FACULTY OF SCIENCES

DEVELOPMENT OF GENOMIC RESOURCES FOR PERENNIAL RYEGRASS

Elisabeth Veeckman

Supervisors: Prof. Dr. Klaas Vandepoele, Dr. Tom Ruttink, Prof. Dr. Peter Dawyndt

Submitted to the Faculty of Sciences of Ghent University,
In Fulfilment of the Requirements for the Degree of Doctor in Bioinformatics

Academic year 2018 – 2019

GHENT UNIVERSITY

ILVO
Institute for Agricultural and Fisheries Research

Examination committee

Prof. Dr. Steven Maere (chair)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Isabel Roldán-Ruiz (secretary)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Flanders Research Institute for Agriculture, Fisheries and Food (ILVO)

Prof. Dr. Klaas Vandepoele (promoter)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Dr. Tom Ruttink (co-promoter)

Flanders Research Institute for Agriculture, Fisheries and Food (ILVO)

Prof. Dr. Peter Dawyndt (co-promoter)

Faculty of Sciences, Department of Applied Mathematics, Computer Science and Statistics, Ghent University

Prof. Dr. Torben Asp

Department of Molecular Biology and Genetics, Aarhus University, Denmark

Dr. Philippe Barre

French National Institute for Agricultural Research (INRA), France

Dr. Lieven Sterck

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Summary

Thanks to the development of Next-Generation Sequencing technologies, the field of genomics has changed rapidly and dramatically in the past decades. This leads to a vast expansion of an array of genomic resources. In this PhD thesis, I used and developed bioinformatics methods to generate and improve three genomic resources for perennial ryegrass (*Lolium perenne* L.): (i) a reference genome sequence, (ii) a complete structurally and functionally annotated gene set, and (iii) a comprehensive overview of the genome sequence diversity within this species. These resources are of great use in understanding the evolution and biology of perennial ryegrass, in explaining species-specific morphology and physiology, as well as within-species phenotypic variation within a species.

Perennial ryegrass is the most widely cultivated grass species in temperate regions. It is of interest for grazing, hay and silage production as it has a long growing season, and high yield potential and high nutritive value. The development of genomic resources for perennial ryegrass suffers from several major challenges. First, the genome size and repetitive content hampers obtaining a complete and contiguous reference genome sequence. Currently the best available draft genome sequence represents only half of the actual genome size, is highly fragmented and only partially anchored onto pseudo-chromosomes (Byrne et al., 2015). Second, the corresponding gene annotation set is purely evidence-based and our analyses showed that this gene set is incomplete. Moreover, this gene set lacks a high-quality functional annotation. Finally, the perennial ryegrass genome is highly heterogeneous because of its outbreeding nature. This leads to difficulties in the identification of genomic sequence variation.

Despite the strong increase in the number of plant genome sequences, we found that no clearly defined measures to assess completeness both at the genome assembly and gene space level existed at the time of the start of this PhD. With hundreds of plant genome sequences available and many more to come, and little consensus in the literature on how to report their quality, we investigated how completeness should be estimated. We therefore defined the concept of the 'expected' genome size and gene space, to which completeness can be expressed. The expected gene space can either be based on experimental evidence, or on evolutionary conservation

principles. These measures can now be used to assess whether the fraction of missing genes reported by the analysis, are indeed missing from the genome sequence, or were missed during gene annotation.

Completeness analysis showed that the current publicly available perennial ryegrass draft genome sequence (Byrne et al., 2015) contained a very large fraction of the gene space, but 1,709 core gene families (24%) were missing from the gene annotation set. We have improved the gene annotation set by integrating evidence-based and *ab initio* gene models, and have calibrated parameters using a set of 503 manually curated, high quality gene models. The resulting genes were functionally annotated, and incorporated into the PLAZA comparative genomics platform that offers a framework for functional, evolutionary, and comparative plant genomics (Proost et al., 2009; Van Bel et al., 2012; Proost et al., 2015; Van Bel et al., 2018).

A complete and contiguous reference genome sequence is required to study genome architecture and evolution. In a collaborative effort with Aarhus University and the University of Tübingen, we assembled a novel, chromosome-scale genome reference sequence for perennial ryegrass, using a combined approach consisting of the most recent technologies. The repetitive fraction of the genome was assembled using additional PacBio SMRT sequencing; hybrid scaffolding with optical mapping further improved scaffold length; and chromosomal conformation capture (Hi-C) was a final step to order and orient the scaffolds into pseudo-chromosomes. This brings the perennial ryegrass genome sequence quality on par with other major grass Poaceae genomes, such barley (5 Gb) and wheat (17 Gb), that were only recently obtained using a similar approach (Mascher et al., 2017; Appels et al., 2018). Together with an annotated gene set, this genome sequence can be used in future studies on genome evolution and species-specific biology through comparative genomics using the PLAZA platform.

We were interested in the identification of genomic sequence variation in 503 candidate genes involved in the control of plant development and architecture. This was done for a collection of 743 individuals, derived from breeding material, current cultivars and natural accessions to comprehensively represent the perennial ryegrass germplasm. To overcome challenges in variant calling, two complementary strategies were used to obtain a complete and reliable variant set.

First, four variant calling pipelines were used and automatically integrated to reach maximal sensitivity. We compared hard filtering with precision-based filtering to obtain a high quality variant set. Highly multiplex amplicon sequencing was used as an independent genotyping technology to empirically estimate an appropriate precision threshold. Second, *de novo* assembly followed by overlap-layout-consensus clustering was used to reconstruct divergent alleles that were missed using variant calling based on short read mapping onto a single reference genome sequence. This approach is broadly applicable to other highly diverse outbreeding species and provides important insights in the pitfalls and solutions of bioinformatics analysis for population-scale resequencing.

Genomic sequence variants may affect gene function and/or regulation and could therefore result in phenotypic variation. Predicting the effect of genomic sequence variants using structural information is a reverse genetics approach, as it first identifies mutations based on DNA variants and then studies the associated phenotype. This allows for the identification of carriers of rare defective alleles that can be used for validation of causal polymorphisms and functional gene analysis, or carriers can be incorporated into the breeding program. For instance, we identified naturally occurring premature stop codons in one-third of the 503 candidate genes investigated, including the single copy genes *GIGANTEA* (*LpGI-01*) and *ENHANCED RESPONSE TO ABSCISIC ACID 1* (*LpERA1-01*). After annotating 18 members of the complete FLOWERING LOCUS T gene family in the *L. perenne* genome, we illustrate how molecular knowledge on essential amino acid residues can also be taken into account, and we have identified several genotypes with sequence variation in the external loop and at key residues of the anion ligand binding site.

Alternatively, association genetics studies aim to statistically associate phenotypic variation with genotypic variation. This is a useful forward-genetics approach for the dissection of quantitative and complex traits that are regulated by multiple genes. We have designed a high-throughput genotyping assay based on multiplex amplicon sequencing and have screened breeding populations for associations with two important traits for perennial ryegrass breeding: flowering time and leaf elongation. Out of 28 candidate genes, *LpFT-03* was found to associate with heading date, and we confirmed the existence of haplotypes previously identified by Skot et al. (2011). *LpMADS1* controls vernalization-induced flowering, and was found to associate with maximal leaf

length after autumn growth. Although the association was not found with other leaf elongation traits such as spring growth rate, this is an interesting candidate marker as the associated polymorphism is located in the first intron, which is involved in regulation of *LpMADS1* expression (Yan et al., 2003; Fu et al., 2005; Hemming et al., 2009).

Samenvatting

De ontwikkeling van *next-generation sequencing* technologieën heeft het *genomics* onderzoeksveld sterk veranderd tijdens de afgelopen decennia, en leidt tot een steeds snellere ontwikkeling van diverse genomische hulpbronnen. In dit proefschrift gebruikte en ontwikkelde ik bio-informatica methoden om drie genomische hulpbronnen voor Engels raaigras (*Lolium perenne* L.) te genereren of te verbeteren: (i) een referentie genoomsequentie, (ii) een volledige structureel en functioneel geannoteerde genen set, en (iii) een uitgebreid overzicht van de diversiteit van de genoomsequentie binnen dit species. Deze hulpbronnen zijn in de toekomst van groot nut bij het begrijpen van de evolutie en biologie van Engels raaigras, in het verklaren van soort-specifieke morfologie en fysiologie, evenals fenotypische variatie binnen deze soort.

Engels raaigras is de meest gecultiveerde grassoort in gematigde streken. Het is van belang voor begrazing en hooi- en kuilvoerproductie, dankzij een lang groeiseizoen, een hoog opbrengstpotentieel en een hoge voedingswaarde. De ontwikkeling van genomische hulpbronnen voor Engels raaigras kent verschillende grote uitdagingen. Ten eerste is het verkrijgen van een volledige en continue genoomsequentie moeilijk door de genoomgrootte en de fractie repetitieve sequenties. De best beschikbare genoomsequentie vertegenwoordigt slechts de helft van de werkelijke genoomgrootte, is sterk gefragmenteerd en slechts gedeeltelijk verankerd op pseudo-chromosomen (Byrne et al., 2015). Ten tweede is de bijbehorende gen annotatie set niet compleet en missen deze genen een functionele annotatie. Tenslotte is het genoom van Engels raaigras zeer heterozygoot vanwege zijn uitkruisende aard. Dit leidt tot problemen bij de identificatie van genomische sequentievariatie.

Ondanks de steeds sterkere toename van het aantal beschikbare plantengenoomsequenties, stelden we bij het begin van dit doctoraat vast dat er geen duidelijk gedefinieerde maten beschikbaar waren om de volledigheid ervan te beoordelen, zowel voor de genoom *assembly* als de gen annotatie. Daarom hebben we onderzocht hoe volledigheid geschat moet worden, en definieerden het concept van de 'verwachte' genoomgrootte en gen inhoud om volledigheid uit te drukken. Zowel experimenteel bewijs als evolutionaire principes kunnen aangewend worden om de verwachte gen inhoud te bepalen. Zo kunnen de maten voor de volledigheid van de

genoom *assembly* en gen annotatie set kunnen gebruikt worden om te beoordelen of de gerapporteerde fractie ontbrekende genen, afwezig was in de genoomsequentie, of dat gemist werden bij de gen annotatie.

Analyse van de volledigheid van de huidig beschikbare genoomsequentie en gen annotatie set (Byrne et al., 2015) toonde aan dat de gen inhoud goed vertegenwoordigd is in de genoomsequentie, maar dat 1,709 *core gene families* (24%) ontbreken in de gen annotatie set. We verbeterden deze gen annotatie set door een integratie van de beschikbare gen modellen met *ab initio* gen modellen, en kalibreerden parameters met behulp van een set van 503 manueel gecureerde gen modellen. De resulterende genen werden functioneel geannoteerd en opgenomen in het *PLAZA comparative genomics platform* dat een raamwerk biedt voor functionele, evolutionaire en vergelijkende genoom analyse voor planten (Proost et al., 2009; Van Bel et al., 2012; Proost et al., 2015; Van Bel et al., 2018).

Een volledige en continue referentie genoomsequentie is vereist om de architectuur en de evolutie van het genoom te bestuderen. In samenwerking met Universiteit Aarhus en Universiteit Tübingen, hebben we een nieuwe referentie genoomsequentie voor Engels raaigras samengesteld, door gebruik te maken van een combinatie van de meest recente technologieën. De repetitieve fractie van het genoom werd geassembleerd met behulp van aanvullende *PacBio SMRT-sequencing*; de combinatie met *optical mapping* verbeterde de *scaffold* lengte; en *chromosomal conformation capture* (Hi-C) was een laatste stap om de *scaffolds* te ordenen en te oriënteren in pseudo-chromosomen. Dit brengt de kwaliteit van de genoomsequentie van Engels raaigras op hetzelfde niveau als andere belangrijke gewassen, zoals gerst (5 Gb) en tarwe (17 Gb), die pas recent werden bekomen met een vergelijkbare strategie (Mascher et al., 2017; Appels et al., 2018). Samen met een geannoteerde genen set kan deze genoomsequentie in de toekomst worden gebruikt om de genoomevolutie en de biologie van Engels raaigras te bestuderen via vergelijkende genoom analyse met behulp van het PLAZA-platform.

We waren geïnteresseerd in de identificatie van genomische sequentievariatie in 503 kandidaat-genen die betrokken zijn bij de controle van plantontwikkeling en architectuur. Dit werd gedaan voor 743 individuen, afkomstig van veredelingspopulaties, huidige cultivars en natuurlijke

accessies, als representatie van de genenpool van Engels raaigras. Twee complementaire strategieën werden gebruikt om een complete en betrouwbare set van varianten te verkrijgen, en zo de moeilijkheden van het bepalen van genomisch sequentievariatie te overkomen. Ten eerste werden vier *variant calling pipelines* gebruikt en automatisch geïntegreerd om maximale sensitiviteit te bereiken. *Hard filtering* werd vergeleken met *precision-based filtering* voor het bekomen van een set van varianten van hoge kwaliteit. *Multiplex amplicon sequencing* werd gebruikt als een onafhankelijke genotyperingstechnologie om de *precision threshold* empirisch in te schatten. Ten tweede werd *de novo assembly*, gevolgd door *overlap-layout-consensus clustering* gebruikt voor het reconstrueren van sterk divergente allelen die gemist worden wanneer gebruik gemaakt wordt van *variant calling* gebaseerd op het aligneren van korte *reads* op een referentie genoomsequentie. Deze aanpak is ook van toepassing op andere, zeer diverse soorten en biedt belangrijke inzichten in de valkuilen en oplossingen van bio-informatica analyses bij het sequeneren van populaties.

Genomische sequentievarianten kunnen de gen functie en/of regulatie beïnvloeden en kunnen daarom resulteren in fenotypische variatie. Het voorspellen van het effect van genomische sequentievarianten met behulp van structurele informatie is een *reverse genetics*-benadering, omdat het eerst mutaties op basis van DNA-varianten identificeert en vervolgens het bijbehorende fenotype bestudeert. We identificeerden bijvoorbeeld van nature voorkomende premature stop codons in één derde van de 503 kandidaat genen, waaronder *GIGANTEA* (*LpGI-01*) en *ENHANCED RESPONSE TO ABSCISIC ACID 1* (*LpERA1-01*). Na annotatie van de 18 leden van de volledige *FLOWERING LOCUS T* genfamilie in het *L. perenne*-genoom, illustreren we hoe moleculaire kennis van essentiële aminozuurresiduen ook in rekening gebracht kan worden, en hebben we verschillende genotypes geïdentificeerd met sequentievariatie in de externe lus en bij belangrijke residuen van de anion ligandbindingsplaats. De dragers van deze interessante allelen kunnen gebruikt worden voor de validatie van polymorfismen die causaal zijn voor fenotypische variatie, voor functionele gen analyse, en voor incorporatie in een veredelingsprogramma.

Daartegenover staan associatie genetica studies, die zijn gericht op het statistisch associëren van fenotypische variatie met genotypische variatie. Dit is een *forward genetics*-benadering voor het

ontleden van kwantitatieve en complexe kenmerken die door meerdere genen worden gereguleerd. Bloeitijdstip en bladelongatie zijn twee belangrijke eigenschappen voor de veredeling van Engels raaigras. We ontwikkelden een kost-efficiënte *high-throughput* genotyperingsmethode op basis van *multiplex amplicon sequencing*, en zochten in veredelingspopulaties naar associaties met deze twee eigenschappen. Uit 28 kandidaat-genen bleek *LpFT-03* te associëren met de bloeitijdstip en we bevestigden het bestaan van haplotypes die eerder al geïdentificeerd werden door Skot et al. (2011). *LpMADS1* controleert de door vernalisatie geïnduceerde bloei en associeert met maximale bladlengte na herfstgroei. Hoewel de associatie niet werd gevonden met andere bladelongatiekenmerken, zoals de groeisnelheid tijdens de lente, is dit een interessante kandidaat-merker omdat het geassocieerde polymorfisme zich in het eerste intron bevindt, dat betrokken is bij regulatie van *LpMADS1*-expressie (Yan et al., 2003; Fu et al., 2005; Hemming et al., 2009).

Table of Contents

List of Abbreviations

AM	Association mapping
CG	Candidate Gene
CNV	Copy Number Variation
coreGF	PLAZA Core Gene Family
EP	Estimated Precision
EST	Expressed Sequence Tag
EVM	EVidence Modeler
GATK	Genome Analysis Toolkit
GBS	Genotyping-by-Sequencing
GDD	Growing Degree Days
GO	Gene Ontology
GS	Genomic Selection
GWAS	Genome Wide Association Study
HD	Heading Date
Indel	Insertion & Deletion
IQR	Interquartile Range
LD	Linkage Disequilibrium
LOF	Loss Of Function
MAF	Minor Allele Frequency
NGS	Next Generation Sequencing
OLC	Overlap-Layout-Consensus
PCA	Principal Component Analysis
PEBP	phosphatidylethanolamine-binding proteins
QTL	Quantitative Trait Locus
SI	Self-Incompatibility
SGR	Spring Growth Rate
SMAP	Stack Mapping Anchor Point
SNP	Single Nucleotide Polymorphism
TE	Transposable Element
VC	Variant Calling

1 Introduction

In this PhD thesis, I used and developed bioinformatics methods to generate or improve genomic resources for perennial ryegrass (*Lolium perenne* L.). Genomic resources, such as a reference genome sequence, a complete structurally and functionally annotated gene set, and a comprehensive overview of the genome sequence diversity within a species, are of great use in understanding the evolution and biology of an organism, to help explain species-specific morphology and physiology, as well as phenotypic variation within the species. Genomics is the study of whole genomes of organisms, using high-throughput DNA sequencing methods and bioinformatics to sequence, assemble and analyze the structure and function of genomes, including the gene content. Thanks to the development of Next-Generation Sequencing (NGS) technologies, the field of genomics has changed rapidly and dramatically in the past decades.

1.1 Perennial ryegrass

Perennial ryegrass is the most widely cultivated grass species in temperate regions. It is native to Europe, temperate Asia and North Africa, and is widely distributed throughout the world, including North and South America, Europe, New Zealand and Australia (Hannaway et al., 1999). It is a major constituent of amenity grass mixtures for lawns and sports turfs, but more importantly, it is used as forage crop. It is of interest for grazing, hay and silage production as it has a long growing season, a high yield potential and high nutritive value. Perennial ryegrass is a grass from the family Poaceae, the fourth largest family of flowering plants. It belongs to the Pooideae subfamily, together with major cereals of the tribe Triticeae (including wheat, barley and oat) (Figure 1.1), and other lawn and pasture grasses, such as Italian ryegrass and Festuca. The perennial ryegrass genome consist of seven chromosomes and is naturally diploid. Because perennial ryegrass is allogamous (outbreeding), individual plants are highly heterozygous and the perennial ryegrass genome is highly diverse both within and across natural accessions and breeding populations. This is established through a gametophytic self-incompatibility (SI) mechanism that is widespread in the grass family (Baumann et al., 2000).

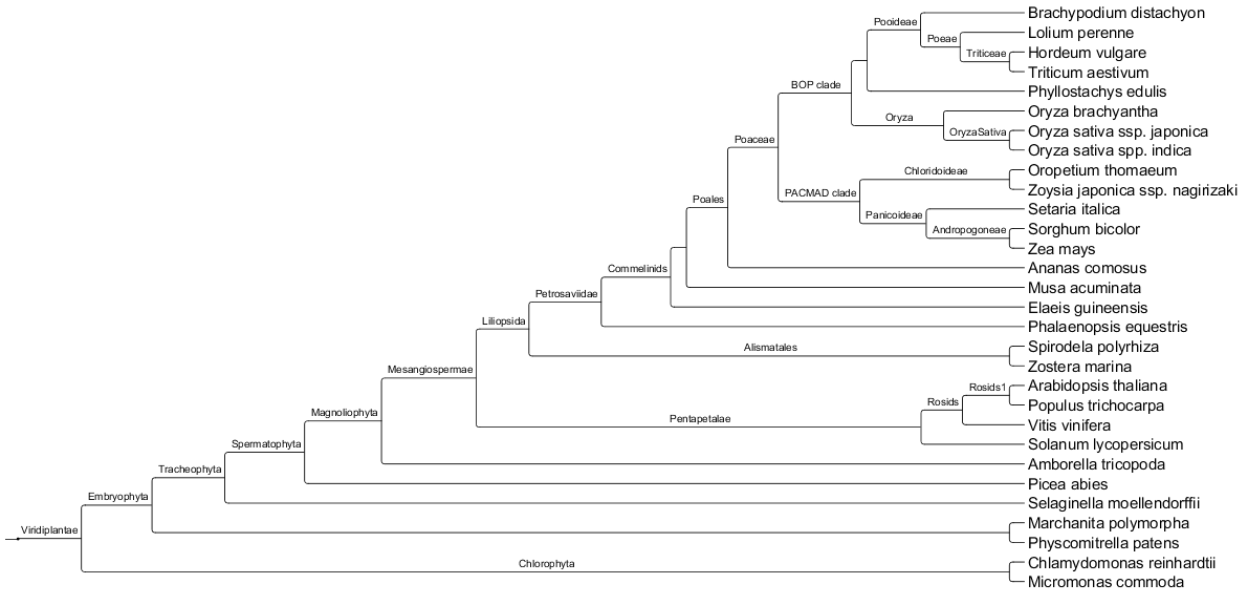


Figure 1.1 Species tree including perennial ryegrass (*L. perenne*). Based on the species tree available for PLAZA 4.0 Monocots (Van Bel et al., 2018).

The ultimate goal of a breeding program is to create new varieties that meet the requirements of the end-users. For a grassland species, such as perennial ryegrass, this means to improve its characteristics to withstand adverse environmental conditions and to increase its overall growth and vigour. At ILVO, perennial ryegrass breeding started in 1932 (then called *Rijksstation voor Plantenveredeling*). Overall, forage breeding has resulted in increased yield, persistency and rust resistance and decreased aftermath heading (Sampoux et al., 2010). Dry matter yield per unit of nitrogen input continues to be an important primary objective in making the most efficient use of land resources to meet increasing global demands for food, feed, fiber and fuel. As perennial ryegrass is a major feed source for livestock, high digestibility and an optimal nutritional composition of the forage help to minimize input costs through increased forage use (Humphreys et al., 2010). Climate change and increased pressure for adopting more sustainable agricultural practices are creating new challenges for ryegrass breeding (Helgadóttir et al., 2016). In Belgium and other western European regions, winters are getting warmer and summers drier, meaning that forage crops will have to use nutrients and water more efficiently to maximize yield.

Thanks to advances in molecular biology and high-throughput genotyping technology, the focus of plant breeding is gradually shifting from phenotype-based to genotype-based selection. The introduction of NGS technologies leads to a vast expansion of an array of genomic resources that

can be used for a wide range of applications. Examples include the study of species-specific biology and genome evolution by comparing the genome sequences of closely related species (Pfeifer et al., 2013); synteny-based transfer of QTL information between closely related species such as rice, barley, and members of the Festuca-Lolium complex (Armstead et al., 2004; Curley et al., 2004); identification of candidate genes involved in physiological or morphological processes (Skot et al., 2007); identification of functional alleles and their carriers (Shinozuka et al., 2012); development of genetic markers and the use of genome sequence variants for high-throughput marker systems (Byrne et al., 2013; Blackmore et al., 2016). These applications are essential to further develop new breeding strategies such as genomic selection that will improve and accelerate the breeding process (Fe et al., 2016).

Definition box

Contig
First result of genome assembly, contiguous sequence derived from overlapping short DNA fragments.

Scaffold
Scaffolds are created by chaining contigs together, and contain information on the relative position and orientation of the contigs. Gaps between contigs are represented by NNN's.

Reference genome sequence
Backbone DNA sequence representative for the haploid genome, ideally on chromosome scale.

Gene structure prediction
Identification of the location and structure of a gene.

Gene function prediction
Assigning functions to genes, such as a biological or biochemical role.

Synteny
Conserved gene content and order among species evolved from a common ancestral genome.

Gene family
Group of genes descended from an ancestral gene by duplication and speciation events.

Genomic sequence variation
Differences in chromosomal DNA sequence across individuals of the same species.

The main genomic resources for perennial ryegrass that are central to this PhD are shown in Figure 1.2. A reference genome sequence, gene annotation set and genomic diversity are highly interconnected, as they are often used together. The quality and completeness of these central resources, therefore, have a major impact on the development of the other genomic resources and analyses. Yet, the development of these genomic resources for perennial ryegrass suffers from several major challenges. Obtaining a complete and contiguous reference genome sequence is challenging because the perennial ryegrass genome is large and highly repetitive. Currently the best available draft genome sequence represents only half (1.1 Gbp) of the actual genome size (2.1 Gbp), is highly fragmented (48k scaffolds) and only partially anchored onto pseudo-chromosomes (Byrne et al., 2015). A second problem is that the corresponding gene annotation set (28k genes) is purely evidence-based, and while it contains high quality gene models, our analyses (Veeckman et al., 2016) show that this gene set is incomplete (see Chapter 2). Moreover, this gene set lacks a high-quality functional annotation. Finally, the perennial ryegrass genome is highly heterogeneous because of its outbreeding nature. This leads to

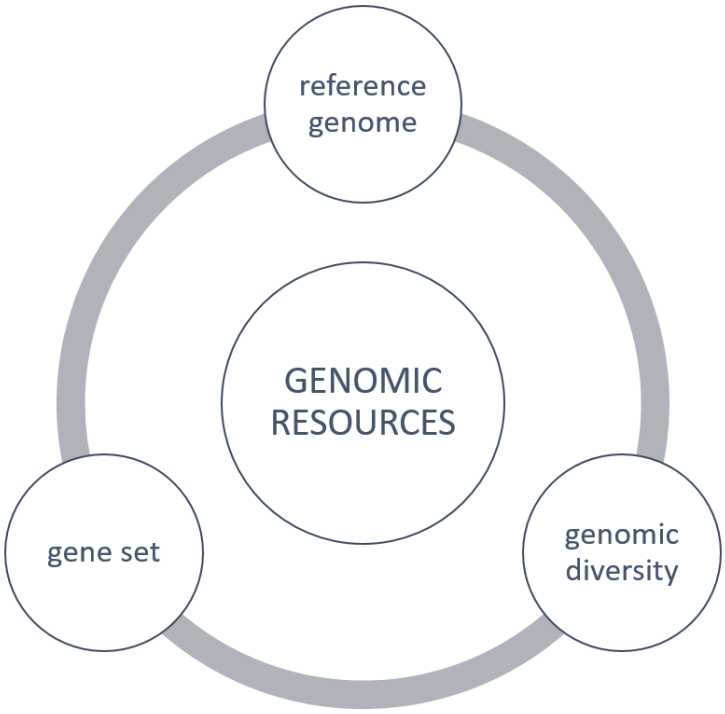


Figure 1.2 Overview of genomic resources that are central to this PhD. A reference genome is the representative DNA sequence of the genome of an organism and is the backbone sequence used for many downstream analyses. The gene set contains information on the location and structure of the genes in the reference genome. Resequencing individuals of the same species gives insight in the genomic diversity that exists within a species.

difficulties in the identification of genomic sequence variation when applying commonly used variant calling algorithms that were developed for species with much lower levels of sequence diversity.

In the remainder of this chapter, I will elaborate on the use of these resources in general, their state at the start of this PhD, and highlight the methods and approaches that were developed during this PhD to improve them.

1.2 Reference genome assembly and gene annotation set

Figure 1.3 shows the relation between the reference genome and gene set, and their respective applications. The reference genome sequence is the basic genomic resource for many applications. This is the backbone used for gene structure prediction that identifies the location and structure of genes. By combining both reference genome sequence and gene set, the gene content and gene order is known. Genome evolution can then be studied through comparative genomics, as the gene order is conserved across closely related species (synteny), because they are derived from a common ancestral genome. Similarly, gene content can be compared across

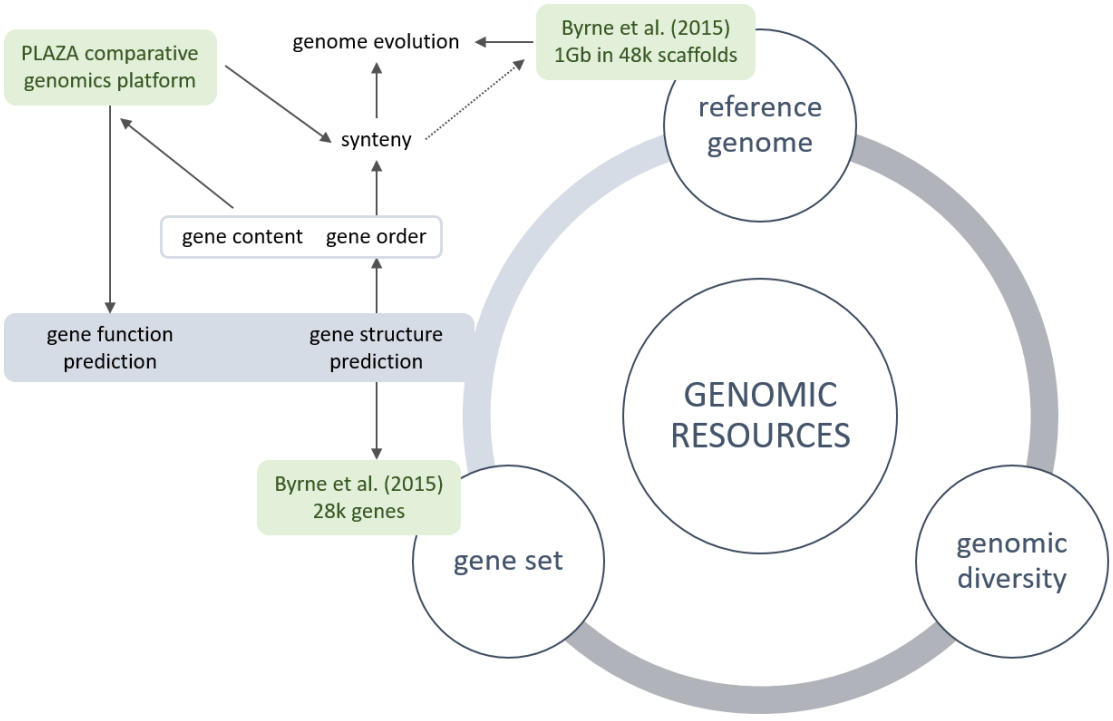


Figure 1.3 Overview of genomic resources and applications related to the reference genome and set of annotated genes.

species to identify gene family expansion, contraction, gene loss and gain that further define species-specific biology. Comparative genomics analyses become possible by integration of perennial ryegrass into the PLAZA comparative genomics platform (Proost et al., 2009). Additionally, PLAZA also provides a framework for gene function prediction by projecting functional annotations through orthology within gene families.

Plant genome assembly

A genome is the genetic material of an organism and consists of a DNA sequence containing all of the hereditary instructions for creating and maintaining life, as well as instructions for reproduction, to build a body, and to respond to the environment. The development of high-throughput DNA sequencing technologies has made genome sequencing much cheaper and easier, and the number of complete genome sequences is growing rapidly (Figure 1.4). A reference genome sequence represents a complete and contiguous genome sequence of a single representative individual of a species of interest, with a cumulative scaffold length equal to the haploid genome size.

The genome of *Arabidopsis thaliana* was the first plant genome to be fully sequenced and assembled in the year 2000, and was still entirely based on Sanger sequencing. This was a major milestone not only for plant research, but also for genome sequencing. Since then, there has

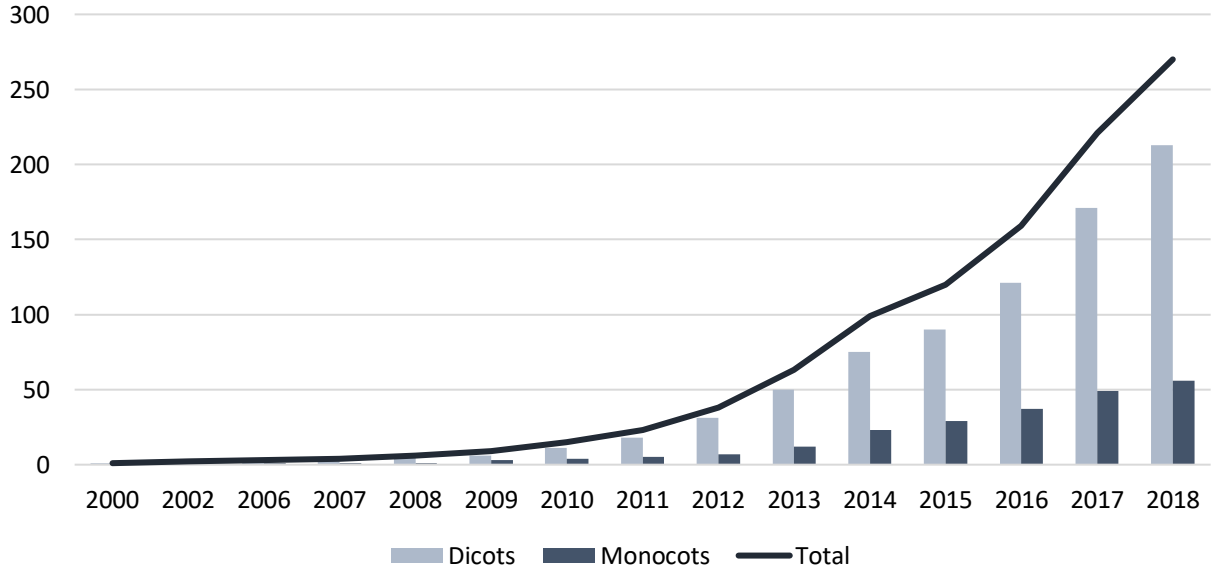


Figure 1.4 Exponential increase of the number of sequenced plant genomes. Source: PlabiPD (www.plabipd.de)

been an exponential increase in the number of available plant genomes, thanks to the shift to NGS technologies. Genome sequencing of other model species and crops followed, such as rice (Yu et al., 2002), maize (Schnable et al., 2009) and wheat (Mayer et al., 2014; Appels et al., 2018), and currently genome sequences of more than 250 plant species are available (Figure 1.4). The completeness, contiguity and sequence accuracy of a reference genome sequence determine its quality, which is highly variable among newly sequenced plant genomes. Very often, improved versions are released in the years following the first draft genome sequence, illustrating that a reference genome sequence is dynamic and subject to continuously ongoing research.

The main goal of genome assembly is to create a genome sequence with the longest possible contiguous sequences and the smallest number of mis-assemblies, thus representing the actual DNA sequence of the chromosomes. Before starting to assemble a new genome, it is important to first estimate the genome size, repetitive content, level of heterozygosity and ploidy, and GC content. The larger the genome, the more sequencing data is required to cover all genome positions at sufficient read depth. Ideally, the genome size is estimated by flow cytometry, but the genome size of related species can also give a first indication. Second, it is important to estimate the repetitive sequence content. Repeats are sequences that occur in multiple copies in the genome, at different locations. Both the amount and dispersion of repeats affect the assembly process as repeats are typically collapsed leading to a fragmented assembly, with contigs that often end where a repeat region starts (Phillippy et al., 2008). Highly heterozygous regions will be reconstructed as independent sequences, leading to redundant sequences for a single chromosomal locus. Conversely, highly homologous sequences in polyploid genomes will collapse. Finally, Illumina sequencing is biased against sequences with low or extremely high GC-content, resulting in low or even no coverage, thus effectively excluding such regions from the assembly (Dominguez Del Angel et al., 2018).

New technologies and strategies for the assembly of large and complex plant genomes

The assembly of large and complex plants genomes is still a big challenge. The genome size and large amount of repetitive sequences are the main reasons why the genome assembly for *L. perenne* and other closely related species such as barley, wheat and oat require new strategies

and a combination of innovative sequencing and assembly technologies to obtain a chromosome-scale assembly.

The short read length of NGS technologies limits contiguous *de novo* genome assembly, due to repetitive elements and large structural variations that are common in plant genomes. A new generation of sequencing methods has been developed as single-molecule sequencing technology, also known as the 'third' generation. The most popular platforms are the single-molecule real time (SMRT) sequencing technology of Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). These platforms sequence directly individual DNA molecules and produce reads: the PacBio platform sequences a DNA template multiple times to generate a contiguous long read with average length > 10 kb and up to 60kb (Weirather et al., 2017), while with ONT the maximal read length is theoretically only limited by the length of the DNA fragment, resulting in read length of several hundreds kb (Jiao and Schneeberger, 2017). Although third generation sequencing platforms overcome the read length limitation, they come with other limits, such as higher sequencing cost per base, lower throughput and higher sequencing error rate (Carneiro et al., 2012). Therefore, hybrid assembly is now the more common approach for plant genome assembly, using long-read sequencing data for scaffolding and contiguity, combined with error correction by Illumina short-read data with high read depth and relatively low error rate (Badouin et al., 2017; Jiao et al., 2017; Zou et al., 2017).

Optical mapping is a technique for constructing an ordered, genome-wide restriction map and was originally developed by Dr. David C. Schwartz in the 1990s (Schwartz et al., 1993). An endonuclease creates single-strand nicks in long DNA molecules at a specific recognition site, wherever it occurs in the genome. Fluorescently labeled nucleotides are incorporated at these sites, resulting in a characteristic fingerprint for each DNA molecule. Integration of sequence contigs with optical maps (hybrid scaffolding) in which the distance between restriction sites is estimated helps bridging repetitive regions, and is therefore often used to improve assembly contiguity. It also allows to order and orient scaffolds, to identify and correct possible chimeric joints in the assembly and to estimate gap sizes (Zhou et al., 2009; Chamala et al., 2013; Tang et al., 2014). Optical mapping is very useful in assisting the assembly of polyploid and highly heterozygous plant genomes.

A third technique is based on the conformation of chromatin in the nucleus. Hi-C probes this three-dimensional architecture of whole genomes, by coupling proximity-based ligation with massively parallel sequencing (Belton et al., 2012). First, adjacent chromatin segments are covalently linked. Next, the chromatin is digested with a restriction enzyme, and the DNA fragments that are covalently linked together are ligated. Meanwhile, a biotin-labelled nucleotide is incorporated at the ligation junction, enabling selective purification followed by deep paired-end sequencing. The primary application for which Hi-C was developed is to detect and study chromatin interactions. However, as most interactions are intra-chromosomal, and the number of interactions decreases with increasing distance, the information that is embedded in Hi-C can also be used to assemble scaffolds into a chromosomal context (Oddes et al., 2018). Several Hi-C-based scaffolding methods have been developed, such as LACHESIS (Burton et al., 2013), GRAAL and 3D DNA (Dudchenko et al., 2017). The first step of Hi-C scaffolding is grouping scaffolds into chromosomes. Next, the genomic order of scaffolds within each chromosome is determined, as well as the distance between neighboring scaffolds. Finally, the scaffolds are oriented with respect to other scaffolds in the chromosome. The results are typically represented in a genome-wide interaction matrix, reflecting the interaction frequencies between genomic loci, also called a contact probability map. Most interactions are intra-chromosomal, resulting in a high interaction frequency along the diagonal, representing the chromosomal order of scaffolds.

Genome annotation

The next step in a genome project is annotation of protein coding genes, as well as other features such as non-coding RNAs and regulatory and repetitive sequences. Gene prediction comprises the identification of the location and structure of genes (gene structure prediction) and their function (gene function prediction) (Figure 1.3).

There are two fundamentally different methods for gene structure prediction: (i) intrinsic (*ab initio*) methods use only the features embedded in the genome sequence, such as coding potential (including translational start and stop sites, open reading frame length and codon usage), and splice site prediction, to computationally predict gene structures; or (ii) extrinsic

methods use transcript mapping and protein homology with closely related species to predict transcribed regions based on experimental evidence. Each method is associated with inherent advantages and disadvantages. While intrinsic methods rely on statistical models that need to be trained and optimized, they allow for predicting fast evolving species-specific genes and genes with specific expression behavior that are unlikely to be captured by experimental evidence. Whereas extrinsic methods are more universally applicable, they require a large amount of high quality protein sequences and transcripts derived from a broad range of conditions to account for variation in gene expression.

Once the location and structure of the genes is defined, biologically relevant information is assigned to the gene and corresponding mRNA and protein sequences. This process is known as gene function prediction. Functional elements may include putative names for protein-coding genes, gene ontology terms, functional sites and protein domains. The assignment of functional elements to genes allows for further understanding of specific genome properties and similarities to closely related species. Second, this is an additional quality check for the gene annotation set, by detection of genes that are annotated with terms associated with transposable elements, annotated with suspicious domains, or lack functional elements (Dominguez Del Angel et al., 2018).

Using conserved gene order to study genome evolution

As increasingly more plant genome sequences become available across the whole plant kingdom, the comparison of genes and genomes from different species becomes possible. This allows for the study of plant genome architecture, a better understanding of the function and location of genes, and provides insight in evolutionary relationships. Consequently, the field of comparative genomics is rapidly evolving and has uncovered the complexity of plant genomes and their evolutionary history, which often involves extensive duplications, reshuffling and gene reduction (Maere et al., 2005; Cui et al., 2006; Flagel and Wendel, 2009; Magadum et al., 2013).

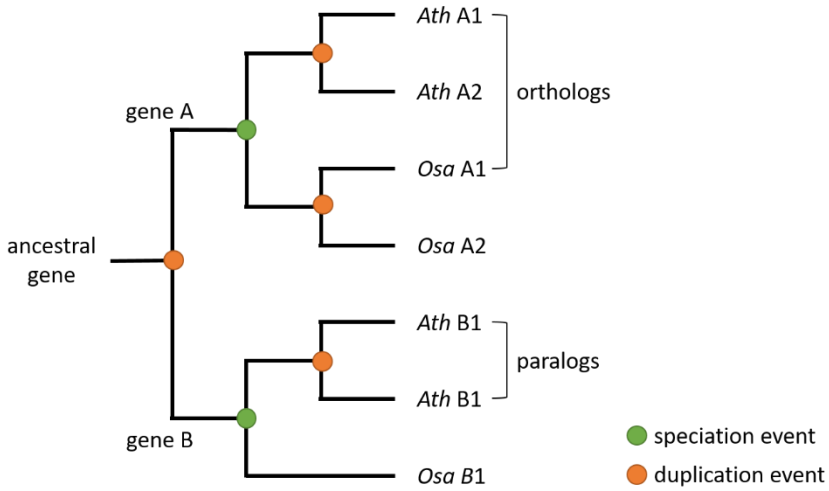
According to the evolution theory, closely related species are derived from a common ancestral species, hence share a common ancestral genome. Evolution theory further predicts that genomes from closely related species contain a similar gene content and gene order, which can

be recognized through DNA sequence similarity between the extant species. Species-specific genomic changes, such as chromosomal rearrangements, gene duplication and gene loss, that occurred since a speciation event, may have led to species-specific biology.

Syntenic regions are chromosomal regions that share a common order of homologous genes that are derived from a common ancestral genome (Tang et al., 2011). Synteny provides a framework that gives insight in the evolutionary processes leading to diversity in chromosome number and chromosome structure, e.g. the chromosome translocation involving the long arms of chromosomes 4 and 5 that is characteristic for some Triticeae species, such as barley, is absent in perennial ryegrass (Pfeifer et al., 2013). Additionally, synteny allows for the transfer of genetic markers and generation of genetic maps if there is no genome sequence available. For instance, synteny was used to identify QTLs for heading date in perennial ryegrass based on the chromosomal position of the heading date locus in rice (Armstead et al., 2004).

Using gene content to study species-specific biology

Homologous sequences descend from a common ancestral sequence. Sequence homology among DNA or protein sequences is inferred from sequence similarity through sequence alignment. Clustering genes based on sequence similarity results in the construction of gene families. This is a group of related genes that share a common ancestral gene. Two sequences



can share ancestry because of different scenarios (Figure 1.5): (i) orthologs are related sequences

Figure 1.5 Example of homologous relationships between genes. Orthologs are related sequences in different organisms as a result of a speciation event. Paralogs are related sequences within a single organism as a result of a gene duplication event. *Ath*: *Arabidopsis thaliana*, *Osa*: *Oryza sativa*.

in different organisms as a result of a speciation event, while (ii) paralogs are related sequences within a single organisms as a result of a gene duplication event. They share similar sequence characteristics, and often have a similar structure and function. Different molecular events can alter the number of gene family members in a given species: (1) gene duplication increases gene family size, (2), gene deletion decreases family size, and (3) creation of new genes creates new gene families. As gene duplication increases the gene copy number, initial functional redundancy can lead to sub-functionalization, neo-functionalization or non-functionalization. These changes may contribute to the distinguishable characteristics that differentiate a species from descendants of the same common ancestor.

Current draft genome sequence and gene annotation set in perennial ryegrass

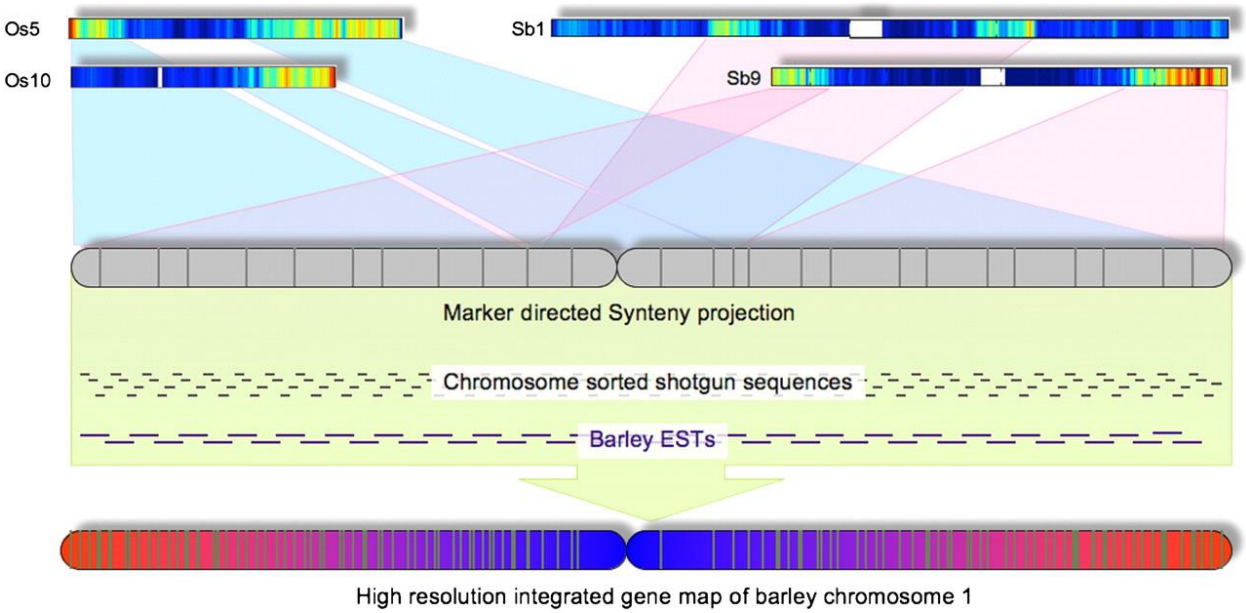
As many other grass genomes, the perennial ryegrass genome is large and complex: the total genome size was estimated to be 2068 Mbp, and 76% of the genome was estimated to be repetitive (Byrne et al., 2015). Sequencing and assembling the genome is therefore a challenging task. As low copy number regions tend to assemble well, a gene-space assembly is much easier to obtain and will contain the bulk of gene-coding regions. Multiple draft genome sequences of the ryegrass genome exist (Mollison et al., 2016), but the synteny-based draft genome published by Byrne et al. (2015) has currently the highest quality in terms of completeness and contiguity. This assembly was generated using *de novo* shotgun assembly of Illumina reads and a nine-fold coverage PacBio sequence reads to fill the gaps. This resulted in an assembly of 1128 Mbp consisting of 48,128 scaffolds with a minimum length of 1 Kbp and an N50 of 70.1 Kbp. The completeness of the gene space assembled in the backbone genome sequence was estimated using CEGMA: 239 out of 248 CEGMA proteins were present at partial or complete level (96.4%), suggesting that a very large fraction of the gene space was covered. The gene annotation set contained 28,455 genes and was generated with a conservative evidence-based approach using the MAKER2 annotation pipeline (Holt and Yandell, 2011). There were on average 2.1 genes per scaffold as the genes originated from 13,725 scaffolds, accounting for 796 Mbp of the assembly (70.6%).

As most ongoing studies are based on these central genomic resources (reference genome sequence and gene set), we analyzed if their quality is essentially fit-for-purpose and investigated how these resources could be improved.

Highly fragmented and incomplete reference genome sequence cannot be used to study genome evolution

The draft genome sequence of perennial ryegrass represents only half of the total genome size, and the assembly is fragmented in 48k scaffolds. This means that both completeness and contiguity need to improve to study genome evolution and species-specific biology through comparison of gene order and gene content with closely related species such as *Brachypodium distachyon* and barley (*Hordeum vulgare*). Detailed studies of gene order sheds light on genome evolutionary processes and is the basis for transfer of QTL information between species, one of the practical applications of a genome assembly.

Genetic mapping experiments have established a remarkable conservation of gene content and order in the Poaceae family , although genome sizes vary as much as 40-fold between some of



the species, and despite the fact that they diverged as long as 60 million years ago (Moore et al.,

Figure 1.6 Schematic representation of the GenomeZipper approach to obtain a gene map for barley chromosome 1H. Genetically anchored barley markers have been integrated with rice and sorghum genes located in syntenic regions to give an enriched tentative ancestral gene scaffold. WCA1H sequence reads as well as barley EST sequences have been associated with this chromosome matrix and give rise to an ordered integrated gene map of barley chromosome 1H. Source: Mayer et al. (2009)

1995). The GenomeZipper uses a reverse engineering approach to obtain an ordered gene map exploiting synteny with closely related grass species rice, sorghum and *B. distachyon* (Mayer et al., 2009) (Figure 1.6). Even before the draft genome sequence became available, the GenomeZipper had been implemented in perennial ryegrass and was proven useful for map-based cloning and QTL fine mapping (Brazauskas et al., 2013; Pfeifer et al., 2013; Arojju et al., 2016).

The GenomeZipper was also used to project 13,411 scaffolds and 11,311 genes onto a putative chromosomal position and led to a synteny-based linear scaffold order of the perennial ryegrass draft genome sequence (Byrne et al., 2015). However, any research aimed at investigating genome evolution, gene order and gene content must be based on a primary assembly that is essentially independent from any synteny-based assumptions.

Are all genes present and annotated in the draft genome assembly?

CEGMA was used to estimate the completeness of the gene space in draft genome sequence (Byrne et al., 2015). The resulting completeness score of 96.4% suggests that a very large fraction of the gene space was covered. CEGMA is based on the presence of 248 single copy, core eukaryotic genes (Parra et al., 2007), a small set compared to the total number of genes in a plant genome. Furthermore, this core eukaryotic gene set does not account for plant-specific genes, neither for the fact that many plant genes are duplicated. Gene annotation of the draft genome assembly was purely evidence-based, implicating that fast-evolving genes and genes with specific expression behavior are most likely to be missed.

Only by improving both the reference genome sequence, and the gene annotation set, it will become possible to dissect the perennial ryegrass genome, study gene and genome evolution, and define the species-specific biology in comparison with closely related species such as barley and other grasses of the Festuca-Lolium complex.

1.3 *De novo* variant discovery by short-read resequencing

Figure 1.7 shows the relation between the reference genome and genomic diversity. Sequencing cost has decreased drastically with the development of NGS technologies, and enables genome resequencing on a large scale. Unbiased discovery of *a priori* unknown genomic sequence

variants relies on *de novo* resequencing of individuals and aligning short read sequencing data to a common central reference genome sequence, followed by variant calling. The resulting catalog of genomic sequence variants provides insight in the genome sequence diversity that is present within a species.

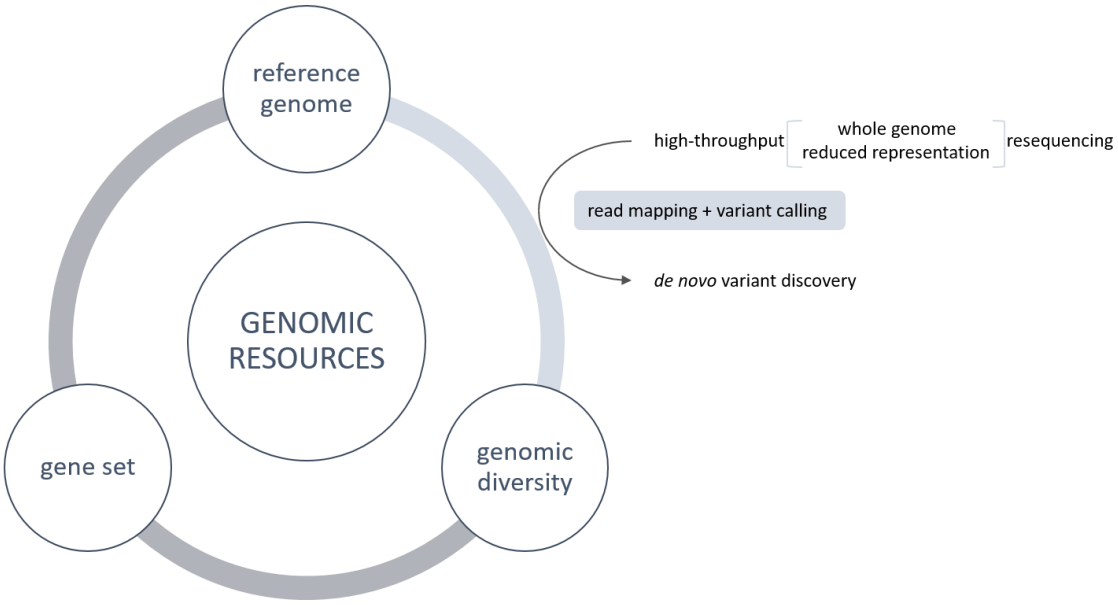


Figure 1.7 Overview of genomic resources and applications related to the identification of genomic sequence variation.

Access to genomic diversity through population-scale resequencing

After the first reference genome sequence had been established for various crops and model organisms, the next logical consequence of the dramatic reduction in sequencing costs using NGS technologies was to sequence large numbers of additional individuals of the same species. A catalog of *A. thaliana* whole-genome sequence variation was generated within the 1,001 Genomes Project by resequencing at least 1,001 strains (Weigel and Mott, 2009a; Cao et al., 2011). The 3,000 Rice Genomes Project is a similar project, on an even larger scale (The 3,000 Rice Genomes Project, 2014). Resequencing hundreds or thousands of individuals of the same species has revealed a huge genomic diversity, indicating that a single reference genome is not sufficient as a representation of the species.

Genomic sequence variants

Genomic sequence variation comprises the differences in chromosomal DNA sequence. The major sources of genomic variation are mutation and reshuffling of mutations through recombination. A mutation is a permanent change in DNA sequence, occurring during DNA replication or induced by external factors, such as chemicals and radiation. Mutations can be neutral (no effect), deleterious (harmful) or beneficial. Genomic sequence variation is therefore the basis of phenotypic variation and adaptation, and the ultimate driving force of evolutionary change.

Genomic sequence variants range from small-scale to large-scale variants, depending on the length of the affected sequence. There are two types of small-scale genomic sequence variants. The most abundant type is a single nucleotide polymorphism (SNP), which is the substitution of a single nucleotide. Insertions and deletions (indels) affect multiple nucleotides, as a stretch of up to several hundreds of nucleotides is inserted into or deleted from the genome. Structural variations describe genomic sequence variation on a large scale, and comprises copy number variations (CNV), large-scale indels and chromosomal rearrangements, such as translocations. These variants are likely to disrupt local gene order (micro-synteny), or affect substantial sections of chromosomal arms (macro-synteny).

Using short-read data to identify small-scale genomic sequence variation

While the sequencing cost continues to decrease, and the volume of resequencing data increases accordingly, the analysis and interpretation of large-scale sequencing data remains challenging and is the main bottleneck to valorization of NGS data. Variant calling (VC) is a bioinformatics method to identify variants from DNA sequence data. VC is a multistep procedure, organized in two phases (Figure 1.8): (1) reads are aligned to determine their corresponding location on the reference genome sequence, and (2) variants are identified by locally comparing the read sequence to the reference sequence. Alleles are different versions that occur at the same variant position. False positive variants are introduced by mistakes during read alignment, and by considering sequencing errors as a sequence variant. The latter is challenging, as true alleles should still be detected in regions with low read depth. During the last decade, many read

aligners and variant callers have been developed and composed into diverse pipelines for high-quality variant identification (Liu et al., 2013).

Figure 1.8 shows a general overview of a common VC pipeline. The first part involves pre-processing of raw NGS reads to prepare read alignment files. This involves mapping reads to the reference genome and indel realignment. Re-aligning reads near detected indels improves identification of indels, as indels can lead to incorrect sequence read mapping, causing false negative and positive SNP and indel calls (Alkan et al., 2011). Sometimes an additional step is necessary to mark PCR duplicates. DNA fragments can be PCR-amplified during library preparation. PCR duplicates arise from multiple PCR products derived from the same template molecule binding on the flow cell. These duplicates need to be removed, as they can lead to false positive variant calls (Ebbert et al., 2016). The second part involves the actual variant discovery. Most VC tools are able to identify variants in multiple samples simultaneously. A VC identifies the variant sites and then assigns a genotype call for each variant site to each sample. This produces a raw variant set that still needs post-processing to filter out false positive variant positions and to obtain a high-quality variant set. Variant calling is a multistep procedure, and each step is associated with its own biases and uncertainties. First, identification of variants from NGS data is challenging because of the base call errors during sequencing (0.1-10%). Variant callers incorporate statistical methods to model the read errors and to identify real differences between

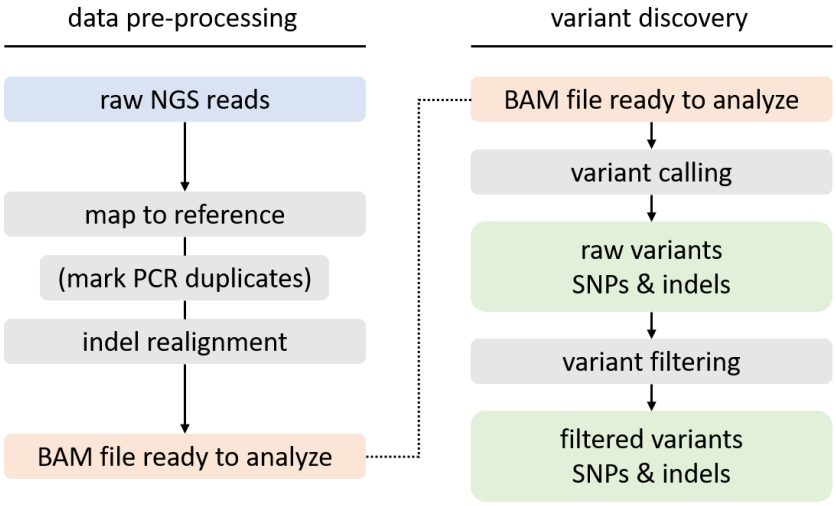


Figure 1.8 General workflow for variant calling, starting from raw NGS data to obtain a set of variants and corresponding genotype calls.

reads and the reference genome sequence. Second, there are biases related to read mapping: the accuracy of read mapping can vary significantly with the read quality, alignment errors frequently occur in regions with small indels and reads can fail to align to regions of high divergence (Bertels et al., 2014). Third, different variant callers have different biases towards specific types of genomic sequence variants, and towards the reference or alternative allele (Hwang et al., 2015). Finally, filtering variants to obtain a high quality variant set is not straightforward. The VCF format is the standard format to report variant positions and corresponding genotype calls for each sample. Commonly used VC pipelines all employ this format to report the final variant set, but all slightly differ on the quality score annotations of the variants. As a consequence, there is no general variant filtering protocol and some callers come with their own filtering parameters and criteria (Li, 2014).

Challenges and current strategies for variant discovery in perennial ryegrass

Because the perennial ryegrass genome is large and consist of 76% repetitive sequences, reduced representation libraries are often used for *de novo* variant discovery and genotype calling in large collections of samples.

Before a reference genome sequence became available, transcriptome sequencing (RNA-Seq) was the most efficient strategy to get access to genomic sequence variation in transcribed genomic regions. However, transcript assembly is a major challenge, because of the high levels of heterozygosity. The large amount of polymorphisms hamper *De Bruijn* graph assembly, causing transcript fragmentation and redundant assembly of allelic contigs and has led to the development of new strategies to create a reference transcriptome (Ruttink et al., 2013). Despite the challenges in transcript profiling and assembly, RNA-Seq has become an important method for high-throughput discovery of gene-associated SNPs that are important for genetic analyses and genome-assisted breeding approaches in perennial ryegrass (Studer et al., 2012; Ruttink et al., 2013; Farrell et al., 2014; Shinozuka et al., 2017).

Genotyping-by-Sequencing (GBS) is a fast and robust approach for reduced-representation sequencing and genome-wide SNP discovery, by sequencing the ends of genomic fragments obtained by digestion with restriction enzymes. Repetitive regions can be avoided by using

methylation sensitive restriction enzymes, thereby simplifying bioinformatics analysis and improving the accuracy of SNP calling (Gore et al., 2009). GBS is increasingly being used in perennial ryegrass for SNP discovery (Byrne et al., 2013) to characterize the genetic diversity, for genetic linkage mapping and QTL analysis (Hegarty et al., 2013) and genome-wide association studies (GWAS) (Kopecký and Studer, 2014). There is, however, no *a priori* control over which genes are tagged as only one-third of the GBS tags are be located in the gene space. Moreover, GBS tags typically span only 100-300 bp, and with about 60,000 loci, GBS covers about 0.1% of the genome.

Identification of genomic sequence variants is associated with specific challenges in variant calling, as the presence of many SNPs and small indels hampers read alignment, and distinguishing true variants from read errors is more difficult at low depth regions. Moreover, different VC pipelines are reported to produce low concordant variant sets (O'Rawe et al., 2013) and there is no high quality, reference variant set for perennial ryegrass available to optimize VC parameters and variant filtering.

1.4 Identification of interesting genomic sequence variants

A genotype is the genetic make-up of an individual, while a phenotype describes all the physical traits and characteristics of an individual and is the combined result of the expression of the genotype and the interaction between genotype and environment. Genomic sequence variants may have an effect on gene function and/or regulation and could therefore result in phenotypic variation. The detection and exploitation of genomic variation that is responsible for variation in a phenotypic trait is of great interest for plant breeders, as these are heritable characteristics that can be selected for.

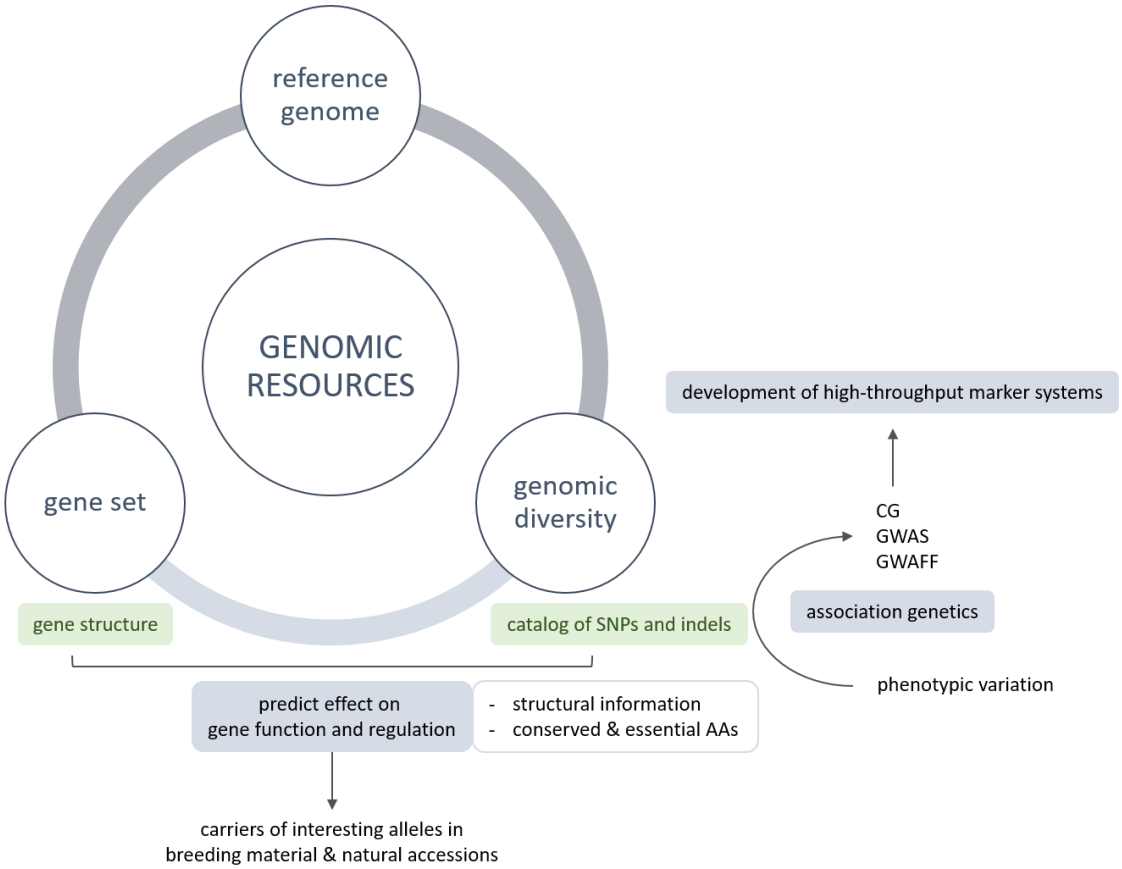


Figure 1.9 Overview of genomic resources and applications related to genomic sequence variation. AA: amino acid, CG: candidate gene, GWAS: genome-wide association study, GWAFF: genome-wide allele frequency fingerprint.

There are two main strategies to detect the most interesting variants from a large collection of genomic sequence variants (Figure 1.9). The effect of sequence variation on gene function and regulation can be computationally predicted using gene structure information, combined with positions of conserved and/or essential amino acids. This allows for the identification of carriers of interesting alleles in breeding populations and natural accessions that can be used for further functional analysis, or can be incorporated into the breeding program. Alternatively, association genetics studies aim to statistically associate phenotypic variation with genotypic variation. This is a useful forward genetics approach for the dissection of quantitative and complex traits that are regulated by multiple genes.

Variant effect prediction based on structural sequence features

Predicting the effect of genomic sequence variants using gene structure information is a reverse genetics approach, as it first identifies mutants based on DNA variants and then studies the

associated phenotype. The effect of a small-scale genomic sequence variant (SNP or indel) depends on its position with respect to coding genes:

- If a variant occurs within a coding region, it is categorized based on the possible effects on the resulting protein sequence. A variant can affect a codon in different ways. Because the genetic code is redundant, not all changes in nucleotide sequence result in the change of an amino acid. These nucleotide changes are called synonymous substitutions. There are three types of non-synonymous variants, i.e. variants affecting the encoded amino acid. If a coding codon is converted into a stop codon, the variant is classified as a nonsense variant. Conversely, a stop codon can be turned into a coding codon, resulting in a longer protein. If the variant results in a change in amino acid, the variant is called missense. The effect on protein function or 3D structure can be predicted based on the evolutionary conservation level and the chemical differences between the amino acids (Flanagan et al., 2010). Indels with a length divisible by three in coding regions will cause insertions and deletions of amino acids into the protein, but may also result in a nonsense variant. If the length of the indel is not divisible by three, this will cause a frame shift where all codons downstream are shifted to an other reading frame. This often results in disrupted protein domains, a malformed protein or nonsense-mediated decay, where the mRNA is eliminated because it contains a premature stop codon.
- If a variant is not located in a coding region, i.e. noncoding variants, the effects are more difficult to predict computationally. These variants can disrupt the DNA sequence motifs that are located in promoters and enhancers, such as transcription factor binding sites. This may create new motifs or disrupt an existing motif, or influence affinity for transcription factor binding, thereby affecting gene regulation. Variants located in introns may affect splicing, resulting in gain or loss of splice donor and acceptor sites. In addition, variants located in 5'UTR, introns, or 3'UTR may affect folding or stability of the mRNA, thus affecting post-transcriptional regulation.

Association genetics studies link phenotypic variation to genotypic variation

Combining high-throughput genotypic and phenotypic data has enabled large-scale marker-trait association analysis to dissect the genetic architecture of plant traits. Association genetics is a forward genetics approach to identify interesting variants in large collections of genomic sequence variation that explains a particular phenotype. Association mapping (AM), also known as linkage disequilibrium (LD) mapping, is a commonly used method to link phenotypic variation to genotypic variation by exploiting historical recombination events at the population level. This is especially useful for the dissection of complex traits that are quantitative and may be controlled by multiple genes. AM has been introduced in plants in 2001 (Thornsberry et al., 2001) and has since then been applied in a variety of crops (Huang et al., 2010; Tadesse et al., 2015; Gyawali et al., 2017). AM studies can be divided into two categories: (i) candidate gene AM, which relates variants in selected genes that are involved in biological, physiological or morphological processes related to phenotypic traits of interest, and (ii) genome-wide AM, to find signals of association for various complex traits starting from genome-wide genotyping data.

GWAS have now been carried out successfully in many crops, to unravel the genetic architecture of agronomically relevant traits and identify candidate loci for subsequent validation. Rice and maize are two major models for crop GWAS, as thousands of inbred lines have been genotyped and multiple trials have been conducted for several traits (Huang and Han, 2014). GWAS have been successfully extended to genetic studies in other crops, such as barley, sorghum and wheat. Nevertheless, sequence-based GWAS is still rare in wheat as most studies use more affordable technologies such as diversity arrays technology (DArT) (Jia et al., 2018).

AM studies in perennial ryegrass are often candidate gene based, because of the rapid LD decay in allogamous species (Skot et al., 2005; Skot et al., 2007; Asp et al., 2011; Yu et al., 2015; van Parijs et al., 2016). Additionally, GBS is increasingly being used in Genome-Wide Association Studies (GWAS) (Kopecký and Studer, 2014). Although GBS is mainly used for genotyping single individuals, it is more and more used for Genome-Wide Allele Frequency Fingerprinting of perennial ryegrass populations (GWAFF) (Byrne et al., 2013; Fe et al., 2016). The output of an AM study may have a dual purpose: (i) the identification of an associating variant can lead to a

candidate gene involved in the regulation of expression of a trait of interest, providing insight in the underlying molecular mechanism, and (ii) variants linked to phenotype can be used as molecular markers for marker-assisted selection.

The conversion of genomic sequence variants into genetic markers allows for cost-effective marker-assisted selection strategies and genome-wide fingerprinting. SNPs have become the most popular markers in breeding programs because of their abundance, stability over generations and the ability to detect them with high-throughput genotyping methods. Indels are often not taken into account because of their low confidence identification with short reads. By developing more accurate sequencing technologies or better variant calling algorithms, indel variation could be used more often if indels could be identified as easy and accurate as SNPs.

Molecular markers can be integrated in a breeding program using marker-assisted selection. This works fine for simple traits, by selecting individuals with QTL-associated markers that have major effects, and not significantly associated markers are neglected. Improving complex quantitative traits using QTL-associated marker detection has been unsuccessful due to the difficulty of finding the same QTL across multiple environments (due to QTL by environment interactions) or in different genetic backgrounds. Additionally, the main disadvantage of AM is the low power of detecting rare variants that may be associated with relevant traits (Crossa et al., 2017). In contrast, genomic selection (GS) aims to predict breeding and/or genetic values, by combining all molecular markers and phenotypic data in a training population to obtain genomic estimated breeding values of individuals in a testing population that have been genotyped but not phenotyped (Meuwissen et al., 2001). GS is better suited to improve complex traits with low heritability as it is typically based on models incorporating information from all available markers, while significantly reducing the cost per breeding cycle and time required to develop a new variety. GS often results in a good prediction at the expense of low interpretability, thereby generating a trade-off between model interpretability and model complexity (Gianola and van Kaam, 2008). The implementation of GS in perennial ryegrass breeding has been tested, and showed good perspectives. Problems resulting from low LD can be reduced by the intentional inclusion of structure and related families in the breeding population (Fè et al., 2015; Faville et al., 2016). GS in perennial ryegrass is gaining interest, and has already successfully been

performed for key agronomic traits, such as heading date, rust resistance and seasonal biomass yield (Guo et al., 2018; Pembleton et al., 2018).

1.5 Genomic resources in related crop species

This final paragraph of the introduction provides an overview of the development and usage of different genomic resources in related crop species, such as rice, maize, and barley and wheat, which are closely related to perennial ryegrass. It provides a perspective on the current status of genomics in these crops, as well as the opportunities that genomic resources provide to study the biology of a crop in relation to its evolution, diversity and biology.

Rice (*Oryza sativa*) is one of the most important crops in the world. It is the smallest of the major cereal crop genomes and estimated genome size is 400 to 430 Mb. It was the first sequenced crop genome (International Rice Genome Sequencing and Sasaki, 2005), following the publication of the *A. thaliana* genome in 2000. Because of the remarkable conservation of gene content and order of grass species (Poaceae) (Gale and Devos, 1998), rice has served as a model for other crops with larger genomes, such as maize and wheat. Population-scale genomic studies have unraveled introgression of genes during domestication and diversification of Asian rice (Sweeney and McCouch, 2007; Choi et al., 2017). More recently, the 3000 Rice Genomes Project resequenced more than 3,000 rice accessions from 89 countries. A pan-genome analysis discovered 29 million SNPs, 2.4 million short indels, over 90,000 structural variants and more than 10,000 novel genes. Phylogenetic analysis has revealed five varietal groups (Wang et al., 2018). These resources serve as the foundation for the discovery of novel alleles through bioinformatics and/or genetic approaches, as well as to understand the genomic diversity within *O. sativa*. It is a great help for advancing rice breeding technology for future rice improvement (The 3,000 Rice Genomes Project, 2014).

The maize genome is much larger (2.4 Gbp) and more complex compared to the rice genome, creating new challenges for assembling a reference genome sequence. The first maize reference sequence was based on Sanger sequencing and published in 2009 (Schnable et al., 2009). Although this assembly consisted of 100k contigs that were arbitrarily ordered and oriented, it has enabled a rapid progress in maize genomics (Edwards et al., 2013). The most recent version

of the maize reference genome was obtained using third-generation sequencing and optical mapping, enabling the characterization of the repetitive portion of the genome and the identification of lineage-specific expansions (Jiao et al., 2017). As maize is one of the most diverse crops in the world and breeders have exploited this genetic diversity to create the highest yielding grain crop in the world (Whitt et al., 2002). The maize HapMap project is a large-scale resequencing project, covering pre-domestication and domesticated maize varieties from all over the world, and has revealed a tremendous amount of genetic variation (Gore et al., 2009; Chia et al., 2012; Bukowski et al., 2018). Structural variations were found to be enriched at loci associated with important traits. The identification of small sequence variants required an entirely new computational pipeline to resolve genotyping errors derived from incorrect mapping of short reads. Additionally, the B73 genome reference represents 91% of the genome and captures only 70% of the low-copy gene fraction of all maize inbred lines (Gore et al., 2009). Mapping reads from each individual onto a common reference genome to identify sequence variations resulted in incorrect mapping, either to the paralogous loci or highly repetitive regions (Bukowski et al., 2018). This limits the use of a genome from a single individual as a reference, and this has become a limiting factor to study the genetic diversity within a species.

The innovation in NGS technologies and third-generation sequencing methods has greatly accelerated the assembly of plant genomes. However, large genome sizes, high repetitiveness due to transposable elements are characteristic for some major crops, thereby displaying unique challenges for genome assembly (Schatz et al., 2012). Good examples are barely (5 Gbp), rye (7.9 Gbp) and wheat (17 Gbp), all members of the Triticeae family and closely related to perennial ryegrass. First attempts to assemble these genomes resulted in draft genome sequences that are highly fragmented, and do not represent the full genome size because of the repetitive content (The International Barley Genome Sequencing et al., 2012; Jia et al., 2013; Ling et al., 2013; Mayer et al., 2014; Bauer et al., 2017). Nevertheless, genomic resources have been successfully developed for barley and wheat, and have been applied in GWAS and marker-assisted breeding. Examples are high-density SNP arrays that allow genotyping large populations (Cavanagh et al., 2013; Bayer et al., 2017), and *de novo* genotyping GBS-based platforms that can be used

regardless of prior knowledge of genomics, genome size, organization or ploidy (Poland et al., 2012).

2 Thesis Outline & Objectives

Innovation in genomics has already proven its strength for many major crops, such as rice, maize, barley and wheat. High-throughput sequencing methods and bioinformatics are instrumental to characterize plant genomes and genetic diversity, and ultimately to accelerate crop improvement. In this PhD thesis, I used and developed bioinformatics methods to generate and improve three genomic resources for perennial ryegrass (*L. perenne* L.): (i) a reference genome sequence, (ii) a complete structurally and functionally annotated gene set, and (iii) a comprehensive overview of the genome sequence diversity within the genepool. The experimental chapters of this thesis each deal with objectives related to these genomic resources.

Objective 1 – Defining completeness measures for plant genome projects and application for perennial ryegrass.

Completeness, or rather incompleteness, of a reference genome sequence and gene annotation set has a great impact on further downstream analyses, such as comparative genome analysis to study genome architecture and evolution. Despite the strong increase in the number of genome sequences, we found that no clearly defined measures for genome and gene space completeness existed at the start of this PhD. As completeness is expressed as the ratio of the 'observed' and the 'expected', the determination of what is expected in terms of genome size or gene space is fundamental in the definition of a completeness measure. In **Chapter 3,** we investigated how completeness should be estimated in plant genome projects, for both genome assemblies and gene annotation sets, and formulated guidelines for future plant genome projects. Hence, we estimated the completeness of the draft genome sequence of *L. perenne* and the corresponding gene annotation set, to assess whether the gene space was fully assembled and annotated.

Objective 2 – Developing a strategy to generate a complete and reliable catalog of sequence variation for L. perenne from Illumina short read sequencing data.

The identification of genomic sequence variation in perennial ryegrass is challenging because of three reasons: (i) the genome of perennial ryegrass individuals is highly heterozygous because of

the outbreeding nature of this species, (ii) the high levels of polymorphisms in transcribed regions (Ruttink et al., 2013; Farrell et al., 2014; Paina et al., 2014) can interfere with accurate read mapping, and (iii) commonly used VC pipelines generate low concordant variant sets. There is currently no straightforward way to *de novo* identify genomic sequence variants from short read data. Moreover, at the start of this PhD, there was no high quality variant set available that was representative for the perennial ryegrass genepool that could be used to calibrate VC parameters and quality control. Our objective was to develop a strategy to generate a complete and reliable catalog of sequence variation from Illumina short read data for *L. perenne* (**Chapter 4**). For this, we resequenced 503 candidate genes putatively involved in plant growth and development in a collection of 736 individuals derived from natural accessions, breeding populations and current cultivars. This strategy is broadly applicable to other highly diverse outbreeding species and provides important insights in the pitfalls and solutions of bioinformatics analysis of population-scale genomic resequencing studies. The resulting catalog of genomic sequence variation is currently the most comprehensive gene-anchored variant set in *L. perenne* that is functionally annotated and can be mined for interesting variants, either through reverse or forward genetics approaches. In the future, this catalog can be used for association genetics studies with phenotypic traits related to plant architecture and cell wall digestibility.

Objective 3 – Mining interesting sequence variants in the L. perenne genepool as candidates for a reverse genetics approach.

In outbreeding species, defective alleles occur in natural populations at low frequency and usually occur in a recessive heterozygous state (Marroni et al., 2011). Identification of carriers of these rare defective alleles allows the validation of causal polymorphisms and functional gene analysis by generation of lines that are homozygous through dedicated crosses. The catalog of genomic sequence variation is a rich resource to mine for rare defective alleles, given an appropriate detection method. An association genetics approach is typically insensitive to detect the effect of alleles with low frequency in the genotype collection, as the number of replicate observations is too low, leading to a decreased statistical power. In contrast, the effect of SNPs and indels on protein function or activity can be computationally predicted. We first tested this in a straightforward approach relying on the genetic code for protein translation, and screened

for high-impact effects, such as frameshifts and premature stop codons (**Chapter 4**). A complementary, more advanced approach relies on evolutionary conservation of amino acid residues, or experimental evidence of amino acids critical for protein function or activity obtained from orthologs in related species. This approach was the objective of **Chapter 5**, where we demonstrate how to mine for sequence variants at critical amino acid residues for the well-studied FLOWERING LOCUS T gene family.

Objective 4 – Screening breeding populations for associations with two important agronomic traits.

Flowering time and leaf elongation are regulated by multiple genes, and are important agronomic traits for perennial ryegrass breeding. As genetic variation is the foundation of phenotypic variation, we want to test whether variations in heading date and leaf length observed in a population can be explained by the genotypic variation in candidate genes controlling these traits. In **Chapter 6**, a candidate gene association mapping study was performed in five breeding populations and a natural accession, to detect alleles associating with heading date and leaf length. A high-throughput genotyping assay for 28 candidate genes was developed, based on the catalog of genomic sequence variation generated in Chapter 4, to accurately determine the alleles present in each genotype. Given a sufficient phenotypic range within a population, accurate genotyping, and statistical power through a balanced representation of genotypic classes (i.e. presence of different alleles), we want to identify alleles responsible for a difference in phenotype, and to determine their effect on the traits of interest. This approach is a forward genetics approach, and contrasting to the approach of Objective 3. Moreover, this is an example of how we can exploit the catalog of genomic sequence variation to design a targeted, high-throughput genotyping assay.

Objective 5 – Improving the annotation of the L. perenne gene space, as well as generating a chromosome-scale reference genome sequence.

As closely related species are derived from a common ancestral species, their genomes contain a similar gene content and gene order. Species-specific genomic changes, such as chromosomal rearrangements, gene duplication and gene loss, that occurred since a speciation event, may have led to species-specific biology. The detection of these signals through comparative genomics requires complete and contiguous reference sequences and corresponding annotation sets for all comparator species. The draft genome sequence of *L. perenne* represents half of the genome size and covers the full gene space (Byrne et al., 2015), but gene annotation was incomplete (see Chapter 3). In **Chapter 7**, we present currently unpublished results on the improvement of this gene annotation set, as well as a functional annotation. This improves the value of the draft genome sequence, as it will be better suited for the identification of candidate genes and the study of the protein coding capacity. However, the draft genome sequence is highly fragmented, incomplete, and anchored to pseudo-chromosomes through projection of synteny with genomes of related species (Byrne et al., 2015). It is not suited for the study of genome architecture and evolution, as the chromosomal assembly itself is not independent from any synteny-based assumptions. **Chapter 7** presents a novel complete chromosome-scale draft genome sequence for perennial ryegrass, making use of the most recent advances in sequencing and assembly of complex plant genomes, namely the combination of long-read sequencing, optical mapping and chromosomal conformation capture (Hi-C) sequencing. This reference sequence, together with a corresponding annotation set, is an important advance for genomics in perennial ryegrass. Finally, this chapter covers a general overview and discussion of the genomic resources developed, and their further applications and value for the community.

3 Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences[1]

Genome sequencing is becoming cheaper and faster thanks to the introduction of next-generation sequencing techniques, leading to a considerable increase in the number of available genome sequences. In the last few years, dozens of new plant genome sequences have been released, ranging from small to gigantic repeat-rich, or polyploid genomes. Most genome projects have a dual purpose: delivering a contiguous and complete genome assembly, together with a full catalog of correctly predicted genes. Frequently, the completeness of a species' gene catalog is measured using a set of marker genes that are expected to be present. It is vital to understand that this expectation can be defined along an evolutionary gradient, ranging from highly conserved genes to species-specific genes. Furthermore, large-scale population resequencing studies have revealed that gene space is fairly variable even between closely related individuals. This clearly limits the definition of the 'expected' gene space, and, consequently, the accuracy of different estimates used to assess genome and gene space completeness. We argue that, based on the desired applications of a genome sequencing project, different completeness scores for the genome assembly and/or gene space should be determined. In addition, based on examples from recent literature comprising several dicot and monocot genomes, we outline some pitfalls and recommendations as to which methods are most suitable to estimate the completeness during the subsequent steps of genome assembly and annotation.

[1] This chapter is based on **Veeckman, E., Ruttink, T., and Vandepoele, K.** (2016). Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. Plant Cell **28,** 1759-1768. Author contribution, see page 49.

3.1 Introduction

The ever-decreasing cost together with the expanding capacity of genome sequencing using NGS techniques leads to a fast increase in the number of available genome sequences. As of 2016, over 100 plant genomes have been sequenced, ranging from small (e.g. *Utricularia gibba*, 80 Mbp) to gigantic, repeat-rich, or polyploid genomes (e.g. *Triticum aestivum*, 17 Gbp), with many more expected in the years to come (Weigel and Mott, 2009b; Chia et al., 2012; Michael and Jackson, 2013; Li et al., 2014). Ideally, a genome assembly represents a complete and contiguous genome sequence with a cumulative scaffold length equal to the haploid genome size (Figure 3.1, box A). In addition, a complete set of annotated genes offers a starting point for a detailed characterization of gene functions, biochemical and regulatory pathways, or QTLs. Genes are the nodes in a biological network, which offers valuable insights into protein complexes, regulatory

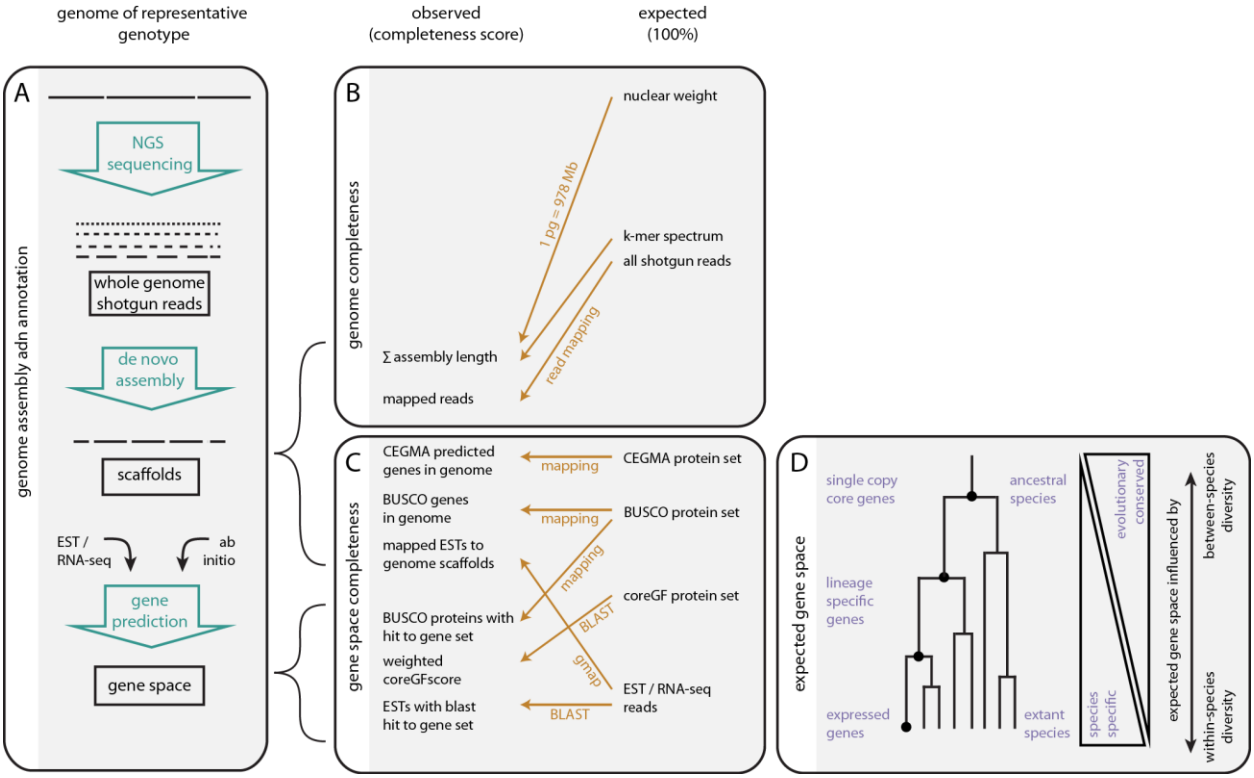


Figure 3.1 Framework for genome assembly and gene space completeness estimation. Box A shows the workflow for genome assembly and annotation. A representative genotype is selected for sequencing and the whole-genome shotgun reads are assembled into incrementally longer contiguous scaffolds. In a final step, gene prediction provides the description of repetitive regions and the annotation of genes. Boxes B and C represent the estimation of genome assembly and gene space completeness, respectively. Measures for the expected and observed size of the genome assembly and gene space are shown, connected by specific methods. Box D shows how the expected gene space can be estimated along an evolutionary scale, ranging from evolutionarily highly conserved to species-specific genes.

interactions and metabolic processes that determine the physiological and biochemical properties of a cell, an organ or an organism (Bassel et al., 2012).

Clearly, comparative genomics and evolutionary studies thus require complete genomes and gene sets. Well-assembled genome sequences are necessary to characterize different classes of repetitive elements, to identify large-scale gene co-linearity across related species, and to reconstruct the organization and evolution of transposable elements (Bennetzen and Wang, 2014). Moreover, a complete catalog is required to test if the gain or loss of biochemical or signaling pathways in specific plant species can explain the structural and physiological adaptations required to survive in extreme environments. The absence of specific genes in the genome, and not just the assembly, should be independently confirmed using e.g. de novo assembled transcripts (Olsen et al., 2016) or hybridization-based molecular techniques.

The N50 is a commonly used contiguity measure denoting that 50% of the total assembly length is contained in scaffolds of length N50 or longer. Over the last 15 years, genome assemblies display a large range of N50 values and indicate low contiguity even for relatively small genomes (Supplemental Figure 1), suggesting that fragmented draft genomes are generated for many plants. As this wealth of new plant genome sequences and gene catalogs expands, so does the variety of methods to measure their quality and completeness (Earl et al., 2011; Salzberg et al., 2012). Consequently, there are no uniform metrics or standards yet in place to estimate the completeness of a genome assembly or the annotated gene space, despite their importance for downstream analyses.

3.2 Material and Methods

Data

In total, twelve species including rosids and monocots were used to compare gene space completeness measures. Based on an initial list of 18 studies that used CEGMA to assess the completeness of a sequencing project within flowering plants, assembled sequence information could be retrieved for ten species. These datasets covered seven rosid species (*C. rubbella, C. arietinum L., N. nucifera Gaertn., P. veris, P. communics L.* 'Bartlett'*, R. raphanistrum, V. angularis*) and three monocots (*L. perenne*, *P. equestris*, *S. italica*). *A. thaliana* and *O. sativa* were also

included as the oldest, high-quality reference genomes, which were sequenced using a gold-standard BAC-clone based approach and are thoroughly expert curated (Supplemental Table 1).

Core Eukaryotic Genes Mapping Approach (CEGMA)

The CEGMA completeness score reports the number of conserved eukaryotic genes that could be found in the genome assembly using an accurate mapping protocol (Figure 3.1 C). Partial and complete CEGMA scores refer to the presence of a gene fragment or a complete copy, respectively. For the ten species included in the comparison, the complete CEGMA score was extracted from the corresponding genome paper. The CEGMA score of *A. thaliana* is equal to 100%, as this species was one of the six eukaryotic species used to define the CEGMA core gene set. As CEGMA has been frequently used in plant genome sequencing projects, we do not discuss BUSCO in great detail, although most reported features hold for both methods. Whereas CEGMA only works on raw genome or transcript sequences and performs gene prediction prior to the completeness estimation, BUSCO can be applied on a genome sequence as well as on an annotated gene set.

Transcript mapping score

For twelve species, EST sequences were obtained from the NCBI Nucleotide EST database (downloaded on October 12, 2015). The EST sequences were mapped to their respective reference genome using GMAP with default parameters (Wu and Watanabe, 2005). We collected all EST sequences that are publicly available for *A. thaliana* (186 libraries, ranging from 1 to 541,852 ESTs per library, 1,529,700 ESTs in total) and *O. sativa* (220 libraries, ranging from 1 to 53,637 ESTs per library, 987,327 ESTs in total). For *A. thaliana* and *O. sativa*, all ESTs were also mapped on simulated incomplete genomes. To simulate genome fragmentation, the genome was cut into pieces of 10kb and incomplete genomes were constructed by randomly selecting 50%, 75%, 80%, 90%, 95%, and 100% of these fragments. Next, we randomly sampled ESTs, to construct bin sizes containing 100 up to 300,000 ESTs, and for each bin we estimated which fraction was mapped onto the genome. Optionally, an extra filtering step was applied retaining only EST mappings with >90% coverage. For each bin size, the mean and standard deviation of the transcript mapping score was calculated over 100 random replicates per bin size. In a second

approach, each EST was assigned to its original EST library, and the transcript mapping score was calculated per EST library.

CoreGF completeness score

Three sets of PLAZA Core Gene Families (coreGFs) have been defined: green plants, based on conserved genes in 25 species including angiosperms, mosses and green algae (2928 coreGFs); rosids, based on conservation in 12 species (6092 coreGFs); and monocots, based on conservation in 5 species (7076 coreGFs) (Van Bel et al., 2012). A BLAST-based sequence similarity search is applied per set of transcript sequences or predicted proteins to detect the presence of a coreGF, using one representative protein per coreGF. The representation across all individual coreGFs is summarized in a global weighted coreGF score, where large gene families get a smaller weight than single-copy families, as the former have a higher probability to be detected.

Expression and gene function bias of CEGMA and coreGFs

Gene function bias was determined through Gene Ontology enrichment analysis using the PLAZA 3.0 Dicots Workbench (Proost et al., 2015) for the CEGMA core genes and the coreGFs of green plants and rosids. The expression bias was assessed through *A. thaliana* gene expression analysis using the Compendium2 from the CORNET database (De Bodt et al., 2010). Highly similar experiments were removed by clustering the experiments using a 0.95 Pearson Correlation Coefficient threshold and taking into account the sample descriptions available in Gene Expression Omnibus. This resulted in an expression atlas of 75 experiments. Expression bias was determined for all expressed *A. thaliana* genes, the CEGMA core genes and the coreGFs of green plants and rosids. For each gene in these gene sets, we counted the number of experiments in which the gene is expressed (expression value > $2^{7.5}$) and summarized the values in an expression breadth histogram.

3.3 Defining the expected genome size and gene space

The common approach to all reported measures of completeness is simple to grasp (Figure 3.1, box B, C). First, one measures the size of the assembled genome (i.e. total assembly length) or the gene space (i.e. the number of genes), in the following referred to as the ‘observed’. Second, one selects a reference to define the expected genome size or gene space, here referred to as

the ‘expected’. To define the expected genome size, one can use either physical measurements (e.g. nuclear weight) or computational methods that analyze the sequence space (such as k-mer spectra). Furthermore, to define the expected gene space, one can either rely on evolutionary conservation and use the gene space of related species as reference (inter-species comparisons). Alternatively, one can define a species-specific measure of the gene space by transcriptome or Expressed Sequence Tag (EST) sequencing in the species itself (intra-species comparisons; Figure 3.1, box D). Clearly, these methods rely on starkly contrasting assumptions, as further detailed below. Third, the comparison method (e.g. BLAST or read mapping) inherently assumes directionality and sets the external reference as the expected ‘100%’. The observed measure is then expressed as a fraction of the expected, and interpreted as completeness score for the genome assembly or gene space. Given the diversity of approaches, it is important to understand the underlying concepts to provide consistent and realistic measures of genome and gene space completeness.

The genome can be partitioned into two main fractions with contrasting characteristics in terms of assembly and annotation. The repetitive DNA, mostly contained in heterochromatin, is difficult to assemble using short shotgun reads and this partition is commonly collapsed or absent from draft genome assemblies. It generally contains transposable elements and relatively few coding genes. In contrast, the non-repetitive sequence space, mostly contained in euchromatin, is relatively easy to assemble and is commonly assumed to contain the gene-rich partition. It is important to realize that methods to estimate genome or gene space completeness target these partitions of the genome differently, and although completeness scores may seem related, they should not be extrapolated between the two levels.

Here, we will outline the challenges of estimating the completeness of the genome assembly and annotated gene space. We first explain how the ‘expected’ is defined for different measures of completeness and comment on the assumptions made by each method, including their strengths and weaknesses. Next, we will compare different measures of completeness in twelve recently published plant genomes and highlight several cases where dissimilar completeness scores are the consequence of technical issues of assembly or annotation, or due to strong gene function or expression biases in the expected gene space. Finally, we will provide some guidelines to

determine more robust completeness scores and comment on the challenges of future plant genome projects, such as defining the 'expected' gene space in the context of pan genome sequencing.

3.4 Estimating the completeness of a genome assembly

The first step in a genome assembly workflow (Figure 3.1, box A) is selecting an individual that is representative of the species. For this individual, shotgun libraries are constructed with variable insert sizes, ranging from 100 bp to over 100 kb for paired-end, mate-pair, PacBio, Moleculo or BAC libraries. Sequencing will yield reads of variable length, ranging from 100 bp to more than 10 kb, depending on the applied sequencing technology. These reads are then assembled into incrementally longer contiguous sequences in three steps. First, contigs are constructed through *de novo* assembly based on the overlap of short reads or de Bruijn k-mer graphs. Secondly, the contigs are ordered into scaffolds using mate-pairs, BAC-end sequences, or hybrid assembly with long PacBio reads. Finally, the scaffolds are ordered and anchored into pseudomolecules or linkage groups representing chromosomes using optical mapping, cytogenetic mapping, Hi-C sequencing, genetic maps, population sequencing or physical maps such as BAC minimal tiling paths (Mascher et al., 2013; Mendelowitz and Pop, 2014; Flot et al., 2015).

Two main factors affect the completeness and contiguity of the assembly: the level of heterozygosity and the length, abundance, and dispersal of duplicated regions or repetitive sequences (Wendel et al., 2016). Genome assembly algorithms attempt to reconstruct unique sequences in order to separate recently duplicated regions, closely related gene family members, or highly conserved protein domains. As a result, allelic sequences in highly heterozygous species are often also reconstructed as independent sequences, thereby inflating the total assembly length (e.g. *Malus domestica* (Velasco et al., 2010)). Conversely, repeat regions are typically collapsed during assembly of short reads, thereby severely reducing the total assembled genome size and interrupting scaffold contiguity (e.g. *Lolium perenne* (Byrne et al., 2015)). Apart from their effect on the total assembly length, high levels of heterozygosity also reduce contiguity. Highly polymorphic regions disturb sequence alignment during *de novo* assembly, lead to bubbles and branches in de Bruijn graphs, and cause breakpoints when de Bruijn graphs are resolved into

contiguous sequences. Some of these issues may be overcome in the near future using third generation long read sequencing technologies.

The expected genome size of an organism can be measured using the physical properties of the nuclear genome: by reassociation kinetics of high molecular weight genomic DNA (C_ot assay), pulsed field gel electrophoresis or, ideally, flow cytometry after DNA staining. These methods use standards of known molecular weight or reference species with a defined nuclear DNA mass (Zonneveld et al., 2005). The total assembled scaffold length (in Mbp) can then be expressed as a fraction of the molecular weight of the nuclear DNA (in pg), using the standard average molecular weight of 1 pg per 978 Mb for the conversion. Strikingly, closely related species may display considerable variation in genome size, hence limiting the accuracy of inter-species comparative measures of completeness (Garcia et al., 2014). In contrast, flow cytometry-based measurements of genome size turn out to be fairly constant across individuals within a species (Dolezel and Bartos, 2005), thus providing accurate estimates of the expected genome size within that species.

Alternatively, the expected genome size and repetitive sequence content can be estimated using computational methods, such as k-mer frequency spectra of the shotgun sequencing reads (Chor et al., 2009). Furthermore, the percentage of the shotgun reads or BAC-end sequences that map onto the scaffolds yields a genome completeness score that indicates whether the shotgun read sequences have all been incorporated into the scaffolds. The read depth profile may further identify wrongly assembled, collapsed, or duplicated regions (Hunt et al., 2013; Rahman and Pachter, 2013). Conversely, one can control over-assembly by analyzing if all scaffolds are supported by read data. Just as the assembly algorithms are sensitive to genetic diversity and heterozygosity while searching for sequence overlap to build contiguous scaffolds, these assembly completeness methods rely on sequence identity for read mapping, thus, completeness scores are inherently sensitive to mismatch stringency parameters in highly heterozygous genomes (Wendel et al., 2016).

3.5 Estimating the completeness of the annotated gene space

In an ideal scenario, genome annotation describes repetitive regions and the complete set of protein-coding genes and various classes of non-coding RNAs with a correctly identified gene structure. However, gene prediction does become challenging in the absence of extrinsic data, such as EST/RNA-Seq transcript data of the species under investigation or well-characterized proteins from related species. Re-training gene prediction software to learn codon biases or specific splicing motifs is both important to obtain high-quality gene models and to identify species-specific genes lacking homologs in other plant families. Gene prediction benchmarks exist for different eukaryotic model species and automated self-learning gene prediction approaches have been developed (Korf, 2004). However, in the absence of large species-specific transcript databases generic gene prediction tools have been used for several plant genomes, compromising validation of the quality and completeness of the predicted gene catalog. Recently developed methods like MAKER-P and BRAKER1 offer a practical solution for some of these issues, provided that sufficient extrinsic information is available (Campbell et al., 2014; Hoff et al., 2016).

If the N50 is smaller than the average size of a gene, one can expect to annotate many partial gene models due to gene splitting, resulting in an over-estimation of the number of genes in the genome. Clearly, such erroneous gene models will compromise the correct delineation of homologous gene families and orthologous genes, as well as the detection of protein domains This obstructs the interpretation of gene family expansion or gene loss and any other downstream gene-based analysis, such as gene expression quantification through RNA-Seq, annotation of ChIP-Seq binding events or gene network analysis. For several plant genomes, gene catalogs have been created based on incomplete genome assemblies after separating gene-rich regions from repetitive DNA using methyl-filtration or high-C_ot enrichment methods (Rabinowicz et al., 1999; Yuan et al., 2003). However, it is not entirely clear yet to what extent these approaches capture specific gene loci embedded within large repetitive regions.

Defining the expected gene space on a gliding evolutionary scale

One approach to define the expected gene space is to use evolutionarily highly conserved reference gene sets, which are also expected to be present in the newly assembled genome (Figure 3.1, box D). This requires the definition of the taxonomic range over which genes are expected to be conserved, relative to the species under investigation. This approach inherently poses a crucial trade-off between the level of evolutionary conservation and the number of genes in the reference set. We first illustrate this with the Core Eukaryotic Genes Mapping Approach (CEGMA). The CEGMA reference gene set comprises 458 genes that are highly conserved in six eukaryotic species (*Homo sapiens, Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*), and which are assumed to be encoded in essentially all eukaryotic genomes. Notably, CEGMA was originally created to build a robust set of gene annotations to train gene prediction software in the absence of experimental transcriptome data, but it is not meant to provide a complete catalog of genes in a genome. Nevertheless, a subset of 248 single-copy core eukaryotic genes is frequently used to estimate genome completeness, where the CEGMA completeness score expresses the fraction of the 248 genes that can be accurately mapped onto the genome assembly (Figure 3.1, box C). BUSCO, a method similar to CEGMA, recently defined a set of 429 single-copy orthologs to estimate completeness as well as the duplicated fraction of a eukaryotic genome sequence (Simao et al., 2015).

The CEGMA gene set dates back to the last common eukaryotic ancestor, and thus any extrapolation of the completeness score based on such a limited set of highly conserved proteins will fail to account for many genes unique to plant biology. In addition, as most plant genomes encode more than 20,000 genes, any bias present in such a small set of conserved core genes can lead to errors in the estimated completeness scores. We found that more than half of the 248 CEGMA genes from *A. thaliana* are expressed across all the different conditions and organs contained in a non-redundant *A. thaliana* expression atlas (Figure 3.2 A). This reveals that many genes expressed in specific plant organs or developmental stages are missing. Gene Ontology enrichment further demonstrates the gene function bias in the 248 core eukaryotic genes: housekeeping functions (DNA metabolism, RNA processing, translation and glucose metabolism)

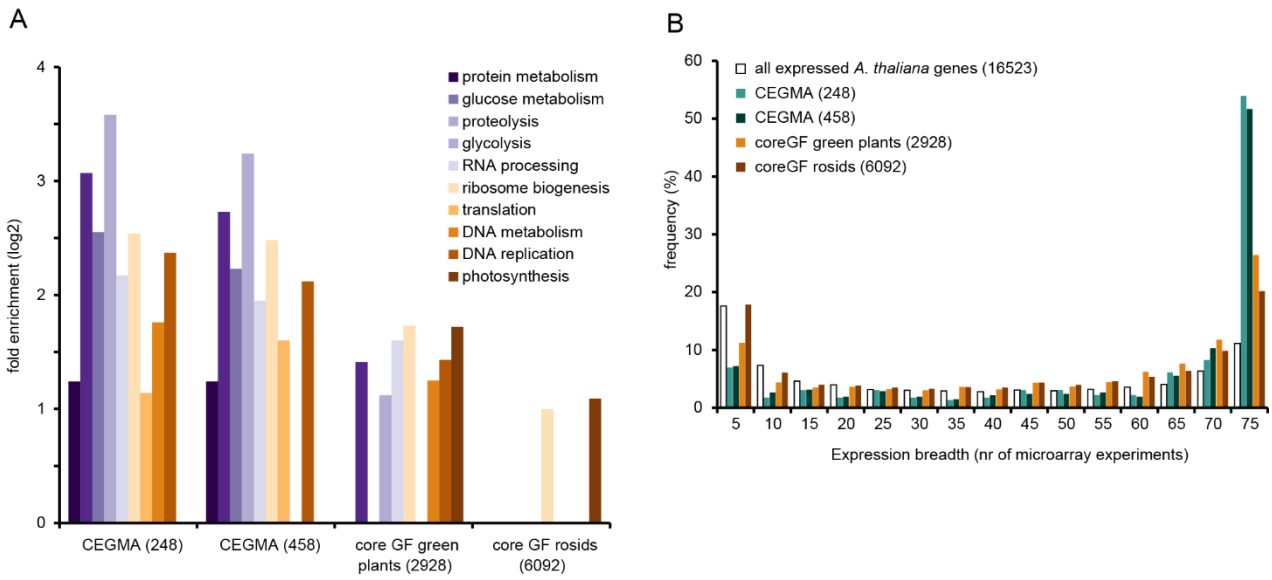


Figure 3.2 Gene function and expression biases associated with CEGMA and coreGFs in *A. thaliana*. Gene function and expression biases were determined for the CEGMA set (248 or 458 single-copy core genes) and the coreGFs for green plants (2928 gene families) and rosids (6092 gene families). (A) Gene function biases were estimated using Gene Ontology (GO) enrichment analysis of the PLAZA 3.0 Workbench. GO terms with at least two-fold enrichment are shown ($p < 0.01$). (B) Expression biases were determined by counting the number of microarray experiments in which a gene is expressed and compared to the expression breadth of the complete gene set of *A. thaliana*.

are over-represented and the CEGMA set does not cover genes functioning in biological processes conserved in green plants such as photosynthesis, biosynthesis of plant-specific hormones, and hormone-mediated signaling (Figure 3.2 B). In summary, comparative genome studies aiming to identify genomic adaptations required for growth in a specific environmental niche (e.g. loss or gain of genes or pathways) should not rely on validating the genome annotation completeness using evolutionarily highly conserved reference sets, because these are blind to lineage-specific genes.

In contrast, transcript mapping is a highly species-specific completeness method which is independent of evolutionary conservation between species. This method uses large-scale EST or RNA-Seq transcript sequencing to estimate how many of the transcribed genes are present in the gene space partition of the genome assembly of a given species (here referred to as 'transcript mapping'). The expected gene space is now defined as the total number of transcript sequences, either specifically generated to guide genome annotation of the sequenced genotype, or derived from public resources. A comparison of transcript mapping at the levels of the genome assembly and the annotated gene catalog indicates the completeness of the gene prediction. Depending

on the library preparation method used, also genes encoding different types of RNA (e.g. rRNAs, tRNAs, snRNA, long non-coding RNAs) can be included.

In reality, *de novo* assembly often first leads to reconstruction of a partition of the genome that contains the euchromatic, gene rich, unique sequences in the genome, and alternative strategies of library preparation and assembly algorithms are needed to reconstruct the heterochromatic, repeat-rich sequence partition. With this in mind, we simulated fragmented and incomplete genomes of *A. thaliana* and *Oryza sativa* to evaluate the influence of transcript mapping parameters on the gene space completeness. In short, we fragmented the genome in 10kb sequences and randomly subsampled genomic fragments to simulate decreasing levels of completeness (50-100%). Random subsampling of a given fraction of the entire genome creates a reference that contains, proportionally, a 'known' fraction of the gene space, independent of whether the repetitive DNA partition is included in the reference or not. We collected 1,5M and 1M publicly available EST sequences for *A. thaliana* and *O. sativa*, respectively, and mapped them onto the partial reference assemblies. We then calculated mean and standard deviation of the transcript mapping score across 100 replicate random subsamples (bins) with varying numbers of ESTs (range 100 - 300,000 ESTs) (see Material and Methods). Finally, we compared the measured gene space completeness scores to the 'known' fraction of the gene space to estimate the influence of EST mapping parameters (such as minimum % coverage), and EST library size and complexity, because these typically vary across the reported completeness estimates. On average, the transcript mapping score is stable (standard deviation < 1%) in bin sizes of at least 3,000 ESTs, for both *A. thaliana* and *O. sativa* (Figure 3.3 A). Transcript mapping estimates the completeness of the gene space at 61%, when only 50% of the *A. thaliana* genome is used as reference, while for more complete genomes, the transcript mapping score converges to 97% (Figure 3.3 A, upper panel). When partial EST mappings were filtered out (90% coverage filter), partial genomes are no longer overestimated, but more complete genomes do seem incomplete (Figure 3.3 A, lower panel). The latter might be related to the challenge of correctly aligning spliced transcript sequences to their corresponding genomic locus, comprising both exons and introns. These results show that it is important to consistently use and report the mapping parameters per comparison method. As stated above, it is important to note that transcript

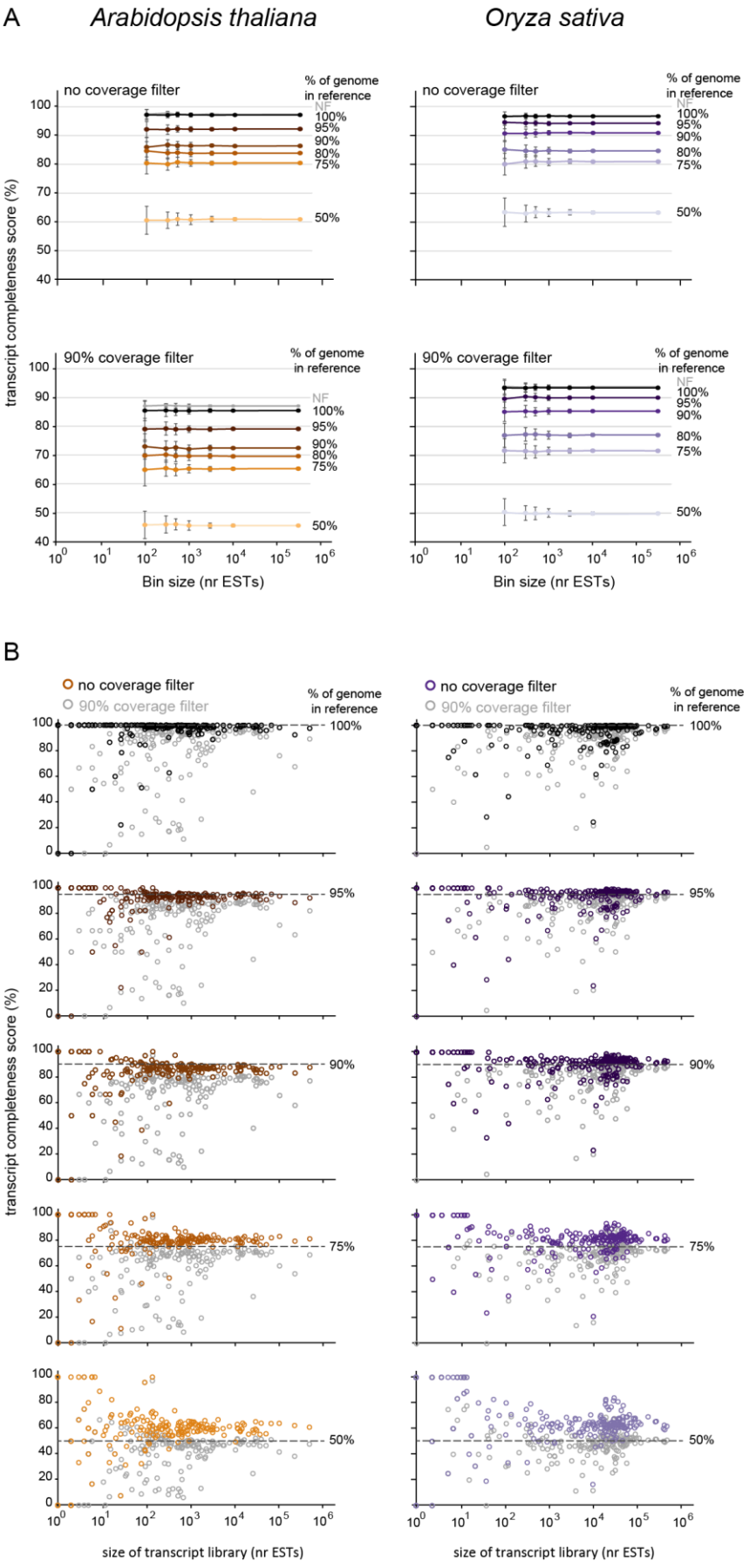


Figure 3.3 Evaluation of transcript completeness scores. To estimate the relationship between transcript completeness score, actual reference genome completeness, and EST library size and complexity, two approaches were compared using *A. thaliana* or *O. sativa*. For each species, the genome was cut into stretches of 10kb and fragments were randomly sampled to create partial genome references containing 50%, 75%, 80%, 90%, 95% and 100% of the original genome sequence (NF: not fragmented). All publicly available EST sequences were mapped onto the respective partial genomes. In a first approach, all ESTs were pooled and random sampling for different EST bin sizes (range from 100 to 300,000) was performed 100 times. The mean and standard deviation of the transcript completeness scores for each bin size and each partial genome is given in Box A. The lower panel shows mean transcript completeness scores and standard deviation counting only mapped ESTs with a length coverage higher than 90%. Box B shows the transcript completeness score for each individual EST library (indicated by a circle) mapped onto the partial genomes. Completeness scores per library based on EST mappings with a length coverage higher than 90% are shown in grey in each panel.

mapping scores should not be extrapolated to the completeness of the total genome assembly, but only apply to the gene space partition, even if the entire genome reference sequence is used for the EST mapping.

We also evaluated transcript mapping scores per library on various simulated genome incompleteness levels for *A. thaliana* and *O. sativa*, to further define the relationship between transcript completeness score, actual completeness, and EST library size and complexity (Figure 3.3 B). Both species display more variation in EST mapping score when smaller libraries are used to define the 'expected' gene space, confirming the results from down-sampling ESTs. If the libraries contain more than 10,000 ESTs, the EST mapping scores for *A. thaliana* libraries converge to the same value as for subsampling bins of >10,000 ESTs. For *O. sativa*, the convergence of EST mapping scores is not as clear. This indicates that the minimum library size needed for a reliable estimate depends on the species, perhaps as function of size and/or complexity of the genome. RNA-seq and *de novo* unigene assembly or PacBio full-length cDNA sequencing generally generates 50,000-100,000 unique transcripts or transcript fragments (Honaas et al., 2016). Several transcript libraries can be generated for a fraction of the cost of the entire genome sequencing project, which suffices to validate the gene space completeness test.

The two methods described above define the expected gene space on two extremes of the evolutionary scale: CEGMA uses highly conserved eukaryotic genes, while transcript mapping is based on the mapping of species-specific cDNA sequences. Although both estimates give a useful insight in the completeness of the gene space in the genome assembly, they also have shortcomings: the CEGMA set does not represent plant-specific gene functions, while transcript mapping is highly dependent on the number of transcript sequences and the complexity captured by the different cDNA libraries. Both methods are dependent on the mapping stringency parameters, which should be adjusted to account for, respectively, divergence between species or genetic diversity within species.

In practice, the underlying assumptions of both approaches can be combined by applying one or more user-defined positions on a branch of the tree of life to define the expected gene space. This approach balances between evolutionary conserved genes or species-specific genes. The

PLAZA Core Gene Families (here referred to as 'coreGFs') are a set of gene families that are highly conserved in a majority of plant species within predefined evolutionary lineages (Van Bel et al., 2012). Three sets of coreGFs have been defined using the PLAZA 2.5 database: green plants (2928 coreGFs), rosids (6092 coreGFs), and monocots (7076 coreGFs), using a parsimony-based selection approach where complete conservation across all species is not required. This approach accounts for the observation that genes are indeed occasionally lost in some species and it tolerates potential annotation errors in a limited number of species. In contrast to CEGMA and BUSCO, coreGFs are not filtered for single-copy genes and can therefore better deal with the frequent occurrence of whole-genome duplications in plants (Van de Peer et al., 2009). Consequently, the number of coreGF genes is five to ten times higher compared to CEGMA or BUSCO gene sets. Similar to BUSCO and transcript mapping, coreGFs can be used to assess the completeness of a gene annotation (for further details on the calculation of the coreGF completeness scores, see Material and Methods). Expression breadth and gene function enrichment analysis reveals that the coreGF gene set is less biased towards ubiquitously expressed genes and does not strongly over-represent specific gene functions (Figure 3.3). Furthermore, because coreGFs sample conserved gene families at different taxonomic levels within green plants, it offers a better representation of the gene function space of flowering plants compared to CEGMA.

Comparison of three gene space completeness measures

The completeness estimates of two methods based on evolutionary conserved gene sets (CEGMA and coreGFs) and transcript mapping were compared (Figure 3.4) using 10 recently published plant genome data sets, including rosids and monocots (Supplemental Table 1). The two high-quality reference genomes of *A. thaliana* and *O. sativa* contain almost all of the CEGMA and coreGFs core genes (completeness scores > 99%; only 50 and 42 missing coreGFs for *A. thaliana* and *O. sativa*, respectively, Figure 3.4). In nine species, the CEGMA score is higher than the coreGF score. So, reporting only CEGMA scores generally leads to an overestimation of the gene space completeness. This difference is at least 5% for more than half of the species while for three species it is even larger than 10%. These missing fractions correspond to the projected

absence of a few hundred to more than a thousand coreGFs genes. The underlying reasons vary and can be illustrated in three specific cases.

First, in *Lolium perenne*, the reported CEGMA score of 96% indicates that the genome assembly is complete, yet 1709 coreGFs are missing from the predicted gene set. For this genome paper, the authors presented a conservative, yet reliable set of annotated genes, by selecting only evidence-based gene models, i.e. supported by *Brachypodium distachyon* protein alignment and transcriptome assemblies (Byrne et al., 2015). The transcript mapping score of 96% on the genome assembly compared to the coreGF score of 76% on the predicted gene set shows that the gene space partition of the genome has been well assembled, but that gene prediction is incomplete. Indeed, mapping of *B. distachyon* proteins on the *L. perenne* genome assembly confirms that at least 924 of the 1709 missing coreGFs can be found using TBLASTN (E-value < 1e-10).

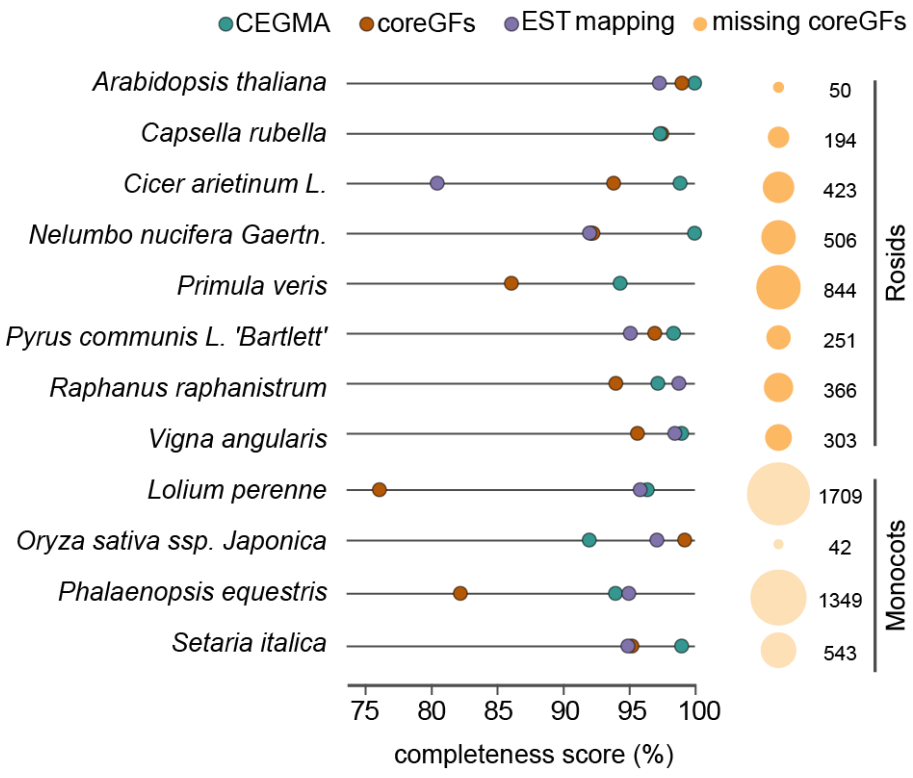


Figure 3.4 Comparison of CEGMA, coreGFs and EST mapping for twelve plant genomes. Twelve genomes within rosids and monocots were analyzed. Left panel: CEGMA, coreGF and EST completeness scores per genome. The reported CEGMA score was obtained from the respective genome publications. We calculated the weighted coreGF score of the respective annotated gene sets, using the rosids or monocots coreGFs according to lineage. The EST mapping completeness score is the percentage of publicly available EST sequences that could be mapped onto the genome. Right panel: the size of the circles and numbers indicate the number of missing coreGFs per genome.

Second, it is important to note that the coreGFs are predefined at three evolutionary levels, rosids, monocots and green plants. Monocot coreGFs were defined only using gene sets from the Poales, which are part of the commelinids. As *Phalaenopsis equestris* belongs to the Asparagales, a sister group to the commelinids, the lower coreGF score could reflect potential gene loss in *P. equestris* and shows the importance of choosing an appropriate phylogenetic level at which an evolutionary conserved gene set is defined.

Third, although coreGF yields a more complete picture of the gene space than CEGMA, it does lack species-specific genes. For most species, the transcript mapping score lies within the same range as the CEGMA and coreGF score. This is not the case for *Cicer arietinum L.*, for which only 89% of the ESTs could be mapped on the genome sequence. More than half of the unmapped sequences are of non-plant origin, mostly from *Fusarium oxysporum*, illustrating how contaminations inflate the expected gene space and lead to an underestimation of the gene space completeness.

Expect the unexpected

Population resequencing studies in *A. thaliana*, rice, potato, and maize have unveiled the extensive individual genomic variation, including structural rearrangements, copy number variations, insertion-deletions, single nucleotide polymorphisms and sequence repeats. This has led to the definition of ‘core’ genome sequences (shared between all members of a species), ‘dispensable’ genome sequences (present in only one or a few members) and ‘pan’ genome sequences (the union, or full genome complement across all members). Hence, the variability of sequence conservation extends to the sub-species or individual organism level (Cao et al., 2011; Hirsch et al., 2014; Marroni et al., 2014). The dispensable genome contains genes with high biological relevance, illustrated by possible roles in adaptation to abiotic and biotic stresses (Hardigan et al., 2016), species diversification and development of novel gene functions (Wang et al., 2006), and agronomic and metabolic traits (Yao et al., 2015). This clearly limits the definition of the 'expected' gene space, and, consequently, the precision and accuracy of completeness estimates of both the genome and the gene space.

3.6 Conclusions and guidelines

A complete genome assembly is essential for the study of chromosome structure and repeat content. Although a complete gene catalog is an important deliverable of a genome sequencing project, the genome assembly should not be restricted to the gene space partition. Here, we discussed different measures to assess genome and gene space completeness and illustrated that large differences in completeness scores for the same genome can be found. Therefore, we advise to assess genome completeness both at the genome assembly and gene space level, to reliably estimate the quality of all steps of assembly and annotation. Based on our observations, we suggest the following guidelines:

1. For genome assembly completeness we suggest to report the estimated genome size based on k-mer statistics of the raw sequence reads, together with the fraction of reads that map onto the assembled genome. In addition, a nuclear weight estimate should also be reported, obtained from an experimental method such as PFGE or flow cytometry using standardized references. Comparison of these measures highlights the fraction of the repeat DNA partition that was not assembled.
2. To assess the completeness of the gene space, we suggest that both inter-species comparisons using a set of conserved core genes as well as intra-species comparisons based on transcript libraries from the sequenced species are performed. Ideally, the core gene set used to model the expected number of genes ought to be defined at various levels of evolutionary conservation, but including a set as large as possible and without strong gene function or expression biases. For transcript mapping preferably different cDNA libraries covering a range of organs and conditions should be included to secure a robust estimate of the expected number of genes. The complexity of EST libraries can be difficult to predict in novel organisms, consequently introducing uncertainty in the ‘expected’ gene space (e.g. Figure 3.3). Assembled transcriptomes, in turn, can be compared against the core gene sets to crosscheck complexity and saturation.
3. Large differences in completeness score between methods based on evolutionary conservation like CEGMA or coreGFs and transcript mapping can point to erroneous assumptions underlying the expected gene space for inter-species comparisons. For

example, in species where large-scale gene loss of specific pathways occurred, application of evolutionary-based methods will result in an underestimated gene space completeness. Similarly, EST datasets contaminated with sequences from other species will underestimate the gene space completeness.

4. The correct structural annotation of species-specific genes and fast-evolving genes, poses big challenges for a full characterization of the gene space. Ideally, gene space completeness estimates should be applied on both the genome assembly and on the annotated gene set, as large score differences can highlight loci in the genome assembly which were missed by the gene prediction. Identifying the missing coreGFs can be used for the targeted investigation of specific gene functions. As such, genes that are truly missing from the genome assembly can point to the discovery of lineage-specific genome evolution, while genes only missing from the predicted gene space indicate that an optimization of the gene prediction algorithms, which frequently suffer from the lack of proper training in a novel organism, is needed.

We believe these pointers will help the next generation of plant scientists to assess the quality of new genome sequences in a transparent and balanced manner and to formulate a standard for delivering better plant genome sequences, which are the templates for new biological discoveries.

3.7 Author Contribution

E.V. retrieved and analyzed the data sets. E.V., T.R., and K.V. contributed to writing the article. K.V. coordinated the project.

4 Overcoming Challenges in Variant Calling: Exploring Sequence Diversity in Candidate Genes for Plant Development in Perennial Ryegrass[2]

Revealing DNA sequence variation within the *Lolium perenne* genepool is important for genetic analysis and development of breeding applications. We reviewed current literature on plant development to select candidate genes in pathways that control agronomic traits, and identified 503 orthologues in *L. perenne*. Using targeted resequencing, we constructed a comprehensive catalog of genomic variation for a *L. perenne* germplasm collection of 736 genotypes derived from current cultivars, breeding material and natural accessions. To overcome challenges of variant calling in heterogeneous outbreeding species, we used two complementary strategies to explore sequence diversity. First, four variant calling pipelines were integrated with the VariantMetaCaller to reach maximal sensitivity. Additional multiplex amplicon sequencing was used to empirically estimate an appropriate precision threshold. Second, a *de novo* assembly strategy was used to reconstruct divergent alleles for each gene. The advantage of this approach was illustrated by discovery of 28 novel alleles of *LpSDUF247*, a polymorphic gene co-segregating with the S-locus of the grass self-incompatibility system. Our approach is applicable to other genetically diverse outbreeding species. The resulting collection of functionally annotated variants can be mined for variants causing phenotypic variation, either through genetic association studies, or by selecting carriers of rare defective alleles for physiological analyses.

[2] This chapter is based on Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., van Parijs, F.R.D., Byrne, S.L., Asp, T., Studer, B., Rohde, A., Roldán-Ruiz, I., Vandepoele, K., and Ruttink, T. (2018). Overcoming challenges in variant calling: exploring sequence diversity in candidate genes for plant development in perennial ryegrass (Lolium perenne). DNA Res, dsy033-dsy033. For author contributions, see page 74.

4.1 Introduction

Perennial ryegrass (*Lolium perenne* L.) is one of the most widely cultivated grass species in Europe. It is of interest for grazing, hay and silage production as it has a long growing season, and relatively high yield and nutritive value. Because of its outbreeding nature, individual plants are highly heterozygous and the diploid perennial ryegrass genome is highly heterogeneous both within and across breeding populations and natural accessions. As genomic variation forms the foundation of phenotypic variation, revealing DNA sequence variation within the genepool is important for genetic analysis and development of breeding applications (Roldán-Ruiz and Kölliker, 2010).

Several studies used a candidate gene-based approach to associate sequence polymorphisms with phenotypic variation. Examples include the association of *Late embryogenesis abundant 3* (*LEA3*) with drought tolerance (Yu et al., 2013), *Brassinosteroid insensitive 1* (*BRI1*) with shoot morphology (Brazauskas et al., 2010), *Gibberellic acid insensitive* (*GAI*) with organ growth (Auzanneau et al., 2007), *Heading date 1* (*HD1*) with carbohydrate content (Skot et al., 2007), and *Flowering locus T* (*FT*) with flowering time (Skot et al., 2007; Fiil et al., 2011; Skot et al., 2011). While these studies show the power of testing gene-trait associations, the limited number of genes per study was mostly due to the high cost of genotyping at the time. However, this approach is not amenable to study complex traits related to plant development and phenology, which are typically regulated by the interaction of many genes. Therefore, we need versatile and cost-efficient methods to characterize the genetic variation in parallel for hundreds of candidate genes and hundreds of genotypes. This enables breeders to perform higher resolution screening of genetic diversity in their material and link genotypic and phenotypic variation.

Single nucleotide polymorphisms (SNPs) are the most prevalent type of genomic variation and are convenient molecular markers. Two complementary SNP genotyping arrays are available for high-throughput screening in perennial ryegrass (Studer et al., 2012; Blackmore et al., 2015). These arrays target SNPs in genic regions, but do not allow discovery of new sequence variants. In contrast, Genotyping by Sequencing (GBS) allows for simultaneous discovery of genome-wide SNPs and genotyping of a large number of individuals or pools, thereby avoiding ascertainment

bias. Therefore, GBS has broad applications in plant breeding and genetics studies, including linkage maps, genome-wide association studies, genomic selection, and genomic diversity studies (Chung et al., 2017). Only short (range about 100-300 bp) fragments are sequenced and there is no *a priori* control over which genes are tagged. In combination with a local short linkage disequilibrium that is typical for an outbreeding species as *L. perenne* (Xing et al., 2007; Brazauskas et al., 2010), it can be very difficult to identify SNPs that are causal for the phenotype of interest.

The large size of the *L. perenne* genome (2 Gbp) and high repetitive sequence content (76%) (Byrne et al., 2015) currently precludes whole genome sequencing at sufficient depth in hundreds of accessions as has been done in e.g. Arabidopsis, rice and soybean. To study trait genetics in forage and turf grasses, we identified hundreds of candidate genes in genetic pathways that control plant development and quality traits, and analysed their genome sequence using probe capture enrichment for targeted resequencing in a large germplasm collection of 736 genotypes. We specifically focussed on genes involved in pathways related to interesting agronomic traits, such as plant growth and architecture (important for biomass yield), development and transition to flowering (important for seasonal control of growth), cell wall biogenesis (important for digestibility) and phytohormone biosynthesis, signalling and response (including abscisic acid, auxin, brassinosteroids, cytokinins, ethylene, gibberellic acid and strigolactones). Identifying sequence variants in these genes provides insights in the range of naturally occurring genomic diversity that can be expected in gene-rich regions of the genome. The variants can be used as markers for association genetics studies (as previously described for *LEA3*, *BRI1*, *GAI*, *HD1*, and *FT3*), and/or to identify alleles with altered amino acid sequence, mRNA splicing, or mRNA stability, hence altered gene function or regulation possibly resulting in an altered physiology and thereby affecting the phenotype.

Multiple bioinformatics methods are available to identify sequence variants using next-generation sequencing (NGS) data, but defining a complete and reliable variant set remains difficult. *De novo* discovery of genomic polymorphisms commonly relies on mapping reads to a single reference genome sequence. Although the GATK best practices are the most commonly

used variant calling (VC) pipeline, there is no single best VC pipeline available with both good sensitivity and precision. Moreover, there is low concordance between VC pipelines, even with the same input data (Liu et al., 2013; O'Rawe et al., 2013; Yu and Sun, 2013; Pirooznia et al., 2014). In addition, each VC pipeline returns different variant annotations that can be used for quality filtering. Choosing the appropriate filtering criteria and thresholds (for instance minimum read depth) is not straightforward as NGS data typically has a non-uniform distribution of coverage (Quail et al., 2012) and estimated quality values may be dataset dependent so that optimal settings need to be calibrated for each dataset. The high density of sequence polymorphisms in the germplasm collection with respect to the *L. perenne* genome sequence could also hamper variant identification. More divergent alleles could contain the most interesting genomic variation, but are also the most difficult to detect as reads that are highly divergent from the reference genome may fail to map if the parameters for short read alignment are too stringent (Bertels et al., 2014). Hence, if capture and/or mapping efficiency of highly divergent sequences precludes their detection, it is to be expected that routine workflows of mapping and variant calling lead to an underestimation of the genetic diversity at highly divergent regions, a known problem in genome resequencing studies (Gan et al., 2011).

Here, we present the identification and annotation of 503 *L. perenne* orthologs of known genes that regulate plant growth and development. These genes were resequenced in a germplasm collection of 736 genotypes to describe the genomic variation in *L. perenne*. Two complementary strategies were used to obtain a reliable and complete catalog of genomic variation. First, four VC pipelines were compared and automatically integrated to reach maximal sensitivity. The influence of mapping algorithms was assessed and hard filtering was compared to precision-based filtering to reach sufficient specificity. Additionally, an alternative strategy consisting of *de novo* assembly followed by overlap-layout-consensus (OLC) clustering was used to circumvent read mapping bias and to construct alternative alleles for each gene. This reference independent allele reconstruction is particularly important for gene families with highly divergent alleles. We demonstrated the benefit of this approach for *LpSDUF247* and identified 28 novel alleles that were not detected using traditional VC pipelines. This approach is broadly applicable to other highly heterozygous outbreeding species. Finally, we used all this information to create a

comprehensive catalog of functionally annotated genetic variation across many pathways that control growth, development and agricultural traits.

4.2 Material and Methods

Candidate gene identification and manual curation

Gene families of *A. thaliana* candidate genes were identified using the comparative genomic platform PLAZA 3.0 Monocots (Proost et al., 2015). *B. distachyon* family members were used to identify homologous loci in the draft genome sequence of *L. perenne* (Byrne et al., 2015) using BLASTx analysis (E-value 10e-5). At each *L. perenne* locus, predicted protein sequences were added to the corresponding PLAZA 3.0 Monocots gene family. Using protein sequences of all gene family members of *A. thaliana, B. distachyon* and *L. perenne*, a phylogenetic tree was built with MUSCLE (v3.8.31) (Edgar, 2004) and PhyML (Guindon et al., 2010) using default settings, and for each *A. thaliana* candidate gene the closest orthologous *L. perenne* gene was selected for further manual curation. The *L. perenne* gene models were evaluated using multiple protein sequence alignments with MUSCLE using orthologous proteins from *B. distachyon*, *O. sativa*, *Z. mays*, and *S. bicolor* according to PLAZA 3.0 Monocots. Additional RNA-seq data (Ruttink et al., 2013; Farrell et al., 2014; Paina et al., 2014) mapped with TopHat (v2.0.13) (Trapnell et al., 2009) using default settings, was used to refine gene models and delineate untranslated regions, or to design a new gene model if required.

Probe design, library construction and sequencing

The coding strand of each of the 503 target regions (gene model and 1,000 bp upstream promoter region) was tiled with 120 bp probes, starting every 40 bp using OligoTiler(Bertone et al., 2006). Probes showing high sequence similarity to non-targets, other probes, repetitive sequences, mitochondrial or chloroplast sequences, or with extreme GC content (<25% or >65%) were removed. Finally, 57,693 SureSelect probes of 120 bp (Agilent) were retained, covering 2.3 Mb of the intended 2.8 Mb target region, at around 3x tiling.

Genomic DNA was extracted from freeze-dried leaf material from 736 *L. perenne* genotypes representing current cultivars, breeding material and natural accessions using the

cetyltrimethylammonium bromide (CTAB) method(Murray and Thompson, 1980). DNA concentration was measured using the Quantus double-stranded DNA assay (Promega, Madison, WI, USA). For each genotype, an indexed shotgun sequencing library was prepared from 100 ng DNA by (i) Adaptive Focused Acoustic fragmentation on a Covaris S2 instrument (Covaris, Inc.), (ii) adapter ligation, and (iii) magnetic bead purification using an adapted protocol of Uitdewilligen et al. (2013). The libraries were pooled without normalization into eight pools, each containing 96 libraries of individual genotypes. Each pool was used for a probe capture hybridization reaction according to the SureSelect protocol (Agilent SureSelectXT2 Target Enrichment for Illumina Paired-End Sequencing Library Protocol, v. 1.0). After PCR amplification of purified enriched pooled libraries, each pool of 96 libraries was sequenced on one lane of a HiSeq2000 using 2x91-PE sequencing (BGI, Shenzhen, China). The raw data is available in the NCBI Sequence Read Archive (BioProject PRJNA434356, Accessions SRR6812717 to SRR6813075).

Read mapping and variant calling

Raw reads were trimmed and quality filtered by Trimmomatic (v0.32) (Bolger et al., 2014) and mapped onto the draft perennial ryegrass genome sequence (Byrne et al., 2015) with default settings of BWA-MEM (version 0.7.8-r455) (Li and Durbin, 2009) and GSNAP (version 2016-09-23) (Wu and Nacu, 2010). Duplicate reads were marked using Picard-tools (release 1.113). Local realignment around indels was performed according to the best practices workflow of the Genome Analysis Toolkit (GATK) (v.3.7) (McKenna et al., 2010; Van der Auwera et al., 2013). Read depth and coverage were calculated on the resulting BAM files using BEDTools (v2.25.0) (Quinlan and Hall, 2010).

Four different VC pipelines were used: SAMtools (version 1.2-115-gb8ff342) (Li et al., 2009), Freebayes (v1.0.2-2-g7ceb532) (Garrison and Marth, 2012), GATK Unified Genotyper (GATK UG) and GATK HaplotypeCaller (GATK HC) (McKenna et al., 2010; Van der Auwera et al., 2013). Multi-allelic variants were removed using VCFtools (v0.1.14) (Danecek et al., 2011). For hard filtering, a custom Python script was used to remove variant positions and genotype calls with a read depth lower than 6 and a genotype quality lower than 30. SNPs and indels were automatically integrated by the VariantMetaCaller (v1.0) (Gézsi et al., 2015), in ten and two partitions,

respectively. The Estimated Precision (EP) was calculated using a custom Python script based on the formulas given in Gézsi et al. (2015). The concordance of SNP and indel sets identified by four VC pipelines was determined using information in the INFO field of the VCF file returned by VariantMetaCaller and visually represented using Upset (Lex et al., 2014), before and after precision-based filtering (EP > 80%). Functional effects of sequence variants were predicted with SnpEff (version 4.3T) (Cingolani et al., 2012). To validate consistency of genotype calls in an F1 segregating population, mendelian inheritance errors (MIE) were defined after precision-base filtering (EP > 80%) using PLINK (v1.90b2t), for two parents and their F1 progeny of 29 individuals. Variants with a missing genotype call in either one of the parents were excluded from analysis, as were MIEs derived from a missing genotype call in one of the 29 F1 progeny.

Hi-Plex amplicon sequencing

To generate an independent variant set, 78 genotypes were selected for resequencing of 171 amplicon regions of 80-140 bp. Of these, 147 amplicons overlap with 28 candidate genes. Primers were designed with Primer3 (Untergasser et al., 2012) and divided into two highly multiplex (Hi-Plex) PCR-reactions according to their amplification efficiency. DNA was extracted using the CTAB method (Murray and Thompson, 1980) and DNA concentration was measured using the Quantus double-stranded DNA assay (Promega, Madison, WI, USA). Per sample, the final DNA concentration was adjusted to 40 ng/µL and the amplicons were PCR-amplified while adding sample specific indices. Libraries were prepared using the KAPA Hyper Prep PCR-free Kit according to manufacturer directions (Kapa Biosystems, USA). Hi-Plex amplification reactions and library preparations were done by Floodlight Genomics LLC (Knoxville, TN, USA). The libraries were sequenced with 2x150 PE on a HiSeq3000 (OMRF, Oklahoma City, OK, USA). Paired-end reads were merged with PEAR (v0.9.8) (Zhang et al., 2014) and adapter sequences were removed. The read data is available in the NCBI Sequence Read Archive (BioProject PRJNA437219, Accessions SRR6813540 to SRR6813585). BWA-MEM was used for read mapping, and variant calling was done by running the four VC pipelines. Bi-allelic variants were extracted using VCFtools and combined by VariantMetaCaller and the EP was calculated as described above.

Identification of divergent alleles of *LpSDUF247*

Per genotype, all reads were used for De Bruijn graph assembly without scaffolding (CLC Genomics Workbench 9.5.3, https://www.qiagenbioinformatics.com). Contigs of at least 200 bp were retained and mapped onto the reference genome with BWA-MEM using default parameters, to group all contigs of the 736 genotypes per candidate gene. Per candidate gene, sequences of overlapping allelic fragments were extracted from the BAM files using BEDtools and clustered with the OLC assembler CAP3 (version date 02/10/15) (Huang and Madan, 1999). Singlet sequences returned by CAP3 were removed from further analysis.

One of the candidate genes of the 503 gene set, *LpSDUF247*, is known to be highly polymorphic and was selected to demonstrate in-depth reconstruction of divergent alleles. The 34 contigs of *LpSDUF247* were aligned using MUSCLE and six highly similar sequences (>98% identity) were removed. The reference gene model of *LpSDUF247* was projected onto the contigs using GenomeThreader (v 1.6.6) (Gremme et al., 2005) to identify CDS regions and corresponding protein sequences.

All *B. distachyon* members of the DUF247 gene family (HOM03M000101) were used in a tBLASTn search against the perennial ryegrass genome sequence, and 25 *LpDUF247* genes were identified and manually annotated. After multiple sequence alignment of all 25 LpDUF247 protein sequences with *B. distachyon* and *H. vulgare* gene family members using MUSCLE, a phylogenetic tree was built with PhyML using 100 rounds of bootstrapping (Supplemental Figure 6). Similarly, a phylogenetic tree was built using the reference protein sequences of LpDUF247-01, LpSDUF247, LpDUF247-03 and LpDUF247-04, the protein sequences of the LpSDUF247 alleles, and five LpSDUF247-02 alleles identified by Manzanares et al. (2016) (Supplemental Figure 7).

The 28 novel alleles were added to the reference genome sequence and read mapping was repeated for all 736 genotypes onto this multi-allelic reference genome. A matrix was created with the average read depth per *LpSDUF247* allele per genotype using BEDtools. This matrix was normalised per genotype, by dividing the read depth *per LpSDUF247* allele by the sum of read depths across *all LpSDUF247* alleles, to identify alleles with the highest relative RD for each

genotype while correcting for differences in library size and capture efficiency across the set of 736 samples.

4.3 Results and Discussion

Identification, classification and curation of target genes

To identify *L. perenne* genes putatively involved in the regulation of plant growth and development, plant architecture, induction of flowering, cell wall biogenesis, and phytohormone biosynthesis, signaling and response, we first searched the literature for *Arabidopsis thaliana* genes with a well-defined molecular and physiological function (Supplemental Table 2). Next, the corresponding 174 gene families were identified with the comparative genomics platform PLAZA 3.0 Monocots (Proost et al., 2015). For each of the *A. thaliana* candidate genes, a comprehensive list of orthologous loci in the draft genome sequence of *L. perenne* (Byrne et al., 2015) was delineated. A phylogenetic tree was built for *A. thaliana* and *Brachypodium distachyon* gene family members of the 174 PLAZA gene families, to select the closest orthologous *L. perenne* sequences of the candidate genes. When no clear one-to-one orthologous pairs were found due to lineage-specific gene duplication or gene loss events, the best two or three *L. perenne* loci were selected from the respective clades. The final selection contained 503 *L. perenne* candidate genes (Supplemental Table 2). For 407 of these loci, an annotated gene model was available (Byrne et al., 2015). For the other 96 loci, a gene model needed to be annotated *ab-initio*, in line with previous observation that the annotated gene space of *L. perenne* is 76% complete (Veeckman et al., 2016). The available gene models were evaluated using multiple protein sequence alignments with all their monocot gene family members according to PLAZA 3.0 Monocots. In addition, mapped RNA-seq data (Ruttink et al., 2013; Farrell et al., 2014; Paina et al., 2014) was used to refine gene models and delineate untranslated regions. Taken together, manual curation of 503 gene models, showed that previously available gene models (Byrne et al., 2015) were correct for 272 loci (54%) and needed small adaptations for 135 loci (27%). A completely new gene model was annotated at 96 loci (19%) using RNA-seq data. The length of the protein sequences corresponds well to that of their closest *B. distachyon* orthologs (Supplemental Figure 2), showing that the 503 manually curated *L. perenne* gene models are of

Table 4.1 Assignment of 503 candidate genes to pathways and distribution of high impact mutations per pathway.

Pathway	Gene families	# candidate genes	Stop gain	Splice site	Frame shift
Plant development and architecture					
Development	BCH1, BRIZ, CBP80, DRM1, HB13, HYL1, ING2, RSM1, SAMDC4	14	2 (14%)	4 (29%)	-
Cell wall	4CL, ALDH, C3H, C4H, CAD, CAD2, CCoAOMT, CCR, CES, COMT, F5H, HCT, HPRGP, IRX, LAC, OFP, PAL, POX, SND, XylS, XylT	121	41 (34%)	16 (13%)	5 (4%)
Cell wall TF	ERF, WRKY	6	2 (33%)	-	1 (17%)
Cell wall TF MYB	MYB	21	3 (14%)	-	1 (5%)
Cell wall TF NAC	NAC	11	2 (18%)	4 (36%)	1 (9%)
Chromatin remodelling	MET1, SWI	4	3 (75%)	2 (50%)	-
Lateral organ initiation	ANT, SLOMO, TOP1A	6	1 (17%)	-	-
Lateral organ patterning morphogenesis	AS, CLF, DOT5, GRF, KAN, NOV, SE, TRN1, YABBY, ZPR1, ZPR3	30	7 (23%)	3 (10%)	2 (7%)
Lateral organ identity	AN3, BOP, HDZIPIII	10	4 (40%)	1 (10%)	-
Light signalling	bHLHABAI, CO1, COP9, CRY, DET1, HY5, LHY, PCI, PFT1, PHYB, PIF, SPA	29	4 (14%)	7 (24%)	-
Shoot apical meristem	BARD1, BLH, CLPS3, FTA, KNAT, OBE1, ULT1, USP1, VEF2, WOX14, WUS	25	8 (32%)	5 (20%)	3 (12%)
Self-incompatibility	DUF247, GK	4	2 (50%)	1 (25%)	1 (25%)
Transition to flowering	CCA, FCA, FIE, FKF1, FLD, FPA, FT, FVE, FWA, FY, GI, LHP1, MBD9, PHP, RAV, SDG8, SPL3, VIL3, VRN1, VRN1-like	45	19 (42%)	12 (27%)	1 (2%)
Flower development	ESD4, HAC3, LFY3, LUG, MADS, RGA, SEU, SUF4, SUP	31	2 (6%)	4 (13%)	-
Transcription factor	BIM2, TCP	8	2 (25%)	-	-
Phytohormone biosynthesis, signalling and response					
ABA biosynthesis	NCED1, PDS1, PDS3	4	1 (25%)	1 (25%)	-
ABA signalling	ABI1, ABI3, ABI5, ABI8, AIP3, DRIP, GBF, GPA, GTG2, HD2C, PSY, SAD1, SIR3, WIG, ZEP	29	10 (34%)	3 (10%)	-
Auxin biosynthesis	TAA1, TAR2, YUC	6	3 (50%)	-	1 (17%)
Auxin signalling	ADA2B, AMP1, ARF, AUXIAA, AXR, AXR1, AXR4, AXR6, CAND1, GH3, TIR1	20	8 (40%)	3 (15%)	-
Auxin transport	AUX1, ENP, PGP4, PID2, PIN1, PIN1like, SPS	12	1 (8%)	-	-
Brassinosteroid biosynthesis	DWF1, DWF3, DWF5, DWF7, SQS	8	2 (25%)	1 (13%)	-
Brassinosteroid signalling	BES1	2	-	-	-
Cytokinin signalling	ARR, CRE, GCR1, RR	11	2 (18%)	1 (9%)	-
Ethylene biosynthesis	ACS	2	1 (50%)	-	-
Ethylene signalling	EBF1, EBF2, EIL3, EIN2, ETO1, ETR1	13	7 (54%)	1 (8%)	1 (8%)
Gibberellin biosynthesis	GAOX	11	4 (36%)	-	2 (18%)
Gibberellin signalling	GID1A, SHI, SPY	5	-	1 (20%)	-
Strigolacton biosynthesis	D14, D27, MAX1, MAX3, MAX4	11	3 (27%)	2 (18%)	-
Strigolacton signalling	MAX2, TB1	4	-	-	-
Total	**180**	**503**	**144**	**72**	**19**

high quality. This was required to delineate regions for probe design and to correctly position variants relative to the reading frame in the CDS to functionally interpret the consequences of sequence polymorphism in the genic regions. Finally, the 503 candidate genes were assigned to biological processes based on the known function of their *A. thaliana* orthologs (Table 4.1 and Supplemental Table 2). This high quality gene set can also be used to train and validate gene prediction algorithms to improve genome-wide gene annotation.

Design and efficacy of targeted resequencing by probe capture enrichment

For each candidate gene, a target region was delineated spanning the curated gene model and an additional 1,000 bp upstream promoter region, as previously described (Ruttink et al., 2015). Probes were designed for a total length of 2.3 Mbp, corresponding to a coverage of 85% of each target region on average, as probes targeting repetitive regions were excluded. Targeted resequencing of 736 genotypes resulted in 3.2 million reads per genotype on average (range 20 thousand – 31 million). After duplicate read removal, a mean of 1.9 million reads was retained per genotype, corresponding to a mean read depth (RD) of 80X per position within the target regions. For variant calling analysis in heterozygous diploid species, a coverage of at least 6-10x is desirable, to avoid false negative heterozygous calls (Song et al., 2016). Saturation curves show a non-linear relationship between number of reads per library and target region coverage at a given RD threshold, as expected for probe capture enriched shotgun sequencing libraries (Figure 4.1). At least 550,000 uniquely mapped reads per genotype were required to reach the probe region coverage plateau at 95% for RD ≥ 1. Further increasing the number of reads per sample did not substantially increase probe region coverage (see Ruttink et al. (2015)). The probe region coverage was slightly lower at higher RD thresholds (89% for RD ≥ 6 and 85% for RD ≥ 10 (Boxplots in Figure 4.1).

Optimization of variant calling pipelines to compile a reliable catalog of sequence variation

To obtain a complete and reliable variant set, we selected two mapping algorithms and four frequently used multi-sample VC pipelines to reach maximal sensitivity. Read mapping algorithm BWA-MEM (Li and Durbin, 2009) was compared to GSNAP (Wu and Nacu, 2010), which is able to

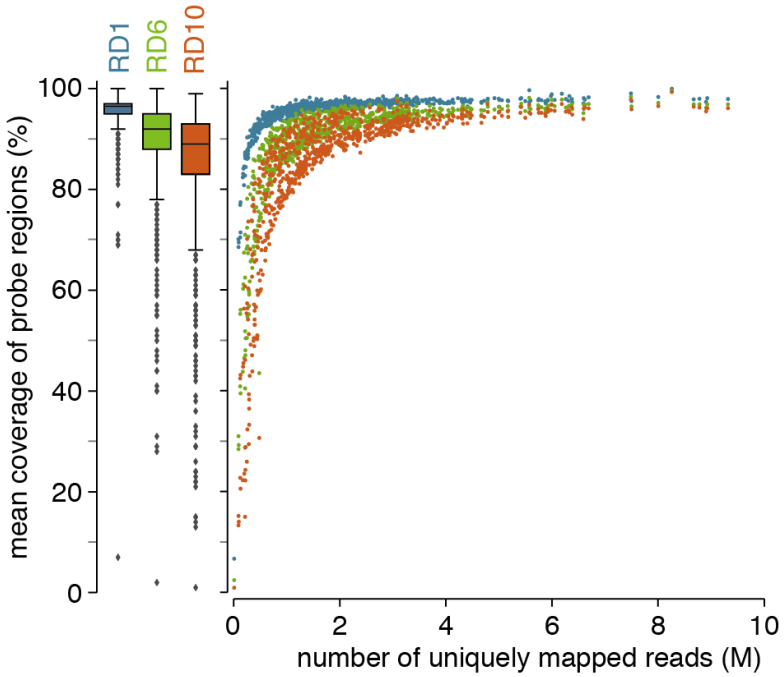


Figure 4.1 Target region coverage per genotype. For each of the 503 candidate genes, the target region was delineated as the gene model and an additional 1000 bp upstream promoter region. The mean fraction of the target region covered per genotype is shown in function of the number of uniquely mapped reads using BWA-MEM after duplicate removal, using different read depth thresholds (RD ≥ 1 (blue), RD ≥ 6 (green) and RD ≥ 10 (orange)).

handle short and long insertions and deletions. Two alignment based VC pipelines were selected for their strength in SNP calling (Tian et al., 2016): GATK Unified Genotyper (GATK UG) (McKenna et al., 2010; Van der Auwera et al., 2013) and SAMtools (Li et al., 2009). Additionally, two haplotype-based VC pipelines were chosen for their strength in indel detection: GATK HaplotypeCaller (GATK HC) and Freebayes (Garrison and Marth, 2012). We compared the resulting variant sets and assessed the performance of hard filtering to improve the precision of variant sets. Finally, the four individual variant sets and corresponding variant quality annotations were merged by the VariantMetaCaller (Gézsi et al., 2015), allowing for precision-based filtering as an alternative to hard filtering.

Influence of read mapping algorithms and variant calling pipelines

For each VC pipeline, the similarity of SNP and indel sets identified using BWA-MEM or GSNAP mappings was calculated using the Jaccard Index (Figure 4.2). The similarity of SNP_{BWA} and SNP_{GSNAP} sets was lowest for Freebayes (0.73) and highest for SAMtools (0.83). The similarity of $indel_{BWA}$ and $indel_{GSNAP}$ sets was lower than that of SNP_{BWA} and SNP_{GSNAP}, independent of the VC pipeline. Jaccard index between $indel_{BWA}$ and $indel_{GSNAP}$ sets ranged from 0.47 (Freebayes) to 0.70

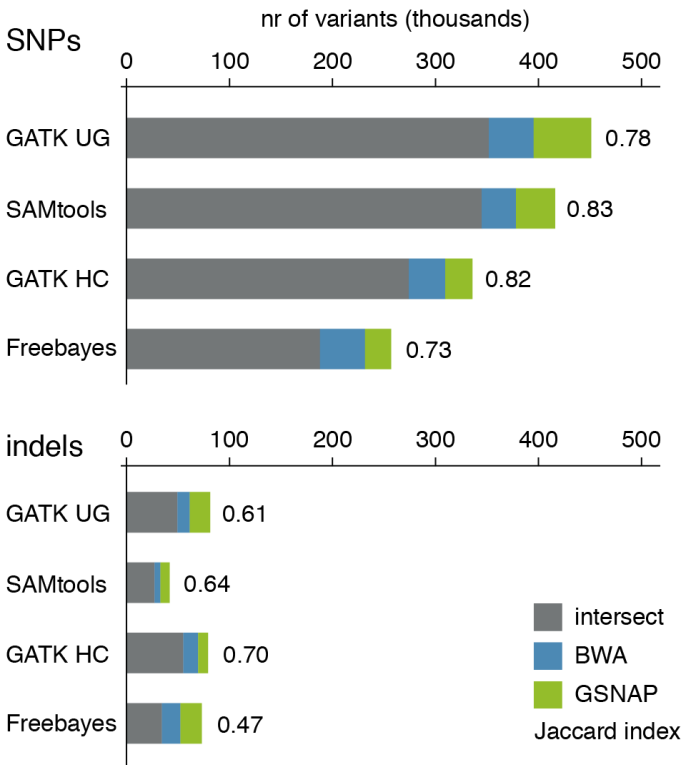


Figure 4.2 Overlap of variant sets generated using BWA-MEM and GSNAP mappings as input for four variant calling pipelines. SNPs and indels were determined for 503 candidate genes in 736 genotypes for BWA-MEM and GSNAP mappings using four variant calling pipelines. The intersect of variants sets was calculated to determine common variants (grey) and uniquely identified variants using BWA-MEM mappings (blue) or GSNAP mappings (green) as input. The Jaccard index value indicates the corresponding similarity.

(GATK HC). On average, 11% of the SNPs were uniquely identified in the SNP_{BWA} set and 9% of the SNPs were uniquely identified in the SNP_{GSNAP} set. Likewise, on average, 17% of the indels were uniquely identified in the $indel_{BWA}$ set and 19% of the indels were uniquely identified in the $indel_{GSNAP}$ set. In summary, the choice of read mapper did not affect the SNP and indel sets as much as the choice of VC pipeline. For results presented below, only variants identified on BWA-MEM mappings are shown, as the same trend was observed for GSNAP mappings.

Concordance of variant sets produced by four variant calling pipelines

Next, the size and concordance of variant sets (bi-allelic SNPs and indels) identified by the four VC pipelines were compared (Figure 4.3). The number of SNPs was highest for GATK UG and SAMtools and considerably lower for Freebayes. The number of indels was at least four times lower than the number of SNPs identified by the same VC pipeline. GATK UG and GATK HC identified the highest number of indels, and SAMtools the least. The concordance of all four VC

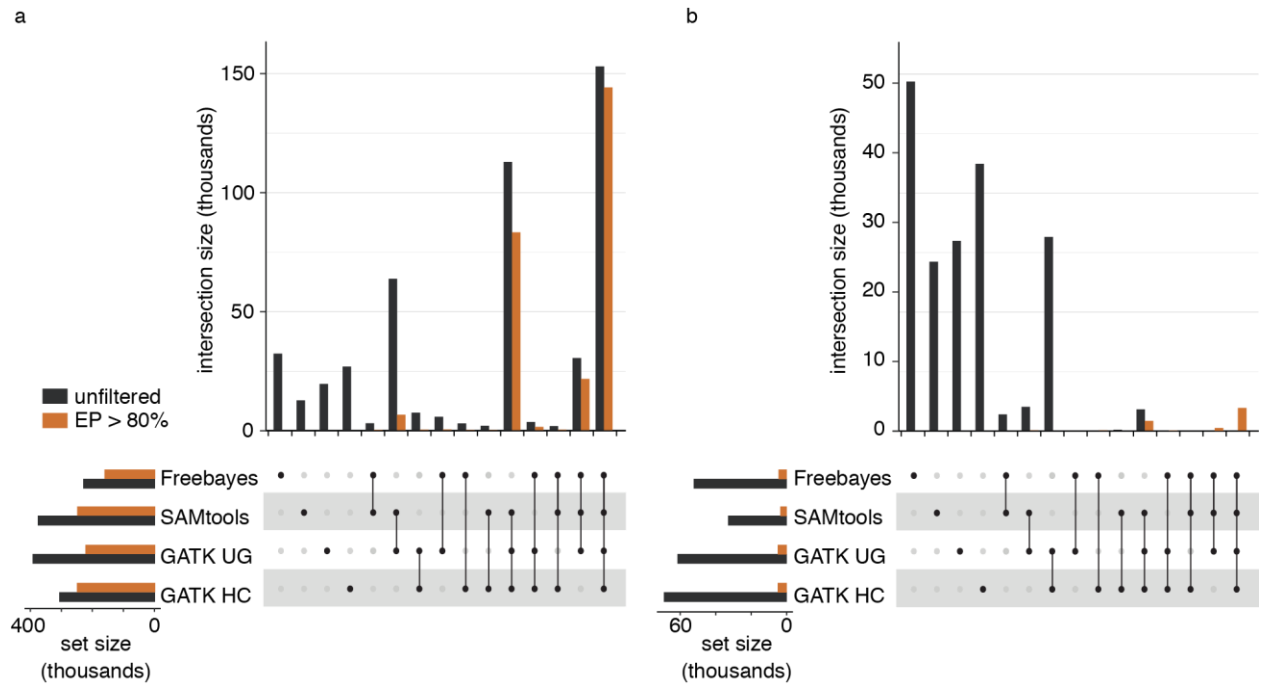


Figure 4.3 Size and concordance of bi-allelic SNP and indel sets of four variant calling pipelines, before and after precision-based filtering. SNPs and indels were identified for 503 candidate genes in 736 genotypes using four VC pipelines and concordance was calculated for bi-allelic SNPs (A) and indels (B). Per Upset plot, the lower left panel shows the total number of variants per VC pipeline; the lower right panel shows the overlap in call sets between the four VC pipelines. The bar graph shows the size per concordance group before (black) and after integration by the VariantMetaCaller and precision-based filtering (EP > 80%) (orange).

pipelines was low: only 150k SNPs (33% of the total number of SNPs identified) and 6.8k indels (5% of the total number of indels identified) were commonly identified, in line with previous reports (O'Rawe et al., 2013).Precision-based filtering is more reliable than hard filteringHard filtering on, e.g., minimal read depth (RD) and genotype quality (GQ) is a commonly used strategy to improve the precision of variant sets. As expected, both number of variant positions and call rate (number of genotype calls per position across 736 genotypes) decreased by filtering on minimal RD of six and minimal GQ score of 30 (Figure 4.4 A and B). Notably, hard filtering did not increase the concordance between VC pipelines (Supplemental Figure 3), indicating that true variants were not necessarily identified by multiple VC pipelines. These results corroborate that it is difficult to build a reliable catalog of sequence variation using a single VC pipeline and applying hard filtering (Park et al., 2014).

As an alternative for hard filtering on individual VC pipelines, the VariantMetaCaller (Gézsi et al., 2015) uses support vector machines to automatically combine multiple information sources (including RD and GQ values) generated by the four VC pipelines, and estimates the probability that a variant is a true genetic variant and not a sequencing artefact. The unfiltered, VMC

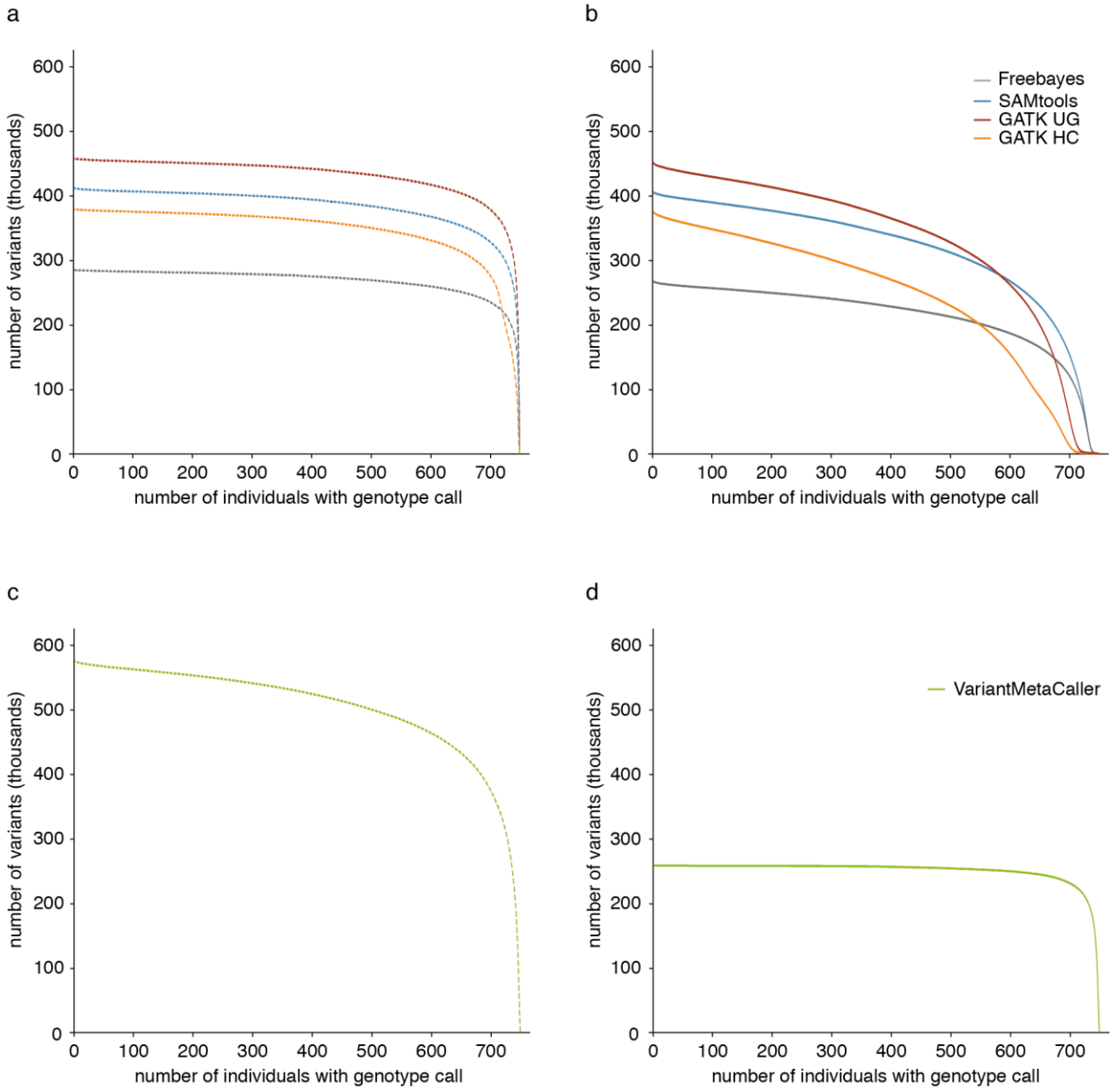


Figure 4.4 Effect of hard filtering and precision-based filtering on the saturation of genotype calls across the 736 genotypes. SNPs and indels were determined for 503 candidate genes in 736 genotypes using four variant calling pipelines and integrated using the VariantMetaCaller (VMC). The genotype call rate was calculated as the number of genotype calls present for each variant, over the total number of genotypes, and plotted cumulatively to estimate the genotype call saturation. This was done for bi-allelic variant sets: (A) before and (B) after hard filtering (read depth > 6, genotype quality > 30) of the variant sets returned by the four variant calling pipelines and (C) before and (D) after precision-based filtering (EP > 80%) of the VariantMetaCaller output.

integrated probe capture variant set contained 444,222 SNPs and 132,766 indels, determined in 736 genotypes and 503 candidate genes. By ordering the variants according to their probability, an Estimated Precision (EP) was calculated for each variant, which can be used for precision-based filtering. In general, variants identified by multiple VC pipelines were assigned higher EP scores. As precision in this context refers to the number of true called variants, choosing an EP threshold is equivalent to finding a dataset- and aim-specific balance between sensitivity and precision of variant calling (Gézsi et al., 2015).

Empirical determination of the EP threshold

Instead of using an arbitrary EP threshold, we reasoned that the EP threshold should be determined empirically, based on the distributions of EP values of high quality and low quality variants. In the absence of a published reference set of variants for the genotypes and genes used in this study, we generated an independent variant set for a subset of 78 genotypes using a Hi-Plex amplicon sequencing assay (Nguyen-Dumont et al., 2013) of 171 amplicons, of which 147 overlap with 28 out of the 503 candidate genes. Hi-Plex amplicon sequencing resulted in 126,000 reads per genotype on average (range 11,000 – 418,000), corresponding to an average read depth of 619 reads per amplicon (range 24 – 20,000).

Using the four VC pipelines integrated by the VariantMetaCaller resulted in a Hi-Plex variant set containing 813 SNPs and 184 indels, compared to 775 SNPs and 246 indels in the probe capture variant set that overlap with these amplicons. In total, 593 SNPs and 60 indels were commonly identified by the two independent sequencing-based genotyping methods. Together, these variants were defined as the High Quality (HQ) subset of variants. Conversely, SNPs and indels that were unique to either set (i.e. non-reproducible and more likely to be random artefacts), were defined as the Low Quality (LQ) subset of variants per genotyping method.

To further validate HQ variants, we compared genotype calls of two methods (probe capture vs Hi-Plex) at the individual genotype level. The mean genotype call consistency, calculated as percentage of identical genotype calls on the total of 593 HQ SNPs and 60 HQ indels, over all 78 genotypes was 97% (range 93% – 100%). This high level of genotype call consistency confirmed the high quality of commonly identified variants. Inconsistent genotype calls are most likely the

result of failed probe-capture, low read depth, the complexity of the region potentially hampering read mapping, allele specific amplification bias in the amplicon sequencing data, or combinations thereof.

Comparison of EP value distributions of HQ and LQ variant positions (Figure 4.5) revealed that EP values of the Hi-Plex variant set were generally lower than those of the probe capture variant set, possibly because of higher read depth and lower complexity of amplicon reads. Furthermore, EP values associated with HQ SNPs were higher than EP values of LQ SNPs for both Hi-Plex and probe capture SNP sets. Taken together, these data show that an EP threshold of 80% differentiates most HQ variants from LQ variants in the probe capture SNP set, whereas the EP threshold needs to be set at 70% to remove LQ variants from the Hi-Plex SNP set. This further indicates that different EP thresholds ought to be used depending on the genotyping method. In contrast to SNPs, there was no clear differentiation between EP values associated with HQ or LQ indels (Figure 4.5). This shows that indel detection remains challenging because of mapping

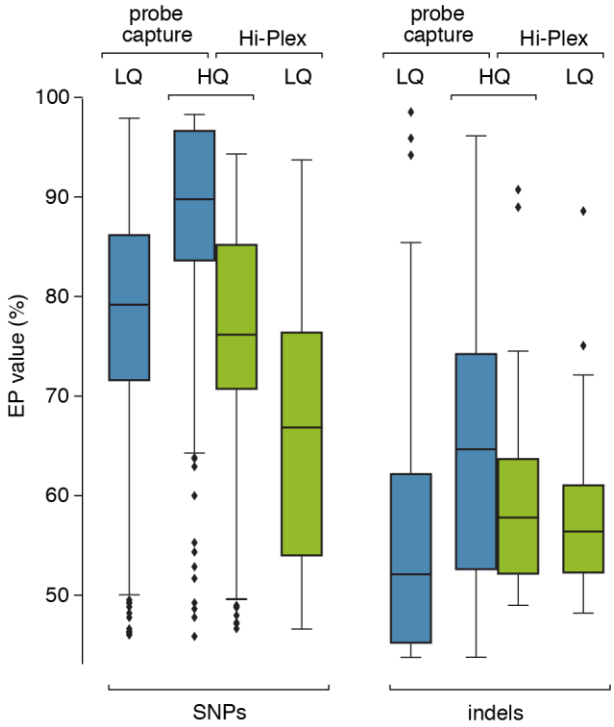


Figure 4.5 Distribution of Estimated Precision values in High Quality and Low Quality variant sets. For variants present in 78 genotypes and 147 amplicons, box plots show the distributions of Estimated Precision values for commonly identified (high quality, HQ) and uniquely identified variants (low quality, LQ) in the probe capture and Hi-Plex variant sets.

and/or realignment errors, and errors near repetitive regions (Li, 2014), thus leading to an incomplete indel set and underestimation of frameshift variants.

Using the empirically validated EP threshold of 80% for the probe capture variant set containing variants of 736 genotypes and 503 candidate genes resulted in 252,406 SNPs and 5,074 indels. The high genotype call consistency indicates that the VariantMetaCaller, at least for SNPs, was able to reliably integrate variant sets without losing genotype call quality. Moreover, using the VariantMetaCaller for precision-based filtering led to a higher genotype call rate compared to hard filtering of individual VC pipelines (Figure 4.4 C and D).

Validation of the resulting variant set in an F1 progeny

Mendelian inheritance in a segregating F1 progeny derived from a bi-parental cross was used as an accuracy measurement for the precision-based filtered variant set: Mendelian inheritance errors (MIEs) are most likely the result of erroneous genotype calls. The set of 736 individuals contained two parents and their respective F1 progeny of 29 individuals. The genotype calls of these individuals were used to calculate the number of MIEs. Out of the 257,480 variants, 10,669 contained a missing genotype call in either one or both parents (4%) and could not be tested. For the 246,881 remaining variants and 29 individuals, 89,789 MIEs were identified, of which 57,326 (63%) were due to a missing genotype call. The other 32,463 MIEs represent a genotype call error in only a fraction of all genotype calls among the 246,881 variants in this F1 progeny (< 0.5%). Moreover, these MIEs corresponded to 9,440 variant positions (4%) of which most had a MIE in a single individual (Supplemental Figure 4).

Effects of sequence variation on gene function

We investigated the consequences of sequence variants on predicted gene function, using the manually curated high quality gene models to annotate the variants with SnpEff (Cingolani et al., 2012). The complete annotation of functional effects for each of 252,406 SNPs and 5,074 indels is now available for the community to use. Out of the 257,480 variants, 65,225 resided in exon regions (25%) and 116,274 in intron regions (45%) corresponding to a density of 8.6 and 10.1 variants per 100 bp, respectively. Among the SNPs in coding regions, 38% were non-synonymous

substitutions, which is consistent with previous observations in *L. perenne* transcriptomes (Ruttink et al., 2013; Farrell et al., 2014; Paina et al., 2014).

A general overview of the abundance of high impact effects on gene function, listed per functional category or pathway is presented in Table 4.1. These include gain of stop codons, frame shifts and alterations in splice sites, as they are most likely to disrupt protein function, possibly leading to loss-of-function (LOF), and causing phenotypic variation. For instance, the variant set contained 256 stop gain variants, affecting 144 out of the 503 candidate genes. The position of each stop gain relative to the total CDS length could indicate the degree to which the protein is affected (Supplemental Figure 5). Most of these stop gain variants occur at low allele frequency across the germplasm collection. Additionally, 72 candidate genes were affected by splice site variants: 40 variants affected donor splice site and 47 variants affected acceptor splice sites. In line with the relatively low number of indels (5,074), only 20 frameshift variants were identified in 19 genes. In summary, naturally occurring LOF alleles could be readily identified in as much as one-third of the genes tested across various pathways that are important for plant growth and development.

This variant catalog can be exploited in a dual fashion: (1) to associate genomic variation with phenotypic variation using an association mapping approach, which we are currently performing for architectural traits and cell wall digestibility, or (2) to mine for rare defective alleles, i.e. variants that disrupt gene function or regulation, and to subsequently select carriers of these variants for detailed phenotypic analysis. For example, we observed naturally occurring alleles for the single copy genes *GIGANTEA* (*LpGI-01*) and *ENHANCED RESPONSE TO ABSCISIC ACID 1* (*LpERA1-01*), in which a premature stop codon truncates translation at 5% and 23% of the protein length, respectively. Crosses with the carriers of these putative null alleles could help to clarify the function of *LpGI-01* in the regulation of flowering time, circadian clock, and/or hypocotyl elongation (Mishra and Panigrahi, 2015) and *LpERA1-01* in meristem organization and the abscisic acid (ABA) mediated signal transduction pathway (Cutler et al., 1996).

Sequence variants were determined in a germplasm collection representing commercial cultivars, breeding populations and natural accessions, to ensure the downstream application in

current breeding programs. For instance, the 170 amplicons used to estimate the EP threshold, and to validate the genotype calls of the probe capture set, were designed to cover the genetic diversity in 28 genes putatively involved in flowering time and other phenotypic traits of interest to breeders. Design and validation of these amplicons is a clear illustration of the application of the variant set. Since a comprehensive set of SNPs and indels are now known for our breeding materials, detailed and customised design of PCR primers targeting specific SNPs in candidate genes spread across the genome, while avoiding polymorphisms in the flanking primer binding site, becomes feasible. Similar methods and criteria apply to the design of hybridisation probes for high density SNP arrays. We are currently using Hi-Plex amplicon sequencing as a very cost- and time-efficient method to screen hundreds of variants simultaneously in a few thousand genotypes, a scale required to screen for putative associations with phenotypic traits in our current breeding populations.

Reconstruction of divergent alleles enables better characterization of genetic variation

The prime goal of targeted resequencing is to *de novo* discover alleles that are divergent from the reference genome sequence. However, the capture efficiency of a divergent sequence is reduced with increasing sequence dissimilarity to the reference sequence for which the probes were designed. Additionally, reads may fail to map to highly divergent regions if the parameters for short read alignment are too stringent (Bertels et al., 2014). To circumvent mapping short reads to a single reference sequence, on which classical VC pipelines rely, we devised a *de novo* assembly strategy to reconstruct full-length alleles. First, *de novo* assembly of the captured reads was performed per individual genotype to reconstruct alleles for each of the 503 candidate genes in parallel. Next, all *de novo* assembled contigs from all 736 genotypes were aligned to the reference genome to sort out and extract all corresponding allelic fragments per candidate gene. Per gene, all contigs were clustered using the Overlap Layout Consensus assembler CAP3 (Huang and Madan, 1999) to collapse allelic redundancy and resolve fragmented gene sequences. This three-step approach results in a collection of alternative alleles assigned to each of the 503 candidate genes (on average 58 contigs per gene, range 4 to 203). The entire set of 29,320 CAP3 contigs is now available for the community to use. The value of this approach is that we can now

characterise genetic variation in regions of high sequence diversity where traditional short read mapping-based VC pipelines fail.

We demonstrate the value of this approach for *LpSDUF247*, but all analyses described below may be repeated for any of the other 502 candidate genes. *LpSDUF247* is a highly polymorphic gene co-segregating with the *S*-locus that determines the grass self-incompatibility system (Manzanares et al., 2016). CAP3 clustered the allelic diversity present in the 736 genotypes into 34 separate contigs. Alignment of CAP3 contigs identified a central region of the protein with high sequence divergence, while this region is virtually free from SNPs in the VC dataset, showing the limitations of read mapping based variant calling. Six contigs displayed high sequence similarity (98%) to the reference sequence or to other contigs and were removed. Within the remaining allelic contigs, a single exon encoded for the DUF247 protein. The translated proteins showed only 73 to 84% global sequence identity with each other (Figure 4.6 A). These data are consistent with the previously reported identification of at least five unique alleles of *LpSDUF247* with 80 to 90% protein sequence identity (Manzanares et al., 2016). Phylogenetic analysis confirmed that *de novo* assembled contigs were indeed novel alleles of *LpSDUF247* at the *S*-locus, and not of any of the 24 other DUF247 paralogs in the *L. perenne* genome (Supplemental Figure 6 and Supplemental Figure 7).

Next, we analyzed the distribution of *LpSDUF247* alleles across the *L. perenne* germplasm collection. The 28 novel alleles were added to the reference genome sequence, thus complementing the reference *LpSDUF247* allele, and giving reads the opportunity for near-perfect mapping at their respective allele. Differential read mapping *across* the alleles in a multi-allelic context was then used to score which alleles are present in each genotype. Mapping reads in a multi-allelic context eliminates the need for variant calling, but only if alleles are sufficiently divergent so that differential read depth can be used to identify which alleles are present per genotype. Near-perfect mapping of the raw reads onto the newly constructed *LpSDUF247* alternative alleles confirmed their existence in the *L. perenne* germplasm collection, except for allele 31 which had no read support (Figure 4.6 C). This also shows that the capture efficiency of

120 bp probes was sufficient to detect alleles with as little as 80% sequence identity to the reference genome sequence.

In the vast majority of genotypes (501 out of 736) the reads almost exclusively mapped onto a combination of two *LpSDUF247* alleles, often at similar RD, and only a minor fraction (<5%) of reads mapped to additional alleles (Figure 4.6 C). There was no bias for combinations of alleles across natural accessions, breeding populations and current cultivars, and clear segregation of alleles was observed in the F1 progeny (n = 29) of a bi-parental cross that was included in the set of 736 individuals (Supplemental Figure 8). Furthermore, 57 genotypes displayed reads mapping only to a single allele, suggesting either homozygosity or the failure to capture and sequence yet undiscovered alternative alleles with even stronger sequence divergence to the reference genome sequence used to design the probes. Finally, 177 genotypes displayed read depth spread over three or more alternative alleles. In 65 of them, the higher allele count could be explained by a consistent segregation of *LpSDUF247*-04 with *LpSDUF247*-28 suggesting a gene duplication, in combination with an additional, variable third allele. In the remaining 112 genotypes, the observation that reads map to more than two alleles in a multi-allelic reference genome, could indicate ambiguity of read mapping between closely related alleles, or the presence of additional alleles derived from cross-over events.

Although *LpSDUF247* was the most extreme case of sequence divergent in alternative alleles, Supplemental Figure 9 presents four other candidate genes with different levels of divergence, global or local. The alternative alleles of *LpMAX3-01* and *LpETR1-01* showed only local sequence divergence, at the introns and 5'UTR regions respectively. The sequence variation of *LpFT-04* was captured in only four contigs, explaining why the variant density across the gene region was low, especially when low frequent SNPs were filtered out. Taken together, the analysis of *LpSDUF247* demonstrates the rich sequence diversity that can be mined for in this catalog of genomic diversity across 503 candidate genes.

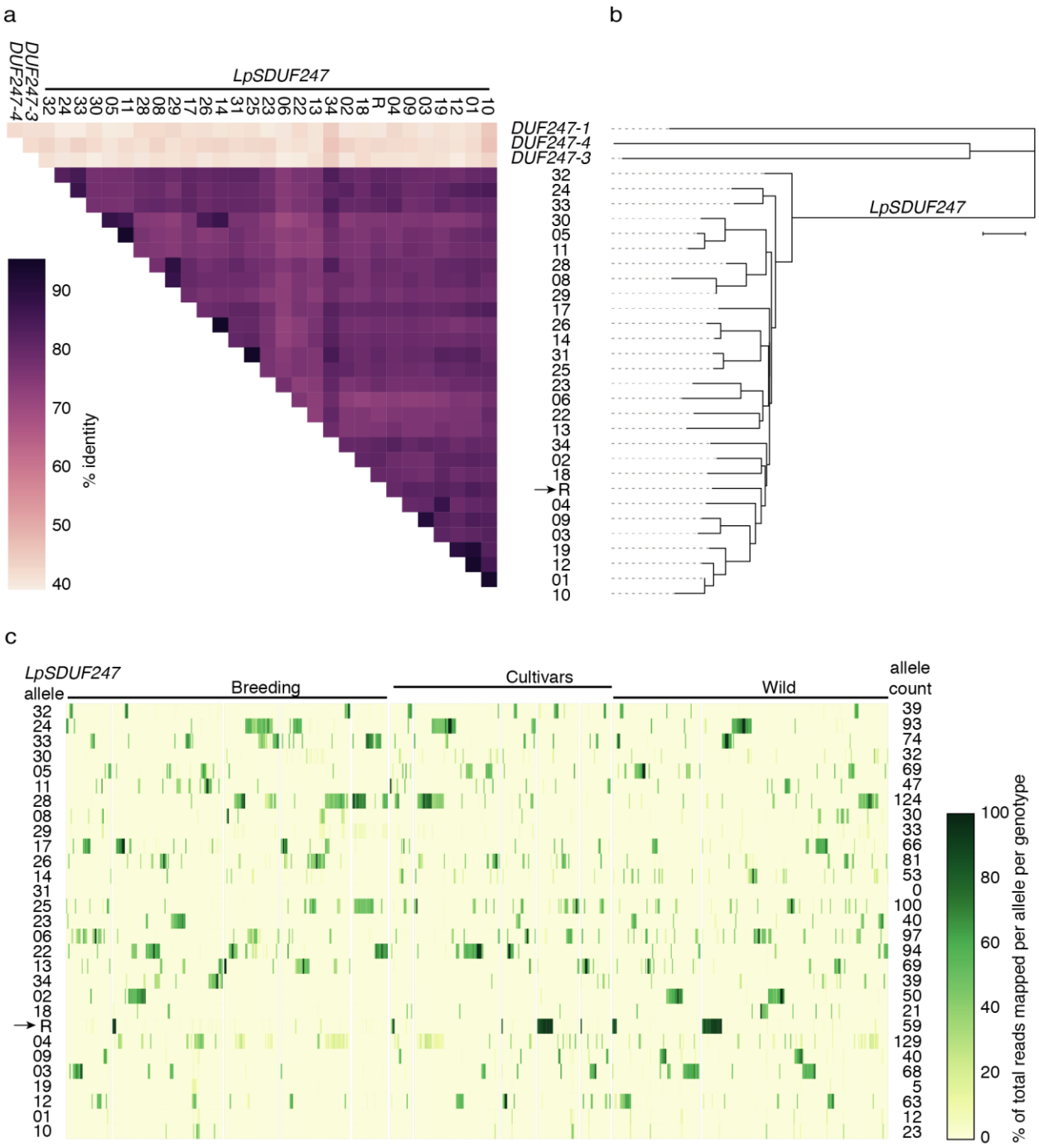


Figure 4.6 Sequence diversity and distribution of 28 newly identified alleles of *LpSDUF247* across breeding populations and natural accessions. A similarity matrix (A) and phylogenetic tree (B) were built using the protein sequences of 28 alleles and reference sequence (R) of *LpSDUF247,* together with three additional *DUF247* genes. Panel C gives an overview of the distribution of the *LpSDUF247* alleles across the gene pool. The alleles present per genotype were identified by mapping the reads to a multi-allelic reference genome, and calculating the ratio of average read depth per allele over the total number of reads mapping to *LpSDUF247* alleles.

4.4 Author Contribution

A.R., T.R., I.R.R., F.v.P. and H.M were involved in experimental design. H.M., F.v.P., A.H., T.R. and E.V. performed phylogenetic analysis and selected the target regions. S.v.G. was responsible for the experimental work. T.A. and S.B. provided genome sequence data. B.S. provided background information on SI. E.V., T.R. and K.V. designed the research methodology. E.V. and T.R. performed data analysis. E.V., T.R., K.V., I.R.R and A.H. contributed to writing the manuscript.

5 Genomic Variation in the FLOWERING TIME Gene Family of Perennial Ryegrass[3]

The transition from vegetative growth to flowering is one of the most important processes in plant development. The genetic control of flowering is of great interest to breeders, as the timing and intensity of flowering affects agronomic traits such as biomass yield and quality (digestibility) and seed yield. The regulatory pathway involves the perception and processing of a diverse range of environmental and internal signals, and converges on genes of the FLOWERING LOCUS T gene family. Rather than selecting for phenotypes, we want to uncover the genomic variation in the FT gene family that could possibly cause variation in flowering time. The FT gene family of *Lolium perenne* contains 18 members, including novel paralogs that may be functionally redundant with previously described genes of this family. Five FT family members were resequenced in 736 genotypes including natural accessions, commercial cultivars and breeding material, revealing high degrees of genomic sequence diversity. Many deletions occur in the promoter and UTR, which may affect transcript expression or stability. We identified several genotypes with sequence variation at positions coding for amino acid residues essential for PEBP protein function. Using the sequence context, haplotypes were reconstructed to estimate their distribution in a broad genotype collection. This case study illustrates how variant effect prediction based on structural sequence features leads to the identification of interesting alleles. Carriers of these alleles in breeding populations and natural accessions can be used for future functional analysis, or can be incorporated into the breeding program.

[3] This chapter is based Veeckman, E., Vandepoele, K., Asp, T., Roldán-Ruiz, I., and Ruttink, T. (2016b). Genomic Variation in the FT Gene Family of Perennial Ryegrass (Lolium perenne). In Breeding in a World of Scarcity, I. Roldán-Ruiz, J. Baert, and D. Reheul, eds (Cham: Springer International Publishing), pp. 121-126. For author contributions, see page 86.

5.1 Introduction

The transition from vegetative growth to flowering is one of the most important processes in plant development, and the timing is strictly controlled to coincide with conditions that enhance production of seeds and fruits. In perennial ryegrass, flowering is induced by a period of vernalization, followed by long days at higher temperatures. Because the timing and intensity of flowering affects agronomic traits such as biomass yield and quality (digestibility) and seed yield, the genetic control of flowering traits is of great interest to breeders. The regulatory network that controls flowering involves the perception and processing of a diverse range of environmental and internal signals and integrates them into a single decision: to flower or not to flower. This network slightly differs across species but always converges on genes of the PEBP (phosphatidylethanolamine-binding proteins) gene family, also known as the FLOWERING LOCUS T (FT) gene family. Phylogenetic analyses have revealed the complex evolutionary history of this gene family, with strong evidence of lineage specific expansion, creating the possibility for functional diversification between clades, as well as functional redundancy of paralogs within clades (Faure et al., 2007; Wickland and Hanzawa, 2015).

Two members of this gene family have been studied intensively: *FT*, a positive regulator of flowering, and *TERMINAL FLOWER 1* (*TFL1*), a negative regulator of flowering. *MOTHER OF FT AND TFL1* (*MFT*) is another FT gene family member, which was found to act redundantly with other members of the FT gene family (Yoo et al., 2004). Only three members of this family, *LpFT3, LpTFL1* and *LpSFT*, have previously been characterized in *L. perenne* and different alleles of *LpFT03* have been shown to associate with variation in heading date (Jensen et al., 2001; King et al., 2006; Fiil et al., 2011; Skot et al., 2011). We used this gene family as a case study to see how we can identify genomic variation and investigated whether sequence variation affected residues that are critical for PEBP protein function, or may be associated with phenotypic variation in flowering time.

Small-scale genomic variation arises by natural processes, and results in single nucleotide polymorphisms (SNPs), insertions and deletions of a short stretch of nucleotides and recombination. The detection of these variants in candidate genes involved in complex biological processes are of interest to plant breeders, as genomic variation could cause interesting phenotypic variation: variants occurring at positions in the genetic code that are essential for gene function may change the gene activity or function, thereby affecting phenotypic characteristics of an individual. This way, the results of the analysis of genomic sequence variation can be directly applied in agriculture, such as the selection of plants with the best agronomic characteristics for growth, yield, quality or optimal flowering based on their genomic sequence.

We are interested to identify genomic variation in perennial ryegrass that putatively underlies phenotypic variation in adaptive traits and traits that are relevant for breeders. This information can be used in association mapping approaches to identify loci and causal genes with relatively large phenotypic effects on agronomic traits (plant architecture and forage quality). Furthermore, we want to use bioinformatics approaches to exploit this catalog of genomic sequence diversity to predict the effect of SNPs and indels on gene functions. For this, we are guided by some key questions: in which gene is the variant present? What is the function of that gene? Is the position of the polymorphism conserved or essential for gene function? Here, we present a case study to illustrate the pipeline from the selection of candidate genes, to the identification and in-depth interpretation of genomic variation.

5.2 Material and Methods

Gene family annotation and phylogenetic analysis

Protein sequences of 18 *Brachypodium distachyon* genes were extracted from the PLAZA 3.0 Monocots database (Proost et al., 2015) and used for a tBLASTn search against a draft assembly of the *L. perenne* reference genome sequence (Byrne et al., 2015), revealing 18 candidate orthologous loci. Overlapping gene models, predicted using MAKER v2.3 (Cantarel et al., 2008), were checked and corrected based on a multiple sequence alignment of the whole gene family (HOM03M000266, PLAZA 3.0 monocots) and RNA-seq data. Annotated gene sequences are

deposited in GenBank under accession numbers KR706144 to KR706161. For phylogenetic analysis, the set of 18 *LpFT* genes was complemented with homologs of *B. distachyon* (18), *Hordeum vulgare* (12), *Oryza sativa* spp. *japonica* (19), *Zea mays* (25) and *Arabidopsis thaliana* (6). A multiple sequence alignment was created using MUSCLE (Edgar, 2004) and a phylogenetic tree was constructed using PhyML (Guindon et al., 2010) applying the JTT substitution model, and bootstrap values were calculated using 1000 replicates. Human phosphatidylethanolamine-binding protein 1 preproprotein (gi|4505621, NP_002558.1) was used to root the tree.

SNP and indel discovery

The plant collection used for variant discovery comprised 736 genotypes, including natural accessions, breeding material and current cultivars. From that collection, targeted resequencing of 503 genes of interest was performed using SureSelect probe capture enrichment of indexed genome shotgun libraries followed by Illumina HiSeq sequencing (PE 2x100). The raw reads were trimmed, mapped to a the draft genome sequence using BWA-mem (Li and Durbin, 2009) and polymorphisms (SNPs and short indels) were identified using GATK v3.2-2 (McKenna et al., 2010), as described in Ruttink et al. (2015). Multi-allelic SNPs and indels were filtered out and genotype calls were filtered on read depth (DP ≥ 6) and likelihood score (GQ ≥ 30). A total of 1645 SNPs and 505 indels are located in the five *LpFT* genes reported here.

Haplotype reconstruction

To perform haplotype reconstruction, fastPHASE (Scheet and Stephens 2006) was used for imputation of missing genotype calls and phasing, as described in van Parijs (2016). The population structure needed for imputation was determined with fastStructure (Raj et al. 2014), revealing four subpopulations. The largest subpopulation (A) comprises genotypes from Central and Northern Europe and derivations, whereas genotypes from New-Zealand and warmer European regions were assigned to another subpopulation (B). The third subpopulation (C) contains genotypes derived from a commercial breeding program. Finally, two parents and 30 F_1 plants of a QTL mapping population were grouped together. This subpopulation was used as a positive control (e.g. for Mendelian segregation) and was not used for further analyses. To avoid inflation of distinct but rare haplotypes, we excluded variant positions with a minor allele

frequency (< 5% for both SNPs and InDels) and low variant call rate (< 200 genotype calls for SNPs and < 625 genotype calls for InDels). Of the total set of 736 genotypes, only 600 genotypes with sufficient target region coverage were selected to avoid imputation on too much missing data per genotype. Two haplotypes were reconstructed for each diploid genotype and assigned to the corresponding subpopulations. Average distance trees were built for collapsing the unique set of haplotypes per gene into major haplotype classes. For each haplotype class, the most frequent haplotype was selected to build an average distance tree as representation of the major haplotype classes.

5.3 Results and Discussion

Delineation of the FT gene family in perennial ryegrass

The FT gene family is characterized by the PEBP domain, which is represented in the genomes of all three major phylogenetic divisions; eukaryotes, bacteria and archaea. BLAST searches with all 18 PEBP proteins of *B. distachyon* against a draft *L*. *perenne* genome sequence revealed 18 FT gene family members (Figure 5.1), consistent with a similar number of FT genes in other monocot species (see Materials and Methods). *LpFT3, LpSFT,* and *LpTFL1* (here named *LpFT03*, *LpFT01* and *LpFT07*, respectively) have previously been described (Jensen et al., 2004; King et al., 2006; Fiil et al., 2011; Skot et al., 2011).

The FT gene structure typically consists of four highly conserved exons and three introns of variable length. Some genes across the phylogenetic tree lack an intron suggesting lineage specific intron loss: *ZCN2* and *ZCN20* have three exons resulting from the fusion of exon1 and exon2, whereas exon3 and exon4 appear to be fused in the common ancestral gene of *LpFT3*, *HvFT1* and BD1G48830. The specific loss of an intron could be the result of reverse transcription of spliced mRNAs followed by homologous recombination of the cDNA with the genomic copy of the gene (Roy and Gilbert 2006).

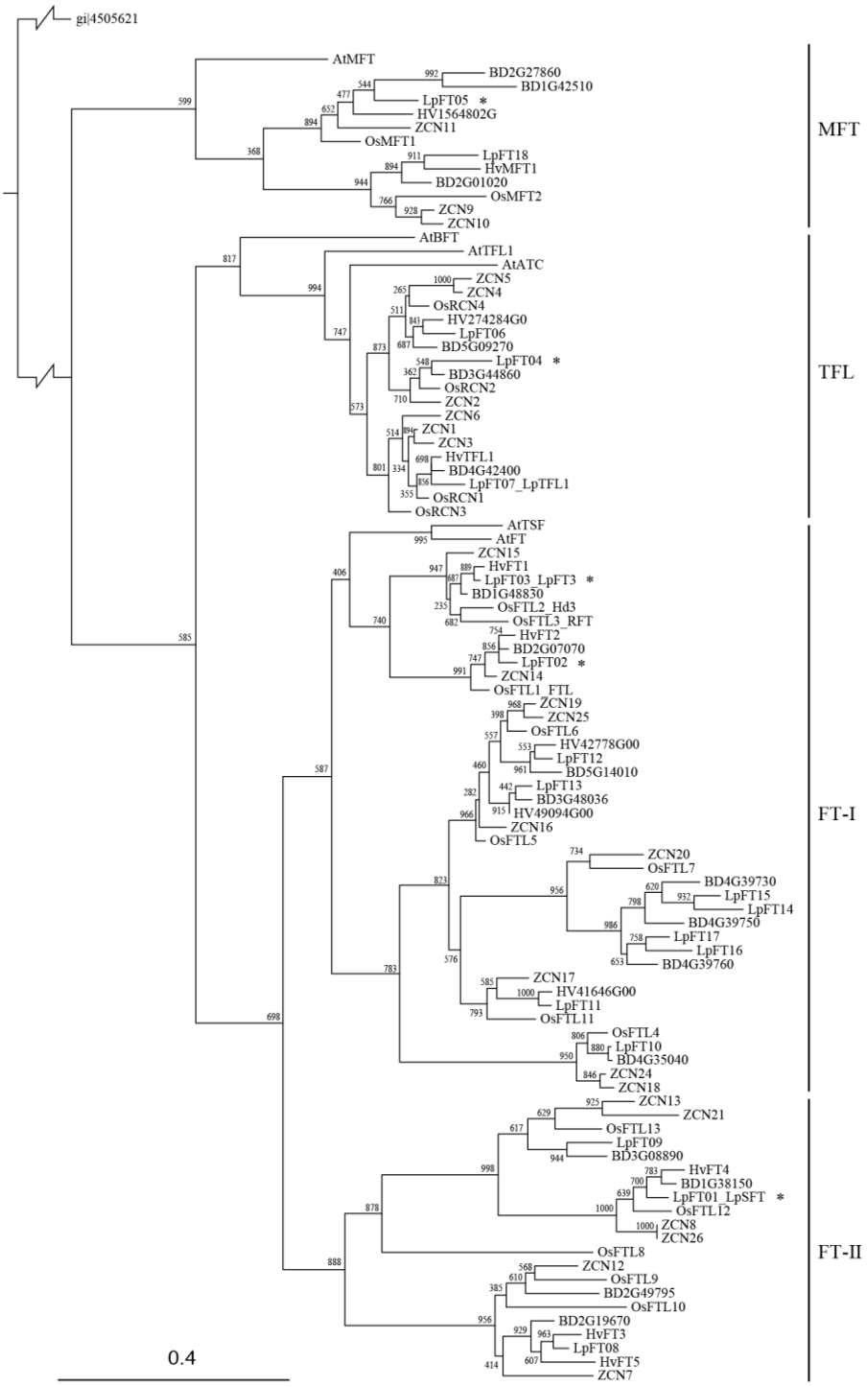


Figure 5.1 Phylogenetic tree of the FT gene family of *L. perenne* (Lp), *B. distachyon* (BD), *H. vulgare* (Hv), *O. sativa spp. japonica* (Os), *Z. mays* (ZCN) and *A. thaliana* (At). Human PEBP1 (gi|4505621) was used to root the tree. Support values for branches are represented by bootstrap values (1000 replicates). Three major clades (MFT-like, TFL1-like, and FT-like) including two subclades (FT-like I and II) are shown. An asterisk indicates LpFT genes selected for targeted resequencing.

Expansion of the FT gene family in grasses

Previously published phylogenetic analyses of the PEBP domains of the FT genes in *H. vulgare*, *O. sativa* spp. *japonica* and *A. thaliana* revealed that the FT gene family contains three major clades (Chardon and Damerval, 2005; Faure et al., 2007). *LpFT05* and *LpFT18* belong to the MFT-like clade, *LpFT04*, *LpFT06* and *LpFT07* belong to the TFL1-like clade and the remaining 13 *LpFT* genes belong to the FT-like clade (Figure 5.1). This last clade can be subdivided into two subclades: the FT-like clade I, containing *AtFT* and *AtTSF*, and the FT-like clade II, containing only grass and cereal-specific genes.

The number of PEBP genes in grasses and cereals is three to four times larger than that in *A. thaliana*, due to several whole-genome duplications and tandem duplications that are specific for the grass lineage. These ancient duplications are revealed by the consistent grouping of orthologs in clades containing members of various grass species. The most parsimonious hypothesis suggests that two *MFT-like*, two *TFL1-like* and at least eight *FT-like* genes were present in the ancestral grass genome (Chardon and Damerval, 2005). The orthologous relationships within subfamilies are often difficult to deduce because genes likely evolved at least partially independently in each taxon by duplication and possible gene loss.

We selected five FT genes representing the four (sub)clades of the phylogenetic tree (indicated with an asterisk in Figure 5.1) for detailed analysis (see below). One of them, *LpFT03*, has previously been shown to associate with heading date (Skot et al., 2011). All *LpFT* genes showed the typical residues at positions that were conserved in the whole gene family (residues coloured purple in Figure 5.2). Moreover, positions that were only conserved within clades were also conserved in the *LpFT* gene family members (results not shown), suggesting that these paralogs share biochemical function.

Identification of sequence variation in five family members

An average of 350 variant positions (SNPs and indels) were identified for each of the five target genes, but strong differences were found among genes (Figure 5.1). As expected, coding sequences contain fewer variants than promoter, untranslated regions or intron sequences.

Despite having sufficient read data for *LpFT04*, this gene shows remarkably little sequence variation.

Next, we investigated sequence variation in the external loop and key residues of the anion ligand binding site (Ahn et al., 2006; Danilevskaya et al., 2008) (Figure 5.2). Projection of non-synonymous SNPs and indels onto the protein sequence alignment of the five selected genes (Figure 5.2) revealed that most of them code for non-essential residues. In *LpFT03*, however, some variation in critical residues was observed, including the P136 in the external loop and the Y151 of the LYN triad, which are normally completely conserved in *FT-like* genes (Ahn et al., 2006). *TFL1-like* and *MFT-like* genes in other species also showed variation in the external loop, which may affect the surface charge around the ligand-binding pocket and thus may affect protein function or activity. On the other hand, no variation was found in the external loop of

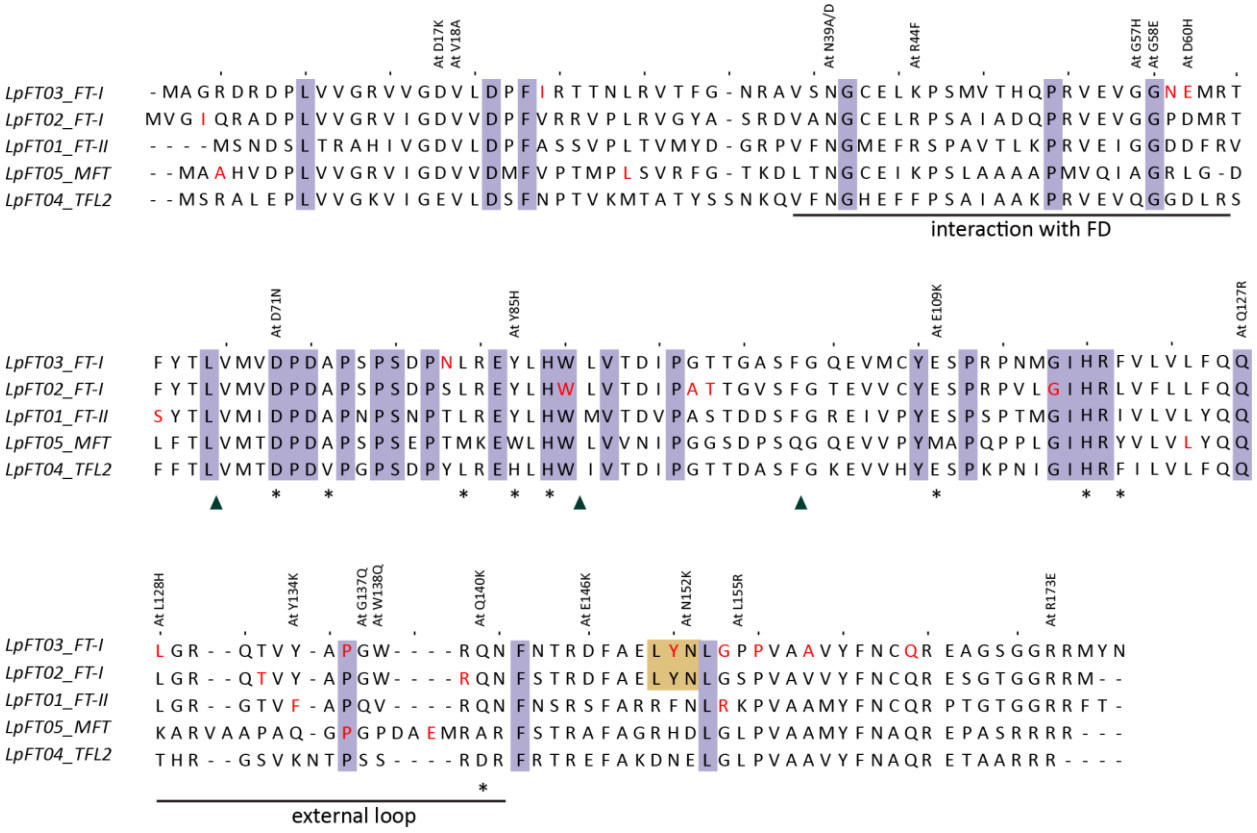


Figure 5.2 Protein sequence alignment of five *LpFT* genes and delineation of conserved residues across the gene family (purple). Exon borders are indicated by black triangles. Key residues of the anion ligand binding pocket are indicated with an asterisk (Ahn et al., 2006; Danilevskaya et al., 2008). The LYN triad sequence of the FT-like clade I is coloured orange (Ahn et al., 2006). Residues that are essential for molecular function of FT or for interaction with FD or TCP proteins are given on top of the alignment (Ho and Weigel 2014). Residues with non-synonymous substitutions found in the genepool collection are colored red.

LpFT04. For *LpFT02*, a SNP introduces a premature stop-codon at W89 at the end of the second exon. This mutation occurs only in heterozygous state in 11 of the resequenced genotypes, and can be considered a rare defective allele.

Ho and Weigel (2014) identified residues that were critical for the molecular function of FT and residues that are essential for interaction with FD or TCP proteins (Figure 5.2, residue numbers refer to amino acid positions in *AtFT*). Only for two of these residues a non-synonymous SNP was detected: an alternative residue at D60 could lead to differential interactions with FD and sequence variation at L128 may affect the molecular function of *LpFT3*.

Distribution of haplotypes across a broad genotype collection

Sequence diversity within a broad genotype collection may lead to phenotypic differences in flowering traits. Most SNPs and indels however, lie in non-coding regions (such as promoter, UTR, intron regions) (Table 5.1), making it difficult to directly deduce their effect on gene function or activity. Skøt et al. (2011) have detected some deletions in the promoter region that could be associated with flowering time. We coupled the alleles at SNP and indel positions throughout the gene to reconstruct phased haplotypes, which may improve interpretation of sequence variation. To reconstruct representatives of major haplotypes, we removed low quality and low frequency polymorphisms from the set of variants. Next, we used FastPHASE (Scheet and Stephens 2006) to impute missing data and reconstruct full-length haplotypes. The resulting set of unique haplotypes per gene was clustered into a phylogenetic tree, and highly similar branches were manually collapsed to summarize major haplotype classes. *LpFT04* has the fewest major haplotypes (4), consistent with the low degree of sequence variation in this gene (Table 5.1). For the other genes more major haplotype classes could be delineated (8 for *LpFT01*, 10 for *LpFT02*, 11 for *LpFT03* and 9 for *LpFT05*).

Table 5.1 Distribution of SNPs and indels for five *LpFT* genes in coding and non-coding regions at minor allele frequency (MAF) of 1% across all 736 genotypes with positive observations. Non-synonymous SNPs in the CDS are given between brackets.

		LpFT01	**LpFT02**	**LpFT03**	**LpFT04**	**LpFT05**
Coding	Length (bp)	519	528	531	513	534
	SNPs / indels	17(3) / -	24(7) / -	27(11) / -	- / -	23(5) / -
Non-coding	Length	2396	3032	1537	2365	1759
	SNPs / indels	223 / 62	253 / 65	114 / 71	27 / 3	158 / 41

Next, we investigated whether the genetic structure of the genotype collection coincided with the distribution of the haplotypes. FastStructure (Raj et al. 2014) was used to group the 736 genotypes, revealing three subpopulations: genotypes originating from Central and Northern Europe and derivations (A), from New-Zealand and warmer European regions (B) and genotypes derived from a commercial breeding program (C). Some genes show a similar distribution across subpopulations per haplotype, eg. *LpFT04* and *LpFT05* (horizontally aligned pie charts in Figure 5.3). Notably, for *LpFT03* haplotype 10 and 11 are much more abundant in subpopulation A. In contrast, haplotype 9, which is closely related to haplotype 10, is more abundant in subpopulation B, as are haplotypes 1 to 6.

The abundance of the major haplotype classes per subpopulation was also determined, as shown in the vertically aligned distribution pie charts in Figure 5.3. For *LpFT02*, *LpFT04* and *LpFT05* these distributions look alike, indicating that sequence variation of these genes is equally spread over the three subpopulations. Haplotypes of *LpFT01* and *LpFT03* on the other hand, display differential representation between the subpopulations. For *LpFT01* the profile of subpopulation C differs from the profiles of subpopulations A and B. For *LpFT03,* differentiation is more pronounced and all three subpopulations show a different composition of major haplotype classes.

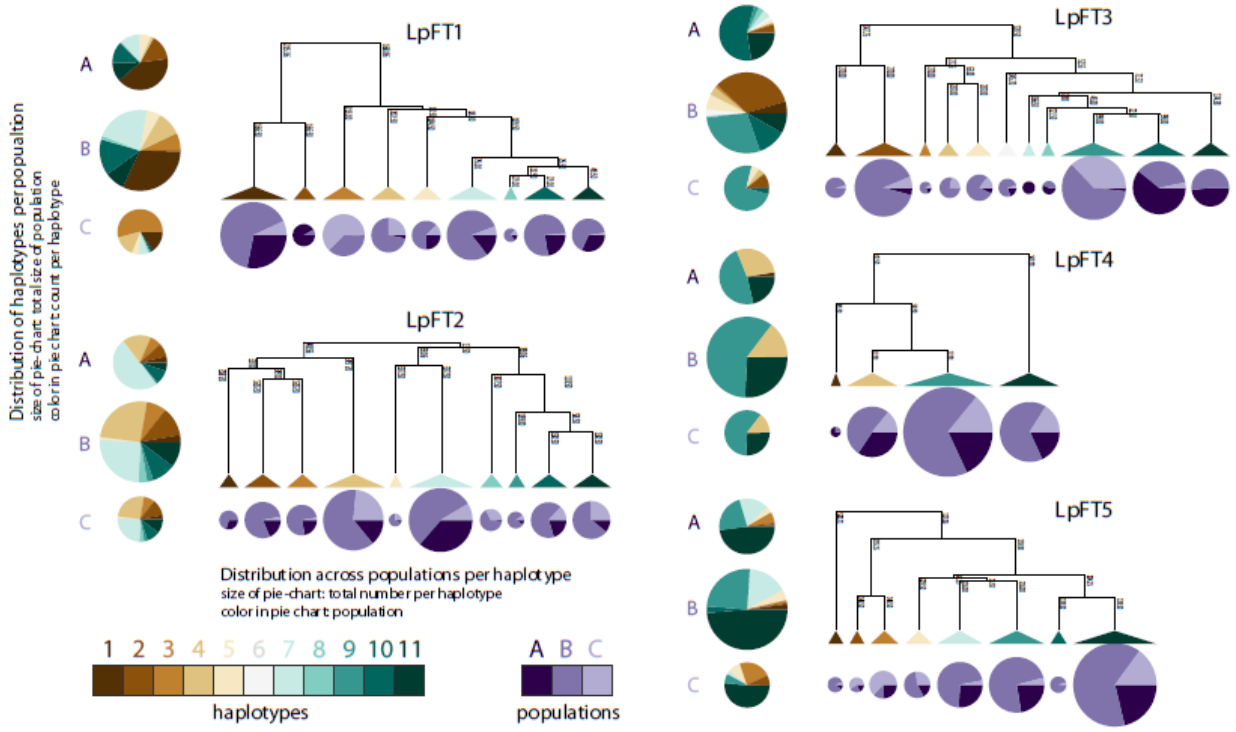


Figure 5.3 Distribution of haplotypes of five *LpFT* genes across subpopulations. For each gene haplotypes were reconstructed using imputed and phased SNPs and indels. After clustering in a phylogenetic tree, branches were collapsed in order to only present the major haplotypes. For each haplotype the distribution across the three subpopulations (A: Northern and Central Europe, B: New Zealand and warmer European regions, C: commercial breeding program) is shown in a pie-chart (purple) below the phylogenetic tree. The area of the circles corresponds to the number of times a certain haplotype is observed. Additionally, for each subpopulation (A, B, C) the distribution of the major haplotypes is shown at the left of the phylogenetic trees. The area of the circles corresponds to the number of genotypes within each subpopulation.

5.4 Conclusion

We have completed the FT gene family of *L. perenne* by identifying 18 family members in the draft genome sequence. All newly characterized genes contain amino acid residues that are conserved throughout the whole gene family, suggesting that all encode functional PEBP proteins. Furthermore, *L. perenne* members contain the clade-specific residues that differentiate between MFT, TFL and FT functions. In this way we have identified several novel paralogs that may be functionally redundant with previously described genes in the TFL1-like clade and the LpFT-like clade I.

Five FT family members were resequenced in 736 genotypes, revealing high degrees of genomic sequence diversity. The sequence variation across the *L. perenne* genepool may be used in two

complementary ways. First, we have identified several genotypes with sequence variation at critical residues. Non-synonymous substitutions that reside in the external loop may cause changes in the surface charge and have consequences for the biochemical function. Furthermore, the defective W89* allele of *LpFT02* could be useful to study the role of *LpFT02* and its putative functional redundancy with *LpFT03*. In addition, many deletions occur in the promoter and UTR, which may affect transcript expression or stability, but the effects are more difficult to predict by bioinformatics alone. Haplotype reconstruction could be exploited to combine sequence variation in coding and non-coding regions, and to perform association mapping with reduced multiple testing correction. The distribution of haplotypes across natural accessions, breeding populations, and current cultivars may further reveal whether breeding has selected for particular alleles.

5.5 Author Contribution

E.V. and T.R. designed the research methodology and performed the data analysis. T.A. provided genome sequence data. E.V., K.V, T.A., I.R.R and T.R. contributed to writing the manuscript.

6 Screening Breeding Populations for Variants Associating with Heading Date and Plant Height[4]

Association genetics is a forward-genetics approach to link phenotypic variation to genotypic variation and is a powerful tool to dissect the genetic basis of complex traits. We aim to identify alleles associating with heading date and leaf elongation, two important traits for perennial ryegrass breeding. We describe a high-throughput genotyping assay for 28 candidate genes that was designed for efficient screening of five breeding populations and a natural accession. Heading date was determined for each population, and leaf length was monitored throughout one growing season to estimate leaf elongation rates. A single SNP in *LpFT-03* was significantly associated with heading date. The associating substitution in the first exon confirmed the existence of haplotypes previously identified by Skot et al. (2011), and their relation to earlier or later heading. For leaf elongation, a single SNP in *LpMADS-01* was significantly associated with leaf length after autumn growth. Although no associations were detected between this polymorphism in *LpMADS-01* and other leaf elongation traits, such as spring growth rate, this is an interesting candidate marker, as it is located in the first intron that is involved in regulation of LpMADS-01 expression.

[4] For author contributions, see page 106.

6.1 Introduction

Plant breeding often relies on phenotypic selection: a breeder selects in his/her opinion the best individuals to progress. As genotypic variation lead to variation in phenotype, a breeder is inherently selecting for a (combination of) beneficial allele(s). The identification of alleles that are linked to a difference in phenotype is mostly done through association genetics. Association mapping (or linkage disequilibrium (LD) mapping) is a method to link phenotypes to genotypes, by exploiting historical recombination events at a population level that maintain LD between a polymorphic marker and a specific phenotypic trait (Nordborg and Weigel, 2008). This is a powerful tool to dissect the genetic basis of complex traits and offers a high genetic resolution, but the detection of associations is statistically difficult and depends on the LD extent determined by the population structure (Auzanneau et al., 2007). There are two main approaches: association studies based on candidate genes (CG) or based on testing the entire genome (Genome-wide association study, GWAS). The latter allows for the identification of novel genes underlying a phenotypic trait, but have low power owing to the number of independent tests performed (McCarthy et al., 2008). CG association studies tend to have a high statistical power, but are directed by the choice of candidate gene(s) and are therefore not capable of discovering new (combinations of) genes underlying a phenotypic trait (Amos et al., 2011).

Because perennial ryegrass is allogamous, LD decays much faster compared to inbreeding species (Smith et al., 2009; Brazauskas et al., 2010; Fiil et al., 2011). With rapid LD decay, CG association mapping is better suited to relate sequence variations in selected genes to specific traits of interest (Zhu et al., 2008). This has been successful in perennial ryegrass for genetically well-described traits such as flowering time (Skot et al., 2005; Skot et al., 2007), vernalization requirement (Asp et al., 2011), lignin content and cell wall digestibility (van Parijs et al., 2016), and even for complex traits such as drought tolerance and winter survival and spring regrowth (Yu et al., 2015).

Flowering time and leaf elongation are regulated by multiple genes, and are important agronomic traits for perennial ryegrass breeding. Flowering time has a major influence on feed quality and farm management practices. It is induced by a period of short days and low temperatures

(vernalization), followed by longer days and higher temperatures. As plants start to flower, fiber content increases with corresponding reduced digestibility. Flowering time is also correlated with seasonal herbage yield, plant height, leaf length, rate of tiller turnover, seed yield, abiotic stresses resistance, and persistence (Kemp et al., 1989; Laidlaw, 2004, 2005; Skot et al., 2007). Leaf elongation is an important factor contributing to plant growth and yield production (Horst et al., 1978). Therefore, fast growing perennial ryegrass cultivars are desirable in forage. Leaf elongation is controlled by cell elongation and cell division rates, but the cellular and molecular factors accounting for the genetic variation in leaf elongation are still not well understood (Xu et al., 2016)

As genetic variation is the foundation of phenotypic variation, we want to test whether variations in flowering time and leaf elongation observed in perennial ryegrass populations can be explained by the genotypic variation in candidate genes controlling these traits. Here, we present a CG association mapping study in five breeding populations and a natural accession, to detect alleles associating with variation in flowering time and leaf elongation. Heading date was determined for each plant as measure for flowering time, and leaf length was monitored throughout one growing season as a measure for leaf elongation. As flowering time is a well-studied trait in crops and the underlying genes are well described in model species, candidate genes were selected and orthologs were identified in the *L. perenne* genome. A high-throughput genotyping assay for 28 candidate genes was developed, to accurately determine the alleles present in each population. The molecular knowledge that can be derived from the associating variants and corresponding genes can be used to reveal the genetic basis of flowering time and leaf elongation and to guide future breeding programs.

6.2 Material & Methods

Description of plant materials

Six populations were screened in total (Table 6.1): one natural accession from Spain, ba12990, and five breeding populations. Population 1853-7 is the F2 of a single component of a poly-cross with seven parents that are intermediate heading. The other four breeding populations are the F2 of one or two crossing cells. In this scenario, a crossing cell contains two parents that can only

pollinate each other. Seeds were harvested from each parental plant, and grown in a tray. In total, 80 individuals of the F1 were further grown in a random setup on a field plot surrounded by barley. The seeds that resulted from cross-pollination across the F1 individuals were harvested and gave rise to the F2. The breeding population asturionxWAR10 is the result of a single crossing cell containing two late heading parents. The breeding populations 5297xWAR10 and ba12990xplenty were derived from two crossing cells, containing late and intermediate heading parents, respectively. The last population, ba12990x5554 was also derived from two crossing cells, containing one common parent. All parents were intermediate heading.

Table 6.1 Overview and background of the populations

Population	Background	Heading	# parents	# individuals
1853-7	F1 of one component of polycross	Intermediate	1+6	210
5297xWAR10	F2 of two crossing cells	Late	4	338
asturionxWAR10	F2 of 1 crossing cell	Late	2	357
ba12990	Natural accession, Spanish origin	-	-	142
ba12990x5554	F2 of 2 crossing cells with one common parent	Intermediate	3	324
ba12990xplenty	F2 of 2 crossing cells	Intermediate	4	358

A total of 1729 individuals were sown in August 2016. The plants were transferred to individual containers of 20 cm diameter with drip irrigation, and moved outside on February 20, 2017. Plants were cut a first time after all plants of the same population started flowering. Three additional cuts were performed simultaneously for all six populations, on July 4 (C1), August 17 (C2) and September 25 (C3), 2017, to simulate the ryegrass cutting regime and assess regrowth. Fertilizer (NPKMgO 16:8:22:3) was applied on April 21, 2017 and immediately after each cut.

Phenotypic traits

An overview of phenotypic traits is presented in Table 6.2. For each individual, leaf length was determined as the length of the longest leaf, measured from the base of the plant. In spring, leaf length was measured weekly from February 28 until May 29, 2017 or until the plant started flowering. For the linear range of growth, spring growth rate (SGR) was calculated as the slope of

the linear fit to the growth curve for each plant using growing degree days (GDD) with base temperature 0 °C, counting from January 1, 2017. Heading date was determined for each plant as the GDD on which a plant showed three ears. The plants were cut three times and leaf length was measured two and six weeks after each cut as a measure of regrowth. From September 25 until November 13, 2017, leaf length was measured weekly, and the mean of the last three measurements was used as maximal leaf length after autumn growth. For each trait and per population, outliers were detected based on interquartile ranges (IQR), and treated as missing values. Outliers were defined as observations that fall below the first quartile minus 1.5 times the IQR or above the third quartile added 1.5 times the IQR. Pearson correlation coefficients between HD and leaf elongation traits were calculated with Scipy (v0.19.0). SGR was corrected for HD effect per population (SGR_{HDcorr}). Median SGR was determined at each HD, and subtracted for each individual with the corresponding HD. The mean SGR at the median HD was added to obtain results that can be interpreted as the SGR as if the individual was flowering at median HD of the corresponding population.

Table 6.2 Overview of phenotypic traits used to calculate associations. GDD: growing degree days.

Trait	Measure
HD	Heading date GDD on which a plant showed three ears (with base temperature 0 °C, counting from January 1, 2017)
SGR_{HDcorr}	Spring growth rate corrected for HD. Slope of the linear fit to the spring growth curve using GDD (base temperature 0 °C, counting from January 1, 2017). Correction for HD was performed per population.
C1-W2	Leaf length measured two weeks after the first cut (July 4, 2017)
C1-W6	Leaf length measured six weeks after the first cut (July 4, 2017)
C2-W2	Leaf length measured two weeks after the second cut (August 17, 2017)
C2-W6	Leaf length measured six weeks after the second cut (August 17, 2017)
C3-W2	Leaf length measured two weeks after the third cut (September 25, 2017)
C3-max	Maximal leaf length after autumn growth, estimated as the mean of the leaf length measured on October 30, November 6 and November 13, 2017.

Hi-Plex amplicon sequencing

Primers were designed for 171 amplicon regions of 80-140 bp with Primer3 (Untergasser et al., 2012) and divided into two highly multiplex (Hi-Plex) PCR-reactions according to their amplification efficiency. This Hi-Plex assay has been designed and validated in Chapter 3. DNA

was extracted using the CTAB method (Murray and Thompson, 1980) and DNA concentration was measured using the Quantus double-stranded DNA assay (Promega, Madison, WI, USA). Per sample, the final DNA concentration was adjusted to 40 ng/µL and the amplicons were PCR-amplified while adding sample specific indices. Libraries were prepared using the KAPA Hyper Prep PCR-free Kit according to manufacturer directions (Kapa Biosystems, USA). Hi-Plex amplification reactions and library preparations were done by Floodlight Genomics LLC (Knoxville, TN, USA). The libraries were sequenced with 2x150 PE on a HiSeq3000 (OMRF, Oklahoma City, OK, USA). Paired-end reads were merged with PEAR (v0.9.8) (Zhang et al., 2014) and adapter sequences were removed.

Read mapping and variant calling

The reads were mapped onto the draft genome sequence (Byrne et al., 2015) with default settings of BWA-MEM (version 0.7.8-r455) (Li and Durbin, 2009). Local realignment around indels was performed according to the best practices workflow of the Genome Analysis Toolkit (GATK) (v.3.7) (McKenna et al., 2010; Van der Auwera et al., 2013). Read depth and coverage were calculated on the resulting BAM files using BEDTools (v2.25.0) (Quinlan and Hall, 2010).

Four different variant calling pipelines were used: SAMtools (version 1.2-115-gb8ff342) (Li et al., 2009), Freebayes (v1.0.2-2-g7ceb532) (Garrison and Marth, 2012), GATK Unified Genotyper (GATK UG) and GATK HaplotypeCaller (GATK HC) (McKenna et al., 2010; Van der Auwera et al., 2013). The resulting variants and genotype calls were automatically integrated by the VariantMetaCaller (v1.0) (Gézsi et al., 2015). The Estimated Precision (EP) was calculated using a custom python script based on the formulas given in Gézsi et al. (2015) and an EP threshold of 70% was used to retain high-quality variants (Veeckman et al., 2018).

Candidate gene association mapping

The minor allele frequency (MAF) and observed heterozygosity were calculated for each population using VCFtools (v0.1.14) (Danecek et al., 2011). Variants with a MAF smaller than 1% were removed. Population structure within each population was determined using an Principal Component Analysis (PCA) with the R-package Adegenet (2.1.1) (Jombart, 2008; Jombart and Ahmed, 2011).

The 28 target genes were projected onto their corresponding chromosome (genome assembly v2.6.1, see Chapter 6) using GenomeThreader (1.6.6) (Gremme et al., 2005). For each marker, association analysis was performed with model P = G + K, with the kinship (K) matrix calculated using GAPIT (v2) (Tang et al., 2016) using only variants residing on other chromosomes. The effect of population structure (Q) was not taken into account, as it was considered to be redundant with K. This strategy was designed to reduce influence of markers located on the same chromosome in the determination of K, as they are more likely to be inherited together.

6.3 Results

Description of phenotypic diversity

During the growing season of 2017, heading date (HD) and leaf length were monitored for 1729 plants, originating from five breeding populations and one natural accession (Table 6.1).

Heading date

Figure 6.1 shows the distribution of HD for the six populations. Two populations showed early heading: 1853-7 (mean HD at 989.70 GDD) and ba12990 (999.41 GDD), two populations showed intermediate heading: ba12990x5554 (1093.14 GDD) and ba12990xplenty (1077.98 GDD), and

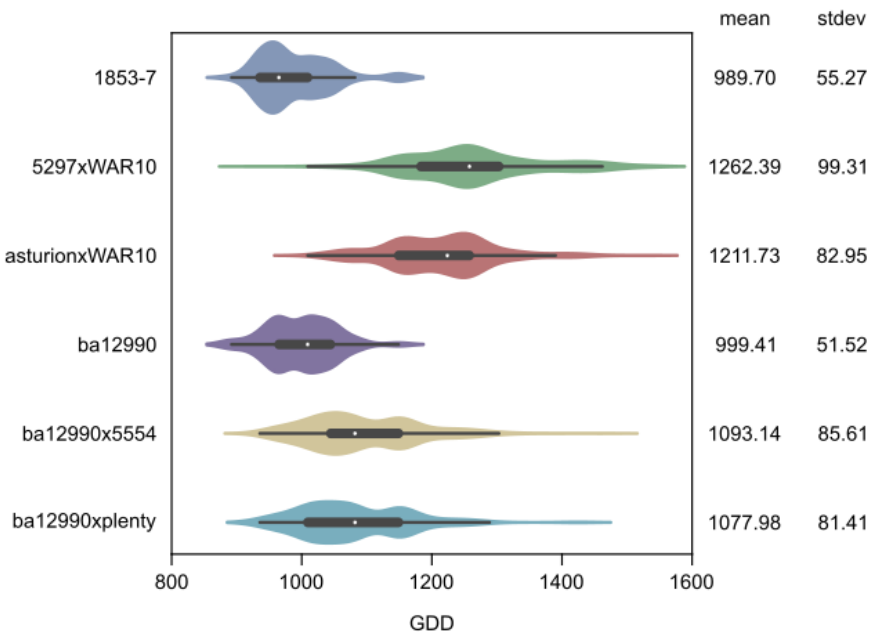


Figure 6.1 Distribution of heading date determined in six populations. Heading date was determined for each plant as the growing degree days (GDD) (base temperature 0°C, counting from January 1, 2017) on which a plant showed three ears. For each population, mean and standard deviation (stdev) are given.

two populations showed late heading: 5297xWAR10 (1262.39 GDD) and asturionxWAR10 (1211.73 GDD). Populations derived from parents with a common background showed a similar HD: 5297xWAR10 and asturionxWAR10 are both progeny of crossings cells of ILVO and Eurograss breeding material, ba12990x5554 and ba12990xplenty are both progeny of crossing cells of ILVO breeding material and a member of the Spanish accession ba12990 and showed intermediate heading. Some populations showed a higher variation in HD than others, with populations 1853-7 and the natural accession ba12990 having the smallest standard deviation corresponding to a difference in HD of two days, and 5297xWAR10 having the largest standard deviation, corresponding to a difference of four days. The observed standard variation within each population is comparable to other association studies in perennial ryegrass (Barre et al., 2009; Arojju et al., 2016).

Leaf elongation

Leaf growth was monitored making use of weekly leaf length measurements during spring, and a growth curve was generated per plant and a linear fit was used to estimate SGR (Barre et al., 2016). All populations, except 1853-7, showed a similar distribution of SGR and average SGR of

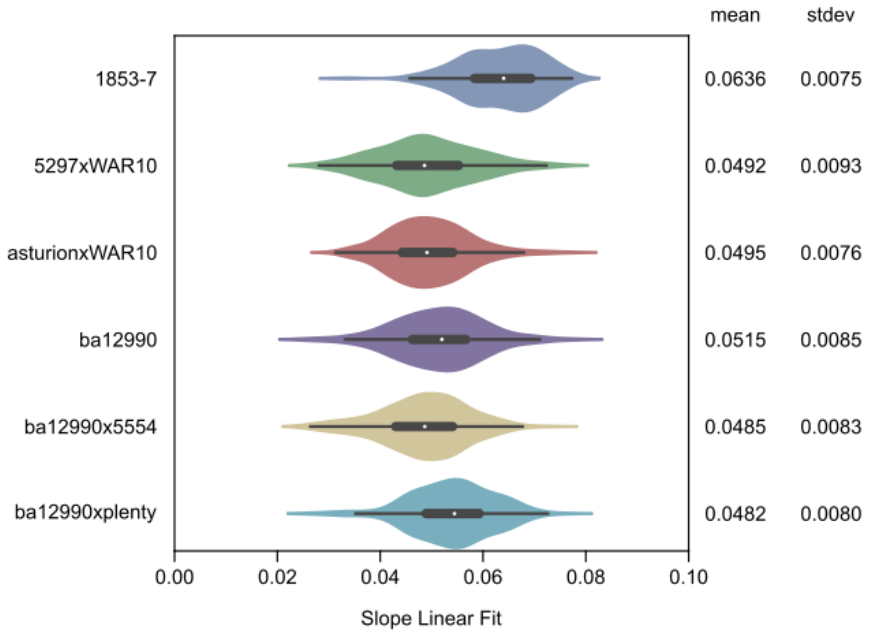


Figure 6.2 Distribution of spring growth rate, calculated as the slope of the linear fit to the growth curve. In spring, leaf length was measured weekly from February 28 until May 29, 2017 or until the plant started flowering. For the linear range of growth, spring growth rate was calculated as the slope of the linear fit to the growth curve for each plant (cm per GDD). For each population, mean and standard deviation (stdev) are given.

0.05 cm per GDD (Figure 6.2). The average SGR in 1853-7 was higher compared to the other populations (0.06 cm per GDD). During the growing season, plants were cut every six weeks starting from July 4, 2017. Leaf length was measured two and six weeks later, to estimate regrowth in a cutting regime. After the third cut, leaf length was measured until plants stopped growing to calculate the maximum leaf length after autumn growth. Corresponding distributions and figures are presented in Figure 6.3 and Table 6.3. Regrowth after the first cut, and the maximum length after autumn growth showed the widest distributions. Average leaf length measured two weeks after the third cut (C3-W2) was smaller compared to the first two cuts. Overall, individuals of population 5297xWAR10 had shorter leaves for each of these traits, and the narrowest distribution.

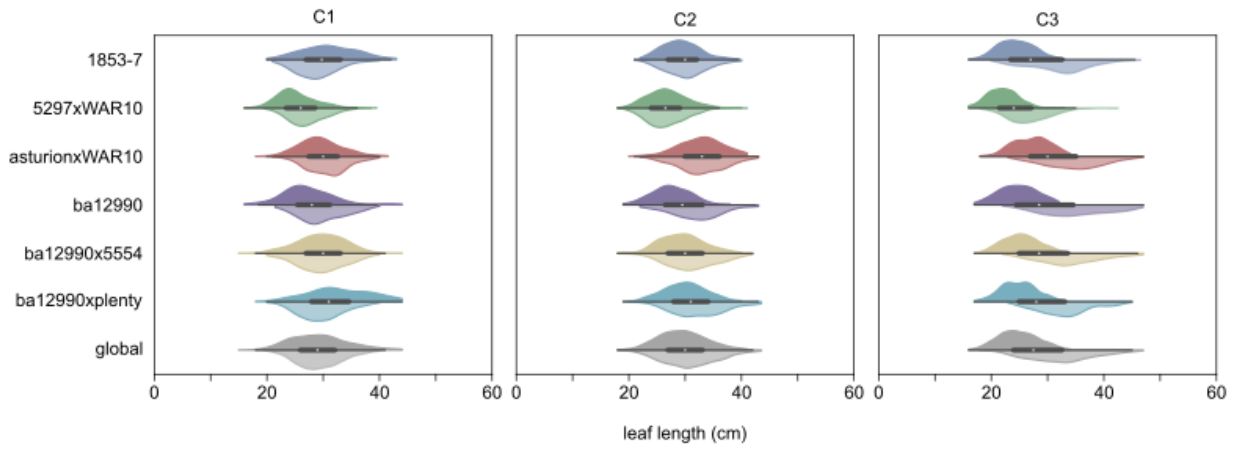


Figure 6.3 Distributions of leaf length measured two and six weeks after each cut and of the maximal leaf length after autumn growth. The top and bottom part of violin plots for each cut (C1, C2 and C3) represent leaf length measured two and six weeks after the cut, respectively. For the third cut, the maximal leaf length after autumn growth is shown, instead of leaf length six weeks after C3.

Table 6.3 Average leaf length after regrowth. For each cut (C1, C2 and C3), mean and standard deviation of leaf length was calculated (cm) for measurements two and six weeks (W2 and W6, respectively) after cut. For the third cut, the mean maximal leaf length and standard deviation after autumn growth is shown.

Population	C1-W2	C1-W6	C2-W2	C2-W6	C3-W2	C3-max
1853-7	31±5	29±4	29±3	30±4	24±3	32±5
5297xWAR10	25±3	27±3	27±4	27±3	22±3	27±4
asturionxWAR10	30±4	31±4	33±4	33±4	27±3	34±5
Ba12990	27±5	30±4	28±3	32±5	25±3	35±6
Ba12990x5554	30±4	30±4	29±4	31±4	25±3	34±5
Ba12990xplenty	33±6	30±4	30±4	33±5	25±3	33±5

Correlation between phenotypic traits

For each of the six populations, correlation between the HD and leaf length measurements were calculated (Table 6.4). A significant negative correlation was observed between HD and SGR in all populations. Early flowering individuals are taller than late-flowering individuals at the same time in spring, because their stem elongation is more advanced during the vegetative phase (Hazard et al., 2006). SGR corrected for HD was therefore used for further association analysis (see Material and Methods) to separate the effects of HD and inherent differences in elongation growth.

Table 6.4 P-values of Pearson Correlations calculated between HD and leaf elongation traits per population.

	1853-7	5297xWAR10	Asturionx WAR10	Ba12990	Ba12990x 5554	Ba12990x plenty
SGR	-0.35***	-0.67***	-0.60***	-0.49***	-0.43***	-0.42***
C1W2	-0.09 ^{NS}	-0.01 ^{NS}	0.01 ^{NS}	-0.15 ^{NS}	-0.18**	-0.07 ^{NS}
C1W6	-0.05 ^{NS}	-0.13*	0.10 ^{NS}	-0.05 ^{NS}	-0.01 ^{NS}	0.16**
C2W2	-0.03 ^{NS}	0.05 ^{NS}	0.02 ^{NS}	-0.10 ^{NS}	-0.01 ^{NS}	0.05 ^{NS}
C2W6	0.02 ^{NS}	0.01 ^{NS}	0.01 ^{NS}	0.06 ^{NS}	-0.06 ^{NS}	0.09 ^{NS}
C3W2	-0.10 ^{NS}	-0.02 ^{NS}	0.09 ^{NS}	0.03 ^{NS}	-0.10 ^{NS}	0.06*
C3max	-0.05 ^{NS}	-0.05 ^{NS}	0.07 ^{NS}	-0.02 ^{NS}	-0.09 ^{NS}	0.12*

^{NS}, $P > 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Description of genotypic diversity

A Hi-Plex amplicon sequencing array of 171 amplicons (Veeckman et al., 2018) was used to genotype the 1729 individual plants. The amplicons covered 28 candidate genes, involved in light signaling, transition to flowering and hormone biosynthesis and signaling (Supplemental Table 3). This resulted in 253,000 reads per genotype on average, corresponding to an average read depth of 1,343 reads per amplicon.

Four VC pipelines integrated by the VariantMetaCaller followed by precision-based filtering resulted in a high-quality variant set containing 952 SNPs and 261 indels. Variants with MAF<1% per population were removed to reduce the number of false positive variants. The number of polymorphic sites differs between populations (Table 6.1): asturionxWAR10 contained the fewest variants. The natural accession, ba12990, contained the most variants of which 245 were

uniquely present in this population. The call rate was high, with an average call rate of 98% per variant.

Table 6.5 Overview of the number of variants and mean observed heterozygosity per population.

Population	Number of variants	Number of variants MAF > 1%	Mean observed heterozygosity
1853-7	1165	370	0.42
5297xWAR10	1171	416	0.28
asturionxWAR10	1174	289	0.38
ba12990	1176	694	0.20
ba12990x5554	1173	548	0.37
ba12990xplenty	1186	508	0.29

The mean level of observed heterozygosity is a valuable parameter to estimate the degree of genetic variation in each population (Table 6.5). This resembled the number of parental alleles present in each population (Table 6.1). Population 1853-7 is a half-sib family from a poly-cross with seven components, i.e. all individuals are half-sibs derived from seed harvested on one component, the mother plant 1853-7, pollinated by 6 possible fathers. This explains the highest level of heterozygosity. AsturionxWAR10 was derived from a single cross of two parents, and

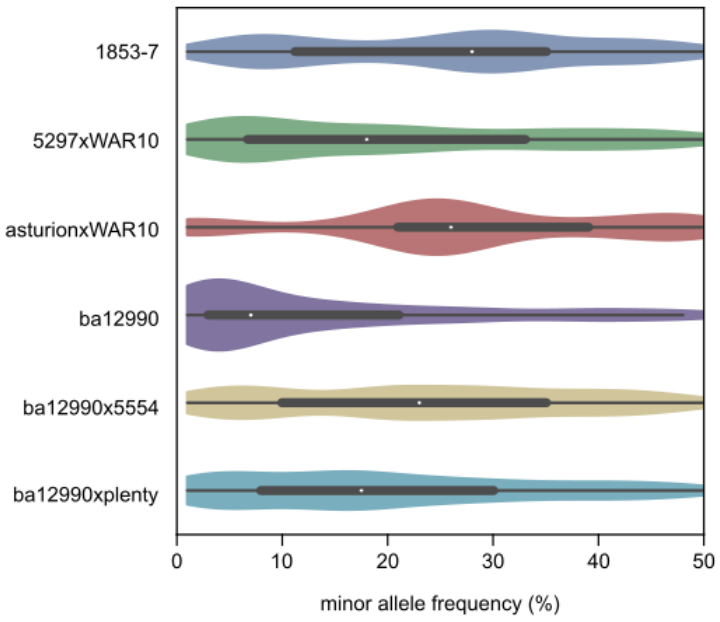


Figure 6.4 Distribution of the minor allele frequencies per population. Variants with minor allele frequency < 1% per population were removed.

individuals showed the second highest level of heterozygosity on average. Populations 5297xWAR10 and ba12990xplenty (four parents) show lower levels of heterozygosity than ba12990x5554 (three parents). Individuals of the wild accession ba12990 showed the lowest level of heterozygosity, indicating that many alleles occur at low frequency (Figure 6.4).

Figure 6.5 shows population structure (Q) within each population as assessed with a PCA. The populations ba12990, ba12990x5554 and ba12990xplenty showed more structure compared to the other three populations. This could not be explained by the number of parental genotypes, or the setup of the poly-cross for the generation of the F2 for populations ba12990x5554 and ba12990xplenty. Association analysis was performed taking only the effect of kinship (K) into account, as the effect of population structure (Q) was considered redundant.

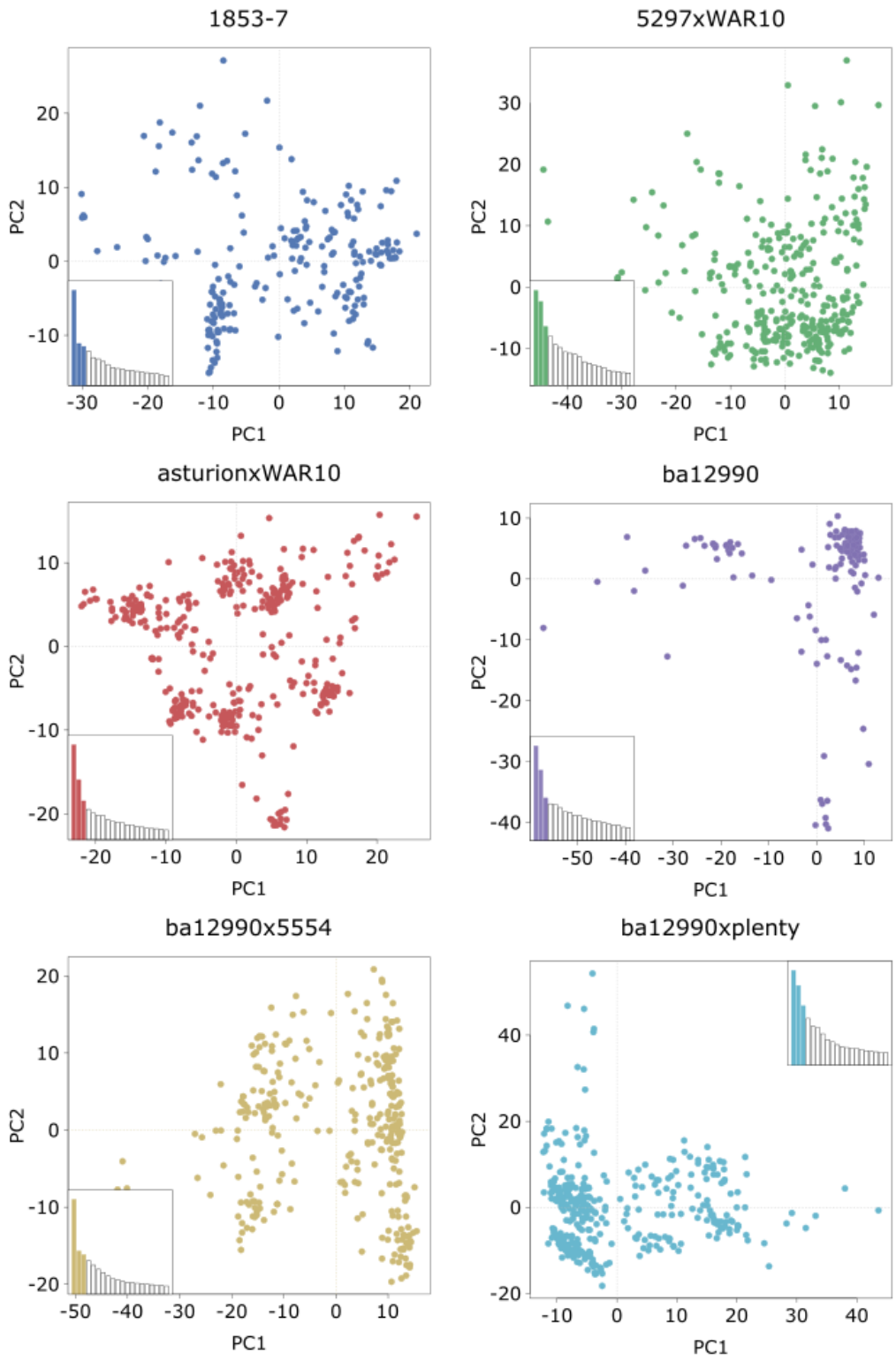


Figure 6.5 Principal Component Analysis to reveal population structure within each population. PCA was performed with the R-package Adegenet (MAF > 1%). For each population, the individuals are positioned based on the first two principal components (PC1, PC2).

LpFT-03 associates with heading date

Association mapping was performed for each population and trait separately. For HD, a single significant association with *LpFT-03* was found in population ba12990xplenty ($p < 1e-05$) (Table 6.6). This polymorphism explained 7% of the phenotypic variation in HD present in this population. *LpFT-03* resides on chromosome 7, in a QTL that has previously been described to regulate flowering in perennial ryegrass (Armstead et al., 2004). This gene is part of the PEBP (phosphatidylethanolamine-binding proteins) gene family, also known as the FLOWERING LOCUS T (FT) gene family. Only three members of this family, *LpFT3, LpTFL1* and *LpSFT*, have previously been characterized in *L. perenne* and different alleles of *LpFT-03* have been shown to associate with variation in HD (Jensen et al., 2001; King et al., 2006; Fiil et al., 2011; Skot et al., 2011).

Table 6.6 Significant association with heading date. (FDR: false discovery rate, PVE: phenotypic variation explained)

Scaffold	Position	Gene	Chromosome	Population	p-value (FDR adjusted)	PVE
5073	38,536	FT-03	7	ba12990xplenty	6.02E-05	7.39%

The association corresponds to an adenosine-thymine substitution in the first exon of the *LpFT-03* gene and is synonymous: codons GGT and GGA both code for a glycine residue. The substitution itself is unlikely to be causal, and may therefore be in LD with a polymorphism that is directly causal for variation in HD.

For all 358 individuals of population ba12990xplenty a genotype call was present: 114 individuals were homozygous for the reference allele (adenosine), 41 individuals were homozygous for the alternative (thymine) and 203 individuals were heterozygous at the associating position. In the other five populations, the association was not significantly associated. Three possible reasons may explain this: not all three genotypic classes were adequately represented (Supplemental Table 4), the associating SNP is not in LD with the causative polymorphism in the existing alleles in these populations, or the range of phenotypic variation was too small.

Allelic effect

The allelic effect was estimated to be -79.99 GDD indicating that the alternative allele, i.e. a thymine, corresponds to later heading. Skot et al. (2011) have validated the existence of seven haplotypes for *LpFT-03* and their association with flowering time. The alternative allele (thymine) of the associating SNP was characteristic for haplotypes A and G described in Skot et al. (2011), which are identical except for one nucleotide. These haplotypes contain a deletion in the promoter region of *LpFT-03* in the proximity of a GGACAT motif, which could affect transcription factor binding efficiency, and, in turn, the expression of the *LpFT-03* transcript, thereby leading to later flowering. Additional weaker associations with polymorphisms in *LpFT-03* in population ba12990xplenty confirmed the existence of haplotype A/G in this population and the association with later heading.

In population 5297xWAR10, the SNP at position 38,536 was not significantly associating (p-value 3.24E-02), but this population showed an opposite allelic effect of +54.84 GDD, indicating that the alternative allele corresponded to earlier heading compared to the reference allele.

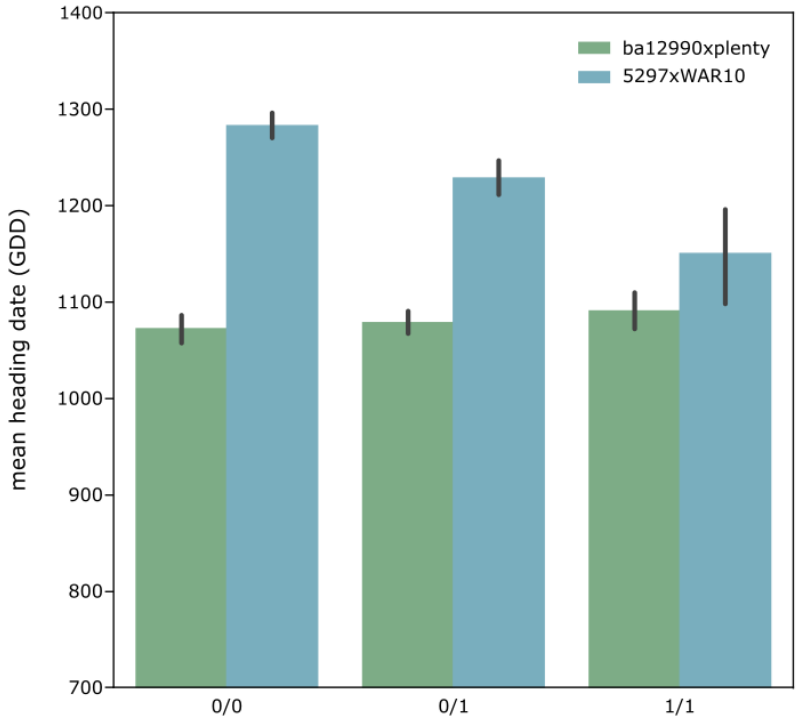


Figure 6.6 Distribution of heading date per genotypic class for the association with *LpFT-03* (scaffold 5073, position 38,536). For the populations ba12990xplenty and 5297xWAR10, the mean heading date (in GDD) was calculated per genotypic class. Error bars indicate 95% confidence intervals.

Distribution of HD per genotypic class confirmed this contrasting trend for this SNP between populations ba12990xplenty and 5297xWAR10 (Figure 6.6). Individuals that are homozygous for the alternative allele in both ba12559xplenty and 5297xWAR10 showed a similar HD, while the reference allele corresponded to earlier and later heading, respectively. This suggests that multiple alleles exist across the different test populations, in which the causative polymorphisms are in alternative phase with neighboring SNPs covered in our Hi-Plex assay.

A comparison of the haplotypes identified by Skot et al. (2011) and phased variants across amplicons of *LpFT-03* in populations ba12990xplenty and 5297xWAR10 confirmed that the alternative allele at position 38,536 corresponded to the presence of haplotype A/G in both populations, thus explaining a similar HD in individuals homozygous for the alternative allele. On the other hand, a reference allele at position 38,536 corresponded to at least two new haplotypes in these populations that were different from the seven haplotypes previously identified by Skot et al. (2011).

LpMADS-01 associates with maximal leaf length after autumn growth

Leaf elongation was monitored throughout one growing season, and different traits were used in the association analysis: SGR, leaf length two weeks and six weeks after cutting (C1, C2, C3), and the maximal leaf length at the end of autumn growth. Only for this latter trait, a single significant association with *LpMADS-01* was found in population 5297xWAR10 ($p < 1e\text{-}05$) (Table 6.7). This polymorphism explained 8% of the phenotypic variation in maximal leaf length at the end of autumn growth present in this population. *LpMADS-01* co-localizes with the *VERNALIZATION1* (VRN1) locus, which encodes for a MADS-box transcription factor and controls vernalization-induced flowering in cereals. It is related to genes that promote flowering in other plant species: phylogenetic analysis of grass MADS-box genes together with *A. thaliana* homologs has shown that *LpMADS-01* clusters in the APETALA1 (AP1) subgroup (Danyluk et al., 2003; Trevaskis et al., 2003; Yan et al., 2003; Petersen et al., 2004).

For 329 out of 338 individuals of population 5297xWAR10 a genotype call was present: 128 individuals were homozygous for the reference allele (cytosine), none of the individuals were homozygous for the alternative allele (guanine) and 201 individuals were heterozygous at the

associating position. Although the distribution among genotypic classes was similar for other populations (Supplemental Table 4), no significant association was found.

The allelic effect of the association was found to be -2.97 cm in population 5297xWAR10. In other words, the alternative allele (guanine) at this position corresponds to shorter leaves after autumn growth. The association corresponds to a cytosine-guanine substitution in the first intron of the *LpMADS-01* gene.

Table 6.7 Significant association with leaf length. (FDR: false discovery rate, PVE: phenotypic variation explained)

Scaffold	Position	Gene	Chromosome	Population	p-value (FDR adjusted)	PVE
312	70351	MADS-01	4	5297xWAR10	3,60E-05	8.13%

6.4 Discussion and Conclusion

For 28 selected candidate genes, 170 amplicons were designed for a Hi-Plex genotyping assay to identify the genetic variation across 1729 individuals. Hi-Plex amplicon sequencing proved to be an efficient genotyping method at the scale of breeding populations of several thousands of plants. The resulting variant set was of high quality, containing only 2% missing data. This is a good example of how the catalog of sequence variation present in 503 candidate genes that was generated for 743 individuals derived from natural accessions and current cultivars and breeding materials, is a good representation for the variation present in the *L. perenne* genepool. This variant collection is a reliable resource for primer design for the 503 candidate genes, a critical step in the development of a high-throughput and targeted genotyping assay.

A single SNP in *LpFT-03* was significantly associated with HD and a single SNP in *LpMADS-01* was significantly associated with leaf length after autumn growth, explaining 7% and 8% of the phenotypic variation respectively. Both associations were found in a single breeding population. The range of variation in HD and especially the small range of variation in leaf length was a first limiting factor to find associations. This was reflected in the small allelic effect size of the significant associations. The number of target genes and/or small genotypic diversity in each population is a second limiting factor to find associations. This illustrates the importance of selecting parents based on contrasting phenotypes and/or the presence of divergent alleles for

candidate genes. Thirdly, the association study was performed for each population separately, possible leading to false negative associations, i.e. true associations that were not significantly associating within a population because of reduced statistical power due to a small range in phenotypic variation or suboptimal representation of genotypic classes.

Taken this last argument, analyzing all populations jointly would reduce the number of false negative associations. We chose to analyze populations separately because of two reasons. First, performing the performing the association analysis using all populations together requires accurate estimation of the population structure, as population stratification is one of the major sources for false positive associations. Given the fact that the genotypic assay covered 28 genes representing only five of the seven chromosomes of the *L. perenne* genome, inference of the population structure is not straightforward for this study. Second, analysis of the *LpFT-01* alleles suggested that multiple haplotypes exist across the different test populations, meaning that neighboring SNPs are in alternative phase. Using single binary markers (reference versus alternative allele) may confound the association analysis, as they represent different haplotypes across populations.

LpFT-03 is a central regulator of the genetic network that controls flowering, more specifically, the transition from vegetative to generative growth. Different alleles of *LpFT-03* have been found to associate with variation in HD (Jensen et al., 2001; King et al., 2006; Fiil et al., 2011; Skot et al., 2011). Phasing variants across amplicons of *LpFT-03* confirmed that the associating substitution in the first exon was specific for haplotype A/G previously identified by Skot et al. (2011). However, the reference allele at this position was linked to at least two new haplotypes in populations ba12990xplenty and 5297xWAR10, with a contrasting phenotype. This illustrates how transfer of markers across different populations is not straightforward, as single polymorphisms are member of different haplotypes and therefore not always in LD with a sequence variant causal for phenotypic variation, thus confounding the analyses. Further analyses are required to untangle the existing haplotypes for *LpFT-03* and their relationship to flowering time: reconstruction of full haplotypes will provide a way to identify sequence variation causal for phenotypic variation, as well as higher power and increased precision for association analyses compared to single genetic markers such as SNPs.

LpMADS-01 is associating with maximal leaf length after autumn growth, one of the five representative traits for leaf elongation. This gene is also involved in the genetic network controlling flowering time. It is related to the *A. thaliana APETALA1* gene and an orthologue of the wheat *VRN1* gene. This substitution does not lead to a difference in protein sequence, but is located in the first intron that plays an important role in *VRN1* regulation (Yan et al., 2003; Fu et al., 2005). The first intron contains a core region required for repression of *VRN1* before winter, together with regions at the 5'end of the intron. Alleles containing larger deletions are more active and are associated with earlier flowering without vernalization (Hemming et al., 2009). Histone modifications at the first intron and promoter are also important for *VRN1* activity: histone-3-lysine-27-trimethylation (H3K27Me3) induces an inactive chromatin state and might contribute to repression of *VRN1* before winter (Oliver et al., 2009). After vernalization, the level of H3K27Me3 decreases and makes space for histone modifications associated with active chromatin state. The associating substitution in the first intron could be involved in the regulation of the expression level of *LpMADS-01*, either epigenetically or by influencing binding of repressors. It is, however, also likely that (i) this association is linked to a causal association in a neighboring gene on chromosome 4 or (ii) this is a false positive association, as only leaf length after autumn growth was found to be associated but none of the other leaf length traits, such as spring growth rate. Although *LpMADS-01* is involved in the flowering regulatory network and the association was not found with other elongation traits such as spring growth rate, this is an interesting candidate marker, as it is located in the first intron that is involved in regulation of *LpMADS-01* expression.

With this candidate gene association study, we aimed to detect alleles in breeding populations responsible for phenotypic variation in flowering time and leaf elongation. Although there were no significant associations identified in the natural accession, including wild material is valuable to identify novel alleles that are not yet represented in the breeding material. Using a conservative association mapping approach, a single association was identified for each trait of interest. The corresponding polymorphisms can be developed into molecular markers that can be used to direct future breeding programs. However, it should be noted that the associating variant is not necessarily the causal mutation for the corresponding phenotype, and is only

significantly associating because it is genetically linked to the causal mutation. Additionally, as shown for the associating SNP in *LpFT-03*, a single marker is not capable of distinguishing between different haplotypes that exist for this gene. As such, a haplotype-based association mapping approach will be more valuable to capture the full genetic diversity present within and across populations, and may provide more power to identify the causal mutation.

6.5 Author Contribution

Elisabeth Veeckman, Tom Ruttink, An Ghesquière and Hilde Muylle were involved in selection of the populations. Tom Ruttink, Sabine van Glabeke, Elisabeth Veeckman and Kurt Lamour designed the Hi-Plex genotyping assay. Elisabeth Veeckman, Hilde Muylle, Isabel Roldán-Ruiz and Tom Ruttink designed the research methodology. Elisabeth Veeckman performed the data analysis. Elisabeth Veeckman, Hilde Muylle, Isabel Roldán-Ruiz and Tom Ruttink contributed to writing the chapter.

7 Valorization, Outreach and Conclusion[5]

The focus of this PhD was the development of genomic resources for perennial ryegrass. In the previous chapters, some of the major challenges that we encountered in doing so were described. Here, I give an overview of the resources that are now available, focusing on what was needed to create or improve these resources while remaining perceptive for possible pitfalls and imperfections. Most resources described in this chapter are currently unpublished data and approaches, developed in collaboration with partners of the international ryegrass research community. First, I describe how the gene space completeness measures developed in Chapter 3 were used to develop an annotation pipeline resulting in a more complete and accurate gene annotation set for the draft genome sequence published by Byrne et al, 2015 (further referred to as the v1.4 genome sequence). Using the PLAZA workflow, functional annotations were generated for the resulting gene set. Furthermore, we illustrate how the completeness of this gene set and the corresponding functional annotations are relevant for the ryegrass research community using studies that are still ongoing and are already using the newly generated resources. Next, I describe our joint collaborative efforts to generate a novel, chromosome-scale reference sequence assembly (further referred to as the v2.6.1 genome sequence), to overcome the limitations of the current available reference genome sequence. Finally, we discuss the challenges on identifying genomic diversity in perennial ryegrass and the possible applications of the catalog of sequence diversity in the dissection of complex traits and marker-assisted breeding.

[5] This chapter contains currently unpublished data obtained in close collaboration with partners from Aarhus University, Teagasc, University Tübingen, and INRA-Lusignan. For author contributions, see page 129.

7.1 Structural and functional gene annotation of the draft genome sequence

The completeness of a gene annotation set and accuracy of the gene models are two important prerequisites for comparative genomics and evolutionary studies. Manual curation of 503 candidate genes that were selected for resequencing (Chapter 4) showed that 20% did not have a gene model predicted, clearly indicating that the gene annotation set of the v1.4 genome sequence was incomplete. This can be explained by the use of a conservative gene annotation approach, resulting in 28,182 reliable gene models (Byrne et al., 2015). Only evidence-based gene models, i.e. supported by extensive transcriptome assemblies and *B. distachyon* protein alignments, were retained and genes are missing from the final gene annotation set because no *ab initio* gene prediction was included.

Need for gene space completeness measures

This raised the question on how to measure the completeness of the gene space in both the genome assembly and gene annotation set. In order to answer this question, it was important to first define the 'expected' gene space, in order to express completeness as a fraction of the expected.

In Chapter 3, we introduced the concept of defining the expected gene space along the evolutionary scale, as this accommodates using different gene space completeness measures. All measures have their own strength, weaknesses, underlying assumptions and potential biases, and the results should be interpreted accordingly (Veeckman et al., 2016). To define the expected gene space, one can either rely on evolutionary conservation and use the gene space of related species as a reference. Examples are CEGMA, BUSCO and the PLAZA core gene families (coreGFs). Alternatively, one can define a species-specific measure of the gene space using transcripts or EST sequences.

CEGMA, BUSCO and transcript mapping can be used to estimate completeness of the gene space in the genome assembly, while BUSCO and the coreGFs can be used to estimate completeness of the gene annotation set. For the v1.4 genome sequence, a CEGMA score of 96% was reported, and the transcript mapping score of 96% confirmed that the gene space was fully represented in

the v1.4 genome sequence. However, the coreGF score of 76%, corresponding to 1,709 missing coreGFs, confirmed our suspicion that the v1.4 annotation set is not complete.

All studies relying on the v1.4 annotation without further gene prediction will therefore suffer from the missing unpredicted genes, for instance, one may falsely conclude that 1,709 core gene families are lost in perennial ryegrass during genome evolution. As measures for gene space completeness are not yet a common standard in genome publications, it is often overlooked that an assembly or the gene annotation set could be incomplete. Therefore, we strongly encourage using clear and objective measures indicating completeness of both assembly and annotation set, so that end-users are well aware of possible limitations and are less prone to false interpretations.

Improving the gene annotation using EVidenceModeler

As the v1.4 genome sequence is currently the best assembled genome sequence available, and many studies are using this version as a reference, it is important to improve the current gene annotation to obtain a more complete set of gene models while maintaining gene model accuracy.

The EVidenceModeler (EVM) is a tool for automated gene structure annotation, and combines evidence from secondary sources, such as *ab initio* gene predictions and various forms of sequence homologies (Haas et al., 2008). We used the EVM to improve completeness of the gene annotation set of the v1.4 genome sequence, without losing gene model accuracy. For this, the current gene annotation set was complemented with a less conservative set of gene predictions, orthology-guided transcript assemblies (Ruttink et al., 2015) and aligned proteomes of closely related species (*B. distachyon,* rice*,* maize and sorghum). The EVM then computes a gene model consensus based on the types of evidence available and their corresponding weight values.

Finding a combination of weights that provides the best consensus prediction accuracy is an important goal, but very different weight settings can lead to similar levels of performance. Therefore, tuning the EVM weights intuitively is a straightforward way to obtain a high-quality gene annotation set and combinations of assigned weights in the following form provides adequate consensus prediction accuracy (Haas et al., 2008):

(ab initio predictions) ≤ (protein & transcript alignments) < (evidence based gene prediction)

Given the set of 503 candidate genes that were manually curated in the framework of identification of sequence variation (Chapter 4), it was possible to estimate the accuracy of the individual input tracks as well as to evaluate the consensus gene models returned by the EVM using these manually curated gene models as a gold standard. Predicted exons were compared to the gold standard exons at single nucleotide level. The F1 score is the harmonic mean of precision and recall, and was used as a quality measure for the predicted exons (Table 7.1).

The EVM gene annotation set for the v1.4 genome sequence contained 39,967 genes. The completeness was estimated at 92.6% ([S:89.0%, D:3.6%], F:2.5%, M:4.9%, n:1440) using BUSCO (Simao et al., 2015) and 89% using the PLAZA 2.5 monocots core gene families (Van Bel et al., 2012). Only 3% of the candidate genes were missing, compared to 20% in the original gene annotation set (Table 7.1). As the EVM gene set is more complete, it can be used for gene content analysis, as evidence for gene prediction on future genome assemblies, and it forms a better basis for estimating gene expression levels in transcriptome profiling experiments (see below). However, some genes may still be absent in the EVM annotation set. BLAST searches of the missing core gene families could indicate their absence/presence in the v1.4 genome sequence, thereby revealing the strengths/weaknesses of our approach to improve the annotation of the v1.4 genome sequence using the EVM.

Accurate gene models are important for the prediction of the effects of sequence variants on gene function (Chapter 5 and 6). The novel gene models proved to be 10% more accurate (Table 7.1). Setting weights for the EVM was mainly guided by testing the accuracy of the gene models using a set of 503 manually curated genes, representing 180 gene families. This could have generated a bias to correctly predicting gene models for genes resembling genes of the 180 selected gene families. However, setting equal weights for the input tracks resulted in the best accuracy, thanks to the large amount of evidence tracks and the high quality of each input track individually.

Table 7.1 Comparison of gene annotation sets for the v1.4 genome sequence.

	V1.4 gene set (Byrne et al., 2015)	**EVM gene set**
Number of genes	28,182	39,967
Accuracy (F1 measure)	87.40%	97.01%
BUSCO (n = 1,440)	Complete: **81.6%** [Single: 61.6%,Duplicated: 20.0%] Fragmented: 2.5% Missing: 15.9%	Complete: **92.6%** [Single: 89.0%, Duplicated: 3.6%] Fragmented: 2.5% Missing: 4.9%
PLAZA CoreGF (n = 7,076)	76.87% 1,709 missing	**89.42%** 538 missing

From trait-associated SNP to the closest gene

The EU-project GrassLandscape aims to screen the natural diversity of perennial ryegrass to discover genetic variability involved in environmental adaptation, more specifically in climatic adaptation. Within the project, 550 natural populations were sampled across the whole area of primary expansion of perennial ryegrass (Europe, Northern Africa and Near East) and genotyped using GBS (Blanco-Pastor et al., 2018). These populations were also phenotyped in the field at three different locations to record agronomic and eco-physiological traits. Association models between genomic variants and environmental variation are being used to map the spatial distribution of genomic markers linked to adaptive diversity in present climatic conditions, an approach called 'landscape genomics'. The aim of the study is to identify genomic markers in the populations linked to the environmental and climatic characteristics of their location of origin, i.e. markers carrying signatures of selection.

In order to add more biological meaning to these signatures of selection, the gene context of associating variants can be assessed. In an early stage analysis for the GrassLandscape project, 51,695 SNPs were identified as possible candidates for signatures of selection. Using the published gene annotation set (28,182 genes) 15% of the SNPs resided on a scaffold where no gene was annotated (Table 7.2). Compared to the EVM gene set, this number decreased by half, with only 4,220 SNPs residing on a scaffold without an annotated gene. Moreover, the number

of SNPs located within a gene increased, from 45% to 52% of the SNPs, and the average distance to the closest gene decreased.

The EVM gene set is more complete compared to the published gene set and should therefore be used in future studies as the standard annotation set for the genome v1.4 genome sequence. This example has already illustrated that using the EVM gene set increased the resolution for interpreting associating variants, so that the biological signal becomes clearer, and this helps to explain why a certain region in the genome associates with a trait of interest.

Table 7.2 Comparison of two gene annotations sets of the v1.4 genome sequence for the identification of the closest gene for trait-associated SNPs.

	V1.4 gene set (n = 28,182)	**EVM gene set (n = 39,967)**
# SNPs on scaffold without genes	8,014	4,220
# SNPs within gene	23,065 6,194 unique genes	26,868 7,499 unique genes
Mean distance to closest gene	9,869 bp	7,476 bp

Generation of gene function annotations using the PLAZA comparative genomics platform

The improved gene annotation set of the v1.4 genome sequence generated by the EVM consists of 39,967 genes. Only for a few genes of *L. perenne* a functional description is available in public resources such as Genbank. We generated functional annotation for the v1.4 EVM consensus gene models by adding *L. perenne* to a private version of PLAZA, based on version 4.0 Monocots (https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v4_monocots/) that contains 29 comparator species (Figure 1.1).

PLAZA is an access point for plant comparative genomics centralizing genomic data produced by different genome sequencing initiatives. It integrates plant genome sequence data and comparative genomics methods and provides an online platform to perform evolutionary analyses and data mining within the green plant lineage (*Viridiplantae*) (Proost et al., 2009; Van Bel et al., 2012; Proost et al., 2015; Van Bel et al., 2018). Using this platform, it is possible to transfer knowledge about molecular functions from model plant species, such as *A. thaliana* and

rice, to crop species. Functional annotation is expressed using the Gene Ontology (GO) syntax, and added using InterProScan and projection of functions to orthologous genes (Burge et al., 2012).

First, gene families were created based on an all-against-all BLAST of all protein coding genes, and a multiple sequence alignment and phylogenetic tree was generated for each gene family. As a result, 160,545 homology groups and 236,721 orthogroups were created.

For each protein sequence, the InterPro database was searched for matching protein signatures. If the InterPro entry was associated with a GO term describing the conserved molecular function, biological process, or cellular location, the corresponding gene was also assigned this GO term. This resulted in the assignment of 40,825 GO terms to 18,412 *L. perenne* genes. The PLAZA workflow offers the possibility of GO projection: functional annotation is exchanged between orthologs, using a set of rules based on the phylogenetic trees and sets of orthologs. In total, 45,304 GO terms were assigned to 20,224 genes (51%), derived from either InterProScan (90% of the GO terms) or GO projection (remaining 10%) (Figure 7.1). Figure 7.2 shows the number of GO terms per gene per GO category: 11,201 genes are associated with a Biological Process, of which 1,095 through projection; 18,104 genes are associated with a Molecular Function, of which 2,340 through projection; and 3,977 genes are associated with a Biological Process, of which 308 through projection.

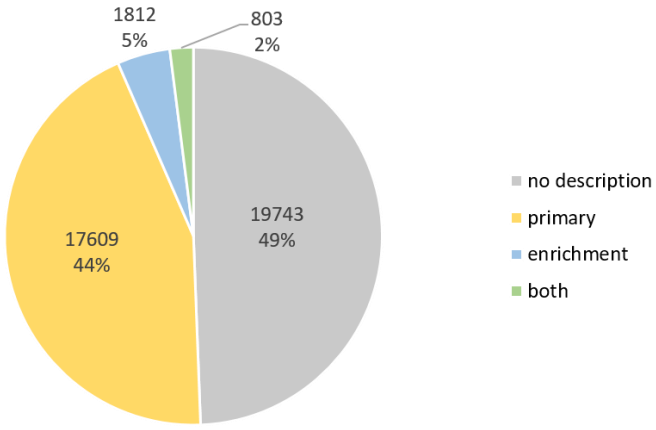


Figure 7.1 Number of EVM gene models with associated functional annotation derived from InterProScan and PLAZA GO projection.

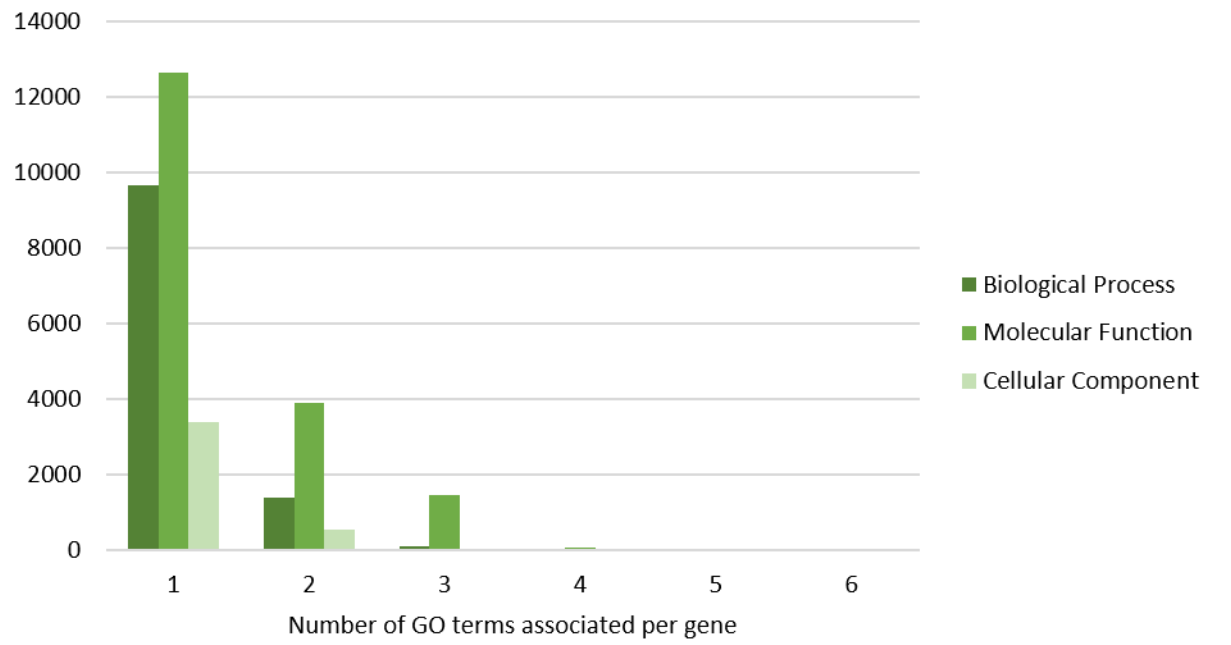


Figure 7.2 Number of GO terms associated per EVM gene per GO category. The number of GO terms per gene was counted per GO category, after removing parental GO terms.

GO terms and InterPro domains are very useful as they are part of a controlled vocabulary, allow for functional interpretation through enrichment analysis and are easily transferable across species. However, a free-text gene description is often easier to interpret and biologically more relevant. Using AnnoMine, a homology-based text-mining approach, the genes were functionally annotated complementary to the well-structured InterPro and GO annotations (Van Landeghem, 2014). In total, 19,301 *L. perenne* genes were assigned a short gene description, of which 3,655 genes did not have a GO or InterPro annotation.

Taking both GO, InterPro and AnnoMine annotations into account, a functional description is now available for 23,879 *L. perenne* genes (59.8%). Many genes did not get any functional description, but also for well-studied model species *O. sativa* or *B. distachyon*, the fraction of genes without GO term is also relatively large, 29% and 38%, respectively (Figure 7.3). This mainly reflects the global state of the art of plant molecular biology: the function of many genes is not yet elucidated in any of the well-studied model species (Rhee and Mutwil, 2014), and therefore this information cannot be projected to homologous *L. perenne* genes.

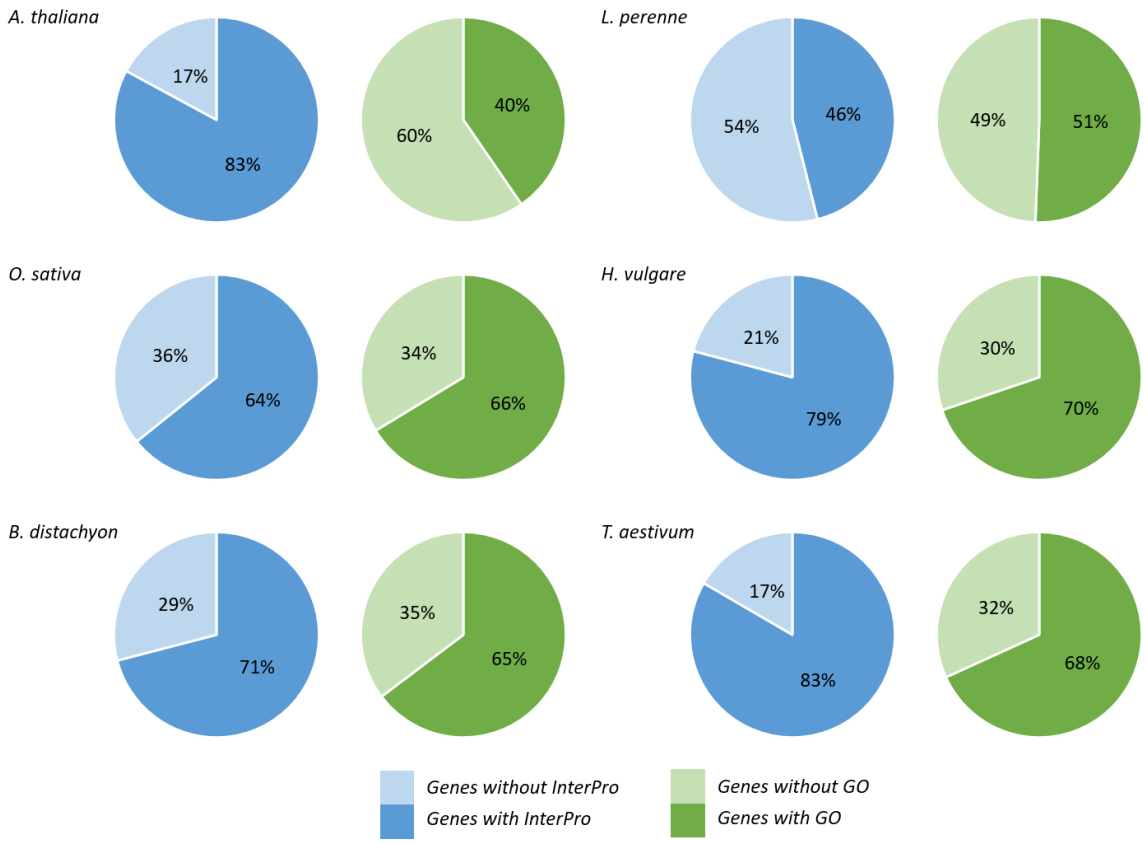


Figure 7.3 PLAZA data overview for functional annotations for six selected species.

Usage of the functional annotation in the context of cold tolerance

The improved gene set generated using the EVidenceModeler is now functionally annotated. This is a valuable asset for the interpretation of trait-associated SNPs, as described above, but also for other functional studies in perennial ryegrass.

Dr. Stephen Byrne and colleagues at Teagasc, Ireland, are investigating the transcriptome response to cold stress in six perennial ryegrass cultivars. The selected cultivars are highly heterogeneous, and RNA sampling was done in bulk for each cultivar to obtain the average transcriptome response of the cultivar, as opposed to the response of individual genotypes within the cultivar. The study aims to compare the transcriptomes of each cultivar under cold stress to their matched control samples and identify differentially expressed genes or altered biochemical pathways. On a second level, the differentially expressed genes or pathways were compared across the six cultivars.

The incompleteness of the v1.4 gene set as a reference for differential expression analysis, and the limited functional annotation of these gene models hampered interpretation of the results. Using the EVM gene set instead, ca. 70% of the RNA-Seq reads could be aligned to known gene models. Genes that are commonly regulated in response to both short periods of cold and longer periods of cold acclimation could be identified. In the future, this knowledge will be exploited in predictive modelling for forage yield in early spring, using approaches such as genomic feature best linear unbiased predictor (GFBLUP) to improve the accuracy of genomic prediction.

7.2 Towards a chromosome-scale reference genome for *Lolium perenne*

In a collaborative effort with Aarhus University and the University of Tübingen, this PhD contributed to a novel assembly of the ryegrass genome sequence, including gene annotation (see below). Here, we shortly describe the efforts to *de novo* assemble an entirely new version of the *L. perenne* reference genome sequence. Repetitive regions remained the main disruptive factor to obtain a chromosome-scale assembly representative for the total haploid genome size, as well as the fact that previously, there was no technology available in the 10-20kb linkage range to anchor and orient scaffolds in the final assembly stage. Anchoring scaffolds using synteny impedes further evolutionary and comparative genomics studies, and genetic linkage maps provide low resolution, as crossover events are rare in centromeric regions. An alternative strategy, implementing state of the art techniques implemented in the most optimal order, was needed to obtain a chromosome-scale genome assembly for *L. perenne.*

Integration of third-generation sequencing, optical mapping and Hi-C results in a chromosome-scale assembly

Figure 7.4 shows an overview of the strategy that was used to obtain a chromosome-scale genome assembly for *L. perenne*. This was done using a combination of Illumina short read sequencing with PacBio long read sequencing, optical mapping, and Hi-C. New PacBio SMRT sequencing was used compared to the v1.4 genome sequence. Long read assembly and polishing resulted in an assembly that is representative for the total genome size and therefore also contains the repetitive fraction of the genome (in contrast to the v1.4 genome sequence), but is still highly fragmented (> 40k contigs). Because optical maps can bridge repetitive regions, hybrid

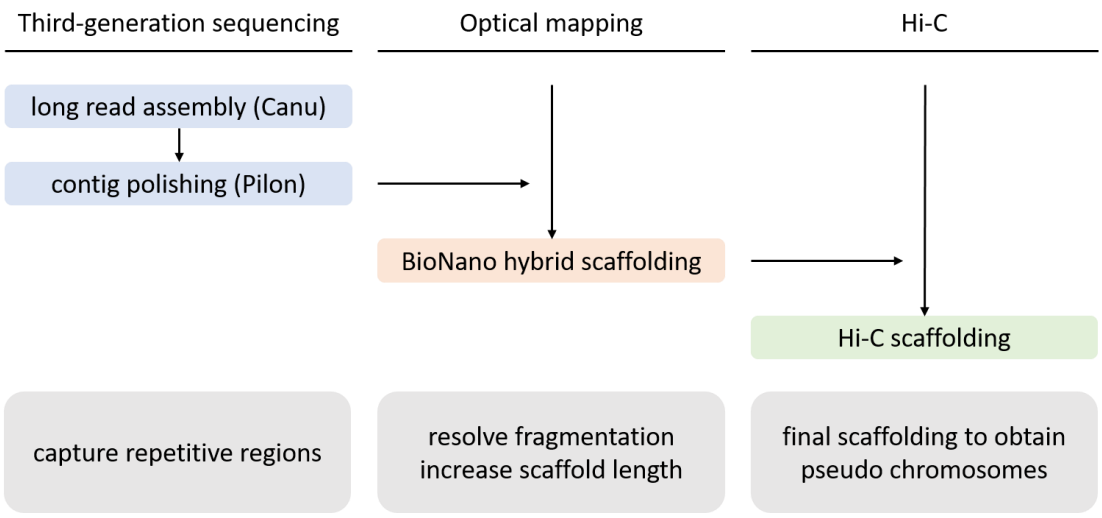


Figure 7.4 Overview of the assembly workflow to obtain a chromosome-scale assembly for *Lolium perenne*.

scaffolding with Bionano optical maps improved the assembly contiguity and reduced the number of contigs by half. Ultimately, the incorporation of the spatial context of chromosomal conformation capture delivered by Hi-C sequencing resulted in incorporation of 92% of the scaffolds into seven pseudo-chromosomes. After manual curation of pseudo-chromosomes and unanchored scaffolds, the final chromosome-scale genome assembly for perennial ryegrass comprised 2.5 Gbp, of which 90.46% was incorporated in seven pseudo-chromosomes.

It was only by combining NGS with third-generation sequencing, optical mapping and Hi-C that a chromosome-scale assembly for the large and highly repetitive *L. perenne* genome could be obtained. The v1.4 genome sequence contains half of the genome size, represented in 48k scaffolds and as many gaps. In other words, the average gap size is equal to the average scaffold size, namely 20kb, which, in turn, corresponds to the limit of effective sequence range distance that is connected by Hi-C crosslinks. Beyond that range, Hi-C assembly loses signal strength. Using Hi-C to anchor the scaffolds of the v1.4 genome sequence, therefore, did not result in an assembly of pseudo-chromosomes, as bridging gaps that are larger than 20kb is not possible.

The combination of large sequence contigs, optical maps, and Hi-C scaffolding data provides a very powerful set of resources for genome assembly. Only recently, a similar approach has been used to obtain a high-quality reference sequence for barley (5 Gbp) and wheat (17 Gbp) (Mascher et al., 2017; Appels et al., 2018). This clearly illustrates our innovative strategy for the new

assembly of the perennial ryegrass genome, having a similar quality compared to these major crop genomes. However, obtaining a contiguous genome assembly remains challenging, requires a great effort in both time and resources, and is still subject to technical artefacts. The quality and the length of the backbone contigs used for optical mapping and Hi-C scaffolding greatly affects the resulting assembly. Hi-C was originally developed to detect intra- and inter-chromosomal interactions, meaning that Hi-C scaffolding data includes linkages between loci co-localized on the same chromosome as well as structural DNA linkages, because telomeric and centromeric regions of different chromosomes are co-localized in the nucleus (Belton et al., 2012). This creates a risk of incorporating inter-chromatin interactions into the linear assembly. Validation of the assembly should be done during hybrid scaffolding and Hi-C scaffolding, to identify discrepancies that require adjustments or corrections (Udall and Dawe, 2018). Some genomic regions are easily corrected, e.g. the mitochondrial genome was manually removed from pseudo-chromosome 7 of *L. perenne*, while others require multiple iterations to resolve or are likely to remain unresolved and require local reassembly of the underlying DNA sequence.

The most important advantage of the v2.6.1 genome sequence is that the chromosome-scale sequence was obtained without relying on synteny with closely related grass species. There were no prior assumptions on gene content and order used, so it is now possible to answer questions in the context of comparative and evolutionary studies. For instance, it is now possible to define the precise breakpoints of the large-scale translocation between the long arms of chromosomes 4 and 5 of perennial ryegrass and Triticeae species, such as barley, and the full set of genes that were translocated.

Annotation of the chromosome-scale genome sequence of *Lolium perenne*

From improving the gene annotation set for the v1.4 genome sequence, we have compiled a useful strategy for gene annotation in *L. perenne* resulting in the best gene annotation set possible with high accuracy and completeness. Using this knowledge, a similar strategy was used to generate a gene annotation set for the v2.6.1 genome sequence.

A first gene annotation set was generated with Mikado (Venturini et al., 2017). Mikado provides a framework for integrating transcripts from multiple sources into a consolidated set of gene

annotations. It defines gene loci, scores transcripts and defines a representative model for each locus. Finally, transcripts that are chimeric, fragmented or that have short or disrupted coding sequences are removed, returning a filtered set of gene models according to the requirements. The input data consisted of aligned proteomes from four related species (*B. distachyon, O. sativa, Z. mays* and *S. bicolor*), four sets of aligned transcripts (GMAP, Wu and Watanabe (2005)) and corresponding TransDecoder bed files, and RNA-Seq alignments of seven different tissues (leaves, roots, meristems, leaf sheets and inflorescence) and corresponding high quality splice junctions (Portcullis, Mapleson et al. (2017)). Mikado turned out to perform well for *L. perenne*, leading to annotation of 101,876 genes (Table 7.3). A second annotation set was generated using Maker (Campbell et al., 2014), with six rounds of training to include *ab initio* gene predictions. The resulting gene annotation set contained 321,636 genes, three times more compared to the Mikado annotation set, and consists of a large portion of false positive gene annotations. This result is similar to the performance of Maker followed by eight rounds of training on the v1.4 genome sequence.

The Maker annotation set is more complete than the Mikado annotation set, with only 247 missing coreGFs (Table 7.3). The gene model accuracy was calculated using a projection of the 503 manually curated gene models onto the v2.6.1 genome sequence and retaining 437 genes with an exact match. The F1 measure denoting nucleotide-level gene model accuracy was slightly higher for the Mikado annotation set compared to Maker. However, only 52% of the Mikado transcripts lead to a correct protein sequence (including start and stop codon, and not containing internal stop codons), while for Maker this was almost 90%. Although Maker leads to a strong over-prediction of the total number of genes, the gene models seem to be more accurate compared to Mikado.

We decided to use the EVM to integrate Mikado and Maker annotations in order to benefit from the strengths of both methods: Mikado is strongly evidence-based, while Maker also includes *ab initio* gene predictions with high quality gene model structures. The resulting EVM annotation set contained 169,635 genes, and Maker gene models that were not supported by any other type of evidence were removed, leaving 139,003 genes in the final EVM annotation set (Table 7.3). Both BUSCO and CoreGF completeness measures were comparable to Maker, and 85% of the

transcripts lead to correct protein sequences. This means that the accuracy of the gene models did improve compared to the Mikado gene models, a large portion of false positive gene annotations are already removed compared to Maker, and all of this not at the expense of the completeness of the annotated gene space estimated with BUSCO and coreGFs.

Table 7.3 Comparison of gene annotation sets for the v2.6.1 genome sequence.

	Mikado	**Maker**	**EVM**	**EVM$_{HC}$**
Number of genes	101,876	321,636	139,003	48,812
Accuracy (F1 measure)	96.01%	94.35%	93.30%	93.24%
BUSCO (n = 1,440)	Complete: **87.4%** [Single: 63.0%, Duplicated: 24.4%] Fragmented: 4.6% Missing: 8.0%	Complete: **93.5%** [Single: 88.8%, Duplicated: 4.7%] Fragmented: 3.9% Missing: 2.6%	Complete: **93.1%** [Single: 89.7%, Duplicated: 3.4%] Fragmented: 3.1% Missing: 3.8%	Complete: **93.1%** [Single: 89.7%, Duplicated: 3.4%] Fragmented: 3.1% Missing: 3.8%
PLAZA CoreGF (n = 7,076)	82.18% 951 missing	**95.21%** 247 missing	**93.74%** 315 missing	**92.63%** 370 missing

The number of genes in the EVM gene set (139,003) is much higher compared to closely related, diploid species (34k in *B. distachyon* (JGI v3.1) and 26k high confidence genes in barley (Ensembl Genomes ASM32608v1). The higher number of predicted genes is because no repeat masking of the v2.6.1 genome sequence was performed prior to the gene annotation. Instead, the EVM gene set was filtered into a high and low confidence gene set, based on different selecting criteria. First, functional descriptions were generated using PLAZA, as described for the v1.4 EVM gene annotation set. Genes were considered high confidence, if they showed homology to reference proteins of related species (*A. thaliana*, *O. sativa*, *S. bicolor* and *B. distachyon*), did not overlap with a repeat region, and had no functional description resembling a transposable element (TE). The final high-confidence gene set (EVM$_{HC}$) contained 48.812 genes, and had similar accuracy and completeness compared to the EVM gene set (Table 7.3). This inverse approach of identifying TEs after gene prediction is more liberal, and allows for the identification of novel TE-mediated genes that play a role in e.g. tolerance to abiotic stresses (Sahebi et al., 2018).

An accurate and complete gene annotation set for the chromosome-scale v2.6.1 genome sequence will contribute to the interpretation of trait-associated SNP markers as described above. The full chromosomal context has become available, leading to a more accurate identification of the closest gene both up and downstream, compared to the v1.4 genome assembly in which only 2.1 genes are joint per scaffold on average.

7.3 Insights in the genomic sequence diversity of perennial ryegrass

Because of its outbreeding nature, individual plants are highly heterozygous and the diploid perennial ryegrass genome is highly diverse both within and across breeding populations and natural accessions. Previous studies have reported that there are on average five SNPs every 100 bp in the *L. perenne* transcriptome (Studer et al., 2012; Ruttink et al., 2013). The high density of sequence variants creates a challenge for establishing a comprehensive catalog of genomic sequence variation, as there is no universal standard variant calling pipeline to use, nor a reference variant set available to compare with, or to use for variant calling calibration.

Challenges in identification of genomic variation in *L. perenne* using standard variant calling pipelines

At first, a single variant calling pipeline was used to identify sequence variation for 503 candidate genes in 743 genotypes. The GATK HaplotypeCaller is the most commonly used variant calling pipeline, and can identify SNPs and indels simultaneously via local *de novo* assembly of haplotypes in an active region. By completely re-assembling reads in a region showing sequence variation, the GATK HaplotypeCaller is more accurate in regions where different types of variants are present in close proximity, and is also more accurate in the identification of indels compared to position-based callers like the GATK UnifiedGenotyper.

However, when using different VC pipelines, the concordance of the resulting variant sets was surprisingly low. This has lead us to the conclusion that the identification of *de novo* genomic sequence variants is not straightforward in a highly heterogeneous species, such as perennial ryegrass. Therefore, we developed two complementary strategies. First, variants sets generated by four VC pipelines were automatically integrated to reach maximal sensitivity. To reach maximal precision, highly multiplex amplicon sequencing was used as an independent

genotyping technology to empirically estimate an appropriate precision threshold. Second, an alternative strategy based on *de novo* assembly followed by overlap-layout-consensus clustering was used to reconstruct sets of divergent alleles for each candidate gene. This is required for the detection of highly divergent alleles that are missed during variant calling relying on alignment of short reads to a single reference genome sequence. This approach is applicable to other highly diverse outbreeding species and provides important insights in the pitfalls and solutions of bioinformatics analyses of populations-scale genome resequencing studies.

Implications of high variant density for allele frequency profiling with GBS

The observation that in the perennial ryegrass germplasm there is more than one SNP every 100 bp has some serious complications for standard analyses being used for genotypic screening of ryegrass populations. Allele frequencies of molecular markers in populations are the basis for many population genetics analyses. The GrassLandscape project makes use of GBS pool-Seq to efficiently profile allele frequencies in 550 ryegrass natural populations (Blanco-Pastor et al., 2018).

Using haplotypes in association genetics studies can improve the inference of population structure and provide higher power and precision, as they exploit LD information from multiple markers (Lorenz et al., 2010). Inferring haplotypes from GBS data currently relies on an *in silico* restriction digest of the reference genome in order to determine were stacks of reads will align. Not surprisingly, one of the main reasons that GBS suffers from data incompleteness is that SNPs and indels occur at restriction sites, the molecular basis of all classical AFLP analyses. These polymorphisms may cause gain, loss or shift of restriction enzyme recognition sites, leading to gain or loss of amplifiable fragments and absence/presence polymorphisms of divergent alleles per GBS locus. Additionally, detailed analysis of read mapping profiles revealed that local variation in read mapping position occurred consistently, and fairly frequently, within GBS loci of heterozygous *L. perenne* individuals, leading to partially overlapping reads, instead of perfectly aligned read stacks flanked by restriction sites in the reference genome.

In order to increase the accuracy for allele frequency profiling based on pooled GBS data, we have developed a novel method to delineate the locations of stacks mapping to the genome, and to identify variation in stack start and end positions across individuals and pools. Figure 7.5 illustrates the workflow on how to delineate Stack Mapping Anchor Points (SMAPs). The input files consist of BAM files, generated by mapping reads per sample to the reference genome. Per sample, stacks are delineated as reads mapping to the reference with exactly the same start and end mapping position. Overlapping stacks are then merged into stack clusters, meanwhile keeping track of all start and end positions within a stack cluster (i.e. SMAPs). Next, stack clusters are integrated across all samples based on positional overlap, thereby combining all possible SMAPs per locus across all samples. Finally, the SMAPs are combined with variant positions, and haplotypes can be created using combinations of the SMAPS as well as the intermittent SNPs.

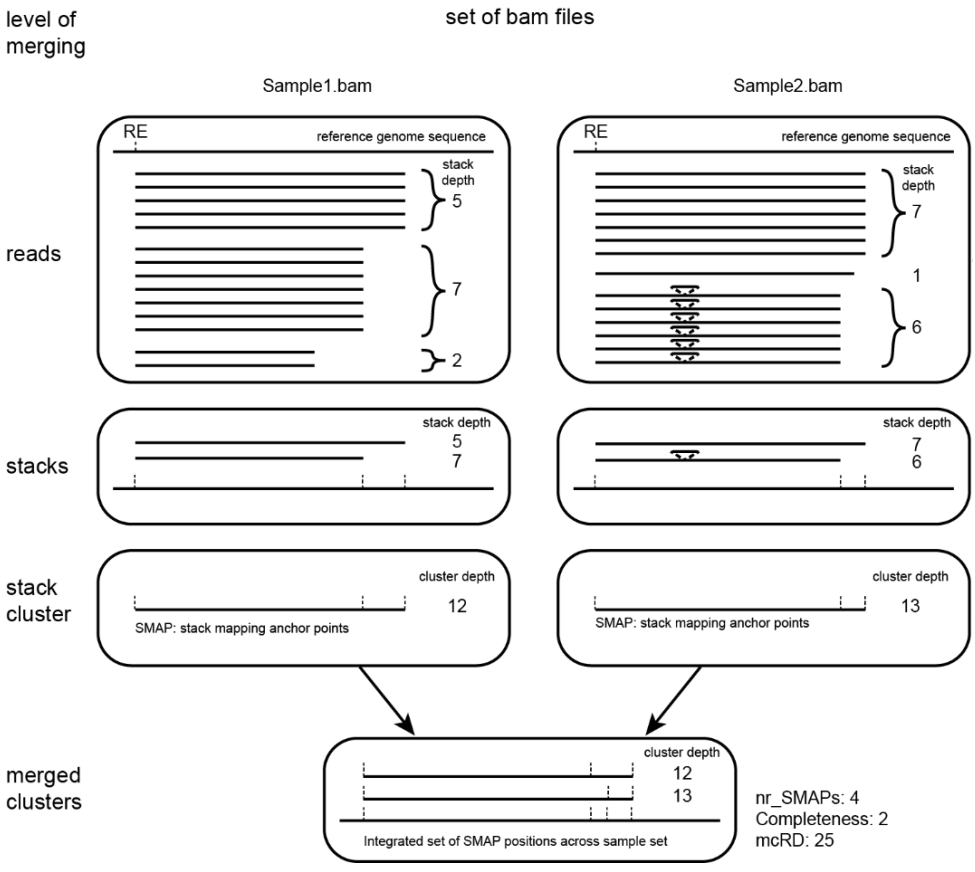


Figure 7.5 Overview of the delineation of stack mapping anchor points (SMAPs).

This procedure can be applied for GBS using both single and double enzyme digests, and can analyze read mapping data of both individual and pooled samples as well as single-end and paired-end read data. As the delineation of the read mapping positions is independent from an *in silico* restriction digest, this method is also more robust and reduces the fraction of missing data in a highly heterozygous species, such as *L. perenne*. Moreover, by capturing read mapping polymorphisms, a new type of molecular marker is generated that reflect underlying sequence variation, including indels that are otherwise difficult to detect. These markers can be used in a similar way as traditional SNP and indel data for allele frequency profiling and association studies.

Possible applications for the catalog of sequence diversity of 503 candidate genes

Using targeted resequencing for 503 candidate genes involved in pathways that control biological processes underlying agronomic traits, we constructed a comprehensive catalog of genomic variation for a *L. perenne* germplasm collection of 736 genotypes derived from current cultivars, breeding material and natural accessions. The final variant set consisted out of 252,406 SNPs and 5,074 indels (EP > 80%) in 2.3 Mbp, corresponding to up to 10 variants per 100 bp.

The catalog of sequence variation is useful for many purposes:

- In an association genetics study to uncover genes involved in important agronomic traits. The candidate genes were chosen based on orthology with genes that are known to be involved in plant growth and development from experimental studies in model species such as *A. thaliana.* These genes were resequenced in 736 genotypes that are representative for the *L. perenne* germplasm, and each of these genotypes was extensively phenotyped for different traits related to plant architecture (heading date, number of tillers, leaf length etc.) and cell wall digestibility. Moreover, some variants may be causal for a certain phenotype, or may be strongly linked to a causal variant.
- To identify genomic sequence variants that affect gene function and regulation, using the manually curated gene model for each of the 503 candidate genes. Our analysis showed that naturally occurring LOF alleles could be readily identified in as much as

one-third of the genes (Chapter 4). As illustration, we identified premature stop codons in ERA and GI, and in Chapter 5 for the FLOWERING LOCUS T gene family. Variants that disrupt gene function are often rare, which makes it very difficult to identify them using a classic association mapping study because they lack statistical power. This approach operates in the opposite direction and is thereby complementary.

- To select carriers of interesting variants. These variants may be derived from association mapping studies or from the prediction of their effect on gene function. Using the corresponding genotypes, it becomes possible to validate associations, or to study their effect on a phenotypic trait in more detail.
- For the design of genotyping platforms to screen breeding populations for alleles associating with important agronomic traits. As described in Chapter 6, highly multiplex amplicon sequencing can be used as a time- and cost-efficient method to screen thousands of individuals. Prior knowledge is necessary for the primer design for the amplicons needed to screen a locus of interest, as primers should be designed in a sequence region free of SNPs to avoid interference with primer binding.

7.4 Conclusion & Perspectives

This PhD has contributed to the development of genomic resources for perennial ryegrass. An overview of the three resources central in this these is presented in Figure 7.6.

We have improved the gene annotation set for the current draft reference genome sequence (Byrne et al., 2015), and generated functional annotations using the PLAZA comparative genomics platform. A new chromosome-scale reference genome sequence was obtained by integration of third-generation sequencing, optical mapping, and Hi-C, together with a high quality gene annotation set. This provides a high-quality framework for synteny-based transfer of QTL markers and the identification of candidate genes involved in physiological or morphological processes related to important agronomic traits. The quality of this reference sequence is now on par with golden standard reference genomes of important food crops, such as barley and wheat. Further comparative genomics studies with other species have become

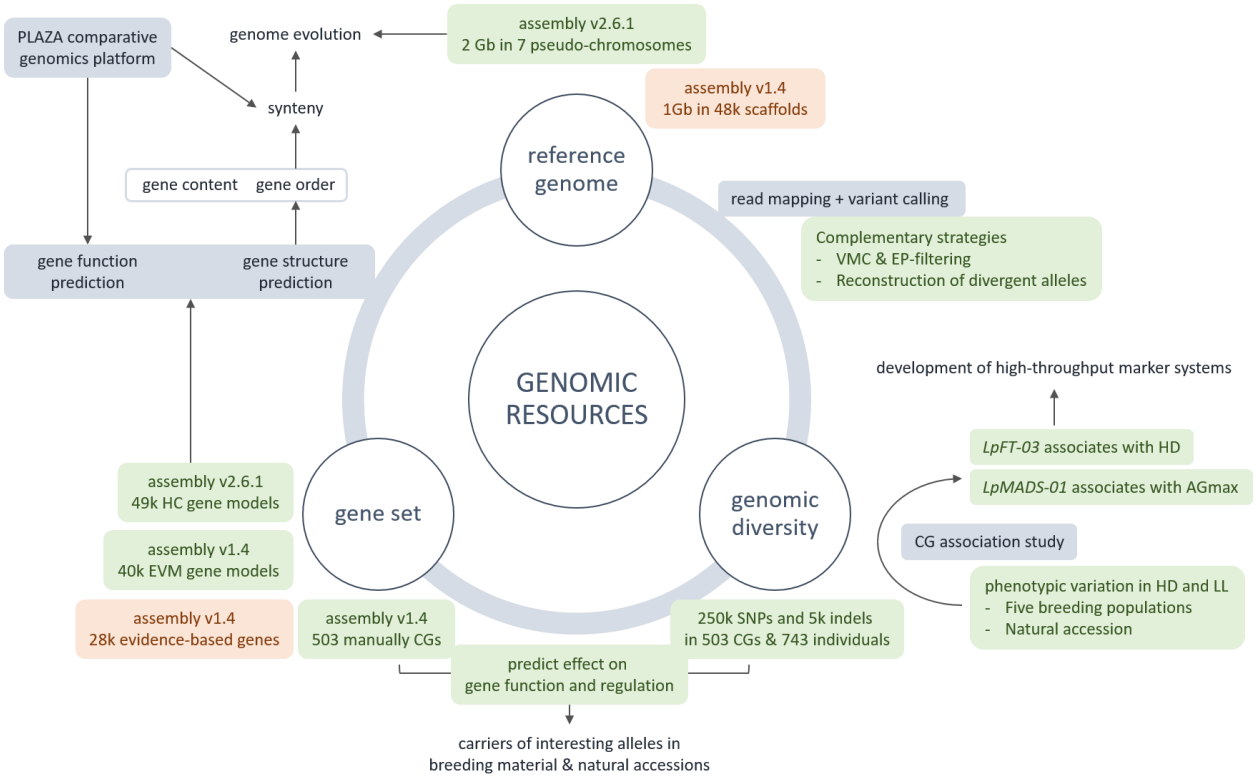


Figure 7.6 Overview of genomic resources that were developed for *L. perenne*. Resources available at the start of this PhD are indicated in orange, and resources generated during this PhD are indicated in green. (VMC: VariantMetaCaller, EP: estimated precision, EVM: EVidence Modeler, CG: candidate gene, HD: heading date, LL: leaf length)

possible and allow the study of evolution and architecture of the perennial ryegrass genome, in order to unravel species-specific biology explaining the typical characteristics of this major cultivated grass species.

At the beginning of this PhD, using a single variant calling pipeline was the most common approach for *de novo* identification of genomic sequence variants in perennial ryegrass. However, using different variant calling pipelines generated low concordant variant sets. Additionally, using hard filtering criteria to identify true sequence variants is not straightforward. In this thesis, we have presented complementary strategies to generate a complete and reliable variant set, overcoming challenges related to the highly polymorphic nature of this species. The newly developed strategy based on the integration of four different variant calling pipelines and precision-based filtering resulted in a highly accurate variant set with high genotype call rate. Reconstruction of divergent alleles for highly polymorphic genes, such as *LpSDUF247*, clearly illustrated the blind spot of variant calling pipelines that rely on mapping short reads to a single

reference sequence. A combination of both strategies should be used for future *de novo* identification of genomic sequence variation in *L. perenne* and other highly heterozygous outbreeding species, to get insights in the full extent of the genomic sequence variation.

The catalog of sequence variants for 503 candidate genes that control plant growth and architecture has provided insight in the genotypic diversity present in the *L. perenne* germplasm, and will be used for further genetic association studies with related traits. Furthermore, as regions with low SNP density can be distinguished from regions with high SNP density for primer design, this collection of sequence variants can be used to design high-throughput molecular marker assays for the 503 candidate genes.

The causal relationship between genomic sequence variants and the phenotypic differences observed across individuals is of fundamental biological interest. This relationship can be tested in a forward genetics approach, using candidate gene association studies, genome-wide association studies (GWAS) and genome-wide allele frequency fingerprinting (GWAFF). Ultimately, perennial ryegrass breeding programs will benefit from the development of molecular markers followed by marker-assisted selection. More recently, genomic selection is gaining interest, as it is better suited to improve complex traits with low heritability (Fè et al., 2015; Faville et al., 2016; Guo et al., 2018; Pembleton et al., 2018). These studies require both high-quality phenotypic and genotypic data. The latter can be provided using the newly developed strategies to identify genomic sequence variants. The delineation of SMAPs provides a novel type of molecular markers that can be used to increase the accuracy for allele frequency profiling based on pooled GBS data. Finally, the candidate gene association study performed in Chapter 6, indicated that single markers fail to distinguish between different alleles present in a population. This information can be captured by the reconstruction of haplotypes, which can also be used in an association genetics study (Yang et al., 2013; Yano et al., 2016). Haplotype-based association mapping has some benefits over single-marker association mapping, as they may be in closer LD with a causal variant, or haplotypes themselves are the causal variant of interest (Stram and Seshan, 2012). Haplotype analysis can therefore complement the well-established quantitative genetics frameworks in crops, such as quantitative trait analysis and genomic selection (Qian et al., 2017). Ideally, improved genotyping strategies lead to better identification

of genomic sequence variation underlying phenotypic variation. Prioritizing markers in genomic selection can also attribute to the prediction power and to the identification of causal variants (Chang et al., 2018). A combination of genome-wide markers derived from GBS data, and highly predictive markers screened using a targeted amplicon sequencing assay can contribute to the development of future genomic selection strategies in perennial ryegrass.

In outbreeding species, defective alleles occur in natural populations at low frequency and usually occur in a recessive heterozygous state (Marroni et al., 2011). An association genetics study is typically insensitive to detect the effect of alleles with low frequency in the genotype collection, as the number of replicate observations is too low, leading to a decreased statistical power. However, rare alleles may account for a substantial fraction of the unexplained phenotypic variation present. Using bioinformatics tools, a collection of sequence variation can be queried for polymorphisms that severely disrupt gene function, such as premature stop codons and frameshift mutations. This reverse genetics approach should be used complementary to an association genetics study (forward genetics approach). As such, a possible LOF mutation was detected in one third of the 503 candidate genes. Using additional information on conserved and essential amino acid residues was illustrated for five members of the FLOWERING LOCUS T gene family: non-synonymous sequence variants were identified affecting amino acid residues in the external loop and ligand binding pocket. The number of genes with a LOF mutation is higher compared to previous observations in perennial ryegrass (Ruttink et al., 2013) and other plant species (Clark et al., 2007; McNally et al., 2009). It is likely that observed number of LOF mutations is an overestimation: when taking the full sequence context into account, it can be expected that other variants can be identified that compensate the LOF mutations (Gan et al., 2011). This can be assessed by performing *de novo* assembly of the full gene per genotype, similar to our strategy to reconstruct divergent alleles, and performing gene model prediction on the resulting contigs. The rare defective alleles should be validated in a future experiment, for instance by crossing two heterozygous plants to yield a homozygous line, in which the effect of the LOF mutation in a gene of interest can be studied. Carriers of the LOF mutations in *LpGI-01* and *LpERA-01* are good candidates, as these genes are member of single-copy gene families.

The high level of genomic sequence diversity that is inherent to outbreeding crops, such as perennial ryegrass, is a rich resource of genetic variation that underlies phenotypic variation and forms the core unit of breeding. However, the identification of genomic sequence variation presented in this thesis was based on the use of a single reference genome sequence, representing a single individual. This does not reflect the diversity in genome content and organization that exists across individuals. Moreover, copy number variations and presence/absence variations greatly contribute to intra-species genetic variation (Zmienko et al., 2014; Bai et al., 2016). The concept of a pan-genome was introduced to describe the variation among closely related bacterial strains belonging to the same species, and has been extended to the plant kingdom thanks to technological advances and reduced sequencing technology costs. A pan-genome refers to the non-redundant set of sequences present across individuals of the same species. It consists of two sets of sequences: those present in every individual (the core genome), and those present in only a subset of individuals (the dispensable genome). Pan-genome analyses have been applied to a number of model and crop species such as *A. thaliana* (Cao et al., 2011), rice (Yao et al., 2015), maize (Hirsch et al., 2014), and wheat (Montenegro et al., 2017). The availability of additional reference genomes for perennial ryegrass will greatly facilitate structural variation characterization and lead to a better understanding of the perennial ryegrass genome. However, some challenges still need to be overcome before these new genomic resources can be effectively used: (i) the definition of a reference genome needs to be reconsidered: is it the genome of a selected individual, the consensus sequence of a population, or is it equal to the pan-genome?, (ii) new bioinformatics workflows are required to translate coordinates and compare genome features between assemblies, and (iii) should we abandon the concept of single, linear reference genome and move towards a graph-based approach? (Computational Pan-Genomics Consortium, 2016)

7.5 Author Contribution

Elisabeth Veeckman, Klaas Vandepoele and Tom Ruttink developed the annotation pipeline for the 1.4 genome sequence. Torben Asp (Aarhus University, Denmark) and Stephen Byrne (Teagasc, Ireland) provided input annotation tracks. Elisabeth Veeckman, Klaas Vandepoele and Michiel Van Bel were involved in the PLAZA build for functional annotation. Torben Asp and Istvan

Nagy (Aarhus University, Denmark) performed hybrid assembly using PacBio and optical mapping. Chang Liu (Tübingen University, Germany) and Tom Ruttink were involved in Hi-C analysis and scaffolding. Torben Asp, Istvan Nagy and Elisabeth Veeckman were involved in the annotation of the v2.6.1 genome sequence. Thomas Keep, José-Luis Blanco-Pastor, Jean-Paul Sampoux and Philippe Barre (INRA, France) provided trait-associating variants identified in the EU-project GrassLandscape. Stephen Byrne provided figures on the transcriptome analysis using the EVM annotation and functional annotation generated by PLAZA. Elisabeth Veeckman and Tom Ruttink designed and implemented the SMAP algorithm. Elisabeth Veeckman, Tom Ruttink and Klaas Vandepoele contributed to writing this chapter.

A Curriculum Vitae

Personal information	Elisabeth Veeckman Driesstraat 9, 9050 Ghent, Belgium elisabeth.veeckman@hotmail.com +32 4 74 65 72 76 Date of birth: 28/07/1991
Education	**2019** PhD in Bioinformatics, Ghent University, Belgium **2014** Master of Science: Biochemistry & Biotechnology, Ghent University, Belgium **2012** Bachelor of Science: Biochemistry & Biotechnology, Ghent University, Belgium
Technical skills	· Languages: Python, Perl, Bash, R, VBA · Operating systems: Microsoft operating systems, Linux · Database systems: MySQL · Cluster computing
Language skills	Dutch – native speaker English – full professional proficiency French – professional working proficiency
Selected conferences	**BIG N2N annual symposium**, May 21, Ghent, Belgium. Poster presentation. **EUCARPIA Symposium Section Fodder Crops and Amenity Grasses: Breeding in a World of Scarcity**, September 13-17, 2015, Ghent, Belgium. Oral presentation. **Applied Bioinformatics in Life Sciences**, March 17-18, 2016, Leuven, Belgium. Poster presentation. **Symposium: Genomics in Agriculture, Fisheries and Food (ILVO)**, April 21, 2015, Melle, Belgium. Oral presentation. **BIG N2N annual symposium**, May 19, 2015, Ghent, Belgium. Oral presentation. **International Plant and Animal Genome Conference**, January 14-18, 2017, San Diego, USA. Oral presentation. **12th Congress of the International Plant Molecular Biology**, August 5-10, 2018, Montpellier, France. Poster presentation. **EUCARPIA Section Biometrics in Plant Breeding 2018**, September 3-6, 2018, Ghent, Belgium.

Selected training	**Computationele biowetenschappen** (by Prof. dr. Peter Dawyndt), September – December 2014, Ghent, Belgium **N2N multidisciplinary Seminar Series on Bioinformatics,** September 2014 – June 2015, Ghent, Belgium **Summer School Wetenschapscommunicatie 'Zeg't eens'**, July 2015, Leuven, Belgium **Effective Scientific Communication** (by Jean-Luc Dumont), November 13 and 27, and December 4, 2015, Ghent, Belgium **Introduction Day for new PhD students**, February 4, 2016, Ghent, Belgium **Project Management** (by Tom Jacobs), January 11, 18 and 25, 2016, Ghent, Belgium **Meeting Skills** (by Johan De Bruycker), November 24-25, 2016, Ghent, Belgium **Career Guidance for PhD students** (by Lucia Smit), October – December 2017, Ghent, Belgium
Publications	**Veeckman, E., Ruttink, T., and Vandepoele, K.** (2016). Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. Plant Cell **28,** 1759-1768. **Veeckman, E., Vandepoele, K., Asp, T., Roldán-Ruiz, I., and Ruttink, T.** (2016). Genomic Variation in the FT Gene Family of Perennial Ryegrass (*Lolium perenne*). In Breeding in a World of Scarcity, I. Roldán-Ruiz, J. Baert, and D. Reheul, eds (Cham: Springer International Publishing), pp. 121-126. **Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., van Parijs, F.R.D., Byrne, S.L., Asp, T., Studer, B., Rohde, A., Roldán-Ruiz, I., Vandepoele, K., and Ruttink, T.** (2018). Overcoming challenges in variant calling: exploring sequence diversity in candidate genes for plant development in perennial ryegrass (*Lolium perenne*). DNA Res, dsy033.

B Supplemental Tables and Figures

Supplemental Table 1 Datasets used to evaluate genome assembly and gene space completeness measures.

Species	Taxonomic clade	Size (Mbp)	Number sequences	N50 (Kbp)	Number ESTs	Reference
Arabidopsis thaliana	Rosids	125	7	23,460	1,529,700	Parra et al. (2007)
Capsella rubella	Rosids	219		15,100	NA	Haudry et al. (2013); Slotte et al. (2013)
Cicer arietinum L.	Rosids	738	181,462	39,990	44,618	Parween et al. (2015)
Nelumbo nucifera Gaertn.	Rosids	929	3334	3400	2207	Ming et al. (2013)
Primula veris	Rosids	302		164	NA	Nowak et al. (2015)
Pyrus communis L. 'Bartlett'	Rosids	265	142,083	27,400	450	Chagne et al. (2014)
Raphanus raphanistrum	Rosids	254	68,331	10	81,524	Moghe et al. (2014)
Vigna angularis	Rosids	443	3387	703	11,199	Kang et al. (2015)
Lolium perenne	Monocots	1128	48,415	70	19,774	Byrne et al. (2015)
Oryza sativa	Monocots	389	16	29,895	987,327	Parra et al. (2007)
Setaria italica	Monocots	510	37,854	47,600	66,027	Zhang et al. (2012)
Phalaenopsis equestris	Monocots	1086	236,185	359	5604	Cai et al. (2015)

* CEGMA score reported in Figure 2.4 was obtained from this reference.

Supplemental Table 2

Gene name	*A. thaliana* candidates	Reference	Gene family	Nr *B. distachyon* genes	Nr *L. perenne* candidates
Development					
BCH1	AT5G52570 AT4G25700	Fiore et al. (2006)	HOM03M002075	2	1
BRIZ	AT2G26000 AT2G42160	Hsia and Callis (2010)	HOM03M002863	2	1
CBP80	AT2G13540	Kuhn et al. (2007)	HOM03M004935	1	1
DRM1	AT1G28330 AT2G33830	Gonzali et al. (2006)	HOM03M005389	2	1
HB13	AT1G69780 AT5G03790	Silva et al. (2016)	HOM03M000110	24	1
HYL1	AT1G09700	Liu et al. (2011)	HOM03M000640	4	2
ING2	AT1G54390 AT3G24010	Lee et al. (2009)	HOM03M002363	3	2
RSM1	AT2G21650	Hamaguchi et al. (2008)	HOM03M000100	25	2
SAMDC4	AT5G18930	Cui et al. (2010)	HOM03M001070	3	3
Cell wall					
4CL	AT1G51680 AT3G21240	Hamberger and Hahlbrock (2004); Heath et al. (2002); Li et al. (2015); van Parijs et al. (2015)	HOM03M000192	17	4
ALDH	AT1G23800 AT1G54100 AT1G74920 AT1G79440 AT2G14170 AT2G24270 AT3G24503 AT3G48000 AT3G48170 AT3G66658	Skibbe et al. (2002)	HOM03M000214	11	11
IRX	AT1G27440 AT2G28110 AT3G57630 AT5G22940 AT5G61840	Brown et al. (2009)	HOM03M000476	8	6
C3H, C4H, F5H	AT2G30490 AT2G40890 AT4G36220	Raes et al. (2003); Vanholme et al. (2012)	HOM03M000011	118	8
CAD	AT4G34230	Eudes et al. (2006); Kim et al. (2004); Raes et al. (2003); van Parijs et al. (2015)	HOM03M000073	36	2

CAD2	AT3G19450	Eudes et al. (2006); Kim et al. (2004); Raes et al. (2003); van Parijs et al. (2015)	HOM03M000256	8	10
CCoAOMT	AT4G34050	Raes et al. (2003); Vanholme et al. (2012)	HOM03M000557	8	7
CCR	AT1G15950 AT1G80820	McInnes et al. (2002); Tu et al. (2010); van Parijs et al. (2015)	HOM03M000073	36	7
CES	AT5G44030 AT5G17420 AT4G18780	Endler and Persson (2011); Persson et al. (2007)	HOM03M000097	26	8
COMT	AT5G54160	Heath et al. (1998); Raes et al. (2003); Tu et al. (2010); van Parijs et al. (2015)	HOM03M000105	22	5
HCT	AT5G48930	Vanholme et al. (2012)	HOM03M000043	63	22
HPRGP	AT2G18910	Johnson et al. (2017a); Johnson et al. (2017b)	HOM03M004130	1	1
LAC	AT2G38080 AT5G60020	Berthet et al. (2011); Zhao et al. (2013)	HOM03M000056	47	2
OFP	AT1G06920 AT2G18500 AT2G30400 AT2G36026 AT3G52525 AT4G18830 AT5G01840 AT5G19650 AT5G22240 AT5G58360	Li et al. (2011a); Wang et al. (2011)	HOM03M000249	17	1
PAL	AT2G37040 AT3G53260 AT5G04230 AT3G10340	Raes et al. (2003); Vanholme et al. (2012)	HOM03M000409	8	13
POX	AT5G66390 AT5G42180 AT5G51890	Novo-Uzal et al. (2013)	HOM03M000017	142	3
SND	AT1G28470	Zhong et al. (2008)	HOM03M000590	7	1
XylS	AT1G27600 AT2G37090	Wu et al. (2010)	HOM03M000746	8	6

XylT	AT2G03360 AT2G03370 AT2G41640 AT3G10320 AT3G18170 AT3G18180 AT3G57380	Voiniciuc et al. (2015)	HOM03M000169	21	4
Cell wall TF					
ERF	AT1G12980 AT1G24590 AT5G18560	Chandler and Werr (2011); Mehrnia et al. (2013); Nakano et al. (2006)	HOM03M000014	107	1
WRKY	AT2G44745 AT1G29860 AT2G38470	Eulgem et al. (2000); Guo and Qin (2016)	HOM03M000024	84	5
Cell wall TF, lateral organ identity or polarity, lateral organ initiation, lateral organ patterning and morphogenesis					
MYB	AT3G49690 AT5G23000 AT1G69560 AT2G37630 AT5G14750	Chen et al. (2006); Ehrenreich et al. (2007); Guo and Gan (2011); Ikezaki et al. (2010); Keller et al. (2006); Lee et al. (2009); Muller et al. (2006); Schmitz et al. (2002)	HOM03M000013	100	21
Cell wall TF, lateral organ patterning and morphogenesis, shoot apical meristem					
NAC	AT1G56010 AT5G53950 AT1G56010 AT3G15170 AT1G76420	Hasson et al. (2011); Hibara et al. (2006); Ooka et al. (2003); Raman et al. (2008); Vroemen et al. (2003)	HOM03M000023	73	11
Chromatin remodelling					
MET1	AT4G08990 AT4G13610 AT4G14140 AT5G49160	Finnegan et al. (1998); Li et al. (2011b)	HOM03M002194	2	3
SWI	AT3G06400 AT5G18620	Huanca-Mamani et al. (2005); Li et al. (2012)	HOM03M000104	28	1
Lateral organ initiation					
ANT	AT4G37750 AT4G36920	Ehrenreich et al. (2007)	HOM03M000117	24	4
SLOMO	AT4G33210	Lohmann et al. (2010)	HOM03M004562	1	1
TOP1A	AT5G55300	Laufs et al. (1998)	HOM03M003833	2	1
Lateral organ patterning and morphogenesis					

AS	AT1G65620 AT4G00220 AT5G63090	Ikezaki et al. (2010); Jun et al. (2010)	HOM03M000103	20	4
CLF	AT2G23380	Lopez-Vernaza et al. (2012)	HOM03M001993	2	2
DOT5	AT1G13290	Petricka et al. (2008)	HOM03M000656	7	2
GRF	AT2G22840 AT2G36400 AT3G13960	Kim et al. (2003)	HOM03M000362	12	4
KAN	AT5G42630 AT4G17695 AT1G32240 AT5G16560	Kerstetter et al. (2001); Pires et al. (2014)	HOM03M000071	40	6
NOV	AT4G13750	Tsugeki et al. (2009)	HOM03M002764	2	2
SE	AT2G27100	Prigge and Wagner (2001)	HOM03M001979	3	3
TRN1	AT5G55540	Cnops et al. (2006)	HOM03M004520	1	1
YABBY	AT2G45190	Sarojam et al. (2010)	HOM03M000503	8	4
ZPR1	AT2G45450	Kim et al. (2008); Wenkel et al. (2007)	HOM03M008264	1	1
ZPR3	AT3G52770	Kim et al. (2008); Wenkel et al. (2007)	HOM03M001755	3	1
Lateral organ identity					
AN3	AT4G00850 AT1G01160 AT5G28640	Vercruyssen et al. (2014)	HOM03M001522	3	3
BOP	AT2G41370 AT3G57130	Ha et al. (2007); Jun et al. (2010)Ha 2007; Hyung Jun 2010	HOM03M001913	2	2
HDZIPIII	AT1G30490 AT1G52150 AT2G34710 AT5G60690	Ehrenreich et al. (2007); Green et al. (2005); McConnell et al. (2001); Prigge et al. (2005); Talbert et al. (1995); Zhong and Ye (2001)	HOM03M000777	4	5
Light signalling					
bHLHABAI	AT1G32640 AT2G46510	Gangappa et al. (2013)	HOM03M000211	13	2
CO1	AT5G15840	Suarez-Lopez et al. (2001); Wang et al. (2013)	HOM03M000253	14	6
COP9	AT4G26430	Peng et al. (2001)	HOM03M003255	1	1
CRY	AT4G08920	Gao et al. (2015)	HOM03M001198	4	2

DET1	AT4G10180	Fernando and Schroeder (2015); Song and Carre (2005)	HOM03M004583	1	1
HY5	AT3G17609 AT5G11260	Ciolfi et al. (2013); Jang et al. (2013)	HOM03M001986	3	3
LHY	AT1G01060	Schaffer et al. (1998); Song and Carre (2005)	HOM03M007132	1	1
PCI	AT1G02090	Dessau et al. (2008)	HOM03M003249	1	1
PFT1	AT1G25540	Rival et al. (2014)	HOM03M004643	1	1
PHYB	AT2G18790	Finlayson et al. (2010); Reed et al. (1993)	HOM03M001131	4	2
PIF	AT1G09530 AT1G18400 AT1G26260 AT2G46970 AT3G62090 AT4G34530 AT5G61270	Bours et al. (2015); Nozue et al. (2011); Sun et al. (2013); Wei et al. (2017); Zhang et al. (2013)	HOM03M000047	54	6
SPA	AT1G53090 AT2G32950 AT2G46340 AT3G15354 AT4G11110 AT5G23730 AT5G52250	Komatsu et al. (2003)	HOM03M000659	4	3
Shoot apical meristem					
BARD1	AT1G04020	Han et al. (2008)	HOM03M001623	3	3
BLH	AT1G75410 AT2G23760 AT2G27990 AT2G35940	Kumar et al. (2007)	HOM03M000222	15	5
CLPS3	AT3G04680	Xing et al. (2008)	HOM03M004452	1	1
FTA	AT3G59380	Running et al. (2004)	HOM03M005386	2	1
KNAT	AT1G23380 AT4G08150 AT1G62360	Belles-Boix et al. (2006); Ehrenreich et al. (2007); Khan et al. (2012); Li et al. (2011a); Townsley et al. (2013)	HOM03M000310	12	6
OBE1	AT3G07780	Saiga et al. (2008)	HOM03M001177	4	2
ULT1	AT4G28190	Pires et al. (2014)	HOM03M003027	1	1
USP1	AT5G10790	Liu et al. (2008)	HOM03M000305	9	1
VEF2	AT5G51230	Yoshida et al. (2001)	HOM03M002064	1	2
WOX14	AT1G20700	Denis et al. (2017); Etchells et al. (2013)	HOM03M001281	4	1

WUS	AT2G17950 AT2G28610 AT2G33880	Long et al. (1996); Wang et al. (2017); Wu et al. (2005)	HOM03M000493	5	2
Self-incompatibility					
DUF247	AT3G50170 AT3G50120	Manzanares et al. (2016)	HOM03M000101	31	3
GK	AT1G80460	Manzanares et al. (2016)	HOM03M003925	3	1
Transition to flowering					
CCA	AT5G52660	Lu et al. (2011)	HOM03M000510	7	4
FCA	AT1G03457 AT4G03110 AT4G16280	Macknight et al. (1997)	HOM03M001246	3	4
FIE	AT3G20740	Chanvivattana et al. (2004)	HOM03M003503	4	1
FKF1	AT1G68050 AT2G18915 AT5G57360	Nelson et al. (2000); Sawa et al. (2007)	HOM03M001149	5	2
FLD	AT3G10390	Yu et al. (2011)	HOM03M000331	8	1
FPA	AT2G43410	Schomburg et al. (2001)	HOM03M006372	1	1
FT	AT1G18100 AT1G65480 AT2G27550 AT5G03840 AT5G62040	Ehrenreich et al. (2007); Niwa et al. (2013); Skot et al. (2007); Skot et al. (2011)	HOM03M000266	18	5
FVE	AT2G19520	Baek et al. (2008)	HOM03M000152	20	1
FWA	AT4G25530	Ikeda et al. (2007)	HOM03M000193	15	5
FY	AT4G15900 AT5G13480	Simpson (2003)	HOM03M000141	20	2
GI	AT1G22770	Oliverio et al. (2007)	HOM03M004581	1	1
LHP1	AT5G17690	Rizzardi et al. (2011); Valdes et al. (2012)	HOM03M005863	1	1
MBD9	AT3G01460	Peng et al. (2006); Yaish et al. (2009)	HOM03M002872	1	1
PHP	AT3G22590	Park et al. (2010)	HOM03M005329	1	1
RAV	AT1G13260 AT1G25560	Hu et al. (2004)	HOM03M000293	12	5
SDG8	AT1G77300	Cazzonelli et al. (2009)	HOM03M001151	3	2
SPL3	AT1G53160 AT2G33810 AT3G15270	Jiao and Meyerowitz (2010); Schwarz et al. (2008); Wang et al. (2009)	HOM03M000136	18	3
VIL3	AT2G18880	Sung et al. (2006)	HOM03M001250	5	3

VRN1	AT3G18990 AT4G01580	Levy et al. (2002)	HOM03M000432	14	1
VRN1-like	AT3G18990 AT4G01580	Levy et al. (2002)	HOM03M007993	2	1
Flower development and flower organ identity					
ESD4	AT3G06910 AT4G00690 AT4G15880	Hermkes et al. (2011)	HOM03M002166	3	2
HAC3	AT3G54610	Kim et al. (2015)	HOM03M005475	1	1
LFY3	AT5G61850	Ehrenreich et al. (2007)	HOM03M006114	1	1
LUG	AT2G32700 AT4G32551	Stahle et al. (2009)	HOM03M000732	5	4
MADS	AT1G69120 AT2G45660 AT3G54340 AT4G18960 AT4G24540 AT5G10140	Ehrenreich et al. (2007); Yanofsky et al. (1990)	HOM03M000042	52	9
RGA	AT1G14920 AT1G55580	Ehrenreich et al. (2007); Greb et al. (2003); Haywood et al. (2005); Raman et al. (2008)	HOM03M000051	43	6
SEU	AT1G43850	Bao et al. (2010); Ehrenreich et al. (2007)	HOM03M001061	3	3
SUF4	AT1G30970	Kim et al. (2006)	HOM03M003695	1	1
SUP	AT3G23130	Bowman et al. (1992)	HOM03M000478	5	4
ABA biosynthesis					
NCED1	AT3G63520	Finkelstein et al. (2002)	HOM03M000349	13	1
PDS1	AT1G06570	Norris et al. (1998)	HOM03M003395	2	2
PDS3	AT4G14210	Qin et al. (2007)	HOM03M002179	2	1
ABA signalling					
ABI1	AT4G26080	Finkelstein et al. (2002)	HOM03M000064	41	4
ABI3	AT3G24650	Ehrenreich et al. (2007); Finkelstein et al. (2002)	HOM03M000742	5	3
ABI5	AT2G36270 AT4G35900 AT2G17770	Finkelstein et al. (2002)	HOM03M000245	15	2
ABI8	AT3G08550	Finkelstein et al. (2002)	HOM03M002458	2	2
AIP3	AT1G08780	Kurup et al. (2000)	HOM03M005806	1	1
DRIP	AT1G08780	Kurup et al. (2000)	HOM03M000770	6	3
GBF	AT2G46270 AT4G01120	Lu et al. (1996)	HOM03M000451	12	4

GPA	AT2G26300	Pandey and Assmann (2004)	HOM03M000896	6	1
GTG2	AT4G27630	Jaffe et al. (2012)	HOM03M004596	1	1
HD2C	AT2G27840 AT5G22650 AT3G44750	Wu et al. (2000)	HOM03M001261	3	3
PSY	AT5G17230	Rodriguez-Villalon et al. (2009)	HOM03M001609	3	1
SAD1	AT5G48870	Xiong et al. (2001)	HOM03M006261	1	1
SIR3	AT1G16540	LeonKloosterziel et al. (1996)	HOM03M004196	1	1
WIG	AT5G40280	Ehrenreich et al. (2007)	HOM03M005784	1	1
ZEP	AT5G67030	Barrero et al. (2006)	HOM03M002490	1	1
Auxin biosynthesis					
TAA1	AT1G70560	He et al. (2011); Stepanova et al. (2011); Won et al. (2011)	HOM03M004340	1	1
TAR2	AT4G24670	He et al. (2011); Stepanova et al. (2011); Won et al. (2011)	HOM03M000928	3	1
YUC	AT4G32540 AT4G13260 AT5G25620 AT1G48910 AT1G21430 AT5G43890	Cheng et al. (2007)	HOM03M000262	16	4
Auxin signalling					
ADA2B	AT4G16420	Vlachonasios et al. (2003)	HOM03M002292	1	1
AMP1	AT3G54720	Chin-Atkins et al. (1996); Ehrenreich et al. (2007); Helliwell et al. (2001)	HOM03M001436	6	1

ARF	AT2G33860 AT1G19220 AT1G30330 AT5G62000 AT2G28350 AT1G19850 AT5G37020 AT1G59750 AT2G46530 AT1G34310 AT1G34170 AT1G35540 AT1G35520 AT3G61830 AT1G35240 AT1G34410 AT1G34390 AT1G43950 AT5G60450 AT4G23980	Dharmasiri and Estelle (2002); Ehrenreich et al. (2007)	HOM03M000112	25	6
AUXIAA	AT3G23050 AT5G25890 AT1G51950 AT1G04240 AT1G52830	Chen et al. (2013); Ehrenreich et al. (2007); Ploense et al. (2009)	HOM03M000126	26	3
AXR	AT1G55000	Hobbie (2006)	HOM03M006810	1	1
AXR1	AT1G05180	Aguilar-Martinez et al. (2007); Ehrenreich et al. (2007); Leyser et al. (1993); Stirnberg et al. (1999)	HOM03M003416	1	1
AXR4	AT1G54990	Hobbie (2006)	HOM03M004068	1	1
AXR6	AT4G02570	Ehrenreich et al. (2007)	HOM03M000391	9	2
CAND1	AT2G02560	Cheng et al. (2004)	HOM03M003522	1	1
GH3	AT4G27260	Park et al. (2007)	HOM03M000269	9	2
TIR1	AT3G62980	Ehrenreich et al. (2007)	HOM03M000440	9	1
Auxin transport					
AUX1	AT2G21050 AT2G38120	Petrasek and Friml (2009); Swarup et al. (2008); Ugartechea-Chirino et al. (2010)	HOM03M000858	3	2
ENP	AT4G31820 AT5G10250	Furutani et al. (2007); Treml et al. (2005)	HOM03M000108	25	3
PGP4	AT2G36910 AT2G47000 AT3G28860	Terasaka et al. (2005)	HOM03M000080	36	1

PID2	AT2G34650 AT1G53700 AT3G14370 AT2G26700	Cheng et al. (2007); Christensen et al. (2000); Ehrenreich et al. (2007); Friml et al. (2004); Pressoir et al. (2009)	HOM03M000079	30	2
PIN1	AT1G73590	Ehrenreich et al. (2007); Guenot et al. (2012); Petrasek and Friml (2009)	HOM03M000341	11	2
PIN1like	AT1G73590	Ehrenreich et al. (2007); Guenot et al. (2012); Petrasek and Friml (2009)	HOM03M004220	2	1
SPS	AT1G18350	Dai et al. (2006); Ehrenreich et al. (2007); Tantikanjana et al. (2004); Tantikanjana et al. (2001); Zhang et al. (2008)	HOM03M000059	45	1
Brassinosteroid biosynthesis					
DWF1	AT3G19820	Takahashi et al. (1995)	HOM03M003235	2	1
DWF3	AT3G50660 AT5G05690	Guo et al. (2010)	HOM03M000086	29	2
DWF5	AT1G50430	Silvestro et al. (2013)	HOM03M005311	1	1
DWF7	AT3G02580	Choe et al. (1999)	HOM03M003295	1	3
SQS	AT4G34640 AT4G34650	Kribii et al. (1997)	HOM03M002846	2	1
Brassinosteroid signalling					
BES1	AT1G19350 AT1G75080	Yin et al. (2002)	HOM03M000759	5	2
Cytokinin signalling					
ARR	AT1G19050 AT1G74890 AT3G48100 AT5G62920	D'Agostino et al. (2000); Efroni et al. (2013); Meng et al. (2017); Taniguchi et al. (1998); Wang et al. (2005); Wang et al. (2014)	HOM03M000363	11	4
CRE	AT2G01830	Ehrenreich et al. (2007)	HOM03M000267	6	4
GCR1	AT1G48270	Chen et al. (2004); Colucci et al. (2002)	HOM03M006708	1	1
RR	AT4G31920 AT3G16857	Argyros et al. (2008)	HOM03M000170	18	2
Ethylene biosynthesis					

ACS	AT3G61510 AT1G01480 AT5G28360 AT2G22810 AT5G65800 AT4G26200 AT4G37770 AT3G49700 AT1G62960 AT4G08040 AT4G11280 AT5G51690	Bours et al. (2015); Vandenbussche et al. (2003)	HOM03M000616	4	2
Ethylene signalling					
EBF1	AT2G25490	Binder et al. (2007)	HOM03M000829	4	3
EBF2	AT5G25350	Binder et al. (2007)	HOM03M007381	1	1
EIL3	AT1G73730	Wawrzynska and Sirko (2016)	HOM03M000703	6	2
EIN2	AT5G03280	Fischer et al. (2006)	HOM03M002491	3	1
ETO1	AT3G51770	Wang et al. (2004)	HOM03M001610	3	3
ETR1	AT1G66340	Bakshi et al. (2015)	HOM03M000531	7	3
Gibberellin biosynthesis					
GAOX	AT1G15550 AT1G78440 AT4G25420 AT5G51810	Luo et al. (2015)	HOM03M000022	84	11
Gibberellin signalling					
GID1A	AT3G05120 AT3G63010 AT5G27320	Marin-de la Rosa et al. (2011)	HOM03M000067	52	1
SHI	AT5G12330 AT5G66350	Fridborg et al. (2001)	HOM03M000773	4	2
SLY1	AT4G24210	Dill et al. (2004)	HOM03M007351	1	1
SPY	AT3G11540 AT3G04240	Greenboim-Wainberg et al. (2005)	HOM03M001501	4	1
Strigolactone biosynthesis					
D14	AT4G37470 AT3G03990	Arite et al. (2009)	HOM03M000849	5	2
D27	AT1G03055 AT4G01995 AT1G64680	Lin et al. (2009)	HOM03M001399	3	3
MAX1	AT2G26170	Bennett et al. (2006); Booker et al. (2005); Ehrenreich et al. (2007)	HOM03M001789	5	3
MAX3	AT2G44990	Bennett et al. (2006); Booker et al. (2005); Ehrenreich et al. (2007)	HOM03M003927	1	1

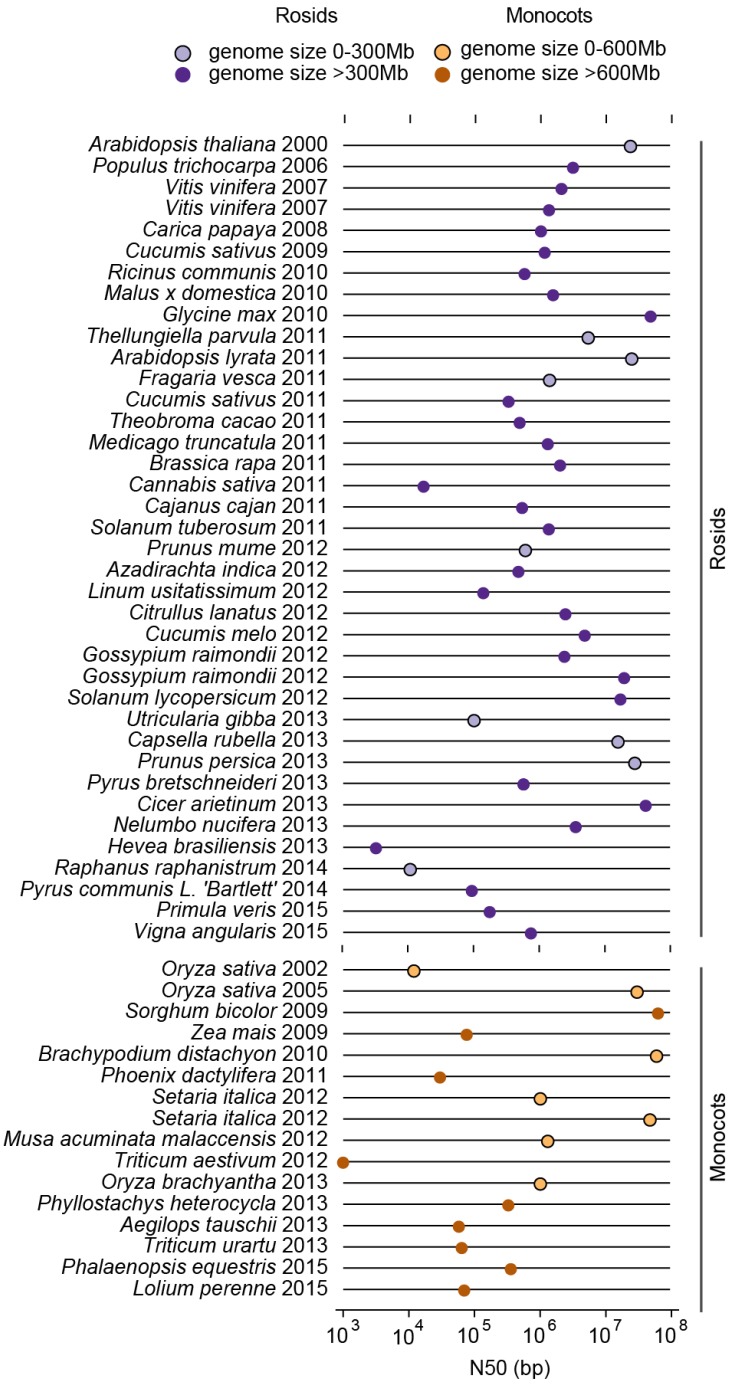
MAX4	AT4G32810	Arite et al. (2009); Bennett et al. (2006); Ehrenreich et al. (2007); Hayward et al. (2009)	HOM03M002869	1	2
Strigolactone signalling					
MAX2	AT2G42620	Bennett et al. (2006); Ehrenreich et al. (2007); Shen et al. (2007); Stirnberg et al. (2007); Stirnberg et al. (2002); Woo et al. (2001)	HOM03M005154	1	1
TB1	AT3G18550 AT1G68800 AT1G67260	Aguilar-Martinez et al. (2007); Cubas et al. (1999); Danisman et al. (2012); Efroni et al. (2013); Gonzalez-Grandio et al. (2013); Guo et al. (2010); Kosugi and Ohashi (2002); Koyama et al. (2010); Martin-Trillo and Cubas (2010)	HOM03M000311	10	3

Supplemental Table 3

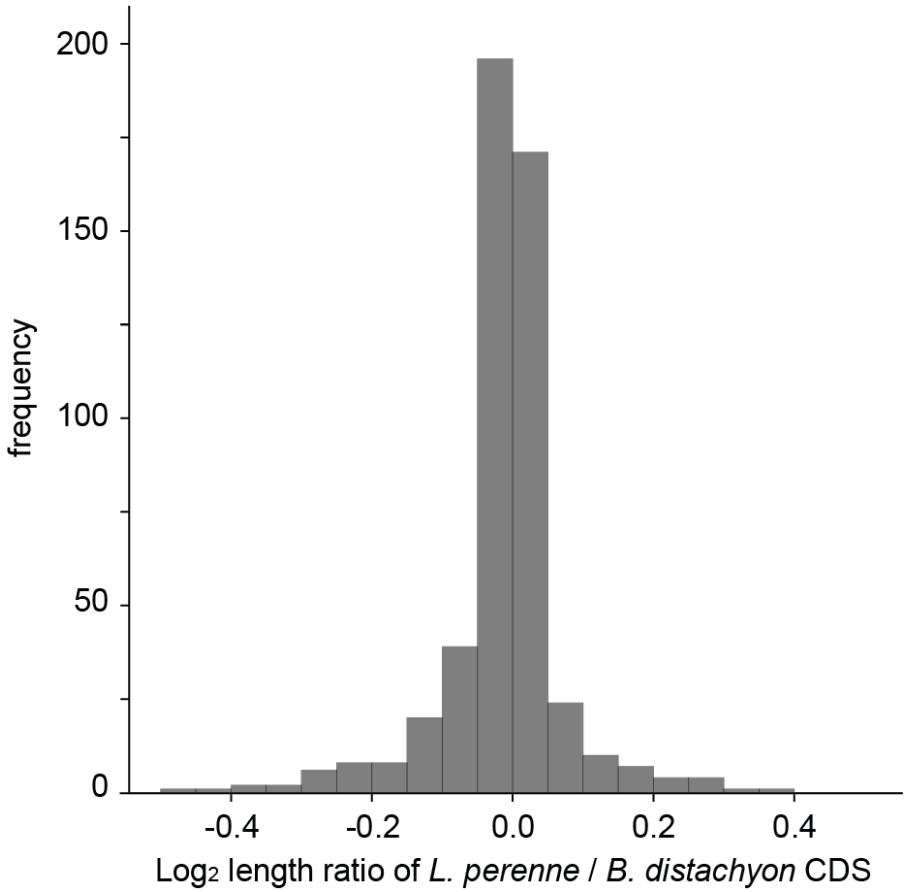
Gene	Pathway	Number of amplicons
MADS-01	Flower development and flower organ identity	34
MADS-05	Flower development and flower organ identity	3
AS-05	Lateral organ morphogenesis	3
CCA-02	Transition to flowering	2
CIB5-01	Light signaling	18
CO1-01	Light signaling	8
COL-01	Light signaling	3
COP9-01	Light signaling	2
CRY-01	Light signaling	4
CRY-02	Light signaling	3
FKF1-01	Transition to flowering	3
FLC-01	Transition to flowering	2
FLC-02	Transition to flowering	2
FLD-01	Transition to flowering	4
FPF1-01	Transition to flowering	1
FPF1-02	Transition to flowering	2
FT-03	Transition to flowering	24
GA2Ox-02	Gibberellin biosynthesis	11
GBF-03	ABA signaling	2
GCR1-01	Cytokinin signaling	9
PHYB-01	Light signaling	2
PHYB-02	Light signaling	2
PIF3	Light signaling	3
PIF3like2-01	Light signaling	3
SPL3-02	Transition to flowering	7
SPY-04	Gibberellin signaling	5
TOP1A-01	Lateral organ initiation	3
VRN2-01	Transition to flowering	6

Supplemental Table 4

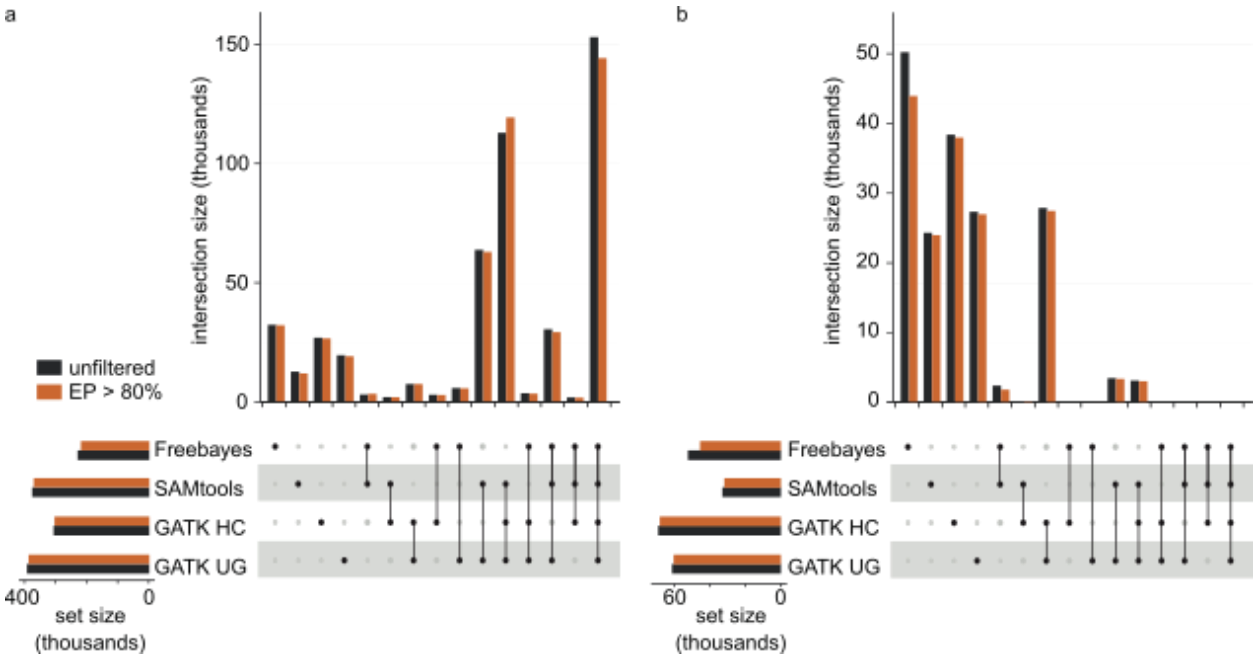
Population	p-value (FDR adjusted)	MAF	Genotypic classes (0/0 – 0/1 – 1/1)	Allelic effect (GDD and cm)
Significant association of HD with *LpFT-03* (scaffold 5073, position 58,536)				
1853-7	6.58E-01	30%	86 – 124 – 0	-20.55
5297xWAR10	2.92E-02	18%	224 – 105 – 8	54.84
Ba12990	8.25E-01	2%	136 – 5 – 0	-14.53
Ba12990x5554	1.00E+00	14%	311 – 7 – 0	23.35
Ba12990xplenty	6.02E-05	40%	114 – 203 – 41	-79.99
Significant association of leaf length with *LpMADS-01* (scaffold 312, position 70,351)				
1853-7	1.00E+00	33%	73 – 123 – 0	-0.63
5297xWAR10	3.10E-05	31%	128 – 201 – 0	-2.97
Ba12990	8.21E-01	37%	37 – 104 – 0	-2.83
Ba12990x5554	1.00E+00	43%	75 – 213 – 34	0.50
Ba12990xplenty	1.00E+00	25%	178 – 170 – 0	0.27



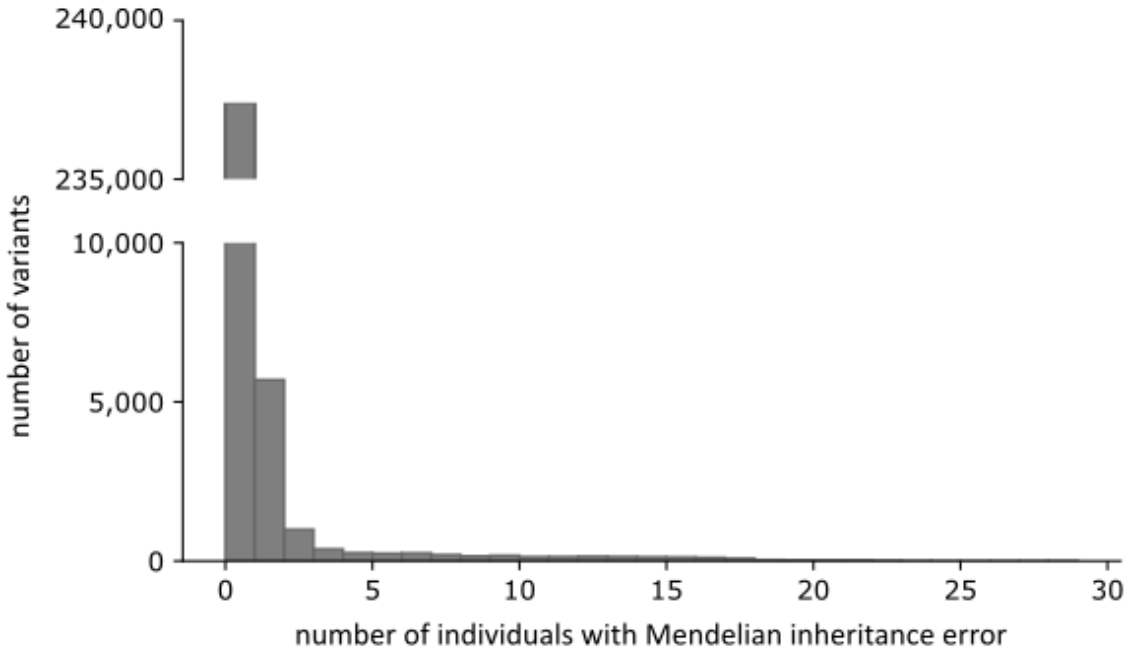
Supplemental Figure 1 N50 values for plant genomes published over the last 15 years. The N50 values of the first 50 published plant genomes were collected from Michael and Jackson (2013), complemented with the ten species used in the comparison of measures for genome and gene space completeness. The species are ordered according to their lineage (Rosids, purple; Monocots, orange) and publication date.



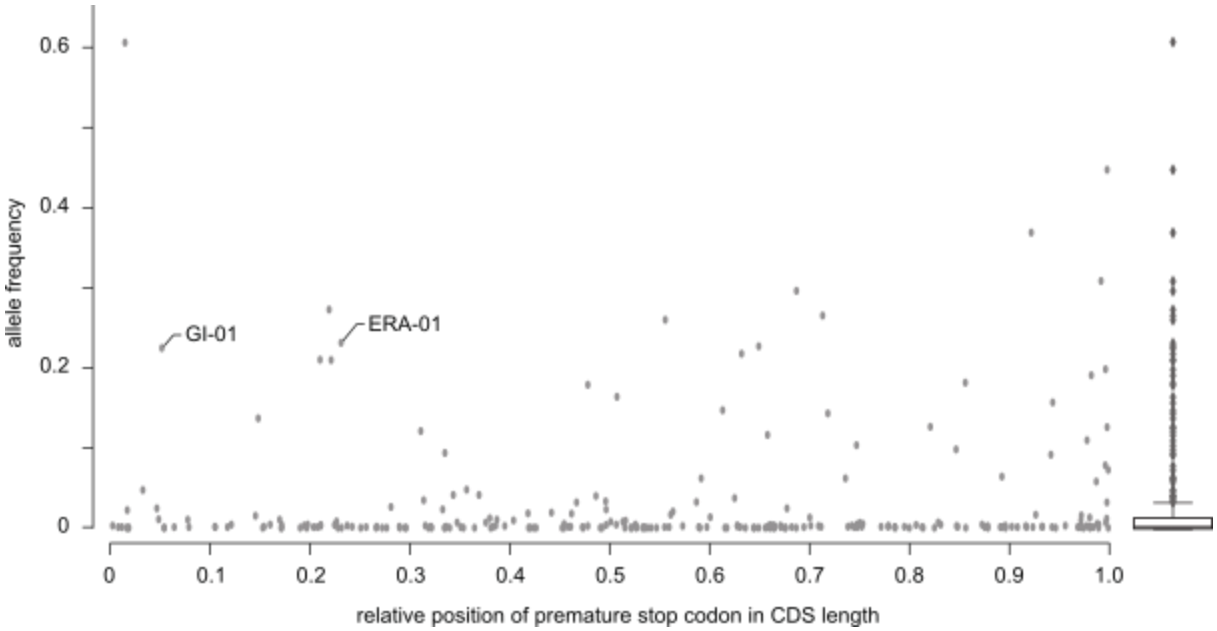
Supplemental Figure 2 Distribution of the protein length ratio of 503 target genes and their best *Brachypodium distachyon* BLASTp hit.



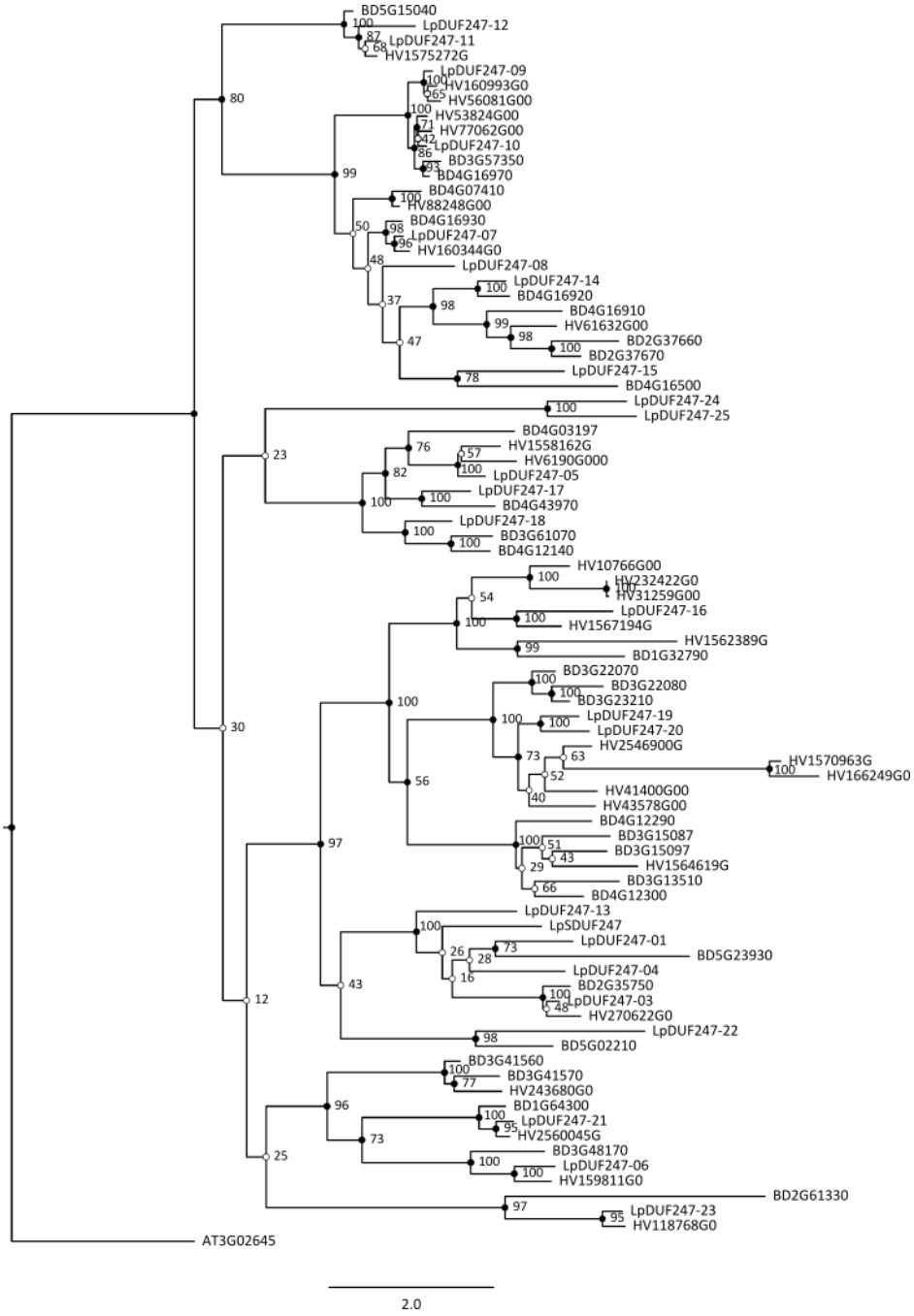
Supplemental Figure 3 Size and concordance of bi-allelic SNP and indel sets of four variant calling pipelines, before and after hard filtering. SNPs and indels were identified for 503 candidate genes in 736 genotypes using four VC pipelines: SAMtools, Freebayes, GATK UG and GATK HC (mapping with BWA-MEM). Concordance between the four variant sets was calculated for (a) bi-allelic SNPs and (b) bi-allelic indels. Per Upset plot, the lower left panel shows the total number of variants per VC pipeline; the lower right hand panel shows the overlap in call sets between the four VC pipelines. Concordance groups are ordered by increasing overlap, from left to right: variants unique to a VC pipeline; overlap of two and three VC pipelines; common to all four VC pipelines. The upper right hand panel shows the size per concordance group after integration of variant sets of the four VC pipelines using the VariantMetaCaller. Black bars: before hard filtering, orange bars: after hard filtering on minimal read depth 6 and genotype quality 30.



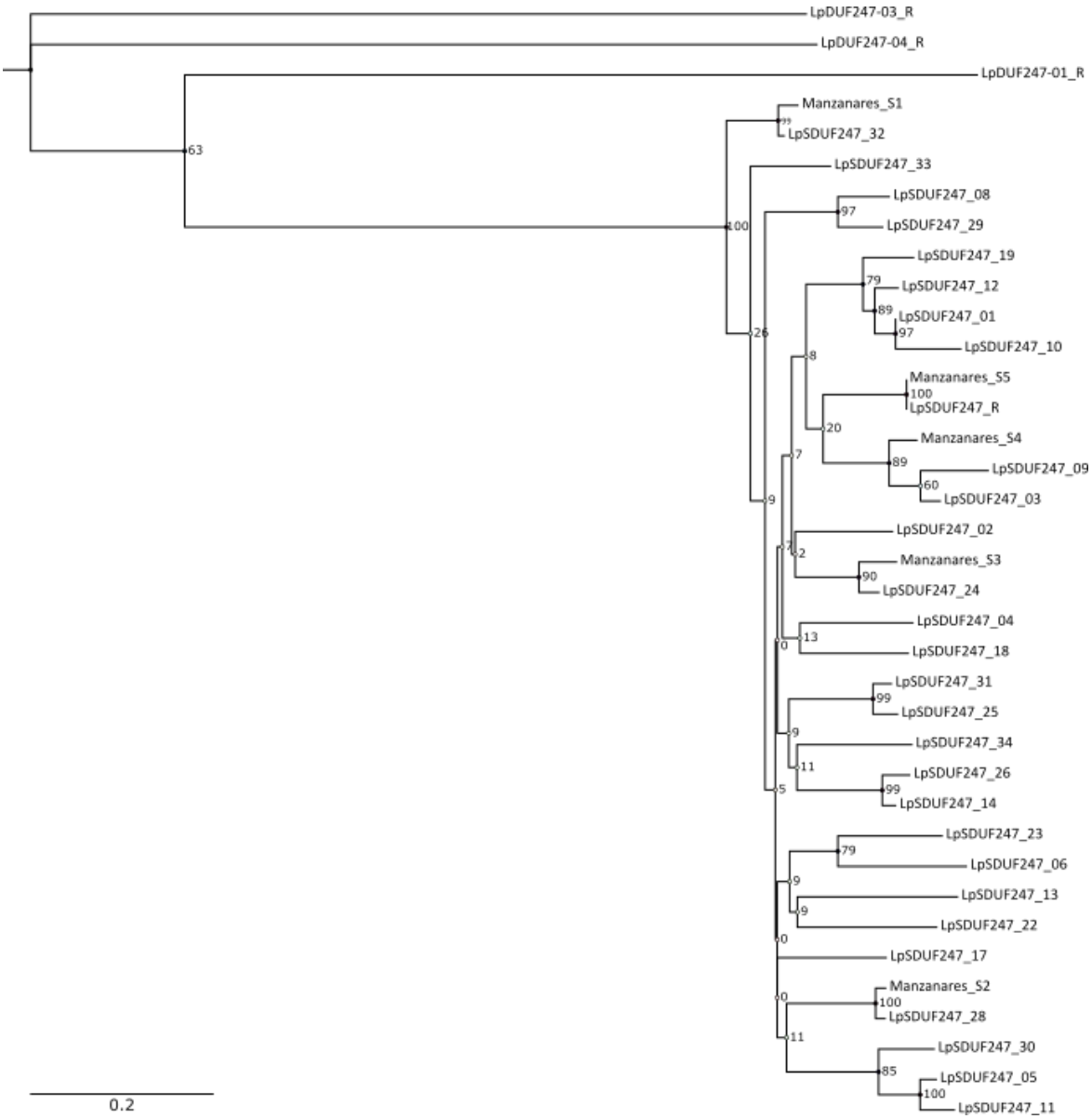
Supplemental Figure 4 Per-variant Mendelian inheritance error rate determined in an F1 progeny. Variants were identified using the VariantMetaCaller. After precision-based filtering (EP > 80%), Mendelian inheritance errors were identified for two parental individuals and a respective F1 progeny of 29 individuals.



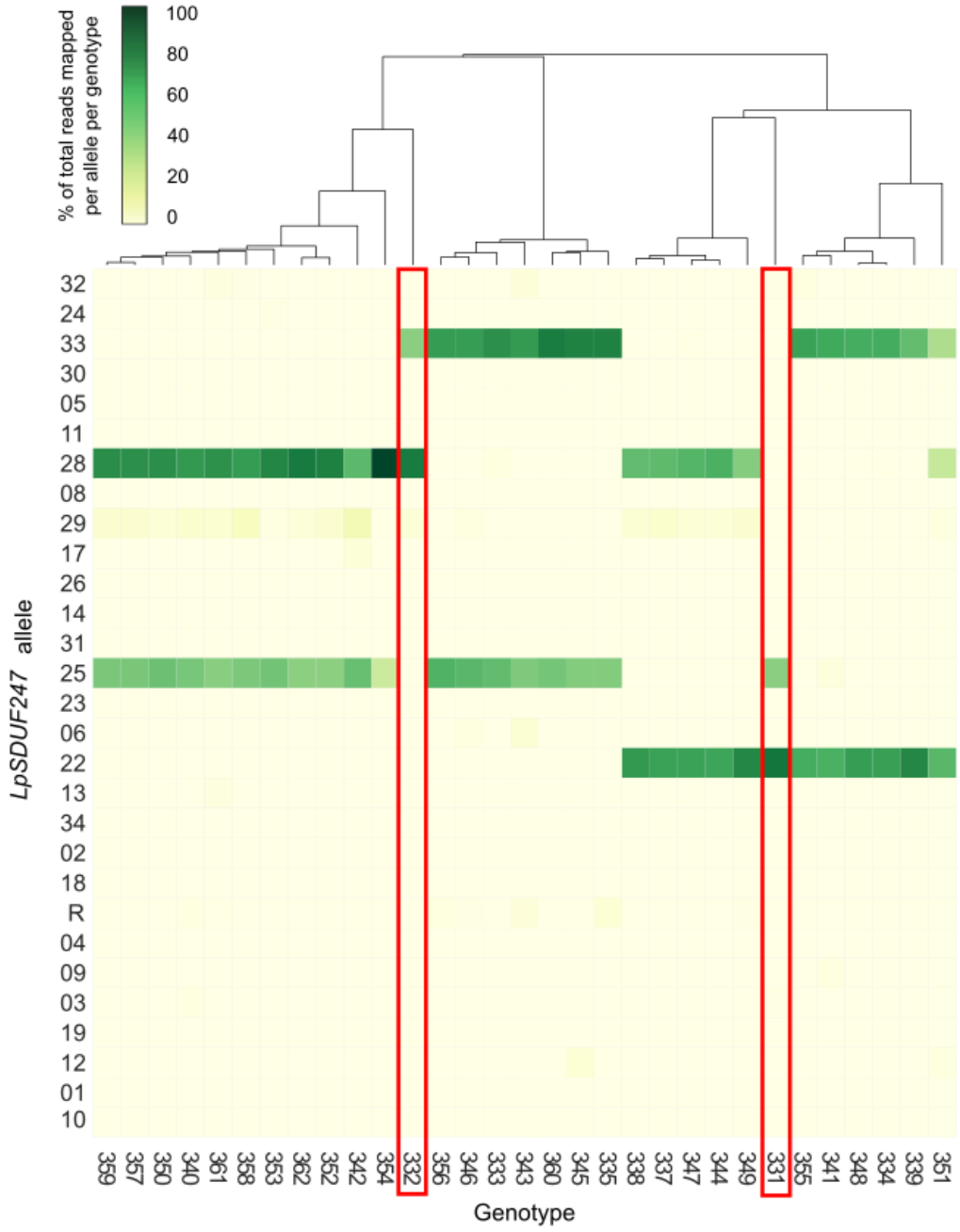
Supplemental Figure 5 Relative position and frequency of stop gain mutations identified in 503 candidate genes and 736 genotypes. Variants were identified using the VariantMetaCaller. After precision-based filtering (EP > 80%), effects were predicted with SnpEff (Cingolani et al., 2012). For each of the 256 stop gain mutations, the relative position of the stop gain mutation to the total gene length (x-axis) and the allele frequency (y-axis) were calculated. A boxplot of the allele frequencies is shown at the right side.



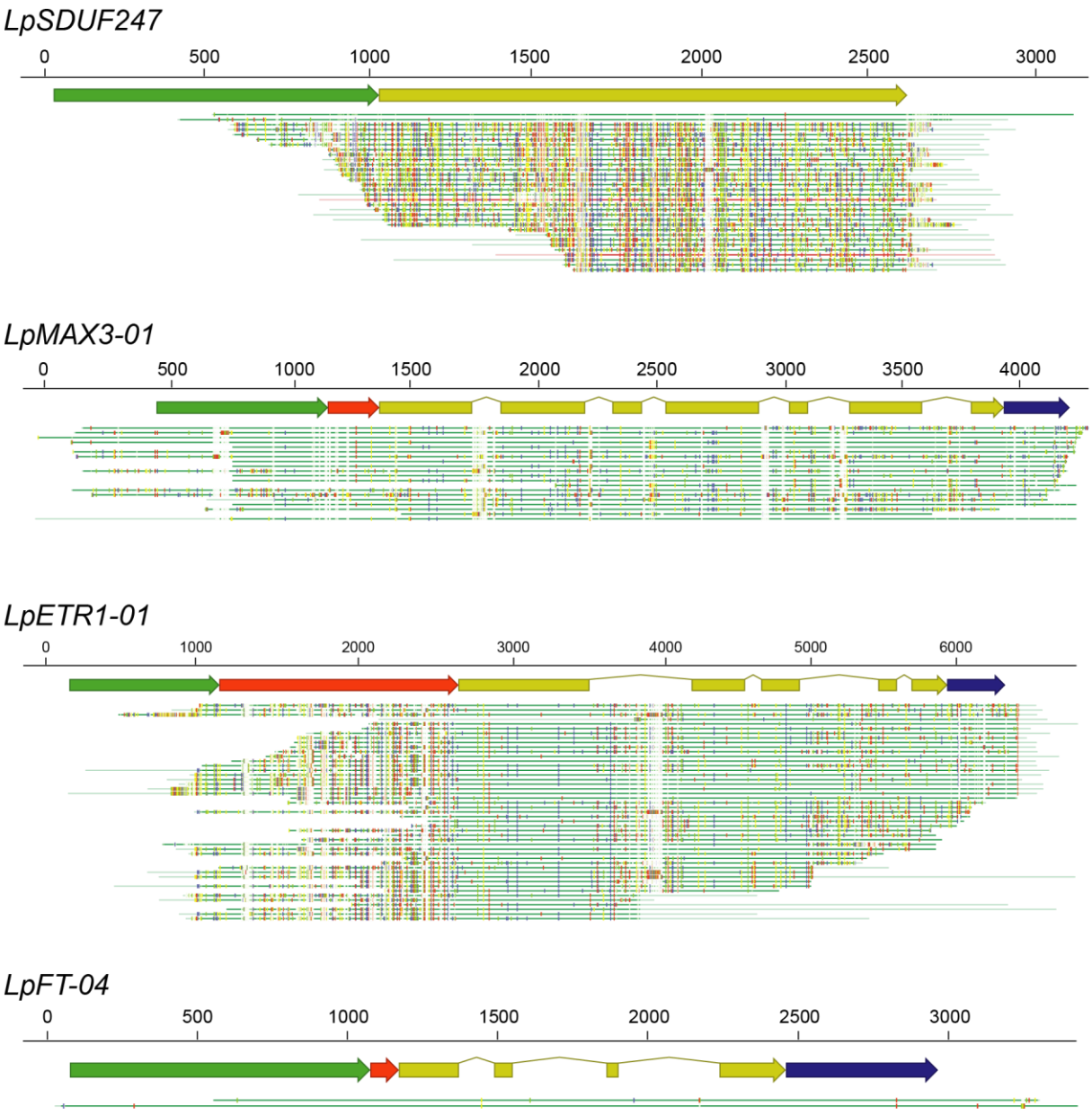
Supplemental Figure 6 Phylogenetic tree of DUF247 gene family members of *Brachypodium distachyon*, *Hordeum vulgare* and *Lolium perenne*. After multiple sequence alignment of all 25 LpDUF247 protein sequences and PLAZA HOM03M000101 gene family members of *B. distachyon* and *H. vulgare* with MUSCLE, a phylogenetic tree was built with PhyML using 100 rounds of bootstrapping. Bootstrapping values smaller and greater than 70 are indicated with empty and black nodes, respectively. An *A. thaliana* gene containing a DUF247 domain was used as outgroup.



Supplemental Figure 7 Phylogenetic tree of DUF247-01, -03 and -04 together with 5 LpSDUF247 alleles identified by (Manzanares et al., 2016) and 28 newly identified LpSDUF247 alleles. After multiple sequence alignment of the reference protein sequences of LpDUF247-01, LpSDUF247, LpDUF247-03 and LpDUF247-04 and the protein sequences of the newly identified alleles of LpSDUF247, complemented by the five LpSDUF247-02 alleles identified by Manzanares et al. (2016) a phylogenetic tree was built with PhyML using 100 rounds of bootstrapping. Bootstrapping values smaller and greater than 70 are indicated with empty and black nodes, respectively. Reference sequences of LpDUF247-01, LpSDUF247, LpDUF247-03 and LpDUF247-04 are indicated with R.



Supplemental Figure 8 Segregation of *LpSDUF247* alleles in an F1 progeny of 29 individuals. The alleles present per genotype were identified by mapping the reads to a multi-allelic reference genome, and calculating the ratio of average read depth per allele over the total number of reads mapping to *LpSDUF247* alleles. Parental genotypes are indicated by red rectangles.



Supplemental Figure 9 Four genes containing different levels of allelic diversity across the gene regions. The curated gene model is displayed for four genes, consisting of a promoter region (green), CDS (yellow), and 5' UTR (red) and 3'UTR (blue) if available. Below the gene model, CAP3 assembled contigs are aligned with CLCbio Genomics Workbench. Yellow, blue and red colors indicate sequence differences from the reference sequence.

C Bibliography

(2014). The 3,000 rice genomes project. GigaScience **3,** 7-7.

(2016). Computational pan-genomics: status, promises and challenges. Briefings in bioinformatics **19,** 118-135.

Ahn, J.H., Miller, D., Winter, V.J., Banfield, M.J., Lee, J.H., Yoo, S.Y., Henz, S.R., Brady, R.L., and Weigel, D. (2006). A divergent external loop confers antagonistic activity on floral regulators FT and TFL1. Embo J **25,** 605-614.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). APPLICATIONS OF NEXT-GENERATION SEQUENCING Genome structural variation discovery and genotyping. Nat Rev Genet **12,** 363-375.

Amos, W., Driscoll, E., and Hoffman, J.I. (2011). Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? Proceedings. Biological sciences **278,** 1183-1188.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Sharpe, A.G., Pozniak, C., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A.-A., Himmelbach, A., Balfourier, F., Gutierrez-Gonzalez, J., Hayden, M., Koh, C., Muehlbauer, G., Pasam, R.K., Paux, E., Rigault, P., Tibbits, J., Tiwari, V., Spannagl, M., Lang, D., Gundlach, H., Haberer, G., Mayer, K.F.X., Ormanbekova, D., Prade, V., Šimková, H., Wicker, T., Swarbreck, D., Rimbert, H., Felder, M., Guilhot, N., Kaithakottil, G., Keilwagen, J., Leroy, P., Lux, T., Twardziok, S., Venturini, L., Juhász, A., Abrouk, M., Fischer, I., Uauy, C., Borrill, P., Ramirez-Gonzalez, R.H., Arnaud, D., Chalabi, S., Chalhoub, B., Cory, A., Datla, R., Davey, M.W., Jacobs, J., Robinson, S.J., Steuernagel, B., van Ex, F., Wulff, B.B.H., Benhamed, M., Bendahmane, A., Concia, L., Latrasse, D., Alaux, M., Bartoš, J., Bellec, A., Berges, H., Doležel, J., Frenkel, Z., Gill, B., Korol, A., Letellier, T., Olsen, O.-A., Singh, K., Valárik, M., van der Vossen, E., Vautrin, S., Weining, S., Fahima, T., Glikson, V., Raats, D., Číhalíková, J., Toegelová, H., Vrána, J., Sourdille, P., Darrier, B., Barabaschi, D., Cattivelli, L., Hernandez, P., Galvez, S., Budak, H., Jones, J.D.G., Witek, K., Yu, G., Small, I., Melonek, J., Zhou, R., Belova, T., Kanyuka, K., King, R., Nilsen, K., Walkowiak, S., Cuthbert, R., Knox, R., Wiebe, K., Xiang, D., Rohde, A., Golds, T., Čížková, J., Akpinar, B.A., Biyiklioglu, S., Gao, L., N'Daiye, A., Kubaláková, M., Šafář, J., Alfama, F., Adam-Blondon, A.-F., Flores, R., Guerche, C., Loaec, M., Quesneville, H., Condie, J., Ens, J., Maclachlan, R., Tan, Y., Alberti, A., Aury, J.-M., Barbe, V., Couloux, A., Cruaud, C., Labadie, K., Mangenot, S., Wincker, P., Kaur, G., Luo, M., Sehgal, S., Chhuneja, P., Gupta, O.P., Jindal, S., Kaur, P., Malik, P., Sharma, P., Yadav, B., Singh, N.K., Khurana, J., Chaudhary, C., Khurana, P., Kumar, V., Mahato, A., Mathur, S., Sevanthi, A., Sharma, N., Tomar, R.S., Holušová, K., Plíhal, O., Clark, M.D., Heavens, D., Kettleborough, G., Wright, J., Balcárková, B., Hu, Y., Salina, E., Ravin, N., Skryabin, K., Beletsky, A., Kadnikov, V., Mardanov, A., Nesterov, M., Rakitin, A., Sergeeva, E., Handa, H., Kanamori, H., Katagiri, S., Kobayashi, F., Nasuda, S., Tanaka, T., Wu, J., Cattonaro, F., Jiumeng, M., Kugler, K., Pfeifer, M., Sandve, S., Xun, X., Zhan, B., Batley, J., Bayer, P.E., Edwards, D., Hayashi, S., Tulpová, Z., Visendi, P., Cui, L., Du, X., Feng, K., Nie, X., Tong, W., and Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**.

Armstead, I.P., Turner, L.B., Farrell, M., Skot, L., Gomez, P., Montoya, T., Donnison, I.S., King, I.P., and Humphreys, M.O. (2004). Synteny between a major heading-date QTL in perennial ryegrass (Lolium perenne L.) and the Hd3 heading-date locus in rice. Theor Appl Genet **108,** 822-828.

Arojju, S.K., Barth, S., Milbourne, D., Conaghan, P., Velmurugan, J., Hodkinson, T.R., and Byrne, S.L. (2016). Markers associated with heading and aftermath heading in perennial ryegrass full-sib families. Bmc Plant Biol **16**.

Asp, T., Byrne, S., Gundlach, H., Bruggmann, R., Mayer, K.F.X., Andersen, J.R., Xu, M., Greve, M., Lenk, I., and Lübberstedt, T. (2011). Comparative sequence analysis of VRN1 alleles of Lolium perenne with the co-linear regions in barley, wheat, and rice. Mol Genet Genomics **286,** 433-447.

Auzanneau, J., Huyghe, C., Julier, B., and Barre, P. (2007). Linkage disequilibrium in synthetic varieties of perennial ryegrass. Theor Appl Genet **115,** 837-847.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière, C., Owens, G.L., Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N.C., Bowers, J.E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., Boniface, M.-C., Brunel, D., Catrice, O., Chaidir, N., Claudel, C., Donnadieu, C., Faraut, T., Fievet, G., Helmstetter, N., King, M., Knapp, S.J., Lai, Z., Le Paslier, M.-C., Lippi, Y., Lorenzon, L., Mandel, J.R., Marage, G., Marchand, G., Marquand, E., Bret-Mestries, E., Morien, E., Nambeesan, S., Nguyen, T.,

Pegot-Espagnet, P., Pouilly, N., Raftis, F., Sallet, E., Schiex, T., Thomas, J., Vandecasteele, C., Varès, D., Vear, F., Vautrin, S., Crespi, M., Mangin, B., Burke, J.M., Salse, J., Muños, S., Vincourt, P., Rieseberg, L.H., and Langlade, N.B. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature **546,** 148.

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R.A., and Chen, M. (2016). The impact and origin of copy number variations in the Oryza species. Bmc Genomics **17,** 261.

Barre, P., Moreau, L., Mi, F., Turner, L., Gastal, F., Julier, B., and Ghesquiere, M. (2009). Quantitative trait loci for leaf length in perennial ryegrass (Lolium perenne L.). Grass Forage Sci **64,** 310-321.

Barre, P., Ruttink, T., Muylle, H., Lootens, P., Sampoux, J.P., Rohde, A., Combes, D., and Roldan-Ruiz, I. (2016). Natural diversity in vegetative and reproductive investments of perennial ryegrass is shaped by the climate at the place of origin. Grass Forage Sci **73,** 193-205.

Bassel, G.W., Gaudinier, A., Brady, S.M., Hennig, L., Rhee, S.Y., and De Smet, I. (2012). Systems Analysis of Plant Functional, Transcriptional, Physical Interaction, and Metabolic Networks. Plant Cell **24,** 3859-3875.

Bauer, E., Schmutzer, T., Barilar, I., Mascher, M., Gundlach, H., Martis, M.M., Twardziok, S.O., Hackauf, B., Gordillo, A., Wilde, P., Schmidt, M., Korzun, V., Mayer, K.F.X., Schmid, K., Schön, C.-C., and Scholz, U. (2017). Towards a whole-genome sequence for rye (Secale cereale L.). The Plant Journal **89,** 853-869.

Baumann, U., Juttner, J., Bian, X.Y., and Langridge, P. (2000). Self-incompatibility in the grasses. Ann Bot-London **85,** 203-209.

Bayer, M.M., Rapazote-Flores, P., Ganal, M., Hedley, P.E., Macaulay, M., Plieske, J., Ramsay, L., Russell, J., Shaw, P.D., Thomas, W., and Waugh, R. (2017). Development and Evaluation of a Barley 50k iSelect SNP Array. Front Plant Sci **8**.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. Methods **58,** 268-276.

Bennetzen, J.L., and Wang, H. (2014). The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. Annual Review of Plant Biology, Vol 65 **65,** 505-530.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., and van Nimwegen, E. (2014). Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. Mol Biol Evol **31,** 1077-1088.

Bertone, P., Trifonov, V., Rozowsky, J.S., Schubert, F., Emanuelsson, O., Karro, J., Kao, M.Y., Snyder, M., and Gerstein, M. (2006). Design optimization methods for genomic DNA tiling arrays. Genome Res **16,** 271-281.

Blackmore, T., Thomas, I., McMahon, R., Powell, W., and Hegarty, M. (2015). Genetic-geographic correlation revealed across a broad European ecotypic sample of perennial ryegrass (Lolium perenne) using array-based SNP genotyping. Theor Appl Genet **128,** 1917-1932.

Blackmore, T., Thorogood, D., Skøt, L., McMahon, R., Powell, W., and Hegarty, M. (2016). Germplasm dynamics: the role of ecotypic diversity in shaping the patterns of genetic variation in Lolium perenne. Sci Rep-Uk **6,** 22603.

Blanco-Pastor, J.L., Manel, S., Barre, P., Roschanski, A.M., Willner, E., Dehmer, K.J., Hegarty, M., Muylle, H., Ruttink, T., Roldan-Ruiz, I., Ledauphin, T., Escobar-Gutierrez, A., and Sampoux, J.-P. (2018). Pleistocene climate changes explain large-scale genetic variation in a dominant grassland species, Lolium perenne L. bioRxiv.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30,** 2114-2120.

Brazauskas, G., Pasakinskiene, I., Asp, T., and Lubberstedt, T. (2010). Nucleotide diversity and linkage disequilibrium in five Lolium perenne genes with putative role in shoot morphology. Plant Sci **179,** 194-201.

Brazauskas, G., Xing, Y., Studer, B., Schejbel, B., Frei, U., Berg, P.R., and Lübberstedt, T. (2013). Identification of genomic loci associated with crown rust resistance in perennial ryegrass (Lolium perenne L.) divergently selected populations. Plant Sci **208,** 34-41.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., Fan, L., Gao, S., Xu, X., Zhang, G., Li, Y., Jiao, Y., Doebley, J.F., Ross-Ibarra, J., Lorant, A., Buffalo, V., Romay, M.C., Buckler, E.S., Ware, D., Lai, J., Sun, Q., and Xu, Y. (2018). Construction of the third-generation Zea mays haplotype map. GigaScience **7,** gix134-gix134.

Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S.Y., Mulder, N., and Hunter, S. (2012). Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database-Oxford.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol **31,** 1119.

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., and Asp, T. (2013). Genome Wide Allele Frequency Fingerprints (GWAFFs) of Populations via Genotyping by Sequencing. Plos One **8,** e57438.

Byrne, S.L., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., Mayer, K., Campbell, J.D., Czaban, A., Hentrup, S., Panitz, F., Bendixen, C., Hedegaard, J., Caccamo, M., and Asp, T. (2015). A synteny-based draft genome sequence of the forage grass Lolium perenne. Plant J **84,** 816-826.

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.C., Liu, K.W., Chen, L.J., He, Y., Xu, Q., Bian, C., Zheng, Z.J., Sun, F.M., Liu, W.Q., Hsiao, Y.Y., Pan, Z.J., Hsu, C.C., Yang, Y.P., Hsu, Y.C., Chuang, Y.C., Dievart, A., Dufayard, J.F., Xu, X., Wang, J.Y., Wang, J., Xiao, X.J., Zhao, X.M., Du, R., Zhang, G.Q., Wang, M.N., Su, Y.Y., Xie, G.C., Liu, G.H., Li, L.Q., Huang, L.Q., Luo, Y.B., Chen, H.H., de Peer, Y.V., and Liu, Z.J. (2015). The genome sequence of the orchid Phalaenopsis equestris. Nat Genet **47,** 65-+.

Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Current Protocols in Bioinformatics **48,** 4.11. 11-14.11. 39.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18,** 188-196.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet **43,** 956-U960.

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., and DePristo, M.A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. Bmc Genomics **13**.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G.L., Akhunova, A., See, D., Bai, G., Pumphrey, M., Tomar, L., Wong, D., Kong, S., Reynolds, M., da Silva, M.L., Bockelman, H., Talbert, L., Anderson, J.A., Dreisigacker, S., Baenziger, S., Carter, A., Korzun, V., Morrell, P.L., Dubcovsky, J., Morell, M.K., Sorrells, M.E., Hayden, M.J., and Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proceedings of the National Academy of Sciences **110,** 8057-8062.

Chagne, D., Crowhurst, R.N., Pindo, M., Thrimawithana, A., Deng, C., Ireland, H., Fiers, M., Dzierzon, H., Cestaro, A., Fontana, P., Bianco, L., Lu, A., Storey, R., Knabel, M., Saeed, M., Montanari, S., Kim, Y.K., Nicolini, D., Larger, S., Stefani, E., Allan, A.C., Bowen, J., Harvey, I., Johnston, J., Malnoy, M., Troggio, M., Perchepied, L., Sawyer, G., Wiedow, C., Won, K., Viola, R., Hellens, R.P., Brewer, L., Bus, V.G.M., Schaffer, R.J., Gardiner, S.E., and Velasco, R. (2014). The Draft Genome Sequence of European Pear (Pyrus communis L. 'Bartlett'). Plos One **9**.

Chamala, S., Chanderbali, A.S., Der, J.P., Lan, T.Y., Walts, B., Albert, V.A., Depamphilis, C.W., Leebens-Mack, J., Rounsley, S., Schuster, S.C., Wing, R.A., Xiao, N.Q., Moore, R., Soltis, P.S., Soltis, D.E., and Barbazuk, W.B. (2013). Assembly and Validation of the Genome of the Nonmodel Basal Angiosperm Amborella. Science **342,** 1516-1517.

Chang, L.Y., Toghiani, S., Ling, A., Aggrey, S.E., and Rekaya, R. (2018). High density marker panels, SNPs prioritizing and accuracy of genomic selection. Bmc Genet **19,** 4.

Chardon, F., and Damerval, C. (2005). Phylogenomic analysis of the PEBP gene family in cereals. J Mol Evol **61,** 579-590.

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., Gore, M., Guill, K.E., Holland, J., Hufford, M.B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B.M., Pyhäjärvi, T., Rong, T., Sekhon, R.S., Sun, Q., Tenaillon, M.I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S.M., Ross-Ibarra, J., McMullen, M.D., Buckler, E.S., Zhang, G., Xu, Y., and Ware, D. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet **44,** 803.

Choi, J.Y., Platts, A.E., Fuller, D.Q., Hsing, Y.-I., Wing, R.A., and Purugganan, M.D. (2017). The Rice Paradox: Multiple Origins but Single Domestication in Asian Rice. Mol Biol Evol **34,** 969-979.

Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic DNA k-mer spectra: models and modalities. Genome Biol **10**.

Chung, Y.S., Choi, S.C., Jun, T.H., and Kim, C. (2017). Genotyping-by-Sequencing: a Promising Tool for Plant Genetics Research and Breeding. Hortic Environ Biote **58,** 425-431.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) **6,** 80-92.

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., Chen, H.M., Frazer, K.A., Huson, D.H., Schoelkopf, B., Nordborg, M., Raetsch, G., Ecker, J.R., and Weigel, D. (2007). Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science **317,** 338-342.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., and Varshney, R.K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends in Plant Science **22,** 961-975.

Cui, L.Y., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis, C.W. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res **16,** 738-749.

Curley, J., Sim, S.C., Jung, G., Leong, S., Warnke, S., and Barker, R.E. (2004). QTL Mapping of Gray Leaf Spot Resistance in Ryegrass, and Synteny-based Comparison with Rice Blast Resistance Genes in Rice (Dordrecht: Springer Netherlands), pp. 37-46.

Cutler, S., Ghassemian, M., Bonetta, D., Cooney, S., and McCourt, P. (1996). A protein farnesyl transferase involved in abscisic acid signal transduction in Arabidopsis. Science **273,** 1239-1241.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., and Grp, G.P.A. (2011). The variant call format and VCFtools. Bioinformatics **27,** 2156-2158.

Danilevskaya, O.N., Meng, X., Hou, Z.L., Ananiev, E.V., and Simmons, C.R. (2008). A genomic and expression compendium of the expanded PEBP gene family from maize. Plant Physiol **146,** 250-264.

Danyluk, J., Kane, N.A., Breton, G., Limin, A.E., Fowler, D.B., and Sarhan, F. (2003). TaVRT-1, a putative transcription factor associated with vegetative to reproductive transition in cereals. Plant Physiol **132,** 1849-1860.

De Bodt, S., Carvajal, D., Hollunder, J., Van den Cruyce, J., Movahedi, S., and Inze, D. (2010). CORNET: A User-Friendly Tool for Data Mining and Integration. Plant Physiol **152,** 1167-1179.

Dolezel, J., and Bartos, J.A.N. (2005). Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size. Ann Bot-London **95,** 99-110.

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J., Vlasova, A., Leskosek, B., Soler, L., Binzer-Panchal, M., and Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation [version 1; referees: 2 approved]. F1000Research **7**.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science **356,** 92-95.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D.W., Fass, J., Hung, O.K.Y., Buffalo, V., Zerbino, D.R., Diekhans, M., Nguyen, N., Ariyaratne, P.N., Sung, W.K., Ning, Z.M., Haimel, M., Simpson, J.T., Fonseca, N.A., Birol, I., Docking, T.R., Ho, I.Y., Rokhsar, D.S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M.C., Kelley, D.R., Phillippy, A.M., Koren, S., Yang, S.P., Wu, W., Chou, W.C., Srivastava, A., Shaw, T.I., Ruby, J.G., Skewes-Cox, P., Betegon, M., Dimon, M.T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F.F., Luo, R.B., Li, Z.Y., Xie, Y.L., Liu, B.H., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Yin, S.Y., Sharpe, T., Hall, G., Kersey, P.J., Durbin, R., Jackman, S.D., Chapman, J.A., Huang, X.Q., DeRisi, J.L., Caccamo, M., Li, Y.R., Jaffe, D.B., Green, R.E., Haussler, D., Korf, I., and Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res **21,** 2224-2241.

Ebbert, M.T.W., Wadsworth, M.E., Staley, L.A., Hoyt, K.L., Pickett, B., Miller, J., Duce, J., for the Alzheimer's Disease Neuroimaging, I., Kauwe, J.S.K., and Ridge, P.G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. Bmc Bioinformatics **17,** 239.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32,** 1792-1797.

Edwards, D., Batley, J., and Snowdon, R.J. (2013). Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet **126,** 1-11.

Farrell, J.D., Byrne, S., Paina, C., and Asp, T. (2014). De Novo Assembly of the Perennial Ryegrass Transcriptome Using an RNA-Seq Strategy. Plos One **9,** e103567.

Faure, S., Higgins, J., Turner, A., and Laurie, D.A. (2007). The FLOWERING LOCUS T-like gene family in barley (Hordeum vulgare). Genetics **176,** 599-609.

Faville, M.J., Ganesh, S., Moraga, R., Easton, H.S., Jahufer, M.Z.Z., Elshire, R.E., Asp, T., and Barrett, B.A. (2016). Development of Genomic Selection for Perennial Ryegrass (Cham: Springer International Publishing), pp. 139-143.

Fe, D., Ashraf, B.H., Pedersen, M.G., Janss, L., Byrne, S., Roulund, N., Lenk, I., Didion, T., Asp, T., Jensen, C.S., and Jensen, J. (2016). Accuracy of Genomic Prediction in a Commercial Perennial Ryegrass Breeding Program. Plant Genome-Us **9**.

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B.H., Pedersen, M.G., Roulund, N., Asp, T., Janss, L., Jensen, C.S., and Jensen, J. (2015). Genomic dissection and prediction of heading date in perennial ryegrass. Bmc Genomics **16,** 921.

Fiil, A., Lenk, I., Petersen, K., Jensen, C.S., Nielsen, K.K., Schejbel, B., Andersen, J.R., and Lubberstedt, T. (2011). Nucleotide diversity and linkage disequilibrium of nine genes with putative effects on flowering time in perennial ryegrass (Lolium perenne L.). Plant Sci **180,** 228-237.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol **183,** 557-564.

Flanagan, S.E., Patch, A.M., and Ellard, S. (2010). Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. Genet Test Mol Bioma **14,** 533-537.

Flot, J.F., Marie-Nelly, H., and Koszul, R. (2015). Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. Febs Lett **589,** 2966-2974.

Fu, D.L., Szucs, P., Yan, L.L., Helguera, M., Skinner, J.S., von Zitzewitz, J., Hayes, P.M., and Dubcovsky, J. (2005). Large deletions within the first intron in VRN-1 are associated with spring growth habit in barley and wheat. Mol Genet Genomics **273,** 54-65.

Gale, M.D., and Devos, K.M. (1998). Comparative genetics in the grasses. Proceedings of the National Academy of Sciences **95,** 1971-1974.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E.J., Harberd, N.P., Kemen, E., Toomajian, C., Kover, P.X., Clark, R.M., Rätsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature **477,** 419.

Garcia, S., Leitch, I.J., Anadon-Rosell, A., Canela, M.A., Galvez, F., Garnatje, T., Gras, A., Hidalgo, O., Johnston, E., de Xaxars, G.M., Pellicer, J., Siljak-Yakovlev, S., Valles, J., Vitales, D., and Bennett, M.D. (2014). Recent updates and developments to plant genome size databases. Nucleic Acids Res **42,** D1159-D1166.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Gézsi, A., Bolgár, B., Marx, P., Sarkozy, P., Szalai, C., and Antal, P. (2015). VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. Bmc Genomics **16,** 875.

Gianola, D., and van Kaam, J.B.C.H.M. (2008). Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics **178,** 2289-2303.

Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., Ware, D.H., and Buckler, E.S. (2009). A First-Generation Haplotype Map of Maize. Science **326,** 1115-1117.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. Inform Software Tech **47,** 965-978.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol **59,** 307-321.

Guo, X., Cericola, F., Fè, D., Pedersen, M.G., Lenk, I., Jensen, C.S., Jensen, J., and Janss, L.L. (2018). Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing. Front Plant Sci **9**.

Gyawali, S., Otte, M.L., Chao, S., Jilal, A., Jacob, D.L., Amezrou, R., and Verma, R.P.S. (2017). Genome wide association studies (GWAS) of element contents in grain with a special focus on zinc and iron in a world collection of barley (Hordeum vulgare L.). Journal of Cereal Science **77,** 266-274.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol **9**.

Hannaway, D., Fransen, S., Cropper, J., Teel, M., Chaney, M., Griggs, T., Halse, R., Hart, J., Cheeke, P., and Hansen, D. (1999). Perennial Ryegrass. Pacific Northwest Ext. Pub. PNW-503. Available: http://eesc. orst. edu/agcomwebfile/edmat/html/pnw/pnw503/complete. html. Accessed Feb **26,** 2003.

Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Manrique-Carpintero, N.C., Newton, L., Pham, G.M., Vaillancourt, B., Yang, X.M., Zeng, Z.X., Douches, D.S., Jiang, J.M., Veilleux, R.E., and Buell, C.R. (2016). Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. Plant Cell **28,** 388-405.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., Dewar, K., Stinchcombe, J.R., Schoen, D.J., Wang, X.W., Schmutz, J., Town, C.D., Edger, P.P., Pires, J.C., Schumaker, K.S., Jarvis, D.E., Mandakova, T., Lysak, M.A., van den Bergh, E., Schranz, M.E., Harrison, P.M., Moses, A.M., Bureau, T.E., Wright, S.I., and Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet **45,** 891-U228.

Hazard, L., Betin, M., and Molinari, N. (2006). Correlated response in plant height and heading date to selection in perennial ryegrass populations. Agron J **98,** 1384-1391.

Hegarty, M., Yadav, R., Lee, M., Armstead, I., Sanderson, R., Scollan, N., Powell, W., and Skøt, L. (2013). Genotyping by RAD sequencing enables mapping of fatty acid composition traits in perennial ryegrass (Lolium perenne (L.)). Plant Biotechnology Journal **11,** 572-581.

Helgadóttir, Á., Østrem, L., Collins, R., Humphreys, M., Marshall, A., Julier, B., Gastal, F., Barre, P., and Louarn, G. (2016). Breeding forages to cope with environmental challenges in the light of climate change and resource limitations. In Breeding in a World of Scarcity (Springer), pp. 3-13.

Hemming, M.N., Fieg, S., Peacock, W.J., Dennis, E.S., and Trevaskis, B. (2009). Regions associated with repression of the barley (Hordeum vulgare) VERNALIZATION1 gene are not required for cold induction. Mol Genet Genomics **282,** 107-117.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Penagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., de Leon, N., Kaeppler, S.M., and Buell, C.R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. Plant Cell **26,** 121-135.

Ho, W.W.H., and Weigel, D. (2014). Structural Features Determining Flower-Promoting Activity of Arabidopsis FLOWERING LOCUS T. Plant Cell **26,** 552-564.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics **32,** 767-769.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. Bmc Bioinformatics **12**.

Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y.T., Edger, P.P., Altman, N.S., Pires, J.C., Leebens-Mack, J.H., and dePamphilis, C.W. (2016). Selecting Superior De Novo Transcriptome Assemblies: Lessons Learned by Leveraging the Best Plant Genome. Plos One **11**.

Horst, G., Nelson, C., and Asay, K. (1978). Relationship of Leaf Elongation to Forage Yield of Tall Fescue Genotype 1. Crop Sci **18,** 715-719.

Huang, X., and Han, B. (2014). Natural Variations and Genome-Wide Association Studies in Crop Plants. Annu Rev Plant Biol **65,** 531-551.

Huang, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., and Li, M. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet **42,** 961.

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

Humphreys, M., Feuerstein, U., Vandewalle, M., and Baert, J. (2010). Ryegrasses. In Fodder Crops and Amenity Grasses, B. Boller, U.K. Posselt, and F. Veronesi, eds (New York, NY: Springer New York), pp. 211-260.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. (2013). REAPR: a universal tool for genome assembly evaluation. Genome Biol **14**.

Hwang, S., Kim, E., Lee, I., and Marcotte, E.M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep-Uk **5**.

International Rice Genome Sequencing, P., and Sasaki, T. (2005). The map-based sequence of the rice genome. Nature **436,** 793.

Jensen, C.S., Salchert, K., and Nielsen, K.K. (2001). A TERMINAL FLOMER1-Like gene from perennial ryegrass involved in floral transition and axillary meristem identity. Plant Physiol **125,** 1517-1528.

Jensen, C.S., Salchert, K., Gao, C.X., Andersen, C., Didion, T., and Nielsen, K.K. (2004). Floral inhibition in red fescue (Festuca rubra L.) through expression of a heterologous flowering repressor from Lolium. Mol Breeding **13,** 37-48.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K.F.X., Li, D., Pan, S., Zheng, F., Hu, Q., Xia, X., Li, J., Liang, Q., Chen, J., Wicker, T., Gou, C., Kuang, H., He, G., Luo, Y., Keller, B., Xia, Q., Lu, P., Wang, J., Zou, H., Zhang, R., Xu, J., Gao, J., Middleton, C., Quan, Z., Liu, G., Wang, J., International Wheat Genome Sequencing, C., Yang, H., Liu, X., He, Z., Mao, L., and Wang, J. (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature **496,** 91.

Jia, M., Guan, J., Zhai, Z., Geng, S., Zhang, X., Mao, L., and Li, A. (2018). Wheat functional genomics in the era of next generation sequencing: An update. The Crop Journal **6,** 7-14.

Jiao, W.-B., and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol **36,** 64-70.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K.L., Wolfgruber, T.K., May, M.R., Springer, N.M., Antoniou, E., McCombie, W.R., Presting, G.G., McMullen, M., Ross-Ibarra, J., Dawe, R.K., Hastie, A., Rank, D.R., and Ware, D. (2017). Improved maize reference genome with single-molecule technologies. Nature **546,** 524.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics **24,** 1403-1405.

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics **27,** 3070-3071.

Kang, Y.J., Satyawan, D., Shim, S., Lee, T., Lee, J., Hwang, W.J., Kim, S.K., Lestari, P., Laosatit, K., Kim, K.H., Ha, T.J., Chitikineni, A., Kim, M.Y., Ko, J.M., Gwag, J.G., Moon, J.K., Lee, Y.H., Park, B.S., Varshney, R.K., and Lee, S.H. (2015). Draft genome sequence of adzuki bean, Vigna angularis. Sci Rep-Uk **5**.

Kemp, D.R., Eagles, C.F., and Humphreys, M.O. (1989). Leaf Growth and Apex Development of Perennial Ryegrass During Winter and Spring. Ann Bot-London **63,** 349-355.

King, R.W., Moritz, T., Evans, L.T., Martin, J., Andersen, C.H., Blundell, C., Kardailsky, I., and Chandler, P.M. (2006). Regulation of flowering in the long-day grass Lolium temulentum by Gibberellins and the FLOWERING LOCUS T gene. Plant Physiol **141,** 498-507.

Kopecký, D., and Studer, B. (2014). Emerging technologies advancing forage and turf grass genomics. Biotechnology Advances **32,** 190-199.

Korf, I. (2004). Gene finding in novel genomes. Bmc Bioinformatics **5**.

Laidlaw, A.S. (2004). Effect of heading date of perennial ryegrass cultivars on tillering and tiller development in spring and summer. Grass Forage Sci **59,** 240-249.

Laidlaw, A.S. (2005). The relationship between tiller appearance in spring and contribution to dry-matter yield in perennial ryegrass (Lolium perenne L.) cultivars differing in heading date. Grass Forage Sci **60,** 200-209.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. Ieee T Vis Comput Gr **20,** 1983-1992.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics **30,** 2843-2851.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25,** 2078-2079.

Li, J.Y., Wang, J., and Zeigler, R.S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience **3**.

Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., Gao, C., Wu, H., Li, Y., Cui, Y., Guo, X., Zheng, S., Wang, B., Yu, K., Liang, Q., Yang, W., Lou, X., Chen, J., Feng, M., Jian, J., Zhang, X., Luo, G., Jiang, Y., Liu, J., Wang, Z., Sha, Y., Zhang, B., Wu, H., Tang, D., Shen, Q., Xue, P., Zou, S., Wang, X., Liu, X., Wang, F., Yang, Y., An, X., Dong, Z., Zhang, K., Zhang, X., Luo, M.-C., Dvorak, J., Tong, Y., Wang, J., Yang, H., Li, Z., Wang, D., Zhang, A., and Wang, J. (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. Nature **496,** 87.

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.-Z. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. Plos One **8,** e75619.

Lorenz, A.J., Hamblin, M.T., and Jannink, J.-L. (2010). Performance of Single Nucleotide Polymorphisms versus Haplotypes for Genome-Wide Association Analysis in Barley. Plos One **5,** e14079.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. P Natl Acad Sci USA **102,** 5454-5459.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. J Genet **92,** 155-161.

Manzanares, C., Barth, S., Thorogood, D., Byrne, S.L., Yates, S., Czaban, A., Asp, T., Yang, B.C., and Studer, B. (2016). A Gene Encoding a DUF247 Domain Protein Cosegregates with the S Self-Incompatibility Locus in Perennial Ryegrass. Mol Biol Evol **33,** 870-884.

Mapleson, D., Venturini, L., Kaithakottil, G., and Swarbreck, D. (2017). Efficient and accurate detection of splice junctions from RNAseq with Portcullis. bioRxiv**,** 217620.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol **18,** 31-36.

Marroni, F., Pinosio, S., Di Centa, E., Jurman, I., Boerjan, W., Felice, N., Cattonaro, F., and Morgante, M. (2011). Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation Ecotilling. Plant J **67,** 736-745.

Mascher, M., Muehlbauer, G.J., Rokhsar, D.S., Chapman, J., Schmutz, J., Barry, K., Munoz-Amatriain, M., Close, T.J., Wise, R.P., Schulman, A.H., Himmelbach, A., Mayer, K.F.X., Scholz, U., Poland, J.A., Stein, N., and Waugh, R. (2013). Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J **76,** 718-727.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., Bayer, M., Ramsay, L., Liu, H., Haberer, G., Zhang, X.-Q., Zhang, Q., Barrero, R.A., Li, L., Taudien, S., Groth, M., Felder, M., Hastie, A., Šimková, H., Staňková, H., Vrána, J., Chan, S., Muñoz-Amatriaín, M., Ounit, R., Wanamaker, S., Bolser, D., Colmsee, C., Schmutzer, T., Aliyeva-Schnorr, L., Grasso, S., Tanskanen, J., Chailyan, A., Sampath, D., Heavens, D., Clissold, L., Cao, S., Chapman, B., Dai, F., Han, Y., Li, H., Li, X., Lin, C., McCooke, J.K., Tan, C., Wang, P., Wang, S., Yin, S., Zhou, G., Poland, J.A., Bellgard, M.I., Borisjuk, L., Houben, A., Doležel, J., Ayling, S., Lonardi, S., Kersey, P., Langridge, P., Muehlbauer, G.J., Clark, M.D., Caccamo, M., Schulman, A.H., Mayer, K.F.X., Platzer, M., Close, T.J., Scholz, U., Hansson, M., Zhang, G., Braumann, I., Spannagl, M., Li, C., Waugh, R., and Stein, N. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature **544,** 427.

Mayer, K.F.X., Taudien, S., Martis, M., Šimková, H., Suchánková, P., Gundlach, H., Wicker, T., Petzold, A., Felder, M., Steuernagel, B., Scholz, U., Graner, A., Platzer, M., Doležel, J., and Stein, N. (2009). Gene Content and Virtual Gene Order of Barley Chromosome 1H. Plant Physiol **151,** 496-505.

Mayer, K.F.X., Rogers, J., Dolezel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J., Sourdille, P., Endo, T.R., Dolezel, J., Kubalakova, M., Cihalikova, J., Dubska, Z., Vrana, J., Sperkova, R., Simkova, H., Rogers, J., Febrer, M., Clissold, L., McLay, K., Singh, K., Chhuneja, P., Singh, N.K., Khurana, J., Akhunov, E., Choulet, F., Sourdille, P., Feuillet, C., Alberti, A., Barbe, V., Wincker, P., Kanamori, H., Kobayashi, F., Itoh, T., Matsumoto, T., Sakai, H., Tanaka, T., Wu, J.Z., Ogihara, Y., Handa, H., Pozniak, C., Maclachlan, P.R., Sharpe, A., Klassen, D., Edwards, D., Batley, J., Olsen, O.A., Sandve, S.R., Lien, S., Steuernagel, B., Wulff, B., Caccamo, M., Ayling, S., Ramirez-Gonzalez, R.H., Clavijo, B.J., Steuernagel, B.,

Wright, J., Pfeifer, M., Spannagl, M., Mayer, K.F.X., Martis, M.M., Akhunov, E., Choulet, F., Mayer, K.F.X., Mascher, M., Chapman, J., Poland, J.A., Scholz, U., Barry, K., Waugh, R., Rokhsar, D.S., Muehlbauer, G.J., Stein, N., Gundlach, H., Zytnicki, M., Jamilloux, V., Quesneville, H., Wicker, T., Mayer, K.F.X., Faccioli, P., Colaiacovo, M., Pfeifer, M., Stanca, A.M., Budak, H., Cattivelli, L., Glover, N., Martis, M.M., Choulet, F., Feuillet, C., Mayer, K.F.X., Pfeifer, M., Pingault, L., Mayer, K.F.X., Paux, E., Spannagl, M., Sharma, S., Mayer, K.F.X., Pozniak, C., Appels, R., Bellgard, M., Chapman, B., Pfeifer, M., Pfeifer, M., Sandve, S.R., Nussbaumer, T., Bader, K.C., Choulet, F., Feuillet, C., Mayer, K.F.X., Akhunov, E., Paux, E., Rimbert, H., Wang, S.C., Poland, J.A., Knox, R., Kilian, A., Pozniak, C., Alaux, M., Alfama, F., Couderc, L., Jamilloux, V., Guilhot, N., Viseux, C., Loaec, M., Quesneville, H., Rogers, J., Dolezel, J., Eversole, K., Feuillet, C., Keller, B., Mayer, K.F.X., Olsen, O.A., Praud, S., and Iwgsc. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science **345**.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet **9,** 356.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20,** 1297-1303.

McNally, K.L., Childs, K.L., Bohnert, R., Davidson, R.M., Zhao, K., Ulat, V.J., Zeller, G., Clark, R.M., Hoen, D.R., Bureau, T.E., Stokowski, R., Ballinger, D.G., Frazer, K.A., Cox, D.R., Padhukasahasram, B., Bustamante, C.D., Weigel, D., Mackill, D.J., Bruskiewich, R.M., Ratsch, G., Buell, C.R., Leung, H., and Leach, J.E. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. P Natl Acad Sci USA **106,** 12273-12278.

Mendelowitz, L., and Pop, M. (2014). Computational methods for optical mapping. Gigascience **3**.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics **157,** 1819-1829.

Michael, T.P., and Jackson, S. (2013). The First 50 Plant Genomes. Plant Genome-Us **6**.

Ming, R., VanBuren, R., Liu, Y.L., Yang, M., Han, Y.P., Li, L.T., Zhang, Q., Kim, M.J., Schatz, M.C., Campbell, M., Li, J.P., Bowers, J.E., Tang, H.B., Lyons, E., Ferguson, A.A., Narzisi, G., Nelson, D.R., Blaby-Haas, C.E., Gschwend, A.R., Jiao, Y.N., Der, J.P., Zeng, F.C., Han, J., Min, X.J., Hudson, K.A., Singh, R., Grennan, A.K., Karpowicz, S.J., Watling, J.R., Ito, K., Robinson, S.A., Hudson, M.E., Yu, Q.Y., Mockler, T.C., Carroll, A., Zheng, Y., Sunkar, R., Jia, R.Z., Chen, N., Arro, J., Wai, C.M., Wafula, E., Spence, A., Han, Y.N., Xu, L.M., Zhang, J.S., Peery, R., Haus, M.J., Xiong, W.W., Walsh, J.A., Wu, J., Wang, M.L., Zhu, Y.J., Paull, R.E., Britt, A.B., Du, C.G., Downie, S.R., Schuler, M.A., Michael, T.P., Long, S.P., Ort, D.R., Schopf, J.W., Gang, D.R., Jiang, N., Yandell, M., dePamphilis, C.W., Merchant, S.S., Paterson, A.H., Buchanan, B.B., Li, S.H., and Shen-Miller, J. (2013). Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome Biol **14**.

Mishra, P., and Panigrahi, K.C. (2015). GIGANTEA – an emerging story. Front Plant Sci **6,** 8.

Moghe, G.D., Hufnagel, D.E., Tang, H.B., Xiao, Y.L., Dworkin, I., Town, C.D., Conner, J.K., and Shiu, S.H. (2014). Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. Plant Cell **26,** 1925-1937.

Mollison, E.M.B., Barth, S., Milbourne, D., Milne, L., Halpin, C., McCabe, M., Creevey, C., and Marshall, D.F. (2016). De novo Genome Sequencing and Gene Prediction in Lolium perenne, Perennial Ryegrass. In Breeding in a World of Scarcity, I. Roldán-Ruiz, J. Baert, and D. Reheul, eds (Cham: Springer International Publishing), pp. 127-131.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.K., Visendi, P., Lai, K., Dolezel, J., Batley, J., and Edwards, D. (2017). The pangenome of hexaploid bread wheat. Plant J **90,** 1007-1013.

Moore, G., Devos, K.M., Wang, Z., and Gale, M.D. (1995). Cereal Genome Evolution - Grasses, Line up and Form a Circle. Curr Biol **5,** 737-739.

Murray, M.G., and Thompson, W.F. (1980). Rapid Isolation of High Molecular-Weight Plant DNA. Nucleic Acids Res **8,** 4321-4325.

Nguyen-Dumont, T., Pope, B.J., Hammet, F., Mahmoodi, M., Tsimiklis, H., Southey, M.C., and Park, D.J. (2013). Cross-platform compatibility of Hi-Plex, a streamlined approach for targeted massively parallel sequencing. Anal Biochem **442,** 127-129.

Nordborg, M., and Weigel, D. (2008). Next-generation genetics in plants. Nature **456,** 720.

Nowak, M.D., Russo, G., Schlapbach, R., Huu, C.N., Lenhard, M., and Conti, E. (2015). The draft genome of Primula veris yields insights into the molecular basis of heterostyly. Genome Biol **16**.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K., and Lyon, G.J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med **5,** 28.

Oddes, S., Zelig, A., and Kaplan, N. (2018). Three invariant Hi-C interaction patterns: Applications to genome assembly. Methods **142,** 89-99.

Oliver, S.N., Finnegan, E.J., Dennis, E.S., Peacock, W.J., and Trevaskis, B. (2009). Vernalization-induced flowering in cereals is associated with changes in histone methylation at the VERNALIZATION1 gene. P Natl Acad Sci USA **106,** 8386-8391.

Olsen, J.L., Rouze, P., Verhelst, B., Lin, Y.C., Bayer, T., Collen, J., Dattolo, E., De Paoli, E., Dittami, S., Maumus, F., Michel, G., Kersting, A., Lauritano, C., Lohaus, R., Topel, M., Tonon, T., Vanneste, K., Amirebrahimi, M., Brakel, J., Bostrom, C., Chovatia, M., Grimwood, J., Jenkins, J.W., Jueterbock, A., Mraz, A., Stam, W.T., Tice, H., Bornberg-Bauer, E., Green, P.J., Pearson, G.A., Procaccini, G., Duarte, C.M., Schmutz, J., Reusch, T.B.H., and Van de Peer, Y. (2016). The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. Nature **530,** 331-+.

Paina, C., Byrne, S.L., Domnisoru, C., and Asp, T. (2014). Vernalization Mediated Changes in the Lolium perenne Transcriptome. Plos One **9,** e107365.

Park, M.-H., Rhee, H., Park, J.H., Woo, H.-M., Choi, B.-O., Kim, B.-Y., Chung, K.W., Cho, Y.-B., Kim, H.J., Jung, J.-W., and Koo, S.K. (2014). Comprehensive Analysis to Improve the Validation Rate for Single Nucleotide Variants Detected by Next-Generation Sequencing. Plos One **9,** e86664.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. Bioinformatics **23,** 1061-1067.

Parween, S., Nawaz, K., Roy, R., Pole, A.K., Suresh, B.V., Misra, G., Jain, M., Yadav, G., Parida, S.K., Tyagi, A.K., Bhatia, S., and Chattopadhyay, D. (2015). An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). Sci Rep-Uk **5**.

Pembleton, L.W., Inch, C., Baillie, R.C., Drayton, M.C., Thakur, P., Ogaji, Y.O., Spangenberg, G.C., Forster, J.W., Daetwyler, H.D., and Cogan, N.O.I. (2018). Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik **131,** 1891-1902.

Petersen, K., Didion, T., Andersen, C.H., and Nielsen, K.K. (2004). MADS-box genes from perennial ryegrass differentially expressed during transition from vegetative to reproductive growth. J Plant Physiol **161,** 439-447.

Pfeifer, M., Martis, M., Asp, T., Mayer, K.F.X., Lubberstedt, T., Byrne, S., Frei, U., and Studer, B. (2013). The Perennial Ryegrass GenomeZipper: Targeted Use of Genome Resources for Comparative Grass Genomics. Plant Physiol **161,** 571-582.

Phillippy, A.M., Schatz, M.C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. Genome Biol **9,** R55.

Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R., and Zandi, P.P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics **8,** 14.

Poland, J.A., Brown, P.J., Sorrells, M.E., and Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. Plos One **7,** e32253.

Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y., and Vandepoele, K. (2009). PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. Plant Cell **21,** 3718-3731.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inze, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res **43,** D974-D981.

Qian, L., Hickey, L.T., Stahl, A., Werner, C.R., Hayes, B., Snowdon, R.J., and Voss-Fels, K.P. (2017). Exploring and Harnessing Haplotype Diversity to Improve Yield Stability in Crops. Front Plant Sci **8,** 1534.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. Bmc Genomics **13,** 341-341.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26,** 841-842.

Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R., and Martienssen, R.A. (1999). Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet **23,** 305-308.

Rahman, A., and Pachter, L. (2013). CGAL: computing genome assembly likelihoods. Genome Biol **14**.

Rhee, S.Y., and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. Trends in Plant Science **19,** 212-221.

Roldán-Ruiz, I., and Kölliker, R. (2010). Marker-assisted selection in forage crops and turf: a review. In Sustainable use of genetic diversity in forage and turf breeding (Springer), pp. 383-390.

Ruttink, T., Sterck, L., Rohde, A., Bendixen, C., Rouze, P., Asp, T., Van de Peer, Y., and Roldan-Ruiz, I. (2013). Orthology Guided Assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in Lolium perenne. Plant Biotechnol J **11,** 605-617.

Ruttink, T., Haegeman, A., van Parijs, F., Van Glabeke, S., Muylle, H., Byrne, S., Asp, T., and Roldán-Ruiz, I. (2015). Genetic Diversity in Candidate Genes for Developmental Traits and Cell Wall Characteristics in Perennial Ryegrass (Lolium perenne). In Molecular Breeding of Forage and Turf: The Proceedings of the 8th International Symposium on the Molecular Breeding of Forage and Turf, H. Budak and G. Spangenberg, eds (Cham: Springer International Publishing), pp. 93-109.

Sahebi, M., Hanafi, M.M., van Wijnen, A.J., Rice, D., Raffi, M.Y., Azizi, P., Osman, M., Taheri, S., Abu Bakar, M.F., Isa, M.N.M., and Noor, Y.M. (2018). Contribution of transposable elements in the plant's genome. Gene **665,** 155-166.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marcais, G., Pop, M., and Yorke, J.A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res **22,** 557-567.

Sampoux, J.-P., Métral, R., Ghesquière, M., Baudouin, P., Bayle, B., Béguier, V., Bourdon, P., Chosson, J.-F., de Bruijn, K., Deneufbourg, F., Galbrun, C., Pietraszek, W., Tharel, B., and Viguié, A. (2010). Genetic Improvement in Ryegrass (Lolium perenne) from Turf and Forage Breeding Over the Four Past Decades (Dordrecht: Springer Netherlands), pp. 325-330.

Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biol **13,** 243.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.-T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting,

G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326,** 1112-1115.

Schwartz, D.C., Li, X.J., Hernandez, L.I., Ramnarain, S.P., Huff, E.J., and Wang, Y.K. (1993). Ordered Restriction Maps of Saccharomyces-Cerevisiae Chromosomes Constructed by Optical Mapping. Science **262,** 110-114.

Shinozuka, H., Cogan, N.O.I., Spangenberg, G.C., and Forster, J.W. (2012). Quantitative Trait Locus (QTL) meta-analysis and comparative genomics for candidate gene prediction in perennial ryegrass (Lolium perenne L.). Bmc Genet **13**.

Shinozuka, H., Cogan, N.O.I., Spangenberg, G.C., and Forster, J.W. (2017). Reference transcriptome assembly and annotation for perennial ryegrass. Genome **60,** 1086-1088.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31,** 3210-3212.

Skot, L., Humphreys, J., Humphreys, M.O., Thorogood, D., Gallagher, J., Sanderson, R., Armstead, I.P., and Thomas, I.D. (2007). Association of candidate genes with flowering time and water-soluble carbohydrate content in Lolium perenne (L.). Genetics **177,** 535-547.

Skot, L., Sanderson, R., Thomas, A., Skot, K., Thorogood, D., Latypova, G., Asp, T., and Armstead, I. (2011). Allelic Variation in the Perennial Ryegrass FLOWERING LOCUS T Gene Is Associated with Changes in Flowering Time across a Range of Populations. Plant Physiol **155,** 1013-1022.

Skot, L., Humphreys, M.O., Armstead, I., Heywood, S., Skot, K.P., Sanderson, R., Thomas, I.D., Chorlton, K.H., and Hamilton, N.R.S. (2005). An association mapping approach to identify flowering time genes in natural populations of Lolium perenne (L.). Mol Breeding **15,** 233-245.

Slotte, T., Hazzouri, K.M., Agren, J.A., Koenig, D., Maumus, F., Guo, Y.L., Steige, K., Platts, A.E., Escobar, J.S., Newman, L.K., Wang, W., Mandakova, T., Vello, E., Smith, L.M., Henz, S.R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., Andolfatto, P., Hu, T.T., Blanchette, M., Clark, R.M., Quesneville, H., Nordborg, M., Gaut, B.S., Lysak, M.A., Jenkins, J., Grimwood, J., Chapman, J., Prochnik, S., Shu, S.Q., Rokhsar, D., Schmutz, J., Weigel, D., and Wright, S.I. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet **45,** 831-U165.

Smith, K.F., Dobrowolski, M.P., Cogan, N.O.I., Spangenberg, G.C., and Forster, J.W. (2009). Utilizing Linkage Disequilibrium and Association Mapping to Implement Candidate Gene Based Markers in Perennial Ryegrass Breeding (New York, NY: Springer New York), pp. 259-274.

Song, K., Li, L., and Zhang, G. (2016). Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. Sci Rep-Uk **6,** 35736.

Stram, D.O., and Seshan, V.E. (2012). Multi-SNP haplotype analysis methods for association analysis. Methods Mol Biol **850,** 423-452.

Studer, B., Byrne, S., Nielsen, R.O., Panitz, F., Bendixen, C., Islam, M.S., Pfeifer, M., Lubberstedt, T., and Asp, T. (2012). A transcriptome map of perennial ryegrass (Lolium perenne L.). Bmc Genomics **13,** 140.

Sweeney, M., and McCouch, S. (2007). The complex history of the domestication of rice. Ann Bot-London **100,** 951-957.

Tadesse, W., Ogbonnaya, F.C., Jighly, A., Sanchez-Garcia, M., Sohail, Q., Rajaram, S., and Baum, M. (2015). Genome-Wide Association Mapping of Yield and Grain Quality Traits in Winter Wheat Genotypes. Plos One **10,** e0141339.

Tang, H., Lyons, E., Pedersen, B., Schnable, J.C., Paterson, A.H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. Bmc Bioinformatics **12,** 102.

Tang, H.B., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A.N., Zhou, S.G., Gentzbittel, L., Childs, K.L., Yandell, M., Gundlach, H., Mayer, K.F.X., Schwartz, D.C., and Town, C.D. (2014). An improved genome release (version Mt4.0) for the model legume Medicago truncatula. Bmc Genomics **15**.

Tang, Y., Liu, X.L., Wang, J.B., Li, M., Wang, Q.S., Tian, F., Su, Z.B., Pan, Y.C., Liu, D., Lipka, A.E., Buckler, E.S., and Zhang, Z.W. (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. Plant Genome-Us **9**.

The International Barley Genome Sequencing, C., Mayer, K.F.X., Waugh, R., Langridge, P., Close, T.J., Wise, R.P., Graner, A., Matsumoto, T., Sato, K., Schulman, A., Muehlbauer, G.J., Stein, N., Ariyadasa, R., Schulte, D., Poursarebani, N., Zhou, R., Steuernagel, B., Mascher, M., Scholz, U., Shi, B., Langridge, P., Madishetty, K., Svensson, J.T., Bhat, P., Moscou, M., Resnik, J., Close, T.J., Muehlbauer, G.J., Hedley, P., Liu, H., Morris, J., Waugh, R., Frenkel, Z., Korol, A., Bergès, H., Graner, A., Stein, N., Steuernagel, B., Scholz, U., Taudien, S.,

Felder, M., Groth, M., Platzer, M., Stein, N., Steuernagel, B., Scholz, U., Himmelbach, A., Taudien, S., Felder, M., Platzer, M., Lonardi, S., Duma, D., Alpert, M., Cordero, F., Beccuti, M., Ciardo, G., Ma, Y., Wanamaker, S., Close, T.J., Stein, N., Cattonaro, F., Vendramin, V., Scalabrin, S., Radovic, S., Wing, R., Schulte, D., Steuernagel, B., Morgante, M., Stein, N., Waugh, R., Nussbaumer, T., Gundlach, H., Martis, M., Ariyadasa, R., Poursarebani, N., Steuernagel, B., Scholz, U., Wise, R.P., Poland, J., Stein, N., Mayer, K.F.X., Spannagl, M., Pfeifer, M., Gundlach, H., Mayer, K.F.X., Gundlach, H., Moisy, C., Tanskanen, J., Scalabrin, S., Zuccolo, A., Vendramin, V., Morgante, M., Mayer, K.F.X., Schulman, A., Pfeifer, M., Spannagl, M., Hedley, P., Morris, J., Russell, J., Druka, A., Marshall, D., Bayer, M., Swarbreck, D., Sampath, D., Ayling, S., Febrer, M., Caccamo, M., Matsumoto, T., Tanaka, T., Sato, K., Wise, R.P., Close, T.J., Wannamaker, S., Muehlbauer, G.J., Stein, N., Mayer, K.F.X., Waugh, R., Steuernagel, B., Schmutzer, T., Mascher, M., Scholz, U., Taudien, S., Platzer, M., Sato, K., Marshall, D., Bayer, M., Waugh, R., Stein, N., Mayer, K.F.X., Waugh, R., Brown, J.W.S., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., Wise, R.P., and Stein, N. (2012). A physical, genetic and functional sequence assembly of the barley genome. Nature **491,** 711.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler Iv, E.S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet **28,** 286.

Tian, S.L., Yan, H.H., Neuhauser, C., and Slager, S.L. (2016). An analytical workflow for accurate variant discovery in highly divergent regions. Bmc Genomics **17,** 703.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25,** 1105-1111.

Trevaskis, B., Bagnall, D.J., Ellis, M.H., Peacock, W.J., and Dennis, E.S. (2003). MADS box genes control vernalization-induced flowering in cereals. P Natl Acad Sci USA **100,** 13099-13104.

Udall, J.A., and Dawe, R.K. (2018). Is It Ordered Correctly? Validating Genome Assemblies by Optical Mapping. Plant Cell **30,** 7-14.

Uitdewilligen, J.G.A.M.L., Wolters, A.-M.A., D'hoop, B.B., Borm, T.J.A., Visser, R.G.F., and van Eck, H.J. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. Plos One **8,** e62355.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3-new capabilities and interfaces. Nucleic Acids Res **40,** e115-e115.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., and Vandepoele, K. (2012). Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. Plant Physiol **158,** 590-600.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res **46,** D1190-D1196.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends in Plant Science **14,** 680-688.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., and DePristo, M.A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics **43,** 11 10 11-33.

Van Landeghem, S. (2014). AnnoMine: Creating a consensus homology-based functional annotation through text mining (GitHub repository).

van Parijs, F.R.D. (2016). Cell Wall Digestibility of Perennial Ryegrass : an Association Mapping Approach. In Faculty of Bioscience Engineering (Ghent, Belgium: Ghent University), pp. XVI, 295 pages.

van Parijs, F.R.D., Ruttink, T., Haesaert, G., Roldán-Ruiz, I., and Muylle, H. (2016). Association Mapping of LpCCR1 with Lignin Content and Cell Wall Digestibility of Perennial Ryegrass (Cham: Springer International Publishing), pp. 219-224.

Veeckman, E., Ruttink, T., and Vandepoele, K. (2016). Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. Plant Cell **28,** 1759-1768.

Veeckman, E., Van Glabeke, S., Haegeman, A., Muylle, H., van Parijs, F.R.D., Byrne, S.L., Asp, T., Studer, B., Rohde, A., Roldán-Ruiz, I., Vandepoele, K., and Ruttink, T. (2018). Overcoming challenges in variant calling:

exploring sequence diversity in candidate genes for plant development in perennial ryegrass (Lolium perenne). DNA Res, dsy033-dsy033.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagne, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.E., Gutin, A., Bumgarner, R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., and Viola, R. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). Nat Genet **42,** 833-+.

Venturini, L., Caim, S., Kaithakottil, G., Mapleson, D.L., and Swarbreck, D. (2017). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. bioRxiv.

Wang, W., Zheng, H.K., Fan, C.Z., Li, J., Shi, J.J., Cai, Z.Q., Zhang, G.J., Liu, D.Y., Zhang, J.G., Vang, S., Lu, Z.K., Wong, G.K.S., Long, M.Y., and Wang, J. (2006). High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell **18,** 1791-1802.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K.C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., Zhao, X., Shen, F., Cui, X., Yu, H., Li, Z., Chen, M., Detras, J., Zhou, Y., Zhang, X., Zhao, Y., Kudrna, D., Wang, C., Li, R., Jia, B., Lu, J., He, X., Dong, Z., Xu, J., Li, Y., Wang, M., Shi, J., Li, J., Zhang, D., Lee, S., Hu, W., Poliakov, A., Dubchak, I., Ulat, V.J., Borja, F.N., Mendoza, J.R., Ali, J., Li, J., Gao, Q., Niu, Y., Yue, Z., Naredo, M.E.B., Talag, J., Wang, X., Li, J., Fang, X., Yin, Y., Glaszmann, J.-C., Zhang, J., Li, J., Hamilton, R.S., Wing, R.A., Ruan, J., Zhang, G., Wei, C., Alexandrov, N., McNally, K.L., Li, Z., and Leung, H. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature **557,** 43-49.

Weigel, D., and Mott, R. (2009a). The 1001 Genomes Project for Arabidopsis thaliana. Genome Biology **10**.

Weigel, D., and Mott, R. (2009b). The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol **10,** 107.

Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K.F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research **6,** 100-100.

Wendel, J.F., Jackson, S.A., Meyers, B.C., and Wing, R.A. (2016). Evolution of plant genome architecture. Genome Biol **17**.

Whitt, S.R., Wilson, L.M., Tenaillon, M.I., Gaut, B.S., and Buckler, E.S.t. (2002). Genetic diversity and selection in the maize starch pathway. P Natl Acad Sci USA **99,** 12959-12962.

Wickland, D.P., and Hanzawa, Y. (2015). The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms. Mol Plant **8,** 983-997.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21,** 1859-1875.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26,** 873-881.

Xing, Y., Frei, U., Schejbel, B., Asp, T., and Lubberstedt, T. (2007). Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in Lolium perenne. Bmc Plant Biol **7,** 43.

Xu, Q., Krishnan, S., Merewitz, E., Xu, J., and Huang, B. (2016). Gibberellin-Regulation and Genetic Variations in Leaf Elongation for Tall Fescue in Association with Differential Gene Expression Controlling Cell Expansion. Sci Rep-Uk **6,** 30258.

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., and Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene VRN1. P Natl Acad Sci USA **100,** 6263-6268.

Yang, Q., Li, Z., Li, W., Ku, L., Wang, C., Ye, J., Li, K., Yang, N., Li, Y., Zhong, T., Li, J., Chen, Y., Yan, J., Yang, X., and Xu, M. (2013). CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci U S A **110,** 16969-16974.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.C., Hu, L., Yamasaki, M., Yoshida, S., Kitano, H., Hirano, K., and Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet **48,** 927-934.

Yao, W., Li, G.W., Zhao, H., Wang, G.W., Lian, X.M., and Xie, W.B. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol **16**.

Yoo, S.Y., Kardailsky, I., Lee, J.S., Weigel, D., and Ahn, J.H. (2004). Acceleration of flowering by overexpression of MFT (MOTHER OF FT AND TFL1). Mol Cells **17,** 95-101.

Yu, G., Sauchyn, D., and Li, Y.F. (2013). Drought changes and the mechanism analysis for the North American Prairie. J Arid Land **5,** 1-14.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. (2002). A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). Science **296,** 79-92.

Yu, X.Q., and Sun, S.Y. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. Bmc Bioinformatics **14,** 274.

Yu, X.Q., Pijut, P.M., Byrne, S., Asp, T., Bai, G.H., and Jiang, Y.W. (2015). Candidate gene association mapping for winter survival and spring regrowth in perennial ryegrass. Plant Sci **235,** 37-45.

Yuan, Y.N., SanMiguel, P.J., and Bennetzen, J.L. (2003). High-Cot sequence analysis of the maize genome. Plant J **34,** 249-255.

Zhang, G.Y., Liu, X., Quan, Z.W., Cheng, S.F., Xu, X., Pan, S.K., Xie, M., Zeng, P., Yue, Z., Wang, W.L., Tao, Y., Bian, C., Han, C.L., Xia, Q.J., Peng, X.H., Cao, R., Yang, X.H., Zhan, D.L., Hu, J.C., Zhang, Y.X., Li, H.N., Li, H., Li, N., Wang, J.Y., Wang, C.C., Wang, R.Y., Guo, T., Cai, Y.J., Liu, C.Z., Xiang, H.T., Shi, Q.X., Huang, P., Chen, Q.C., Li, Y.R., Wang, J., Zhao, Z.H., and Wang, J. (2012). Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nat Biotechnol **30,** 549-+.

Zhang, J.J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics **30,** 614-620.

Zhou, S.G., Wei, F.S., Nguyen, J., Bechner, M., Potamousis, K., Goldstein, S., Pape, L., Mehan, M.R., Churas, C., Pasternak, S., Forrest, D.K., Wise, R., Ware, D., Wing, R.A., Waterman, M.S., Livny, M., and Schwartz, D.C. (2009). A Single Molecule Scaffold for the Maize Genome. Plos Genet **5**.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. (2008). Status and Prospects of Association Mapping in Plants. The Plant Genome **1,** 5-20.

Zmienko, A., Samelak, A., Kozlowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. Theor Appl Genet **127,** 1-18.

Zonneveld, B.J.M., Leitch, I.J., and Bennett, M.D. (2005). First nuclear DNA amounts in more than 300 angiosperms. Ann Bot-London **96,** 229-244.

Zou, C., Chen, A., Xiao, L., Muller, H.M., Ache, P., Haberer, G., Zhang, M., Jia, W., Deng, P., Huang, R., Lang, D., Li, F., Zhan, D., Wu, X., Zhang, H., Bohm, J., Liu, R., Shabala, S., Hedrich, R., Zhu, J.-K., and Zhang, H. (2017). A high-quality genome assembly of quinoa provides insights into the molecular basis of salt bladder-based salinity tolerance and the exceptional nutritional value. Cell Res **27,** 1327.