Published July 28, 2016

ORIGINAL RESEARCH

Orthology Guided Transcriptome Assembly of Italian Ryegrass and Meadow Fescue for Single-Nucleotide Polymorphism Discovery

Štěpán Stočes, Tom Ruttink, Jan Bartoš, Bruno Studer, Steven Yates, Zbigniew Zwierzykowski, Michael Abrouk, Isabel Roldán-Ruiz, Tomasz Książczyk, Elodie Rey, Jaroslav Doležel, and David Kopecký*

Abstract

Single-nucleotide polymorphisms (SNPs) represent natural DNA sequence variation. They can be used for various applications including the construction of high-density genetic maps, analysis of genetic variability, genome-wide association studies, and map-based cloning. Here we report on transcriptome sequencing in the two forage grasses, meadow fescue (*Festuca pratensis* Huds.) and Italian ryegrass (*Lolium multiflorum* Lam.), and identification of various classes of SNPs. Using the Orthology Guided Assembly (OGA) strategy, we assembled and annotated a total of 18,952 and 19,036 transcripts for Italian ryegrass and meadow fescue, respectively. In addition, we used transcriptome sequence data of perennial ryegrass (*L. perenne* L.) from a previous study to identify 16,613 transcripts shared across all three species. Large numbers of intraspecific SNPs were identified in all three species: 248,000 in meadow fescue, 715,000 in Italian ryegrass, and 529,000 in perennial ryegrass. Moreover, we identified almost 25,000 interspecific SNPs located in 5343 genes that can distinguish meadow fescue from Italian ryegrass and 15,000 SNPs located in 3976 genes that discriminate meadow fescue from both *Lolium* species. All identified SNPs were positioned in silico on the seven linkage groups (LGs) of *L. perenne* using the GenomeZipper approach. With the identification and positioning of interspecific SNPs, our study provides a valuable resource for the grass research and breeding community and will enable detailed characterization of genomic composition and gene expression analysis in prospective *Festuca* × *Lolium* hybrids.

Core Ideas

- Transcriptomes of *F. pratensis* and *L. multiflorum* were sequenced and assembled
- We present a catalogue of SNPs for ancestry analysis and future breeding of grasses
- We defined interspecific SNPs to study parental-genome specific gene expression in hybrids
- We positioned SNPs on linkage groups for high-resolution genome constitution analysis

GRASSES of the genera *Festuca* and *Lolium* belong to the Pooideae subfamily within the grass family (Poaceae) and are widely cultivated in temperate regions. They include species such as perennial ryegrass, Italian ryegrass, meadow fescue, and tall fescue (*F. arundinacea* Schreb.). Perennial ryegrass is a highly tillering species

Š. Stočes, J. Bartoš, M. Abrouk, E. Rey, J. Doležel, and D. Kopecký, Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 78371 Olomouc, Czech Republic; T. Ruttink and I. Roldán-Ruiz, Institute for Agricultural and Fisheries Research (ILVO), Plant Sciences Unit–Growth and Development, Caritasstraat 39, 9090 Melle, Belgium; B. Studer, and S. Yates, Institute of Agricultural Sciences, Forage Crop Genetics, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland; Z. Zwierzykowski and T. Książczyk, Dep. of Environmental Stress Biology, Institute of Plant Genetics of the Polish Academy of Sciences, Strzeszyńska 34, 60479 Poznań, Poland. Received 16 Feb. 2016. Accepted 6 June 2016. *Corresponding author (kopecky@ueb.cas.cz).

Abbreviations: AD, minimal allele count score; CDS, coding DNA sequence; DBG, De Bruijn graph; DP, total read depth; GQ, genotype quality score; HQ, high quality; HRM, high-resolution melting curve analysis; LG, linkage group; OGA, Orthology Guided Assembly; PCR, polymerase chain reaction; PGAS, parental genome allele specific; QTL, quantitative trait loci; RNA-seq, RNA sequencing; SNP, single-nucleotide polymorphism.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2016.02.0017

suitable for permanent pastures and meadows and is a dominant species for turf applications in Europe. Italian ryegrass is used for hay and silage production in temporary grasslands. Both ryegrasses are highly nutritious and high seed yielding (reviewed in Humphreys et al., 2010). However, they are sensitive to abiotic stress conditions. Meadow fescue is widely grown as a pasture crop in Northern temperate regions (Ergon et al., 2006) because of its good persistence and high tolerance to environmental stresses such as cold, freezing, and drought. Tall fescue is an important cool-season perennial forage and turf grass species with excellent drought tolerance (Rognli et al., 2010).

Intergeneric hybrids between *Festuca* and *Lolium* species, referred to as Festuloliums, combine the beneficial agronomic traits from both genera (Ghesquiere et al., 2010). During the last decades, Festuloliums have become a valuable source of new grasses for cultivation (Yamada et al., 2005). They are highly productive, nutritious, and resilient and, as such, are widely used for agricultural and amenity purposes (reviewed in Ghesquiere et al., 2010). Festuloliums exhibit significant diversity for many interesting traits, mainly as a result of their variable genome constitution, which differs markedly even among plants of the same cultivar (Kopecký et al., 2011). Identification of parental chromosomes and characterization of genomic constitution of these interspecific hybrids is typically done using cytogenetic methods such as fluorescence and genomic in situ hybridization (FISH and GISH) (Zwierzykowski et al., 1998a,b; Canter et al., 1999; Kopecký et al., 2005, 2006; Książczyk et al. 2015). However, these methods only allow the detection of large chromosomal segments and are not applicable for high-throughput screening (Kopecký et al., 2006).

In contrast to cytogenetic approaches, DNA sequence polymorphisms offer the opportunity to determine genomic constitution of Festuloliums at much higher resolution. Various DNA marker types distinguishing *Lolium* and *Festuca* genomes have been reported such as microsatellites (Pašakinskiene et al., 2000; Studer et al., 2006) and Diversity Array Technology markers (Kopecký et al., 2009). However, SNPs are the most prevalent type of polymorphisms and thus informative on account of their relatively even dispersal across the genome and their high density especially in allogamous species such as fescues and ryegrasses (Birrer et al., 2014; Wang et al., 2014; Czaban et al., 2015). Single-nucleotide polymorphisms within one species (here referred to as intraspecific SNPs) can be used for quantitative trait loci (QTL) mapping, map-based cloning, genome-wide association studies, and marker-assisted or genomic selection. Single-nucleotide polymorphisms distinguishing the progenitor species (here referred to as interspecific SNPs) can be used to study genomic constitution of interspecific hybrids such as Festuloliums or hybrid ryegrasses with a high resolution. Moreover, if identified in transcribed genome regions, interspecific SNPs can also be used to profile parental genome allele specific (PGAS) expression levels in subsequent generations of interspecific hybrids. Thus, the underlying causes of the superior phenotype could be better understood and perhaps used to improve *Festuca*, *Lolium*, and Festulolium cultivars.

RNA sequencing (RNA-seq, Mortazavi et al., 2008) is suitable to identify SNPs in the gene space of a given species, particularly in the absence of a complete genome sequence (Denoeud et al., 2008; Varshney et al., 2009; Garvin et al., 2010). While a draft set of scaffolds of the *L. perenne* genome has recently been published (Byrne et al., 2015), there is no high-quality genome sequence available, which could serve as a reference for read mapping and identification of polymorphisms in Italian ryegrass and meadow fescue, the two species used in this study.

Considering the available genomic resources, we set out to create a reference transcriptome for Italian ryegrass and meadow fescue based on RNA-seq on a small panel of genotypes. Illumina sequencing technology has been widely used for transcriptome sequencing because of its high throughput and low error rate (Gonçalves da Silva et al., 2014; Peng et al., 2014; Wu et al., 2014). This technology yields short reads (up to 300 bp), which can be used to de novo reconstruct the transcriptome. Several algorithms exist for de novo assembly of short Illumina reads based on De Bruijn graphs (DBG), including Trans-ABySS (Simpson et al., 2009; Robertson et al., 2010), Trinity (Grabherr et al., 2011), Velvet (Zerbino and Birney 2008), Oases (Schulz et al., 2012), and CLCBio Genomics Workbench (CLC Bio-Qiagen). However, a high degree of heterozygosity, typical for outcrossing species such as fescues and ryegrasses (Modrek and Lee 2002; Ruttink et al., 2013; Farrell et al., 2014), can result in transcript fragmentation and allelic redundancy in contigs that are assembled using DBG algorithms. This problem can be resolved by a second round of clustering with overlap–layout–consensus assemblers such as CAP3 (Huang and Madan, 1999). A quick and targeted strategy to perform such secondary clustering on a gene-by-gene basis for an entire transcriptome is known as Orthology Guided Assembly (OGA; Ruttink et al., 2013). The OGA strategy uses the proteome of a closely related species to guide the clustering procedure and to generate a nonredundant and annotated consensus sequence per gene (Ruttink et al., 2013).

The purpose of the experiments presented here was to identify a large and robust set of SNP markers useful to determine genomic constitution of Italian ryegrass × meadow fescue hybrids. By focusing on the transcriptome, we circumvented the lack of a reference genome sequence from the two species. The newly developed set of SNPs will also be useful to quantify PGAS expression levels in interspecific hybrids. In this paper we report on (i) the reconstruction of reference transcriptomes for meadow fescue and Italian ryegrass and their comparison with the previously published perennial ryegrass transcriptome assembled with the same OGA strategy; (ii) the identification of transcript-anchored intraspecific and interspecific SNPs in the three species; (iii) the identification of SNPs that discriminate parental genome

specific alleles in pairwise crosses between meadow fescue and Italian ryegrass; and (iv) the identification of putative genomic locations of the SNPs, based on the perennial ryegrass GenomeZipper (Pfeifer et al., 2013). This large-scale marker set we have obtained constitutes an important resource to advance genomic research in meadow fescue and Italian ryegrass, two of the most important forage grass species, and their hybrids.

Materials and Methods

Plant Material and RNA Extraction

Six genotypes of Italian ryegrass (Lm) were used for transcriptome sequencing: two genotypes of the tetraploid cultivar 'Mitos' (Lm51 and Lm52) and one genotype of each of the diploid Italian ryegrass cultivars 'Abercomo' (Lm53), 'Fox' (Lm54), 'Prolog' (Lm55) and 'Sikem' (Lm56). Six meadow fescue (Fp) genotypes were used for transcriptome sequencing: two genotypes of the tetraploid cultivar 'Westa' (Fp52 and Fp53), one genotype of the tetraploid cultivar 'Patra' (Fp54), and one genotype of each of the diploid cultivars 'Skawa' (Fp51), 'Fure' (Fp55), and 'WSC' (Fp56). The plants were grown in a growth chamber (Weiss-Gallenkamp) at 50% relative humidity (day, 14 h of 20,000 lux light intensity at 24°C; night, 10 h at 20°C). After establishment, the plants were cut to a height of ~5 cm above ground level to stimulate tillering. After 3 wk of regrowth, 100 mg of young leaves were harvested and frozen in liquid nitrogen. Three independent samples were collected from the genotypes Lm51, Lm52, Fp52, and Fp53 and one sample from each of the remaining genotypes. Total RNA was extracted from ground samples using the RNeasy Plant Mini Kit (Qiagen, Inc.). The quality, quantity, and integrity of RNA were checked using an RNA Pico 6000 chip on a Bioanalyzer 2100 (Agilent Technologies, Inc.). Samples with RNA integrity number >7 were used for RNA sequencing.

RNA Sequencing Library Construction and Illumina Sequencing

High-throughput sequencing of Italian ryegrass and meadow fescue libraries was performed at Istituto di Genomica Applicata (IGA Technology Services, Udine, Italy). RNA sequencing libraries were generated using the TruSeq RNA Sample Prep kit v2 according to the manufacturer's protocol (Illumina Inc.). Adapters were ligated to the complementary DNA and 200 ± 25 bp fragments were gel purified and amplified by polymerase chain reaction (PCR). DNA libraries were quantified using a Bioanalyzer 2100 (Agilent Technologies), pooled in equimolar amounts for sequencing using Illumina HiSeq2000 or HiSeq2500 instruments to produce paired-end 101 bp reads.

De Novo Transcriptome Assembly

The FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) was used to trim the 5-bp index and to discard reads with a Phred quality score <25. Reads of each sample were then used for independent de novo assembly using CLCBio Genomics Workbench v8.4. Only contigs longer than 200 bp were retained. Scaffolding was switched off to avoid long stretches of Ns in the contigs, which would interfere with the construction of a consensus sequence by CAP3 during downstream OGA assembly. The total number of reads, number of contigs, and N50 length are listed in Supplemental Table S1.

Orthology Guided Assembly

To resolve transcript fragmentation and allelic redundancy, an OGA strategy was performed combining the contigs from each of the 10 independent de novo assemblies per species. Briefly, we used 26,632 *Brachypodium distachyon* (L.) Beauv. proteins (The International Brachypodium Initiative, 2010) as a reference to assemble consensus transcript sequence as described previously (Ruttink et al., 2013). Transcripts matching particular *B. distachyon* proteins were assembled using CAP3. We then trimmed the transcripts to the coding DNA sequence (CDS), determined the remaining redundancy per *B. distachyon* protein, selected the longest representative sequence per *B. distachyon* protein (minimal CDS length of 102 bp), and thus created novel, nonredundant reference transcriptome assemblies for meadow fescue and Italian ryegrass. All transcripts were assigned to orthologous protein families of the PLAZA 3.0 Monocots database (Proost et al., 2014) as an entry point for full comparative genomics and functional annotation. The previously assembled reference transcriptome from perennial ryegrass (Ruttink et al., 2013) was included in the subsequent steps to compare the effect of using a reference transcriptome from a closely related species on read mapping and SNP calling.

For more accurate and cross-species analyses, we also identified near full-length transcripts (length of the transcript is between 85 and 115% of the respective orthologous *B. distachyon* CDS). An overview per orthologous *B. distachyon* protein is available in Supplemental Table S2. The OGA assembled transcriptome sequences are available as FASTA file and GFF with annotations in Zenodo (http://dx.doi.org/10.5281/zenodo.55304).

Read Mapping and Single-Nucleotide Polymorphism Calling

Here, we describe the analysis procedure for a single reference transcriptome. However, all analyses were performed separately on the two other reference transcriptomes in parallel. Quality trimmed RNA-seq reads of all genotypes were aligned against one of the three nonredundant reference transcriptomes using BWA (Li and Durbin, 2009). After local realignment with GATK IndelRealigner, duplicated reads were removed with MarkDuplicates. In parallel to the new reads obtained from meadow fescue and Italian ryegrass, the previously published RNA-seq reads of 14 genotypes of perennial rygrass (Supplemental Table S1) were mapped onto the reference and all BAM files were used for combined SNP analyses

across the three species. The SNPs were called using the UnifiedGenotypeCaller of GATK (McKenna et al., 2010; DePristo et al., 2011). The identified genotype calls per sample were further filtered using custom Python scripts. Only biallelic SNPs with genotype quality score GQ $\geq$ 30 and minimal allele count AD $\geq$ 4 were retained. We used different total read depth for tetraploid (DP $\geq$ 16) and for diploid (DP $\geq$ 10) genotypes as a trade-off between assuring a high probability of a correct genotype call and retrieving enough data for a similar number of genes given varying total library size per sample. For some perennial ryegrass samples, the library size was much lower than the average library size of the novel obtained data from meadow fescue and Italian ryegrass. Although missing data in some of the 14 perennial ryegrass genotypes may lead to some underestimation of the number of SNPs (see Fig. 4C of Ruttink et al., 2013), we required SNP call data in at least 10 out of 14 perennial ryegrass genotypes during interspecific SNP identification. With these settings, over 7000 genes exceeded the threshold number of genotypes required to call interspecific SNP in each of the species (six in Fp, six in Lm, and 10 in Lp).

For each of the tetraploid genotypes Fp52, Fp53, Lm51, and Lm52, we tested the robustness of SNP calling by comparing the genotype calls across their three biological replicates. Genotype calls present in at least two of three replicates were compared after quality filtering (GQ $\geq$ 30, DP $\geq$ 16, AD $\geq$ 4). If all genotype calls per SNP were consistent (two identical calls and one missing call or three identical calls), these were merged into a single genotype call per sample and used for all downstream analyses. Otherwise, the genotype call for that SNP was labeled as missing data.

Classification of Intraspecific and Interspecific Single-Nucleotide Polymorphisms

All analyses were performed in parallel on the three nonredundant reference transcriptomes of meadow fescue, Italian ryegrass, or perennial ryegrass, respectively. For each reference transcriptome, SNP calls from the different meadow fescue, Italian ryegrass, or perennial ryegrass genotypes were first classified as homozygous reference allele, heterozygous, homozygous alternative allele, or missing data. Next, for each SNP, the following criteria were used to classify it as one of following three classes: (i) intraspecific SNPs (INTRA), (ii) interspecific SNPs in two-way comparison (INTER-2W), and (iii) interspecific SNPs in three-way comparison (INTER-3W) (See Fig. 1 for the classification scheme).

(i) If two or more alleles of a given SNP were detected in any number of genotypes within a species (i.e., at least one heterozygous genotype or at least one genotype each for the homozygous reference allele and the homozygous alternative allele), the SNP position was classified as an intraspecific SNP and indexed for that species (e.g., $INTRA_{Fp}$).

(ii) If a SNP position was monomorphic in all genotypes of one species and carried monomorphic alternative allele in all genotypes of other species, the SNP position was called as interspecific SNP of type INTER-2W for two-way species comparison (e.g., $INTER\text{-}2W_{Fp\text{-}Lm}$). Such interspecific SNPs were identified for all three combinations of two-way species comparison (Fp–Lm, Fp–Lp, and Lm–Lp). In case of meadow fescue and Italian ryegrass, all six genotypes had to have a SNP call. In case of perennial ryegrass, some of the 14 genotypes had low read depth (see Supplemental Table S1) resulting in a relatively low number of SNPs with calls in all 14 genotypes. Thus, we requested at least 10 out of 14 genotypes to have SNP calls.

(iii) If all six Fp genotypes and all six Lm genotypes or at least 10 out of 14 Lp genotypes were monomorphic in one species and carried the monomorphic alternative allele in the two other species, the SNP position was called interspecific SNP of type INTER-3W for three-way species comparison (e.g., $INTER\text{-}3W_{Fp}$ is different in Fp vs. both Lm and Lp; see Fig. 1B).

Classification of Interspecific Single-Nucleotide Polymorphisms in Pairwise Comparisons

Pairwise comparisons were made between tetraploid genotypes of meadow fescue and tetraploid genotypes of Italian ryegrass, to simulate biparental hybrid crosses. If the meadow fescue genotype contained a homozygous reference allele, and the Italian ryegrass genotype contained the homozygous alternative allele (or vice versa), the SNP was called interspecific SNP of INTER-PW type for pairwise comparison (e.g., $INTER\text{-}PW_{Fp52\text{-}Lm51}$). Such pairwise comparisons were made for all six combinations of three tetraploid meadow fescue genotypes (Fp52, Fp53, and Fp54) and two tetraploid Italian ryegrass genotypes (Lm51 and Lm52). For each pairwise combination, a unique HQ gene set (which is different from the near full-length gene set described above) was selected after three filtering steps (see Fig. 1C). Each OGA-assembled transcript in the separate transcriptomes is annotated to the central *B. distachyon* gene set, thus delineating orthologs across the three transcriptomes. Therefore, the effect of the reference bias can be analyzed on a gene-per-gene level. For Filter 1, we used only 8094 near full-length transcripts (85–115% of the *B. distachyon* CDS length in both species) shared between the meadow fescue and Italian ryegrass references. For Filter 2, to avoid read mapping bias, we retained only transcripts for which the ratio of average read depth for meadow fescue reference transcript and Italian ryegrass reference transcript varied between 0.5 and 2 for both meadow fescue and Italian ryegrass reads mapped on both respective transcriptomes. For Filter 3, we retained only those genes where the difference in the number of SNPs identified in meadow fescue vs. Italian ryegrass orthologous reference transcript was $\leq$2 or <20% of the average. Together, these three filters strongly select for transcripts with little or no reference sequence specific bias and allow studying the PGAS expression in interspecific hybrids using any of the two reference transcriptomes for read mapping. For each

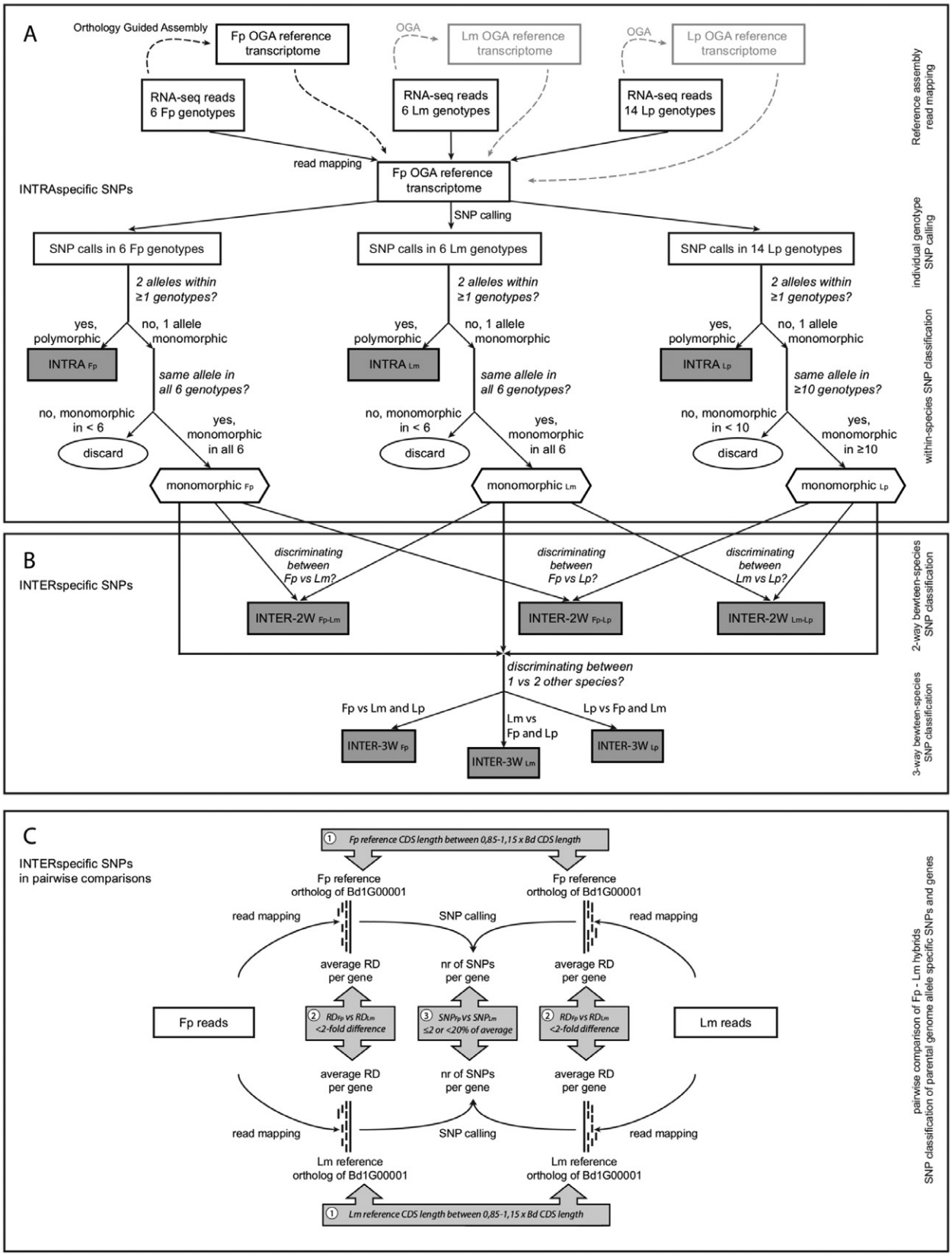

Fig. 1. Classification of intraspecific and interspecific single-nucleotide polymorphisms (SNPs). Selection criteria are indicated in italics. (A) De novo assembly and Orthology Guided Assembly generate a reference transcriptome for each of the three species. Intraspecific SNPs (INTRA) are identified by mapping all RNA-seq data from *Festuca pratensis*, *Lolium multiflorum*, and *L. perenne* onto a single reference transcriptome. This analysis is performed in parallel for each reference transcriptome. (B) Interspecific SNPs are identified between two species (INTER-2W) or by comparing all three species (INTER-3W). (C) Filtering scheme for the selection of high-quality reference transcripts in pairwise genotype comparisons (INTER-PW).

pairwise comparison, a specific HQ set of transcripts was defined because these depend on the specific read depth (Filter 2) and SNP sets (Filter 3) for a given combination of parental genotypes (see Fig. 1C).

SNPhylo

A phylogenetic tree comparing SNP profiles of the 26 genotypes using the meadow fescue reference transcriptome was calculated in SNPhylo (v.20140430; Lee et al., 2014) using 100 bootstraps. Only SNPs without missing data in any of the 34 samples (including RNA-Seq replicates) and with minor allele frequency >10% (default settings) were used over the whole set of genotypes. The SNP sets identified with the three different reference transcriptomes gave nearly identical results (data not shown).

Positioning of Genes using the Perennial Ryegrass GenomeZipper

We positioned orthologous genes of meadow fescue, Italian ryegrass, and perennial ryegrass containing various classes of SNPs on seven LGs of perennial ryegrass using the GenomeZipper approach (Mayer et al., 2009). This approach is based on the collinearity of the gene order shared between a model species (e.g., *B. distachyon*) and the three species of our interest. The genetic map of perennial ryegrass (Studer et al., 2012) was used as a backbone to determine syntenic chromosome blocks between *B. distachyon* and perennial ryegrass. A similar virtual gene order was previously published for perennial ryegrass (Pfeifer et al., 2013). We placed additional genes onto the existing virtual gene order of perennial ryegrass based on orthologous relationships with *B. distachyon* genome using in-house Perl scripts. Previously unanchored meadow fescue, Italian ryegrass, and perennial ryegrass transcripts were positioned between orthologs of neighboring *B. distachyon* genes with a known position in the GenomeZipper.

We used Circos (Krzywinski et al., 2009) for circular visualization of orthologous blocks of *B. distachyon* chromosomes with perennial ryegrass LGs, distribution of all Italian ryegrass transcripts, distribution of near full-length transcripts assembled in all three species, the number of intraspecific SNPs per gene $INTRA_{Fp}$, $INTRA_{Lm}$, $INTRA_{Lp}$, and two examples of interspecific SNPs in pairwise comparisons of potential Italian ryegrass × meadow fescue hybrids.

Validation of Single-Nucleotide Polymorphism by High-Resolution Melting Curve Analysis

High-resolution melting curve analysis (HRM) was employed to validate the SNPs identified in silico. Ten SNPs (from 10 genes) discriminating the alleles of meadow fescue and Italian ryegrass were selected. Primers for HRM analysis were designed to amplify 43 to 67 bp DNA fragments with a melting temperature of 60°C. High-resolution melting curve analysis–PCR was performed as described by Studer et al. (2009) in a total volume of 10 μL using 1× LightScanner High Sensitivity Genotyping MasterMix containing LCGreen PLUS (Idaho Technology, Inc.). Each reaction contained 20 ng of genomic DNA and 0.10 μM of forward and reverse primers and was covered with 15 μL mineral oil to avoid sample evaporation during PCR and the melting process. The PCR amplification was conducted in a Thermoblock 96 Cycler (SensoQuest) under the following conditions: initial denaturation for 2 min at 95°C, 40 cycles of denaturation for 30 s at 94°C, annealing for 30 s at T_a (Supplemental Table S3), and elongation for 30 s at 72°C. The melting curve analysis was performed at a temperature range from 60 to 95°C in steps of 0.05°C using a LightScanner Instrument and the LightScanner and Call-IT software modules (Idaho Technology).

Table 1. Number of intraspecific single-nucleotide polymorphisms (italic) in *Festuca pratensis* (Fp), *Lolium multiflorum* (Lm), and *L. perenne* (Lp) and corresponding genes (bold) identified using three different reference transcriptomes with all genes or the 6475 near full-length gene set (in parentheses).

Reference transcriptome	$INTRA_{Fp}$	$INTRA_{Lm}$	$INTRA_{Lp}$
Fp	*250,929 (113,934)*	*715,572 (370,524)*	*529,367 (288,958)*
	12,910 (5,552)	**13,864 (5,966)**	**13,472 (5,940)**
Lm	*248,271 (115,390)*	*750,463 (385,592)*	*556,027 (303,890)*
	13,420 (5,750)	**15,320 (6,279)**	**14,717 (6,258)**
Lp	*249,469 (114,545)*	*786,791 (388,114)*	*634,918 (314,057)*
	13,670 (5,768)	**15,368 (6,288)**	**15,617 (6,321)**

Results

De Novo Transcriptome Assembly in Meadow Fescue and Italian Ryegrass

Transcriptome sequencing was performed on 10 samples (corresponding to six genotypes) of meadow fescue and 10 samples (corresponding to six genotypes) of Italian ryegrass, yielding 38.6 million reads per sample on average (ranging from 14.7 to 73.5 million). After quality trimming, reads of each sample were used separately for de novo assembly using CLCBio Genomics Workbench v8.4. This yielded between 41,110 and 100,314 contigs per sample with N50 ranging from 558 to 1025 bp (Supplemental Table S1). The N50 value was markedly lower in Italian ryegrass (666 bp on average) than meadow fescue (941 bp on average) suggesting a higher level of sequence variation and heterozygosity in Italian ryegrass. Using the OGA approach, 18,952 nonredundant Italian ryegrass transcripts were assembled by combining the contigs of all six genotypes based on orthology with the *B. distachyon* proteome. Similarly, 19,036 nonredundant Italian ryegrass transcripts were assembled and annotated (Table 1).

In total, 17,455 orthologous transcripts were shared between the transcriptomes of meadow fescue and Italian ryegrass. Out of these, 16,613 transcripts overlapped with the previously published perennial ryegrass transcriptome containing 19,279 nonredundant transcripts (Ruttink et al., 2013) generated using the same OGA

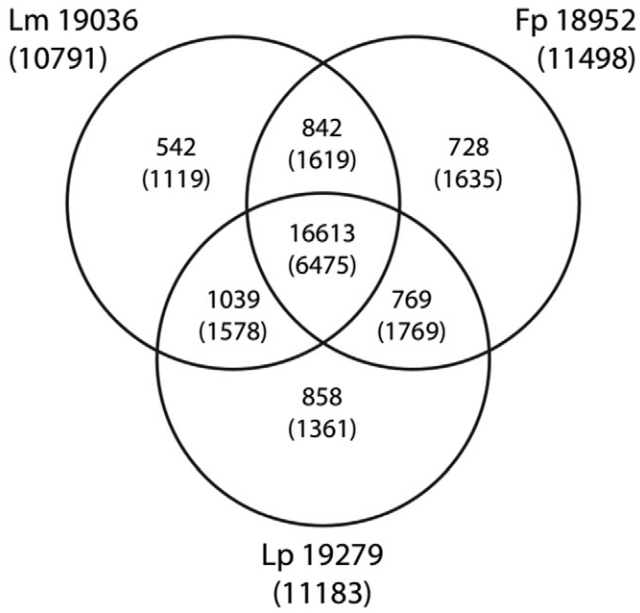

Fig. 2. Venn diagram of the overlap in the number of transcripts in three different reference transcriptomes (*Festuca pratensis*, *Lolium multiflorum*, and *L. perenne*). The number of near full-length transcripts is given in parentheses.

procedure. Only 728, 542, and 858 genes were uniquely assembled in the meadow fescue, Italian ryegrass, and perennial ryegrass reference transcriptomes, respectively (Fig. 2). Most likely, this was due to low expression levels in the other species. From a total of 11,498; 10,791; and 11,183 near full-length transcripts (length of the transcript CDS varied between 85 and 115% of the respective orthologous *B. distachyon* CDS) in meadow fescue, Italian ryegrass, and perennial ryegrass, respectively, 8094 were shared between Italian ryegrass and meadow fescue, and 6475 transcripts were common for all three species (Fig. 2). Estimation of the number of SNPs per transcript depends on the length of the assembled transcript, which varied between species. Thus, to make accurate comparisons of SNP counts and SNP densities between species, we only compared orthologous transcripts that were near full-length in all three references (linked through their OGA based annotation). This is represented by a set of 6475 genes per reference transcriptome (Table 1, 2).

Single-Nucleotide Polymorphism Discovery and Robustness of Genotype Calling

The robustness of SNP calling was tested by comparing genotype calls across the three biological replicates of two genotypes of meadow fescue and two genotypes of Italian ryegrass. The results indicated high-quality SNP data as 98.6, 98.9, 99.1, and 99.2% of the genotype calls for the tetraploid genotypes Fp52, Fp53, Lm51, and Lm52, respectively, were identical among replicates.

Phylogenetic relationships between all genotypes were determined using SNPhylo. Expectedly, genotypes from the three species formed separate clusters and both Italian and perennial ryegrass clustered relatively close to each other. The lowest level of genetic diversity was found among the meadow fescue genotypes (Fig. 3) and the highest among the 14 perennial ryegrass genotypes, which originate from three cultivars and 11 natural accessions from across Europe.

Intraspecific Single-Nucleotide Polymorphisms per Species

We first identified SNPs that were polymorphic within species irrespective of their genotype calls in the other two species (Fig. 1A; Table 1). Similar numbers of genes with at least one SNP were identified in all three species (Table 1), while Italian ryegrass displayed the highest number of SNPs per gene (49.8 SNPs per gene), followed by perennial ryegrass (37.8 SNPs per gene) and meadow fescue (18.5 SNPs per gene) using the Italian ryegrass reference transcriptome. There were only slight differences among the results using three different reference transcriptomes (Supplemental Fig. S2). Because the perennial ryegrass reference transcriptome contains slightly more genes, higher numbers of SNPs and corresponding genes were identified in all three species using this reference transcriptome.

Interspecific Single-Nucleotide Polymorphisms in Group Comparisons

Interspecific SNPs (Fig. 1B) were identified by two-way comparison (when just two species have to be distinguishable) and by three-way comparison (when one species carries a different allele at a given position as compared with the other two species). This analysis was performed separately per reference transcriptome, and data was presented on either the complete reference transcriptome or on the subset of 6475 near full-length genes per reference transcriptome (Table 2) for comparisons among references.

In total, 21,992 SNPs located in 4962 genes of the Italian ryegrass reference transcriptome, which can distinguish meadow fescue from Italian ryegrass, were identified. Of these, 13,373 SNPs were located in 2668 genes of the 6475 near full-length Italian ryegrass gene set. This set of SNPs can be used to develop a platform for genotyping a broader genepool of meadow fescue and Italian ryegrass to validate its potential to discriminate between species. Validated SNPs can further be used to characterize genomic constitution of meadow fescue × Italian ryegrass hybrids without a prior knowledge of parents from which they derived. Similar number of SNPs was identified for perennial ryegrass–meadow fescue combinations (Table 2). In contrast, only 354 interspecific SNPs located in 109 genes of the Italian ryegrass reference transcriptome that consistently discriminate perennial ryegrass from Italian ryegrass were identified. Similarly, a three-way comparison identified 13,260 SNPs located in 3611 genes of the Italian ryegrass reference transcriptome, which discriminated meadow fescue from both *Lolium* species (Table 2). Because of the low number of interspecific SNPs distinguishing perennial ryegrass from Italian ryegrass, only 122 SNPs located in 45 genes of the Italian ryegrass reference transcriptome were

Table 2. Number of interspecific single-nucleotide polymorphisms (italic) in *Festuca pratensis* (Fp), *Lolium multiflorum* (Lm), and *L. perenne* (Lp) and corresponding genes (bold) in two-way and three-way comparisons identified using three different reference transcriptomes with all genes or the 6475 near full-length gene set (in parentheses).

Reference transcriptome	Fp–Lm	Fp–Lp	Lm–Lp	Fp-specific	Lm-specific	Lp-specific
Fp	*20,840 (12,517)*	*20,223 (12,607)*	*358 (231)*	*11,603 (7,341)*	*134 (83)*	*130 (82)*
	4,554 (2,510)	**4,172 (2,348)**	**106 (69)**	**3,183 (1,843)**	**43 (26)**	**58 (40)**
Lm	*21,992 (13,373)*	*22,540 (14,181)*	*354 (241)*	*13,260 (8,406)*	*122 (85)*	*127 (83)*
	4,962 (2,668)	**4,648 (2,564)**	**109 (66)**	**3,611 (2,030)**	**45 (25)**	**62 (39)**
Lp	*24,765 (13,761)*	*26,397 (14,952)*	*472 (274)*	*15,204 (8,720)*	*179 (100)*	*178 (98)*
	5,343 (2,712)	**5,138 (2,641)**	**139 (81)**	**3,976 (2,084)**	**59 (33)**	**80 (47)**

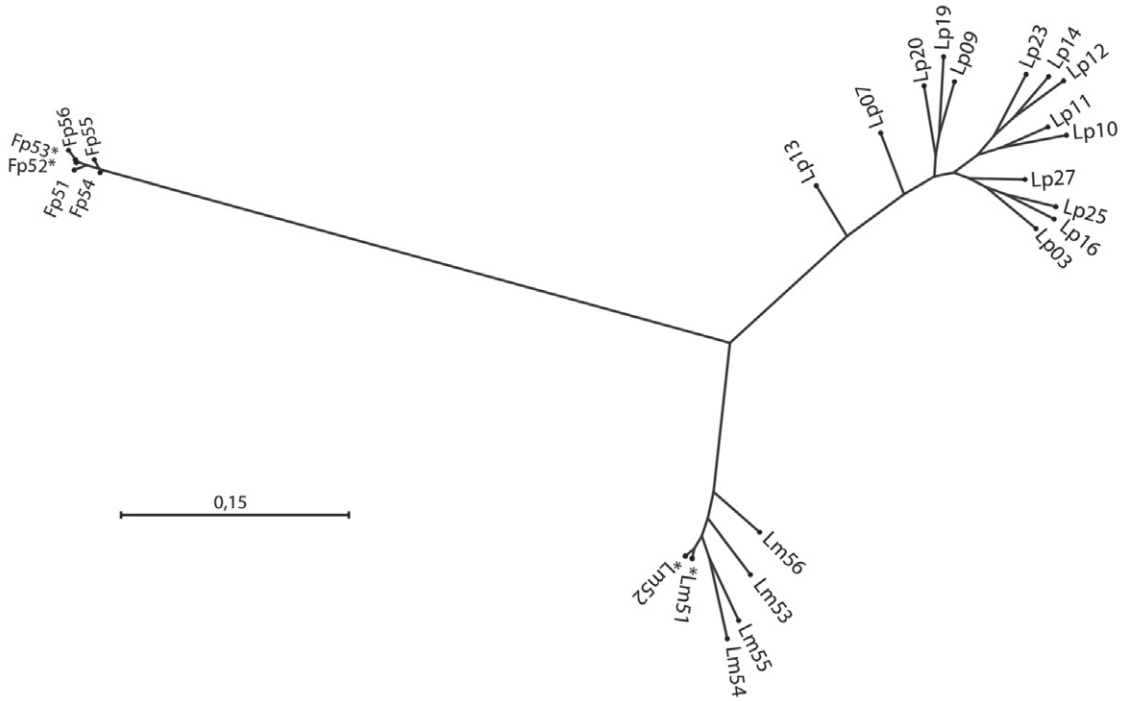

Fig. 3. SNPhylo phylogenetic tree based on 33,725 single-nucleotide polymorphisms shared among 26 genotypes of *Lolium perenne*, *L. multiflorum*, and *Festuca pratensis*. Asterisk (*) indicates overlapping positions in the phylogenetic tree of the three replicates for Fp52, Fp53, Lm51, and Lm52.

classified as species-specific for Italian ryegrass. One hundred twenty-seven SNPs located in 62 genes of the Italian ryegrass reference transcriptome were classified as species-specific for perennial ryegrass.

Because all three reference transcriptomes were annotated based on orthology with the *B. distachyon* proteome, we intersected the gene sets based on common *B. distachyon* orthologs and found that many of the genes with interspecific SNPs were commonly identified using the alternative reference transcriptomes. The numbers of SNPs identified in the set of 6475 near full-length orthologous transcripts present in all three reference transcriptomes were consistent (Table 2), indicating a limited bias of the reference transcriptomes in these analyses. A detailed comparison at the gene-by-gene level further showed high consistency in the number of SNPs called per gene across the 6475 near full-length gene sets (Supplemental Fig. S3). It is important to note that these correlations indicate reproducibility of read mapping and subsequent SNP calling and do not concern the general sequence similarity between orthologous reference sequences and also do not take into account the number of intraspecific SNPs per gene. Taken together, these analyses illustrate that, in general, the three reference sequences per gene can be used interchangeably to identify interspecific SNPs.

Interspecific Single-Nucleotide Polymorphisms in Pairwise Comparisons

Single-nucleotide polymorphisms and corresponding genes that enabled allele discrimination in Italian ryegrass × meadow fescue hybrids derived from known tetraploid parental genotypes were identified in a set of pairwise genotype comparisons. For each pair of parents, two complete reference transcriptomes (meadow fescue and Italian ryegrass) were used. The 62,052 interspecific SNPs located in 8370 genes in the combination Fp54–Lm51 and 93,967 interspecific SNPs located in 10,918 genes in

Table 3. The number of interspecific single-nucleotide polymorphisms (italic) and corresponding genes (bold) in pairwise comparison of selected *Festuca pratensis* (FP) and *Lolium multiflorum* (Lm) genotypes (Fp52, Fp53, or Fp54 vs. Lm51 or Lm52) identified using two different transcriptome references with all genes or with the HQ (high-quality) gene sets per pairwise comparison (in parentheses).

Reference transcriptome		Fp52	Fp53	Fp54
Fp	Lm51	*88,660 (49,261)*	*83,215 (46,518)*	*60,592 (33,228)*
		10,072 (5,090)	**9,667 (4,951)**	**7,737 (4,026)**
	Lm52	*79,744 (43,818)*	*75,810 (41,814)*	*61,957 (34,096)*
		9,277 (4,688)	**8,988 (4,587)**	**7,753 (4,021)**
Lm	Lm51	*93,967 (49,920)*	*86,697 (46,638)*	*62,052 (33,050)*
		10,918 (5,090)	**10,422 (4,951)**	**8,370 (4,026)**
	Lm52	*85,456 (44,633)*	*80,180 (42,332)*	*63,687 (34,046)*
		10,168 (4,688)	**9,775 (4,587)**	**8,374 (4,021)**

the combination Fp52–Lm51 were identified using the Italian ryegrass reference transcriptome (Table 3). Thus, the pairwise comparison strategy enabled scoring higher numbers of interspecific SNPs and respective genes than general group comparisons (INTER-2W and INTER-3W). However, these SNPs were specific to a particular combination of parents. Slightly higher numbers of SNPs were detected when the Italian ryegrass reference was used as compared with the meadow fescue reference.

Because interspecific SNPs in pairwise comparisons discriminated between alleles derived from specific parental genomes, they can be used to quantify PGAS expression and to analyze genes preferentially expressed from one parental genome in hybrid progeny. Since our genotypes were highly polymorphic (Table 1), the possibility was raised that intraspecific SNPs interfere with read mapping to the common consensus reference obtained from different genotypes during OGA. Furthermore, read mapping bias depends on the combination of the reference and the particular hybrid genotype under study as each genotype carries different alleles leading to genotype-specific read–reference mismatch patterns. We tested this potential bias on a gene-by-gene basis (see Fig. 1C) and for each biparental combination. We found that the number of discriminatory SNPs per gene identified using a given meadow fescue reference transcript was not always equal to the number of SNPs identified using the orthologous Italian ryegrass reference transcript (Fig. 4). Therefore, after filtering, as described in Material and Methods and Fig. 1C, only sets of genes without signs of read mapping bias, the so-called HQ gene sets, were kept. This reduced the number of SNPs and corresponding genes to 52 to 56% and 46 to 52%, respectively (Table 3).

Validation of Interspecific Single-Nucleotide Polymorphisms

A set of 10 in silico identified interspecific SNPs for pairwise comparisons located in 10 nuclear genes were validated by HRM analyses. Amplification products sufficiently specific for HRM analysis were obtained for all 10 SNPs. The HRM results were in agreement with the results of the in silico SNP discovery pipeline. The only exception was SNP 1G15070_322, where a clear assignment of the melting profiles into two groups failed, possibly from the presence of an intron between the two primer pairs (Supplemental Table S3).

Genome Location of Genes Containing Various Classes of Single-Nucleotide Polymorphisms

Using the *Lolium* GenomeZipper, 17,768 genes were computationally ordered into the seven perennial ryegrass LGs based on the presumed synteny of perennial ryegrass with *B. distachyon* and using the genetic map of perennial ryegrass as a backbone (Pfeifer et al., 2013). Here, we positioned additional genes into the virtual gene order of perennial ryegrass using the GenomeZipper with *B. distachyon* as reference species. This expanded the GenomeZipper to 23,501 anchored genes distributed across all seven perennial ryegrass chromosomes, ranging from 2488 genes in LG6 to 4276 genes in LG2. Subsequently, we could position various classes of SNPs and their corresponding genes onto the seven perennial ryegrass LGs. This gave an overview of the global spatial distribution of various SNPs (Fig. 5). Only small differences in the frequency of intraspecific, as well as interspecific SNPs, were detected along the chromosomes. Moreover, we did not observe any SNP hotspots or large blocks of genes without intra- or interspecific SNPs. This indicates that any part of the genome can be targeted in breeding or in genomic studies and also enables the detection of even very small introgressions in Festulolium hybrids. Additional information on the position of SNPs and corresponding genes is provided in Supplemental Table S2.

Discussion

The advent of next-generation sequencing technologies has been a driving force in the expansion of genomic resources and the knowledge on plant genome structure, function, evolution, and diversity. While next-generation sequencing technologies readily provide vast amounts of short-read sequence data, it is challenging to assemble sequence reads into longer contigs especially in large and complex genomes with a high fraction of repetitive DNA (Alkan et al., 2011). RNA-seq targets transcribed sequences and avoids sequencing repetitive parts of the genome. Thus, this approach supports gene discovery and directs the identification of SNP markers toward the functional genes. In this study, we extensively sequenced transcriptomes of six meadow fescue and six Italian ryegrass genotypes and applied an OGA strategy for transcriptome assembly and annotation. The transcriptomes were compared with previously published RNA-seq data from 14 perennial ryegrass genotypes and used for large-scale transcript-anchored SNP discovery.

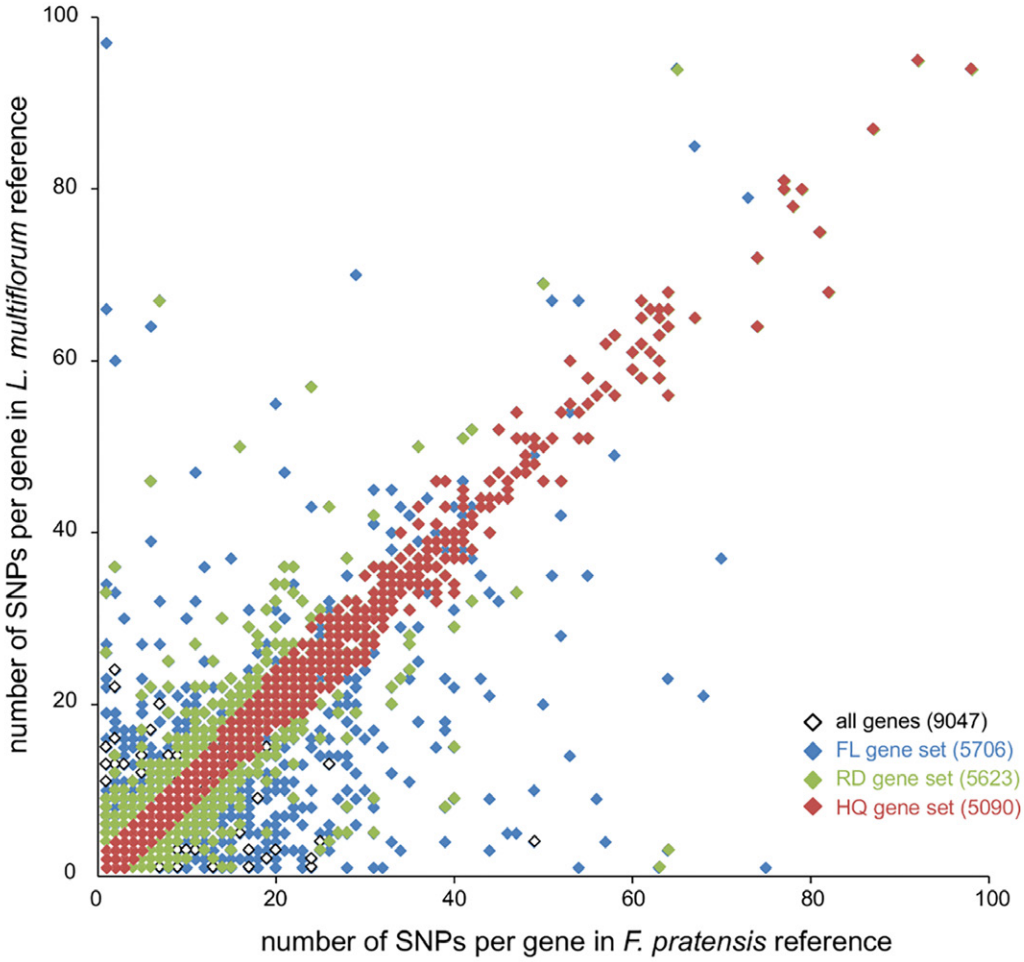

Fig. 4. Selection of single-nucleotide polymorphisms (SNPs) suitable for discrimination between parental genome-specific alleles in progeny derived from a cross between *Festuca pratensis* genotype Fp52 and *Lolium multiflorum* genotype Lm51. Comparison of selected genes sets carrying such SNPs using various filtering criteria. FL, filtered for near full-length transcript length; RD, filtered for average read depth per transcript; HQ (high quality), filtered for number of SNPs per transcript. The numbers of transcripts that pass the filtering criteria are given in parentheses. Only transcripts with at least one SNP per gene in both transcript references are shown.

De Novo Assembly of the Transcriptome of Meadow Fescue, Italian Ryegrass, and Perennial Ryegrass

Using OGA, we assembled 18,952 and 19,036 nonredundant transcripts of meadow fescue and six Italian ryegrass genotypes, respectively, which represent orthologs of ~70% of the *B. distachyon* proteome. This is highly consistent with the previously assembled perennial ryegrass transcriptome (Ruttink et al., 2013). Out of these, 11,498; 10,791; and 11,183 represent near full-length transcripts in meadow fescue, Italian ryegrass, and perennial ryegrass, respectively. The numbers are greater than the 9646 and 10,430 near full-length transcripts of perennial ryegrass (Czaban et al., 2015) but lower than 28,455 gene models identified by Byrne et al. (2015) for perennial ryegrass. This might be at least partially explained by the fact that we used only one tissue for RNA isolation, and thus we did not capture the entire transcriptome.

The main goal of this study was to develop a robust strategy to identify gene-anchored, intra- and interspecific SNPs across three closely related species. However, previous studies on genetic diversity in *Festuca* and *Lolium* revealed great variation in SNP densities across different genes within a species (Ruttink et al., 2013) as well as varying degrees of evolutionary conservation across different genes between species (Czaban et al., 2015). Clearly, SNP identification, as well as read-count-based quantification of allele-specific expression levels, critically depends on the way the reads are mapped onto a representative reference sequence. For this reason, we used multiple SNP and gene filtering to avoid read-mapping bias. The numbers of SNPs identified using different references were compared on a gene-by-gene basis. We proved that application of appropriate filtering ensures comparable results independent on the reference used.

Genetic Variability within Species Predicted by Intraspecific Single-Nucleotide Polymorphisms

A two- to three-fold difference in the number of intraspecific SNPs per gene was found between the species. Compared with both ryegrasses, meadow fescue had the lowest number of intraspecific SNPs per gene, indicating

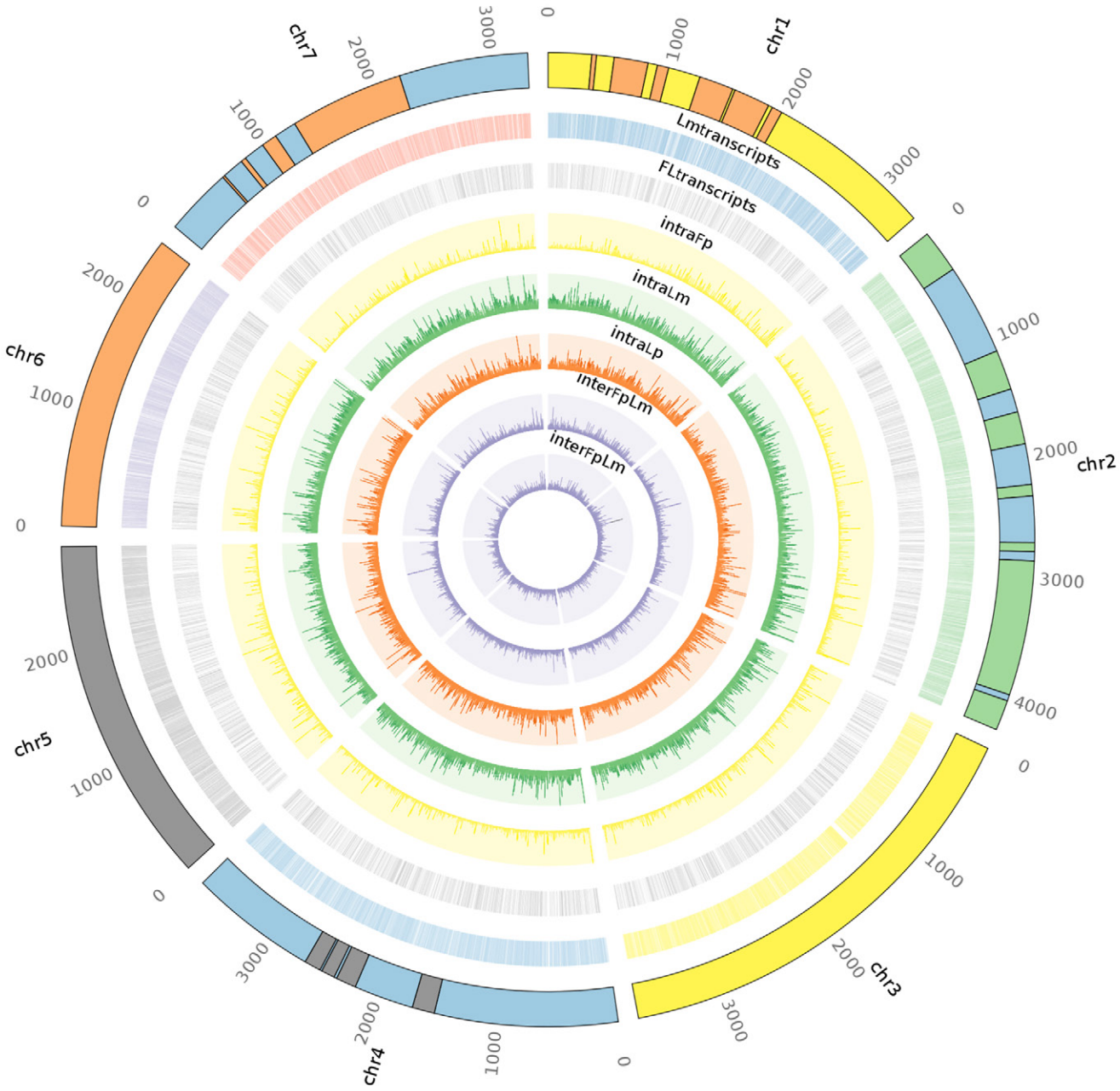

Fig. 5. Circular map of the seven *Lolium perenne* genetic linkage groups. From outer to inner circle: orthologous blocks of *Brachypodium distachyon* chromosomes with *Lolium perenne* (each color represents one *B. distachyon* chromosome: Bd1 blue, Bd2 yellow, Bd3 orange, Bd4 gray, and Bd5 green); distribution of all *L. multiflorum* transcripts; distribution of 6475 near full-length (FL) transcripts assembled in all three species; the number of intraspecific SNPs per gene $INTRA_{Fp}$ (yellow), $INTRA_{Lm}$ (green), $INTRA_{Lp}$ (orange) with maximum 400 SNPs per gene; the number of interspecific SNPs per gene in pairwise comparisons (violet) with a maximum of 150 SNPs per gene ($INTER\text{-}PW_{Fp52\text{-}Lm51}$ and $INTER\text{-}PW_{Fp54\text{-}Lm51}$). Note that the circular maps show linear order of genes but not physical distances between them.

lower genetic variability within this species. This observation, although based on the analysis of a relatively limited number of genotypes per species, correlates with previous findings of Kopecký et al. (2009) and Kölliker et al. (1999). The low level of intraspecific genetic variability is probably a consequence of the genetic bottleneck that meadow fescue has undergone since the last glaciation (Fjellheim et al., 2006).

Interspecific Single-Nucleotide Polymorphisms Enable the Analysis of the Genomic Composition and Gene Expression Analysis of Interspecific Hybrids

Although not novel, SNP genotyping to discriminate plant species and distinguish interspecific hybrids and allopolyploids remains a challenging approach. Curk et al. (2015) employed 454 amplicon sequencing of 57 gene

fragments to mine 105 SNPs distinguishing parental species of modern *Citrus* varieties. High numbers of species-specific SNPs (~33,000) identified in the Illumina Infinium *Brassica* SNP 60K array enabled discrimination of the A and C genomes in *Brassica* allopolyploids (Mason et al., 2015).

The approach presented here allowed us to identify transcript-anchored SNPs that are spread evenly across the genome and can be used to distinguish between the three species (i.e., interspecific SNPs). Moreover, interspecific SNPs can be used to characterize the genomic constitution of interspecific Festulolium hybrids at a high level of resolution. Most of the recent Festulolium cultivars originated from crosses made decades ago followed by several steps of polycrosses and backcrosses (Ghesquiere et al., 2010). Consequently, the information on parental genotypes is lacking for most Festulolium cultivars. For those cultivars, thousands of interspecific SNPs that are monomorphic within species and polymorphic between species (Table 2) can be instrumental to reconstruct their ancestry and relatedness. If new interspecific crosses are made with known parental genotypes, like the ones suggested here, the interspecific SNPs in pairwise comparisons (Table 3) will enable to study genome stability and fine-scale chromosomal rearrangements in subsequent generations of Festulolium (Le Scouarnec and Gribble, 2012). The great advantage of our approach is the even spread of SNP markers across the genomes and a possibility to predict the expected resolution in a given chromosomal region based on the analysis of parental transcriptomes. This may guide the choice of specific parents for interspecific crosses. Another application of selected SNPs in specific HQ gene sets (Table 3) is PGAS expression analysis. We present evidence of gene-specific read mapping bias and resolve this issue by suggesting well-defined controls to avoid reference transcriptome dependent mapping bias, which may affect read-count-based estimations of expression levels (Li and Jiang 2012).

Genotyping Platform for Large-Scale Screening

To facilitate large-scale screening of thousands to tens of thousands of plants in grass breeding programs, a high-throughput genotyping platform needs to be developed with a set of highly informative markers. Customized array-based genotyping platforms were developed for *Lolium* and used to construct genetic maps in biparental crosses (Studer et al. 2012), identify cultivars of Italian and perennial ryegrasses (Wang et al. 2014), and study genetic diversity and the demographic history of natural accessions of perennial ryegrass (Blackmore et al., 2015). Our SNP sets contain complete positional information (putative genome location, neighboring SNPs) on three different types of SNPs suitable for the development of novel probe sets for genes spread across the genome.

Alternatively, targeted resequencing of selected regions of a genome containing intra- and interspecific SNPs could decrease the cost of genotyping. Several methods for target enrichment of genomic DNA have been developed (reviewed in Teer and Mullikin, 2010). Because of a high level of target and sample multiplexing, amplicon resequencing seems to be the most promising approach to genotype hundreds of loci in thousands of plants involved in breeding programs, as it allows simultaneous screening for known SNPs and to discover novel polymorphisms.

Apart from the results from a recent *Lolium–Festuca* study (Czaban et al., 2015), our data represent a valuable resource for ecological and evolutionary genomic research in these species. The two *Lolium–Festuca* transcriptome studies complement each other well. Czaban et al. (2015) used RNA-seq for comparative gene family analysis, phylogenetic analysis, and identification of genes under positive selection pressure. In contrast, our study provides a valuable tool for precise characterization of genomic constitution and gene expression in prospective Festulolium hybrids. Once efficient screening methods have been developed, our set of intraspecific SNPs may be used for QTL fine mapping, map-based cloning of genes involved in traits of agricultural, ecological, and evolutionary interest; genome-wide association studies; demographic history and selection analyses; ecopopulation profiling; and seed contamination screening. Thus, this study provides an outstanding resource for the grass research community and forage grass breeders. Moreover, the versatility of our approach makes it suitable for use in genomic and evolutionary studies of any interspecific hybrids and allopolyploids.

Supplemental Information Available

Supplemental information is available with the online version of this manuscript.

Acknowledgments

This work was supported by Czech Science Foundation (award P501/11/0504), National Program of Sustainability (award no. LO1204), IGA (award no. Prf/2012/001) and Sciex-NMS^{ch}, a scientific exchange program between Switzerland and the new member states of the European Union (Project Code 14.099). Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the program "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is greatly appreciated. We thank Sabine van Glabeke for excellent bioinformatics support.

References

Alkan, C., S. Sajjadian, and E.E. Eichler. 2011. Limitations of next-generation genome sequence assembly. Nat. Methods 8:61–65. doi:10.1038/nmeth.1527

Birrer, M., R. Kölliker, Ch. Manzanares, T. Asp, and B. Studer. 2014. A DNA marker assay based on high-resolution melting curve analysis for distinguishing species of the *Festuca–Lolium* complex. Mol. Breed. 8:421–429. doi:10.1007/s11032-014-0044-0

Blackmore, T., I. Thomas, R. McMahon, W. Powell, and M. Hegarty. 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. doi:10.1007/s00122-015-2556-3

Byrne, S.L., I. Nagy, M. Pfeifer, I. Armstead, S. Swain, B. Studer, et al. 2015. A synteny-based draft genome sequence of the forage grass *Lolium perenne*. Plant J. 84:816–826. doi:10.1111/tpj.13037

Canter, P.H., I. Pašakinskiene, R.N. Jones, and M.W. Humphreys. 1999. Chromosome substitutions and recombination in the amphiploid

Lolium perenne × *Festuca pratensis* cv. Prior (2*n* = 4*x* = 28). Theor. Appl. Genet. 98:809–814. doi:10.1007/s001220050087

Curk, F., G. Ancillo, F. Ollitrault, X. Perrier, J.P. Jacquemoud-Collet, A. Garcia-Lor, et al. 2015. Nuclear species-diagnostic SNP markers mined from 454 amplicon sequencing reveal admixture genomic structure of modern citrus varieties. PLoS One 10:e0125628. doi:10.1371/journal.pone.0125628

Czaban, A., S. Sharma, S.L. Byrne, M. Spannagl, K.F.X. Mayer, and T. Asp. 2015. Comparative transcriptome analysis within the *Lolium/Festuca* species complex reveals high sequence conservation. BMC Genomics 16:249. doi:10.1186/s12864-015-1447-y

Denoeud, F., J.M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, et al. 2008. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 9:R175. doi:10.1186/gb-2008-9-12-r175

DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498. doi:10.1038/ng.806

Ergon, A., C. Fang, O. Jorgensen, T.S. Aamlid, and O.A. Rognli. 2006. Quantitative trait loci controlling vernalisation requirement, heading time and number of panicles in meadow fescue (*Festuca pratensis* Huds.). Theor. Appl. Genet. 112:232–242. doi:10.1007/s00122-005-0115-z

Farrell, J.D., S. Byrne, C. Paina, and T. Asp. 2014. *De Novo* Assembly of the perennial ryegrass transcriptome using an RNA-seq strategy. PLoS One 9:e103567. doi:10.1371/journal.pone.0103567

Fjellheim, S., O.A. Rognli, K. Fosnes, and C. Brochmann. 2006. Phylogeographical history of the widespread meadow fescue (*Festuca pratensis* Huds.) inferred from chloroplast DNA sequences. J. Biogeogr. 33:1470–1478. doi:10.1111/j.1365-2699.2006.01521.x

Garvin, M.R., K. Saitoh, and A.J. Gharrett. 2010. Application of single nucleotide polymorphisms to non-model species: A technical review. Mol. Ecol. Resour. 10:915–934. doi:10.1111/j.1755-0998.2010.02891.x

Ghesquiere, M., M.W. Humphreys, and Z. Zwierzykowski. 2010. Festulolium. In: B. Boller, U.K. Posselt, and F. Veronesi, editors, Fodder crops and amenity grasses, book series: Handbook of plant breeding 5. Springer Science+Business Media, Berlin. p. 293–316.

Gonçalves da Silva, A., W. Barendse, J.W. Kijas, W.C. Barris, S. McWilliam, R.J. Bunch, et al. 2014. SNP discovery in nonmodel organisms: Strand bias and base-substitution errors reduce conversion rates. Mol. Ecol. Resour. 15:723–736. doi:10.1111/1755-0998.12343

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29:644–652. doi:10.1038/nbt.1883

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877. doi:10.1101/gr.9.9.868

Humphreys, M., U. Feuerstein, M. Vandewalle, and J. Baert. 2010. Ryegrasses. In: B. Boller, U.K. Posselt, and F. Veronesi, editors, Fodder crops and amenity grasses, book series: Handbook of plant breeding 5. Springer Science+Business Media, Berlin. p. 211–260.

Kölliker, R., F.J. Stadelmann, B. Reidy, and J. Nösberger. 1999. Genetic variability of forage grass cultivars: A comparison of *Festuca pratensis* huds., *Lolium perenne* L. and *Dactylis glomerata* L. Euphytica 106:261–270. doi:10.1023/A:1003598705582

Kopecký, D., J. Bartoš, P. Christelová, V. Černoch, A. Kilian, J. Doležel. 2011. Genomic constitution of *Festuca* × *Lolium* hybrids revealed by the DArTFest array. Theor. Appl. Genet. 122:355–363. doi:10.1007/s00122-010-1451-1

Kopecký, D., J. Bartoš, A.J. Lukaszewski, J.H. Baird, V. Černoch, R. Kölliker, et al. 2009. Development and mapping of DArT markers within the *Festuca–Lolium* complex. BMC Genomics 10:473. doi:10.1186/1471-2164-10-473

Kopecký, D., J. Loureiro, Z. Zwierzykowski, M. Ghesquière, and J. Doležel. 2006. Genome constitution and evolution in *Lolium* × *Festuca* hybrid cultivars (Festulolium). Theor. Appl. Genet. 113:731–742. doi:10.1007/s00122-006-0341-z

Kopecký, D., A.J. Lukaszewski, and J. Doležel. 2005. Genomic constitution of Festulolium cultivars released in the Czech Republic. Plant Breed. 124:454–458. doi:10.1111/j.1439-0523.2005.01127.x

Krzywinski, M., J. Schein, I. Birol, J. Conners, R. Gascoyne, D. Horsman, S.J. Jones, and M.A. Marra. 2009. Circos: An information aesthetic for comparative genomics. Genome Res. 19:1639–1645. doi:10.1101/gr.092759.109

Książczyk, T., E. Zwierzykowska, K. Molik, M. Taciak, P. Krajewski, Z. Zwierzykowski. 2015. Genome-dependent chromosome dynamics in three successive generations of the allotetraploid *Festuca pratensis* × *Lolium perenne* hybrid. Protoplasma 252:985–996. doi:10.1007/s00709-014-0734-9

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324

Li, W., and T. Jiang. 2012. Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics 28:2914–2921. doi:10.1093/bioinformatics/bts559

Lee, T.H., H. Guo, X. Wang, C. Kim, and A.H. Paterson. 2014. SNPhylo a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162. doi:10.1186/1471-2164-15-162

Le Scouarnec, S., and S.M. Gribble. 2012. Characterising chromosome rearrangements: Recent technical advances in molecular cytogenetics. Heredity 108:75–85. doi:10.1038/hdy.2011.100

Mason, A.S., J. Takahira, C. Atri, B. Samans, A. Hayward, W.A. Cowling, J. Batley, and M.N. Nelson. 2015. Microspore culture reveals complex meiotic behaviour in a trigenomic *Brassica* hybrid. BMC Plant Biol. 15:173. doi:10.1186/s12870-015-0555-9

Mayer, K.F.X., S. Taudien, M. Martis, H. Šimková, P. Suchánková, H. Gundlach, et al. 2009. Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. 151:496–505. doi:10.1104/pp.109.142612

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. doi:10.1101/gr.107524.110

Modrek, B., and C. Lee. 2002. A genomic view of alternative splicing. Nat. Genet. 30:13–19. doi:10.1038/ng0102-13

Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5:621–628. doi:10.1038/nmeth.1226

Pašakinskiene, I., C.M. Griffiths, A.J.E. Bettany, V. Paplauskiene, and M.W. Humphreys. 2000. Anchored simple-sequence repeats as primers to generate species-specific DNA markers in *Lolium* and *Festuca* grasses. Theor. Appl. Genet. 100:384–390. doi:10.1007/s001220050050

Peng, Y., X. Gao, R. Li, and G. Cao. 2014. Transcriptome sequencing and de novo analysis of *Youngia japonica* using the Illumina platform. PLoS One 9:e90636. doi:10.1371/journal.pone.0090636

Pfeifer, M., M. Martis, T. Asp, K.F.X. Mayer, T. Lübberstedt, S. Byrne, U. Frei, and B. Studer. 2013. The perennial ryegrass GenomeZipper: Targeted use of genome resources for comparative grass genomics. Plant Physiol. 161:571–582. doi:10.1104/pp.112.207282

Proost, S., M. Van Bel, D. Vaneechoutte, Y. Van de Peer, D. Inzé, B. Mueller-Roeber, and K. Vandepoele. 2014. PLAZA 3.0: An access point for plant comparative genomics. Nucleic Acids Res. 43:D974–D981. doi:10.1093/nar/gku986

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, et al. 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7:909–912. doi:10.1038/nmeth.1517

Rognli, O.A., M.C. Saha, S. Bhamidimarri, and S. van der Heijden. 2010. Fescues. In: B. Boller, U.K. Posselt, and F. Veronesi, editors, Fodder crops and amenity grasses, book series: Handbook of plant breeding 5. Springer Science+Business Media, Berlin. p. 261–292.

Ruttink, T., L. Sterck, A. Rohde, C. Bendixen, P. Rouzé, T. Asp, Y. Van de Peer, and I. Roldan-Ruiz. 2013. Orthology guided assembly in highly heterozygous crops: Creating a reference transcriptome to uncover genetic diversity in *L. perenne*. Plant Biotechnol. J. 11:605–617. doi:10.1111/pbi.12051

Schulz, M.H., D.R. Zerbino, M. Vingron, and E. Birney. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092. doi:10.1093/bioinformatics/bts094

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J.M. Jones, and İ. Birol. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117–1123. doi:10.1101/gr.089532.108

Studer, B., B. Boller, D. Herrmann, E. Bauer, U.K. Posselt, F. Widmer, R. Kölliker. 2006. Genetic mapping reveals a single major QTL for

bacterial wilt resistance in Italian ryegrass (*Lolium multiflorum* Lam.). Theor. Appl. Genet. 113:661–671. doi:10.1007/s00122-006-0330-2

Studer, B., S. Byrne, R.O. Nielsen, F. Panitz, C. Bendixen, M.S. Islam, M. Pfeifer, T. Lübberstedt, and T. Asp. 2012. A transcriptome map of perennial ryegrass (*Lolium perenne* L.). BMC Genomics 13:140. doi:10.1186/1471-2164-13-140

Studer, B., L.B. Jensen, A. Fiil, and T. Asp. 2009. "Blind" mapping of genic DNA sequence polymorphisms in *Lolium perenne* L. by high resolution melting curve analysis. Mol. Breed. 24:191–199. doi:10.1007/s11032-009-9291-x

Teer, J.K., and J.C. Mullikin. 2010. Exome sequencing: The sweet spot before whole genomes. Hum. Mol. Genet. 19:R145–R151. doi:10.1093/hmg/ddq333

The International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768. doi:10.1038/nature08747

Varshney, R.K., S.N. Nayak, G.D. May, and S.A. Jackson. 2009. Next generation sequencing technologies and their application for crop genetics and breeding. Trends Biotechnol. 27:522–530. doi:10.1016/j.tibtech.2009.05.006

Wang, J., L.W. Pembleton, R.C. Baillie, M.C. Drayton, M.L. Hand, M. Bain, et al. 2014. Development and implementation of a multiplexed single nucleotide polymorphism genotyping tool for differentiation of ryegrass species and cultivars. Mol. Breed. 33:435–451. doi:10.1007/s11032-013-9961-6

Wu, T., S. Luo, R. Wang, Y. Zhong, X. Xu, Y. Lin, X. He, B. Sun, H. Huang. 2014. The first Illumina-based de novo transcriptome sequencing and analysis of pumpkin (*Cucurbita moschata* Duch.) and SSR marker development. Mol. Breed. 34:1437–1447. doi:10.1007/s11032-014-0128-x

Yamada, T., J.W. Forster, M.W. Humphreys, and T. Takamizo. 2005. Genetics and molecular breeding in *Lolium/Festuca* grass species complex. Grassland Science 51:89–106. doi:10.1111/j.1744-697X.2005.00024.x

Zerbino, D.R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Conserv. Genet. Resour. 18:821–829.

Zwierzykowski, Z., A.J. Lukaszewski, A. Lesniewska, and B. Naganowska. 1998a. Genomic structure of androgenic progeny of pentaploid hybrids, *Festuca arundinacea* × *Lolium multiflorum*. Plant Breed. 117:457–462. doi:10.1111/j.1439-0523.1998.tb01973.x

Zwierzykowski, Z., R. Tayyar, M. Brunell, and A.J. Lukaszewski. 1998b. Genome recombination in intergeneric hybrids between tetraploid *Festuca pratensis* and *Lolium multiflorum*. J. Hered. 89:324–328. doi:10.1093/jhered/89.4.324