A synteny-based draft genome sequence of the forage grass
Lolium perenne

Stephen L. Byrne1, Istvan Nagy1, Matthias Pfeifer2†, Ian Armstead3, Suresh Swain3, Bruno Studer4, Klaus Mayer2, Jacqueline D. Campbell1‡, Adrian Czaban1, Stephan Hentrup1, Frank Panitz5, Christian Bendixen5, Jakob Hedegaard5§, Mario Caccamo6 and Torben Asp1,*
1Department of Molecular Biology, Genetics, Aarhus University, Forsøgsvej 1, Slagelse 4200, Denmark, 2Plant Genome and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstrasse 1, Neuherberg 85764, Germany, 3Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3DA, UK, 4Institute of Agricultural Sciences, ETH Zurich, Universitätsstraße 2, 8092 Zürich, Switzerland, 5Department of Molecular Biology and Genetics, Research Centre Foulum, Aarhus University, Blichers Alle 20, 8830 Tjele, Denmark, 6The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, UK

Received 22 July 2015; revised 4 September 2015; accepted 18 September 2015; published online 26 September 2015.
*For correspondence (e-mail torben.asp@mbg.au.dk).
†Present address: Roche Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany.
‡Present address: Department of Agronomy, Iowa State University, Ames, IA 50010, USA.
§Present address: Department of Molecular Medicine, Aarhus University Hospital, Skejby, Palle Juul-Jensens Boulevard 99, DK-8200 Aarhus N, Denmark.

SUMMARY
Here we report the draft genome sequence of perennial ryegrass (Lolium perenne), an economically important forage and turf grass species that is widely cultivated in temperate regions worldwide. It is classified along with wheat, barley, oats and Brachypodium distachyon in the Pooidae sub-family of the grass family (Poaceae). Transcriptome data was used to identify 28 455 gene models, and we utilized macro-co-linearity between perennial ryegrass and barley, and synteny within the grass family, to establish a synteny-based linear gene order. The gametophytic self-incompatibility mechanism enables the pistil of a plant to reject self-pollen and therefore promote out-crossing. We have used the sequence assembly to characterize transcriptional changes in the stigma during pollination with both compatible and incompatible pollen. Characterization of the pollen transcriptome identified homologs to pollen allergens from a range of species, many of which were expressed to very high levels in mature pollen grains, and are potentially involved in the self-incompatibility mechanism. The genome sequence provides a valuable resource for future breeding efforts based on genomic prediction, and will accelerate the development of new varieties for more productive grasslands.

Keywords: Lolium perenne, perennial ryegrass, genome sequence, self-incompatability, pollen allergens.

INTRODUCTION
Ryegrasses (Lolium spp.) and fescues (Festuca spp.) are the principle forage grasses underpinning forage-based meat and dairy production throughout the temperate world. The Lolium genus consists of nine closely related species that share a close evolutionary relationship to a number of broad-leaf fescues (sub-genus Schenlonorus, also frequently referred to as Festuca spp.). Species within the Lolium/Festuca complex are partially interfertile, form a well-defined ploidy series, and incorporate a wide range of variation in terms of phenology, agronomy and specific adaptive traits (Humphreys et al., 2006). Perennial ryegrass (2n = 14) has particular importance as a forage grass in temperate climate zones, and may be utilized as hay, silage and pasture. Its agricultural value lies in its rapid establishment, high yields, long growing season, tolerance of grazing, and high palatability and digestibility for ruminant...
animals. Formal perennial ryegrass breeding is estimated to be little more than 100 years old, and the genetic gains achieved in forage yield have been modest in comparison to genetic gains for seed yield in grain crops (Humphreys, 1997). Thus, there are considerable opportunities to increase the rate of genetic gain for perennial ryegrass through molecular breeding approaches such as genomic prediction (Fe et al., 2014) and through hybrid breeding (Arias et al., 2012). Efficient implementation of such genomics-based breeding strategies depends on the availability of a reference genome, and an understanding of the genetic control of the self-incompatibility (SI) mechanism in grasses. Successful implementation will have consequences not only for agricultural production, but also for the many roles of ryegrasses in contributing to the multiple ecosystems services of perennial grasslands, such as biodiversity reserves, carbon sequestration and water catchments.

Here we report a draft assembly of the perennial ryegrass genome, the first species within the Festuca/Lolium complex to be sequenced. The high level of macro-co-linearity between perennial ryegrass and barley was confirmed and exploited to identify syntenic relationships between ryegrass and high-quality reference genomes from the grass family (Poaceae). We have characterized the transcriptome in pollen and stigma, and during a compatible and incompatible pollination reaction. Genes with homology to allergens were expressed to very high levels in mature pollen grains, and we used the draft genome assembly to identify 630 genes encoding potential pollen allergens that have been partitioned into 17 protein families. The draft genome presented here will support studies to link phenotype and genotype, and help advance breeding efforts within the Festuca/Lolium complex to produce more productive and sustainable grasslands.

RESULTS

Perennial ryegrass genome assembly and annotation

We generated a shotgun sequence assembly of the genome using a mixture of Illumina HiSeq 2000 paired-end and mate-pair reads (Table S1). We sequenced a predominantly homozygous genotype (P226/135/16) that has been inbred for six generations. The estimated genome size from a k-mer spectrum analysis (k = 25 bp) (Figure S1) was 2068 Mb, and 76% of the genome was estimated to be repetitive. A similar estimate, 1998 Mb, was determined using a larger k-mer size (k = 31 bp) and an approach that takes into account sequencing error (Simpson, 2014). De novo assembly of these data using ALLPATHS-LG (Gnerre et al., 2011) yielded a shotgun assembly of 1128 Mb in scaffolds with a minimum length of 1 kb, which represents 54% of the estimated genome size of perennial ryegrass. Where possible, captured gaps were filled using PBJelly (English et al., 2012), using a ninefold coverage of the genome on long PacBio sequence reads. This resulted in a larger N50 and a reduced number of captured gaps (Table S2). The genome assembly consisted of 67 024 contigs and 48 128 scaffolds (>1 kb), with N50 sizes of 163 and 70.1 kb, respectively (Tables 1 and S3). The shotgun assembly was evaluated for gene representation by performing CEGMA analysis (Parra et al., 2007), which showed that 91.1 and 96.4% of core genes were present as complete hits and partial hits, respectively. If only the most highly conserved core eukaryotic genes (group 1) were considered, then 96.9% are present as complete hits (Table S4). The assembly was further evaluated by aligning a de novo transcriptome assembly of P226/135/16 (Farrell et al., 2014) against the genome assembly. Seventy-seven per cent of the transcripts had a match, and the coverage of all mappings was 98.7%. On average, there were only 1.14 scaffolds per mapped transcript. Using de novo assembled transcripts as evidence, we predicted 28 455 genes (encoding 40 068 proteins) using the MAKER2 annotation pipeline (Holt and Yandell, 2011). The gene set originated from 13 725 scaffolds that collectively accounted for 796 Mb of the assembly. The N50 length of the gene-containing scaffolds was 88.8 kb, and there were an average 2.1 genes per scaffold.

A synteny-based linear order of the ryegrass genome

The assembled perennial ryegrass whole-genome shotgun scaffolds, annotated genes and publicly available perennial ryegrass EST sequences were ordered along the seven linkage groups using the GenomeZipper approach, a comparative genomics-based bioinformatics strategy that integrates genetic marker maps and synteny information among related grass genomes (Mayer et al., 2009, 2011; Pfeifer et al., 2013). Briefly, this method makes use of 762 perennial ryegrass gene-derived markers that have been mapped in a two-way pseudo-testcross population at a mean marker distance less than 0.9 cM (Studer et al., 2012), of which 342 matched barley full-length cDNAs (see figure 1 in Pfeifer et al., 2013). Therefore, in the absence of chromosomal information for the assembled ryegrass genome sequences, the high level of macro-co-linearity between perennial ryegrass and barley (Pfeifer et al., 2013)

Table 1 Summary statistics of perennial ryegrass genome assembly

| Contigs (>1 kb) | Scaffolds (>1 kb) |
|----------------|------------------|
| Sum (Mb)       | 718              | 1128             |
| Total number   | 67 024           | 48 415           |
| N50 (bp)       | 16 370           | 70 062           |
| N90 (bp)       | 5500             | 14 249           |
| Maximum (bp)   | 99 259           | 512 939          |
| Captured gaps (Mb)| 292               |
Bio reads that had been error-corrected with an ABySS contrast to the masking of a 1 masker.org) with a Poaceae database (Table 2). This was in 10 464 unambiguously assigned to a linkage group; Data S1–S7.

Repetitive elements in the genome

The estimated repeat content from k-mer spectrum analysis (k = 25 bp) was 76%. The high repeat content was also demonstrated by looking at the number of repeat-induced branches in a de Bruijn assembly graph (Simpson, 2014). In comparison with other species, the frequencies of repeat branches are both high, and were not reduced as rapidly as the k-mer length increased (Figure S4). This makes genome assembly from short-read sequence data challenging. Only 17.4% of the final perennial ryegrass genome assembly was masked using RepeatMasker (http://www.repeatmasker.org) with a Poaceae database (Table 2). This was in contrast to the masking of a 1x genome coverage in PacBio reads that had been error-corrected with an ABySS assembly (Table S6) (Simpson et al., 2009), where 34.1% of the sequence was masked. The ratios of retroelements to DNA transposons were 4:1 and 8:1 in the assembly and PacBio reads, respectively. The low amount of repetitive content in the assembly is not surprising given the difficulties in assembling repetitive DNA with current sequencing technologies.

Genes differentially regulated during compatible and incompatible pollination reactions

The gametophytic self-incompatibility system in perennial ryegrass is controlled by at least two independent and multi-allelic loci, S and Z (Thorogood et al., 2002). The self-incompatibility reaction occurs when both the S and Z alleles of the haploid pollen are matched by the S and Z alleles in the diploid pistil. At present, the genes at S and Z have not been cloned, although candidate genes have been proposed (Yang et al., 2008; Shinozuka et al., 2010). Semi in vivo pollinated pistils were flash frozen 2 h after pollination with either compatible or incompatible pollen (Figure 1), and we compared their transcriptomes to the transcriptomes of unpollinated stigma collected at the same time. A greater number of genes were significantly up-regulated than down-regulated in both cases (Figures S6 and S7), and many of these genes were found to be regulated in the same manner, regardless of whether the pollination reaction was performed using compatible or incompatible pollen (Figure 1). We selected out genes that had at least a fourfold change in expression, and an FDR-adjusted P value of <0.001 (Data S9–S14), and focused on those genes that were uniquely differentially expressed under either a compatible or an incompatible reaction.

In a compatible pollination reaction, we observed significant accumulation of a gene with sequence homology to obtusifoliol 14α-demethylase (fold change > 240), a sterol 14α-demethylase and a member of the cytochrome monoxygenase superfamily. It has already been shown in a species of wild potato (Solanum chacoense) that obtusifoliol 14α-demethylase accumulates in the ovary after pollination, and that only sustained levels (more than 3 days) may be maintained under a compatible reaction (O’Brien et al., 2005). Here, we observed accumulation of obtusifoliol 14α-demethylase in the stigma within 2 h of pollination, and such accumulation was specific to pollination with compatible pollen. We also observed up-regulation of 12 genes with sequence homology to serine/threonine protein kinases, two with homology to G-type lectin S-receptor-like kinases, three with homology to wall-associated receptor kinases, and two with homology to L-type lectin domain-containing receptor kinases during a compatible reaction. An S-locus receptor-like kinase determines SI specificity at the stigma in sporophytic self-incompatibility systems, and was first identified in Brassica oleracea (Stein et al., 1991). Using GenomeZipper, we putatively place the

| Repeat          | Assembly Number | Assembly % | EC PacBio reads Number | EC PacBio reads % |
|-----------------|-----------------|------------|------------------------|-------------------|
| Retroelements   | 235 669         | 13.03      | 1 155 138               | 28.82             |
| SINEs           | 6236            | 0.10       | 7745                   | 0.05              |
| LINEs           | 42 271          | 3.01       | 55 712                 | 1.34              |
| RTE/Bov-B       | 1572            | 0.09       | 3233                   | 0.05              |
| L1/CIN4         | 40 699          | 2.92       | 52 479                 | 1.29              |
| LTR elements    | 187 072         | 9.92       | 1 091 681              | 27.43             |
| Ty1/Copia       | 57 106          | 3.87       | 198 182                | 7.05              |
| Gypsy/DIRS1     | 127 754         | 5.98       | 887 744                | 20.33             |
| DNA transposons | 98 229          | 3.13       | 196 172                | 3.53              |
| hobo-Activator  | 8959            | 0.22       | 11 146                 | 0.1               |
| Tc1-LS630-Pogo  | 19 090          | 0.30       | 26 902                 | 0.16              |
| En-Spm          | 2366            | 0.24       | 7618                   | 0.35              |
| Tourist/Harbiner| 16 427          | 0.62       | 21 671                 | 0.32              |
| Unclassified    | 7080            | 0.17       | 8904                   | 0.08              |
| Small RNA       | 4625            | 0.09       | 13 597                 | 0.48              |
| Satellites      | 6152            | 0.08       | 18 136                 | 0.2               |
| Simple repeats  | 149 747         | 0.82       | 373 282                | 0.86              |
| Low complexity  | 29 361          | 0.19       | 71 397                 | 0.18              |

The search was performed against the RepeatMasker (RM) Poaceae database (http://www.repeatmasker.org, version 20130433.).
two G-type lectin S-receptor-like kinases (up-regulated approximately 12-fold) proximal to the Z locus on perennial ryegrass linkage group 2 (Thorogood et al., 2002; Manzanares, 2013), one co-locating with the locus, and the other approximately 14 cM from the locus. Another up-regulated gene with homology to a leucine-rich repeat receptor-like serine/threonine protein kinase also co-locates with the Z locus.

In an incompatible reaction, we identified two genes with sequence homology to α-expansins that were both more than 200 fold up-regulated within 2 h of pollination. Using the zipper-anchored assembly, we found that both of these genes are co-located on linkage group 1, together with the S locus. Another gene co-located at the S locus, and up-regulated over 50 fold, shared sequence homology with the endoribonuclease Dicer homolog 3b-like (DCL3b). DCL3b (also known as DCL5) is part of the RNase III family, and plays a role in producing mature miRNAs that function in regulating gene expression. Interestingly, rice DCL3b was found to function in the production of stamen-specific miRNAs.

Figure 1. Self-incompatibility expression study.
Top: Design of self-incompatibility transcriptome study.
Bottom: Overview of differentially expressed genes for the various pairwise comparisons.
24 nt phased small RNAs (Song et al., 2012), and a role for these in reproductive development, meiosis or gamete formation has been suggested (Fei et al., 2013). Seed yield in perennial ryegrass is generally regarded as being low and unreliable (Elgersma, 1990). This affects commercial seed producers and the viability of a variety. It has already been demonstrated that the presence of large amounts of incompatible pollen in the pollen cloud leads to reduced seed yield per panicle and per plant (Studer et al., 2008). Molecular markers at S and Z may enable any SI-mediated loss in seed yield to be avoided.

The pollen transcriptome and characterization of allergen homologs

The proteins functioning as the male component in the SI mechanism need to be functional at pollen maturity. We therefore examined the relative expression of genes within the pollen transcriptome. Transcriptional activity in mature pollen is quite distinct from transcriptional activity in leaves, particularly with respect to how expression of individual genes contributes to the overall expression levels (Figure S5). For example, the expected fragment counts of the 100 most expressed genes in leaves contribute 13.9% of the total fragment count. This is in contrast with mature pollen, where the expected fragment counts of the 100 most expressed genes contribute 66.3% of the total fragment count. The 100 most highly expressed genes in mature pollen are enriched for genes involved in biological processes such as cell-wall organization and protein phosphorylation (Figure 2d). These included three pectinesterase inhibitors, a methyl esterase inhibitor, and three pectinesterase proteins, plus many genes that shared homology to well characterized allergens from a range of species.

![Classification of allergen genes](image)

Figure 2. Classification of allergen genes. (a) Classification and phylogenetic tree of allergen genes from various species that were used as references to build the hidden Markov model profiles. They are classified into 17 grass pollen allergen groups. (b) Phylogenetic analysis of 78 group 1 and seven group 2 grass pollen allergens that include α-expansin, β-expansin, expansin-like A and expansin-like B sub-classes of the expansin protein family. (c) Protein domains shared between Lol p1, Lol p2 and Lol p3, and the level of sequence similarity. (d) Gene ontology annotation graph of the 630 Lolium perenne genes from the 17 grass pollen allergen groups. Nodes with fewer than 30 sequences were filtered out.
Members of the grass family are wind-pollinated, and release huge amounts of pollen during anthesis, making allergies from grass pollen one of the most common pollen allergies worldwide. Susceptible individuals express immunoglobulin E (IgE) in response to proteins on the exterior coat of inhaled pollen, and these attach themselves to mast cells, triggering the release of histamine and causing an allergic reaction. We developed hidden Markov model profiles representing pollen allergens from various species (Figure 2a and Table S7), and used them to search the perennial ryegrass gene predictions for homologs. This resulted in the identification of 630 non-redundant hits, belonging to 17 protein families (Figure 2a and Table 3). The largest protein family with pollen allergen homologs is the polcalcin family (EF-hand calcium-binding proteins) with 145 members. Other large protein families represented include glycosyl hydrolases, expansins, phytocyanins and FAD-binding domain (berberine-like) proteins. N-terminal signal peptides were found in ten of the 17 protein families, indicating that these proteins are secreted across the endoplasmic reticulum. Major pollen allergens belonging to protein families without secretion signals (such as polcalcins, profilins or Bet v I allergens) were present in several plant species.

Using the pollen RNA sequencing data, we found that 60% of the pollen allergen homologs had detectable expression in mature pollen (Data S8), and 26 had more than 1000 transcripts per million (TPM) (Table S8). Of the pollen allergens with detectable expression levels within mature pollen, five genes were within the 20 with the highest relative expression levels. These comprised three members of the expansin family (two Lol p1 signature proteins and one Lol p2-like protein), a profilin protein and a pectate lyase, altogether accounting for 6.4% of the total expected counts.

### Expansin gene family

Expansins are secreted proteins that play a role in plant cell enlargement, by acting on the cell wall to promote its extensibility (McQueen-Mason et al., 1992; Cosgrove and Li, 1993). It is hypothesized that expansins break non-covalent bonds between cell-wall polysaccharides, thereby permitting pressure-dependent expansion of the cell (McQueen-Mason and Cosgrove, 1994). Expansins typically have two conserved protein domains: a cysteine-rich domain showing similarity to family 45 endoglucanases (EG45-like domain) and a cellulose-binding domain. They are divided into four sub-classes: expansin A, expansin B, expansin-like A and expansin-like B. We found members of all four sub-classes within group 1 (78 genes) and group 2 (seven genes) (Figure 2b). Lol p1 is a group 1 grass pollen allergen belonging to the β-expansin family, and is the most prominent pollen allergen in grasses, with approximately 90% of allergic patients showing Lol p1-specific IgE in their sera (Laffer et al., 1996). Lol p2 and Lol p3 are group 2 grass pollen allergens, and were identified as minor allergens in perennial ryegrass more than 40 years ago (Johnson and Marsh, 1966). Group 1 and 2 pollen allergens share sequence similarity at their C-terminal part (Figure 2c), which covers nearly 90% of the group 2 allergen sequences, and is identified as a cellulose-binding-like domain. In the case of group 1 pollen allergens, protein domain analysis and prediction show an additional conserved domain in the N-terminal part of their sequences, the so-called RlpA-like double-psi beta-barrel domain (Figure 2c). Five of the seven group 2 pollen allergens were expressed relatively highly in mature pollen (>100 TPM), and to a very low levels in leaves (<1 TPM), possibly indicating a specialized function in pollen grains. One group 1 pollen allergen was expressed to a very high level specifically in mature pollen (LOL1_050, 13 635 TPM). It is hypothesized that group 1 grass pollen allergens have a role to

### Table 3 Proteins identified in perennial ryegrass with homology to previously identified pollen allergens

| Protein group | Description | Code | Total | Signal peptide |
|---------------|-------------|------|-------|----------------|
| 1             | Expansins, Lol p1 signatures | LOL1 | 78    | +              |
| 2             | Expansins, Lol p2 signatures | LOL2 | 72    | +              |
| 3             | Copper/zinc superoxide dismutase (SODC) family | OLE5 | 6     | –              |
| 4             | Glycosyl hydrolase family 17 | OLE9 | 91    | –              |
| 5             | Pectate lyases, Amb family signatures | AMBR | 19    | +              |
| 6             | Ribonucleases, Poo pl signatures | POAP | 6     | +              |
| 7             | Glycosyl hydrolase family 28, parallel β-helix repeats | PARA | 53    | +              |
| 8             | Berberine-like gene family | BERB | 37    | –              |
| 9             | Ole e6 pollen allergens | OLE6 | 3     | –              |
| 10            | Phytocyanin domain profile family | AMB3 | 60    | –              |
| 11            | Cysteine-rich secretory protein family | ART2 | 18    | +              |
| 12            | γ-thionin family, defensin-like proteins | ART1 | 15    | +              |
| 13            | Pathogenesis-related proteins, Bet v I family signatures | BETV | 26    | –              |
| 14            | Lol p1-like pollen allergens, Ole e1 family signatures | OLE1 | 28    | +              |
| 15            | Pathogenesis-related proteins, thaumatin family signatures | JUN3 | 34    | +              |
| 16            | Polcalcins, EF-hand calcium-binding domain profile proteins | EFHA | 145   | –              |
| 17            | Profilins | PROF | 4     | –              |
| Total         |             |      | 630   |                |

The protein group numbers correspond to those in Figure 2a.

© 2015 The Authors
The Plant Journal © 2015 John Wiley & Sons Ltd, The Plant Journal, (2015), 84, 816–826
play in softening maternal cell walls to allow the pollen to penetrate the stigma and style (Cosgrove et al., 1997; Cosgrove, 2000).

**DISCUSSION**

We have sequenced and assembled a draft genome of perennial ryegrass, the first species from the *Lolium/Festuca* complex for which this has been achieved. This complex encompasses many of the species used in temperate agricultural grasslands that underpin dairy and meat production, of which perennial ryegrass is the most commercially important. The assembly was generated using Illumina paired-end and mate-pair reads, and short to medium-sized gaps were filled using PacBio reads. The resulting assembly covers 1128 Mb of the perennial ryegrass genome, and consists of 28 455 gene models supported by transcript data. Completeness of genic sequences was verified by mapping a de novo assembled transcriptome to the genome assembly, and the GenomeZipper approach was used to construct a synteny-based gene order for 11 311 annotated genes. The assembly will facilitate future genome research within the *Lolium/Festuca* complex, and will accelerate map-based cloning.

Self-incompatibility prevents self-pollination in plants, and ensures that inbreeding is restricted and diversity is preserved. Grasses utilize a unique gametophytic self-incompatibility mechanism that is controlled by two independent and multi-allelic loci, S and Z, which have already been mapped in perennial ryegrass (Thorogood et al., 2002). However, the genes acting at S and Z have yet to be elucidated. We have used our annotated draft sequence assembly to study gene expression in the stigma during pollination with either compatible or incompatible pollen. This enabled us to identify genes co-locating with S and Z that were differentially expressed during pollination, including homologs to S-receptor-like kinases at Z, and a homolog to DCL3b at S. The availability of the draft genome sequence should now accelerate use of map-based cloning strategies to identify S and Z, and enable the design of functional markers to reduce any SI-mediated seed loss.

The male determinant in the self-incompatibility mechanism must be functional at pollen maturity, because the SI reaction occurs within minutes after pollination. Many of the genes that are highly expressed in ryegrass pollen share sequence homology to previously identified pollen allergens. Although their exact biological function remains unknown, various roles have been proposed for these proteins, including involvement in pollen–stigma recognition (Cosgrove et al., 1997). Allergic rhinitis (commonly known as hayfever) is estimated to affect between 10 and 30% of the world’s population (http://www.worldallergy.org/UserFiles/file/WAO-White-Book-on-Allergy_web.pdf), and is the result of inhaling allergens to which the body has become sensitized. Grass pollen is one of the major causes of allergic rhinitis worldwide, and some of the first pollen allergens characterized were identified over 50 years ago in perennial ryegrass (Johnson and Marsh, 1965). Group 1 allergens are glycoproteins that are recognized by over 90% of allergic patients, and Lol p1, the group 1 allergen of perennial ryegrass has previously been cloned and sequenced (Griffith et al., 1991). We identified a small cluster within the larger expansin protein family containing five homologs to the published Lol p1 protein, of which four were relatively highly expressed in mature pollen grains. One of these genes was ranked 12th in terms of relative expression in mature pollen grains. We also identified homologs of many other important allergens that are expressed to high levels in mature pollen grains of perennial ryegrass. In addition to treating the symptoms of allergies, e.g. with antihistamines, there may be opportunities to use hypoallergenic recombinant expressed allergens for patient-tailored immunotherapy (Singh and Bhalla, 2003). Characterization of the allergen complex in the pollen of common grass species will facilitate this.

The lack of a genome sequence for perennial ryegrass has prevented the application of molecular markers in traditional breeding programs. The draft genome assembly presented here will enable efficient discovery and characterization of genetic variation, and a deeper understanding of the genetic control of important traits such as self-incompatibility. This will accelerate the implementation of genomic prediction and hybrid breeding within forage grasses, leading to better varieties, and more productive and sustainable grasslands.

**EXPERIMENTAL PROCEDURES**

**DNA preparation and whole-genome shotgun sequencing**

High-quality genomic DNA was isolated from leaves of the genotype P226/135/16 using CTAB extraction and passed through a DNAeasy plant spin column (Qiagen, https://www.qiagen.com/) to remove contaminants. The genotype originated from the Institute of Biological, Environmental and Rural Sciences (Abertywth University, UK), and has undergone inbreeding for six generations. Illumina (http://www.illumina.com/) paired-end libraries with mean fragment lengths of 140 and 550 bp were prepared from genomic DNA. Illumina mate-pair libraries were also prepared, with mean insert sizes of 1.8, 3.4 and 8.6 kb. DNA was fragmented by shearing using nebulizers using an Illumina mate pair library kit version 2. The paired-end libraries were prepared using an NEBNext DNA sample preparation kit (New England Biolabs, https://www.neb.com/) with Illumina adaptors according to the NEBNext instructions. Libraries were sequenced using either an Illumina GAIIx or a HiSeq 2000 using TruSeq chemistry according to the manufacturer’s instructions. HiSeq 2000 sequence data (FASTQ format) were extracted using CASAVA version 1.8 (https:// support.illumina.com/content/dam/illumina-support/documents/myillumina/a557afc4-bf0e-4dad-9e59-9c740d1e751/casava_user guide_15011196d.pdf). Long jump distance libraries with insert sizes of 7.7, 26.2 and 29.7 Kb were also prepared by Eurofins...
Genome assembly and evaluation

The Illumina data were assembled using ALLPATHS-LG (release 47760) (Gnerre et al., 2011) on a machine with 1009 GB RAM and 24 cores (four Intel®/Westmere® E7-4807 CPUs at 1.87 GHz, 6 cores per CPU), making use of the different libraries in specific ways (Table S9). The PacBio data was used for gap filling using PBJelly (version 12.3.14) (English et al., 2012) after aligning the reads to the ALLPATHS-LG assembly using BLASR (Chaisson and Tesler, 2012). The consensus sequence of aligned PacBio reads was used to fill in the gaps, and the mean number of reads mapping to closed gaps was 8.7 (Figure S8). In order to evaluate the completeness of the assembly in terms of gene content, we performed CEGMA analysis (Parra et al., 2007). We also aligned a de novo transcriptome assembly to the assembled genome using BLAT (Kent, 2002), and evaluated the alignments using IsoBLAT (https://github.com/josephryan/isoblat).

Genome annotation

Multiple RNA-seq datasets were available from various expression studies performed on multiple plant sections, multiple genotypes, and at multiple time points in various environments (Table S10). This included RNA-seq data from the same inbred genotype used for the genome assembly (Farrell et al., 2014). Reads were error-corrected before assembly using the ErrorCorrectReads.pl script provided with ALLPATHS-LG (Gnerre et al., 2011). We performed individual de novo transcriptome assemblies for each genotype using Trinity (Grabherr et al., 2011). After assembly, the error-corrected reads were aligned back against the assembled transcripts, and abundance estimates were made using RSEM (Li and Dewey, 2011). We filtered out transcripts with less than 1% of the per component (IsoPct) expression level. PASA (Haas et al., 2003) was then used to perform alignment assemblies (using all de novo transcriptome assemblies together) with the genome assembly as a reference to generate a comprehensive transcriptome database that contained 178 589 transcripts. The 178 589 transcripts were used as evidence to directly infer gene predictions using MAKER (version 2.31.6) (Holt and Yandell, 2011). We also provided a protein dataset from Brachypodium distachyon (International Brachypodium Initiative, 2010) as evidence, and repeat masking was performed using RepeatMasker 3.3.0 (http://www.repeatmasker.org) using a Poaceae repeat database. Functional annotation of the resulting gene predictions was performed using InterProScan V5 (Jones et al., 2014).

Repeat content

We performed an assembly of the Illumina data using ABySS (Simpson et al., 2009) with a k-mer size of 73 bp. The total contig length was 2144 Mb, and this assembly was used to error correct the PacBio reads using ECTools (Lee et al., 2014). We randomly sampled a 1x genome coverage in reads greater than 2 kb from the error-corrected reads (Table S11). We used RepeatMasker 3.3.0 (http://www.repeatmasker.org) and a Poaceae repeat database to annotate and quantify repeats in this sample of error-corrected PacBio reads.

GenomeZipper

On the basis of the identified syntenic relationships to the high-quality reference grass genomes of perennial ryegrass, whole-genome shotgun scaffolds assigned to individual chromosomes were ordered using the GenomeZipper approach (Mayer et al., 2011; Pfeifer et al., 2013). Briefly, guided by a genetic marker map, GenomeZipper determines orthologous gene relationships and syntenic regions between a target genome (i.e. perennial ryegrass) and high-quality reference genome sequences, e.g. Brachypodium distachyon (International Brachypodium Initiative, 2010), O. sativa (International Rice Genome Sequencing Project, 2005) and S. bicolor (Paterson et al., 2009). This comparative genome framework constitutes the backbone for linearly ordering target gene sequences of interest (i.e. annotated high-quality perennial ryegrass proteins) along individual chromosomes. Stringent best bidirectional BLAST hit (bbh) searches were used to construct seven chromosome scaffolds by merging available perennial ryegrass genetic marker sequences (Studer et al., 2012) and annotated proteins for the three reference genomes. Furthermore, the annotated perennial ryegrass genes were integrated into the genome backbone as determined by bbh searches against the marker sequences and coding sequences of the anchored reference proteins [BLASTN with E < 1e-10, minimum alignment length 100 bp, and minimum alignment identity of 80% (marker), 75% (Brachypodium) and 70% (rice and sorghum)]. Then the chromosome-assigned perennial ryegrass whole-genome shotgun assemblies were positioned in the GenomeZipper backbones via sequence homology to the respective elements (i.e. genetic markers, reference proteins and annotated perennial ryegrass genes). Additionally, publicly available perennial ryegrass EST sequences (Studer et al., 2012; Pfeifer et al., 2013) were ordered and attached to the GenomeZipper framework [BLASTN with E < 1e-10, minimum alignment length 100 bp, and minimum alignment identity of 80% (marker), 75% (Brachypodium) and 70% (rice and sorghum)].

Gene expression analysis

We used genotypes whose S and Z allele composition is known and that represent a diverse set of S-locus alleles. The plant F1-30, which is heterozygous S1S2Z1Z2 (the allele numbers are allocated at random), was complemented with pollen from the variety Foxtrot S0S1Z0Z4. Three biological replicates of pollen and stigma samples were collected from the various genotypes. The pollen was harvested into clear bags and transferred into micro-centrifuge tubes; the pollen volume was approximately 0.1 ml per sample. The stigmas were collected from unopened florets with non-dehiscent anthers. Non-pollinated pistils at various stages of maturity were sampled, and the ovary was removed using a razor blade. Approximately 50 pistils were sampled to create one stigma sample. In addition to pollen and stigma tissue samples, self-incompatible and self-compatible pollen-stigma interactions were analyzed. Using in vivo pollinations, three biological replicates from self-pollinations (self-incompatible) and two replications from cross-pollinations (self-compatible) were collected. The semi in vivo pollinated pistils were collected approximately 2 h...
after pollination, the ovaries were removed and the stigma samples were placed in liquid nitrogen. Total RNA was extracted from the pistils, compatible pollinated pistils and incompatible pollinated pistils using an RNeasy™ plant mini kit (Qiagen) according to the manufacturer’s instructions. RNA from the pollen samples was extracted using a PicoPure™ RNA isolation kit (Arcturus Bioscience Inc., http://www.arctur.com/) according to the manufacturer’s instructions. RNA integrity was measured using an RNA 6000 Nano Labchip™ on an Agilent 2100 Bioanalyzer (Agilent Technologies, http://www.agilent.com/home). The sequencing libraries were generated using a TruSeq Stranded mRNA sample preparation kit (Illumina) according to the manufacturer’s instructions. The cDNA libraries were sequenced by Illumina HiSeq 2000 sequencing, and 76 bp paired-end reads were generated for each of them. All libraries were pooled into two groups, and each group was distributed over two lanes, with an expectancy minimum of 30 million reads/sample (AROS Applied Biotechnology, http://arosab.com/). Leaf samples were taken from genotype NGB09 grown under ambient conditions, and flash frozen in liquid nitrogen. Total RNA was extracted using an RNeasy™ plant mini kit (Qiagen) according to the manufacturer’s instructions. RNA integrity was measured using an RNA 6000 Nano Labchip™ on an Agilent 2100 Bioanalyzer (Agilent Technologies). Libraries for sequencing were prepared as described above. The number of paired-end reads for the stigma and pollen samples varied between 19 and 33 million pairs. The numbers of read pairs for the three leaf replicates were 13.79, 12.98 and 13.79 million.

Reads were aligned to the gene predictions using Bowtie2 (Langmead and Salzberg, 2012) with the following settings (-sensitive -dpad 0 -gbar 99999999 -mp 1,1 -np 1 -score-min L0,-0,1 -i 1 -X 1000 -no-mixed -no-discordant -k 200). RSEM (Li and Dewey, 2011) was used to calculate the number of transcripts per million (TPM) for each gene, and edgeR (Robinson et al., 2010) was used to compare gene expression in two pairwise comparisons: (i) unpollinated stigma versus stigma pollinated with compatible pollen, and (ii) unpollinated stigma versus stigma pollinated with incompatible pollen.

Identification of allergen homologs

Publicly available protein sequences of pollen allergens were collected from the National Center for Biotechnology Information/UniProt protein sequence databases (http://www.ncbi.nlm.nih.gov/protein). In addition to known grass pollen allergens such as Lol p1, Lol p2, Lol p3, Lol p5 and Lol p11, prevalent pollen allergen sequences from trees, conifers and weeds were also collected, resulting in a set of 74 reference sequences belonging to 17 protein families (Table S7). Reference sequences were used as queries for BLASTP searches (Altschul et al., 1990) against a non-redundant database containing Viridiplantae protein sequences. Sequences representing significant BLAST hits (at least 80% similarity level, scores > 80) were collected and aligned to the reference sequences of each protein family. After visual inspection, problematic sequences were removed from the alignments, retaining 5–50 sequences in the core alignments representing each protein family, depending on the availability and diversity of the family. Where possible, monocot sequences were preferred to dicot sequences during selection. The remaining sequences were re-aligned using Clustal Omega (Sievers and Higgins, 2014) with default parameters. The resulting Stockholm format alignments were used to construct hidden Markov model profile matrices using the hmmbuild program of the HMMER package (version 3.1b1, http://hmmer.org). The profile matrices were used to scan the protein sequences from the perennial ryegrass genome, using the hmmscan program of the HMMER package. Predictive information on protein functions and conserved sequence elements was obtained by sending all hits and reference sequences through a stand-alone InterProScan 5 pipeline (Jones et al., 2014) by scanning the PANTHER (http://pantherdb.org/), PROSITE profiles (http://prosite.expasy.org/), Pfam (http://pfam.xfam.org/) and SUPERFAMILY databases (http://supfam.org/SUPERFAMILY/). This pipeline was also used for prediction of transmembrane topology and signal peptides by integrating the Phobius (Kall et al., 2004) and SignalP (Emanuelsson et al., 2007) utilities.

Accession numbers

The perennial ryegrass genome and transcriptome assemblies may be downloaded in fasta format, and accessed through a genome browser at http://185.45.23.197:5080/ryegrassgenome. The complete SI expression study data have been deposited in ArrayExpress with accession number E-MTAB-3760.

ACKNOWLEDGMENTS

This work was supported by the Godtfred Birkedal Hartmann Foundation, and the UK Biotechnology and Biological Sciences Research Council. The genome browser is supported by a grant from Promilleleafgiftsfonden. We thank the Duke Center for Genomic and Computational Biology Genome Sequencing Shared Resource (Durham, NC), which provided the PacBio sequencing service.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. K-mer spectra analysis.
Figure S2. Synteny between ryegrass and barley.
Figure S3. Coverage of the Brachypodium genome by assembled Lolium whole-genome shotgun scaffolds.
Figure S4. Estimated repeat branch rate as a function of k-mer length.
Figure S5. TPM for leaf and pollen samples.
Figure S6. MA-plots and volcano plots for differentially expressed genes during a compatible pollination reaction.
Figure S7. MA-plots and volcano plots for differentially expressed genes during an incompatible pollination reaction.
Figure S8. Histogram of the number of PacBio reads addressing gaps.
Table S1. Illumina sequencing data overview.
Table S2. Summary statistics for gap filling.
Table S3. Summary statistics for assembly.
Table S4. CEGMA summary.
Table S5. GenomeZipper summary.
Table S6. ABySS assembly summary statistics.
Table S7. Pollen allergens identified in various species.
Table S8. TPM of ryegrass pollen allergen homologs in leaf and mature pollen.
Table S9. Assignment of libraries for assembly.
Table S10. Summary of data used in transcriptome assemblies.
Table S11. Summary of PacBio data.

Data S1–S7. Spreadsheets of the perennial ryegrass GenomeZipper data, one file for each of the seven linkage groups.

© 2015 The Authors
The Plant Journal © 2015 John Wiley & Sons Ltd, The Plant Journal, (2015), 84, 816–826
Data S8. Spreadsheet showing the expression levels (TPM) of the 630 genes with sequence homology to known pollen allergens, in both mature pollen and leaf samples.

Data S9. Spreadsheet showing genes that were significantly down-regulated in both a compatible and an incompatible pollination reaction.

Data S10. Spreadsheet showing genes that were significantly down-regulated in a compatible pollination reaction.

Data S11. Spreadsheet showing genes that were significantly up-regulated in both a compatible and an incompatible pollination reaction.

Data S12. Spreadsheet showing genes that were significantly up-regulated in a compatible pollination reaction.

Data S13. Spreadsheet showing genes that were significantly up-regulated in an incompatible pollination reaction.

Data S14. Spreadsheet showing genes that were significantly up-regulated in a compatible pollination reaction.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Arias, A., Studer, B., Frei, U. and Lubberstedt, T. (2012) Prospect for hybrid breeding in bioenergy grasses. Bioenergy Res. 5, 10-19.

Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics, 13, 238.

Cosgrove, D.J. (2000) Loosening of plant cell walls by expansins. Nature, 407, 321-326.

Cosgrove, D.J. and Li, Z.C. (1993) Role of expansin in cell enlargement of oat coleoptiles (analysis of developmental gradients and photocontrol). Plant Physiol. 103, 1321-1328.

Cosgrove, D.J., Bedinger, P. and Durachko, D.M. (1997) Group I allergens of grass pollen as cell wall-loosening agents. Proc. Natl Acad. Sci. USA, 94, 6599-6604.

Elgersma, A. (1990) Genetic variation for seed yield in perennial ryegrass (Lolium perenne L). Plant Breeding, 105, 117-125.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2, 953-971.

English, A.C., Richards, S., Han, Y. et al. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. Plant Cell, 24, 2395-2406.

Farrell, J.D., Byrne, S., Paina, C. and Asp, T. (2012) Prospect for hybrid breeding in bioenergy grasses. Bioenergy Res. 5, 10-19.

Fei, D., Ashraf, B., Byrne, S. et al. (2014) Prospects for introducing genomic selection into forage grass breeding. In EGF at 50: The Future of European Grasslands. Proceedings of the 25th General Meeting of the European Grassland Federation, Aberystwyth, Wales, 7-11 September 2014 (Hopkins, A., Collins, R.P., Fraser, M.D., King, V.R., Lloyd, D.C., Moorthy, J.M. and Robson, P.R.H., eds). Wallingford, UK: Centre for Agriculture and Biosciences International, pp. 830-832.

Fei, Q., Xia, R. and Meyers, B.C. (2013) Phased, secondary, small interfering RNAs in postranscriptional regulatory networks. Plant Cell, 25, 2400-2415.

Gnerre, S., MacCallum, I., Przybylski, D. et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl Acad. Sci. USA, 108, 1513-1518.

Grabherr, M.G., Haas, B.J., Yassour, M. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644-652.

Griffith, I.J., Smith, P.M., Pollock, J. et al. (1991) Cloning and sequencing of Lol pl, the major allergenic protein of rye-grass pollen. FEBS Lett. 279, 210-215.

Haas, B.J., Delcher, A.L., Mount, S.M. et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654-5666.

Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome database management tool for second-generation genome projects. BMC Bioinformatics, 12, 491.

Humphreys, M.O. (1997) The contribution of conventional plant breeding to forage crop improvement [WWW document]. URL http://www.internationalgrasslands.org/files/ipc/publications/1997/iii-071.pdf [accessed on 6 October 2015].

Humphreys, M.W., Yadav, R.S., Cairns, A.J., Turner, L.B., Humphreys, J. and Skot, L. (2006) A changing climate for grassland research. New Phytol. 169, 9-26.

International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature, 463, 763-768.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature, 436, 793-800.

Johnson, P. and Marsh, D.G. (1965) ‘Isoallergens’ from rye grass pollen. Nature, 206, 935-937.

Jones, P., Blins, D., Chang, H.Y. et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics, 30, 1236-1240.

Kall, L., Krogh, A. and Sonnhammer, E.L. (2004) A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338, 1027-1036.

Kent, W.J. (2002) BLAT – the BLAST-like alignment tool. Genome Res. 12, 656-664.

Laffar, S., Spitzauer, S., Susani, M. et al. (1996) Comparison of recombinant timothy grass pollen allergens with natural extract for diagnosis of grass pollen allergy in different populations. J. Allergy Clin. Immunol. 98, 652-658.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357-359.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McMicking, W.R. and Schatz, M. (2014) Error correction and assembly complexity of single molecule sequencing reads. bioRxiv, doi: 10.1101/006395.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12, 323.

Manzanoares, C. (2013) Genetics of self-incompatibility in perennial ryegrass (Lolium perenne L.). PhD Thesis, University of Birmingham, UK.

Mayer, K.F., Taudien, S., Martis, M. et al. (2009) Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. 151, 496-505.

Mayer, K.F., Martis, M., Hedley, P.E. et al. (2011) Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell, 23, 1249-1263.

McQueen-Mason, S. and Cosgrove, D.J. (1994) Disruption of hydrogen bonding between plant cell wall polymers by proteins that induce wall extension. Proc. Natl Acad. Sci. USA, 91, 6574-6578.

McQueen-Mason, S., Durachko, D.M. and Cosgrove, D.J. (1992) Two endogenous proteins that induce cell wall extension in plants. Plant Cell, 4, 1425-1433.

O’Brien, M., Chantha, S.C., Rahier, A. and Matton, D.P. (2005) Lipid signaling in plants. Cloning and expression analysis of the obtusifoliol 14-demethylase from Solanum chacoense Btt., a pollination- and fertilization-induced gene with both obtusifoliol and lanosterol demethylase activity. Plant Physiol. 139, 734-749.

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

Paterson, A.H., Bowers, J.E., Bruggmann, R. et al. (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551-556.

Pfeifer, M., Martis, M., Asp, T., Mayer, K.F., Lubberstedt, T., Byrne, S., Frei, U. and Studer, B. (2013) The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. Plant Physiol. 161, 571-582.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edger: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139-140.

Shinozuka, H., Cogan, N.O., Smith, K.F., Spangenberg, G.C. and Forster, J.W. (2010) Fine-scale comparative genetic and physical mapping supports map-based cloning strategies for the self-incompatibility loci of perennial ryegrass (Lolium perenne L.). Plant Mol. Biol. 72, 343-355.

Sievers, F. and Higgins, D.G. (2014) Clustal Omega. Curr. Protoc. Bioinformatics, 46, 3.
Simpson, J.T. (2014) Exploring genome characteristics and sequence quality without a reference. Bioinformatics, 30, 1228-1235.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res, 19, 1117-1123.

Singh, M.B. and Bhalla, P.L. (2003) Hypoallergenic derivatives of major grass pollen allergens for allergy vaccination. Immunol. Cell Biol. 81, 86-91.

Song, X., Li, P., Zhai, J. et al. (2012) Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J. 69, 462-474.

Stein, J.C., Howlett, B., Boyes, D.C., Nasrallah, M.E. and Nasrallah, J.B. (1991) Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of Brassica oleracea. Proc. Natl Acad. Sci. USA, 88, 8816-8820.

Studer, B., Jensen, L.B., Hentrup, S., Brazauskas, G., Kelliker, R. and Lubberstedt, T. (2008) Genetic characterisation of seed yield and fertility traits in perennial ryegrass (Lolium perenne L.). Theor. Appl. Genet. 117, 781-791.

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

Thorogood, D., Kaiser, W.J., Jones, J.G. and Armstead, I. (2002) Self-incompatibility in ryegrass 12. Genotyping and mapping the S and Z loci of Lolium perenne L. Heredity, 88, 385-390.

Yang, B., Thorogood, D., Armstead, I. and Barth, S. (2008) How far are we from unravelling self-incompatibility in grasses? New Phytol. 178, 740-753.