First assembly of the gene-space of Lolium multiflorum and comparison to other Poaceae genomes

Author(s):
Knorst, Verena; Yates, Steven; Byrne, Stephen; Asp, Torben; Widmer, Franco; Studer, Bruno; Kölliker, Roland

Publication date:
2019-04

Permanent link:
https://doi.org/10.3929/ethz-b-000309533

Originally published in:
Grassland Science 65(2), https://doi.org/10.1111/grs.12225

First assembly of the gene-space of *Lolium multiflorum* and comparison to other Poaceae genomes

Verena Knorst[1,2], Steven Yates[1], Stephen Byrne[3], Torben Asp[4], Franco Widmer[2], Bruno Studer[1], Roland Kölliker[1,2,]

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

[2] Molecular Ecology, Agroscope, 8046 Zurich, Switzerland

[3] Teagasc, Crops Science Department, Oak Park, R93 XE12 Carlow, Ireland

[4] Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, 4200 Slagelse, Denmark

Corresponding author:

Roland Kölliker

Molecular Plant Breeding

ETH Zürich c/o Agroscope

Reckenholzstr. 191

8046 Zürich, Switzerland

Phone: +41 58 468 7345, Fax: +41 58 468 7201

e-mail: roland.koelliker@usys.ethz.ch

Running title: Gene space of *Lolium multiflorum*

Abstract

Italian ryegrass (*Lolium multiflorum* Lam.) is one of the most important forage grasses of temperate regions. It is characterized by a high level of within cultivar diversity, making the fixation of traits in breeding programs difficult. The aim of this study was to develop a functional resource for *L. multiflorum* that will greatly assist the development of genomics-assisted breeding strategies.

The genome of a diploid, heterozygous genotype was sequenced and assembled into 130k scaffolds with a total assembly size of 574 Mb. 61k gene models were annotated and 80.5% of BUSCOs were complete, indicating a good representation of the gene content. Simple sequence repeat (SSR) mining identified 29,840 SSR loci for *L. multiflorum*, with 4,601 being shared with *L. perenne*. Proteome comparison with other Poaceae revealed several protein clusters including the SAUR-like auxin-responsive protein family, glutathione S-transferases, BTB-POZ and the photosystem II light harvesting complex gene B1B2 to be enriched in *Lolium* spp., while 86 clusters were more and 293 less abundant in *L. multiflorum* than in *L. perenne*.

Here we present the first sequence, assembly, and annotation of the gene-space of *L. multiflorum*, providing a valuable resource for marker identification for breeding programs, genome-wide association studies or evolutionary studies.

Keywords

Gene space; Lolium multiflorum; ryegrasses

Introduction

The ever-growing world population increasingly demands for meat and dairy products. Efficient and sustainable ways to produce these products are therefore required. Grassland-based livestock production offers a promising alternative to high-input confined feed-lot operations (O'Mara 2012). However, in order to meet the demand for high quality roughage from grassland, high yielding, persistent and disease resistant forage grasses are needed.

Italian ryegrass (*Lolium multiflorum*, Lam.) is one of the most important forage grasses, widely grown in temperate regions worldwide. It is particularly valued for a number of agronomic traits including fast establishment, good palatability and high yields (Humphreys *et al.* 2010). In addition, *L. multiflorum* may be used for phytoremediation or to monitor contaminations, due to its association with endophytic bacteria such as *Pseudomonas* and *Rhodococcus* spp. that are able to degrade alkanes (e.g. diesel; Andria *et al.* 2009, Yousaf *et al.* 2010).

L. multiflorum belongs to the tribe Poeae in the Poaceae grass family, which contains major crop species such as wheat (*Triticum aestivum)*, rice (*Oryza sativa*) and maize (*Zea mays*), as well as the model grass *Brachypodium distachyon* (Soreng *et al.* 2015). It is a diploid (2n=2x=14), allogamous species. The genome size is expected to be comparable to the one of its close relative *L. perenne* (~2.5E+09 base pairs [bp]) and also to contain a large amount of repetitive sequences (Byrne *et al.* 2015). As is characteristic for allogamous species, *Lolium* spp. display a very high level of heterozygosity (Byrne *et al.* 2015).

The genus *Lolium* contains different species with contrasting flowering patterns. While *L. perenne* is a perennial and only flowers once per growing season, *L. multiflorum* has the ability to produce flowers repeatedly throughout the season (i.e. after each cut). *L. multiflorum* can be further divided into the subspecies Westerwolds ryegrasses (*L. multiflorum* ssp. *westerwoldicum*), which is an annual species with high yield potential

but low persistency, and Italian ryegrass (*L. multiflorum* ssp. *italicum*), which is a bi- to pluriannual species with increased persistence (Humphreys *et al.* 2010).

Focused on its primary use in hay or silage production for ruminant nutrition, breeding targets for *L. multiflorum* include high dry matter yield (DMY), high nutritive value and high seed yield as well as resistance against diseases such as crown rust caused by *Puccinia coronata* f.sp. *lolii* and bacterial wilt caused by *Xanthomonas translucens* pv. *graminis* (Humphreys *et al.* 2010). Breeding of forage grasses mainly builds on recurrent phenotypic selection, based on either population improvement or the production of synthetic progeny using intercrosses of a limited number of parental plants (Posselt 2010). Consequently, *L. multiflorum* cultivars usually consist of many different genotypes and are characterized by a high level of within cultivar diversity. This may be advantageous in terms of adaptability across a broad range of environments, but may also impair breeding progress, as fixation of desired traits is often difficult.

Molecular genetic and genomic tools have been shown to valuably complement phenotypic selection, allowing for developing highly efficient and targeted breeding strategies (Xu *et al.* 2017). Although forage grasses have not received as much attention as major crops such as rice and wheat, where high quality reference genomes have become available, a number of genomic resources have been developed for *L. multiflorum*. For example, a number of high quality linkage maps have been constructed using predominantly AFLP (Studer *et al.* 2006), SSR (Hirata *et al.* 2006) or DArT (Bartoš *et al.* 2011) markers. Traits for which linked genetic markers have been identified in *L. multiflorum* and closely related species include disease resistance (Studer *et al.* 2007, Studer *et al.* 2006), lodging tolerance (Inoue *et al.* 2004), seed yield (Studer *et al.* 2008) or drought tolerance (Humphreys *et al.* 2005). In addition, genomic resources for *L. perenne* include chloroplast sequences (Diekmann *et al.* 2009), mitochondrial sequences (Islam *et al.* 2013), a transcriptional map (Studer *et al.* 2012) and a draft

genome sequence consisting of 1,128 Mb and 48,415 scaffolds with a minimum size of 1 kb with a total of 28,455 genes encoding 40,068 proteins (Byrne *et al.* 2015).

Genomic research in *Lolium* spp. has substantially profited from genetic resources of wheat (Brenchley *et al.* 2012), maize (Schnable *et al.* 2009), rice (Goff *et al.* 2002) or *Brachypodium* (International Brachypodium Initiative 2010), based on the high synteny and collinearity of Poaceae genomes (Jones *et al.* 2002). For example, using the GenomeZipper (Pfeifer *et al.* 2013), candidate genes for map-based cloning of QTL can be identified in *Lolium* spp. based on synteny to the available reference genomes. However, the availability of a high quality reference genome for the species under study is indispensable for the efficient development and implementation of genomics-assisted breeding strategies. This is particularly true for genome wide association mapping (Kole *et al.* 2015) or genomic prediction (Grinberg *et al.* 2016), which no longer rely on the correlation between single markers and traits but on a large number of genetic loci covering the entire genome. Genome sequences can also facilitate SNP mapping in genotyping by sequencing (GBS) studies (Poland and Rife 2012), provide information of candidate genes when targeting induced local lesions in genomes (TILLING; Till *et al.* 2003) or enable the identification of a large number of genetic markers such as SSRs (Perez-de-Castro *et al.* 2012). Moreover, genome editing methods like the CRISPR/Cas9 system are dependent on sequence information provided by reference genomes (Cong *et al.* 2013).

Therefore, the aims of this study were to sequence, assemble, and annotate the gene-space of *L. multiflorum.* This will assist the development of both marker and genome-wide assisted breeding strategies.

Materials and Methods

DNA preparation and sequencing

For establishing the first draft genome sequence of *L. multiflorum*, we selected a highly heterozygous genotype from advanced breeding germplasm, which was also used as a parental plant in a mapping population segregating for disease resistance (M2289; Studer *et al.* 2006). DNA was extracted from freeze-dried leave material using the Qiagen DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) and quality and quantity were assessed by spectrophotometry (Nanodrop; Thermo Fisher Scientific, Waltham, Massachusetts, USA) and through gel electrophoresis.

The Illumina HiSeq2000 platform (Illumina Inc., San Diego, USA) was used for sequencing. Two Illumina paired-end libraries with mean fragment lengths of 300bp and 800 bp were prepared from genomic DNA, using the NEBNext DNA sample preparation kit (New England Biolabs, https://www.neb.com/) with Illumina adaptors according to the NEBnext instructions. The genomic DNA was sequenced on two lanes.

Genome assembly

For the genome assembly, the program ABySS (Version: 1.3.5; Simpson *et al.* 2009) was used with a *k*-mer size of 55 bp, raw data from sequencing both libraries (300 and 800 bp), a minimum base quality of 30 and a minimum contig size of 300bp to build a scaffold. BUSCO V3 was used to generate a quantitative measure of genome completeness (Simão *et al.* 2015) with the Embryophyta odb9 lineage (creation date 2016-02-13, number of species 30, number of BUSCOS: 1440). Since organelle DNA was not removed prior to library construction, organelle sequences were assembled into contiguous blocks which did not interfere with downstream analysis.

SSR mining

SSRs were identified in the *L. multiflorum* genome and in the *L. perenne* genome of Byrne et al. (2015) using the MIcroSAtellite (MISA) software tool (Thiel *et al.* 2003) and considering SSR loci with repeat motifs of 2bp or more. Conserved SSRs between both species were identified by selecting 200 bp regions up- and downstream of the SSR motif. A BLAST database was then created from the *L. multiflorum* SSR loci (as this species contained fewer SSRs than *L. perenne*), which the *L. perenne* SSR loci were searched against using BLASTN. Based on BLASTN, loci with < 299 bp alignment were rejected, and only SSR loci with a single significant BLAST match were retained. The data from MISA were then merged with the remaining BLASTN pairs. Using the BLASTN results (start and end values) the orientation of the sequence was then corrected and the SSR motif was reverse complemented, if applicable. Additionally we observed that two SSR motifs might differ but potentially be the same (for example TC and CT), for this reason, we moved the last base to the first position. From these data, we then considered pairs of SSR loci to be informative if the SSR motif matched (exactly with either the original or the re-positioned last base) and the number of repeats differed. For these analyses, custom Perl and R scripts were used.

Genome annotation

To identify transcribed regions of the genome and corresponding functional coding DNA sequences (CDS), the following RNA-seq datasets were used: (i) data from six different tissues from *L. perenne* (Bioproject: PRJNA222646; Farrell *et al.* 2014); (ii) five *L. multiflorum* datasets from meristem samples (SRR3100250-4; Stoces *et al.* 2016); (iii) data from pollen and stigma samples from *L. perenne* (Manzanares *et al.* 2016); and (iv) an in-house data set of 48 *L. perenne* meristem samples, taken at three time-points (data unpublished). The reads were aligned to the transcriptome using Tophat (version 2.0.11) and Bowtie2 (version: 2.1.0; Trapnell *et al.* 2012, Langmead 2010) for all samples. Isoforms of genes were identified using Cufflinks (version: 2.2.0; Trapnell *et al.* 2012)

179 producing a genomic feature format file (GFF). The individual GFF files were then
180 merged using the cuffmerge command, default settings.
181
182 *Coding sequence identification*
183 To identify coding sequences (CDS), the spliced exons for each GFF transcript were
184 retrieved using gffread (part of the Tuxedo tool suite). To identify the correct open reading
185 frames (ORF) for protein sequences the program ORFpredictor (version: 3.0; Min *et al.*
186 2005) was used. For frame selection, the transcripts were first BLASTX (Altschul *et al.*
187 1990) searched against a protein database consisting of the proteomes from *Arabidopsis*
188 *thaliana* (TAIR version 10; Swarbreck *et al.* 2008), *O. sativa* (downloaded from Ensembl;
189 Kersey *et al.* 2016), *Glycine max* (Ensembl), *Populus trichocarpa* (Ensembl) and *Manihot*
190 *esculenta* (cassava, v4.1; Prochnik *et al.* 2012). This database, although not exhaustive,
191 provided a broad basis of existing plant proteins. ORFpredictor was then used to identify
192 CDS by use of the best BLAST hits frame selection. In the absence of a homologous
193 BLAST hit, ORFpredictor selected the longest ORF. These results were then used to
194 annotate the GFF file created by Cufflinks for CDS using scripts kindly provided by
195 Palmieri *et al*. (2012).
196
197 *Gene annotation*
198 For functional annotation of genes, three synergistic methods were employed, based on
199 protein sequences. First, the protein sequences were search against the *A. thaliana*
200 TAIR10 proteome using BLASTP. Second, the proteins were search against the Swiss-
201 Prot non-redundant protein database (http://www.uniprot.org/downloads; downloaded
202 14/03/2016), again using BLASTP. In both cases, the functional annotation of the best
203 BLAST hit (based on E-value, maximum 1e-15) protein was used to assign annotations
204 for functional description and gene ontology (GO); from Swiss-Prot an InterPro domain
205 was also assigned where possible. In the third step, the protein sequences were scanned

against InterPro's signatures using InterProScan (version: 5.16-55; Jones *et al.* 2014). From this, a number of assignments could be made including HAMAP (High-quality Automated and Manual Annotation of Proteins; Pedruzzi *et al.* 2015), Pfam (Finn *et al.* 2016) and PIRSF (Protein Information Resource Super Family; Nikolskaya *et al.* 2007). For the aforementioned, the corresponding GO annotation was also retrieved from http://geneontology.org/external2go/ (downloaded 27/06/2016). The three sources of annotation were then combined into a single table and the GO terms from each were concatenated into a non-redundant list.

Protein clustering

To identify orthologous clusters of proteins in the *Lolium* genus and in wheat (*T.aestivum*, TGACv1), maize, (*Z. mays*, AGPv4), *Brachypodium* (*B. distachyon*, v1.0) and rice (*O. sativa*, IRGSP-1.0) the proteomes of these grasses were compared. First, both *Lolium* proteomes were merged *in silico* to produce a single representative proteome of the *Lolium* genus. Additionally, the longest amino acid sequence (aa) per gene locus was retained using an in-house Perl script. For the other grasses, the proteomes were downloaded from Plant Ensembl FTP (ftp://ftp.ensemblgenomes.org/) and the longest protein per gene loci was retained.

To cluster the protein sequences into orthologous clusters, the offline version of OrthoMCL (Li *et al.* 2003) was used, as described by Sykes et al. (2017). Briefly, the protein names within a fasta file (per species) were first adapted for consistency and to eliminate problems arising from special characters and name similarities. The resulting fasta file was then formatted to make it compliant with the OrthoMCL algorithm (a short species-specific prefix was added to each name for subsequent species identification). The sequences were then filtered for low quality, based on sequence length (>30 aa) and percentage of stop codons (<10%). From these high quality proteins, an all-vs-all BLASTP was run where all proteins were searched against all proteins (maximum E-

value 1e-5). The results of the BLASTP were collated and then parsed before loading into a local MySQL orthoMCL database. In the next step, pairs of proteins that are potentially orthologs, in-paralogs or co-orthologs were identified using the OrthoMCL algorithm (Li *et al.* 2003), where protein pairwise connections were normalised for ortholog pairs between and within species. The resulting potential pairs were then organized in clusters using the MCL algorithm (Enright *et al.* 2002) and the names were reverted to their original values for subsequent work. Protein clusters were counted for *L. perenne* and *L. multiflorum* separately, and protein clusters were defined as enriched for one species if they were not present in the other species or if the cluster had a +/- 2.5x occurrence after log2 transformation.

To identify protein clusters unique to the *Lolium* genus, we then selected orthologous clusters from OrthoMCL, which only included *Lolium* proteins. Furthermore, we selected only clusters, which contained both *L. perenne and L. multiflorum* proteins. For each cluster, we then extracted the protein names within it and looked for enriched groups by using the GO information previously described. For this, the annotations of *L. perenne and L. multiflorum* were concatenated into a single file. To identify enriched groups of proteins, based on function, gene ontology (GO) analysis was implemented using TopGO (Alexa and Rahnenfuhrer 2016) in R (R Core Team 2016). A one-sided Fisher`s exact test was used to identify enriched GO terms with a minimum $P < 0.05$ and a minimum of five genes found with the GO term. Go terms were then summarized using a cluster algorithm relying on semantic similarity and visualized using the Revigo online tool (Supek *et al.* 2011).

Results and Discussion

Italian ryegrass genome assembly

The raw data consisted of 6.89E+08 100 base pair (bp) reads. Assuming a genome size of approximately 2.5E+09 bp (as it was observed for the close relative L. perenne; Byrne *et al.* 2015), this corresponds to a 28x genome coverage by the raw reads. The reads were quality filtered and assembled into 129,579 scaffolds of a minimum length of 2kb. A total of 574 Mb were included in this assembly with an N50 scaffold length of 4,949 bp. Although this represents a draft genome suitable for many downstream applications, the overall quality of the draft genome was lower when compared to the *L. perenne* assembly, where a 1.13 Gb assembly was produced consisting of 48,415 scaffolds with an N50 of 70'062 bp. However, this was generated using a considerably larger sequencing effort including multiple paired-end and mate-pair libraries in addition to PacBio sequencing (Byrne *et al.* 2015). In addition, the *L. perenne* genome assembly was based on a highly homozygous genotype derived from multiple generations of inbreeding. We chose a highly heterozygous *L. multiflorum* for our assembly for two reasons: first, *L. multiflorum* suffers from severe inbreeding depression and advanced inbred lines are not available for the species. Second, the plant was used as resistance donor in a bi-parental cross used extensively for the characterization of disease resistance in *L. multiflorum* (Rechsteiner *et al.* 2006, Studer *et al.* 2006, Wichmann *et al.* 2011, Studer *et al.* 2007). Thus, the draft assembly presented here will greatly facilitate the discovery of candidate resistance genes in *L. multiflorum*. The assembly provides a good representation of the gene space. This was demonstrated by the BUSCO analysis using the Embyrophyta lineage, which provides a quantitative assessment of the completeness of the genome in terms of the expected gene content. Out of 1,440 BUSCO groups searched, 80.5% were complete (59.7% single-copy BUSCOs, 20.8% duplicated BUSCOs), 8.5% were fragmented and 11% were missing.

Mining for SSR markers characteristic for *Lolium* ssp.

The development of high-throughput genotyping techniques based on next generation sequencing such as genotyping by sequencing (GBS; Elshire *et al.* 2011) now allows to easily generate a large number of data points in a short time and with moderate effort. However, sequence specific multiallelic markers such as SSRs still represent the markers of choice for specific applications such as routine screening of a limited number of loci in large germplasm collections or diversity analyses in diverse genetic backgrounds. Therefore, we scanned the *L. multiflorum* and the *L. perenne* (Byrne *et al.* 2015) genomes for SSR motifs. In total, 29,840 and 46,780 SSRs (with repeat motifs of 2 bp or larger) in both the *L. multiflorum* and the *L. perenne* genomes were found, respectively (Table 1). Of these, 88 to 92 % were di- and tri-nucleotide repeats. Based on BLAST homology, we found 4076 SSRs common in both species (Table 1). Among these, 1,171 and 818 repeats were di- or tri-nucleotide repeats, respectively, which differed in repeat number between the two species. The relatively low number of common SSRs identified was mainly due to the conservative approach used for identification. The common loci provide valuable tools to study common characteristics of the two species, while the remaining SSRs build a resource to develop species specific SSRs for species identification.

Annotation of the *L. multiflorum* draft genome

For annotation of gene models we chose to identify transcribed regions, using experimental data derived from existing mRNA sequencing projects (for details see material and methods). In total, 61k gene models were identified using the Tuxedo suite of tools (Table 2).

Of the 61k gene models, 60k had an open reading frame and a total of 80k splice variants (minimum length10 amino acids [aa]), with a mean length of 176 aa were identified. We then annotated the gene models for functional description using the predicted coding sequences using BLASTP based methods against the TAIR10 proteome and the Swiss-

Prot non-redundant protein database. In addition, we used motif scanning based on InterPro signatures. From these data, a functional description was assigned to 35K transcripts, of which 30K were assigned at least one Gene Ontology (GO) term.

To compare the results found for *L. multiflorum* with *L. perenne*, we implemented the same protocol to annotate the *L. perenne* genome of Byrne et al. (Byrne *et al.* 2015). In *L. perenne*, we found 55k gene models with 102k splice variants, with mean length of 1,577bp, N50 values of 2,082bp. Using the methods described above for open reading frame identification, we found a corresponding open reading frame in 54k gene models and in 101k splice variants; with mean and N50 of 247 and 273, respectively.

A similar number of gene models was found in both species: 61k for *L. multiflorum* and 55k for *L. perenne*. From this, we conclude that most of the gene models from *L. perenne* were present in the *L. multiflorum* draft assembly, although their length, on average, was shorter.

Despite the clearly larger number of scaffolds and a lower N50 value for the *L. multiflorum* draft genome assembly, we were able to identify a large number of the *L. perenne* genes in *L. multiflorum*. Thus, the dataset will be valuable for downstream applications like identification of candidate genes or development of markers for marker-assisted breeding programs.

Protein clusters enriched in the genus *Lolium*

The two *Lolium* proteomes (from *L. perenne* and *L. multiflorum*) generated in the previous step were merged *in silico* to produce a single representative proteome of the *Lolium* genus. This dataset was compared to the proteomes of wheat, maize, *Brachypodium* and rice to identify orthologous protein clusters enriched in the genus *Lolium*.

In total, 315,182 proteins of these five genomes were used for clustering with the OrthoMCL algorithm to identify orthologous protein clusters. Of these, 223,756 proteins

were clustered into 35,739 orthologous clusters. Unsurprisingly, most of the genes (54%: 121,349) were common between the five species, within 11,694 orthologous clusters (Figure 1). The investigated species shared a core set of 121,349 genes related to a number of common characteristics and general metabolic processes, which is in line with previous studies (Davidson *et al.* 2012). In this core set, genes for photosynthesis or general metabolic processes were included.

The *Lolium* genus was characterized by the highest number of unique proteins (19,335), clustered into 7611 orthologous protein clusters. *Brachypodium* was characterized by the lowest number of only 600 unique orthologous protein clusters (Figure 1). On the other hand, *Brachypodium*, rice wheat and maize shared 12,076 genes, which were absent in the combined proteome of the two *Lolium* spp. This may be partially explained by the inferior quality of the genome assemblies of the two *Lolium* spp. when compared to the other species. More complete *Lolium* spp. genomes and multiple reference assemblies would be required for more detailed comparison of gene contents among the different species.

Moreover, while the other crops are annual species primarily selected for grain yield, *Lolium* spp. are annual, bi-annual or perennial species primarily grown and selected for biomass production. This could have favoured the selection of different gene families. From the 7611 orthologous clusters specific to the *Lolium* genus, only those present in both species (i.e. *L. perenne* and *L. multiflorum*) were retained, resulting in 5171 orthologous clusters. Gene ontology (GO) enrichment in these clusters was tested using Fisher`s exact test and the complexity of the data was reduced using semantic word space and the Revigo webtool (Figure 2). Eleven GO terms were found to be significantly enriched in *Lolium* ssp. and to contain at least five protein families. Four protein families seem to have expanded in the *Lolium* genus in comparison to the other four Poaceae genomes, including BTB-POZ, the SAUR-like auxin-responsive protein family, glutathione S-transferases, and the photosystem II light harvesting complex gene B1B2.

366 The BTB-POZ protein cluster contains proteins, which are part of zinc-finger proteins.
367 These proteins serve a variety of functions by interacting with DNA. For example, the
368 disease resistance protein NPR1 of Arabidopsis contains a BTB/POZ domain, which
369 interacts with the repression domain of TGA2 (Boyle *et al.* 2009). SAUR-like auxin
370 responsive genes may be involved in disease resistance against *X. arboricola* pv. *pruni*
371 in peach (Socquet-Juglard *et al.* 2013), but they have also been shown to be involved in
372 root development in Arabidopsis (Markakis *et al.* 2013). Species-specific pathogens,
373 together with a particular root development in perennial species may partly explain the
374 enrichment of these gene clusters in *Lolium* spp. Glutathione S-transferases have been
375 shown to confer resistance against herbicides in *L. multiflorum* and *L. rigidum* (Del Buono
376 and Ioli 2011, Cummins *et al.* 2013). The enrichment of this protein cluster may partly
377 explain the high level of herbicide resistance often observed in *Lolium* spp. (Mahmood
378 *et al.* 2016, Han *et al.* 2016). Finally, the photosystem II light harvesting complex B1B2
379 proteins are a central part of the chloroplast membrane and serve in harvesting energy
380 from sunlight and transferring it to downstream proteins (Kuhlbrandt *et al.* 1994). *Lolium*
381 spp. are mostly grown as perennial plants in temperate regions and therefore more
382 heavily exposed to varying day lengths than crop species such as wheat, maize or rice,
383 which are only grown for one growing season. *Lolium* spp. may therefore depend more
384 heavily on an efficient light harvesting system to make use of short days in early spring
385 and late autumn.
386
387 Unique properties of the *L. multiflorum* and *L. perenne* proteome
388 To identify unique properties of the *L. multiflorum* and the *L. perenne* proteome, the
389 number of gene models in each protein cluster were compared. Only 86 protein clusters
390 were found to be larger in *L. multiflorum* when compared *to L. perenne,* and 293 were
391 larger in *L. perenne* when compared to *L. multiflorum*. Most of the corresponding gene
392 models contained no functional annotation and these were omitted from further analysis.

In *L. multiflorum*, 31 gene models did contain annotation and for about half of these, gene models were assigned a corresponding TAIR 10 based annotation (Supplementary Table 1). In *L. perenne*, 125 gene models could be identified and 84 of those contained an annotation (Supplementary Table. 2). These annotations were grouped according to the assigned functional groups based on the TAIR10 database (Tab. 3).

The list was summarized into nine categories. The category with the most entries was "general biological processes" where no specific function could be assigned. Second was "stress response", which included all proteins involved in disease resistance and abiotic stress response. The group "protein modification" mainly contained proteins related to protein phosphorylation. The groups "hormones" and "reproduction" showed the lowest level of enrichment.

Both genomes show significant gene expansion. For half of those genes an annotation was available because they showed homologies to know proteins. Several disease resistance proteins found to be enriched in *L. perenne*, such as OC_22035 (CC-NBS-LRR class), OC_258 (CC-NBS-LRR class), OC_369 (PR5-like receptor kinase) or OC_9754 (dirigent-like protein). The proteins found were described as proteins generally involved in disease resistance (Tan *et al.* 2007, Wang *et al.* 1996, Ralph *et al.* 2007).

Interestingly, a high number of enriched clusters contain genes involved in metal ion binding like OC_1089 in *L. perenne* that is annotated as "copper transport protein family" or OC_1619 in *L. multiflorum* annotated as "heavy metal transport/detoxification superfamily protein". These may be particularly important for the phytoremediation capacities shown for *L. multiflorum* is (Mugica-Alvarez *et al.* 2015).

Among the proteins enriched in *L. perenne*, one cluster contained genes annotated as an AGAMOUS-like gene (Supplementary table 2). The protein AGAMOUS-like 19 (AGL19) is a transcription factor shown to be involved in flowering in *Arabidopsis*, especially in the Flowering Locus C (FLC)-independent vernalisation pathway which does not require VRN2 (Schonrock *et al.* 2006). In *L. perenne,* a number of AGAMOUS-

like genes such as AGL24 were reported to be involved in flowering (Ciannamea *et al.* 2006), but their role and interaction with other *L. perenne* flowering genes such as VRN1 or VRN2 (Andersen *et al.* 2006) is not well established. A main difference between *L. multiflorum* and *L. perenne* is their flowering behaviour with *L. multiflorum* being able to flower throughout the season while *L. perenne* only flowers once. Therefore, the AGAMOUS-like genes could play a role in the particular flowering pattern of *L. perenne*. Interestingly, in *L. perenne* the two protein clusters OC_5274 (BTB-POZ) and OC_5278 (glutathione s-transferase) were found to be enriched as they were already found in the comparison of the *Lolium* proteomes with other Poaceae proteomes. Most differences observed concern protein clusters involved in basic processes and may concern general adaptation rather than specific evolutionary responses to particular selection pressures through biotic or abiotic factors. However, in general the differences observed concerned only relatively few gene clusters, which emphasizes the close relationship between the two species.

Conclusion

Here we present the first draft assembly of the gene space of the forage grass *L. multiflorum.* This is a valuable resource for the development of molecular genetic tools to be used in breeding programs, as a basis for extended comparative genome studies in the important family of Poaceae.

Acknowledgments

We thank Stephan Hentrup, Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, 4200 Slagelse, Denmark for technical support. This work was supported through grant no: 31003A_138358 of the Swiss National Science Foundation.

References

Alexa A, Rahnenfuhrer J (2016) topGo: enrichment analysis for gene ontology. *R package version 2.32.0*.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *J Mol Biol* 215:403-410.

Andersen J, Jensen L, Asp T, Lübberstedt T (2006) Vernalization response in perennial ryegrass (*Lolium perenne* L.) involves orthologues of diploid wheat (*Triticum monococcum*) *VRN1* and Rice (*Oryza sativa*) *Hd1*. *Plant Mol Biol* 60:481.

Andria V, Reichenauer TG, Sessitsch A (2009) Expression of alkane monooxygenase (alkB) genes by plant-associated bacteria in the rhizosphere and endosphere of Italian ryegrass (*Lolium multiflorum* L.) grown in diesel contaminated soil. *Environ Pollut* 157:3347-3350.

Bartoš J, Sandve S, Kölliker R, et al. (2011) Genetic mapping of DArT markers in the *Festuca–Lolium* complex and their use in freezing tolerance association analysis. *Theor Appl Genet* 122:1-15.

Boyle P, Le Su E, Rochon A, et al. (2009) The BTB/POZ domain of the *Arabidopsis* disease resistance protein NPR1 interacts with the repression domain of TGA2 to negate its function. *Plant Cell* 21:3700-3713.

Brenchley R, Spannagl M, Pfeifer M, et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705-710.

Byrne SL, Nagy I, Pfeifer M, et al. (2015) A synteny-based draft genome sequence of the forage grass *Lolium perenne*. *Plant J* 84:816-826.

Ciannamea S, Kaufmann K, Frau M, et al. (2006) Protein interactions of MADS box transcription factors involved in flowering in Lolium perenne. *J Exp Bot* 57:3419-3431.

Cong L, Ran FA, Cox D, et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339:819-823.

474 Cummins I, Wortley DJ, Sabbadin F, et al. (2013) Key role for a glutathione transferase
475 in multiple-herbicide resistance in grass weeds. *Proc Natl Acad Sci U S A*
476 110:5812-5817.
477 Davidson RM, Gowda M, Moghe G, et al. (2012) Comparative transcriptomics of three
478 *Poaceae* species reveals patterns of gene expression evolution. *Plant J* 71:492-
479 502.
480 Del Buono D, Ioli G (2011) Glutathione S-transferases of italian ryegrass (*Lolium*
481 *multiflorum*): activity toward some chemicals, safener modulation and
482 persistence of atrazine and fluorodifen in the shoots. *J Agric Food Chem*
483 59:1324-1329.
484 Diekmann K, Hodkinson TR, Wolfe KH, Van Den Bekerom R, Dix PJ, Barth S (2009)
485 Complete chloroplast genome sequence of a major allogamous forage species,
486 perennial ryegrass (*Lolium perenne* L.). *DNA Res* 16:165-176.
487 Elshire RJ, Glaubitz JC, Sun Q, et al. (2011) A robust, simple genotyping-by-
488 sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379.
489 Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale
490 detection of protein families. *Nucleic Acids Res* 30:1575-1584.
491 Farrell JD, Byrne S, Paina C, Asp T (2014) De novo assembly of the perennial
492 ryegrass transcriptome using an RNA-Seq strategy. *PLoS One* 9:e103567.
493 Finn RD, Coggill P, Eberhardt RY, et al. (2016) The Pfam protein families database:
494 towards a more sustainable future. *Nucleic Acids Res* 44:D279-285.
495 Goff SA, Ricke D, Lan TH, et al. (2002) A draft sequence of the rice genome *(Oryza*
496 *sativa* L. ssp. *japonica*). *Science* 296:92-100.
497 Grinberg NF, Lovatt A, Hegarty M, et al. (2016) Implementation of Genomic Prediction
498 in Lolium perenne (L.) Breeding Populations. *Front Plant Sci* 7:133.

499 Han H, Yu Q, Owen MJ, Cawthray GR, Powles SB (2016) Widespread occurrence of
500 both metabolic and target-site herbicide resistance mechanisms in *Lolium*
501 *rigidum* populations. *Pest Manag Sci* 72:255-263.
502 Hirata M, Cai H, Inoue M, et al. (2006) Development of simple sequence repeat (SSR)
503 markers and construction of an SSR-based linkage map in Italian ryegrass
504 (*Lolium multiflorum* Lam.). *Theor Appl Genet* 113:270-279.
505 Humphreys J, Harper JA, Armstead IP, Humphreys MW (2005) Introgression-mapping
506 of genes for drought resistance transferred from *Festuca arundinacea* var.
507 *glaucescens* into *Lolium multiflorum*. *Theor Appl Genet* 110:579-587.
508 Humphreys M, Feuerstein U, Vandevalle M, Baert J (2010) Ryegrasses. In: *Fodder*
509 *Crops and Amenity Grasses.* (Eds Boller B, Posselt UK & Veronesi F), Springer
510 Science + Business Media, New York, 211-260.
511 Inoue M, Gao Z, Hirata M, Fujimori M, Cai H (2004) Construction of a high-density
512 linkage map of Italian ryegrass (*Lolium multiflorum* Lam.) using restriction
513 fragment length polymorphism, amplified fragment length polymorphism, and
514 telomeric repeat associated sequence markers. *Genome* 47:57-65.
515 International Brachypodium Initiative (2010) Genome sequencing and analysis of the
516 model grass *Brachypodium distachyon*. *Nature* 463:763-768.
517 Islam MS, Studer B, Byrne SL, et al. (2013) The genome and transcriptome of
518 perennial ryegrass mitochondria. *BMC Genomics* 14:1-21.
519 Jones ES, Mahoney NL, Hayward MD, et al. (2002) An enhanced molecular marker
520 based genetic map of perennial ryegrass (*Lolium perenne*) reveals comparative
521 relationships with other *Poaceae* genomes. *Genome* 45:282-295.
522 Jones P, Binns D, Chang HY, et al. (2014) InterProScan 5: genome-scale protein
523 function classification. *Bioinformatics* 30:1236-1240.
524 Kersey PJ, Allen JE, Armean I, et al. (2016) Ensembl Genomes 2016: more genomes,
525 more complexity. *Nucleic Acids Res* 44:D574-580.

526 Kole C, Muthamilarasan M, Henry R, et al. (2015) Application of genomics-assisted
527 breeding for generation of climate resilient crops: progress and prospects. *Front*
528 *Plant Sci* 6:563.
529 Kuhlbrandt W, Wang DN, Fujiyoshi Y (1994) Atomic model of plant light-harvesting
530 complex by electron crystallography. *Nature* 367:614-621.
531 Langmead B (2010) Aligning short sequencing reads with Bowtie. *Curr Protoc*
532 *Bioinformatics* Chapter 11:Unit 11.17.
533 Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for
534 eukaryotic genomes. *Genome Res* 13:2178-2189.
535 Mahmood K, Mathiassen SK, Kristensen M, Kudsk P (2016) Multiple herbicide
536 resistance in *Lolium multiflorum* and identification of conserved regulatory
537 elements of herbicide resistance genes. *Front Plant Sci* 7:1160.
538 Manzanares C, Barth S, Thorogood D, et al. (2016) A gene encoding a duf247 domain
539 protein cosegregates with the s self-incompatibility locus in perennial ryegrass.
540 *Mol Biol Evol* 33:870-884.
541 Markakis MN, Boron AK, Van Loock B, et al. (2013) Characterization of a small auxin-
542 up RNA (SAUR)-like gene involved in *Arabidopsis thaliana* development. *PLoS*
543 *One* 8:e82596.
544 Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding
545 regions in EST-derived sequences. *Nucleic Acids Res* 33:W677-680.
546 Mugica-Alvarez V, Cortés-Jiménez V, Vaca-Mier M, Domínguez-Soria V (2015)
547 Phytoremediation of mine tailings using *Lolium multiflorum*. *Int J Environ Sci*
548 *Dev* 6:246.
549 Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH (2007) PIRSF family
550 classification system for protein functional and evolutionary analysis. *Evol*
551 *Bioinform Online* 2:197-209.

552 O'mara FP (2012) The role of grasslands in food security and climate change. *Ann Bot*
553 110:1263-1270.
554 Palmieri N, Nolte V, Suvorov A, Kosiol C, Schlotterer C (2012) Evaluation of different
555 reference based annotation strategies using RNA-Seq - a case study in
556 *Drososphila pseudoobscura*. *PLoS One* 7:e46415.
557 Pedruzzi I, Rivoire C, Auchincloss AH, et al. (2015) HAMAP in 2015: updates to the
558 protein family classification and annotation system. *Nucleic Acids Res*
559 43:D1064-1070.
560 Perez-De-Castro AM, Vilanova S, Canizares J, et al. (2012) Application of genomic
561 tools in plant breeding. *Curr Genomics* 13:179-195.
562 Pfeifer M, Martis M, Asp T, et al. (2013) The perennial ryegrass GenomeZipper:
563 targeted use of genome resources for comparative grass genomics. *Plant*
564 *Physiol* 161:571-582.
565 Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics.
566 *Plant Genome* 5:92-102.
567 Posselt U (2010) Breeding methods in cross-pollinated species. In: *Fodder Crops and*
568 *Amenity Grasses.* (Eds Boller B, Posselt U & Veronesi F), Springer Science +
569 Business Media, New York, 39-88.
570 Prochnik S, Marri PR, Desany B, et al. (2012) The Cassava Genome: Current
571 Progress, Future Directions. *Trop Plant Biol* 5:88-94.
572 R Core Team (2016) R: A language and environment for statistical computing. R
573 Foundation for Statistical Computing. Vienna, Austria.
574 Ralph SG, Jancsik S, Bohlmann J (2007) Dirigent proteins in conifer defense II:
575 Extended gene discovery, phylogeny, and constitutive and stress-induced gene
576 expression in spruce *(Picea* spp.). *Phytochemistry* 68:1975-1991.

577 Rechsteiner MP, Widmer F, Kölliker R (2006) Expression profiling of Italian ryegrass
578 (*Lolium multiflorum* Lam.) during infection with the bacterial wilt inducing
579 pathogen *Xanthomonas translucens* pv. *graminis*. *Plant Breed* 125:43-51.
580 Schnable PS, Ware D, Fulton RS, et al. (2009) The B73 maize genome: complexity,
581 diversity, and dynamics. *Science* 326:1112-1115.
582 Schonrock N, Bouveret R, Leroy O, et al. (2006) Polycomb-group proteins repress the
583 floral activator AGL19 in the FLC-independent vernalization pathway. *Genes*
584 *Dev* 20:1667-1678.
585 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO:
586 assessing genome assembly and annotation completeness with single-copy
587 orthologs. *Bioinformatics* 31:3210-3212.
588 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a
589 parallel assembler for short read sequence data. *Genome Res* 19:1117-1123.
590 Socquet-Juglard D, Kamber T, Pothier JF, et al. (2013) Comparative RNA-seq analysis
591 of early-infected peach leaves by the invasive phytopathogen *Xanthomonas*
592 *arboricola* pv. *pruni*. *PLoS One* 8:e54196.
593 Soreng RJ, Peterson PM, Romaschenko K, et al. (2015) A worldwide phylogenetic
594 classification of the Poaceae (Gramineae). *J Syst Evol* 53:117-137.
595 Stoces S, Ruttink T, Bartos J, et al. (2016) Orthology guided transcriptome assembly of
596 italian ryegrass and meadow fescue for single-nucleotide polymorphism
597 discovery. *Plant Genome* 9.
598 Studer B, Boller B, Bauer E, Posselt UK, Widmer F, Kölliker R (2007) Consistent
599 detection of QTLs for crown rust resistance in Italian ryegrass (*Lolium*
600 *multiflorum* Lam.) across environments and phenotyping methods. *Theor Appl*
601 *Genet* 115:9-17.

602 Studer B, Boller B, Herrmann D, et al. (2006) Genetic mapping reveals a single major
603 QTL for bacterial wilt resistance in Italian ryegrass (*Lolium multiflorum* Lam.).
604 *Theor Appl Genet* 113:661-671.
605 Studer B, Byrne S, Nielsen R, et al. (2012) A transcriptome map of perennial ryegrass
606 (*Lolium perenne* L.). *BMC Genomics* 13:140.
607 Studer B, Jensen LB, Hentrup S, Brazauskas G, Kölliker R, Lubberstedt T (2008)
608 Genetic characterisation of seed yield and fertility traits in perennial ryegrass
609 (*Lolium perenne* L.). *Theor Appl Genet* 117:781-791.
610 Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes
611 long lists of gene ontology terms. *PLoS One* 6:e21800.
612 Swarbreck D, Wilks C, Lamesch P, et al. (2008) The Arabidopsis Information Resource
613 (TAIR): gene structure and function annotation. *Nucleic Acids Res* 36:D1009-
614 1014.
615 Sykes T, Yates S, Nagy I, Asp T, Small I, Studer B (2017) In silico identification of
616 candidate genes for fertility restoration in cytoplasmic male sterile perennial
617 ryegrass (*Lolium perenne* L.). *Genome Biol Evol* 9:351-362.
618 Tan X, Meyers BC, Kozik A, et al. (2007) Global expression analysis of nucleotide
619 binding site-leucine rich repeat-encoding and related genes in Arabidopsis.
620 *BMC Plant Biol* 7:56.
621 Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the
622 development and characterization of gene-derived SSR-markers in barley
623 (Hordeum vulgare L.). *Theor Appl Genet* 106:411-422.
624 Till BJ, Reynolds SH, Greene EA, et al. (2003) Large-scale discovery of induced point
625 mutations with high-throughput TILLING. *Genome Res* 13:524-530.
626 Trapnell C, Roberts A, Goff L, et al. (2012) Differential gene and transcript expression
627 analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*
628 7:562-578.

629 Wang X, Zafian P, Choudhary M, Lawton M (1996) The PR5K receptor protein kinase
630 from *Arabidopsis thaliana* is structurally related to a family of plant defense
631 proteins. *Proc Natl Acad Sci U S A* 93:2598-2602.
632 Wichmann F, Asp T, Widmer F, Kölliker R (2011) Transcriptional responses of Italian
633 ryegrass during interaction with *Xanthomonas translucens* pv. *graminis* reveal
634 novel candidate genes for bacterial wilt resistance. *Theor Appl Genet* 122:567-
635 579.
636 Xu Y, Li P, Zou C, et al. (2017) Enhancing genetic gain in the era of molecular
637 breeding. *J Exp Bot*:1-26.
638 Yousaf S, Andria V, Reichenauer TG, Smalla K, Sessitsch A (2010) Phylogenetic and
639 functional diversity of alkane degrading bacteria associated with Italian ryegrass
640 (*Lolium multiflorum*) and birdsfoot trefoil (*Lotus corniculatus*) in a petroleum oil-
641 contaminated environment. *J Hazard Mater* 184:523-532.
642

643

Figure Legends

Figure 1: Shared gene models between *Lolium* and cereal species. Five way Venn diagram showing the distribution of shared gene models between the grasses wheat (yellow: Tae, *Triticum aestivum*, TGACv1), maize, (purple: Zma, *Zea mays*, AGPv4), *Brachypodium* (blue: Bdi, *Brachypodium distachyon*, v1.0), rice (red: Osa, *Oryza sativa*, IRGSP-1.0) and the merged proteomes of the two *Lolium* species (green: Lol, *L. multiflorum* and *L. perenne*). First are the numbers of orthologous groups and in brackets the number of genes.

Figure 2: Protein clusters enriched in the combined proteome of *Lolium multiflorum* and *L. perenne*. Enriched clusters were identified using gene ontology analysis and graphed using the Revigo webtool based on semantic word space (Supek *et al.* 2011). Log10_p_values are based on a Fisher`s exact test. Colours indicate levels of enrichment from high (blue) to low (red).

Tables

Table 1: Number of SSR loci identified in the genomes of *L. multiflorum* and *L*. *perenne*. Common SSR loci were identified by comparing 200 bp regions up- and downstream of the SSR locus.

Genome	Number of SSR loci		
	Di-nucleotide repeats	Tri-nucleotide repeats	Total
L. multiflorum	13,946	13,539	29,840
L. perenne	20,855	20,202	46,780
Common in both	2,013	1,943	4,076

Table 2: Summary annotation statistics of the *L. multiflorum* draft genome

	Gene models	Splice variants
Total number	61,153	81,244
Mean length	951 bp	1,175 bp
Length N50	1,558 bp	1,862 bp

Table 3: Number of gene models enriched in protein clusters of *L. perenne* or *L. multiflorum* when compared to each other. Clusters were grouped according to their functions assigned in the TAIR database and summarized according to functional processes.

Functional processes	*L. perenne*	*L. multiflorum*
General biological processes	22	2
Stress response	15	3
Protein modification	11	1
Regulation of transcription and translation	11	1
Metabolism	11	3
Metal ion binding	7	1
Oxidaten-reduction process	7	0
Hormones	4	2
Reproduction	2	1

Data Availability

The *L. multiflorum* genome assembly, as well as SSR sequences, genomic feature files, annotation data, canonical proteins and canonical transcripts for *L. multiflorum* and *L. perenne* as well as a list of orthologous clusters used for clustering are available on zenodo.org (doi:10.5281/zenodo.832654).

Supporting Information

SupplementaryTables.pdf (Protein clusters enriched in L. multiflorum [Supplementary Table 1] and L. perenne [Supplementary Table 2])