

Received Date : 22-Jul-2015

Revised Date : 04-Sep-2015

Accepted Date : 18-Sep-2015

Article type : Resource

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

Stephen L. Byrne¹(stephen.byrne@mbg.au.dk), Istvan Nagy¹(istvan.nagy@mbg.au.dk), Matthias Pfeifer^{2,9}(matthiaspfeifer@gmx.net), Ian Armstead³(ipa@aber.ac.uk), Suresh Swain³ (deceased), Bruno Studer⁴ (bruno.studer@usys.ethz.ch), Klaus Mayer²,(k.mayer@helmholtz-muenchen.de), Jacqueline D. Campbell^{1,7},(jacqueline.danielle@gmail.com), Adrian Czaban¹(adrian.czaban@mbg.au.dk), Stephan Hentrup¹ (stephan.hentrup@mbg.au.dk), Frank Panitz⁵(frank.panitz@mbg.au.dk), Christian Bendixen⁵ (christian.bendixen@mbg.au.dk), Jakob Hedegaard^{5,8} (jakob.hedegaard@clin.au.dk), Mario Caccamo⁶ (Mario.Caccamo@tgac.ac.uk), Torben Asp^{1*}(torben.asp@mbg.au.dk).

¹Dept. of Molecular Biology and 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, United Kingdom. ⁴Institute of Agricultural Sciences, ETH Zurich, Universitätstrasse 2, 8092 Zürich, Switzerland. ⁵Department of Molecular Biology and Genetics, Aarhus University, Research Centre Foulum, Blichers Allé 20, 8830 Tjele, Denmark. ⁶The Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, United Kingdom. ⁷Present address: Department of Agronomy, Iowa State University, Ames, IA 50010 USA. ⁸Present address: Dept. of Molecular Medicine (MOMA), Aarhus University Hospital, Skejby, Palle Juul-Jensens Boulevard 99, DK-8200 Aarhus N, Denmark. ⁹Present address: Roche Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1111/tbj.13037

This article is protected by copyright. All rights reserved.

***Corresponding author:** Torben Asp [email: torben.asp@mbg.au.dk]

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

Running title: Genome sequence of perennial ryegrass.

Accession Numbers: E-MTAB-3760.

Summary

Here we report the draft genome sequence of perennial ryegrass (*Lolium perenne*), an economically important forage and turf grass species widely cultivated in temperate regions worldwide. It is classified along with wheat, barley, oats and *Brachypodium distachyon* in the Pooideae sub-family of the grass family (Poaceae). Transcriptome data was used to identify 28,455 gene models, and we utilize 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 (SI) mechanism enables the pistil of a plant to reject self-pollen and therefore promote outcrossing. We have used the sequence assembly to characterise transcriptional changes in the stigma during pollination with both compatible and incompatible pollen. Characterisation 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 potentially involved in the SI mechanism. The genome sequence provides a valuable resource for future breeding efforts based on genomic prediction, and will accelerate the development of varieties for more productive grasslands.

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 (subgenus *Schendonorus* 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, M. W. et al. 2006). Perennial ryegrass (2n=14) has particular importance as a forage grass in temperate climate zones, and can be utilised 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, M. O. 1999). Thus, there are considerable opportunities to increase the rate of genetic gain for perennial ryegrass through molecular breeding approaches such as genomic prediction (Fe, D. et al. 2014), and through hybrid breeding (Arias, A et al. 2012). Efficient implementation of such genomics based breeding strategies depends on a reference genome, and an understanding of the genetic control of the SI mechanism in grasses. Success in this 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 characterised 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 have 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 to 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 kmer (k=25 bp) spectrum analysis (Figure S1) was 2,068 Mb and 76 percent of the genome was estimated to be repetitive. A similar estimate, 1,988 Mb was determined with a larger kmer size (k=31 bp), and using an approach that takes into account sequencing error (Simpson, J. T. 2014). De novo assembly of these data with ALLPATHS-LG (Gnerre, S. et al. 2011) yielded a shotgun assembly of 1,128 Mb in scaffolds with a minimum length of 1 Kb, which represents 54 percent of the estimated genome size of perennial ryegrass. Where possible captured gaps were filled with PBJelly (English, A. C. et al. 2012) using a 9 x coverage of the genome on long sequence reads (PacBio). This resulted in a larger contig 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 16.3 Kb and 70.1 Kb, respectively (Table 1, Table S3). The shotgun assembly was evaluated for gene representation by performing CEGMA analysis (Parra, G. et al. 2007), which showed that 91.1 percent and 96.4 percent of core genes were present as complete hits and partial hits, respectively. If we were to consider only the most highly conserved core eukaryotic genes (group 1 CEGs), then 96.92 percent are present as complete hits (Table S4). The assembly was further evaluated by aligning a *de-novo* transcriptome assembly of P226/135/16 (Farrell, J. D. et al. 2014) against the genome assembly. Seventy seven percent of the transcripts had a match, and the coverage of all mappings was 98.7 percent. On average there were only 1.14 scaffolds per mapped transcript. Using

de-novo assembled transcripts as evidence, we predicted 28,455 gene encoding loci (encoding 40,068 proteins) using the MAKER2 annotation pipeline (Holt, C. and Yandell, M. 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 was on average 2.1 genes per scaffold.

A synteny-based linear order of the ryegrass genome. The assembled Perennial ryegrass whole genome shotgun (WGS) 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, K. F. et al. 2009, Mayer, K. F. et al. 2011, Pfeifer, M. et al. 2013). Briefly, this makes use of 762 perennial ryegrass gene-derived markers that have been mapped in a two-way pseudo-testcross population at an average marker distance less than 0.9 cM (Studer, B. et al. 2012), of which 342 matched barley full-length cDNAs (see Figure 1 in Pfeifer, M. et al. 2013). Therefore, in absence of chromosomal information for the assembled ryegrass genome sequences, the high level of macro-co-linearity between Perennial ryegrass and barley (Pfeifer, M. et al. 2013, Figure S2) was exploited to identify syntenic relationships between ryegrass and the high-quality reference grass genomes *Brachypodium* (2010), rice (2005) and sorghum (Paterson, A. H. et al. 2009) (Figure S3). Based on the determined syntenic regions, between 2,887 and 4,153 Perennial ryegrass WGS scaffolds were assigned to individual chromosomes (Table S5). This allowed constructing a synteny-based linear gene order of the ryegrass genome and positioning in total 13,411 scaffolds and 11,311 annotated genes (10,464 unambiguously assigned to a LG; Supporting Information Data S1-S7).

Repetitive elements in the genome. The estimated repeat content from kmer spectrum (k=25 bp) analysis was 76 per cent. The high repeat content was also demonstrated by looking at the number of repeat-induced branches in a de-Bruijn assembly graph (Simpson, J. T. 2014). In comparison to other species the frequency of repeat branches

are both high, and not reduced as rapidly as we increase the kmer length (Figure S4). This makes genome assembly from short read sequence data challenging. Only 17.4 per cent of the final perennial ryegrass genome assembly was masked using RepeatMasker (Smith, A. F. A. et al. 2008) with a poaceae database (Table 2). This was in contrast to the masking of a one times genome coverage in PacBio reads that had been error-corrected with an abyss assembly (Table S6), where 34.1 per cent of the sequence was masked. The ratio of retroelements to DNA transposons was 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, D. et al. 2002). The self-incompatibility reaction occurs when both the *S* and the *Z* allele 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 (Shinozuka, H. et al. 2010, Yang, B. et al. 2008). Semi *in-vivo* pollinated pistils were flash frozen two hours 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. There were a greater number of genes significantly up-regulated than down-regulated in both cases (Figure S6-S7), and many of these genes were found to be regulated in the same manner, regardless of whether the pollination reaction was carried out with a compatible or incompatible pollen (Figure 1). We selected out genes that had at least a four-fold change in expression, and an FDR adjusted P-value of <0.001 (Supporting Information Data S9-S14), and focussed on those genes that were uniquely differentially expressed under either a compatible or an incompatible reaction.

Accepted Article

In a compatible pollination reaction we see the significant accumulation of a gene with sequence homology to obtusifolius 14 α -demethylase (fold change > 240), a sterol 14 α -demethylase and member of the cytochrome monooxygenase superfamily. It has already been shown in a species of wild potato that obtusifolius 14 α -demethylase accumulates in the ovary after pollination, and that only sustained levels (more than 3 days) can be maintained under a compatible reaction (O'Brien, M. et al. 2005). Here, we see an accumulation of obtusifolius 14 α -demethylase in the stigma within 2 hrs of pollination, an accumulation specific to pollination with compatible pollen. We also observed twelve 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 that were upregulated during a compatible reaction. An S-locus receptor like kinase (SRK) determines SI specificity at the stigma in sporophytic self-incompatibility systems, and was first identified in *Brassica oleracea* (Stein, J. C. et al. 1991). Using the GenomeZipper we can putatively place the two upregulated (~12 fold) g-type lectin s-receptor like kinases proximal to the z-locus on perennial ryegrass linkage group two (Manzanares, C 2013, Thorogood, D. et al. 2002), one co-locating with the locus, and the other ~14 cM from the locus. Another upregulated gene with homology to an Irr-like 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 hrs of pollination. Using the zipper anchored assembly, we see that both of these genes are co-located on linkage group 1 (LG1) together with the S-locus. Another gene co-located at the S-locus, and up-regulated over 50 fold shared sequence homology to an 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 24-nucleotide phased small RNAs (Song, X. et al. 2012), and a role for these in reproductive development, meiosis, or gamete formation has already been speculated (Fei, Q. et al. 2013). Seed yield in perennial ryegrass is generally regarded as being low and

unreliable(Elgersma, A 1990). This affects seed producers and impacts the commercial viability of a variety. It has already been demonstrated that large amounts of incompatible pollen in the pollen cloud leads to reduced seed yield per panicle and per plant(Studer, B. et al. 2008). Functional markers at S and Z may enable any SI-mediated loss in seed yield to be avoided.

The pollen transcriptome and characterisation of allergen homologs. The proteins functioning as the male component in the SI mechanism need to be functional at pollen maturity. We therefore looked at the relative expression of genes within the pollen transcriptome. Transcriptional activity in mature pollen is quite distinct to transcriptional activity in leaves, particularly in relation 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 to 13.9 percent 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 percent 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 organisation and protein phosphorylation (Figure 2d). These included three pectinesterase inhibitors, a methyl esterase inhibitor, and three pectinase proteins. It also included many genes that shared homology to well characterised allergens from a range of species.

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 raise the antibodies Immunoglobulin E (IgE) in response to proteins on the exterior coat of inhaled pollen, and these attach themselves to mast cells, triggering them to release histamine and causing an allergic reaction. We developed Hidden Markov Model (HMM) profiles representing pollen allergens from various species (Figure 2a, Table S7), and used them to scan the Perennial ryegrass gene predictions for homologs. This resulted in the identification of 630 non-redundant hits, belonging to 17 protein families (Figure 2a, Table 3). The largest protein family with pollen

allergen homologs is the Polcalcin family (EF-hand CA-binding proteins) with 145 members. Other large protein families represent 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 will be secreted and directed across the endoplasmatic reticulum. Major pollen allergens belonging to protein families without secretion signals (like Polcalcins, Profilins or Bet V I allergens) were reported in several plant species.

Using the pollen RNA sequencing data we found that 60 percent of the pollen allergen homologs had detectable expression in mature pollen (Supporting Information Data S8), and 26 had greater than 1000 transcripts per million (TPM, Table S8). Of the pollen allergens with detectable expression levels within mature pollen, five genes fell within the twenty with the highest relative expression levels. These comprised three members of the expansin family (two Lol pl signature proteins, one Lol pII-like protein), a profilin protein and a pectate lyase, altogether accounting for 6.4 percent 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 (Cosgrove, D. J. and Li, Z. C. 1993, McQueen-Mason, S. et al. 1992). It is hypothesised that expansins break non-covalent bonds between cell-wall polysaccharides, thereby permitting pressure dependent expansion of the cell (McQueen-Mason, S. and Cosgrove, D. J. 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 (CBD). They are divided into four subclasses: expansin A, expansin B, expansin-like A and expansin-like B. We find members of all four subclasses within our group 1 (78 genes) and group 2 (7 genes) proteins (Figure 2b). Lol pl is a Group 1 grass pollen allergen belonging to the β -expansin family, and is the most prominent pollen allergen in grasses with ~90 percent of allergic patients showing Lol pl specific IgE in their sera (Laffer, S. 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, P. and Marsh, D. G. 1965a). Group 1 and 2 pollen allergens share sequence similarity at their C-terminal part (Figure 2c), which covers nearly 90 percent of the Group 2 allergen sequences and is discerned as a Cellulose-binding-like domain (CBD). In 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 Barwin-like double psi beta-barrel structure (DPBB) domain. 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 specialised 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 hypothesised that Group 1 grass pollen allergens have a role to play in softening maternal cell walls to allow the pollen penetrate the stigma and style (Cosgrove, D. J. et al. 1997, Cosgrove, D. J. 2000).

Discussion

We have sequenced and assembled a draft genome of perennial ryegrass, the first species from the *Lolium/Festuca* complex to be associated with a published genome assembly. 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 with PacBio reads. The resulting assembly covers 1,128 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 both inbreeding is restricted, and that diversity is preserved. Grasses utilise 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, D. 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 draft genome sequence should now accelerate the 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 happens within minutes after pollination. Many of the genes 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 an involvement in pollen-stigma recognition (Cosgrove, D. J. et al. 1997). Allergic rhinitis (commonly known as hayfever) is estimated to affect between 10 and 30 per cent (WAO, www.worldallergy.org) of the world's population, and is the result of inhaling allergens to which the body has become sensitised. Grass pollen is one of the major causes of allergic rhinitis worldwide, and some of the first pollen allergens characterised were identified over 50 years ago in perennial ryegrass (Johnson, P. and Marsh, D. G. 1965b). Group 1 allergens are glycoproteins that are recognised by over 90 per cent of allergic patients, and Lol p 1, the Group 1 allergen of perennial ryegrass has previously been cloned and sequenced (Griffith, I. J. et al. 1991). Our search of the genome identified a small cluster within the larger expansin protein family containing five homologs to the published Lol p 1 protein, of which four were relatively highly expressed in mature pollen grains. One of these genes was ranked twelfth in terms of relative expression in mature pollen grains. We have also identified homologs of many other important allergens that are expressed to high levels in

Accepted Article

mature pollen grains of perennial ryegrass. Besides treating the symptoms of allergies, e.g. with antihistamines, there may be opportunities to use hypoallergenic recombinant expressed allergens for patient-tailored immunotherapy (Singh, M. B. and Bhalla, P. L. 2003). Characterisation 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 characterisation 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 standard methods and passed through a DNeasy plant spin column (Qiagen) to remove contaminants. The genotype originated from IBERS, Aberystwyth University, UK, and has undergone inbreeding for six generations. Illumina paired-end (PE) libraries with mean fragment lengths of 140 bp and 550 bp were prepared from genomic DNA. Illumina mate-pair (MP) libraries were also prepared with mean insert sizes of 1.8 Kb, 3.4 Kb, and 8.6 Kb. DNA was fragmented by shearing using nebulisers according to the Illumina Mate Pair Library Kit v2. The PE libraries were prepared using NEBNext DNA Sample Preparation Kit with Illumina adaptors according to the manufacturer's specification. Libraries were sequenced on either an Illumina GAIIx or HiSeq 2000 using TruSeq chemistry according to Illumina recommendation. HiSeq 2000 sequence data (fastq) was extracted using CASAVA v1.8. Long Jump Distance (LJD) libraries with insert sizes of 7.7, 26.2, and 29.7 Kb were also prepared as a custom service provided by Eurofins Genomics (Ebersberg, Germany) and

sequenced on an Illumina 2000 system. The sequencing effort for each library varied (Table S1). For PacBio whole genome shotgun sequencing, 40 µg of DNA was extracted from P226/135/16, and four independent libraries (10 Kb inserts) were prepared for sequencing (C2 chemistry) on 166 SMRT cells. This was carried out at the Genome Sequencing and Analysis Core Resource at Duke University, NC, USA. The total number of subreads was 14,843,437 (18.3 Gb), the N50 read length was 3,461 bp, and the average read length was 1,236 bp.

Genome assembly and evaluation. The Illumina data was assembled with ALLPATHS-LG (release 47760, ref) (Gnerre, S. 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/CPU), making use of the different libraries in specific ways (Table S9). The PacBio data was used for gapfilling with PBJelly (version 12.9.14, ref) (English, A. C. et al. 2012) after aligning the reads to the ALLPATHS-LG assembly with BLASR (Chaisson, M. J. and Tesler, G. 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 (Supplementary Figure S8). In order to evaluate the completeness of the assembly in terms of gene content we performed CEGMA analysis (Parra, G. et al. 2007). We also aligned a *de-novo* transcriptome assembly to the assembled genome using BLAT (Kent, W. J. 2002), and evaluated the alignments with IsoBLAT (<https://github.com/josephryan/isoblat>).

Genome annotation. We had multiple RNA-seq data sets available to us from various expression studies, performed on multiple plant sections, multiple genotypes, and at multiple time points in different environments (Table S10). This included RNA-seq data from the same inbred genotype used for the genome assembly (Farrell, J. D. et al. 2014). Reads were error-corrected before assembly using the ErrorCorrectReads.pl script bundled with ALLPATHS-LG (Gnerre, S. et al. 2011). We performed individual *de-novo* transcriptome assemblies for each genotype using Trinity (Grabherr, M. G. et al. 2011). After assembly the error corrected reads were aligned back against the assembled transcripts and abundance estimates were calculated using RSEM (Li, B. and Dewey, C.

N. 2011). We filtered out transcripts with less than one percent of the per-component (IsoPct) expression level. PASA (Haas, B. J. 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, C. and Yandell, M. 2011). We also provided a protein data set from *Brachypodium distachyon* (2010) as evidence, and repeat masking was performed with RepeatMasker (3.3.0, (Smith, A. F. A. et al. 2008)) using a poaceae repeat database. Functional annotation of the resulting gene predictions was performed with InterProScan V5 (Jones, P. et al. 2014).

Repeat Content. We performed an assembly of the Illumina data using Abyss (Simpson, J. T. et al. 2009) with a kmer size of 73 bp. The total contig length was 2,144 Mb, and this assembly was used to error correct the PacBio reads using ECTools (Lee, Hyan et al. 2014). We randomly sampled a one times genome coverage in reads greater than 2 Kb from the error corrected reads (Table S11). We used RepeatMasker (3.3.0, (Smith, A. F. A. et al. 2008)) and a poaceae repeat database to annotate and quantify repeats in this sample of error corrected PacBio reads.

GenomeZipper. On basis of the identified syntenic relationships to the high-quality reference grass genomes of the Perennial ryegrass whole genome shotgun (WGS) scaffolds, which could be assigned to individual chromosomes, were ordered by using the GenomeZipper approach (Mayer, K. F. et al. 2011, Pfeifer, M. et al. 2013). Briefly, guided by a genetic marker map the GenomeZipper determines orthologous gene relationships and syntenic regions between a target genome (i.e. Perennial ryegrass) and high-quality reference genome sequences (i.e., *Brachypodium distachyon* (2010), *Oryza sativa* (2005) and *Sorghum bicolor* (Paterson, A. H. 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, B. et al. 2012) and proteins annotated for the three used reference genomes. Furthermore, the annotated Perennial ryegrass genes were integrated into the genome backbones as determined by bbh searches against the marker sequences and coding sequences of the anchored reference proteins [BLASTN with $E < 1e-10$ and minimum alignment length 100bp and minimum alignment identity of 80 percent (marker), 75 percent (*Brachypodium*) and 70 percent (rice and sorghum), respectively]. Then, the chromosome-assigned Perennial ryegrass WGS assemblies were positioned in the GenomeZipper backbones via sequence homology to the respective elements (i.e. genetic marker, reference proteins and annotated Perennial ryegrass genes). Additionally, publicly available Perennial ryegrass EST sequences (Pfeifer, M. et al. 2013, Studer, B. et al. 2012) were ordered attached to the GenomeZipper framework [BLASTN with $E < 1e-10$ and minimum alignment length 100bp and minimum alignment identity of 80 percent (marker), 75 percent (*Brachypodium*) and 70 percent (rice and sorghum), respectively].

Gene expression analysis. We used genotypes that are known for their *S* and *Z* allele composition and represent a diverse set of *S*-locus alleles. The plant F1-30, being heterozygous $S_1S_2Z_1Z_2$ (the allele numbers are allocated at random) was complemented with pollen from the variety Foxtrot $S_3S_4Z_3Z_4$. Three biological replicates of pollen and stigma samples were collected from the different genotypes. The pollen was harvested into clear bags and transferred into microcentrifuge 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 different 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-specific samples, self-incompatible and self-compatible pollen-stigma interactions were analysed. Using *in-vivo* pollinations, three biological replicates from both self-pollinations (self-incompatible) and two replications from cross-pollinations (self-compatible) were collected. The semi *in-vivo* pollinations were conducted as described by Thorogood et al., (Thorogood, D. et al. 2002). The genotype F1-30 was used for self-pollination, while F1-30 was pollinated with pollen from an unrelated genotype of the cultivar Foxtrot (DLF-Trifolium, Store Heddinge,

Denmark) for the cross-pollination. In both cases, semi *in-vivo* pollinated pistils were collected approximately two hours 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 the RNeasy™ Plant Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's instructions. RNA from the pollen samples was extracted using the PicoPure™ RNA isolation Kit (Arcturus Bioscience, Inc., CA, USA) following the manufacturer's instructions. RNA integrity was measured with an RNA 6000 Nano Labchip™ on the Agilent 2100 Bioanalyzer™ (Agilent Technologies, Santa Clara, CA). The sequencing libraries were generated using the TruSeq Stranded mRNA Sample Preparation kit (Illumina, San Diego, USA) according to the manufacturer's instructions. The cDNA libraries were sequenced using 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 30M reads/sample (AROS Applied Biotechnology AS, Aarhus, Denmark). Leaf samples were taken from the genotype NGB/09 that was growing under ambient conditions and flash frozen in liquid nitrogen. Total RNA was extracted using RNeasy® Plant Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. RNA integrity was measured with an RNA 6000 Nano Labchip™ on the Agilent 2100 Bioanalyzer™ (Agilent Technologies, Santa Clara, CA). 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, B. and Salzberg, S. L. 2012) with the following settings (--sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min L,0,-0.1 -I 1 -X 1000 --no-mixed --no-discordant -k 200). RSEM (Li, B. and Dewey, C. N. 2011) was used to calculate Transcripts Per Million (TPM) for each gene, and EdgeR (Robinson, M. D. et al. 2010) was used to compare gene expression in two pairwise comparisons; (1) unpollinated stigma vs. stigma pollinated with compatible pollen, and (2) unpollinated stigma vs. stigma pollinated with incompatible pollen.

Identification of allergen homologs. Publicly available protein sequences of pollen allergens were collected from the NCBI/UniProt protein sequence databases. Beside known grass pollen allergens like Lol p1, Lol p2, Lol p3, Lol p5, 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 BLAST (blastp) (Altschul, S. F. et al. 1990) searches against a non-redundant database containing *Viridiplantae* protein sequences. Sequences representing significant BLAST hits (at least 80 percent similarity level, above 80 scores) were collected and aligned to the reference sequences of each protein families. After visual inspection, outscoring sequences were removed from the alignments, keeping 5 to 50 sequences in the core alignments representing each protein family, depending on availability and diversity of the family. When possible, monocot sequences were preferred above dicot sequences during selection. The remaining sequences were re-aligned using Clustal Omega (Sievers, F. and Higgins, D. G. 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 (v3.1b1, <http://hmmer.org>). The profile matrices were used to scan the protein sequences from the perennial ryegrass genome, using the *hmmsearch* 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 standalone InterProScan 5 (Jones, P. et al. 2014) pipeline by scanning the PANTHER, PROSITE profiles, Pfam and SUPERFAMILY databases. This pipeline was also used for prediction of transmembrane topology and signal peptides by integrating Phobius (Kall, L. et al. 2004) and SignalP (Emanuelsson, O. et al. 2007) utilities.

Accession Numbers

The perennial ryegrass genome and transcriptome assemblies can be downloaded in fasta format, and accessed through a genome browser at the following address

<http://185.45.23.197:5080/ryegrassgenome> . The complete SI expression study data has been

deposited in ArrayExpress with the accession number E-MTAB-3760.

Acknowledgements

This work was supported by the Godtfred Birkedal Hartmann Foundation, and the UK Biotechnology and Biological Sciences Research Council. We thank the Duke GCB Genome Sequencing Shared Resource, which provided a PacBio sequencing service.

Short Legends for Supporting Information

Figure S1: Kmer spectra analysis.

Figure S2: Synteny between ryegrass and barley.

Figure S3: Coverage of the *Brachypodium* genome by assembled *Lolium* WGS scaffolds and syntenic.

Figure S4: Estimated repeat branch rate as a function of K .

Figure S5: TPM for leaf and pollen samples.

Figure S6: MA and volcano plots for DE genes during a compatible pollination reaction.

Figure S7: MA and volcano plots for DE genes during an incompatible pollination reaction.

Figure S8: Histogram of PacBio reads addressing gaps.

Table S1: Illumina sequencing data overview.

Table S2: Summary statistics of gap filling.

Table S3: Summary statistics of 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.

Supporting Information Data S1-S7: spreadsheets with the perennial ryegrass GenomeZipper data, one file for each of the seven linkage groups.

Supporting Information Data S8: spreadsheet showing the expression levels (as TPM) of the 630 genes with sequence homology to known pollen allergens, in both mature pollen and leaf samples.

Supporting Information Data S9: spreadsheet showing genes significantly down regulated in both a compatible and incompatible pollination reaction.

Supporting Information Data S10: spreadsheet showing genes significantly down regulated in a compatible pollination reaction.

Supporting Information Data S11: spreadsheet showing genes significantly down regulated in an incompatible pollination reaction.

Supporting Information Data S12: spreadsheet showing genes significantly up regulated in both a compatible and incompatible pollination reaction.

Supporting Information Data S13: spreadsheet showing genes significantly up regulated in a compatible pollination reaction.

Supporting Information Data S14: spreadsheet showing genes significantly up regulated in an incompatible pollination reaction.

Reference List

(2005) The map-based sequence of the rice genome. *Nature* **436**:793-800.

(2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**:763-768.

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 Research* **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., Bedinger,P., and Durachko,D.M. (1997) Group I allergens of grass pollen as cell wall-loosening agents. *Proc.Natl.Acad.Sci.U.S.A* **94**:6559-6564.

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.

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

Emanuelsson,O., Brunak,S., von,H.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., Wang,M., Vee,V., Qu,J., Qin,X., Muzny,D.M., Reid,J.G., Worley,K.C., and

Gibbs,R.A. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**:e47768.

Farrell,J.D., Byrne,S., Paina,C., and Asp,T. (2014) De novo assembly of the perennial ryegrass transcriptome using an RNA-Seq strategy. *PLoS One* **9**:e103567.

Fe, D., Ashraf, B., Byrne, S., Czaban, A., Roulund, N., Lenk, I., Asp, T., Pedersen, M. G., Jensen, C., Janss, L., and Jensen, J. Prospects for introducing genomic selection info forage grass breeding. European Grassland Federation. 2014.

Ref Type: Conference Proceeding

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

Gnerre,S., Maccallum,I., Przybylski,D., Ribeiro,F.J., Burton,J.N., Walker,B.J., Sharpe,T., Hall,G., Shea,T.P., Sykes,S., Berlin,A.M., Aird,D., Costello,M., Daza,R., Williams,L., Nicol,R., Gnirke,A., Nusbaum,C., Lander,E.S., and Jaffe,D.B. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc.Natl.Acad.Sci.U.S.A* **108**:1513-1518.

Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q., Chen,Z., Mauceli,E., Hacohen,N., Gnirke,A., Rhind,N., di,P.F., Birren,B.W., Nusbaum,C., Lindblad-Toh,K., Friedman,N., and Regev,A. (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., Theerakulpisut,P., Avjioglu,A., Davies,S., Hough,T., Singh,M.B., Simpson,R.J., Ward,L.D., and . (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., Wortman,J.R., Smith,R.K., Jr., Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., Salzberg,S.L., and White,O. (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. The contribution of conventional plant breeding to forage crop improvement. *International Grasslands*. *International Grasslands* , 71-78. 1999.

Ref Type: Conference Proceeding

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.

Johnson,P. and Marsh,D.G. (1965a) 'Isoallergens' from rye grass pollen. *Nature* **206**:935-937.

Johnson,P. and Marsh,D.G. (1965b) 'Isoallergens' from rye grass pollen. *Nature* **206**:935-937.

Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., Pesseat,S., Quinn,A.F., Sangrador-Vegas,A., Scheremetjew,M., Yong,S.Y., Lopez,R., and Hunter,S. (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.

Laffer,S., Spitzauer,S., Susani,M., Pairleitner,H., Schweiger,C., Gronlund,H., Menz,G., Pauli,G., Ishii,T., Nolte,H., Ebner,C., Sehon,A.H., Kraft,D., Eichler,H.G., and Valenta,R. (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., McCombie,W.R., and Schatz,M. (2014) Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv*.

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.

Manzanares, C. Genetics of self-incompatibility in perennial ryegrass (*Lolium perenne* L.). 2013.
Ref Type: Thesis/Dissertation

Mayer,K.F., Martis,M., Hedley,P.E., Simkova,H., Liu,H., Morris,J.A., Steuernagel,B., Taudien,S., Roessner,S., Gundlach,H., Kubalakova,M., Suchankova,P., Murat,F., Felder,M., Nussbaumer,T., Graner,A., Salse,J., Endo,T., Sakai,H., Tanaka,T., Itoh,T., Sato,K., Platzer,M., Matsumoto,T., Scholz,U., Dolezel,J., Waugh,R., and Stein,N. (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**:1249-1263.

Mayer,K.F., Taudien,S., Martis,M., Simkova,H., Suchankova,P., Gundlach,H., Wicker,T., Petzold,A., Felder,M., Steuernagel,B., Scholz,U., Graner,A., Platzer,M., Dolezel,J., and Stein,N. (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* **151**:496-505.

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.U.S.A* **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 obtusifolios 14 α -demethylase from *Solanum chacoense* Bitt., a pollination- and fertilization-induced gene with both obtusifolios and lanosterol demethylase activity. *Plant Physiol* **139**:734-749.

Parra,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., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A., Schmutz,J., Spannagl,M., Tang,H., Wang,X., Wicker,T., Bharti,A.K., Chapman,J., Feltus,F.A., Gowik,U., Grigoriev,I.V., Lyons,E., Maher,C.A., Martis,M., Narechania,A., Otiillar,R.P., Penning,B.W., Salamov,A.A., Wang,Y., Zhang,L., Carpita,N.C., Freeling,M., Gingle,A.R., Hash,C.T., Keller,B., Klein,P., Kresovich,S., McCann,M.C., Ming,R., Peterson,D.G., Mehboob,u.R., Ware,D., Westhoff,P., Mayer,K.F., Messing,J., and Rokhsar,D.S. (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* **48**: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.

Smith, A. F. A., Hubley, R., and Green, P. RepeatMasker Open-3.0. 2008.
Ref Type: Computer Program

Song,X., Li,P., Zhai,J., Zhou,M., Ma,L., Liu,B., Jeong,D.H., Nakano,M., Cao,S., Liu,C., Chu,C., Wang,X.J., Green,P.J., Meyers,B.C., and Cao,X. (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.U.S.A* **88**:8816-8820.

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

Studer,B., Jensen,L.B., Hentrup,S., Brazauskas,G., Kolliker,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.

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 (Edinb.)* **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.

Tables

Table 1: Summary statistics of perennial ryegrass genome assembly.

	Contigs (> 1Kb)	Scaffolds (> 1Kb)
Sum (Mb)	718	1,128
Total Number	67,024	48,415
N50 (bp)	16,370	70,062
N90 (bp)	5,500	14,249
Max (bp)	99,259	512,939
Captured Gaps (Mb)		292

Table 2: Repeat content in the final perennial ryegrass genome assembly and in error-corrected PacBio reads (minimum length of 2 Kb, 1 X genome coverage). Search performed against the RM Poaceae database (version 20130433).

	Assembly		EC PacBio Reads	
	No.	%	No.	%
Retroelements	235669	13.03	1155138	28.82
SINEs:	6326	0.10	7745	0.05
LINEs:	42271	3.01	55712	1.34
RTE/Bov-B	1572	0.09	3233	0.05
L1/CIN4	40699	2.92	52479	1.29
LTR elements:	187072	9.92	1091681	27.43
Ty1/Copia	57106	3.87	198182	7.05
Gypsy/DIRS1	127754	5.98	887744	20.33
DNA transposons	98229	3.13	196172	3.53
hobo-Activator	8959	0.22	11146	0.1
Tc1-IS630-Pogo	19090	0.30	26902	0.16

En-Spm	2366	0.24	7818	0.35
Tourist/Harbinger	16427	0.62	21671	0.32
Unclassified	7080	0.17	8904	0.08
Small RNA	4625	0.09	13597	0.48
Satellites	6152	0.08	18136	0.2
Simple repeats	149747	0.82	373282	0.86
Low complexity	29361	0.19	71397	0.18

Table 3: Proteins identified in perennial ryegrass with homology to previously identified pollen allergens. The protein group numbers correspond to those in Figure 4a.

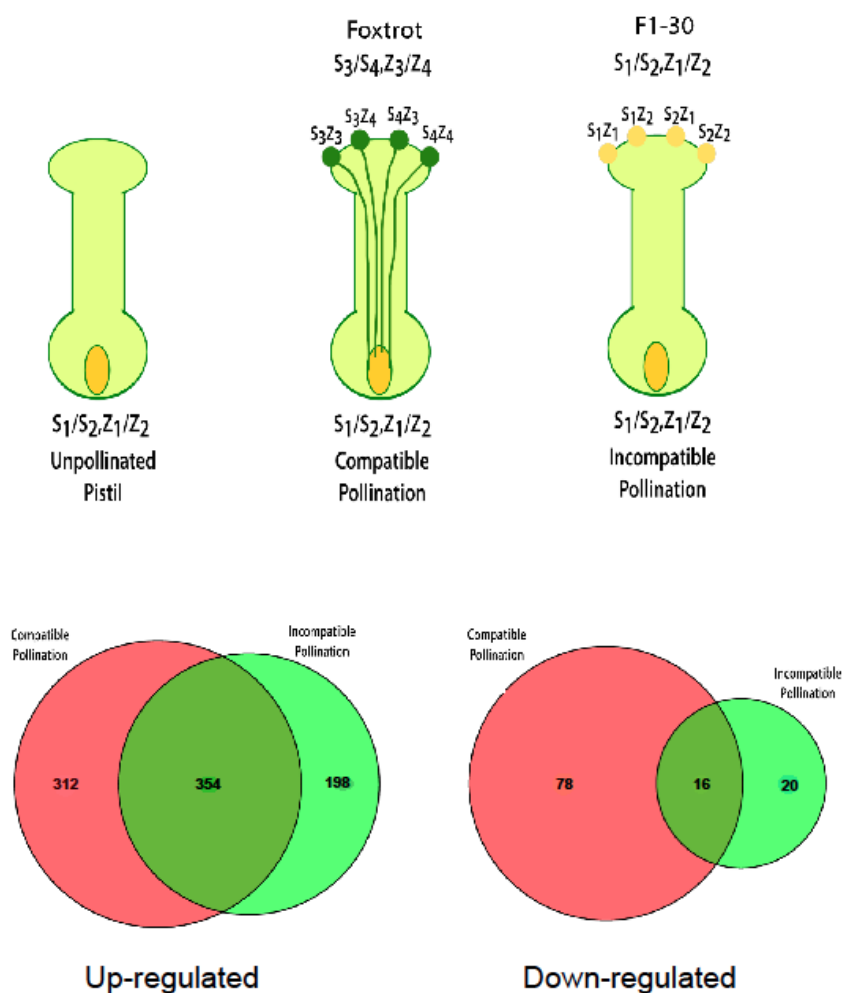
Protein Group	Description	Code	Total	peptide
1	Expansins, Lol p1 signatures	LOL1	78	+
2	Expansins, Lol p2 signatures	LOL2	7	+
3	Copper/zinc superoxide dismutase (SODC) family	OLE5	6	-
4	Glycosyl hydrolases family 17	OLE9	91	-?
5	Pectate lyases, Amb family signatures	AMBR	19	+
6	Ribonucleases, Poa pl signatures	POAP	6	+
7	Glycosyl hydrolases family 28, Parellel beta-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	Gamma-thionin family, Defensin-like proteins	ART1	15	+
13	Pathogenesis-related proteins, Bet v I family signatures	BETV	26	-

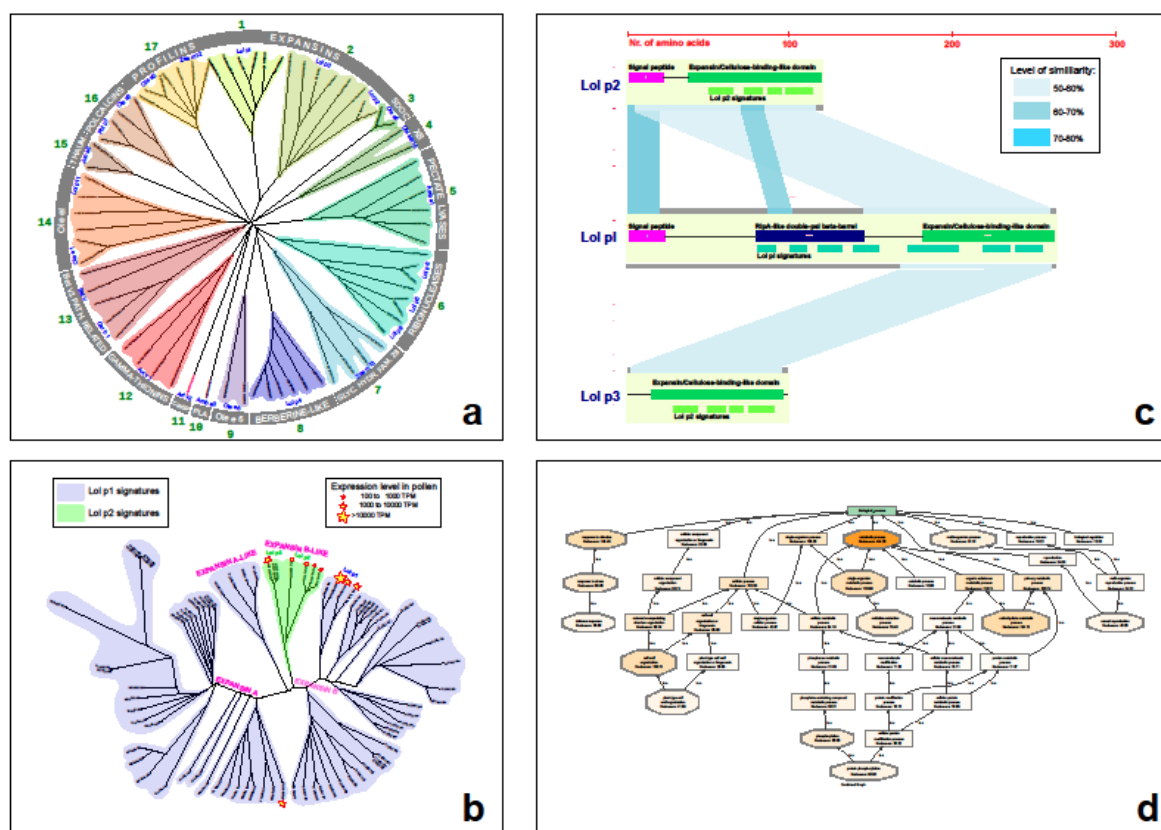
14	Lol p11-like pollen allergens, Ole e1 family signatures	OLE1	28	+
15	Pathogenesis-related proteins, Thaumatin family signatures	JUN3	34	+
16	Polcalcins, EF-hand calcium-binding domain profile proteins	EFHA	145	-
17	Profilins	PROF	4	-
TOTAL			630	

Figure legends

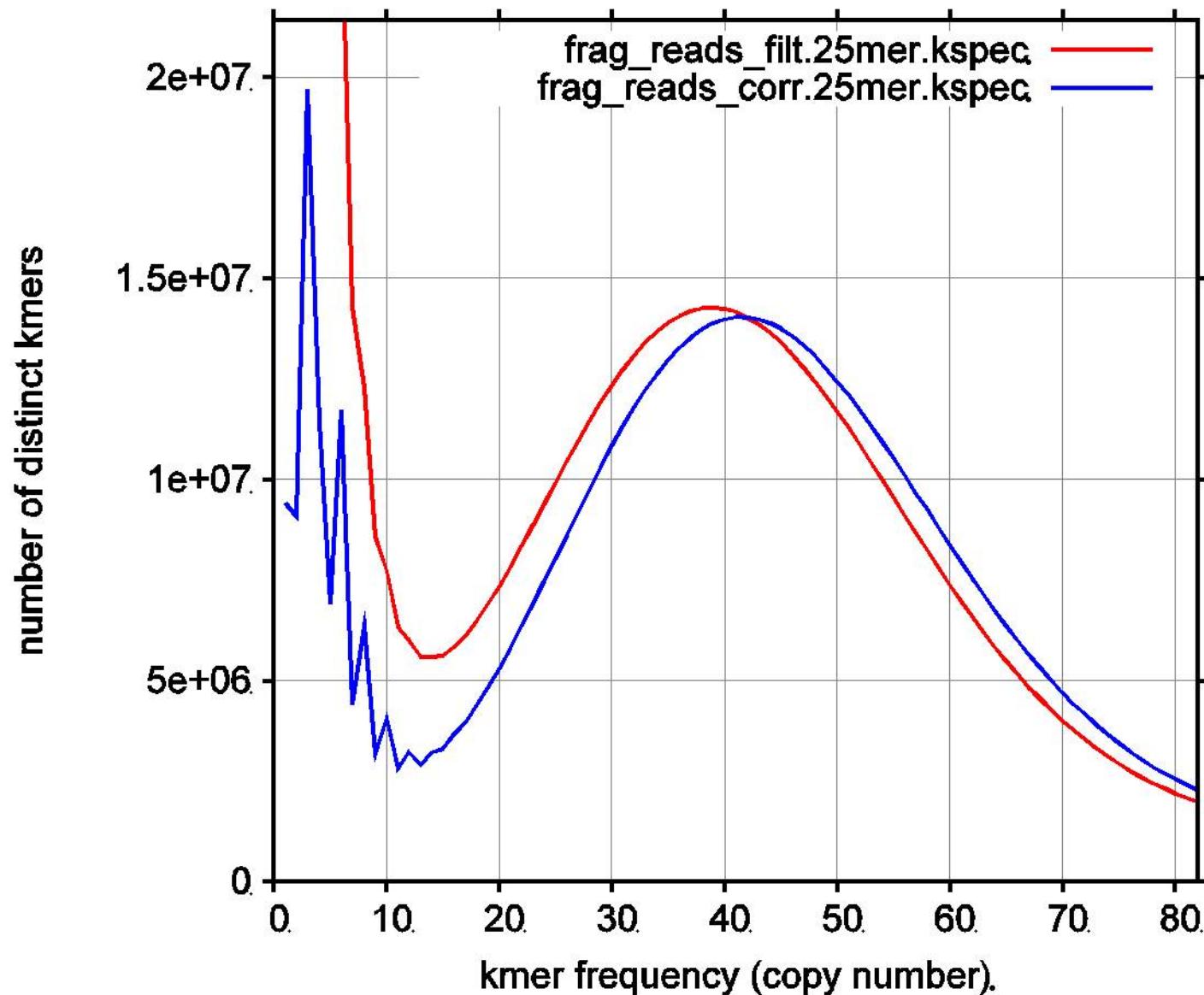
Figure 1: Self-incompatibility expression study. (a) Design of self-incompatibility transcriptome study and (b) overview of differentially expressed genes for the various pairwise comparisons.

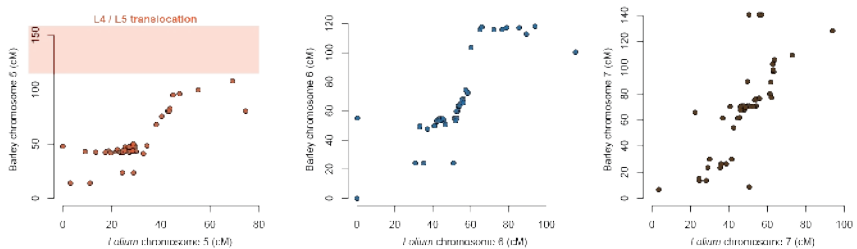
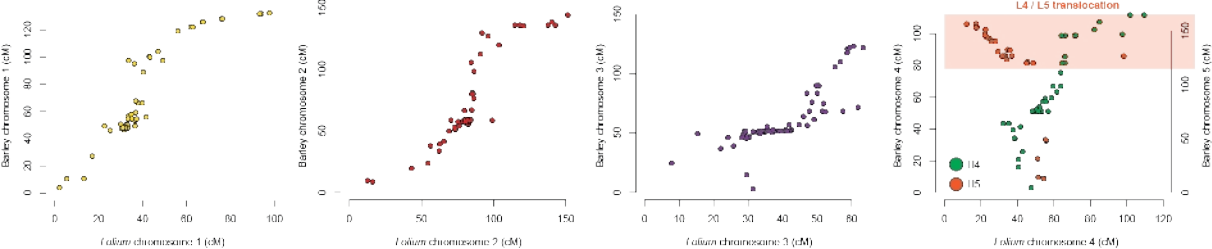
Figure 2: Classification of allergen genes. (a) Classification and phylogenetic tree of allergen genes from various species that were used as references to build the hmm profiles. They are classified into 17 grass pollen allergen groups. (b) Phylogenetic analysis of 78 Group 1 and 7 Group 2 grass pollen allergens that encompasses α -expansin, β -expansin, expansin-like A and expansin-like B subclasses from 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 less than 30 sequences were filtered out.

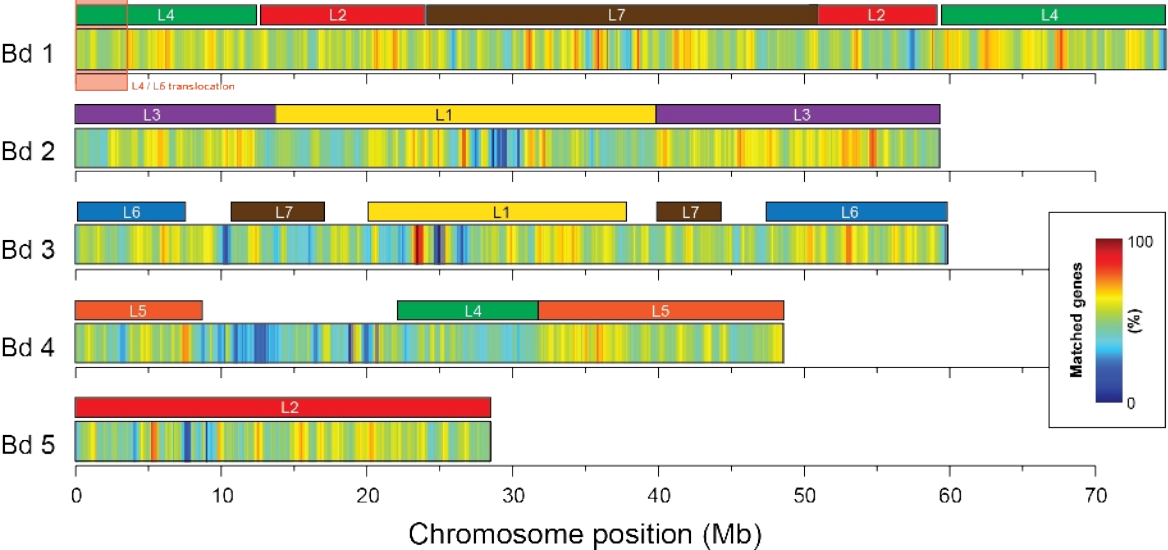




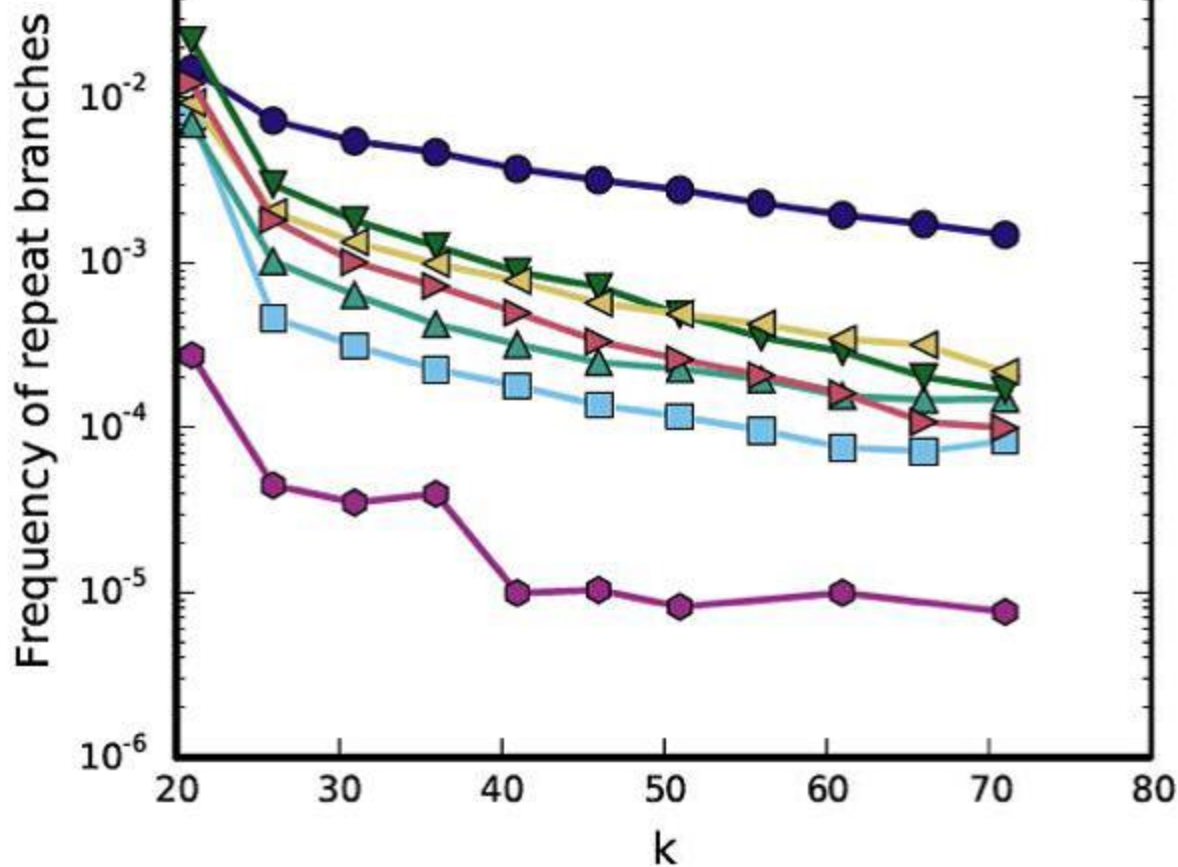
Number of distinct kmers



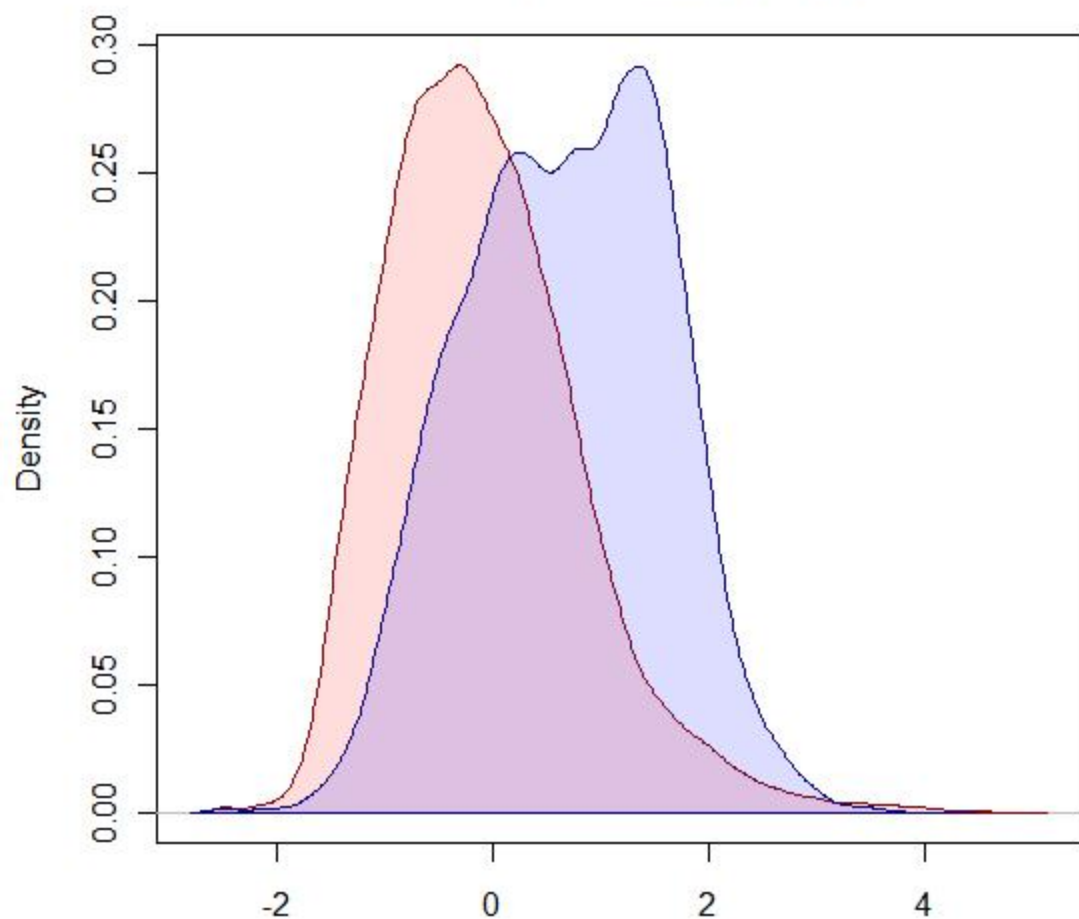




repeat branches in k-de Bruijn graph

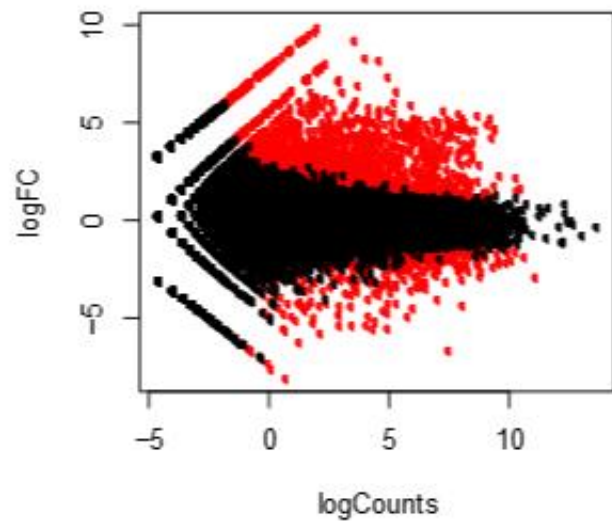


Density Plot $\log_{10}(\text{TPM})$

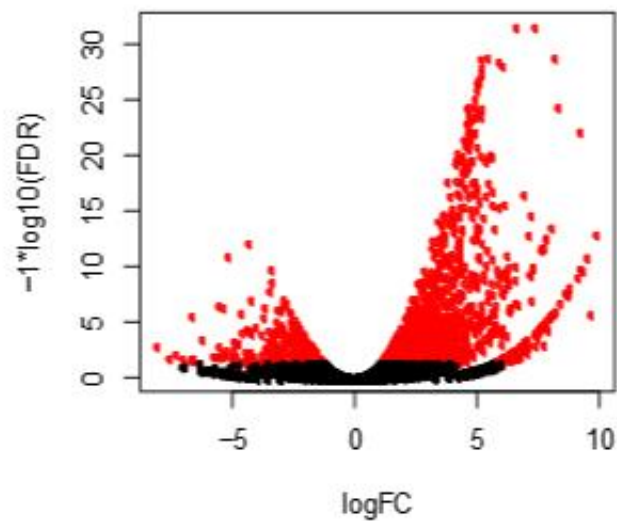


N = 40068 Bandwidth = 0.1039

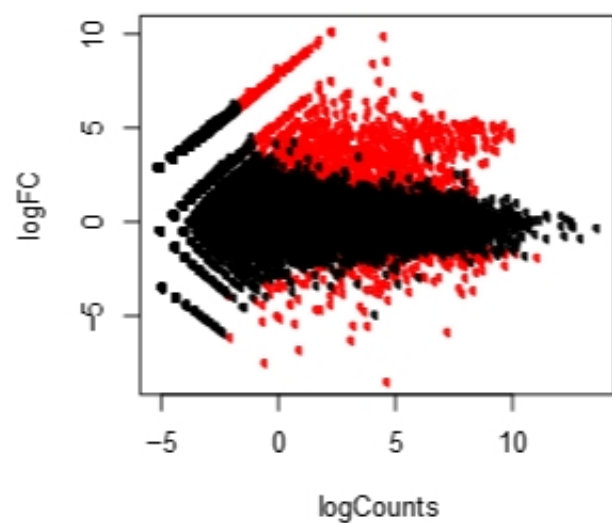
MA plot



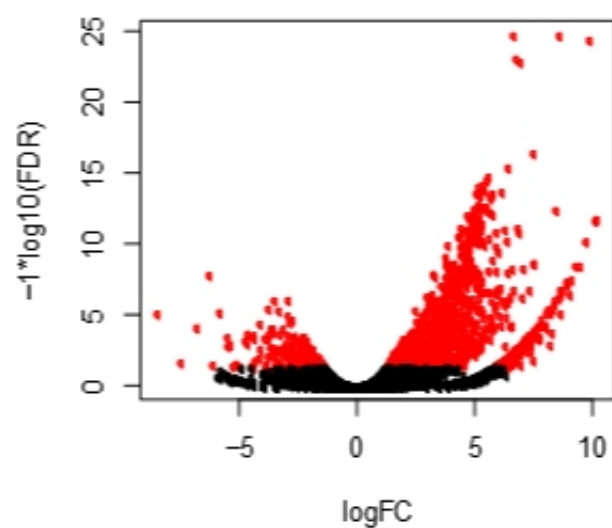
Volcano plot



MA plot



Volcano plot



Histogram of PacBio reads addressing closed gaps

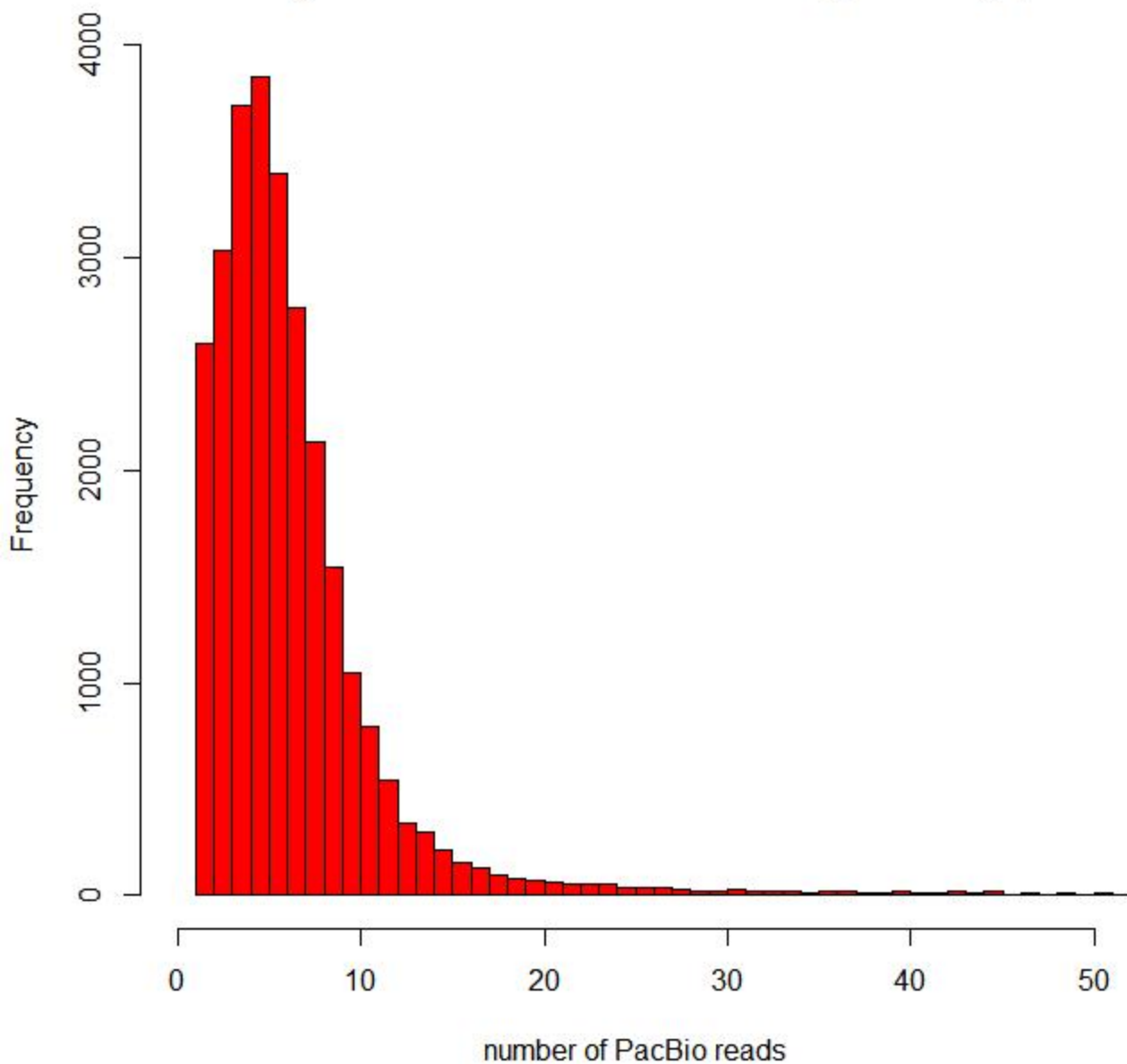


Table S1: Summary of Illumina sequencing data for P226/135/16.

Library Type	Fragment size (bp)	Read length (bp)	Pairs
PE	140 ¹	2 X 100	754,864,255
PE	550 ¹	2 X 100	259,937,464
MP	1,800 ¹	2 X 75	35,662,577
MP	1,800 ¹	2 X 35	88,085,902
MP	3,400 ¹	2 X 75	39,488,199
MP	8,600 ¹	2 X 75	55,911,814
LJD ³	7,765 ²	2 X 100	19,150,772
LJD ³	26,265 ²	2 X 100	26,825,212
LJD ³	29,765 ²	2 X 100	16,203,472

¹fragment size determined after mapping reads to fragment assembly

²expected fragment size from lab measurements

³reads assigned LJD evidentially have an insert size of several kb (depending on specific library type).

PE = paired end, MP = mate pair, LJD = long jumping distance (Eurofins Genomics)

Table S2: Summary of gap, contig, and scaffold statisitcs before and after gap filling with PacBio reads. Gap filling was performed with PBJelly.

	Before	After
<i>Gap statistics</i>		
Number gaps	76,244	48,128
Total gap length (bp)	441,735,717	292,441,192
Mean gap length (bp)	5,793	6,076
Median gap length (bp)	2,000	2,101
Gap N50 (bp)	17,857	16,916
Gap N90 (bp)	2,881	4,189
<i>Contig statistics</i>		
Number contigs	98,159	67,024
Contig N50 (bp)	9,668	16,370
Contig N90 (bp)	2,639	5,500
Total contig length (bp)	592,069,939	718,155,397
<i>Scaffold statistics</i>		
Number scaffolds	48,415	48,415
Scaffold N50 (bp)	69,156	70,062
Scaffold N90 (bp)	13,858	14,249
Total scaffold length (bp)	1,110,739,743	1,128,044,065

Table S3: Summary statistics of the ALLPATHS-LG assembly.

Metric	ALLPATHS-LG Assembly
Minimum contig size (bp)	1000
Number of contigs	130,838
Number of contigs per Mb	118.6
Number of scaffolds	48,415
Total contig length (bp)	667,918,552
Total scaffold length, with gaps (bp)	1,103,168,026
N50 contig size (Kb)	8.3
N50 scaffold size (Kb)	69
Number of scaffolds per Mb	43.89

Table S4: Statistics on the completeness of the genome based on 248 ultra-conserved Core Eukaryotic Genes (CEGs). Results are shown for complete coverage and partial coverage of CEGs. Group 1 CEGs are the least conserved, and group 4 CEGs are the most conserved across species.

	No. Proteins	Completeness (%) ¹	Total ²	Average ³	Ortho (%) ⁴
<i>Complete</i>	226	91.13	455	2.01	62.39
Group 1	59	89.39	110	1.86	55.93
Group 2	49	87.50	87	1.78	48.98
Group 3	55	90.16	117	2.13	63.64
Group 4	63	96.92	141	2.24	77.78
<i>Partial</i>	239	96.37	541	2.26	69.04
Group 1	63	95.45	127	2.02	58.73
Group 2	52	92.86	104	2.00	57.69
Group 3	61	100.00	148	2.43	72.13
Group 4	63	96.92	162	2.57	85.71

¹percentage of 248 CEGs present in the genome

²total number of CEGs present including putative orthologs

³average number of orthologs per CEG

⁴percentage of detected CEGs that have more than one ortholog

Table S5: GenomeZipper statistics for individual linkage perennial ryegrass linkage groups.

	L1	L2	L3	L4	L5	L6	L7	non.-red. Σ
Available <i>Lolium</i> marker	88	125	111	154	79	86	119	762
GenomeZipper loci	2,794	3,497	3,378	3,325	2,364	2,270	3,125	20,783
Syntenic <i>Brachypodium</i> genes	1,752	2,366	2,146	2,198	1,559	1,397	1,781	13,077
Syntenic rice gene	1,417	1,385	1,975	1,774	1,109	1,347	1,674	10,681
Syntenic sorghum genes	1,380	1,848	1,991	1,945	9,38	1,322	1,780	11,186
Integrated <i>Lolium</i> genes	1,591	2,042	2,015	2,069	1,376	1,379	1,744	11,311
Anchored <i>Lolium</i> scaffolds	3,501	4,153	3,966	3,954	2,923	2,887	3,846	13,411
Anchored <i>Lolium</i> ESTs	1,712	1,893	1,867	1,980	1,405	1,465	1,794	4,892

Table S6: Summary statistics of the Abyss unitig assembly of the perennial ryegrass paired-end fragment reads.

	Abyss Assembly
Number of sequences	18,808,785
Total unitig length	1,849,832,347 bp
Minimum unitig	50 bp
Maximum unitig	30,668 bp
N20	765 bp
N50	99 bp

Table S7: Details of pollen allergens from various species. This was the starting point for developing hmm profiles to identify homologs in the perennial ryegrass genome.

Protein Family No.	Seq. Nr.	Accession Nr.	aa	Description
1				Group 1 pollen allergens Expansin B-like Pollen allergen Lol p1 signatures Rare lipoprotein A (RlpA)-like double-psi beta-barrel Expansin, family-45 endoglucanase-like domain profile
	1	P14946.2	263	<i>Pollen allergen Lol p 1 [Lolium perenne]</i>
	2	XP_004986062.1	268	<i>PREDICTED: expansin-B11-like [Setaria italica]</i>
	3	XP_003577577.1	266	<i>PREDICTED: pollen allergen Phl p 1-like [Brachypodium distachyon]</i>
	4	AAP96760.1	264	<i>Group 1 allergen Dac g 1.02 precursor [Dactylis glomerata]</i>
	5	BAK06415.1	262	<i>Predicted protein [Hordeum vulgare subsp. vulgare]</i>
	6	AEW90966.1	270	<i>Expansin B4 F5-2 [Secale cereale x Triticum turgidum subsp. durum]</i>
2				Group 2 pollen allergens Expansin B-like Pollen allergen Lol p2 signatures Expansin, Cellulose-binding-like domain profile.
	7	P14947.1	97	<i>Pollen allergen Lol p 2-A [Lolium perenne]</i>
	8	2103117A	196	<i>Allergen Dac g II [Dactylis glomerata]</i>
	9	AEW67313.1	120	<i>Group 2 grass pollen allergen [Secale cereale x Triticum turgidum subsp. durum]</i>
	10	XP_003566870.1	124	<i>PREDICTED: pollen allergen Lol p 2-A-like [Brachypodium distachyon]</i>
	11	XP_003577434.1	118	<i>PREDICTED: pollen allergen Lol p 2-A-like [Brachypodium distachyon]</i>
	12	P14948.1	97	<i>Pollen allergen Lol p 3 [Lolium perenne]</i>
	13	P93124.1	96	<i>Pollen allergen Dac g 3 [Dactylis glomerata]</i>

Protein Family No.	Seq. Nr.	Accession Nr.	aa	Description
3	14	EMS63733.1	119	<i>Pollen allergen Lol p 3 [Triticum urartu].</i>
	15	ABX54879.1	152	<i>Copper/zinc superoxide dismutase (SODC) family</i> <i>Ole e 5 olive pollen allergen [Olea europaea]</i>
	16	Q8L5E0.2	152	<i>Superoxide dismutase [Cu-Zn] Ole e 5</i>
4				<i>Glycosyl hydrolases family 17</i> <i>X8 domain</i>
	17	Q94G86.1	460	<i>Major pollen allergen Ole e 9 [Olea europaea]</i>
	18	Q84V39.1	123	<i>Major pollen allergen Ole e 10 [Olea europaea]</i>
5				<i>Pectate lyase family</i> <i>Pollen allergen Amb family signatures</i>
	19	P81294.1	367	<i>Pectate lyase 1, Major pollen allergen Jun a 1 [Juniperus ashei]</i>
	20	P27760.1	398	<i>Pectate lyase 1, Antigen Amb a I [Ambrosia artemisiifolia]</i>
	21	P27761.1	397	<i>Pectate lyase 2, Pollen allergen Amb a 1.3 [Ambrosia artemisiifolia]</i>
	22	P28744.1	392	<i>Pectate lyase 3, Pollen allergen Amb a 1.4 [Ambrosia artemisiifolia]</i>
	23	P27762.1	397	<i>Pectate lyase 4, Pollen allergen Amb a 2 [Ambrosia artemisiifolia]</i>
	24	P27759.1	396	<i>Pectate lyase 5, Pollen allergen Amb a 1.1; [Ambrosia artemisiifolia]</i>
6				<i>Group 5 pollen allergens</i> <i>Ribonuclease (pollen allergen) domains</i> <i>Pollen allergen Poa pl signatures</i>
	25	Q40240	307	<i>Major pollen allergen Lol p 5a [Lolium perenne]</i>
	26	EMS45090.1	319	<i>Major pollen allergen Lol p 5a [Triticum urartu]</i>
	27	Q40237	339	<i>Major pollen allergen Lol p 5b [Lolium perenne]</i>
	28	CCI69081.1	301	<i>pollen allergen Ave s 5 (isoallergen B) [Avena sativa]</i>
	29	P43215.1	132	<i>Pollen allergen Phl p 6 [Phleum pratense]</i>
	30	CAA76556.1	138	<i>Phl p6 allergen [Phleum pratense]</i>

Protein Family No.	Seq. Nr.	Accession Nr.	aa	Description
7	31	AAA33405.1	339	<i>Isoform 9 [Lolium perenne]</i>
	32	P22285	333	<i>Pollen allergen KBG 41, Pollen allergen Poa p 9 [Poa pratensis]</i>
				Group 13 pollen allergens Glycosyl hydrolases family 28 Polygalacturonase active site Parallel beta-helix repeats
	33	AIL01320.1	422	<i>Group 13 allergen Sor h 13.01 [Sorghum halepense]</i>
	34	ABD79097.1	404	<i>Zea m 13 allergen, partial [Zea mays]</i>
	35	BAF32130.1	514	<i>Pollen allergen [Cryptomeria japonica]</i>
8	36	Q9FY19.1	507	<i>Major pollen allergen Jun a 2 [Juniperus ashei]</i>
				Group 4 pollen allergens PCMH-type FAD-binding domain profile Berberine and berberine like gene family
	37	AAS02108.1	522	<i>FAD-linked oxidoreductase BG60 [Cynodon dactylon]</i>
	38	ABB78007.1	525	<i>Major pollen allergen Phl p 4 precursor [Phleum pratense]</i>
	39	AEW67310.1	520	<i>Group 4 grass pollen allergen [Secale cereale x Triticum turgidum subsp. durum]</i>
	40	BAF19743.1	531	<i>Os06g0548800 [Oryza sativa Japonica Group]</i>
	41	BAK06312.1	521	<i>Predicted protein [Hordeum vulgare subsp. vulgare]</i>
9	42	EMS61759.1	520	<i>Reticuline oxidase-like protein [Triticum urartu]</i>
				Ole e6 pollen allergens
	43	XP_003632317.1	73	<i>Pollen allergen Ole e 6-like [Vitis vinifera]</i>
10	44	O24172.1	50	<i>Major pollen allergen Ole e 6 [Olea europaea]</i>
				Phytocyanin domain profile family
11	45	P00304.2	101	<i>Pollen allergen Amb a 3 [Ambrosia artemisiifolia var. elatior]</i>
				Cysteine-rich secretory protein family CRISP family signature 1

Protein Family No.	Seq. Nr.	Accession Nr.	aa	Description
12				CRISP family signature 2 Venom allergen 5 signature SCP / Tpx-1 / Ag5 / PR-1 / Sc7 family of extracellular domains.
	46	CAK50834.1	162	<i>Art v 2 allergen [Artemisia vulgaris]</i>
				Gamma-thionin family Knottins Defensin-like proteins
	47	AGT51226.1	78	<i>Defensin [Brassica napus]</i>
	48	CBJ24285.1	147	<i>Ragweed homologue of Art v 1 precursor [Ambrosia artemisiifolia]</i>
13	49	XP_002281189.1	82	<i>PREDICTED: defensin-like protein 19 [Vitis vinifera]</i>
	50	Q84ZX5.1	132	<i>Major pollen allergen Art v, Defensin-like protein 1 [Artemisia vulgaris]</i>
				Pathogenesis-related proteins, Bet v I family signatures
	51	P38948.2	160	<i>Major pollen allergen Aln g 1 [Alnus glutinosa]</i>
	52	CAB02217.1	161	<i>Pollen allergen Car b 1 [Carpinus betulus]</i>
14	53	AFF59689.1	160	<i>Bet v 1 allergen [Quercus suber]</i>
	54	NP_001105279.1	154	<i>Bet v I allergen [Zea mays]</i>
	55	NP_001147371.1	154	<i>Major pollen allergen Car b 1 isoforms 1A and 1B [Zea mays]</i>
				Group 11 pollen allergens, Lol p11-like, Ole e1 family signatures
	56	CAA73036.1	146	<i>Ole e 1 protein [Olea europaea]</i>
	57	CAA73037.1	146	<i>Ole e 1.0103 protein [Olea europaea]</i>
	58	AEW90944.1	167	<i>Group 11 grass pollen allergen F8-9 [Secale cereale x Triticum turgidum subsp. durum]</i>
	59	BAK06698.1	167	<i>Predicted protein [Hordeum vulgare subsp. vulgare]</i>
	60	Q7M1X5.1	134	<i>Major pollen allergen Lol p 11 [Lolium perenne]</i>
	61	XP_004960638.1	172	<i>PREDICTED: pollen allergen Phl p 11-like [Setaria italica]</i>

Protein Family No.	Seq. Nr.	Accession Nr.	aa	Description
15				Pathogenesis-related proteins, Thaumatin family signatures
	62	P81295.1	225	<i>Pathogenesis-related protein, Pollen allergen Jun a 3 [Juniperus ashei]</i>
	63	BAF51970.1	225	<i>Thaumatin-like protein [Cryptomeria japonica]</i>
16				Group 7 pollen allergens Polcalcins EF-hand calcium-binding domain profile
	64	P94092.2	80	<i>Polcalcin, Calcium-binding pollen allergen Cyn d 7 [Cynodon dactylon]</i>
	65	O82040.1	78	<i>Polcalcin, Calcium-binding pollen allergen Phl p 7 [Phleum pratense]</i>
	66	O81701.1	85	<i>Polcalcin, Calcium-binding pollen allergen Aln g 4 [Alnus glutinosa]</i>
	67	O81092.1	84	<i>Polcalcin, Calcium-binding pollen allergen Ole e 3 [Olea europaea]</i>
17				Group 12 pollen allergens Profilins
	68	ABP58625.1	134	<i>Pollen allergen Ole e 2 [Olea europaea]</i>
	69	ABP58624.1	134	<i>Pollen allergen Ole e 2 [Olea europaea]</i>
	70	P35081.1	131	<i>Profilin-1, Pollen allergen Zea m 12 [Zea mays]</i>
	71	A4KA31.1	131	<i>Profilin-4, Pollen allergen Phl p 12 [Phleum pratense]</i>
	72	A4KA41.1	131	<i>Profilin-5, Pollen allergen Cor a 2 [Corylus avellana]</i>
	73	Q64LH1.1	131	<i>Profilin-1, Pollen allergen Amb a 8 [Ambrosia artemisiifolia]</i>
	74	Q8H2C9.3	133	<i>Profilin-1, Pollen allergen Art v 4.01; [Artemisia vulgaris]</i>

Table S8: Average Transcripts Per Million (TPM) for genes with > 1000 TPM in mature pollen grains. Average TPM in leaf samples is also shown.

Nr.	Code	TPM	
		Leaves	Pollen
1	AMBR_151	0.47	30765.05
2	OLE1_056	4.69	28160.01
3	PROF_917	0.65	13949.76
4	LOL1_050	0	13634.86
5	LOL1_117	0.08	9762.51
6	OLE1_121	0	7221.52
7	LOL1_116	0.09	6516.64
8	POAP_111	0	6119.24
9	EFHA_187	3.36	5135.54
10	EFHA_184	0.27	5083.43
11	EFHA_165	0.46	4341.51
12	ART2_329	0	4065.21
13	ART2_327	0.1	3573.75
14	POAP_038	0	2618.21
15	AMB3_005	0	2449.94
16	AMB3_209	0	2341.84
17	EFHA_494	0.17	2319.35
18	PARA_314	0	2225.05
19	AMB3_008	0	2202.19
20	EFHA_236	2.58	1975.55
21	LOL2_102	0	1349.16
22	PARA_312	0.83	1295.35
23	PROF_662	5.15	1251.78
24	OLE9_121	0	1216.64
25	PARA_753	0.03	1204.85
26	LOL1_123	0	1198.52

Table S9: How each library was used in the ALLPATHS-LG assembly, and the percentage of duplicates for each library type.

Library Type	Fragment size (bp)	Library type for ALLPATHS-LG	% Duplicates
PE	140	Fragment	NA
PE	550	Jumping	19.0
MP	1,800	Jumping	
MP	1,800	Jumping	21.2
MP	3,400	Jumping	35.3
MP	8,600	Jumping	76.7
LJD ³	7,765	Jumping	64.8
LJD ³	26,265	Long-jumping	76.8
LJD ³	29,765	Long-jumping	74.8

³Long Jump Distance libraries (Eurofins MWG Operon)

Table S10: The number of Illumina reads used for the nine *de-novo* transcriptome assemblies generated, and the source of the material.

Assembly	Genotype	Plant Structures	Stress Condition	Illumina Reads ¹
1	P226/135/16 (Inbred)	Inflorescence, leaf sheath, mature leaf, meristem, stem, root	NA	36,727,657 ²
2	F130	Inflorescence, leaf sheath, mature leaf, meristem, stem,	NA	31,263,328 ²
3	Mix	Pollen	NA	169,358,368 ³
4	F130	Stigma	NA	240,285,844 ³
5	F139	Mature leaf	NA	11,683,342 ³
6	Falster	Mature leaf	Drought stress	275,821,778 ³
7	Falster	Mature leaf	Cold stress	303,881,606 ³
8	Veyo/09	Mature leaf	Drought stress	306,070,872 ³
9	Veyo/09	Mature leaf	Cold stress	303,635,322 ³

¹Error-corrected reads

²the fragment length was shorter than twice the read length so the pairs were merged using the ALLPATHS-LG error correction module. For P226/135/16 the mean length was 131 bp (std. 17.37bp), and for F130 the mean length was 119 bp (std. 14.66 bp).

³paired reads

Table S11: Summary statistics of PacBio data, before and after error correction, and after sampling a one times genome equivalent.

	Raw reads	After error-correction	Random sample ¹
Number of reads	14,843,437	4,052,697	600,000
Sum (Gb)	18.35	10.04	2.11
Average (bp)	1,236	2,478	3,526
N50 (bp)	3,461	2,942	3,684

¹Random sample of reads above 2 Kb to achieve a one times coverage of the genome.

Figure S1: kmer spectra at K=25 for filtered fragment reads (red) and for fragment reads after error correction (blue). The single peak is indicative of a high degree of homozygosity.

Figure S2: Synteny between ryegrass and barley. Synteny between the ryegrass and barley genomes for each of the seven corresponding chromosomes as revealed on basis of ryegrass genetic marker information and the POPSEQ-based ordering of barley whole genome sequence assemblies. The orange highlighted regions indicate the translocated and inverted genome segment (segment was translocation from the long arm of chromosome L5 to the short arm of chromosome L4).

Figure S3: Coverage of the *Brachypodium* genome by assembled *Lolium* WGS scaffolds and syntenic. The coverage of the *Brachypodium* genome was determined by comparing the assembled ryegrass WGS scaffolds against the *Brachypodium* gene set. The heat maps measure the proportion of detected orthologous *Brachypodium* genes along the five chromosomes (Bd 1 to Bd 5) by using a sliding window approach (window size 0.5Mb and window shift 0.1Mb). The colored boxes above each chromosome indicate the syntenic relationship to *Lolium*. Assembled *Lolium* genome scaffolds matching genes within these regions were assigned to individual chromosomes according to the synteny relationships. The syntenic regions corresponding to the L4/L5 translocation is highlighted by the orange rectangle.

Figure S4: Estimated repeat branch rate as a function of K , for perennial ryegrass (purple) in the context of the repeat branch rate for yeast (*Saccharomyces cerevisiae*, pink), snake (*Boa constrictor constrictor*, red), oyster (*Crassostrea gigas*, yellow), human (dark green),

fish (*Maylandia zebra*, light green), bird (*Melopsittacus undulates*, blue). Perennial ryegrass has the highest frequency of variant branches in the assembly graph.

Figure S5: Comparison of density plots of $\log_{10}(\text{average TPM})$ for leaf (red) and pollen (blue) samples.

Figure S6: MA and Volcano plots showing the differential expression of genes during a compatible pollination reaction in comparison to un-pollinated stigma.

Figure S7: MA and Volcano plots showing the differential expression of genes during an incompatible pollination reaction in comparison to un-pollinated stigma.

Figure S8: Histogram of the number of PacBio reads addressing closed gaps in the *Lolium perenne* genome assembly.

Table S1: Summary of Illumina sequencing data for P226/135/16.

Table S2: Summary of gap, contig, and scaffold statistics before and after gap filling with PacBio reads. Gap filling was performed with PBJelly.

Table S3: Summary statistics of the ALLPATHS-LG assembly.

Table S4: Statistics on the completeness of the genome based on 248 ultra-conserved Core Eukaryotic Genes (CEGs). Results are shown for complete coverage and partial coverage of CEGs. Group 1 CEGs are the least conserved, and group 4 CEGs are the most conserved across species.

Table S5: GenomeZipper statistics for individual linkage perennial ryegrass linkage groups.

Table S6: Summary statistics of the Abyss unitig assembly of the perennial ryegrass paired-end fragment reads.

Table S7: Details of pollen allergens from various species. This was the starting point for developing hmm profiles to identify homologs in the perennial ryegrass genome.

Table S8: Average Transcripts Per Million (TPM) for genes with > 1000 TPM in mature pollen grains. Average TPM in leaf samples is also shown.

Table S9: How each library was used in the ALLPATHS-LG assembly, and the percentage of duplicates for each library type.

Table S10: The number of Illumina reads used for the nine *de-novo* transcriptome assemblies generated, and the source of the material.

Table S11: Summary statistics of PacBio data, before and after error correction, and after sampling a one times genome equivalent.

Supporting Information Data S1-S7: spreadsheets with the perennial ryegrass GenomeZipper data, one file for each of the seven linkage groups.

Supporting Information Data S8: spreadsheet showing the expression levels (as TPM) of the 630 genes with sequence homology to known pollen allergens, in both mature pollen and leaf samples.

Supporting Information Data S9: spreadsheet showing genes significantly down regulated in both a compatible and incompatible pollination reaction.

Supporting Information Data S10: spreadsheet showing genes significantly down regulated in a compatible pollination reaction.

Supporting Information Data S11: spreadsheet showing genes significantly down regulated in an incompatible pollination reaction.

Supporting Information Data S12: spreadsheet showing genes significantly up regulated in both a compatible and incompatible pollination reaction.

Supporting Information Data S13: spreadsheet showing genes significantly up regulated in a compatible pollination reaction.

Supporting Information Data S14: spreadsheet showing genes significantly up regulated in an incompatible pollination reaction.

