to two years.

In a hybrid breeding scheme two genetically distinct and largely homozygous inbred lines are crossed to produce an F1 population. This population, in well-designed hybrids, will show a significant increase in the trait of interest, usually yield. This is due to the phenomenon of heterosis, which despite being poorly understood is widely applied in plant breeding. Hybrid breeding has been put to good use in maize, sorghum, sunflower (*Helianthus annuus* L), rice and many other species, resulting in significant yield increases [27]. As mentioned, CMS is utilised as a pollination control mechanism during hybrid seed production. In practice, this means that when breeders have identified two breeding populations that show favourable hybridisation qualities, they ensure that one of these lines is affected by CMS. Thus, when the two lines are crossed all seed collected from the CMS affected plants is guaranteed to be a result of a hybridisation event between the two populations and not a result of within population

fertilisation. In grain crops, such as maize, the other non-CMS affected line carries a restorer gene, as male fertile hybrids are needed in order to produce grain in the field. This is not true of forage grasses, where biomass is the primary yield target.

It is possible that, as the CMS system under investigation here was induced, the CMS causal mutation is not presenting the usual phenotype of a CMS system: that of being only male sterile. Perennial ryegrass breeders have noted that unrestored hybrids appear "less vigorous" than restored hybrids but as this has not manifested in a significant difference in dry matter yield they have persisted with this system. The RNASeq data presented in chapter 5 shows that there are very large shifts in gene expression in restored leaf samples as compared to the rest of the data set. When analysed, these shifts were observed to be due to a large increase in the expression of genes involved in light harvesting. As this shift in gene expression would likely result in the production of more chlorophyll in restored leaf tissue, chlorophyll measurements comparing restored and non-restored hybrids were taken, revealing that in some populations restored hybrids did indeed contain more chlorophyll. If this result can be verified, which requires the untangling of the effects of heterosis from those of fertility restoration, this is an important observation that will change the way hybrid breeding in forage crops will be undertaken. As discussed above, in current perennial ryegrass breeding programs restorer genes are identified so they can be discarded from breeding populations. If restored hybrids do have an advantage over unrestored hybrids this process will have to be reversed and restorer genes integrated into hybrid breeding lines, making molecular markers for tracking them even more valuable.

**The global potential of hybrid varieties**

Since the 'Green Revolution' of the 1950s-1980s, hybrid breeding has been one of the most important yield increasing innovations – and possibly the most successful application of molecular genetics – in plant breeding. With several key agricultural species yet to have operative hybrid breeding schemes, the potential for hybrid varieties to improve global agricultural yields is still high. The CMS/*Rf* system identified here presents a new approach for the development of one of the molecular tools required for the creation of these hybrid varieties. Not only can this novel CMS/*Rf* system be identified in new species, it also has the potential to be engineered into these species, as demonstrated by its induced origins. This could be achieved through the application of a range of molecular techniques with perhaps the most relevant being genome editing. As it remains to be seen if genome edited crop species will be tolerated within the European Union (EU) [28], the future of any engineered CMS systems applying the novel mechanisms described here may lay outside of the EU. Overall, any advancement of the knowledge base used to create high yielding hybrid varieties will have a positive outcome on food production and agricultural sustainability – the ultimate outcome of which will be a strengthening of global food security.

# References


1. Islam, M.S., Studer, B., Byrne, S.L., Farrell, J.D., Panitz, F., Bendixen, C., Møller, I.M. and Asp, T., 2013. The genome and transcriptome of perennial ryegrass mitochondria. BMC genomics, 14(1), p.202.
2. Chen, L. and Liu, Y.G., 2014. Male sterility and fertility restoration in crops. *Annual review of plant biology*, *65*, pp.579-606.
3. Tang, H., Xie, Y., Liu, Y.G. and Chen, L., 2017. Advances in understanding the molecular mechanisms of cytoplasmic male sterility and restoration in rice. Plant reproduction, pp.1-6.
4. Møller, I.M., 2001. A more general mechanism of cytoplasmic male fertility?. *Trends in plant science*, *6*(12), p.560.
5. Halliwell, B., and Gutteridge, J.M.C., 1999. Free Radicals in Biology and Medicine (3rd edn), Oxford University Press.
6. Noutsos, C., Richly, E. and Leister, D., 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Research, 15(5), pp.616-628.
7. Hazkani-Covo, E., Zeller, R.M. and Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS genetics, 6(2), p.e1000834.
8. Brandvain, Y. and Wade, M.J., 2009. The functional transfer of genes from the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self-fertilization. Genetics, 182(4), pp.1129-1139.
9. Wade, M.J. and Goodnight, C.J., 2006. Cyto-nuclear epistasis: two-locus random genetic drift in hermaphroditic and dioecious species. Evolution, 60(4), pp.643-659.
10. Bock, R., 2017. Witnessing genome evolution: Experimental reconstruction of endosymbiotic and horizontal gene transfer. Annual Review of Genetics.
11. Frank, S.A., 1989. The evolutionary dynamics of cytoplasmic male sterility. *The American Naturalist*, *133*(3), pp.345-376.
12. Miller, I. and Bruns, E., 2016, February. The effect of disease on the evolution of females and the genetic basis of sex in populations with cytoplasmic male sterility. In *Proc. R. Soc. B*(Vol. 283, No. 1824, p. 20153035). The Royal Society.
13. Liu, F., Cui, X., Horner, H.T., Weiner, H. and Schnable, P.S., 2001. Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. *The Plant Cell*, *13*(5), pp.1063-1078.
14. Fujii, S. and Toriyama, K., 2009. Suppressed expression of RETROGRADE-REGULATED MALE STERILITY restores pollen fertility in cytoplasmic male sterile rice plants. *Proceedings of the National Academy of Sciences*, *106*(23), pp.9513-9518.
15. Kitazaki, K., Arakawa, T., Matsunaga, M., Yui-Kurino, R., Matsuhira, H., Mikami, T. and Kubo, T., 2015. Post-translational mechanisms are associated with fertility restoration of cytoplasmic male sterility in sugar beet (Beta vulgaris). *The Plant Journal*, *83*(2), pp.290-299.
16. Touzet, P., 2002. Is rf2 a restorer gene of CMS-T in maize?. *Trends in plant science*, *7*(10), p.434.
17. Liu, F., Cui, X., Horner, H.T., Weiner, H. and Schnable, P.S., 2001. Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. The Plant Cell, 13(5), pp.1063-1078.
18. Barkan, A. and Small, I., 2014. Pentatricopeptide repeat proteins in plants. Annual review of plant biology, 65, pp.415-442.
19. Hu, J., Huang, W., Huang, Q., Qin, X., Yu, C., Wang, L., Li, S., Zhu, R. and Zhu, Y., 2014. Mitochondria and cytoplasmic male sterility in plants. Mitochondrion, 19, pp.282-288.
20. Geddy, R. and Brown, G.G., 2007. Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. *BMC genomics*, *8*(1), p.130.
21. Melonek, J., Stone, J.D. and Small, I., 2016. Evolutionary plasticity of restorer-of-fertility-like proteins in rice. Scientific reports, 6, p.35152.
22. Margulis, L., 1970. Origin of eukaryotic cells: Evidence and research implications for a theory of the origin and evolution of microbial, plant and animal cells on the precambrian Earth. Yale University Press.

23. Bohra, A., Jha, U.C., Adhimoolam, P., Bisht, D. and Singh, N.P., 2016. Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. Plant cell reports, 35(5), pp.967-993.
24. Geyer, M., 2016, April. Improving fertility restoration and seed production efficiency in CMS hybrid wheat. In 2nd HEZagrar PhD Symposium (p. 14).
25. Manzanares, C., Yates, S., Ruckle, M., Nay, M. and Studer, B., 2016. TILLING in forage grasses for gene discovery and breeding improvement. New biotechnology, 33(5), pp.594-603.
26. Sander, J.D. and Joung, J.K., 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology*, *32*(4), pp.347-355.
27. Lee, J., Chin, J.H., Ahn, S.N. and Koh, H.J., 2015. Brief History and Perspectives on Plant Breeding. In Current Technologies in Plant Molecular Breeding (pp. 1-14). Springer Netherlands.
28. Wolt, J.D., Wang, K. and Yang, B., 2016. The Regulatory Status of Genome-edited Crops. Plant biotechnology journal, 14(2), pp.510-518.

## Acknowledgements

First and foremost I would like to thank my supervisor Bruno Studer and the always helpful Steven Yates for their patience and candour. Without their support, especialy during the final weeks, this document would surely not exist. My extended thanks go to the other members of the Molecular Plant Breeding Group at ETH Zurich (as well as its forbearer the Forage Crop Genetics Group) for all their assistance, inspiration and solidarity.

At Aarhus University in Denmark, my thanks go to Torben Asp and Istvan Nagy whose support and input was essential in getting this project underway. Many thanks as well to Wilbert Luesink at Norddeutsche Pflanzenzucht Hans-Georg Lembke for his endless patience while explaining his breeding schemes.

My greatest thanks go to Erin Pobjie, for convincing me that I could achieve this in the first place, and then supporting me throughout. I would not be where I am now without her tireless understanding and encouragement.

Finally, I would like to dedicate this thesis to my Dad, he knows why.

# Appendix I

# Genetic loci governing androgenic capacity in perennial ryegrass (*Lolium perenne* L.)

**Rachel F. Begheyn, Steven A. Yates, Timothy Sykes and Bruno Studer***

ETH Zurich, Institute of Agricultural Sciences (IAS), Zurich, Switzerland

*Corresponding author. Email: bruno.begheyn@usys.ethz.ch

## Abstract

Immature microspores can be induced to switch developmental pathways from gametogenesis to embryogenesis and subsequently regenerate into homozygous, diploid plants in vitro. Such androgenic production of doubled haploids may be a practically feasible method of inbred line production in selfincompatible species. Therefore, increasing the generally low androgenic capacity of perennial ryegrass (Lolium perenne L.) germplasm would enable efficient homozygous line production, so that a more effective exploitation of heterosis through hybrid breeding schemes can be realized. Here, we present the results of a genome-wide association study in a heterozygous, multi-parental perennial ryegrass population (n = 391) segregating for androgenic capacity. Genotyping by sequencing was used to interrogate gene dense genomic regions and revealed over 1100 polymorphic sites. Between 1 and 10 quantitative trait loci (QTL) were identified for anther response, embryo and total plant production, green and albino production and regeneration. Most traits were under polygenic control by several minor QTL, although a major QTL on linkage group 5 was associated with green plant regeneration. Distinct genetic factors seem to affect green and albino plant recovery. Two intriguing candidate genes, encoding chromatin binding domains of the developmental phase transition regulator, Polycomb Repressive Complex 2 (PCR2), were identified. Our results shed the first light on the molecular mechanisms behind perennial ryegrass microspore embryogenesis and enable marker-assisted introgression of androgenic capacity into recalcitrant germplasm of this forage crop of global significance.

**Keywords:** Androgenic capacity · Androgenesis · Anther culture · Doubled haploid (DH) · Perennial ryegrass (Lolium perenne L.) · Genotyping-by-sequencing (GBS) · Albinism · Genome-wide association study (GWAS) · Microspore embryogenesis

## Introduction

In contrast to animals, plant cellular differentiation (cell fate) is both flexible and reversible [1]. In immature male gametophytic cells, a totipotent state can be induced through the application of a stress treatment. Subsequent de-differentiation of such cells into the embryogenic pathway may then be stimulated via their cultivation under suitable in vitro conditions. This process, known as microspore embryogenesis (ME) or androgenesis, ultimately results in the recovery of haploid or, via spontaneous or induced chromosome doubling, diploid completely homozygous individuals [2]. Segregating populations of male gametophytes can thus be transformed into doubled haploids (DHs) in a single generation. These are of great value to fundamental research as well as plant breeding [3]. The practical utility of androgenesis ultimately depends on the efficient production of large numbers of microspore-derived embryos capable of regeneration into green, fertile plants.

The optimum stress and in vitro culture conditions for successful androgenesis are highly species and genotype-dependent [4, 5]. Through decades of empirical research, highly effective isolated microspore culture (IMC) protocols have been developed for barley (*Hordeum vulgare* L.), rapeseed (*Brassica napus* L.) and tobacco (*Nicotiana* spp.). Unfortunately, many economically (Solanaceae, fruit trees) and academically (*Arabidopsis*) important species remain recalcitrant [6]. In monocots, and grasses in particular, high rates of albinism further limit androgenic efficiency [7]. Apart from efforts aimed at establishing which external factors are critical for efficient androgenesis, attempts to uncover the genetic factors controlling ME and plant regeneration have been made.

In many cereal crops, linkage mapping studies have identified chromosomal regions associated with traits related to androgenesis. Quantitative trait loci (QTL) related to embryo production, for example, have been reported in wheat (*Triticum aestivum* L.) [8], barley [9] and triticale (× *Triticosecale* Wittm.) [10, 11]. The combined effect of two QTL on barley chromosomes 5H and 6H explained 51% of variation in green plant recovery [12], although only one QTL on chromosome 3H was implicated in a different study [13]. Two regions on wheat chromosomes 1B and 7B explained 53% of the observed variation in albinism [14], QTL for which have also been reported in barley and triticale [15, 16]. However, due to a lack of protocol uniformity, the diversity of material under study and the high variability inherent to tissue culture, consensus amongst these types of studies is low [17, 18]. In addition, genes underlying any of the reported QTL have not been identified.

Nevertheless, a number of candidate genes have been associated with high levels of ME and plant regeneration by means of gene expression experiments [19]. For example, expression of somatic embryogenesis receptor kinase (SERK) gene *SERK1*, and in some cases *SERK2*, was correlated with embryo production and plant regeneration in species such as Arabidopsis, rapeseed, maize (*Zea mays* L.) and wheat [20-24]. Overexpression of the *APETALA 2* (AP2) transcription factor *BABYBOOM* (BBM), *WUSCHEL* (WUS) and *AGAMOUS-like* (AGL) genes, led to the production of ectopic somatic embryos in Arabidopsis, rapeseed and a number of monocot species and improved in vitro regeneration frequencies [25-27]. Other examples of genes that may control ME are the arabinogalactan-related *EARLY CULTURE ABUNDANT 1* (ECA1) [28], Polycomb Group (PcG) proteins including *FERTILIZATION INDEPENDENT ENDOSPERM* (FIE) [19], BURP-domain proteins like *BnBNM2* [25, 29-31] and the *LEAFY COTYLEDON* (LEC) family of transcription factors [32-34]. Again, similar to the linkage mapping studies, the use of different species, treatments and gene expression platforms as well

as the complexity of the system under study, prohibit conclusive identification of the genes of greatest importance to successful androgenesis [33].

Chromosomal regions or genes associated with androgenic capacity in the most widely grown forage species in temperate agriculture, perennial ryegrass (*Lolium perenne* L.), have not yet been identified. Previous studies concluded that perennial ryegrass' androgenic capacity is under polygenic control, with distinct genetic factors influencing embryo production, plant regeneration and green or albino plant production [35-39]. Additive and dominance effects play a role in embryo and plant production, while green plant production involves dominance effects or the complementation of recessive beneficial alleles. Environmental rather than genetic factors may be the main cause of the high incidence of albinism exhibited by many genotypes [39].

In concert with recent efforts to move towards hybrid perennial ryegrass breeding, the potential of in vitro androgenesis for the efficient production of homozygous lines has been recognized [40-43]. To overcome the problematic recalcitrance of most breeding germplasm, molecular marker-based introgression of beneficial alleles has been proposed [44, 45]. Therefore, the main objective of our study was to identify genetic loci associated with androgenic capacity in a multiparental perennial ryegrass population via a genome-wide association study (GWAS). In addition, we aimed at identifying potential causal genes that may provide clues to the molecular mechanisms behind ME and plant regeneration in this important member of the grass family.

## Materials and Methods

### Plant material and anther culture procedure

A detailed description of most of the plant material and the in vitro AC procedure used here can be found in [39]. Briefly, nine perennial ryegrass genotypes with distinct androgenic capacities were paircrossed as part of a DH induction programme at the DLF A/S research station in Store Heddinge, Denmark (Suppl. Table S5.1). Eleven populations of paircross offspring were grown in 1 L soil filled pots in an unheated greenhouse in Eschikon, Switzerland, vernalized and used as anther donors in 2015 and 2016. Spikes containing microspores in the late-uninucleate stage were harvested and subjected to a 4°C cold stress treatment of 24-72 hours in the dark. After surface sterilization, anthers were aseptically excised and cultured on an adapted 190-2 induction medium [46] in a 90 mm Petri dish, incubated at 26°C with a 16 h photoperiod. After 6-8 weeks, macroscopic embryo-like structures (ELS) were transferred to the regeneration medium for shoot and root induction.

### Phenotypic data collection

To quantify androgenic responses of the anther donor genotypes to in vitro AC, eight phenotypic traits were recorded: (1) anther response as a percentage of anthers producing macroscopic ELS (hereafter 'responding anthers' or RA); (2) embryo production as the number of ELS per 100 anthers cultured (AC); (3) plant, (4) green plant and (5) albino plant production, recorded per 100 AC; and (6) plant, (7) green plant and (8) albino plant regeneration, recorded per 100 ELS cultured. In 2015, a total of 313 genotypes were investigated, while incomplete vernalization prior to 2016 resulted in 116 studied genotypes. A total of 78 genotypes were phenotyped in both years (Suppl. Table S5.1; [39]).

### DNA extraction

Fresh leaf tissue of the anther donor plants was harvested for DNA extraction on a 96-well plate KingFisher Flex Purification System with KingFisher Pure DNA Plant Kits (Thermo Fisher Scientific, Waltham, MA, USA). Genomic DNA was visualized on a 1% agarose gel and quantified with a NanoDrop 8000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

### Genotyping by sequencing library preparation

Genotyping by sequencing (GBS) libraries were prepared by multiplexing single restriction enzyme digested genomic DNA using 192 unique 5-10 bp barcodes, designed with the Deena Bioinformatics online GBS Barcode Generator (http://www.deenabio.com/nl/services/gbsadapters) and synthesized by Microsynth (Balgach, Switzerland).

Per sample, a 20 µL PstI digestion mixture was prepared, containing 10 µL DNA sample (10 ng µL-1), 1 µL PstI (3.5 U µL-1), 2.5 µL barcoded adaptors (0.1 ng µL-1), 2.5 µL common adaptors (0.1 ng µL-1), 2 µL O buffer and 2 µL H2O. Samples were digested for 2 h at 37°C. Ligation with T4 ligase, pooling of 96 samples and purification (Qiagen MinElute PCR Purification Kit; Qiagen, Hilden, Germany) were performed according to Elshire et al. (2011) [47]. Fragments were amplified in volumes of 50 µL, containing 5 µL DNA library, 0.25 µL DreamTaq DNA Polymerase (5 U µL-1), 5 µL 10× DreamTaq Buffer, 5 µL dNTPS (2 mM), 1 µL primers (10 µM; Suppl. Table S2) and 33.75 µL H2O. Thermocycler steps were as follows:

72°C for 5 min, 95°C for 30 s, 21 cycles of 95°C for 10 s, 65°C for 30 s and 72°C for 30 s, with a 5 min final extension at 72°C (GeneAMP PCR System 9700; Thermo Fisher Scientific, Waltham, MA). All enzymes and their associated buffers were purchased from Thermo Fisher Scientific. Purified (as above) fragments were visualized on a 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA) to check for presence of adapter dimers and confirm a majority fragment length of 200-400 bp. If adapter dimers were present, an Agencourt AMPure XP bead purification (Beckman Coulter Inc., Brea, CA, USA) was performed.

**GBS library sequencing**

Two 192-plex and one 39-plex anther donor GBS libraries (423 genotypes in total) were sequenced using 126 bp single-end reads on three lanes of an Illumina HiSeq 2500 platform at the Functional Genomics Center Zurich (FGCZ) of the ETH Zurich, Switzerland.

**GBS data processing, read mapping and variant calling**

Reads were de-multiplexed using sabre (https://github.com/najoshi/sabre) allowing one mismatch. Using Bash commands and custom Perl scripts, reads were trimmed to 100 bp and the frequency (counts) of unique sequences (tags) was summarized per paircross population. Unique tags were backtransformed to FASTQ format. Bowtie v0.12.7 [48] with "--best --strata" and a maximum of two alignments "-m 2" was used to map the FASTQ files to the perennial ryegrass genome v1.0 [49]. Unmapped tags were filtered out using a custom Perl script, resulting in 141,775,689 (20.2% of de-multiplexed) mapped tags. The SAM files as well as the count files were further processed in R v3.3.3 [50].

Numerical factors were set to constrain genotyping to reflect the ploidy level of the genotypes (2n) and the maximum allele number (four) for paircross populations. Cut off values of 100 for the minor allele frequency (minAF) and eight for the minimum allele count (minAC) were used. Unique position identifiers (Upos) were extracted from the SAM files by concatenating the direction (Flag), location (Ref) and position (Pos) data. Low coverage sites were eliminated by retaining only Upos with at least one tag greater than the minAF. From the resulting tags, only those occurring at a frequency greater than 5% were retained.

For genotype calling, all informative, polymorphic nucleotide sites (Isites) across the tags were identified and only informative tags (Itags) with Isites were retained. Two unique alleles at one Isite position were called as heterozygous, while the occurrence of a single allele at one Isite was called as homozygous if its count was greater than the minAC. Informative tags were excluded if the number of unique Isites was greater than the ploidy level, or if the allele number within an Isites was greater than the maximum allele number. Haplotypes were obtained by concatenating alleles at the Isites within each tag, if applicable.

**Genome-wide association mapping (GWAS)**

Population structure was investigated using STRUCTURE v2.3.4 [51], GAPIT v2 [52] as well as the hierarchical clustering hclust() (method = "ward.D") and principal component analysis (PCA) prcomp() functions in R.

Itags were filtered using a minAF threshold of 10% and a minimum of 100 and 50 genotypes in 2015 and 2016, respectively (Suppl. Figure S5.1). Since the phenotypic data did not, and could not be made to fit the criteria for parametric testing [39], the non-parametric, rank-based KruskalWallis (K-W) test was used to detect associations between each segregating haplotype (Itag) and the phenotypic traits [53, 11]. For each of these K-W tests, 10,000

permutations of the phenotypes were run. Associations were considered significant at a K-W LOD of 3.0 or higher and a permutation test threshold of 1%. Bonferroni corrected Dunn's tests ($P \leq 0.05$) were carried out *post hoc* to compare haplotypes' trait values. All statistical analyses were performed using custom scripts in Rstudio v1.0.143 [54], running R v3.3.3 [50]. The R packages ggplot2 [55] and UpSetR [56] were used to generate the figures.

Scaffolds of the perennial ryegrass genome v1.0 [49] containing significant Itags will hereafter be referred to as "significant scaffolds".

**Positioning the significant scaffolds on the GenomeZipper**

Significant scaffolds were compared against the genome sequences of *Brachypodium distachyon*, rice (*Oryza sativa* Japonica Group) and sorghum (*Sorghum bicolor* L.) using a BLASTN search ($E \leq 1e\text{-}5$, sequence identity $\geq 85\%$, match length of $\geq 150$ bp). Matches were compared to the perennial ryegrass GenomeZipper [57] in order to obtain the (approximate) locations of the scaffolds of interest on the linkage groups (LGs).

**Gene annotation**

Gene prediction and annotation has been performed as described in Knorst et al. 2017 (*under revision*; annotation data were deposited at zenodo.org).

## Results

### Phenotypic data

The genotype-dependent response to AC, the wide segregation of androgenic capacity within and the differences between the performance of the bi-parental mapping populations, have been described in detail in Begheyn et al. (2017) [39]. In addition, a further eighteen genotypes were included in this study (populations 12 and 15; Suppl. Table S5.1). A detailed summary of the phenotypic traits can be found in Table 1. A total of 313 and 116 genotypes were subjected to in vitro AC in 2015 and 2016, respectively, with an overlap of 78 genotypes between the two years [39]. While observations ranged from zero to several hundred or even over 1,000 in the case of plant and green plant production, the majority were zeros (mode = 0) or close to zero (medians; Table 1). As a consequence, all of the eight androgenic capacity-related traits were, even upon transformation, not normally distributed [39], which necessitated the use of nonparametric statistics for the GWAS analyses [58].

**Table 1.** Summary of the androgenic capacity-related phenotypic traits under study [39].

| Trait | Min | Max | Median | Interquartile range | Number of genotypes |
|---|---|---|---|---|---|
| 2015 | | | | | |
| RA (%) | 0 | 86 | 7.9 | 27.5 | 313 |
| ELS per 100 AC | 0 | 665 | 21 | 94.9 | 307 |
| Plants per 100 AC | 0 | 1810 | 2.4 | 54 | 305 |
| Plants per 100 EC | 0 | 800 | 38.5 | 95.2 | 229 |
| GP per 100 AC | 0 | 1530 | 0 | 6 | 297 |
| GP per 100 EC | 0 | 335 | 0 | 25 | 229 |
| AP per 100 AC | 0 | 705 | 2 | 28 | 297 |
| AP per 100 EC | 0 | 800 | 21.1 | 52.6 | 229 |
| 2016 | | | | | |
| RA (%) | 0 | 87 | 13 | 18 | 116 |
| ELS per 100 AC | 0 | 933 | 73 | 117 | 116 |
| Plants per 100 AC | 0 | 1609 | 0 | 9 | 116 |
| Plants per 100 EC | 0 | 425 | 0 | 18.3 | 105 |
| GP per 100 AC | 0 | 1203 | 0 | 0 | 115 |
| GP per 100 EC | 0 | 318 | 0 | 0 | 104 |
| AP per 100 AC | 0 | 942 | 0 | 6.6 | 115 |
| AP per 100 EC | 0 | 270 | 0 | 14.4 | 104 |

AC – anthers cultured; AP – albino plants; ELS – embryo-like structures; EC – 100 ELS cultured; GP – green plants; RA – responsive anthers

### Genotyping by sequencing (GBS)

Sequencing of the GBS libraries yielded a total of 884,174,849 raw, or 701,662,007 de-multiplexed reads. Of these, 141,775,689 (20.2%) were mapped to the perennial ryegrass genome assembly v1.0 [49]. After removing non-polymorphic tags (75.6%) and stringent filtering (see Materials and Methods), 1120 and 1079 informative tags of 100 bp, containing a polymorphic SNP or haplotype, could be used for the analysis of the 2015 and 2016 datasets, respectively (Suppl. Figure S5.1). While the majority contained a single SNP, 25.8% (2015)

and 24.2% (2016) of informative tags harboured two or more SNPs. Such sets of SNPs on single tags were treated as haplotypes in subsequent analyses.

Given the multi-parental pedigree of the genotypes used in this study, the necessity for applying a correction for population stratification or structure (kinship) was investigated. No evidence for either was found upon analysis of the genotypic data using STRUCTURE [59], a kinship matrix [60] or hierarchical clustering. In addition, the two principal components of the PCA explained 76.3% and 10.4% of variation, respectively (Suppl. Figure S5.2). It was therefore not deemed necessary to include population structure or relatedness corrections in subsequent analyses.

**Genome-wide association study (GWAS)**

Analysis of the 2015 dataset resulted in the identification of significant associations (LOD ≥ 3.0) between six of the studied traits and nine SNPs as well as five haplotypes. Because two of the tags harbouring these polymorphisms mapped back to the same scaffold (2554) of the perennial ryegrass genome assembly [49], a total of thirteen significant scaffolds were identified (Table 2). No significant associations were found for plant or albino plant regeneration. Analysis of the smaller 2016 dataset yielded seven significant scaffolds (LOD ≥ 3.0) for six traits (Table 2). No significant associations were found for plant production and regeneration and none of the scaffold was significantly associated with a trait in both years given the 3.0 LOD threshold.

Since non-parametric testing does not allow for an estimation of QTL or allelic effects, allele or haplotype medians per significant scaffold and trait, combined with Dunn's tests post hoc to ascertain significant differences ($P \leq 0.05$), are presented instead (Table 2). In the 2015 dataset, for example, differences between the medians of the most and least beneficial SNP or haplotype ranged from 9.7 to 18.1 for percentage responsive anthers, 31.5 to 54.2 ELS per 100 AC and 4.9 to 27 plants per 100 AC. The 2016 dataset included a haplotype (TTTC/TTTC) associated with a median albino plant regeneration of 37.5 compared to 0 for the other haplotypes (CCCG/TTTC and CCCG/CCCG) of the same significant scaffold (3194). The smallest significant differences in median, of less than 1 and 1.2 in the 2015 and 2016 datasets, respectively, were observed for green plant production. Nevertheless, for green plant regeneration, the beneficial allele on scaffold 3723 was associated with a median increase of 62.2 green plants per 100 EC compared to the least beneficial allele (Table 2).

**Table 2.** Overview of the significant scaffolds of the perennial ryegrass genome assembly [49] detected for each trait (LOD ≥ 3.0). Significant differences (P ≤ 0.05) between phenotypic medians are indicated with letters.

| Trait | Scaffold | LG | Position (cM) | LOD | Allele or haplotype | Median | Allele or haplotype | Median | Allele or haplotype | Median |
|---|---|---|---|---|---|---|---|---|---|---|
| **2015** | | | | | | | | | | |
| RA (%) | 815 | 1 | 33.0-33.3 | 3.0 | C/C | 21.0 [a] | C/T | 6.9 [b] | T/T | 6.1 [b] |
| | 233 | 4 | 40.4-40.5 | 3.9 | AC/AC | 17.1 [a] | AC/GT | 10.8 [a] | GT/GT | 1.3 [b] |
| | 16597 | 4 | 52.3-52.4 | 3.4 | GAG/GAG | 19.6 [a] | CGA/CGA | 5.2 [b] | CGA/GAG | 1.5 [b] |
| | 1669 | 5 | 0 | 3.2 | G/G | 14.7 [a] | G/T | 1.3 [b] | | |
| | 2554_2 | 5 | 28.5 | 3.8 | C/C | 11.7 [a] | C/T | 2.0 [b] | | |
| | 2075 | 7 | 43.6-43.7 | 3.3 | GT/GT | 19.4 [a] | TC/TC | 13.8 [a] | GT/TC | 1.3 [b] |
| | 4385 | 7 | 46.5 | 3.1 | TG/TG | 19.0 [a] | GA/TG | 14.2 [a] | GA/GA | 2.4 [b] |
| ELS/100AC | 815 | 1 | 33.0-33.3 | 3.4 | C/C | 73.6 [a] | C/T | 13.1 [b] | T/T | 21.6 [b] |
| | 233 | 4 | 40.4-40.5 | 3.1 | AC/AC | 55.9 [a] | AC/GT | 36.9 [ab] | GT/GT | 1.7 [b] |
| | 16597 | 4 | 52.3-52.4 | 3.9 | GAG/GAG | 62 [a] | CGA/CGA | 8.3 [b] | CGA/GAG | 2.4 [b] |
| | 1669 | 5 | 0 | 3.4 | G/G | 41.9 [a] | G/T | 0.7 [b] | | |
| | 2554_2 | 5 | 28.5 | 4.5 | C/C | 34.9 [a] | C/T | 3.4 [b] | | |
| | 4385 | 7 | 46.5 | 3.9 | TG/TG | 54.6 [a] | GA/TG | 32.8 [a] | GA/GA | 0.8 [b] |
| | 10161 | - | - | 3.5 | C/T | 49.7 [a] | T/T | 47 [a] | C/C | 5.0 [b] |
| Plants/100AC | 16597 | 4 | 52.3-52.4 | 3.0 | GAG/GAG | 27.0 [a] | CGA/CGA | 0.0 [b] | CGA/GAG | 0.0 [b] |
| | 2554_2 | 5 | 28.5 | 4.8 | C/C | 4.9 [a] | C/T | 0.0 [b] | | |
| | 10161 | - | - | 3.3 | C/T | 7.9 [a] | T/T | 3.8 [a] | C/C | 0.0 [b] |
| GP/100AC | 6436 | 2 | 79.6-79.8 | 3.1 | T/T | 1 [a] | C/T | 0.0 [b] | C/C | 0.0 [a] |
| GP/100EC | 3723 | 5 | 4.5-25.4 | 3.1 | C/C | 64.2 [a] | C/T | 2.0 [a] | T/T | 0.0 [b] |
| AP/100AC | 16597 | 4 | 52.3-52.4 | 3.2 | GAG/GAG | 16.3 [a] | CGA/CGA | 0.0 [b] | CGA/GAG | 0.0 [b] |
| | 2554_1 | 5 | 28.5 | 4.0 | G/G | 5.8 [a] | A/G | 0.0 [b] | | |
| | 2554_2 | 5 | 28.5 | 5.3 | C/C | 4.0 [a] | C/T | 0.0 [b] | | |

| Trait | Scaffold | LG | Position (cM) | LOD | Allele or haplotype | Median | Allele or haplotype | Median | Allele or haplotype | Median |
|---|---|---|---|---|---|---|---|---|---|---|
| | 6186 | 7 | 43.6-43.7 | 3.2 | CA/CA | 12.7 [a] | GT/GT | 9.7 [a] | CA/GT | 0.0 [b] |
| | 1607 | 7 | 51.6-51.7 | 3.0 | A/A | 13.1 [a] | A/C | 6.1 [a] | C/C | 0.0 [b] |
| | 123 | 7 | 62.4-62.8 | 3.3 | G/G | 13.8 [a] | A/G | 0.0 [b] | A/A | 0.0 [b] |
| **2016** | | | | | | | | | | |
| RA (%) | 8920 | 4 | 22.2-22.3 | 3.4 | CC/TT | 21.0 [a] | TT/TT | 13.0 [a] | CC/CC | 9.0 [b] |
| | 15142 | 5 | 28.2 | 3.2 | A/G | 22.0 [a] | A/A | 11.0 [ab] | G/G | 8.0 [b] |
| | 60 | 5 | 28.5 | 3.3 | T/T | 34.0 [a] | C/T | 11.0 [b] | C/C | 8.5 [b] |
| ELS/100AC | 8920 | 4 | 22.2-22.3 | 3.3 | CC/TT | 102.0 [a] | TT/TT | 89.0 [a] | CC/CC | 28.5 [b] |
| | 813 | 5 | 28.5-28.5 | 3.2 | A/G | 173.0 [a] | G/G | 36.0 [b] | | |
| GP/100AC | 127 | 1 | 56.1-57.5 | 3.9 | G/G | 1.2 [a] | A/G | 0.0 [b] | A/A | 0.0 [b] |
| GP/100EC | 127 | 1 | 56.1-57.5 | 4.1 | G/G | 1.6 [a] | A/G | 0.0 [b] | A/A | 0.0 [b] |
| AP/100AC | 7045 | 7 | 37.5-38.6 | 3.3 | C/C | 21.1 [a] | C/T | 0.0 [b] | T/T | 0.0 [b] |
| AP/100EC | 3194 | 1 | 30.9-31.1 | 3.0 | TTTC/TTTC | 37.5 [a] | CCCG/TTTC | 0.0 [b] | CCCG/CCCG | 0.0 [b] |
| | 7045 | 7 | 37.5-38.6 | 3.0 | C/C | 19.8 [a] | C/T | 0.0 [b] | T/T | 0.0 [b] |

AP – albino plants; AC – anthers cultured; ELS – embryo-like structures; EC – ELS cultured; GP – green plants; LG – linkage groups; RA – responsive anthers.

Most significant associations were found for the percentage of responsive anthers (ten associations), embryo production (nine) and albino plant production (seven; Figure 1). Using the 2015 dataset, four scaffolds (815, 233, 1669 and 4385) were significant for both the percentage of responsive anthers as well as ELS production, while two scaffolds (16597 and 2554) were significantly associated with percentage responsive anthers and the production of ELS, plants and albino plants. Scaffold 10616 was significantly associated with ELS and plant production. Three scaffolds, 8920, 127 and 7045 were found to be significant for two traits using the 2016 dataset.

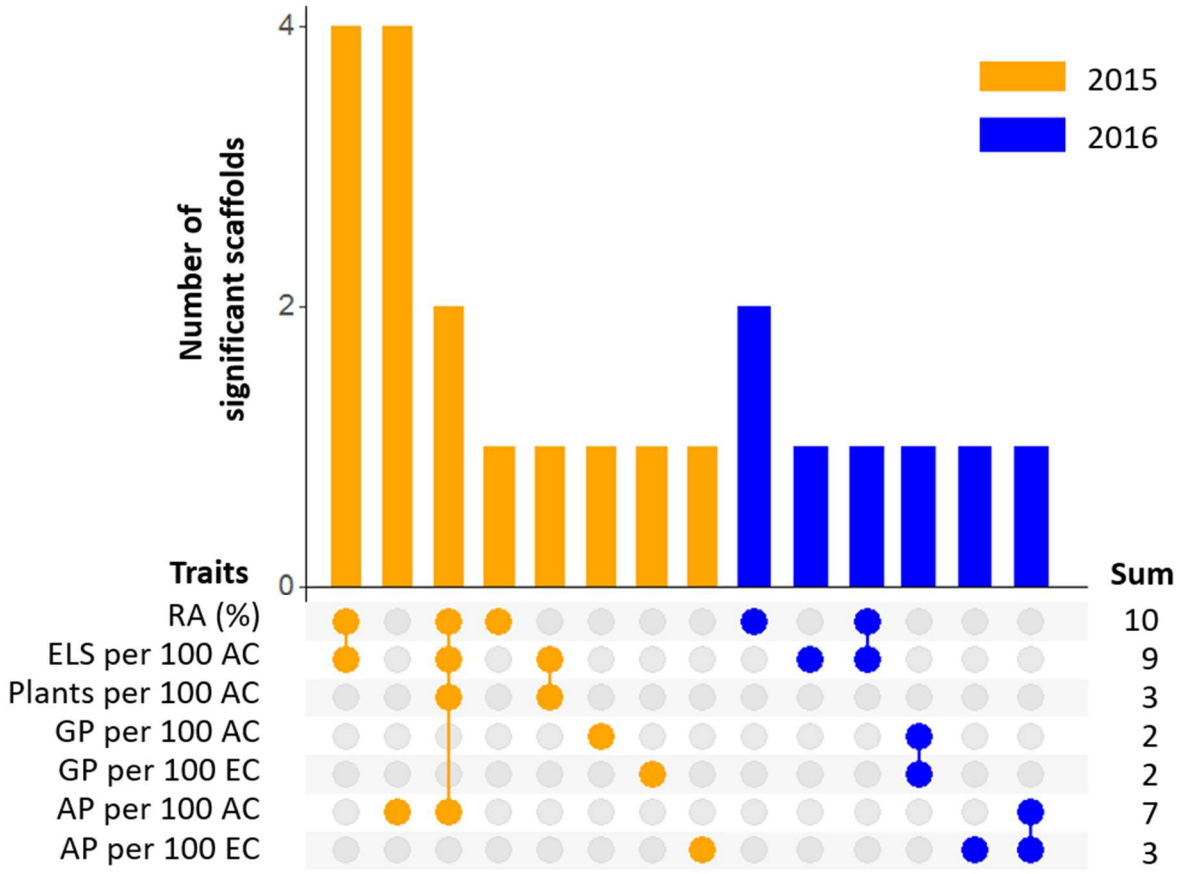



**Figure 1.** Overview of the number of significant scaffolds per trait or, shown with connected dots, per group of traits (bars) and the total number of significant scaffolds per trait (sum). AP – albino plants; AC – anthers cultured; ELS – embryo-like structures; EC – ELS cultured; GP – green plants; RA – responsive anthers.

### Positioning significant scaffolds on the GenomeZipper

By comparing *B. distachyon*, rice and sorghum genes homologs identified on the significant scaffolds with those anchored on the perennial ryegrass GenomeZipper [57], all but one scaffold could be assigned approximate positions on the LGs (Figure 2). Even so, confidence in the positioning varied from case to case. For example, the approximate positions of scaffolds 123, 127, 233, 813, 2075, 3194, 3723, 6186, 15142 and 16597 were resolved via one or several exact gene matches to the same location on the GenomeZipper. Scaffolds 60, 815, 1607, 1669, 2554, 4385, 6436, 7045 and 8920 were positioned (approximately) using three to ten genes that were not anchored on the GenomeZipper, but could be placed between several

genes anchored at the same location. Scaffold 10616 could not be assigned a location because no significant BLASTN hits of sufficient length were obtained.

Even though no scaffold was found to be significant in both years, scaffolds identified in different years were positioned in similar locations on the GenomeZipper LGs (Figure 2). Scaffolds 815 (2015) and 3194 (2016) are approximately 2 cM apart on LG 1 for example, while scaffolds 60, 813 and 15142 (2016) and 2554 (2015) are all positioned within a 0.3 cM region on LG 5. On the lower middle region of LG 7, scaffolds 2075 and 6186 (43.6 to 43.7 cM) and 4383 (46.5 cM) from the 2015 dataset were positioned in close proximity to each other.

No scaffolds were positioned on LGs 3 and 6. Scaffolds associated with the percentage of responsive anthers, ELS production and at least one of the albino plant-related traits were positioned on LGs 1, 4, 5 and 7, mostly relatively close together. Also amidst these, on LGs 4 and 5, were the two plant production-related scaffolds (2554 and 16597) that could be placed on the GenomeZipper. The three scaffolds (127, 3723 and 6463) significantly associated to the green plant-related traits were some distance away from the scaffolds associated to the other traits. In fact, scaffold 6436 was the only scaffold positioned on LG 2.

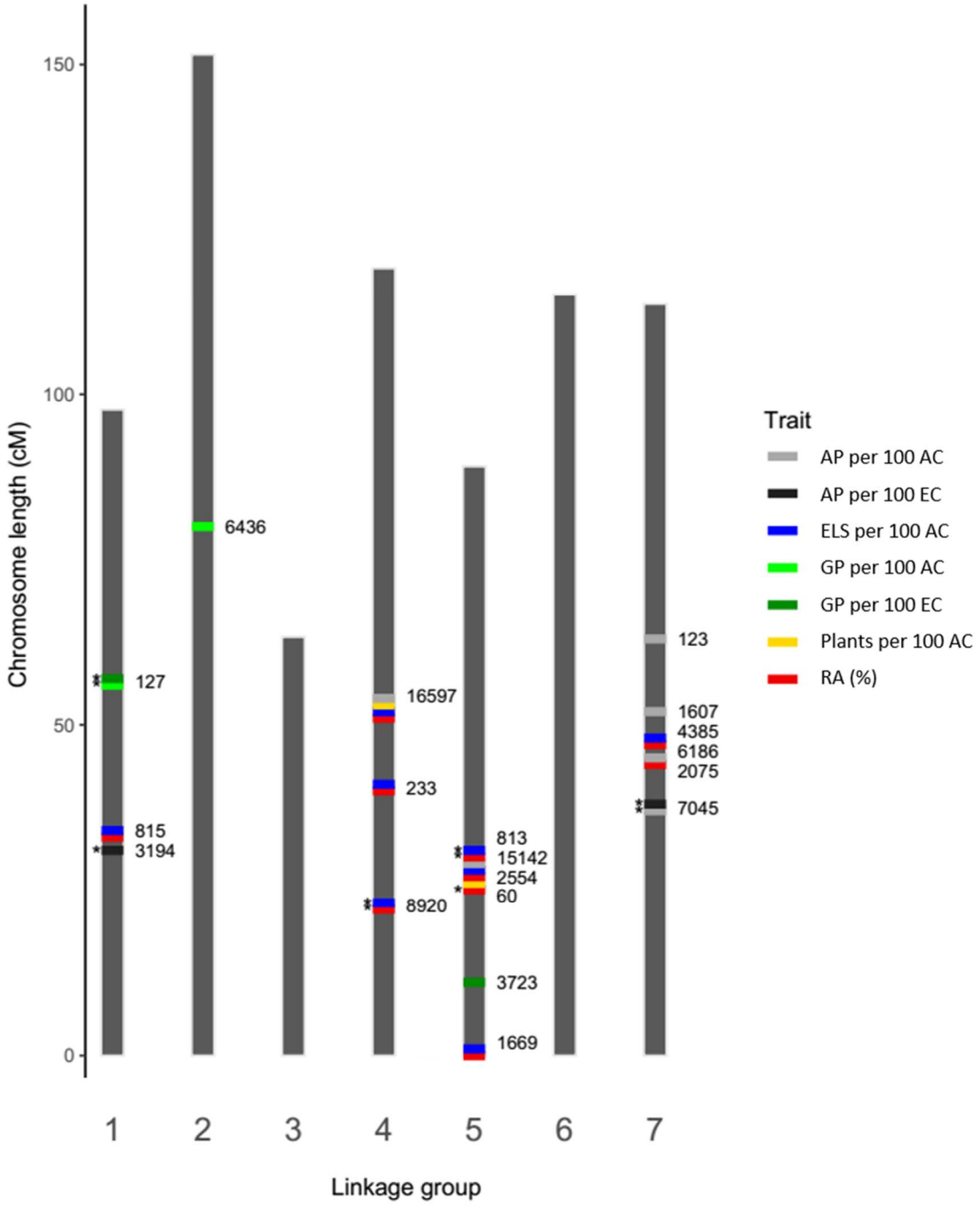



**Figure 2.** Positions of the significant scaffolds detected in 2015 and 2016 (*) on the perennial ryegrass genome as inferred by the perennial ryegrass GenomeZipper [57]. AP – albino plants; AC – anthers cultured; ELS – embryo-like structures; EC – ELS cultured; GP – green plants; RA – responsive anthers.

## Gene annotations

Between one and four predicted genes were annotated for each significant scaffold, with the exception of scaffold 10616 (Suppl. Table S5.2). On scaffold 1607 for example, sequence homology to the Arabidopsis *SERRATE* (SE) gene was found, while homologues of two domains of Polycomb Repressive Complex 2 (PRC2), *FERTILIZATION-INDEPENDENT ENDOSPERM* (FIE) and *CURLY LEAF* (CLF) were identified on scaffolds 4383 and 7045, respectively.

## Discussion

Here, we present the first report of genetic loci associated with in vitro androgenesis in perennial ryegrass. Between two and ten QTL (LOD ≥ 3.0) for anther response percentage, embryo production, total plant production as well as green and albino plant production and regeneration were identified on five of the seven perennial ryegrass LGs. Additionally, several intriguing candidate genes that may be responsible for the observed phenotypic differences were predicted on the QTL-harbouring scaffolds of the perennial ryegrass genome assembly [49]. These results enable the development of the first molecular markers for androgenic capacity in perennial ryegrass, from the identified, polymorphic GBS tags. Their availability will help to realize the long-standing aim of efficient, marker-assisted introgression of good responses to in vitro DH induction into recalcitrant germplasm [44, 45].

### Multi-parental population GWAS in perennial ryegrass

Contrary to previous QTL studies on androgenic capacity, which were based on linkage mapping in bi-parental populations of up to 100 individuals [11, 13, 14], an association mapping approach in a multi-parental population, composed of 391 heterozygous individuals, was applied here. This design increased the presence of distinct alleles, confirmed by the observed phenotypic variation [39], and, due to the recombination between the nine heterozygous parents, ensured high levels of allelic diversity as well as good mapping resolution. Around 1100 polymorphic SNPs and haplotypes, identified using a methylation-sensitive GBS protocol, allowed for the genome-wide interrogation of gene-dense regions within the multi-parental mapping population [61]. Significant population structure was absent, due to the common breeding history of the parental plants used to design the mapping population. This powerful experimental design, combined with robust, non-parametric (K-W) single SNP/haplotype genome-wide analysis and permutation-based validation, was successfully used to detect significant QTL (LOD ≥ 3.0) associated with the component traits of the androgenic response of perennial ryegrass.

### A putative major QTL for green plant regeneration on perennial ryegrass LG 5

Authors have often commented on the difficulty of comparing tissue culture experiments, due to highly genotype-specific responses as well as crucial differences in execution and data collection [17, 18]. Fortunately, comparative genomics studies within the grass family allow for an interspecific comparison of cereal AC and IMC QTL studies, albeit at the chromosomal level [62]. Most homologous grass chromosomes have been associated with all of the androgenicity-related traits at least once, however, and a common pattern is not obvious. One possible exception is a putative locus controlling green plant regeneration, which was identified on Triticeae chromosome group 5 and reported to affect 12-37% of the phenotypic variation in barley, rice (chromosome 9), triticale and wheat [11-13, 63, 64]. Intriguingly, we identified a putative major QTL, associated with a median increase of 62 green plants per 100 AC, on perennial ryegrass LG 5 as well [57]. This locus is therefore of great interest and its further investigation, for example using fine-mapping approaches, may lead to the identification of the gene with a considerable effect on green plant regeneration in the grass family.

### Genetic control of androgenic capacity

A relatively large number of QTL with modest effects were associated with androgenic traits, such as anther response percentage (ten QTL), embryo production (nine QTL) and albino

plant production (seven QTL). In addition, many QTL were shown to affect several traits, confirming the high correlations between, for example, embryo production and anther response as well as plant production observed earlier [39]. Similar results have been reported by other groups [9, 11, 65, 66]. Finally, QTL detected in 2015 were not detected in 2016 and vice versa, although the QTL identified on scaffold 2075 using the 2015 dataset had a LOD of 2.0 using the 2016 dataset for percentage responsive anthers (results not shown). The discrepancy is probably caused by the fact that only 78 genotypes from four bi-parental crosses were subjected to AC in both years and just 45 of those had the same paircross parents (population 1). Allele frequencies of QTL detected using the 2015 dataset were likely too low, or entirely absent, from the 2016 dataset, which in turn harboured distinct beneficial alleles at a high enough frequency for QTL detection. Although a smaller dataset was used in 2016, several QTL of particular interest were detected. For example, a QTL on scaffold 813 was associated with a major median increase in embryo production of 137 ELS per 100 anthers cultured. In addition, the only QTL (on scaffolds 3194 and 7045) associated with albino plant regeneration, connected with an median increase of 19.8 and 37.5 albino plants per 100 ELS cultured, were detected using this dataset.

All of the above findings may be explained by the fact that both ME and albinism during in vitro culture are under complex, polygenic and heterogeneous control [2, 67]. A single genetic master switch for ME has never been identified [19] and albino phenotypes can be caused by mutations in as many as 300 nuclear genes [19]. A significant increase in embryo production may, therefore, be accomplished via the stacking of several genetic loci with modest effect within single genotypes [38, 68]. The production of albinos may be reduced by similar means.

A relatively small number of QTL were associated with plant production, green plant production and green and albino plant regeneration. The three QTL detected for total plant production also affected either embryo production, albino production or both. Conversely, the QTL that influenced green plant production (2 QTL) and regeneration (2 QTL) were not associated with any other traits and positioned at distinct locations on the perennial ryegrass LGs. In addition, only one of the two QTL related to albino plant regeneration was affected a second trait, albino plant production. These results do not only confirm the separate genetic control of green and albino plant production capacity reported previously [10, 11, 39, 63]. They also suggest that total plant production and total plant regeneration, for which no QTL were identified at all, may not be of great use to describe androgenic ability. The three phases of in vitro androgenesis that are commonly distinguished, 1) embryo production, 2) plant regeneration and 3) green plant recovery, can, at least in the grass family, be redefined as 1) embryo production, 2a) green plant recovery and 2b) albino plant recovery. Green plant recovery seems to be controlled by fewer loci than albino plant recovery, although environmental influence on albinism may have masked both green plant production and regeneration capacity as well as the QTL associated with them [39].

**Candidate genes involved in androgenic response**

While the putative function of most candidate genes underlying the QTL identified here has yet to be resolved, several have previously been associated with the regulation of stress response, cell fate change, embryogenesis or organogenesis. The ISOPRENYLCYSTEINE METHEYLESTERASE-LIKE 2 (ICMELIKE2) gene annotated on scaffold 123, for instance, is involved in abscisic (ABA) mediated stress signaling and specifically expressed in

reproductive organs of Arabidopsis [69]. Similarly, the VIP HOMOLOG 1 (VIH1) gene, identified on scaffold 233, is crucial to certain aspects of jasmonate mediated stress signalling and is mainly expressed in Arabidopsis pollen [70]. Phytohormones like ABA and jasmonic acid (JA) have, in fact, been shown to play important roles during androgenesis by ensuring microspore viability through the regulation of stress responses as well as inducing ME via signalling cascades that activate specific gene expression programs [71- 73]. The Arabidopsis *SERRATE* (SE) gene, which is involved in chromatin modification and microRNA-mediated gene expression regulation during organogenesis, was annotated on scaffold 1607 [74, 75]. Embryonic lethality and defective post-embryonic organ formation have been reported in Arabidopsis se mutants, indicating a possible role for SE during plant regeneration after successful ME [74, 76, 77].

Most intriguing, however, was the annotation of orthologs to two genes encoding distinct domains of the Polycomb Repressive Complex 2 (PRC2), a highly conserved and important regulator of developmental processes, on scaffolds 4385 and 7045 [78]. The first, *CURLY LEAF* (CLF), encodes one of three SET domain proteins, the others being *MEDEA* (MEA) and *SWINGER* (SWN), which mediate large-scale chromatin remodelling during embryogenic development [79]. In fact, the mannitol stress treatment used prior to barley IMC was found to induce the upregulation of CLF in anther tissue [80]. The second homolog is a FIE domain which is associated with MEA in the gametophytic- and endosperm-specific configuration of the PRC2. In Arabidopsis, fie as well as clf swn double mutants are unable to terminate the embryogenic phase of germination and proliferate into so-called PcG callus [81, 82]. Furthermore, the PRC2 complex is involved in the negative regulation of the LEC family as well as WUS genes, both of which play key roles in somatic and ME [27, 83]. In fact, LEC1, LEC2 and FUS3 are overexpressed in clf swn double mutants of Arabidopsis [84]. Indeed, LEC1 (over-)expression was shown to negatively affect ME in both rapeseed and rye [32, 34]. Interestingly, a homolog of the MADS box gene AGL26, was annotated along with FIE on scaffold 4385. Several MADS box transcription factors, which are key regulators of developmental processes, are negatively regulated by PRC2 as well [85]. Ultimately, the distinct phases of in vitro androgenesis are likely to require different levels of PRC2 mediated repression of specific genes [78]. Quantification or manipulation of the expression of CLF, FIE, AGL26 or any of the other candidate genes during different stages of perennial ryegrass in vitro AC could confirm their contribution to successful androgenesis and should determine if and when their expression is most beneficial.

**Concluding remarks**

Here, we have demonstrated the effectivity of a multi-parental genome-wide association mapping approach in perennial ryegrass and report the first genetic loci associated with the response to in vitro AC. Elucidation of the exact locations of the QTL detected here will, however, require the availability of a more complete perennial ryegrass genome assembly. It can then be ascertained whether the colocalization of several QTL associated with different traits or detected in different years was, in fact, accurately determined using the GenomeZipper [57]. Future studies on the genetic control of androgenic capacity may then focus on these important regions. Of particular interest is a major QTL for green plant regeneration on LG 5 which, if proven to be effective in different genomic backgrounds, is an excellent candidate for further fine mapping approaches. A second major QTL for embryo production on LG 1 was detected in the smaller of the two datasets that were used here, but nevertheless merits additional investigation. Two of the identified candidate genes, CLF and FIE, are of great potential

interest, given their extensively documented involvement in embryogenesis and organogenesis, although expression studies will have to provide further evidence of their involvement in perennial ryegrass ME [78]. Presently, our results allow for the development of molecular markers which will enable efficient introgression of androgenic capacity into recalcitrant perennial ryegrass germplasm. The availability of an efficient system for homozygous line production will aid in the establishment of a hybrid breeding system, which should increase the rate of genetic gain in this forage crop of global importance.

**Acknowledgements**

This work was supported by ETH Research Grant ETH-34 14-1 and the Swiss National Science Foundation (SNSF Professorship grant No.: PP00P2 138988). We thank Verena Knorst for taking excellent care of the plant material. The sequencing data produced and analysed in this paper were generated in collaboration with the Genetic Diversity Centre (GDC) and the Functional Genomics Center Zurich (FGCZ) of the ETH Zurich, Switzerland. We are indebted to Prof. Dr. Achim Walter and the Crop Science group for having hosted the Molecular Plant Breeding group at the ETH Zurich during most of this project.

# References


1. Walbot V., Evans M. M. S., 2003 Unique features of the plant life cycle and their consequences. Nat. Rev. Genet. 4: 369–79.
2. Seguí-Simarro J. M., Nuez F., 2008 How microspores transform into haploid embryos: changes associated with embryogenesis induction and microspore-derived embryogenesis. Physiol. Plant. 134: 1–12.
3. Forster B. P., Heberle-Bors E., Kasha K. J., Touraev A., 2007 The resurgence of haploids in higher plants. Trends Plant Sci. 12: 368–75.
4. Seguí-Simarro J. M., 2010 Androgenesis revisited. Bot. Rev. 76: 377–404.
5. Dwivedi S. L., Britt A. B., Tripathi L., Sharma S., Upadhyaya H. D., et al., 2015 Haploids: constraints and opportunities in plant breeding. Biotechnol. Adv. 33: 812–829.
6. Seguí-Simarro J. M., 2015 Editorial: doubled haploidy in model and recalcitrant species. Front. Plant Sci. 6: 1–2.
7. Kumari M., Clarke H. J., Small I., Siddique K. H. M., 2009 Albinism in plants: a major bottleneck in wide hybridization, androgenesis and doubled haploid culture. CRC. Crit. Rev. Plant Sci. 28: 393–409.
8. Agache S., Bachelier B., Buyser J. De, Henry Y., Snape J., 1989 Genetic analysis of anther culture response in wheat using aneuploid, chromosome substitution and translocation lines. Theor. Appl. Genet. 77: 7–11.
9. Manninen O. M., 2000 Associations between anther-culture response and molecular markers on chromosomes 2H, 3H and 4H of barley (Hordeum vulgare L.). TAG Theor. Appl. Genet. 100: 57–62.
10. González J. M., Muñiz L. M., Jouve N., 2005 Mapping of QTLs for androgenetic response based on a molecular genetic map of ×Triticosecale Wittmack. Genome 48: 999–1009.
11. Krzewska M., Czyczyło-Mysza I., Dubas E., Gołębiowska-Pikania G., Golemiec E., et al., 2012 Quantitative trait loci associated with androgenic responsiveness in triticale (×Triticosecale Wittm.) anther culture. Plant Cell Rep. 31: 2099–2108.
12. Chen X.-W., Cistué L., Muñoz-Amatriaín M., Sanz M., Romagosa I., et al., 2007 Genetic markers for doubled haploid response in barley. Euphytica 158: 287–294.
13. Muñoz-Amatriaín M., Castillo a. M., Chen X. W., Cistué L., Vallés M. P., 2008 Identification and validation of QTLs for green plant percentage in barley (Hordeum vulgare L.) anther culture. Mol. Breed. 22: 119–129.
14. Nielsen N. H., Andersen S. U., Stougaard J., Jensen A., Backes G., et al., 2015 Chromosomal regions associated with the in vitro culture response of wheat (Triticum aestivum L.) microspores. Plant Breed. 134: 255–263.
15. Bregitzer P., Campbell R. D., 2001 Genetic markers associated with green and albino plant regeneration from embryogenic barley callus. Crop Sci. 41: 173.
16. Krzewska M., Czyczyło-Mysza I., Dubas E., Gołębiowska-Pikania G., Żur I., 2015 Identification of QTLs associated with albino plant formation and some new facts concerning green versus albino ratio determinants in triticale (×Triticosecale Wittm.) anther culture. Euphytica 206: 263–278.
17. Bolibok H., Rakoczy-Trojanowska M., 2006 Genetic mapping of QTLs for tissue-culture response in plants. Euphytica 149: 73–83.
18. Seldimirova O. A., Kruglova N. N., 2015 Androclinic embryoidogenesis in vitro in cereals. Biol. Bull. Rev. 5: 156–165.
19. Hand M. L., Vries S. De, Koltunow A. M. G., Vries S. de, Kotunow A. M. G., 2016 A comparison of in vitro and in vivo asexual embryogenesis. In: Germana M, Lambardi M (Eds.), In vitro embryogenesis in higher plants, Humana Press, New York, NY, pp. 3–23.
20. Hu H., Xiong L., Yang Y., 2005 Rice SERK1 gene positively regulates somatic embryogenesis of cultured cell and host defense response against fungal infection. Planta 222: 107–117.
21. Singla B., Khurana J. P., Khurana P., 2008 Characterization of three somatic embryogenesis receptor kinase genes from wheat, Triticum aestivum. Plant Cell Rep. 27: 833–843.
22. Podio M., Felitti S. A., Siena L. A., Delgado L., Mancini M., et al., 2014 Characterization and expression analysis of SOMATIC EMBRYOGENESIS RECEPTOR KINASE (SERK) genes in sexual and apomictic Paspalum notatum. Plant Mol. Biol. 84: 479–495.

23. Ahmadi B., Masoomi-Aladizgeh F., Shariatpanahi M. E., Azadi P., Keshavarz-Alizadeh M., 2016 Molecular characterization and expression analysis of SERK1 and SERK2 in Brassica napus L.: implication for microspore embryogenesis and plant regeneration. Plant Cell Rep. 35: 185–193.
24. Seifert F., Bössow S., Kumlehn J., Gnad H., Scholten S., 2016 Analysis of wheat microspore embryogenesis induction by transcriptome and small RNA sequencing using the highly responsive cultivar “Svilena.” BMC Plant Biol.: 1–16.
25. Boutilier K., 2002 Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. Plant Cell Online 14: 1737–1749.
26. Muñoz-Amatriaín M., Svensson J. T., Castillo A. M., Close T. J., Vallés M. P., 2009b Microspore embryogenesis: Assignment of genes to embryo formation and green vs. albino plant production. Funct. Integr. Genomics 9: 311–323.
27. Lowe K., Wu E., Wang N., Hoerster G., Hastings C., et al., 2016 Morphogenic regulators Baby boom and Wuschel improve monocot transformation. Plant Cell: tpc.00124.2016.
28. Vrinten P. L., Nakamura T., Kasha K. J., 1999 Characterization of cDNAs expressed in the early stages of microspore embryogenesis in barley (Hordeum vulgare) L. Plant Mol. Biol. 41: 455–63.
29. Tsuwamoto R., Fukuoka H., Takahata Y., 2007 Identification and characterization of genes expressed in early embryogenesis from microspores of Brassica napus. Planta 225: 641–652.
30. Joosen R., Cordewener J., Supena E. D. J., Vorst O., Lammers M., et al., 2007 Combined transcriptome and proteome analysis identifies pathways and markers associated with the establishment of rapeseed microspore-derived embryo development. Plant Physiol. 144: 155 LP-172.
31. Malik M. R., Wang F., Dirpaul J. M., Zhou N., Polowick P. L., et al., 2007 Transcript profiling and identification of molecular markers for early microspore embryogenesis in Brassica napus. Plant Physiol. 144: 134 LP-154.
32. Gruszczyńska A., Rakoczy-Trojanowska M., 2011 Expression analysis of somatic embryogenesis-related SERK, LEC1, VP1 and NiR ortologues in rye (Secale cereale L.). J. Appl. Genet. 52: 1–8.
33. Soriano M., Li H., Boutilier K., 2013 Microspore embryogenesis: establishment of embryo identity and pattern in culture. Plant Reprod. 26: 181–96.
34. Elahi N., Duncan R. W., Stasolla C., 2016 Effects of altered expression of LEAFY COTYLEDON1 and FUSCA3 on microspore-derived embryogenesis of Brassica napus L. J. Genet. Eng. Biotechnol. 14: 19–30.
35. Olesen A., Andersen S. B., Due I. K., 1988 Anther culture response in perennial ryegrass (Lolium perenne L.). Plant Breed. 101: 60–65.
36. Boppenmeier J., Zuchner S., Foroughi-Wehr B., 1989 Haploid production from barley yellow dwarf virus resistant clones of Lolium. Plant Breed. 103: 216–220.
37. Opsahl-Ferstad H. G., Bjørnstad Å., Rognli O. A., 1994 Genetic control of androgenetic response in Lolium perenne L. Theor. Appl. Genet. 89: 133–8.
38. Madsen S., Olesen A., Dennis B., Andersen S. B., 1995 Inheritance of anther-culture response in perennial ryegrass (Lolium perenne L.). Plant Breed. 114: 165–168.
39. Begheyn R. F., Roulund N., Vangsgaard K., Kopecký D., Studer B., 2017 Inheritance patterns of the response to in vitro doubled haploid induction in perennial ryegrass (Lolium perenne L.). Plant Cell, Tissue Organ Cult.: 1–13.
40. Arias Aguirre A., Studer B., Frei U., Lübberstedt T., 2011 Prospects for hybrid breeding in bioenergy grasses. BioEnergy Res. 5: 10–19.
41. Begheyn R. F., Lübberstedt T., Studer B., 2016 Haploid and Doubled Haploid Techniques in Perennial Ryegrass (Lolium perenne L .) to Advance Research and Breeding. Agronomy 6: 1–17.
42. Manzanares C., Barth S., Thorogood D., Byrne S. L., Yates S., et al., 2016 A gene encoding a DUF247 domain protein cosegregates with the S self-incompatibility locus in perennial ryegrass. Mol. Biol. Evol. 33: 870–884.
43. Sykes T., Yates S., Nagy I., Asp T., Small I., et al., 2016 In-silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (Lolium perenne L.). Genome Biol. Evol.: evw047.

44. Halberg N., Olesen A., Tuvesson I. K. D., Andersen S. B., 1990 Genotypes of perennial ryegrass (Lolium perenne L.) with high anther-culture response through hybridization. Plant Breed. 105: 89–94.
45. Andersen S. B., Madsen S., Roulund N., Halberg N., Olesen A., 1997 Haploidy in ryegrass. In: Jain SM, Sopory SK, Veilleux RE (Eds.), In vitro haploid production in higher plants, Springer Netherlands, pp. 133–147.
46. Wang X., Hu H., 1984 The effect of potato II medium for triticale anther culture. Plant Sci. Lett. 36: 237–239.
47. Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., et al., 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.
48. Langmead B., Trapnell C., Pop M., Salzberg S. L., 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.
49. Byrne S. L., Nagy I., Pfeifer M., Armstead I., Swain S., et al., 2015 A synteny-based draft genome sequence of the forage grass Lolium perenne. Plant J. 84: 816–826.
50. R Core Team, 2017 R: a language and environment for statistical computing.
51. Hubisz M. J., Falush D., Stephens M., Pritchard J. K., 2009 Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour. 9: 1322–1332.
52. Lipka A. E., Tian F., Wang Q., Peiffer J., Li M., et al., 2012 GAPIT: Genome association and prediction integrated tool. Bioinformatics 28: 2397–2399.
53. Kiviharju E., Laurila J., Lehtonen M., Tanhuanpää P., Manninen O., 2004 Anther culture properties of oat x wild red oat progenies and a search for RAPD markers associated with anther culture ability. Agric. Food Sci. 13: 151–162.
54. RStudio Team, 2015 Rstudio: integrated development for R.
55. Wickham H., 2009 ggplot2: Elegant graphics for data analysis.
56. Lex A., Gehlenborg N., Strobelt H., Vuillemot R., Pfister H., 2014 UpSet: Visualization of intersecting sets. IEEE Trans. Vis. Comput. Graph. 20: 1983–1992.
57. Pfeifer M., Martis M., Asp T., Mayer K. F. X., Lübberstedt T., et al., 2013 The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. Plant Physiol. 161: 571–82.
58. Rebaï A., 1997 Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. Genet. Res. 69: 69–74.
59. Porras-Hurtado L., Ruiz Y., Santos C., Phillips C., Carracedo Á., et al., 2013 An overview of STRUCTURE: Applications, parameter settings, and supporting software. Front. Genet. 4: 1–13.
60. VanRaden P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–23.
61. Byrne S., Czaban A., Studer B., Panitz F., Bendixen C., et al., 2013 Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. PLoS One 8: e57438.
62. Devos K. M., 2005 Updating the “crop circle.” Curr. Opin. Plant Biol. 8: 155–162.
63. He P., Shen L., Lu C., Chen Y., Zhu L., 1998 Analysis of quantitative trait loci which contribute to anther culturability in rice (Oryza sativa L.). Mol. Breed. 4: 165–172.
64. Torp A. M., Hansen A. L. A., Andersen S. B., 2001 Chromosomal regions associated with green plant regeneration in wheat (Triticum aestivum L.) anther culture. Euphytica 119: 377–387.
65. Murigneux A., Bentolila S., Hardy T., Baud S., Guitton C., et al., 1994 Genotypic variation of quantitative trait loci controlling in vitro androgenesis in maize. Genome 37: 970–976.
66. Beaumont V. H., Rocheford T. R., Widholm J. M., 1995 Mapping the anther culture response genes in maize (Zea mays L.). Genome 38: 968–975.
67. Makowska K., Oleszczuk S., 2014 Albinism in barley androgenesis. Plant Cell Rep. 33: 385–392.
68. Marhic A., Antoine-Michard S., Bordes J., Pollacsek M., Murigneux A., et al., 1998 Genetic improvement of anther culture response in maize: Relationships with molecular, Mendelian and agronomic traits. Theor. Appl. Genet. 97: 520–525.
69. Lan P., Li W., Wang H., Ma W., 2010 Characterization, sub-cellular localization and expression profiling of the isoprenylcysteine methylesterase gene family in Arabidopsis thaliana. BMC Plant Biol. 10: 212.

70. Laha D., Johnen P., Azevedo C., Dynowski M., Weiß M., et al., 2015 VIH2 regulates the synthesis of inositol pyrophosphate InsP8 and jasmonate-dependent defenses in Arabidopsis. Plant Cell 27: 1082–1097.
71. Maraschin S. F., Priester W. De, Spaink H. P., Wang M., 2005 Androgenic switch: An example of plant embryogenesis from the male gametophyte perspective. J. Exp. Bot. 56: 1711–1726.
72. Ahmadi B., Shariatpanahi M. E., Silva J. A. da, 2014 Efficient induction of microspore embryogenesis using abscisic acid, jasmonic acid and salicylic acid in Brassica napus L. Plant Cell, Tissue Organ Cult. 116: 343–351.
73. Żur I., Dubas E., Krzewska M., Janowiak F., 2015a Current insights into hormonal regulation of microspore embryogenesis. Front. Plant Sci. 6: 424.
74. Grigg S. P., Canales C., Hay A., Tsiantis M., 2005 SERRATE coordinates shoot meristem function and leaf axial patterning in Arabidopsis. Nature 437: 1022–1026.
75. Yang L., Liu Z., Lu F., Dong A., Huang H., 2006 SERRATE is a novel nuclear regulator in primary microRNA processing in Arabidopsis. Plant J. 47: 841–850.
76. Prigge M. J., Wagner D. R., 2001 The Arabidopsis SERRATE gene encodes a zinc-finger protein required for normal shoot development. Plant Cell 13: 1263–1279.
77. Lobbes D., Rallapalli G., Schmidt D. D., Martin C., Clarke J., 2006 SERRATE: a new player on the plant microRNA scene. EMBO Rep. 7: 1052–1058.
78. Förderer A., Zhou Y., Turck F., 2016 The age of multiplexity: Recruitment and interactions of Polycomb complexes in plants. Curr. Opin. Plant Biol. 29: 169–178.
79. Liu J., Deng S., Wang H., Ye J., Wu H.-W., et al., 2016 CURLY LEAF regulates gene sets coordinating seed size and lipid biosynthesis. Plant Physiol. 171: 424 LP-436.
80. Muñoz-Amatriaín M., Svensson J. T., Castillo A. M., Cistué L., Close T. J., et al., 2009a Expression profiles in barley microspore embryogenesis. In: Touraev A, Forster BP, Jain SM (Eds.), Advances in Haploid Production in Higher Plants, Springer, Dordrecht, Dordrecht, pp. 127–134.
81. Chanvivattana Y., Bishopp A., Schubert D., Stock C., Moon Y.-H., et al., 2004 Interaction of Polycomb-group proteins controlling flowering in Arabidopsis. Development 131: 5263–76.
82. Bouyer D., Roudier F., Heese M., Andersen E. D., Gey D., et al., 2011 Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. PLoS Genet. 7.
83. Berger N., Dubreucq B., Roudier F., Dubos C., Lepiniec L., 2011 Transcriptional regulation of Arabidopsis LEAFY COTYLEDON 2 involves RLE, a cis-element that regulates trimethylation of histone H3 at lysine-27. Plant Cell 23: 4065–4078.
84. Makarevich G., Leroy O., Akinci U., Schubert D., Clarenz O., et al., 2006 Different Polycomb group complexes regulate common target genes in Arabidopsis. EMBO Rep. 7: 947 LP-952.
85. Masiero S., Colombo L., Grini P. E., Schnittger A., Kater M. M., 2011 The emerging importance of type I MADS box transcription factors for plant reproduction. Plant Cell 23: 865–872.

## Supplementary Data

**Table S1.** Overview of paircross parents and their progeny populations used for phenotypic evaluation of *in vitro* anther culture capacity.

| Population | Parents | | | Number of genotypes evaluated in | | |
|---|---|---|---|---|---|---|
| | Androgenic | | Non-androgenic | 2015 | 2016 | 2015 and 2016 |
| 1 | P2 | × | P133[1] | 49 | 50 | 45 |
| 2 | P2 | × | P10[1] | 43 | - | - |
| 3 | P2 | × | P48[1] | 30 | - | - |
| 4 | P102 | × | P133[1] | 25 | 32 | 18 |
| 6 | P102 | × | P48[1] | 15 | 21 | 8 |
| 7 | P169 | × | P133[1] | 17 | 13 | 7 |
| 8 | P169 | × | P10[1] | 20 | - | - |
| 10 | P2 | × | P144[2] | 48 | - | - |
| 11 | P2 | × | P175[3] | 48 | - | - |
| 12 | P2 | × | P84[2] | 11 | - | - |
| 15 | P102 | × | P84[2] | 7 | - | - |
| Sum | | | | 313 | 116 | 78 |

[1]Many albinos, no green plants; [2]Many embryos, no plants; [3]No embryos, no plants.

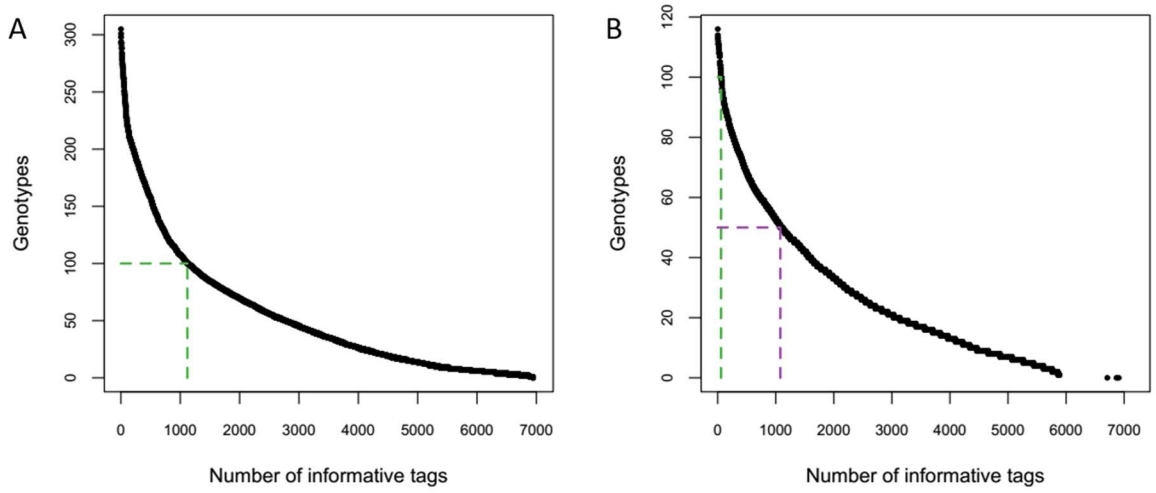



**Figure S1.** Numbers of informative tags per number of genotypes in 2015 (**A**) and 2016 (**B**) with a minor allele frequency of 10%. In this study, a 100 genotype threshold was used for the 2015 data, resulting in 1120 informative tags (green line in graph **A**) and a 50 genotype threshold resulting in 1079 informative tags (purple line in graph **B**) was used for the 2016 data.

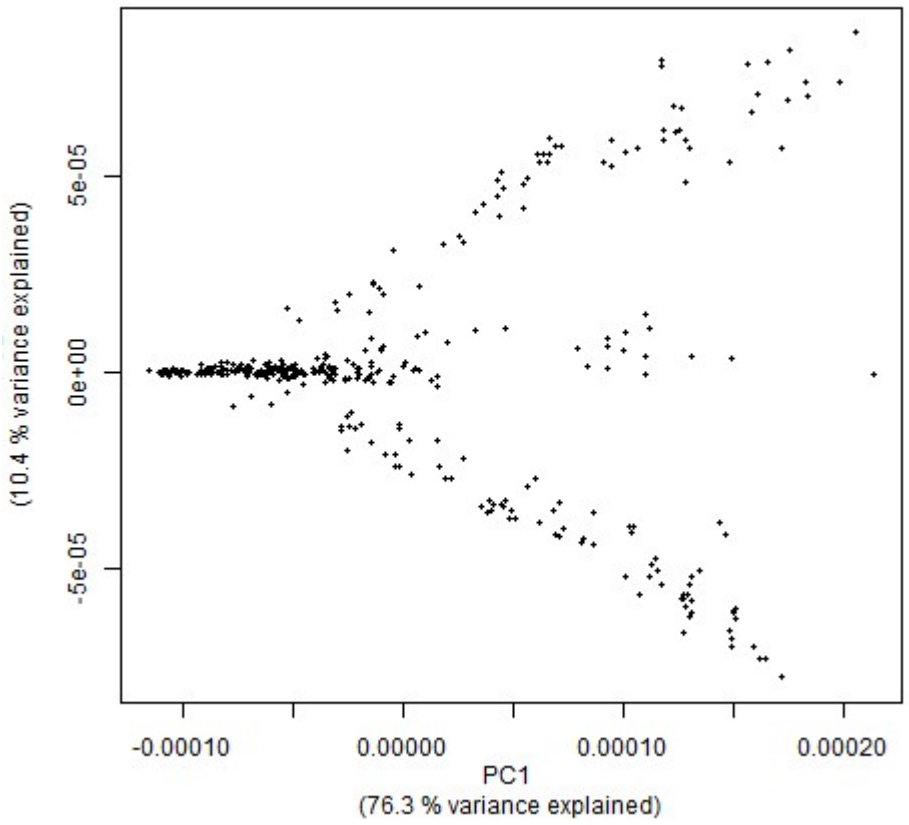



**Figure S2.** The two principal components explaining the greatest variation from a principal component analysis of the genotypic information.

**Table S2.** Gene annotations for each scaffold found to be significantly associated with the response to anther culture.

| Scaffold | Scaffold length (kb) | Start (bp) | End (bp) | Arabidopsis locus | Name(s) or description |
|---|---|---|---|---|---|
| | | | 73719 | AT5G52450 | |
| 60 | 277.8 | 71000 | | | MATE efflux family protein |
| | | 36155 | 38008 | AT5G52450 | MATE efflux family protein |
| | | 41600 | 42255 | AT5G52450 | MATE efflux family protein |
| | | 114910 | 119696 | AT2G15240 | UNC-50 family protein |
| | | 22922 | 26802 | AT3G02410 | ICME-LIKE2, ISOPRENYLCYSTEINE METHYLESTERASE-LIKE 2 |
| 123 | 256.4 | 22922 | 26802 | AT5G15860 | ATPCME, ICME, ISOPRENYLCYSTEINE METHYLESTERASE, PCME, PRENYLCYSTEINE METHYLESTERASE |
| | | 27463 | 31921 | AT3G18040 | MAP KINASE 9, MPK9 |
| | | 197961 | 200766 | AT5G07990 | CYP75B1, CYTOCHROME P450 75B1, D501, TRANSPARENT TESTA 7, TT7 |
| | | 204751 | 213979 | AT5G52450 | MATE efflux family protein |
| | | 38352 | 41786 | AT1G19330 | Histone deacetylase complex subunit |
| 127 | 250.4 | 227351 | 228127 | AT1G29860 | ATWRKY71, EXB1, EXCESSIVE BRANCHES1, WRKY DNA-BINDING PROTEIN 71, WRKY71 |
| | | 13284 | 39350 | AT5G15070 | ARABIDOPSIS HOMOLOG OF YEAST VIP1 2, ATVIP2, VIH1, VIP HOMOLOG 1, VIP1 HOMOLOG 2, VIP2 |
| 233 | 212.6 | 123380 | 130183 | AT5G62670 | AHA11, H(+)-ATPASE 11, HA11 |
| | | 131144 | 134955 | AT1G27150 | Tetratricopeptide repeat (TPR)-like superfamily protein |
| | | 199043 | 204791 | AT3G05545 | RING/U-box superfamily protein |
| | | 34360 | 44241 | AT1G33290 | P-loop containing nucleoside triphosphate hydrolases superfamily protein |
| 813 | 154 | 98603 | 105466 | AT5G14420 | RGLG2, RING DOMAIN LIGASE2 |
| | | 149380 | 153333 | AT1G47830 | SNARE-like superfamily protein |
| | | | 22714 | AT2G20330 | |
| 815 | 148.4 | 75824 | | | Transducin/WD40 repeat-like superfamily protein |
| | | 233726 | 323944 | AT4G21470 | ATFMN/FHY, FMN/FHY, RIBOFLAVIN KINASE/FMN HYDROLASE |
| | | 581499 | 623911 | AT2G01120 | ATORC4, ORC4, ORIGIN RECOGNITION COMPLEX SUBUNIT 4 |
| | | 139557 | 146531 | AT5G43920 | Transducin/WD40 repeat-like superfamily protein |
| | | 28155 | 33804 | AT5G57140 | |
| 1607 | 115.8 | | | | ATPAP28, PAP28, PURPLE ACID PHOSPHATASE 28 |

| Scaffold | Scaffold length (kb) | Start (bp) | End (bp) | Arabidopsis locus | Name(s) or description |
|---|---|---|---|---|---|
| | | 44535 | 50663 | AT2G27100 | SE, SERRATE |
| | | 89740 | 92906 | AT3G07040 | RESISTANCE TO PSEUDOMONAS SYRINGAE 3, RPM1, RPS3 |
| | | 100317 | 104480 | AT3G07040 | RESISTANCE TO PSEUDOMONAS SYRINGAE 3, RPM1, RPS3 |
| 1669 | 116.6 | 36273 | 49336 | AT3G63520 | ATCCD1, ATNCED1, CAROTENOID CLEAVAGE DIOXYGENASE 1, CCD1, NCED1 |
| 2075 | 106.6 | 6177 | 6430 | AT1G70680 | Caleosin-related family protein |
| | | 7022 | 9674 | AT1G70670 | ARABIDOPSIS THALIANA CALEOSIN 4, ATCLO4, CALEOSIN 4, CLO4, PEROXYGENASE 4, PXG4 |
| 2554 | 103 | 2585 | 8463 | AT2G45260 | Myosin-4 protein (DUF641) |
| | | 46779 | 53204 | AT5G35210 | PHD TYPE TRANSCRIPTION FACTOR WITH TRANSMEMBRANE DOMAINS, PTM |
| 3194 | 88.8 | 21212 | 27087 | AT3G46820 | TOPP5, TYPE ONE SERINE/THREONINE PROTEIN PHOSPHATASE 5 |
| | | 36911 | 38396 | AT5G64700 | UMAMIT21, USUALLY MULTIPLE ACIDS MOVE IN AND OUT TRANSPORTERS 21 |
| | | 43210 | 51417 | AT2G47580 | SPLICEOSOMAL PROTEIN U1A, U1A |
| | | 51670 | 56416 | AT3G52860 | MED28, MEDIATOR28 |
| 3723 | 81.3 | 37204 | 45974 | AT1G63810 | Nucleolar protein |
| | | 80302 | 80933 | AT4G23160 | CRK8, CYSTEINE-RICH RLK (RECEPTOR-LIKE PROTEIN KINASE) 8 |
| 4385 | 80.8 | 20744 | 21199 | AT5G26880 | AGAMOUS-LIKE 26, AGL26 |
| | | 35922 | 58157 | AT3G23640 | HETEROGLYCAN GLUCOSIDASE 1, HGL1 |
| | | 67663 | 70662 | AT4G15240 | Glycosyltransferase (DUF604) |
| | | 71970 | 75561 | AT3G20740 | FERTILIZATION-INDEPENDENT ENDOSPERM 1, FIE, FIE1, FIS3 |
| 6186 | 59.2 | 35206 | 36492 | AT1G70090 | GALACTURONOSYLTRANSFERASE-LIKE 9, GATL9, GLUCOSYL TRANSFERASE FAMILY 8, LGT8 |
| | | 47809 | 54233 | AT1G12000 | Phosphofructokinase family protein |
| 6436 | 57 | 45334 | 47951 | AT4G13330 | S-adenosyl-L-methionine-dependent methyltransferases superfamily protein |
| | | 48789 | 50285 | AT5G20190 | Tetratricopeptide repeat (TPR)-like superfamily protein |
| | | 51151 | 57039 | AT5G43960 | Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain-containing protein |

| Scaffold | Scaffold length (kb) | Start (bp) | End (bp) | Arabidopsis locus | Name(s) or description |
|---|---|---|---|---|---|
| 7045 | 53.2 | 1236 | 5326 | AT4G38180 | FAR1-RELATED SEQUENCE 5, FRS5 |
| | | 61229 | 14989 | AT2G23380 | CLF, CURLY LEAF, ICU1, INCURVATA 1, SDG1, SET1, SETDOMAIN 1, SETDOMAIN GROUP 1 |
| | | 39822 | 46005 | AT5G52010 | C2H2-like zinc finger protein |
| | | 39823 | 45693 | AT1G28600 | GDSL-motif esterase/acyltransferase/lipase |
| 8920 | 44.9 | 2729 | 10974 | AT1G75850 | LAZ4, VPS35 HOMOLOG B, VPS35B |
| 15142 | 24.8 | 16759 | 22240 | AT5G08560 | ATWDR26, WD-40 REPEAT 26, WDR26 |
| 16597 | 14.2 | 7880 | 8705 | AT5G54960 | PDC2, PYRUVATE DECARBOXYLASE-2 |

# Appendix II

# Genome editing: scientific opportunities, public interests and policy options in the European Union

**European Academies: Science Advisory Councils report on Genome Editing**

Volker ter Meulen (Chair, Germany), Austin Burt (UK), Baerbel Friedrich (Germany), Goran Hermeren (Sweden), Wlodzimierz Krzyzosiak (Poland), Cecilia Leao (Portugal), Joseph Martial (Belgium), Bert Rima (Ireland), Radislav Sedlacek (Czech Republic), Bruno Studer (Switzerland), **Timothy Sykes** (Switzerland), Miikka Vikkula (Belgium), Kirmo Wartiovaara (Finland), Anna Wedell (Sweden), Detlef Weigel (Germany) and Robin Fears (Secretariat, UK).

(What follows here is an excerpt of the full 43 page report with only the summary, introduction and plant breeding sections shown. The full report can be found at http://www.easac.eu/fileadmin/PDF_s/reports_statements/Genome_Editing/EASAC_Report_31_on_Genome_Editing.pdf)

# Summary


Genome editing, the deliberate alteration of a selected DNA sequence in a cell, using site-specific DNA nuclease enzymes, has become a very important tool in basic research. Genome editing has been described by some as a transformative technology and, certainly, in some areas of research and innovation, it is transforming expectations and ambitions. Genome editing can specifically modify individual nucleotides in the genome of living cells and, together with a growing ability to monitor and reduce off-target effects, it brings new opportunities within range. Because of its general applicability (in microbes, and plant, animal and human cells) it has a very wide range of potential uses in tackling societal objectives. These potential applications include, but are not limited to, gene- and cell-based therapies to control diseases and, in reproduction, approaches to avoid the inheritance of disease traits; the control of vector-borne diseases; improved crop and livestock breeding, including improved animal welfare; modification of animal donors for xenotransplantation; and industrial microbial biotechnology to generate biofuels, pharmaceuticals and other high-value chemicals.

The advent of genome editing has evoked enthusiasm but also controversy. Concerns have been expressed, by some non-governmental organisations (NGOs) for example, that genome editing is 'not natural', that there are too many gaps in our knowledge, that impacts are uncertain and may be inequitable, and that regulation cannot keep pace with the speed of technological innovation.

In this report, EASAC takes a broad perspective on the research advances in editing methods and their applications, policy implications and priorities for EU strategy in promoting innovation and managing regulation. Our report draws on previous work by individual academies in Europe and by other international academy collaborations. Our objectives are to raise awareness of the scientific opportunities and public interest issues: to assess what needs to be done to realise those opportunities and take account of societal concerns.

Current knowledge gaps and uncertainties emphasise the need for more basic research. We expect that research advances will fill many of the current knowledge gaps and that progressive refinement of genome editing tools will further increase their efficiency and specificity, thereby reducing off-target effects. We anticipate that the fast pace of change in research and innovation will continue, so EASAC is willing to return to the subject of this report in due course to review its assessments.

EASAC concludes that policy considerations should focus on the applications in prospect rather than the genome editing procedure itself as an emerging technology. It is important to ensure that regulation of applications is evidence-based, takes into account likely benefits as well as hypothetical risks, and is proportionate and sufficiently flexible to cope with future advances in the science. Our recommendations are as follows.

**Plants**

The increasing precision now possible in plant breeding represents a big change from conventional breeding approaches relying on random, uncontrolled chemical- or radiation-induced mutagenesis and meiotic recombination. In supporting the conclusions from previous EASAC work on new plant breeding techniques, we recommend the following.

- We ask that EU regulators confirm that the products of genome editing, when they do not contain DNA from an unrelated organism, do not fall within the scope of legislation on genetically modified organisms (GMOs).

- We advise that there should be full transparency in disclosing the process used, but that the aim in the EU should be to regulate the specific agricultural trait/product rather than the technology by which it is produced. It follows that new technologies would be excluded from regulation if the genetic changes they produce are similar to, or indistinguishable from, the product of conventional breeding and if no novel, product-based risk is identified.

**Animals**

Research on animals is already subject to stringent regulation. While most genome-edited animals are currently being generated for basic or biomedical research, the technology also provides opportunities for livestock and aquaculture. It should be appreciated that, in addition to potential increases in production, genome editing brings possibilities to enhance animal health and welfare. For specific applications, we recommend the following:

- Livestock breeding in agriculture should also be governed by the same principle as proposed for plant breeding—to regulate the trait rather than the technology and be open and explicit about what is being done.

- With regard to the modification of large animals to serve as a source for xenotransplantation, we urge EU regulators to prepare for the new opportunities coming into range: this may require further discussion of the mechanism for approving medical products relating to cells and tissues, together with assessment of the implications of whether the edited donor, in the absence of additional transgenes, is regarded as a GMO or not.

**Gene drive to modify populations in the wild**

Gene drive applications for vector control and other modifications of target populations in the wild offer significant potential opportunities to help address major public health and conservation challenges. As outlined recently by the US National Academies of Sciences, Engineering, and Medicine, a phased approach to research can enable responsible development and offers sufficient time for considering what amendments are needed to current regulatory frameworks to enable the sound evaluation of a gene-drive-based technology. EASAC supports the recommendations by the US National Academies on gene drive approaches:

- It is essential to continue the commitment to phased research to assess the efficacy and safety of gene drives before it can be decided whether they will be suitable for use.

- This research must include robust risk assessment and public engagement.

- EU researchers must continue to engage with researchers and stakeholders in the countries where gene drive systems are most likely to be applied.

**Micro-organisms**

- We conclude that genome editing in microbes does not raise new issues for regulatory frameworks and is currently subject to the established rules for contained use and deliberate release of GMOs.

- There is a wide range of potential applications, including pharmaceuticals and other high-value chemicals, biofuels, biosensors, bioremediation and the food chain. It is important to recognise this wide range when developing EU strategy for innovation in the bioeconomy.

- Many of the policy issues for microbial genome editing research and innovation fall within the scope of what is regarded as synthetic biology, and we reaffirm the general recommendations

- from previous EASAC work relating to building research capacity, promoting skills development and recognising the need to achieve a balance between protection of innovation and benefit-sharing.

- Concerns have been raised elsewhere about the possibility for genome editing research to be conducted outside regulated laboratory settings. We recommend that the Global Young Academy should assess the issues raised by the expansion of the Do-It-Yourself (DIY) biology community.

- Concerns have also been expressed elsewhere about the potential biosecurity implications of genome editing. We recommend that the scientific community continues to inform and advise policy-makers during review of the Biological and Toxin Weapon Convention.

**Human-cell genome editing**

EASAC endorses the emerging conclusions from other collective academy work (International Summit on Gene Editing and FEAM) and the initiatives of EASAC member academies:

- *Basic and clinical research.* Intensive research is needed and should proceed subject to appropriate legal and ethical rules and standardised practices. If, in the process of research, early human embryos or germline cells undergo genome editing, the modified cells should not be used to establish a pregnancy. EASAC recognises that the decision by the European Commission not to fund research on embryos will be unlikely to change at present.

- *Clinical use: somatic gene editing.* There is need to understand the risks such as inaccurate editing and the potential benefit of each proposed genome modification. These applications can and should be rigorously evaluated within existing and evolving regulatory frameworks for gene and cell therapy by the European Medicines Agency and national agencies.

- *Clinical use: germline interventions.* These applications pose many important issues including the risks of inaccurate or incomplete editing, the difficulty of predicting harmful effects, the obligation to consider both the individual and future generations who will carry the genetic alterations, and the possibility that biological enhancements beyond prevention and treatment of disease could exacerbate social inequities or be used coercively. It would be irresponsible to proceed unless and until

the relevant ethical, safety and efficacy issues have been resolved and there is broad societal consensus.

**General recommendations for cross-cutting issues**

- *Public engagement.* There has to be trust between scientists and the public and, to build trust, there has to be public engagement. Stakeholders, including patients, clinicians, farmers, consumers and NGOs, need to be involved in discussions about risk and benefit, and scientists need to articulate the objectives for their research, potential benefits and risk management practices adopted. There is need for additional social sciences and humanities research to improve public engagement strategies.

- *Enhancing global justice.* There may be risk of increasing inequity and tension between those who have access to the benefits of genome editing applications and those who do not, although the widespread adoption of the technique might facilitate the sharing of benefits. The scientific community must work with others on the determinants to narrow the societal gap: for example, by active knowledge transfer, collaboration between researchers worldwide, open access to tools and education, and education efforts. It is also vital for EU policy-makers to appreciate the consequences, sometimes inadvertent, of EU policy decisions on those outside the EU. There is evidence that previous decisions in the EU (for example, on GMOs) have created difficulties for scientists, farmers and politicians in developing countries. Reforming current regulatory frameworks in the EU and creating the necessary coherence between EU domestic objectives and a development agenda on the basis of partnership and innovation are important for developing countries as well as for Europe.

# Introduction

Genome editing is the alteration of a targeted DNA sequence, achieved by cutting the DNA molecule at a selected point, which activates the cell's own repair system and thus results in small deletions or insertions. This is commonly used to inactivate a target gene or target sequence. When, at the same time, exogenous DNA is introduced, this can support the repair at the target site and enable a predetermined exchange of single or multiple nucleotides (targeted mutagenesis), for example to replicate or rectify a naturally occurring mutation. In this eventuality, the genome-edited organism would be indistinguishable in this specific place of the genome from an organism in which the mutation occurred naturally. The same method can also be used to insert or exchange fragments of foreign DNA at a predetermined site in the genome, generally then resulting in an organism carrying a transgene.

In this report, EASAC takes a broad perspective on the research advances, applications, policy implications and priorities for EU strategy in promoting innovation and managing regulation. The issues reviewed in our report are relevant for policy-makers at the EU level as well as in Member States: we emphasise the importance of developing consistency and coherence in the principles underpinning policy across the EU, with compatibility between different sectors, in support of research and its translation to innovation.

**What are the prospects for genome editing?**

Genome editing to produce selected disruption, correction or integration of genetic material in a cell has significant potential in basic research – including the elucidation of currently poorly understood biological functions of genetic elements – and in wide-ranging fields of application. Genome editing differs from previously employed techniques of genetic engineering in that alterations can be introduced more efficiently and precisely at the molecular level. However, there is more to be done in many cases to understand the biological consequences of those nucleotide changes. Genome editing is a significant scientific advance which, at the same time, may accentuate ethical and social questions associated with some potential applications coming within reach.

The science is advancing rapidly but the technology is already sufficiently mature to warrant assessment of the opportunities and of the challenges for ensuring proportionate, robust and flexible management of research and innovation. There are relevant matters for several EU policy-making departments, relating to the regulation of new products and the avoidance of harm, whether harm is caused inadvertently to human health and the environment, or by intended misuse, with biosecurity consequences.

There are significant strengths in European research in genome editing and it is important that rigorous risk–benefit assessment is part of the regulatory process, that any safety concerns are addressed and that research outputs can be translated into new products and services to fulfil societal needs, underpin the EU bioeconomy and support European competitiveness. Potential benefits include the following: microbial biotechnology, for example in the provision of more efficient pathways for biofuel synthesis, high-value chemicals and pharmaceuticals; new vehicles for drug delivery; sensors and environmental remediation; plant and animal breeding in precision agriculture to tackle issues of food and nutrition security, animal health and a more sustainable agriculture; and a range of other human health applications [1-3]. Tackling disease, genome editing of human cells brings opportunities to treat or avoid monogenic disorders (with recent research in cystic fibrosis, Duchenne muscular dystrophy, diseases affecting the immune system and haemophilia [4] and infectious disease (with first studies in human immunodeficiency virus (HIV)) and diseases that have both a genetic and an environmental component [5]. Examples of prospective benefit and of perceived risks will be discussed later in this report.

**Definition and experimental procedures**

Genome editing refers to DNA mutations that are targeted to a specific region of the genome by site-specific nucleases (SSNs). It does not exclude the possibility that mutations in other regions of the

genome also occur during the genome editing process: to avoid these unintended consequences, tools are being sharpened to prevent off-target effects.

Two forms of mutagenesis need to be distinguished:

- Simple mutagenesis (non-homologous end-joining), resulting either in base-pair substitutions or small insertions or deletions. This form is indistinguishable from spontaneous or induced random mutagenesis.

- Homologous recombination, in which a template of DNA is supplied with the SSN enabling the replacement of a similar sequence in the genome, or insertion of the added DNA in the genome at a pre-specified place. This form is similar to transfer of genetic material from one species to another after conventional crosses, or in cases of a more distantly related donor of the template DNA, similar to naturally occurring lateral/horizontal gene transfer.

A separate consideration is whether genome editing is achieved by insertion of DNA sequences that code for the editing agent (for example, CRISPR–Cas9) into the genome (and later removed by genetic segregation) or whether the editing agent is introduced transiently as DNA, RNA and/or protein without any integration of foreign DNA sequences into the cell.

Further scientific detail about the recent history of genome editing is provided in Box 1.

**Public interests and values**

The outputs from genome editing may have direct or indirect impacts on the well-being and welfare of the public—and the advent of genome editing evokes not only enthusiasm but also controversy. As will be discussed later in this report, when public concerns are elicited, they are usually about the intended use rather than the technology itself. Various queries have been raised about the different applications of genome editing, reflecting field-specific drivers and obstacles, but there are also generic questions that can be asked, as observed in the consultation for the UK Nuffield Council on Bioethics inquiry on genome editing (2015). For example, to what extent can the development of new genome engineering techniques be regarded as distinct from, or continuous with, existing techniques? Does the ease and accuracy of genome editing mean that it is a transformative technology (in either the moral or economic senses) and, therefore, represents a 'tipping point' in the potential of genetic engineering? Should a distinction be made (as it is by some who query these techniques) between directed change and those undirected changes induced, for example, by chemical- or radiation-induced mutagenesis, in conventional plant breeding programmes? There is also a generic technical point that is relevant to the various fields of application. Editing makes only small changes to DNA. At the target site these are easily identified, but off-target changes, which also occur in random mutagenesis, may be difficult to detect without full DNA sequencing. What implications does this have for the regulation of the resulting product?

Potential problems for assessing the products of this emerging technology are compounded in the EU by a legacy of contention and polarisation about the regulation of genetic engineering techniques. Current EU legislative frameworks governing the genetic modification of plants and animals, for example, are controversial; and even when there is an overarching EU policy framework, there is little certainty for researchers and breeders, because individual Member States vary in their implementation or can exercise an 'opt-out'. As critically observed by a recent Member State parliamentary report (UK House of Commons Science and Technology Committee, 2016), '*The regulation of genetic science is an area in which the EU has so far not come close to satisfactorily demonstrating an evidence-based approach to policy making*'.

Responsible innovation requires attending to ethical, legal and societal issues, and seeking to identify common goals important to scientists and the public. Researchers and their funders have a responsibility to engage with the public and to take account of public interests and values. In genome editing these range from the protection of individuals or populations from possible health risks, protection of animals from risks to their health and welfare, to moral and political interests around the acceptable limits to intervening in natural processes [6].

There is a moral obligation to fight disease and relieve humans and animals from suffering. To the extent that genome editing technologies provide useful tools to achieve such purposes, there is an opportunity cost in using them too late or not at all, particularly if they are safer, more effective and cheaper than alternative technologies. Concerns have been expressed about whether regulation can keep pace with the speed of technological innovation, whether scientists (and society) have fully appreciated the implications of what science can deliver and whether it would be possible to reverse undesirable outcomes. Much of the public debate has focused on human germline modification (which means that genetic changes would be heritable), but ethical issues relating to views of nature and ecosystems are also relevant to applications encompassing non-human targets of genome editing [7].

Application-specific issues are discussed in our subsequent chapters. General concerns expressed,

**Box 1 Summary of the science of programmable nucleases**

Genome editing methods take advantage of exogenous programmable nucleases to make double-stranded DNA breaks at selected sites. These breaks activate endogenous repair mechanisms either non-homologous end-joining (NHEJ) or homology-directed repair (HDR). The latter operates when a DNA donor template is provided, and both systems function in all eukaryotic organisms. NHEJ is a more prevalent, error-prone mechanism that often causes mutations (short insertions or deletions), resulting in target gene knockout, when the break is introduced in the coding sequence of a locus; whereas HDR, which functions only in the synthesis (S) and gap 2 (G2) phases of the cell cycle, is the way to knock-in or substitute a desired sequence, for example to replace a mutant DNA fragment for the normal one. The NHEJ efficiency at the site of induced double-stranded DNA break is usually about five- to eight-fold higher than the efficiency of HDR.

The first generation of gene editing tools was based on oligonucleotide-directed mutagenesis (ODM) or microbial meganucleases, possessing long DNA recognition sequences. They were cumbersome to use and often suffered from low efficiency, especially ODM. The desired flexibility in target sequence recognition was achieved with the use of engineered zinc finger nucleases (ZFNs: each finger recognises about three specific nucleotides of DNA) and more recently with transcription activator-like effector nucleases (TALENs: each TALEN recognises short double-stranded specific sequence, typically single nucleotides). In both ZFN [8] and TALEN [9] designs, the DNA recognition module is additionally coupled via a peptide linker to an unspecific DNA cleaving portion, usually the Fok I restriction nuclease domain. As only dimerised Fok I shows DNA cleavage activity, the length of the DNA recognising portion is also doubled by involving two recognition arms, enhancing nuclease specificity. Although TALENs had several advantages over ZFNs, especially in their design, their production is still a laborious process.

Another class of genome editing tool is designer recombinases. Similar to meganucleases, recombinases are difficult to tailor and the generation of enzymes with new DNA-binding specificities is cumbersome and time consuming. However, designer recombinases are highly specific and do not rely on cellular DNA repair as they cut and re-ligate the DNA in a conservative manner. As such, designer recombinases represent interesting alternatives [10], subject to further research.

The revolution in the field of genome editing came in 2012 with the development of the CRISPR–Cas9 system [11], which is much easier to design, produce and use. The acronym CRISPR stands for clustered regularly interspersed short palindromic repeats, and it is considered by some to be a distant bacterial analogue of the RNA interference mechanism in eukaryotes; Cas stands for CRISPR-associated protein nuclease. The system is based on the natural defence mechanism against bacteriophages and plasmids evolved by many bacteria and archaea. Unlike protein meganucleases, ZFNs and TALENs, the new system uses RNA for complementary DNA recognition, and Cas9 protein (or related protein) to recognise a matching target sequence in the DNA, flanked by a short protospacer adjacent motif (PAM), and execute DNA cleavage by its two DNase domains. The RNA component is either composed of two molecules, the CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA) as in the bacteria it derives from, or, what is more common, these two RNAs are fused by researchers into a single guide RNA (gRNA) which is about 100 nucleotides long.

How does the CRISPR–Cas9 system function? In brief, the Cas9 protein is bound to a gRNA and thereby programmed to recognise a target DNA whose sequence is complementary to a ~20 nucleotide segment in the gRNA. Cas9 binds the PAM motif in the target DNA duplex, separates the DNA strands and facilitates base-pairing between the gRNA and the complementary DNA sequence. Subsequently, Cas9 deploys its two DNase domains, RuvC and HNH, to cleave target DNA, generating a double-stranded break. Then, the DNA repair systems, NHEJ or HDR come into action and DNA is either mutated or replaced. The editing process with CRISPR–Cas9 may be multiplexed to inactivate tens of targets at once [12].

The important practical issues in genome editing experiments are the delivery of programmable nucleases into cells, their cleavage efficiency and specificity, in terms of avoiding off-target effects. To minimise the off-target effects, new versions of Cas9 and related proteins have been engineered. Recently, a mutation of three or four amino acids in the Cas9 catalytic domain reduced off-target effects dramatically to levels that were hardly noticeable [13]. Furthermore, in addition to Cas9, other bacterial DNases such as Cpf1 [14], which recognise different PAM sequences, can also be used for genome editing and thus increase the range of targetable sequences in genomes.

Besides genome editing, the CRISPR–Cas9 system has been repurposed for sequence-specific regulation of gene expression, either transcription activation or repression, or specific gene imaging using nuclease-deactivated Cas9 termed dCas9 [15]. The CRISPR–Cas9 system has also been adapted to recognise and track RNA in living cells [16], and a natural RNA-targeting CRISPR system taking advantage of the C2c2 enzyme has been identified [17].

for example by some NGOs, that genome editing is not natural, and that there are too many gaps in our knowledge and that impacts are uncertain, as well as there being issues for global justice, can probably be applied to all emerging technologies in biology and medicine. It is the role of research and of robust regulatory systems to continue to address the uncertainties and fill the knowledge gaps in a transparent way. A cardinal feature of the accuracy of genome editing is that the functional consequences should be more predictable than when using earlier techniques. Of course, there is continuing need to adopt appropriate safety standards, develop risk assessment techniques and to install effective surveillance, monitoring and disclosure systems, whatever the field of application. The recent report from the Nuffield Council on Bioethics (2016) considers further the range of ethical questions to which the recent advances in

genome editing may give rise. These issues and the implications of the 'slippery slope' argument will be dealt with at various places in our report.

Public interest about science and innovation also often refers to the desirability of open science, benefit-sharing and fair competition. There is controversy about competing patent claims for CRISPR–Cas technology [18, 19]. At the same time, CRISPR–Cas9 has become an example of open science, where the development of the procedures has resulted in the sharing of tools from more than 80 laboratories. Patent-related aspects were addressed in a recent statement from ALLEA, the All European Academies (2016) which notes that the use of CRISPR–Cas technology does not require any reforms in patent law: '*EU patent law provides the necessary incentives for further development and use across all fields of life sciences*' and that there will be no patents granted which could offend human dignity and/or integrity.

**Previous work by academies of science and medicine**

There has already been a significant amount of work by academies on the issues elicited by genome editing and our EASAC report draws on this continuing effort:

- *At the national level in Europe,* the German Academies statement [20] on opportunities and limits, covers all applications and emphasises the great scientific potential of genome editing in opening up new scope for basic research. This German statement concludes that it is ethically and legally acceptable in many areas (see Chapter 5 of the present report for further discussion, including a moratorium of genome editing for germline interventions) and that new techniques should not automatically be equated with sporadic cases of improper use or with applications whose ethical and legal ramifications have not yet been assessed. While our EASAC study was in progress, KNAW, the Royal Netherlands Academy of Arts and Sciences (2016), published their national position paper on genome editing. This also covers multiple applications and their recommendations are broadly consistent with the recommendations in the present EASAC report.
- *The International Summit on Human Gene Editing* is led by the US National Academies of Sciences, Engineering and Medicine together with the UK Royal Society and the Chinese Academy of Sciences. This consortium is examining the scientific underpinning as well as the clinical, ethical, legal and social implications of the use of human genome editing technologies in biomedical research and medicine, including editing of the human germline [21].
- *The US National Academies* have also completed investigations of genome editing and gene drive [22], and of genome editing relevant to laboratory animal use.
- *FEAM* organised a workshop in 2016; with support from the InterAcademy Partnership (IAP), to consider the landscape for human genome editing in the EU. This workshop reviewed current scientific and regulatory activity in human genome editing research and clinical applications, to identify where there are significant differences between EU countries and to discuss options for European-level activities [23] The report from this workshop was recently published [24].

The outputs from these other academy activities will be cross-referenced in the following chapters of our report.

**EASAC objectives for this work**

In seeking to add value to the work that has already been done, this report draws on the previous academy publications together with advice and information from a group of experts nominated by EASAC member academies. We take a broad perspective of the science, and our objectives for this report are also wide-ranging in assessing policy and practice:

- To raise awareness across Europe of the scientific opportunities of the new genome editing techniques, and public interest issues, to evaluate what is now needed to realise those opportunities and address those issues, and to consider who should make decisions on governance.
- To identify distinctive aspects confined to particular applications of genome editing, to show where sector-specific outputs are already subject to established policies rules and regulations (at institutional, national and

EU levels) or where changes should now be foreseen.

- To prepare policy-makers to address those issues that have still to be clarified and resolved.
- To serve as an input to global discussions and action on genome editing priorities, alongside the other academy initiatives (that focus on human-cell applications) and for those aspects where global consensus is of particular importance (for example, for biosecurity).

As part of these objectives, we aim to assess what strategic objectives are relevant to the EU level and what is reserved for Member States. EASAC messages are directed to those who make or influence policy in EU institutions, and at Member State level, academies of science in other regions outside the EU, research funding bodies, regulatory authorities, professional societies and others in the scientific community. We recognise the great importance of also engaging with other stakeholders and the community-at-large, and EASAC encourages its member academies to use this report as a resource to disseminate our messages widely.

In the following chapters, we consider particular applications of genome editing and in the final chapter bring together our conclusions and recommendations.

# Plants

For both plants and animals, genome editing has become an essential tool for basic research, to elucidate gene function and to generate model plants and animals. The scientific advances achieved with genome editing, capitalising also on the progress in genome sequencing that is identifying many genes and alleles of interest for agriculture, enhance the potential for tackling a wide range of applications.

There are major global challenges to be faced in addressing issues for food and nutrition security and agriculture, and the opportunities and challenges are discussed more broadly in an ongoing EASAC project that constitutes the European arm of a worldwide IAP project. Current problems of food and nutrition security are compounded by pressures of growing population, climate and other environmental changes, and by economic inequity and insecurity. Setting priorities for increasing agricultural production must also take account of pressures on other critical resources, particularly water, soil and energy, and the continuing imperative to avoid further loss of ecosystems and biodiversity.

**Plant breeding in agriculture**

Plant sciences can do much in continuing to contribute to increased crop quality, for example in developing cultivars with improved water and nitrogen use, better resistance to pests and diseases, or modified crop architecture to reduce waste. Prospects for plant genome editing are discussed widely in the literature [25, 26] and in the recent report from the US National Academies (2016c) [27] which notes the potential of genome editing to introduce more complex changes because multiple genes can be edited simultaneously. Genome editing brings new possibilities to improve plant traits, beyond what has been achieved with the previous generation of genetic modification (mutagenesis) approaches. Molecular targets are being selected and tackled to increase yield, stress- and disease-resistance, elevate nutrient use efficiency and reduce allergens, for example, in broad support of the societal objectives for increased food production, conservation of natural resources, less pollution and healthier food. There are many significant research advances described in the US National Academies report and in other recent publications, for example the induction of targeted heritable mutations in barley and brassica [28] and combatting invading virus DNA in plants [29]. Of particular interest in breeding is the rapid introduction of known natural alleles (genetic variation) into many different genetic backgrounds.

Research advances in plant breeding are now being translated into novel products. There has been recent progress using genome editing in the commercial development of cold-storable potatoes and no-trans-fat soybean oil, but the first organisms to be allowed by the US Government are CRISPR–Cas9-edited mushrooms (with reduced browning by reducing the activity of the endogenous enzyme polyphenol oxidase) and a waxy corn engineered to contain starch composed exclusively of the branched polysaccharide amylopectin (used in processed foods, adhesives and high-gloss paper). These products do not come within US Department of Agriculture regulations [30] although they might still be submitted for voluntary review by the US Food and Drug Administration (FDA).

These rapid advances in research and development accentuate a major underlying question for the EU: to what extent will the regulation of plants/food products developed using genome editing be influenced by previous controversies and current legislation on GMOs? The products of genome editing may contain no foreign DNA, and EASAC has previously advised in the Statement on New Breeding Techniques [31], encompassing genome editing tools and summarised in

**Box 2 Summary of previous EASAC recommendations on new plant breeding techniques**

EU policy development for agricultural innovation should be transparent, proportionate and fully informed by the advancing scientific evidence and experience worldwide.

It is timely to resolve current legislative uncertainties. We ask that EU regulators confirm that the products of new breeding techniques, when they do not contain foreign DNA, do not fall within the scope of GMO legislation.

The aim in the EU should be to regulate the specific agricultural trait and/or product, not the technology by which it was produced.

The European Commission and Member States should do more to support fundamental research in plant sciences and protect the testing in field trials of novel crop variants against vandalism.

Modernising EU regulatory frameworks would help to address the implications of current policy disconnects in support of science and innovation at regional and global levels. At the same time, there is continuing need for wide-ranging engagement on critical issues and this should include re-examination of the appropriate use of the precautionary principle.

Source: EASAC (2015a) [31]

Box 2, that such processes should not be regulated in the same way as GMOs, assuming that there is evidence to demonstrate that any transgene has been segregated away in the final product.

The issues are, however, still contentious. For example, if there is a transient transgenic stage during the plant breeding process, some would assert that this makes the final non-transgenic product still a GMO. However, modern whole-genome sequencing methods allow for unambiguous proof that foreign DNA from transgenes has been completely removed. It should also be noted that many of the agricultural sector-specific public concerns raised by NGOs about genome editing were also raised previously in the early days of genetically modified (GM) crops and were addressed systematically then (for example in the UK GM science review [32], and see EASAC (2013) [33] for further discussion of the GM crop research evidence base).

A European Commission decision on the status of these products is urgent in view of the accelerating pace of research and development and of the regulatory initiatives being undertaken by individual Member States. For example, an oligonucleotide gene-edited canola strain was assessed as non-GMO in Germany [31, 34]. The Swedish Board of Agriculture, a national competent authority, also confirmed that some plants in which the genome had been edited using CRISPR–Cas9 do not fall under the EU GMO definition. Discussion in the EASAC Working Group agreed that a strong case can be made for genome-edited crops to be subject only to the rules and regulations that apply to products of conventional breeding, subject to certain guiding principles [34]:

- Minimising the risk of escape of genome-edited crops from laboratories and fields during the research and development (R&D) phase.
- Demonstrating the absence of foreign sequences if genome engineering proteins were introduced as DNA constructs.
- Documenting DNA sequence changes at the target sites.
- In the case of newly introduced DNA, identifying the phylogenetic relationship between donor and recipient.
- Excluding unintended secondary editing events or off-target sites on the basis of available reference genome information.

Even if a trait-based assessment system did not require specific regulation of a new crop variety, there should still be a legal requirement to disclose the process used, with transparency on why a particular process was used. The alternative regulatory options for genome-edited plants, including product-based approaches, are discussed further in detail by Sprink *et al.* (2016) [35].

Recommendations from the European Commission on what is a GMO are delayed, and continuing discussion with the European Commission, European Parliament and Council of Ministers is expected. There is great need for evidence-based proportionate regulation for next-generation plant breeding (Box 2). EU regulatory frameworks should also take account of best practice outside the EU [31, 33]. For example, reform of the US system for regulation of GMOs and of products using other techniques such as genome editing, which do not currently fall within US GMO regulations, is anticipated in the new US Coordinated Framework for regulating biotechnology. It has been proposed [36] that this new US Framework should be product-based not event-based; novelty-based not method-based; and that modifications that are analogous to what occurs in conventional breeding (but which are more precise and better understood than in conventional breeding) should be exempt, unless a novel product-based risk is identified. It would seem reasonable to consider adopting similar criteria in the EU (and compatible with the recommendations in Box 2), while also taking into account essential features of the responsible governance of agricultural biotechnology [37], including a commitment to candour, recognition of underlying values and assumptions, and a preparedness to respond to new knowledge or concerns.

Recent proposals from the US Government give some indications of how the revised US regulatory system might function. The US Department of Agriculture's Animal and Plant Health Inspection Service [38] set out the criteria by which an organism would not be regarded as genetically engineered. For example, it would not be regarded as a genetically engineered organism if the modification were solely a deletion of any size or a single base-pair substitution that could otherwise be obtained through the use of chemical-or radiation-based mutagenesis. It would also not be

considered a genetically engineered organism if the modification were solely introducing only naturally occurring nucleic acid sequences from a sexually compatible relative that could otherwise cross with the recipient organism and produce viable progeny through traditional breeding (including, but not limited to, marker-assisted breeding, as well as tissue culture and protoplast, cell or embryo fusion). As part of its broader initiative in biotechnology (see subsequently for issues raised for animals and mosquitoes), the FDA has also very recently invited comments on whether genome-edited plants might present new food safety risks and whether they should follow the same pre-market regulatory review at the FDA as transgenic plants currently do. An accompanying commentary emphasises the FDA principle to maintain product-specific, risk-based regulation.

A second international example is provided by Australia, currently conducting a review and public consultation to provide clarity on whether organisms developed using a range of new technologies (including site-directed nuclease techniques) are subject to regulation as GMOs and to ensure that new technologies are regulated in a manner commensurate with the risks they pose [39]. Four options are identified in this Australian review: (1) no amendment to the current regulations; (2) regulate certain technologies (including all site-directed nuclease techniques); (3) regulate some new technologies on the basis of the process used (excluding site-directed nuclease technologies that do not involve application of a DNA template); and (4) exclude certain new technologies from regulation on the basis of the outcomes they produce: that is, exclude if the genetic changes produced are similar to or indistinguishable from the product of conventional breeding (chemical and radiation mutagenesis and natural mutations). This last option, focusing on product rather than process, would again be similar to the recommendations of EASAC for the EU (Box 2): it is important to achieve international coherence in regulation.

## References


1. Hsu, P.D., Lander, E.S. and Zhang, F., 2014. Development and applications of CRISPR-Cas9 for genome engineering. Cell, 157(6), pp.1262-1278.
2. Carroll, D. and Charo, R.A., 2015. The societal opportunities and challenges of genome editing. Genome biology, 16(1), p.242.
3. Barrangou, R. and Doudna, J.A., 2016. Applications of CRISPR technologies in research and beyond. Nature biotechnology, 34(9), pp.933-941.
4. Prakash, V., Moore, M. and Yáñez-Muñoz, R.J., 2016. Current progress in therapeutic gene editing for monogenic diseases. Molecular Therapy, 24(3), pp.465-474.
5. Porteus, M.H., 2015. Towards a new era in medicine: therapeutic genome editing. Genome biology, 16(1), p.286.
6. Nuffield Council on Bioethics (2015). Genome editing consultation. http://nuffieldbioethics.org/project/genome-editing/
7. Charo, R.A. and Greely, H.T., 2015. CRISPR critters and CRISPR cracks. The American Journal of Bioethics, 15(12), pp.11-17.
8. Kim, Y.G., Cha, J. and Chandrasegaran, S., 1996. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proceedings of the National Academy of Sciences, 93(3), pp.1156-1160.
9. Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J. and Voytas, D.F., 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic acids research, 39(12), pp.e82-e82.
10. Karpinski, J., Hauber, I., Chemnitz, J., Schäfer, C., Paszkowski-Rogacz, M., Chakraborty, D., Beschorner, N., Hofmann-Sieber, H., Lange, U.C., Grundhoff, A. and Hackmann, K., 2016. Directed evolution of a recombinase that excises the provirus of most HIV-1 primary isolates with high specificity. Nature biotechnology, 34(4), pp.401-409.
11. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E., 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science, 337(6096), pp.816-821.
12. Yang, L., Güell, M., Niu, D., George, H., Lesha, E., Grishin, D., Aach, J., Shrock, E., Xu, W., Poci, J. and Cortazio, R., 2015. Genome-wide inactivation of porcine endogenous retroviruses (PERVs). Science, 350(6264), pp.1101-1104.
13. Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z. and Joung, J.K., 2016. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. Nature, 529(7587), pp.490-495.
14. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A. and Koonin, E.V., 2015. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell, 163(3), pp.759-771.
15. Dominguez, A.A., Lim, W.A. and Qi, L.S., 2016. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nature Reviews Molecular Cell Biology, 17(1), pp.5-15.
16. Nelles, D.A., Fang, M.Y., O'Connell, M.R., Xu, J.L., Markmiller, S.J., Doudna, J.A. and Yeo, G.W., 2016. Programmable RNA tracking in live cells with CRISPR/Cas9. Cell, 165(2), pp.488-496.
17. Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L. and Severinov, K., 2016. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science, 353(6299), p.aaf5573.
18. Egelie, K.J., Graff, G.D., Strand, S.P. and Johansen, B., 2016. The emerging patent landscape of CRISPR-Cas gene editing technology. Nature biotechnology, 34(10), pp.1025-1031.
19. Nuffield Council on Bioethics (2016). Genome editing: an ethical review. http://nuffieldbioethics.org/wp-content/uploads/Genomeediting-an-ethical-review.pdf
20. Leopoldina German National Academy of Sciences, acatech, Union of German Academies, and DFG (2015). The opportunities and limits of genome editing, http://www.leopoldina.org/nc/en/publications/detailview/?publication%5Bpublication%5D=699&cHash=4d49c84a36e655feacc1be6ce7f98626
21. National Academies (2016a). International summit on human gene editing. http://nationalacademies.org/cs/groups/genesite/documents/webpage/gene_170582.pdf
22. National Academies (2016b). Gene drives on the horizon: Advancing science, navigating uncertainty, and aligning research with public values, http://www.nap.edu/23405
23. Academy of Medical Sciences (2016). The European landscape for human genome editing. www.acmedsci.ac.uk/download.php?f=file&i=34773
24. FEAM (2017). Human genome editing in the EU. Report of a workshop held on 28 April 2016 at the French Academy of Medicine. http://www.feam-site.eu/cms/docs/humangenomeeditingworkshop2016report.pdf
25. Bortesi, L., Zhu, C., Zischewski, J., Perez, L., Bassié, L., Nadi, R., Forni, G., Lade, S.B., Soto, E., Jin, X. and Medina, V., 2016. Patterns of CRISPR/Cas9 activity in plants, animals and microbes. Plant biotechnology journal.
26. Quétier, F., 2016. The CRISPR-Cas9 technology: closer to the ultimate toolkit for targeted genome editing. Plant science, 242, pp.65-76.

27. National Academies (2016c). Genetically engineered crops: experiences and prospects. On www.nap.edu/read/23395
28. Lawrenson, T., Shorinola, O., Stacey, N., Li, C., Østergaard, L., Patron, N., Uauy, C. and Harwood, W., 2015. Induction of targeted, heritable mutations in barley and Brassica oleracea using RNA-guided Cas9 nuclease. Genome biology, 16(1), p.258.
29. Zhang, D., Li, Z. and Li, J.F., 2015. Genome editing: new antiviral weapon for plants. Nature Plants, 1, p.15146.
30. Waltz, E., 2016. CRISPR-edited crops free to enter market, skip regulation. Nature, 34(10), p.582.
31. EASAC (2015a). New breeding techniques. Statement, www.easac.eu/fileadmin/PDF_s/reports_statements/Easac_14_NBT.pdf
32. GM Science Review Panel (2003). GM Science Review First and Second Reports, http://www.gmsciencedebate.org.uk/panel/default.htm
33. EASAC (2013). Planting the future: opportunities and challenges for using crop genetic improvement technologies for sustainable agriculture. Policy Report no. 21, www.easac.eu/fileadmin/Reports/Planting_the_Future/EASAC_Planting_the_Future_FULL_REPORT.pdf
34. Huang, S., Weigel, D., Beachy, R.N. and Li, J., 2016. A proposed regulatory framework for genome-edited crops. Nature genetics, 48(2), pp.109-111.
35. Sprink, T., Eriksson, D., Schiemann, J. and Hartung, F., 2016. Regulatory hurdles for genome editing: process-vs. product-based approaches in different regulatory contexts. Plant cell reports, 35(7), pp.1493-1506.
36. Strauss, S.H. and Sax, J.K., 2016. Ending event-based regulation of GMO crops. Nature biotechnology, 34(5), pp.474-477.
37. Hartley, S., Gillund, F., van Hove, L. and Wickson, F., 2016. Essential features of responsible governance of agricultural biotechnology. PLoS biology, 14(5), p.e1002453.
38. APHIS (2017). Importation, interstate movement and environmental release of certain genetically engineered organisms. Federal Register 82, 7008.
39. Australian Government Department of Health, Office of the Gene Technology Regulator (2016). Technical review of the Gene Technology Regulations 2001 Discussion paper: Options for regulating new technologies. www.ogtr.gov.au/internet/ogtr/publishing.nsf/Content/977EF3D4FDD4552ECA2580B10014663C/$File/Discussion%20Paper%20-%20Review%20of%20the%20Gene%20Technology%20Regulations.pdf

# Appendix III

# The benefits of new plant breeding techniques

**Timothy Sykes** & Bruno Studer

Plant Science News. No 27, Spring 2015.
(http://www.plantsciences.uzh.ch/dam/jcr:00000000-60f2-7d2b-0000-000031c57de9/psc_newsletter_spring_2015-27.pdf)

Plant breeding is an important process that allows agricultural production to adapt to changing environmental conditions, attacks from diseases and pests, and to meet the increased needs of a growing population. Constant improvements of traditional breeding methods as well as innovations in plant breeding are essential to meet these needs.

Through a set of techniques to either speed up the breeding cycle or increase the efficiency in the selection process, it has been possible to dramatically shorten the laborious and time intensive breeding process. While most of these techniques do not include direct modification of the genome, and have been successfully applied by plant breeders for decades, more recent molecular biology tools now allow any gene of interest to be precisely targeted. The novelty of this technique is that the DNA, which is introduced into a genome to specifically edit the target gene, can be removed again, so that the end product – a new crop variety – is indistinguishable from a commercially bred variety.

This raises questions, in Switzerland and beyond, as to where this new technique falls within the current legal regulatory framework. With the key consideration being: Is it the process to generate the final product or the final product itself that is to be regulated?

To answer this question, a professional evaluation based on scientific knowledge and societal requirements is essential to ensure a well-guided evaluation process and to maintain innovative plant breeding. For this, it is imperative that the research community adopts a more active and vocal stance to ensure that the public have access to all facts regarding these new breeding techniques. This is particularly important for Switzerland as although the value of the maintenance of old landraces and the conservation of plant genetic resources for breeding is deeply anchored in society, the importance of modern plant breeding to obtain continuous genetic improvement of crops for food security is not well recognized. In contrast, social and economic benefits are increasingly being recognized in neighbouring countries such as Germany. Indeed, it was modern plant breeding that played the key role in the continuous advancement of crop varieties over the past decades, with achievements including; yield increases, adaptation to changing climate conditions and more environmentally sustainable production systems.

So what can you do as a researcher to ensure modern plant breeding gets the recognition it deserves? Well there are many things, from writing opinion pieces for journals and newspapers to making sure that your friends and family have all the facts before making lifestyle decisions. This debate needs to be seen in the right context, especially given that current food production owes a lot to plant breeding and that these new techniques complement rather than replace traditional plant breeding methods. There are still significant gaps between what we as researchers know and what kind of information is being presented to the general public. If we want plant breeding to continue to achieve all it can - in the increasingly important

and difficult battle to provide food for the ever increasing global population - we need to ensure that innovation and implementation are not stifled before a balanced debate can take place.

# Appendix IV

# EUCARPIA - Plenary Discussion: Innovation vs. Regulation

**Timothy Sykes**

IDP Bridges News - Bridging Plant Sciences and Policy. No 5, 2016.
(http://www.plantsciences.uzh.ch/dam/jcr:3e10707a-3ffb-4e57-a248-dbb40346fef7/idp_NL_05_2016_web.pdf)

During the recent general congress of the European Association for Research on Plant Breeding (EUCARPIA) I had the pleasure of mediating a lively discussion entitled: 'Innovation vs. regulation - Facilitating access to germplasm and release of innovative cultivars.' Broadly this discussion was focussed on emerging plant breeding techniques in the context of reviewing existing regulations along the food production chain, taking into account the interests and concerns of a diverging range of stakeholders: breeders (and other holders of intellectual property rights), variety testers, seed producers and merchants, farmers, consumers and environmental agencies. In order to facilitate a meaningful, diverse and interesting discussion we invited seven panel members from different stake holder groups. The full range of topics covered and the varying opinions of the panel and audience members cannot all be covered here, so rather what follows is a summary of the key points made and who made them.

The first member of the panel, Richard Visser, the incoming president of EUCARPIA and head of the laboratory of plant breeding at Wageningen University, represented academics on the panel. Richard's main hope for the future was that plant breeders would have available to them all new innovations and techniques in order to more efficiently breed new plant varieties. Such a holistic approach to plant breeding was not the focus of Edith Lammerts van Bueren, a senior researcher at the independent Louis Bolk Institute for Organic Agriculture and professor of organic plant breeding at Wageningen University. She stated that within the organic community a different approach to risk perception has led to a unique view on health and environmental concerns, thus the organic community does not intend to use any new plant breeding techniques (NPBTs) that affect a plant on a DNA level. She did make an exception for diagnostic tools such as marker assisted selection. Edith did, however, have a holistic approach to plant breeding regulations, stating the importance that regulations leave space for alternative breeding concepts and not just dominant ideas.

A desire for a collaborative approach to plant breeding regulation was shared by Eva Reinhard, the deputy director of the Swiss Office of Agriculture (FOAG) and head of the production systems and ecosystems directorate, who outlined the FOAGs vision that only a close collaboration between science, farmers, retailers, food industry, and consumers will allow the goal of sustainable agriculture and food security to be reached. She revealed that to this end the FOAG have been working very closely with stakeholders over the last two and a half years on a unified plant breeding strategy for Switzerland. The FOAGs goal of a national plant breeding strategy was also mentioned by Stephan Scheuner, the managing director of Swiss Granum, an umbrella organization concerned with cereals, oilseeds and protein crops that combines organizations of production, collection centres, trade, and fabricators. He called for clarity within this strategy as to the handling of NPBTs, including pointing out difficulties of quality control at the seed testing level given that varieties developed using NPBTs may be indistinguishable from conventionally bred crops.

Another panel member who called for clarity regarding the regulation on NPBTs, albeit on a global level, was Michael Keller, the secretary general of the International Seed Federation (ISF). The ISF represents the interests of the seed industry at a global level and as such is involved in the development of new varieties that can involve up to seven different countries on four different continents. Michael not only highlighted the importance of global consistency with plant breeding regulations across countries, but also pointed out the need for consistency across time, as the breeding process can take many years in some crops and breeders need to know in advance what regulation will be applied.

Likewise, Peter van der Toorn, who as the head of vegetable breeding at Syngenta seeds, similarly works within an international community, also spoke about international regulation. His comments were mainly focussed on the Nagoya Protocol, an international agreement which aims at sharing the benefits arising from the utilization of genetic resources in a fair and equitable way by regulating the movement of genetic resources globally. Peter revealed that, within a commercial setting, the Nagoya Protocol is making it nearly impossible to access some genetic resources and that this was leading his pursuit of NPBTs in order to manufacture the necessary genetic diversity for their breeding programs. He also pointed out, echoing comments from other panel members, that clarity of regulation of NPBTs on a global scale was necessary, even going further to say that if regulations make using these techniques too costly it would make it very difficult for breeders to continue to innovate.

Perhaps the most interesting moment of the discussion occurred when an audience question regarding the applicability of plant variety protection (PVP) laws given the rapid variety turnover seen in plant breeding today, was put to the panel. Stephanie Frank, the CEO of the family owned breeding company Saatzucht Oberlimpurg as well as the president of the Confederation of German Plant Breeders, who is also an expert intellectual property law, spoke about how PVP is affective because not only can other breeders use your varieties but you theirs. This leads to a cycle of innovation which is beneficial to all breeders, as long as there is a diversity of breeders. She also pointed out that neither new plant varieties nor plant related technical innovations are patentable in Europe. This lead to a discussion about patenting plant varieties, where Peter suggested that if a new variety were augmented with specific genes that conferred novel traits that variety would be patentable, and Stephanie insisted that this is just a derived variety and hence covered by PVP.

These are just some of the many important points that were made during the discussion that sit at the heart of the innovation/regulation balance. In this hour-long discussion we did not solve the problem of how plant breeding should be regulated to ensure continuing innovation into the future, but we did manage to highlight the importance of involving all concerned parties, and the main areas where difficulties may arise. The whole concept of how innovation and regulation are interconnected was summed up perfectly by Stephanie when she said that "plant varieties are the vehicle where by innovation comes to the farmer." It is this idea, of beneficial innovation flowing from breeding programs to farmers and eventually consumers through regulatory frameworks, which must be central to any regulation of plant breeding into the future. Failure to keep this point paramount by allowing ideologies, commercial interests, political opinions, research goals or intellectual property rights to become predominant factors, will lead to regulations that do not have the interests of sustainable agriculture and food security at their heart.

# CURRICULUM VITAE

## Timothy Sykes

Frohburgstrasse 311
Zürich 8057, Switzerland
timsykes3@gmail.com

**Tertiary Qualifications**

| | |
|---|---|
| Nov 2013 – Current | **PhD candidate, IDP Bridges 'Science and Policy' fellow**<br>ETH Zurich, Switzerland |
| Oct 2012 – Sep 2013 | **Masters of Science, Biotechnology (distinction),** Phytochromes and the Circadian Clock in Arabidopsis, University of Essex, United Kingdom (Recipient of the Biotechnology Prize) |
| Mar 2003 – Jun 2008 | **Bachelor of Science (Majoring in Genetics, Biochemistry and Molecular Biology)**<br>University of Melbourne, Australia (Second class Honours) |

**Relevant Work Experience**

**Sep 2008 – Mar 2012: Genetic Technologies Ltd,** Melbourne, Australia.

- 2008-2011, Scientist; Animal DNA testing, Forensics and Paternity Testing.
- 2011-2012, Senior Scientist; Team leader for Animal DNA Testing, Molecular Diagnostics.

**Commercial research and development experience**

**Genetic Technologies Ltd:** During my time at Genetic Technologies Ltd. I was involved in developing several DNA-based molecular tests for use in pedigree dog breeding. This involved developing quick and cheap molecular tests to identify individuals that were either carrying a disease allele. In parallel to this, I also managed a small team of scientists and technicians involved in the daily testing of samples using the developed assays.

**PhD candidate:** For my PhD project, I have closely collaborated with the plant breeding company Norddeutsche Pflanzenzucht Hans-Georg Lembke KG where I have gained a thorough understanding of the workings of a commercial plant breeding company. This exposure has allowed me to fully appreciate what is involved in bringing a plant variety to market as well as the knowledge and techniques needed to do so.

**Prizes and Awards**

- **IDP Bridges PhD Fellowship: Bridging Plant Science and Policy,** Marie Curie Initial Training Networks (ITN), European Union's Seventh Framework Programme
- **The Biotechnology Prize 2013,** highest marks during the Biotechnology Masters Thesis, University of Essex
- **Young Scientist Travel Award,** Swiss Society of Agronomy (SGPW/SSA)

**Professional Memberships**

| | |
|---|---|
| Jun 2016 – Mar 2017 | Representative of the Swiss Academy of Sciences (SCNAT) in the European Academies Science Advisory Councils (EASAC) working group on Genome Editing |
| Since 2016 | Member of the European Association for Research on Plant Breeding (EUCARPIA) |

## Current Research

My current research focuses on identifying the gene(s) or genomic region(s) responsible for fertility restoration of cytoplasmic male sterility in perennial ryegrass (*Lolium perenne* L.). This research has the ultimate goal of contributing to the development of hybrid varieties by providing perennial ryegrass breeders with a molecular tool to identify fertility restoring individuals as well as a universal bioinformatics tool for identifying candidate restorer genes.

**Projects:**

- Development of a bioinformatics tool to *in silico* identify candidate genomic region(s) responsible for fertility restoration of cytoplasmic male sterility in multiple species (published).
- Genotyping of a perennial ryegrass population segregating for fertility restoration using genotyping by sequencing and bulk segregant analysis to identify the genomic regions responsible for male fertility restoration.
- Expression analysis of candidate restorer genes identified with the *in silico* pipeline to uncover genetic differences between restoring and maintaining individuals.

## Conferences and Symposia

- **International Plant & Animal Genome XXV - Grasslands workshop,** January 2017, San Diego, USA (Speaker)
- **EUCARPIA - General Congress**, Panel Discussion Moderator of the Plenary Session 'Innovation vs. Regulation', Aug 2016, Zurich, Switzerland
- **Eucarpia Fodder Crops and Amenity Grasses Section Symposium,** Sep 2015 in Ghent, Belgium (Speaker)
- **SwissMito Conference,** Sep 2014 in Kandersteg, Switzerland (Speaker)
- **Molecular Forage Crop Breeding Minisymposium,** Forage Crop Genetics (ETH) and Molecular Plant Ecology (Agroscope), Sep 2014 (Speaker), Apr 2015
- **Zurich-Basel Plant Science Center Symposium,** Nov 2013, Nov 2014 (Poster), Nov 2015 (Poster), Nov 2016 (Poster) in Zürich, Switzerland
- **IDP BRIDGES,** bridging Science and Policy Annual Meeting, Nov 2014 (Speaker), Nov 2015 (speaker), Nov 2016 (Speaker) in Männedorf, Switzerland

## Publications

**Papers:** Sykes, T., Yates, S., Nagy, I., Asp, T., Small, I., & Studer, B. (2016). ***In-silico* identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.).** Genome biology and evolution, evw047.

**Articles:** **'The benefits of new plant breeding techniques.'** Plant Science News. Newsletter of the Zurich-Basel Plant Science Center. No 27, Spring 2015.
**'EUCARPIA - PLENARY DISCUSSION Innovation vs. Regulation'** IDP Bridges News, No. 5, Winter 2016.

**Fiction:** **'Expiration Date'**. Competition winner selected by DSM Nutrition. Published by Kagar publishing in their 'Future of Food' anthology, 2016.

## References

**Prof. Dr. Bruno Studer**
ETH Zurich
Universitätstrasse 2
8092 Zürich, Switzerland
+41 44 632 01 57
bruno.studer@usys.ethz.ch

**Alison Mew**
Centre for Biopharmaceutical Excellence
16 Acacia Place
Victoria 3067, Australia
+61 9445 0845
Alison.mew@cbe-ap.com.au