An ultra-high density genetic linkage map of perennial ryegrass (Lolium perenne) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly

Velmurugan, Janaki; Mollison, Ewan ; Barth, Susanne; Marshall, David; Milne, Linda; Creevey, Christopher; Lynch, Bridget; Meally, Helena; McCabe, Matthew; Milbourne, Dan

Published in:
Annals of Botany

DOI:
10.1093/aob/mcw081

Publication date:
2016

Citation for published version (APA):
Velmurugan, J., Mollison, E., Barth, S., Marshall, D., Milne, L., Creevey, C., Lynch, B., Meally, H., McCabe, M., & Milbourne, D. (2016). An ultra-high density genetic linkage map of perennial ryegrass (Lolium perenne) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly. *Annals of Botany*, *118*(1), 71-87. https://doi.org/10.1093/aob/mcw081

ORIGINAL ARTICLE

An ultra-high density genetic linkage map of perennial ryegrass (*Lolium perenne*) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly

Janaki Velmurugan[1, 2*,] Ewan Mollison [1, 3, 6*] Susanne Barth[1], David Marshall[3], Linda Milne[3], Christopher J Creevey[4,5], Bridget Lynch[2], Helena Meally[1], Matthew McCabe[4], Dan Milbourne[1#]

1. Teagasc, Crops, Environment and Land Use Programme, Oak Park Research Centre, Carlow, Ireland
2. University College Dublin, School of Agriculture and Food Science, Dublin, Ireland
3. Information and Computational Sciences Group, James Hutton Institute, Errol Road, Invergowrie, Dundee, UK
4. Teagasc, Animal and Grassland Research and Innovation Centre, Grange, Ireland
5. Current address: Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, UK
6. Division of Plant Sciences, University of Dundee at the James Hutton Institute, Errol Road, Invergowrie, Dundee, UK

*These authors contributed equally to the manuscript

Running Title: An anchored GBS-based map of perennial ryegrass

Corresponding author email: dan.milbourne@teagasc.ie

Abstract

Background and Aims: High density genetic linkage maps that are extensively anchored to assembled genome sequences of the organism in question are extremely useful in gene discoveryTo facilitate this process in perennial ryegrass (*Lolium perenne* L.)we have developed a high density SNP- and presence/absence variant (PAV)-based genetic linkage map in an F2 mapping population that has been used as a reference population in numerous studies. To provide a reference sequence to align GBS reads to, we created a shotgun assembly of one of the grandparents of the population, a tenth generation inbred line, using Illumina-based sequencing.

Methods: The assembly was based on paired end Illumina reads, scaffolded by mate pair and long-jumping distance reads in the range of 3-40kb, with over 200-fold initial genome coverage. One hundred and sixty nine individuals from an F2 mapping population were used to construct *PstI*-based GBS libraries tagged with unique 4-9 nucleotide barcodes, resulting in 284 million reads, with ~1.6 million reads per individual. A bioinformatics pipeline was employed to identify both SNPs and PAVs. A core genetic map was generated using high confidence SNPs, to which lower confidence SNPs and PAVs were subsequently fitted in a straightforward binning approach.

Key Results: The assembly comprises 424,750 scaffolds, covering 1.11 Gbp of the 2.5 Gbp perennial ryegrass genome, with a scaffold N50 of 25,212 basepairs (bp) and a contig N50 of 3,790 bpIt is available for download and access to a genome browser has been provided. Comparison of the assembly to available transcript and gene-model datasets for perennial ryegrass indicates that approximately 570 Mbp of the gene-rich portion of the genome has been captured. An ultra high-density genetic linkage map with 3092 SNPs and 7260 PAVs was developed, anchoring just over 200 Mb of the reference assembly.

Conclusions: The combined genetic map and assembly, combined with another recently released genome assembly, represent a significant resource for the perennial ryegrass genetics community.

Key words : *Lolium perenne*, perennial ryegrass, genome assembly, genotyping by sequencing, GBS, single nucleotide polymorphism, linkage mapping, presence/absence variation

Introduction

Perennial ryegrass (*Lolium perenne* L.) is an important component species in pastoral based production systems in temperate regions. It is diploid (2n) with 7 chromosomes and has a genome size of ~2.5 Gb (Kopecky et al., 2010). Despite its relative importance it remains poorer in genome-based resources than other grass species, such as members of the closely related Triticeae, lacking at the time of writing a published physical map and genome sequence.

High quality genetic linkage maps remain a cornerstone of discovery genetics in plant species. Despite their numerous drawbacks, including a restricted representation of the true genetic diversity of a species, much progress in discovery genetics continues to be made using "flagship" reference mapping populations over near decadal timescales. The greatest advances can be gained in such reference populations by developing genetic maps that are densely populated with genetic markers located in genic regions and that are sequence-characterised in such a way as to allow anchoring to pre-existing or emerging physical maps, and other important reference maps in the same or other species.

The "F2 Biomass" population has been used to study segregation distortion (Anhalt et al., 2008) and as a basis for several QTL mapping studies for traits including rust resistance (Tomaszewski et al., 2012), biomass yield (Anhalt et al., 2009) polar (A.Foito, JHI, Dundee, UK, unpubl.res) and non-polar metabolites (Foito et al., 2015), Both parents of the F1 parental genotype of this population were originally maintainer lines in a cytoplasmic male sterility (CMS) programme at Teagasc (Connolly and Wrightturner, 1984) and originated from an inter-specific cross between meadow fescue (*Festuca pratensis*) and perennial ryegrass. The initial interspecific hybrid was backcrossed for several generations to the ryegrass parent and recurrently self pollinated for nine (maternal grandparent) or ten (paternal grandparent) generations. The background of the *Lolium* contribution in the pedigree of the

inbred lines was the ryegrass cultivar ‘S24’ (The Institute of Biological, Environmental and Rural Sciences (IBERS)) for the maternal grandparent and the ryegrass cultivar ‘Premo’ (Mommersteeg International BV) for the paternal grandparent. The inbred lines have been subjected to analysis using both fluorescent and genomic in-situ hybridisation (Anhalt et al., 2008) approaches, and no evidence of large intact portions of the fescue parent were evident, indicating that the grandparents largely reflect a perennial ryegrass genetic background. Offspring arising from self pollinationof a single F1 plant from a cross between these two self-compatible lines was used for the basis of the original F2 biomass population of 360 individuals (Tomaszewski et al., 2012, Anhalt et al., 2008). The population is also the basis for an ongoing initiative to develop a recombinant inbred line (RIL) population for perennial ryegrass at Teagasc.

Advances in sequencing technology have allowed the development of approaches to generate extremely large numbers of DNA markers in a quick and cost effective manner (Davey et al., 2011). Although sequencing costs have experienced a continued downward trend over the last several years, it is still relatively expensive and computationally intensive to sequence and assemble whole genomes in order to identify genetic variation/DNA markers for the large numbers of genotypes that tend to comprise experimental populations. As an alternative, numerous strategies have been developed that rely on sequencing reduced subsets of the genomes of different individuals to identify such variation. High throughput polymorphism detection methods like (CRoPS) (van Orsouw et al., 2007), restriction site associated DNA sequencing (RADseq) (Davey and Blaxter, 2011) and genotyping by sequencing (GBS) (Elshire et al., 2011) use genome complexity reduction approaches to target specific regions of the genome and markers are identified by examining DNA variation in similar subsets of genome from different individuals using widely available bioinformatics-based approaches. These methods, combined with the power of next-generation sequencing technology, have

104 radically enhanced our ability to generate thousands of markers in reasonably large
105 experimental populations, opening up a wealth of applications in areas such as discovery
106 genetics and genomics assisted plant breeding.

107 The specific GBS approach described by (Elshire et al., 2011) is increasingly becoming a
108 method of choice for high throughput genotyping applications. This method has a simple
109 protocol for the generation of genotyping libraries, which lacks a specific gel-based size
110 selection step, avoids the use of divergent Y-adapters, and is amenable to parallelisation using
111 either manual or automated liquid handling approaches. Combining these features with a
112 simple in-line barcoding system and the ability to tailor the protocol to suit different
113 organisms and applications by changing the methylation-sensitive restriction enzyme/s
114 employed for the complexity-reduction step makes GBS a powerful but easy to adopt
115 approach for genome-wide marker generation. Because of this, it has been widely adopted in
116 plant species.

117 Although it was conceived primarily as a method for detecting single nucleotide
118 polymorphism (SNP) variation, GBS can also survey other forms of variation including small
119 InDels, SSRs and "presence/absence" variation derived from "anonymous" DNA
120 polymorphisms that cause variation in whether specific DNA fragments are amplified across
121 individuals (Elshire et al., 2011). The GBS approach has been shown to work remarkably well
122 irrespective of the availability of a reference genome (Ward et al., 2013), since fragments
123 produced in individuals of the experimental population can be auto-assembled to produce a
124 reference sequence set to which the same fragments can subsequently be re-aligned in order to
125 identify polymorphisms. However, the approach is also extremely useful when it is coupled
126 with an externally-derived reference sequence, and it is a useful tool for applications such as
127 anchoring shotgun genome assemblies and spotting misassembles in existing reference
128 genome sequences (Mascher et al., 2013).

As previously mentioned, the F2 Biomass population has been used extensively for a variety of purposes, and the densest map available to date for the cross is a Diversity Array Technology (DArT) based map comprising 297 markers, anchored and oriented with 29 SSR markers (Tomaszewski et al., 2012). Despite the fact that the sequences of the DArT markers are available (Bartos et al., 2011), their lack of genomic context and the relatively low density of the map are limitations for the continued use of the population as a platform for genetic analysis. The objective of the current study was to generate a high density genetic linkage map of the F2 Biomass population using GBS in order to increase its utility for genetic mapping studies in the future. Despite the fact that such a map could be developed in the absence of a reference genomic sequence, we decided to increase the utility of the mapping resource by assembling a reference shotgun sequence of the inbred line that was the paternal grandparent of the mapping population. Short read fragments (from the Illumina HiSeq2000 platform) mapped using this approach would thus generally be anchored to larger fragments from the shotgun assembly, providing a better genomic context for each marker mapped – increasing their utility for future applications. This is especially timely in the context of the recent publication of an annotated synteny-based draft genome sequence of another genotype of *Lolium perenne* (Byrne et al. (2015). We present the genetic map, all SNP information and the reference assembly as resources for the forage genetics community.

Materials and Methods

Shotgun sequencing and assembly of reference sequence

Illumina HiSeq and GAII sequencing was used to generate approximately 207-fold raw coverage of the genome of the paternal grandparent of the F2 Biomass population. Libraries were produced from a range of paired-end (<300 bp insert), mate-pair (3 kbp insert) and long-jumping distance (8 kbp, 20 kbp and 40 kbp insert) libraries and with read lengths ranging

153 from 51 – 160 bp. Supplementary File 1 shows details for all libraries used. **[Supplementary**
154 **Information]**

155 For the 300 bp and 3 kbp insert libraries, library production was as follows: DNA of the
156 paternal grandparent (2 μg) was fragmented for 30 minutes with NEBNext® dsDNA
157 Fragmentase (NEB) and purified using a QIAquick PCR purification kit (Qiagen). The
158 NEBNext® End Repair Module was used to blunt end the fragments and purification of the
159 reaction was performed using a QIAquick PCR purification kit (Qiagen). The NEBNext® dA-
160 Tailing Module was used to adenylate the blunt ended fragments and purification of the
161 reaction was performed using a QIAquick PCR purification kit (Qiagen). Illumina standard
162 paired end adapters were ligated onto the adenylated fragments using the Quick Ligation™
163 Kit (NEB) and purification was performed using a QIAquick PCR purification kit (Qiagen).
164 Adapter ligated fragments were then size selected by electrophoresis on an agarose gel,
165 excision of a 2 mm gel slice and extraction of DNA from the agrose using the QIAquick gel
166 extraction kit (Qiagen). PCR enrichment (12 cycles) of the library was performed using
167 Illumina PCR Paired End Primers 1.0 and 2.0 and Phusion™ High-Fidelity PCR Kit
168 (Finnzymes). The library size and absence of adapter dimers were determined with a
169 DNA1000 chip on an Agilent 2100 bioanalyser. Sequencing was performed on either an
170 Illumina HiSeq2000 or GAII platform as outlined in Supplementary File 1. Long jumping
171 distance libraries were constructed by Eurofins Genomics using a proprietary method, and
172 sequenced on an Illumina HiSeq2000 platform.

173 Assembly was carried out using the resulting FASTQ files. Prior to assembly, additional
174 quality control stages were carried out: Sickle (Joshi and Fass, 2011) was used to trim low
175 quality base calls from 5' ends of the reads using a quality cut-off of Q30, equivalent to
176 99.9% confidence in base-calls, and with remaining read length of 50 bp (35 bp in the case of
177 the 3 kbp libraries, as these were sequenced with read length 51 bp); and FastUniq (Xu et al.,

2012) was used to remove redundant read pairs that may have arisen due to PCR duplication. As a result of this filtering, final genome coverage was reduced to approximately 105-fold. Following trimming and de-duplication, paired-end and singleton reads were assembled using CLC Assembly Cell (http://www.clcbio.com/, CLC Bio. Aarhus, Denmark) with a k-mer length of 41 and then scaffolded with the 3kb – 40kb read pairs using SSPACE (Boetzer et al., 2011).

Preliminary annotation of the reference sequence

RepeatMasker version 3.2.8 (Smit et al., 2010) was used to identify common repetitive elements in the scaffolded *Lolium* assembly using a wheat-based model. The widely-used, open-source, gene prediction tool Augustus (Stanke and Waack, 2003) was used for gene prediction using the repeat masked genome assembly, with a wheat-based gene model. BLAST searching (Altschul, 1990) of barley cDNA sequences and publicly available peptide sequences for barley, rice and brachypodium was carried out using default cut-off parameters (E-value = 10), which allows some very dissimilar matches to be returned. This parameter was intentionally left at the default setting to allow identification of distant homologues.

Viewing *Lolium* genomic data

In order to view data generated in this study in a more accessible format, a JBrowse-based genome browser (Skinner et al., 2009) has been set up and made available at https://ics.hutton.ac.uk/jbrowse/lolium and the scaffolded genome assembly is available for download at https://ics.hutton.ac.uk/jbrowse/lolium/data/seq/lolium_scaffolds.zip

Raw reads generated in this study have been deposited with the European Nucleotide Archive under study accession PRJEB12921 (http://www.ebi.ac.uk/ena/data/view/PRJEB12921).

Estimating coverage of the *Lolium* gene complement

Four approaches were taken to estimate the degree to which *Lolium's* gene complement was captured within the assembly.

Byrne et al. (2015) have published models for 28,455 *Lolium* genes, yielding 40,068 transcripts, based on several RNA-Seq studies. Sequences for these genes were compared using BLAST against our *Lolium* assembly (E-value = 10). Completeness of BLAST hits was assessed based on cumulative identity percentage (CIP) and cumulative alignment percentage (CALP) across all high-scoring segment pairs (HSPs) for each match, a method described in (Salse et al., 2008).

Ruttink et al. (2013) used an Orthology Guided Assembly (OGA) approach to creating a reference transcriptome for *Lolium.* For simplicity, we used only the OGA based on *Brachypodium distachyion* in this study. Sequences of over 200 bases from the OGA transcriptome were compared using BLAST against the *Lolium* assembly (E-value = 10) and match strength was evaluated as above.. Another *Lolium* transcriptome assembly is described by (Farrell et al., 2014); all sequences in this transcriptome assembly are over 200 bp in length, so no filtering was required before applying the approach described above.

CEGMA, the Core Eukaryotic Genes Mapping Approach, (Parra et al., 2007) was used to search the *Lolium* assembly for 458 core proteins that are conserved across eukaryotes, with a more highly conserved subset of 248 used to indicate completeness of coverage.

A set of 47 genes associated with control of flowering in rice and *Brachypodium* were selected from Higgins et al., (2010) and an additional gene associated with flowering in barley, CEN, from Comadran et al., (2012). Peptide sequences for those genes were searched against the *Lolium* genome assembly from this study using Exonerate (Slater and Birney, 2005), with the top-ranked hit being considered as the probable *Lolium* orthologue. The same approach was applied to identify the equivalent scaffold from the assembly of Byrne et al. (2015), with the additional step of confirming the equivalency of the scaffold through manual

inspection of BLAST based comparisons of the scaffolds between the assemblies in order to ensure that they represented the orthologous regions in both assemblies.

GBS library construction

One hundred and sixty-nine individuals from the F2 Biomass population were used for the mapping study. For reference purposes, the paternal grandparent and the F1 parental genotype were also used. Unfortunately, the maternal grandparent was no longer extant at the time of the study. Genomic DNA from these individuals was extracted from approximately 3g of flash frozen, fresh leaf material using a variation of the CTAB (cetyltrimethylammoniumbromide) method of (Doyle, 1987). GBS libraries were constructed using an adapted version of the protocol outlined by Elshire et al. (2011), employing the methylation sensitive 6 bp rare cutting restriction enzyme *PstI* instead of Ape*KI*. A set of 48 unique barcode adaptors were generated from complementary sequence with a *PstI* overhang sequence. The barcodes varied from 4-9 nt in length. A common adaptor and PCR primers A and B were generated. Complementary oligos for each of the 48 adaptors at 50 uM were annealed under the following programme: 95°C, 2 minutes: ramp to 25°C by 0.1°C/s: 25°C, 30 minutes: 4°C hold. The annealed adaptors were diluted 1:15 and then a further dilution of 1:100. 100 ul of the 1:1500 diluted barcoded adaptors and the common adaptor were mixed to make the 200ul working stock of 0.6 ng/ul. These were quantified using the Qubit fluorometer.

DNA was digested in 20 ul reactions containing 200-220 ng of genomic DNA, 2 ul 10 X NEB buffer 3, 1.5 ul of BSA, 20 units of *PstI* and 13.5 ul of molecular grade water incubated at 37°C for 2 hours, then deactivated at 80°C for 20 minutes. In the ligation reaction 20 ul of digested product was combined with 12 ng of the working stock of annealed adaptor mix, 5 ul 10X T4 ligase buffer and 400 units T4 ligase in a 50 ul reaction. All ligation reactions were incubated at 22°C for 1 hour and then 65°C for 30 minutes to deactivate the ligase.

The ligation reaction was cleaned up using the Qiagen Qiaquick PCR purification kit and the elution volume was 50 ul. PCR was set up as a 50 ul reaction that included 10 ul of the purified ligation reaction, 25 ul of the NEB 2X Taq master mix, 2 ul of a 3 uM primer 1 and 2 mix and 13 ul of molecular grade water. PCR programme was 72°C for 5 minutes: 98°C for 30 seconds: 18 cycles of 98°C for 10 seconds, 65°C for 30 seconds, 72°C for 30 seconds: 72°C for 5 minutes and 4°C hold. The PCR enriched libraries were purified using the Qiagen MiniElute purification kit and were eluted in 21ul. The quality of the library was checked on the 2100 Bioanalyser from Agilent technologies. The constructed GBS libraries were sequenced in two channels of Illumina HiSeq 2000 (Bentley et al., 2008) for single end 100 bp reads.

Variant Calling Pipeline

The sequenced reads were de-multiplexed and trimmed to 66 bp using process_radtags component of Stacks (Catchen et al., 2013). A sliding window quality metric (-w 0.15) was adopted to discard any reads with low quality scores (i.e. as the sliding window scans the read, if any 15% of a total fraction of the length of the read falls below the phred score value of 10 (-s) the reads were discarded). Also reads with uncalled bases were discarded. The de-multiplexed reads were aligned to the reference set using the Bowtie (Langmead et al., 2009) alignment program, allowing 2 mismatches (-v 2) and allowing only unique mapping to the reference set (-m 1). The resulting alignment files in SAM format were post processed (converting SAM to BAM format, sorting, indexing the BAM files) to create a consensus mpileup file using Samtools-0.1.18 (Li et al., 2009).

For SNP discovery, VarScan.v2.2.11 (Koboldt et al., 2012) was used to call SNP variants from the mpileup file with the settings: (minimum coverage:8; minimum reads:2; minimum variant allele frequency:0.2; minimum average quality:20; p-value threshold:0.05). The

277 resulting variant list file was filtered for SNP markers exhibiting at least one heterozygote to
278 identify the maximum number of possibly segregating markers.

279 To identify presence/absence variants (PAVs), all the individual alignment files in BAM
280 format were merged using the samtools merge command. The merged bam file was then
281 converted to SAM format. The merged SAM file was then parsed into a text file to produce a
282 table with genotypes from the population in columns and independent loci having at least one
283 alignment as rows. A simple UNIX shell script was subsequently used to identify alignments
284 from this table for which between 10% and 50% of the individuals (16) exhibited a Not
285 Called (NC) designation (indicating potential PA variation), and for which the average read
286 depth for individuals exhibiting alignments was eight.

287 **Linkage map construction**

288 Linkage map calculations were performed using R/qtl (Broman et al., 2003) and Joinmap
289 (v.4.1,Kyazma,(Van Ooijen, 2011)). The SNP markers and PAV markers were initially scored
290 respectively as co-dominant marker type [A, H, B] and dominant marker for an F2 population
291 type in R/qtl (Broman et al., 2003). For simplicity, PAVs were all coded in the same way [B
292 (absent), D (present)] regardless of probable grandparental line derivation. Different optimal
293 settings were required to resolve (SNPs and PAVs) into linkage groups so the two datasets
294 were kept separate for this stage. DArT markers from a previous genetic linkage map
295 published in this population (Tomaszewski et al., 2012) were included in order to identify
296 linkage group designations. The pairwise recombinant fractions were estimated between
297 markers in the two datasets using est.rf() function of R/qtl and markers were grouped using
298 formLinkageGroups(). The grouping function resolved the SNPs to the expected 7 linkage
299 groups, but yielded 14 linkage groups for PAV markers, representing 7 paternal and 7

maternal grandparent derived sets of linkage groups. These were subsequently recoded B,D or A,C to reflect grandparental origin.

The output of R/qtl was used to create chromosomally designated locus genotype files for Joinmap 4.1 (DArT markers were not carried forward in the analysis). A framework map was created with only the SNP markers (with less than 40% missing data), using the maximum likelihood algorithm of JoinMap 4.1. To reduce map inflation due to low levels of genotyping error, a single round of imputation based error correction was performed. Graphical genotypes based on the maximum likelihood maps for each linkage group were exported and used as input for the genotype error correction module 'GBS Plumage for F2' (Spindel et al., 2013) with the setting of (-ct 1). This process identified all singletons (double recombinants) in the graphical genotypes and replaced them with missing values. The map order was then recalculated with the maximum likelihood algorithm of JoinMap 4.1 using the error corrected dataset. The final map comprised only non-redundant loci, as all identical loci are grouped into bins automatically during grouping in JoinMap 4.1.

PAV markers (and SNP loci that were excluded due to high missing data) were subsequently fitted to the framework map using a simple binning strategy. Pairwise recombination fractions were calculated between the SNP markers and the PAVs in JoinMap 4.1. PAV markers (and high missing data SNPs) were placed in the SNP bin on the framework linkage map with which they exhibited the lowest recombinant fraction (RF) value, with ties in RF being resolved by referring to the highest LOD score.

Results

Generating a reference sequence for GBS alignment

We generated a draft assembly of the low copy portion of the genome of the inbred paternal grandparent of the F2 Biomass population from Illumina-sequenced paired-end (PE), mate-

pair (MP) and long jumping distance (LJD) libraries. Assembly and scaffolding produced a final assembly of 1.11 Gbp in size, consisting of 424,750 scaffolds, with scaffold N50 of 25,212, contig N50 of 3,790 and GC content of 44.16%. This GC content is consistent with that of barley (Rostoks et al., 2002). The assembly size of 1.11 Gbp reflects only around 40% of *Lolium*'s total genome size, most likely as a result of the limitations of short-read sequencing when assembling complex plant genomes with many repetitive regions being collapsed into a limited number of contigs. Table 1 summarises the assembly statistics.

The shortest contig and scaffold in the assembly are equal in length at 143 bp. For completeness, we have included all sequences above this size in the released version of the assembly from this study. Consequently, short sequences (< 500 bp) are highly abundant, accounting for approximately 50% to 60% of the total number of sequences in the unscaffolded and scaffolded versions of the assembly respectively. However, by length, these sequences comprise only a small part of the assembly. The 254,591 scaffolds <500 bp that account for ~60% of the total of 424,750 scaffolds contain only 7.33% of the sequence, whilst 2.56% of scaffolds account for 50% of the sequence. Figure 1 illustrates the distribution of contig and scaffold lengths according to size range groupings.

Estimate of gene-space coverage

We used four methods to gain an insight into the coverage of the *Lolium perenne* gene-space by the assembly, based respectively on; a core reference set of proteins (CEGMA); the ability to find the majority of genes involved in controllingflowering; representation of both specific public *L. perenne* transcript assembly datasets and of a set of gene models associated with the recently released draft assembly of perennial ryegrass.

CEGMA defines coverage as either "complete" or "partial", based on the length of the aligned region. Complete coverage of a gene is defined as any alignment across over 70% of its

length and partial as less than 70% of the length aligned, but with significant identity. Of the 248 core proteins used by CEGMA to estimate completeness of coverage 239 (96.37%) were found to have complete alignment and 246 (99.19%) were found to have either complete or partial alignment within the *Lolium* assembly.

Because traits related to flowering are important in the utility of perennial ryegrass as a forage crop, we decided to investigate whether we had captured a significant number of the genes involved in the control of this characteristic. Higgins et al. (2010) have identified the probable rice and *Brachypodium* orthologues of ~50 genes involved in the induction of flowering in *Arabidopsis*. Homologues of 47 of the genes described by Higgins et al., (2010) and an additional gene *(CEN*) described by Comadran et al., (2012) in barley, were identified in the *Lolium* assembly by both BLAST and Exonerate methods. The genes were located on 48 individual scaffolds ranging in length from 227 bp to 140,152 bp (perhaps demonstrating the utility of retaining shorter scaffolds in the assembly). The majority of the genes were located on large scaffolds with substantial sequence both up- and downstream of the gene's position. Of the 48 genes, 33 appear to be complete models based on homology with rice, *Brachypodium* or barley, as indicated by Exonerate; 12 genes are classed as partial models due to genome scaffolding around the N-terminal, 2 lie on short scaffolds, and 1 is truncated by scaffolding around the C-terminal. A list of the scaffolds containing the 48 genes is available in Supplementary File 2.**[Supplementary Information]**.

The *Brachypodium*-based OGA transcriptome assembly described in (Ruttink et al., 2013) contained 46,459 sequences, with 41,120 (88.51%) of these being larger than 200 bp. BLAST searching (E-value=10) of these within the draft *Lolium* assembly matched 38,876 sequences (94.54%), with 27,427 (66.67% of total sequences, 70.55% of matched sequences) aligning with at least 95% identity over at least 70% of the query sequence length, based on the calculations for CIP and CALP described by (Salse et al., 2008). The 27,427 strongly matched

transcripts over 200 bp occurred on 11,778 scaffolds ranging from 220 bp – 282,695 bp, Ninety-six percent (11,315) of these scaffolds were over 5 kbp in length and these contained 26,888 (98.03%) of the strongly matched transcripts.

The transcriptome assembly described in (Farrell et al., 2014) contained 185,833 sequences, all over 200 bp in length. BLAST searching (E-value=10) matched 138,028 (74.28%) within the *Lolium* genome assembly; 109,320 (58% of total sequences and 79.20% of matched sequences) aligned with at least 95% identity over at least 70% of the query sequence length. The 109,320 strongly matched transcripts occurred on 14,934 scaffolds ranging from 202 bp – 282,695 bp.Ninety-one percent (13,576) of these scaffolds were over 5 kbp in length and these contained 106,644 (90.91%) of the strongly matched transcripts.

The assembly published by Byrne et al., (2015) contained sequences for 28,455 gene models BLAST searching (E-value=10) of these within the draft *Lolium* assembly matched 28,067 sequences (98.64%), with 22,563 (79.29% of total sequences, 80.39% of matched sequences) aligning with at least 95% identity over at least 70% of the query sequence length. The 22,563 strongly matched gene models occurred on 12,551 scaffolds ranging from 205 bp – 274,411 bp and with a maximum of 45 models on one scaffold; 11,662 scaffolds (92.92%) were over 5 kbp in length and contained 21,461 (95.12%) of the strongly matched models.

Gene models and transcripts from all sets combined were located on a total of 18,135 distinct scaffolds, totalling 570 Mb in length. 5,584 scaffolds were specific to the combined Farrell and Ruttink transcript sets and 1,855 were specific to the set of gene models from Byrne, with 10,696 scaffolds common to both sets. Figure 2 shows the distribution of scaffold sizes along with the numbers of transcripts/gene models aligned to them. Although the number of features differs greatly between the two transcript sets and the gene model set, the distribution of scaffold size bins and number of features in each bin is consistent between the sets, with

larger (over 5 kb) scaffolds being much more prevalent and containing the large majority of transcripts/gene models. In particular, a sharp increase in the number of scaffolds and transcripts/models occurs in the 10 – 20 kb size range, with scaffold count tailing off rapidly, but transcript/model numbers remaining high before beginning to tail off beyond 50 kbp. This trend is reflected in the cumulative totals, with a steep rise in numbers occurring between the 5 kb and 50 kb size ranges and then rapidly levelling out beyond 50 kb.

Accessing the draft assembly

As part of this study we present a genome browser that allows dynamic viewing of the assembly, with tracks for the features described above. In addition, we have also provided additional layers of annotation based on preliminary *de-novo* prediction and homology-based methods involving comparisons to barley, rice and *Brachypodium distachyon*. Tracks for all of the features listed below are also available on the browser.

Prior to preliminary annotation, RepeatMasker version 3.2.8 was used to identify common repetitive elements in the scaffolded *Lolium* assembly using a wheat-based model. The majority of repeats identified in *Lolium* belonged to the retroelement and DNA transposon classes of repeat; this led to 67.77 Mbp of sequence being masked, or 6.09% of the assembled genome sequence. Supplementary File 3 details the repeat content identified **[Supplementary Information]**.

The widely-used, open-source, gene prediction tool Augustus (Stanke and Waack, 2003) was used for gene prediction using the repeat masked genome assembly, with a wheat-based gene model. In total, 188,842 predicted entities were identified from 59,903 scaffolds, with a maximum count of 74 entities on one scaffold. Three scaffolds are likely to be mitochondrial, representing 521.8 Kbp and containing 20 predictions; Augustus did not predict gene models for scaffolds known to be associated with the chloroplast. This prediction of 188,842 entities

423 is clearly a gross overestimate and will reflect a number of confounding factors, including
424 retroelements, pseudogenes, gene fragments and sequencing errors.
425
426 BLAST searching of barley cDNA sequences and publicly available peptide sequences for
427 barley, rice and Brachypodium was carried out using default cut-off parameters (E-value =
428 10). Using this lenient threshold, 99.75% (26,094) of 26,159 barley peptide sequences,
429 98.42% (30,539) of 31,029 Brachypodium peptides and 88.8% (58,905) of 66,338 rice
430 peptides exhibited matches to the alignment. As expected from their phylogenetic and
431 ancestral relationship, a greater proportion of barley peptide sequences were found to have
432 matches within the Lolium assembly. In total 19,477 scaffolds were found to contain
433 homology matches to any of the above datasets.

434 **GBS library construction and sequencing results**

435 GBS libraries were developed for 169 individuals, the F1 parent, and the paternal grandparent
436 of the mapping population following the protocol of (Elshire et al., 2011) using the restriction
437 enzyme *PstI* and sequenced on an Illumina Hiseq2000 to generate single end 100 bp reads. In
438 total, sequencing yielded 284,908,063 reads for the progeny genotypes. Reads were de-
439 multiplexed and, to maintain a consistent read length and quality, the reads were trimmed to
440 66 bp. After cleaning, an average of ~1.7 million reads per individual was obtained.

441 **Alignment of GBS reads to the assembly**

442 The *Lolium* shotgun assembly described above (with 424,750 scaffolds) was used as a
443 reference sequence for SNP variant identification in the F2 Biomass population. The
444 sequences from the de-multiplexed individual fastq files were aligned to the reference set
445 allowing two mismatches. Of the total 284,908,063 reads, 164,285,405 (57.6%) reads had at
446 least one reported alignment. On average, 58% of reads from each individual aligned to the

447 reference genome. Overall, 15,118,076 reads (comprising 5.3% of the total) failed to align to
448 the reference due to the alignment option that allowed alignments only for reads that mapped
449 uniquely to the reference. A further 105,504,582 reads (37.3%) failed to align under the
450 settings used.

451 In total, there were 213,310 *PstI* restriction sites located on 64,977 scaffolds in the assembly.
452 These scaffolds accounted for 75% of the total size of the assembly (834,624,995 bp), with
453 the remaining 359,774 scaffolds accounting for only 25% of the total size of the assembly
454 (277,380,681 bp). Out of the 64,977 scaffolds possessing *PstI* sites, 26,954 (41.5%) have at
455 least one GBS read aligning to them, and these scaffolds contain a total of 111,903 *PstI* sites
456 (52.4% of total *PstI* sites in the assembly).

457 **SNP variant identification**

458 Using Varscan (minimum read depth of 8; minimum 2 reads to call variants, minimum
459 average phred quality base score of 20, variant allele frequency of 0.2), 22,805 SNP positions
460 were reported. This included variants which were monomorphic amongst the progeny
461 individuals, but which differed with the reference nucleotide at that position. Amongst these,
462 a total of 9127 variants exhibited at least one variant in the population and were biallelic. Of
463 the 9127 variants, majority of them were of transition type with C/T and A/G type accounting
464 to 31% and 29% respectively. The remaining SNPs were of transversion type with C/G, G/T,
465 A/C and A/T type accounting to 15%, 9%, 9%, 6% respectively (Table 2). The R/qtl function
466 geno.table() was used to examine the segregation pattern of the markers and 4329 out of 9127
467 markers were eliminated due to severe departure from the expected Mendelian segregation
468 ratio (1:2:1) using a cut-off P-value < 1e-10. The remaining 4798 SNP markers were used for
469 map construction. The identity and location of these SNP variants is provided as a track on
470 the JBrowse of the assembly.

Presence/Absence variant (PAV) identification

Although the GBS approach was originally envisaged primarily as a method for genome-wide SNP discovery, a second type of variation has also been reported in many studies involving its use (Elshire et al., 2011). This variation manifests itself in the presence of alignments at a locus for some individuals versus the lack of alignments for other individuals. Such Presence/Absence variation can arise due to several events (SNPs and small InDels at restriction sites, larger InDels, inversions etc) all of which have the effect of disrupting the formation of *PstI*-site bounded fragments in the size range being selected for by the PCR-amplification step for some alleles at a locus, while such fragments are present for other alleles. The result is the segregation of the presence of the fragment as a dominant marker in the population, with the exact mode dependant on the allelic configuration and population type involved.

Importantly, PA variation can actually far exceed SNP variation in GBS studies (Elshire et al., 2011). Because of its potential to add significant numbers of markers we decided to explore the use of a very simple two-step filtering approach to identify potential PAVs in the F2 Biomass population using a series of UNIX commands (see methods section for details not included below).

There were 111,903 independent loci in the genome that have at least one read aligning to them. For each locus, the individuals with alignments were scored as present and individuals with no alignments were scored as absent. We then filtered this table of variants to identify marker loci that satisfied two criteria (in the following order):

PAVs are expected to exhibit a Mendelian segregation ratio of 3:1 (Presence:Absence) in an F2 population. In the F2 Biomass population, the ideal expected ratio is 127:42. To take into account the known existence of segregation distortion within this population (Anhalt et al.,

2008), data were filtered to identify loci with alignments (potential "Presence" category variants) to between 50%– 90% of the total population. This reduced the number of candidate loci to 20,180.

Lack of read alignment for potential "Absence" variants might be due to segregation of the recessive allele, but might also be due to a technically derived lack of read coverage. This latter class is effectively missing data, but such instances cannot easily be distinguished from "Absence" variants on a case by case basis. To minimise this confounding effect, we screened the remaining marker loci to identify those exhibiting a mean read depth of no fewer than eight alignments per individual to identify loci with an "on-average" reasonable read depth. Of the 20,180 loci from the previous round a total of 7,714 potential PAVs remained after this filtering step.

Construction of a high density SNP and PAV-based genetic linkage map of the F2 Biomass population

A total of 4,798 SNPs and 7,714 PAVs were carried forward for linkage map construction. In order to identify and orient linkage groups, segregation data for 326 DArT markers previously used for map construction in this F2 Biomass population (Anhalt et al., 2008) were included in early rounds of the analysis (grouping and early rounds of mapping prior to error correction), but removed for later rounds.

The SNP and PAV markers were initially assigned to linkage groups using R/qtl. The two subsets of markers (co-dominant SNPs and dominant PAVs) were grouped separately as different optimal settings were required to efficiently resolve the different subsets into linkage groups. At thresholds for recombinant fraction RF/LOD of 0.11/7, SNP markers resolved into seven linkage groups (identified by the presence of DArT markers). Likewise, the PA markers resolved to 14 linkage groups at RF/LOD thresholds of 0.12/10.

519 Of the 4,798 SNP markers, 3,105 grouped into seven large linkage groups and were used for
520 subsequent map calculation. The number of SNP markers per linkage group ranged from 269
521 to 563. Of the 7,714 PAV markers, 7,265 resolved into 7 linkage groups, with a range of
522 from 903 to 1426 markers per linkage group.

523 We adopted a two stage mapping process, using the co-dominant SNP markers to construct a
524 framework map, to which we subsequently fitted the PAV markers using a binning approach.
525 After grouping the markers into linkage groups and removal of non-redundant loci, an initial
526 round of marker ordering was performed using the maximum likelihood algorithm of
527 JoinMap 4.1 for the SNP markers. The resulting linkage groups ranged in size from 687cM to
528 1324cM. Given that the entire map length for the previous DArT marker-based map of the F2
529 Biomass population was 966cM, these map lengths were vastly overinflated. This
530 phenomenon is well established in the production of ultra-high density genetic linkage maps
531 with relatively low resolution, where the cumulative effect of low levels of genotyping error
532 (yielding false recombination events between markers) results in artificial map expansion
533 when analysed with more "traditional" mapping algorithms and approaches such as those
534 implemented in Joinmap (van Os et al., 2006).

535 In order to address this problem, we decided to adopt a conservative approach to correct
536 potential genotyping errors, followed by removal of redundant marker data in order to
537 decrease map length while maintaining accuracy of marker order. From the maximum
538 likelihood maps produced in JoinMap, graphical genotypes were generated for each linkage
539 group in the framework SNP map (Figure 3). These were used as input files for the 'GBS
540 Plumage for F2' utility (Spindel et al., 2013) specifically designed to deal with genotype error
541 correction in F2 population types. Erroneous genotype calls usually manifest themselves as
542 apparent double recombinants. Using the default setting of GBS Plumage, potential double
543 recombinants in progeny linkage groups (rendered as graphical genotypes) were identified

544 and replaced with missing values. After error correction, a second round of ordering was
545 performed in JoinMap 4.1 - pairwise recombinant fractions between all pairs of markers were
546 calculated on the error-corrected and re-ordered linkage group using the maximum likelihood
547 mapping algorithm (Figure 3).

548 In total, 1865 unique bins representing 10,352 markers were used to calculate the map. The
549 total final map length was 952.6cM, which is in keeping with previous map lengths for
550 perennial ryegrass and specifically, for this population, indicating that the error correction and
551 redundancy removal were effective. The number of markers in each chromosome ranges
552 from 845 to 1987 (Table 3). The number of unique bins for each chromosome ranges from
553 179 to 331. Average spacing between unique markers across all the chromosomes was 0.4cM
554 with the maximum spacing of 15.8cM.

555 Markers in the map were defined by alignment to the reference gene-space assembly
556 produced in the paternal grandparent. The 10,352 markers on the map represent 4767 unique
557 scaffolds in the assembly. The majority of the scaffolds anchored were in the size range
558 between 10Kb and 100Kb (Table 4). The total size of the 4,767 scaffolds accounts for 18%
559 (200 Mbp) of the total size of the reference assembly. Supplementary File 4 contains a
560 complete list of the markers, genetic order and identity of anchor markers used to create bins,
561 the bin assignment of the remaining markers, and a list of the unique scaffolds anchored by
562 the markers **[Supplementary Information]**.

563 The number of markers observed per scaffold ranged from 1 to 15. Out of the 4,767 scaffolds
564 that were anchored with GBS markers, 175 scaffolds had 720 markers on them that were
565 mapping to more than one chromosome. Of the remaining 4,591 scaffolds, 1,007 scaffolds
566 comprising 1,590 markers were anchored just by SNP markers, 2,877 scaffolds comprising
567 5331 markers were anchored just by PA markers, and 707 scaffolds comprising 2,711

markers were anchored by both SNP and PA markers. The total space in the assembly anchored by scaffolds mapping to multiple chromosomes accounted for 1% (1,997,625 bp). This could be due to misassembly of the scaffolds, but might also arise from events such as incorrect alignment of fragments to the reference genome.

The GBS-based map of the F2 Biomass population is defined by a considerably larger number of PAVs than SNPs (more than twice as many PAVs than SNPs). Because of their dominant nature, PAV markers are more prone to genotype scoring error, largely due to the difficulty in distinguishing the recessive allelic state (absence of an alignment) from a technically-derived lack of read coverage on a per genotype basis. The existence of 707 scaffolds anchored by both PAV and SNP markers afforded an opportunity to test the accuracy of the PAV markers relative to the more informative SNP markers.

We examined the pairwise recombination fraction (from the JoinMap pairwise data file) between all pairs of SNPs and PAVs occupying the same scaffold for all 707 scaffolds. Given the resolving power of the population and the maximum size of the scaffolds in our assembly, these markers should generally co-segregate. Out of total 1,455 pairwise observations between SNP and PAV markers on the same scaffolds, 430 (30%) had pairwise recombination fractions less than 0.01 and 932 (64%) observations had pairwise recombination fractions between 0.01 and 0.05. A further 60 (4%) observations had pairwise recombination fractions between 0.05 and 0.1, and the remaining 33 (2%) had recombinant fractions exceeding 0.1. In order to test how well the binning strategy to place the PAV markers on the map performed, we also examined the map distance between SNP and PA markers occurring on the same scaffold according to which non-redundant bin they occupied on the final map. Out of the same 1,455 pairwise comparisons, 852 (59%) pairs were separated by less than 1cM, 318 (22%) were separated by between (1 and 5 cM), 132 (9%)

had map distance in between (5 and 10cM) and the rest 153 (10%) had greater than 10cM map distance in between them.

Previous work on the F2 Biomass population showed the presence of unusually high levels of segregation distortion. Anhalt et al. (2008) showed that 63% of the total markers used in an AFLP and SSR based map of the population showed segregation distortion, a level twice that observed in other mapping populations of perennial ryegrass used for comparison in the same study. Linkage groups (LG) 3, 5, 6, and 7 were reported to have high level of segregation distortion and LG 2 and LG 4 with least amount of segregation distortion. In particular, LG 6 was reported to be completely distorted.

The GBS-based SNP map of the population also exhibited significant levels of distortion, but the overall level was much lower than that observed by Anhalt et al. (2008). Out of 10,352 markers and 1865 unique bins on the GBS map, 4,357 (42%) markers and 618 (33%) bins exhibited segregation distortion (P-value < 0.05). This is a twofold discrepancy with the figure found in the previous study. However, the observations of (Anhalt et al., 2008) were based on only 75 markers, with marker densities ranging from only 8-17 per linkage group. To investigate this apparent discrepancy, we placed all of 75 markers from the map of the F2 Biomass population presented by (Anhalt et al., 2008) on to the combined GBS map. As expected, these markers mapped to areas exhibiting segregation distortion in the current map. However, it is apparent that increased marker coverage on the current map is yielding better representation of areas exhibiting lower levels of segregation distortion which were significantly under-represented on the previous map (Figure 4 and Figure5). While segregation distortion is apparent on all chromosomes, the majority of distorted markers are from LG 6 (96% of marker bins distorted) and LG 3 (57% of marker bins distorted), and together these LGs account for over half (57%) of distorted loci on the map. Thus, as well as higher marker density and coverage, the current map of the F2 Biomass population exhibits

617 better representation of both distorted and non-distorted genome regions, which could
618 represent a useful feature in trait mapping experiments in the future.

619 **Homozygosity level of the genotype used for the reference sequence**

620 Both the F1 parent and paternal grandparent were also subject to GBS with the progeny
621 individuals. Unfortunately, the maternal grandparent of the population was no longer in
622 existence at the time of the study, and so could not be examined. However, inclusion of the
623 paternal inbred grandparent, which was also used for the reference sequence assembly,
624 yielded the opportunity to examine the extent of homozygosity of this inbred line. This
625 feature of the paternal grandparent is particularly interesting, since the extent of
626 heterozygosity could yield insights whether there is a requirement to account for extensive
627 presence of biallelic loci in the assembly, or in gene-expression based experiments involving
628 this interesting experimental genotype. Out of 3030 loci from the paternal grandparent
629 mapped using GBS in an F2 population, 3015 of them were of homozygous calls, 11 calls
630 were of heterozygous type and four calls representing alleles from another parent (these may
631 represent "missed" heterozygote calls). Thus, the mapping data indicate that the paternal
632 grandparent and reference sequence genotype is ~99% homozygous (Figure 6).

633 **Anchoring of gene-containing scaffolds**

634 Comparison of the 4,767 GBS-anchored scaffolds with the 18,135 scaffolds that had good
635 matches to the Byrne gene models and the Ruttink and Farrell transcript sets identified 3,679
636 anchored scaffolds (79.32%) that contained transcripts/gene models from any set. The total
637 length of anchored scaffolds containing matches to transcripts or gene-models is 184.36 Mbp,
638 corresponding to 92% of the total cumulative length (200Mb) of anchored scaffolds. In terms
639 of the total proportion of the potential "genespace" of *Lolium perenne* anchored in the study,
640 the 18,135 genic scaffolds cover approximately 570Mb, and we have anchored approximately
641 one-third of this. Use of methylation sensitive enzymes in GBS is expected to target genic

642 areas, and the results for the *Pst1*-based approach used in this study support the veracity of
643 this expectation, with the vast majority of anchored scaffolds showing evidence of being
644 gene-containing.
645
646 In order to gain an insight into the performance of the synteny-based approach for
647 chromosomal anchoring of scaffolds adopted by Byrne et al. (2015) relative to the direct
648 anchoring of scaffolds to chromosomes via genetic mapping in this study, we focused on the
649 48 scaffolds containing flowering related genes that we identified in our assembly. On
650 examination, 22 of the 48 scaffolds were directly anchored to our genetic map
651 (Supplementary File 2) **[Supplementary Information]**. We identified the "equivalent"
652 scaffolds in the assembly of Byrne et al. (2015), defining "equivalent" as a scaffold that
653 contained the probable flowering gene orthologue as identified by Exonerate, but also
654 exhibited a BLAST-based similarity profile on a scaffold level that confirmed that each match
655 represented the orthologous genomic region (for simplicity, we ignored scaffolds from the
656 Byrne et al assembly that overlapped our scaffold, but did not contain the flowering gene).

657 Using this approach, we found 47 equivalent scaffolds in the Byrne et al. (2015) assembly,
658 but were unable to resolve an equivalent for our FRI-containing scaffold due to multiple
659 strong matches (Supplementary File 2). All of the 47 matched genes appear to be represented
660 in the Byrne at al. (2015) assembly by complete models (or in one case a near-complete
661 model) based on homology with rice and *Brachypodium*, and all possessed gene models from
662 the annotation associated with the assembly.

663 Comparing the 22 specific scaffolds for which we have a genetic location to the chromosomal
664 assignment for the equivalent scaffolds in the Byrne et al.(2015) assembly revealed that, in
665 18 cases, the chromosomal assignments agreed, whilst in 4 cases, there were conflicts. We did
666 not compare the relative location within chromosomes of matching results due to the widely

667 differing map lengths of individual linkage groups inour map and the reference map used for
668 anchoring the Genome Zipper. However, at a whole chromosome level, the scaffolds
669 containing *Lolium* homologues of the genes *AP1*, *CEN*, *FCA* and *FIE1* were placed on
670 chromosomes 2, 6, 2 and 1 respectively by Byrne et al. (2015), but were anchored to
671 chromosomes 3, 5, 5 and 3 respectively in our map. (Supplementary File 2). For *CEN*, *FCA*
672 and *FIE1*, these scaffold locations were supported by multiple PA and/or SNP markers in the
673 map, whereas the scaffold containing *AP1* was anchored by a single SNP used to create the
674 framework map. Assuming that, in general, scaffolds directly anchored by multiple markers,
675 or single high confidence markers are robustly assigned, this means that the synteny-based
676 method has resulted in incorrect chromosomal assignments for 18% of these scaffolds.

677 **Discussion**

678 GBS offers a magnitudinal increase in our ability to create densely populated genetic maps in
679 an extremely cost and time effective manner. A genetic linkage map was successfully created
680 containing over 10,000 markers, located in 1865 non-redundant bins, using 169 individuals of
681 the well characterised perennial ryegrass F2 Biomass mapping population. Experience to date
682 suggests that, once the methodology and basic resources are established, dense maps of this
683 sort can be created in a matter of weeks.

684 Although it is possible to perform GBS in the absence of a reference sequence, early pilot
685 experiments using the mapping population suggested that auto assembly of GBS reads to
686 create a reference sequence, as performed in other studies (Russell et al., 2014, Chen et al.,
687 2013), could be problematic, with relatively minor changes in assembly parameters causing
688 relatively large differences in the resulting assemblies (data not shown). Because of this, it
689 was decided to generate a reference sequence to which to align GBS reads. In this case, the
690 only existing inbred grandparental line of the mapping population was used (the paternal
691 grandparent). As outlined previously, this genotype is a tenth generation inbred line

originating from a CMS programme (Connolly and Wrightturner, 1984). Theoretically, a genotype at this level of inbreeding should retain well below 1% heterozygosity, making it an ideal candidate for use in a genome assembly initiative, since only a single haplotype is expected to be present for the majority of the genome. Near complete homozygosity obviates the problems associated with assembly associated with a highly heterozygous species in which SNP densities have been estimated at in the region of 1 SNP every 30 bp (Xing et al., 2007). Inclusion of the paternal grandparent in the GBS experiment allowed us to confirm the expected high levels of homozygosity, with only 0.5% of over 3000 mapped SNP markers present in the paternal grandparent deviating from the expectation of homozygosity.

It is important to note that our study has taken place against the backdrop of the recent release of a more complete synteny-based draft genome sequence of *Lolium perenne* by Byrne et al. (2015). That assembly was generated in a sixth generation inbred line of perennial ryegrass (P226/135/16). Utilising a similar mixture of Illumina paired end, mate pair and long jumping distance library sequencing that we adopted, Byrne et al. (2015) additionally used long read PacBio sequences equivalent to nine-fold coverage of the genome for closure of assembly gaps. Their resulting assembly captured 1128 Mbp of the perennial ryegrass genome in 48415 scaffolds (67024 contigs) with a scaffold N50 of 70,062 bp (contig N50: 16370 bp). These figures account only for scaffolds and contigs in excess of 1kb, and adjusting for this by also only considering sequences in excess of 1kb, by comparison, our assembly captures 977 Mbp of the genome in 90,787 scaffolds (166,217 contigs) with a scaffold N50 of 32,299 bp (contig N50: 5,559 bp). Based on this comparison, our assembly offers slightly lower genome coverage, which is captured in just under double the number of scaffolds. Byrne et al.(2015) also utilised multiple RNA-seq datasets to generate a comprehensive annotation comprising 28455 genes on 13725 scaffolds that accounted for 796 Mbp of their assembly. Subsequently,

using the synteny driven Genome Zipper approach (Pfeifer et al. 2013), they organised a total of 13411 scaffolds (approximately 800 Mbp in total) and 10464 annotated genes into a linear order on the perennial ryegrass genome by virtue of comparison to the reference genomes of *Brachypodium*, rice and Sorghum.

We utilised the gene models generated by Byrne et al. (2015), in addition to extensive transcript datasets generated in *Lolium perenne* by Ruttink et al. (2013) and Farrell et al. (2014) to identify the gene containing portion of our assembly, identifying scaffolds comprising 570Mbp in total length that contain high confidence matches to these *Lolium perenne* sequences. The ~10,000 GBS tags comprising the map of the F2 Biomass population anchor 4767 scaffolds, equivalent to ~200 Mb of the assembly. Although this is considerably lower than the total length assigned a chromosomal location and order by Byrne et al. (2015), the anchoring is more direct in nature, and we used this feature to test the performance of the synteny-based approach by comparing chromosomal assignments for a small subset of genes involved in flowering. The comparison reveals that synteny-based anchoring performs well, with over 80% concordance between genetic mapping and synteny-based results at a "whole chromosome" assignment level. However, the results also demonstrate that, while synteny-based anchoring is a powerful approach, GBS-based genetic mapping in this and other populations may also contribute to the long term goal of producing a more comprehensive, chromosomally anchored pseudomolecule assembly of perennial ryegrass in the future through validating and augmenting synteny based assignments.

Reduced representational sequencing approaches for genotyping are largely based on the concept of characterising the same sequence tag across all individuals in the study population, with variation being detected within the window of sequence covered by the tag. However, a

741 variety of polymorphic events (e.g. SNPs and InDels at the restriction sites being used for
742 complexity reduction) can cause a second type of polymorphism which manifests itself in the
743 form of the differential detection of the presence of aligned tags in different individuals. This
744 presence absence variation (PAV) has in fact been observed at a frequency far higher than the
745 occurrence of SNP variation (Lu et al., 2015). For instance, 80 - 90% of the maize genome is
746 reported to show some PAVs (Chia et al., 2012) and recently 1.1 million PAVs have been
747 mapped to the maize pan genome (Lu et al., 2015). Given the potential for PAVs to add
748 significantly to the marker density of the map (and the extent of genome anchoring), we
749 decided to develop and test some simple procedures to both score and map them in this study.

750 Since PAVs manifest themselves at the alignment stage, we adopted a relatively
751 straightforward approach based on the bowtie-generated alignment files for loci exhibiting the
752 footprint of PA variation. Because lack of alignment at a locus in any individual could come
753 from technical sources such as sequence-under-representation in the GBS libraries, a filtering
754 process based on read depth across all individuals was used to identify loci in which this was
755 not a general problem, followed by the imposition of a requirement to conform to the
756 expected Mendelian segregation pattern for dominant markers in the population. In
757 recognition of the fact that the population exhibits segregation distortion, and that lack of read
758 coverage at individual loci could still contribute to apparent “Absence” variants, a more or
759 less arbitrarily determined window around the expected 3:1 ratio was used, allowing for either
760 a threefold increase or decrease of the “Presence” category, equivalent to ratios between 9:1
761 and 1:1.

762 For the mapping component of the study, early attempts at incorporating the PAVs directly in
763 the map were problematic, probably due to a mixture of incomplete genotype information
764 (heterozygotes {Aa} and homozygotes {AA} are indistinguishable) combined with a higher
765 potential for miscalls, resulting in vastly inflated map distances. To circumvent this, a high

766 quality framework SNP-based map was generated and adopted a very simple binning
767 approach to place PAVs into this (fixed order and distance) framework map. This maintained
768 the integrity of the map produced using the more robust SNP markers, whilst allowing the
769 utilisation of the considerable amount of anchoring information associated with the PAVs.
770 Over half of the total mapped scaffold length (113 Mb of 200Mb) was anchored solely by
771 PAVs and whilst there is an expectation that chromosomal position of markers anchored by
772 PAVs might be inherently less accurate, we felt that inclusion of this information would be
773 beneficial to future gene-discovery applications as long as: 1) the inclusion of the PAVs did
774 not degrade the map, and 2) the potential accuracy range of the PAVs was reasonably well
775 understood.

776 Adopting a binning process addresses the first of the two points above. An attempt to quantify
777 the second was undertaken using the hypotheses that: in general SNP markers were accurately
778 placed relative to PAVs; and that SNPs and PAVs occupying the same scaffolds (given the
779 N50 scaffold size of the assembly is 25 Kb) should theoretically co-segregate in the absence
780 of genotyping error and missing data. Over 90% of SNP-PAV pairs occupying the same
781 scaffold had pairwise recombination fractions of no more than 5% (0.05), and just over 80%
782 of PAVs ended up in a final bin no more than 5cM away from their physically-paired SNP
783 anchor marker. Assuming a low error rate in the SNP dataset, each percentage point of error
784 in genotype calling in the PAVs will be translated into a 1% increase in the recombination
785 fraction between the “reference” SNP and the “query” PAV in question. Our results suggest
786 that our filtering approach is managing to identify PAVs with low error rates (~30% below
787 1% error and ~60% below 5% error), with the binning process placing the majority of the data
788 at a map distance consistent with these recombination fractions. There is no doubt that more
789 sophisticated approaches to both identify and map PAVs in similar studies could be
790 implemented, but this study demonstrates that, even using the relatively straight forward

791 approaches adopted here, PAV markers can contribute significantly to anchoring mapped
792 markers to sequence assemblies in pairwise mapping populations subjected to GBS

793 The F2 Biomass population has, in the past been used as an exemplar for high levels of
794 segregation distortion (SD). Anhalt et al. (2008) indicated the occurrence of levels of SD
795 exceeding 60%. The SNP framework map presented here contrasts with the previous results,
796 with 42% of SNP markers, or 33% of markers representing the signatures for the non-
797 redundant bin-set exhibiting distortion. Placing the markers used in the previous study on this
798 map gives insights into the discrepancy, which seems to be due to a mixture of low marker
799 density and unfortunate distribution of the markers in the previous study. The majority of the
800 distortion appears on LG 6, which exhibits an under-representation of the grandparental-
801 derived genome, and the top two thirds of LG 3, which exhibit similar patterns of distortion.
802 The source of the extensive SD in these regions remains unknown. To our knowledge,
803 extensive distortion of the level observed on LG 6 is unique to the F2 Biomass population.
804 Segregation distortion of LG 3 has previously been associated with the presence of the self-
805 compatibility locus F in the International Lolium Genome Initiative (ILGI) reference mapping
806 population (Thorogood et al., 2002). LG 3 of the F2 Biomass population was the focus of a
807 study by (Manzanares, 2013), who tested the hypothesis that the *F*-locus was responsible for
808 the self-compatibility phenotype in the population, in part based on the observations of the SD
809 on LG 3 in this population. Although the conclusion was based on representation of LG 3 by
810 only 4 markers, the results indicated that LG 3 is not involved in the self-compatibility
811 phenotype of the F2 Biomass population, although the trait is controlled by a single self-
812 compatibility locus. Thus the SD observed on this chromosome seems unrelated to the *F*-
813 locus, and remains unexplained. The existence of a physically anchored, SNP-based map of
814 the F2 Biomass population leads to interesting prospects for establishing the identity of the
815 locus responsible for the self-compatibility phenotype in the population, in order to

816 understand whether one or several such loci exist in the available self-compatible *L. perenne*
817 genotypes available. Regardless of the source of the variation, a more accurate understanding
818 of the distribution of distortion over the map is important in the continuing utility of the F2
819 Biomass population as a key reference population for future trait-mapping and discovery
820 genetics applications.

821 **Conclusion**

822 Our main goal in this study was to produce a high-density, heavily chromosomally anchored
823 genetic map in a key reference mapping population in perennial ryegrass. We adopted a
824 highly inclusive approach, maximising the number of anchored fragments by utilising both
825 SNP and the more frequent PA variation revealed by GBS. Combined with existing and
826 emerging genomic resources such as the recently published synteny-based draft genome
827 sequence of the species released by Byrne et al. (2015), and hopefully more comprehensive
828 assemblies that will be built in the near future,the current map will be a useful tool for
829 understanding the genetic basis of numerous traits for which it segregates. For example, a
830 phenotypic dataset for the segregation of polar secondary metabolites already exists as an
831 extension of the study on the mapping of non-polar metabolites recently published by (Foito
832 et al., 2015), and (interestingly in the context of the flowering-associated genes described) the
833 population also segregates for heading date. Unlike maps produced in the species to date, the
834 high level of direct and indirect anchoring to two perennial ryegrass assemblies yields the
835 potential to routinely identify candidate genes underlying mapped traits. In addition to its
836 future use in trait genetic analysis, the F2 Biomass population is also the source population for
837 the long term goal of the generation of a recombinant inbred line (RIL) population for
838 perennial ryegrass.

839 Finally, the availability of a draft assembly of a second perennial ryegrass genotype, in
840 addition to that of genotype P226/135/16 will allow comparisons that may yield useful

841 insights into intra-specific variation in genome structure in *Lolium perenne*, similar to those
842 that have been enabled by the availability of significant genome-wide sequence information
843 of multiple haplotypes in other recently characterised species(Xu et al., 2011, Deokar et al.,
844 2014, Wilson et al., 2015).

845 **Supplementary Information**

846 The supplementary information for this manuscript comprises four files. Supplementary File 1
847 provides details about the types of libraries used for sequencing for the reference assembly
848 and raw fold-coverage achieved. Supplementary File 2 lists the 48 flowering-associated genes
849 identified in the study, along with associated scaffold ID, location on that scaffold and the
850 chromosomal assignment for the 22 scaffolds we anchored. The corresponding scaffolds from
851 Byrne et al. (2015), and their chromosomal assignment are also shown for comparison, as
852 well as completeness of the gene model from both assemblies. Scaffolds with conflicting
853 chromosomal assignment are highlighted in blue. A direct link to the relevant scaffold
854 location within the genome browser is also given, Supplemenary File 3 gives a summary of
855 repeat types found during repeat masking with wheat-based model. Supplementary File 4 is a
856 multi-tab Excel spreadsheet containing the map positions for all the markers, the non-
857 redundant set of 1865 bins used to calculate the core map, number of markers present in each
858 bin, list of unique scaffold names anchored using the map, and a list of scaffolds that maps to
859 multiple chromosomes.
860

861 **Funding**

862 This study was funded by Teagasc core funding and Teagasc PhD Walsh Fellowships to JV
863 and EM

864 **Acknowledgements**

The authors wish to acknowledge Trinity College Dublin (Elaine Kenny), University College Dublin (Alison Murphy) and the Oslo Sequencing Centre (Lex Nederbragt & Gregor Gilfillan) for their technical expertise in the sequencing for the assembly described in the manuscript. We also thank Dr Tom Ruttink, ILVO, Belgium for access to *Lolium perenne* OGA assembled transcript data, and Dr Stephen Byrne for useful discussions that contributed to the revised draft of the manuscript.

Literature cited

Altschul SG, W.Miller,W.Myers,EW.Lipman,DJ. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology,* **215**: 403-410.

Anhalt U, Heslop-Harrison J, Piepho H, Byrne S, Barth S. 2009. Quantitative trait loci mapping for biomass yield traits in a Lolium inbred line derived F-2 population. *Euphytica,* **170**: 99-107.

Anhalt U, Heslop-Harrison P, Byrne S, Guillard A, Barth S. 2008. Segregation distortion in Lolium: evidence for genetic effects. *Theoretical and Applied Genetics,* **117**: 297-306.

Bartos J, Sandve S, Kolliker R, Kopecky D, Christelova P, Stoces S, Ostrem L, Larsen A, Kilian A, Rognli O, Dolezel J. 2011. Genetic mapping of DArT markers in the Festuca-Lolium complex and their use in freezing tolerance association analysis. *Theoretical and Applied Genetics,* **122**: 1133-1147.

Bentley D, Balasubramanian S, Swerdlow H, Smith G, Milton J, Brown C, Hall K, Evers D, Barnes C, Bignell H, Boutell J, Bryant J, Carter R, Cheetham R, Cox A, Ellis D, Flatbush M, Gormley N, Humphray S, Irving L, Karbelashvili M, Kirk S, Li H, Liu X, Maisinger K, Murray L, Obradovic B, Ost T, Parkinson M, Pratt M, Rasolonjatovo I, Reed M, Rigatti R, Rodighiero C, Ross M, Sabot A, Sankar S,

Scally A, Schroth G, Smith M, Smith V, Spiridou A, Torrance P, Tzonev S, Vermaas E, Walter K, Wu X, Zhang L, Alam M, Anastasi C, Aniebo I, Bailey D, Bancarz I, Banerjee S, Barbour S, Baybayan P, Benoit V, Benson K, Bevis C, Black P, Boodhun A, Brennan J, Bridgham J, Brown R, Brown A, Buermann D, Bundu A Burrows J, Carter N, Castillo N, Catenazzi M, Chang S, Cooley R, Crake N, Dada O, Diakoumakos K, Dominguez-Fernandez B, Earnshaw D, Egbujor U, Elmore D, Etchin S, Ewan M, Fedurco M, Fraser L, Fajardo K, Furey W, George D, Gietzen K, Goddard C, Golda G, Granieri P, Green D, Gustafson D, Hansen N, Harnish K, Haudenschild C, Heyer N, Hims M, Ho J, Horgan A, Hoschler K, Hurwitz S, Ivanov D, Johnson M, James T, Jones T, Kang G, Kerelska T, Kersey A, Khrebtukova I, Kindwall A, Kingsbury Z, Kokko-Gonzales P, Kumar A, Laurent M, Lawley C, Lee S, Lee X, Liao A, Loch J, Lok M, Luo S, Mammen R, Martin J, McCauley P, McNitt P, Mehta P, Moon K, Mullens J, Newington T, Ning Z, Ng B, Novo S, O'Neill M, Osborne M, Osnowski A, Ostadan O, Paraschos L, Pickering L, Pike A, Pinkard D, Pliskin D, Podhasky J, Quijano V, Raczy C, Rae V, Rawlings S, Rodriguez A, Roe P, Rogers J, Bacigalupo M, Romanov N, Romieu A, Roth R, Rourke N, Ruediger S, Rusman E, Sanches-Kuiper R, Schenker M, Seoane J, Shaw R, Shiver M, Short S, Sizto N, Sluis J, Smith M, Sohna J, Spence E, Stevens K, Sutton N, Szajkowski L, Tregidgo C, Turcatti G, van deVondele S, Verhovsky Y, Virk S, Wakelin S, Walcott G, Wang J, Worsley G, Yan J, Yau L, Zuerlein M, Mullikin J, Hurles M, McCooke N, West J, Oaks F, Lundberg P, Klenerman D, Durbin R, Smith A. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* **456**: 53-59.

Boetzer M, Henkel C, Jansen H, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics,* **27**: 578-579.

Broman K, Wu H, Sen S, Churchill G. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics,* **19**: 889-890.

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

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W. 2013. Stacks: an analysis tool set for population genomics. *Molecular Ecology,* **22**: 3124-3140.

Chen Q, Ma Y, Yang Y, Chen Z, Liao R, Xie X, Wang Z, He P, Tu Y, Zhang X, Yang C, Yang H, Yu F, Zheng Y, Zhang Z, Wang Q, Pan Y. 2013. Genotyping by Genome Reducing and Sequencing for Outbred Animals. *Plos One,* **8**.

Chia J, Song C, Bradbury P, Costich D, de Leon N, Doebley J, Elshire R, Gaut B, Geller L, Glaubitz J, Gore M, Guill K, Holland J, Hufford M, Lai J, Li M, Liu X, Lu Y, McCombie R, Nelson R, Poland J, Prasanna B, Pyhajarvi T, Rong T, Sekhon R, Sun Q, Tenaillon M, Tian F, Wang J, Xu X, Zhang Z, Kaeppler S, Ross-Ibarra J, McMullen M, Buckler E, Zhang G, Xu Y, Ware D. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics,* **44**: 803-U238.

Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, Shaw P, Bayer M, Thomas W, Marshall D, Hedley P, Tondelli A, Pecchioni N, Francia E, Korzun V, Walther A, Waugh R. 2012. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nature Genetics* **44:** 1388-1392

Connolly V, Wrightturner R. 1984. Induction of cytoplasmic male-sterility into ryegrass (Lolium perenne). *Theoretical and Applied Genetics,* **68**: 449-453.

Davey J, Blaxter M. 2011. RADSeq: next-generation population genetics (vol 9, pg 416, 2010). *Briefings in Functional Genomics,* **10**: 108-108.

Davey J, Hohenlohe P, Etter P, Boone J, Catchen J, Blaxter M. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics,* **12**: 499-510.

Deokar A, Ramsay L, Sharpe A, Diapari M, Sindhu A, Bett K, Warkentin T, Tar'an B. 2014. Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. *BMC Genomics,* **15**.

Doyle JJ. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem bull,* **19**: 11-15.

Elshire R, Glaubitz J, Sun Q, Poland J, Kawamoto K, Buckler E, Mitchell S. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *Plos One,* **6**.

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

Foito A, Hackett C, Byrne S, Stewart D, Barth S. 2015. Quantitative trait loci analysis to study the genetic regulation of non-polar metabolites in perennial ryegrass. *Metabolomics,* **11**: 412-424.

Higgins J, Bailey P, Laurie D. 2010. Comparative Genomics of Flowering Time Pathways Using Brachypodium distachyon as a Model for the Temperate Grasses. *Plos One,* **5**.

Joshi NA, Fass JN. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Koboldt D, Zhang Q, Larson D, Shen D, McLellan M, Lin L, Miller C, Mardis E, Ding L, Wilson R. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research,* **22**: 568-576.

Kopecky D, Havrankova M, Loureiro J, Castro S, Lukaszewski A, Bartos J, Kopecka J, Dolezel J. 2010. Physical Distribution of Homoeologous Recombination in Individual Chromosomes of Festuca pratensis in Lolium multiflorum. *Cytogenetic and Genome Research,* **129**: 162-172.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* **10**.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* **25**: 2078-2079.

Lu F, Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer8 I, Barad8 O, Buckler ES. 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *NATURE COMMUNICATIONS*.

Manzanares C. 2013. *Genetics of self-incompatibility in perennial ryegrass (Lolium perenne L.),* Ph.D, University of Birmingham, England.

Mascher M, Wu S, St Amand P, Stein N, Poland J. 2013. Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. *Plos One,* **8**.

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

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

Rostoks N, Park Y, Ramakrishna W, Ma J, Druka A, Shiloff B, SanMiguel P, Jiang Z, Brueggeman R, Sandhu D, Gill K, Bennetzen J, Kleinhofs A. 2002. Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. *Functional & Integrative Genomics,* **2**: 51-59.

Russell J, Hackett C, Hedley P, Liu H, Milne L, Bayer M, Marshall D, Jorgensen L, Gordon S, Brennan R. 2014. The use of genotyping by sequencing in blackcurrant (Ribes nigrum): developing high-resolution linkage maps in species without reference genome sequences. *Molecular Breeding,* **33**: 835-849.

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

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi U, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell,* **20**: 11-24.

Skinner M, Uzilov A, Stein L, Mungall C, Holmes I. 2009. JBrowse: A next-generation genome browser. *Genome Research,* **19**: 1630-1638.

Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics,* **6**.

Smit AFA, Hubley R, Green P. 2010. RepeatMasker Open-3.0.http://www.repeatmasker.org.

Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Lorieux M, Ahmadi N, McCouch S. 2013. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theoretical and Applied Genetics,* **126**: 2699-2716.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics,* **19**: II215-II225.

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

Tomaszewski C, Byrne S, Foito A, Kildea S, Kopecky D, Dolezel J, Heslop-Harrison J, Stewart D, Barth S. 2012. Genetic linkage mapping in an F2 perennial ryegrass population using DArT markers. *Plant Breeding,* **131**: 345-349.

Van Ooijen J. 2011. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research,* **93**: 343-349.

van Orsouw N, Hogers R, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk M. 2007. Complexity Reduction of Polymorphic Sequences (CRoPS (TM)): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *Plos One,* **2**.

van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan G, Caromel B, Ghareeb B, Isidore E, de Jong W, van Koert P, Lefebvre V, Milbourne D, Ritter E, van der Voort J, Rousselle-Bourgeois F, van Vliet J, Waugh R, Visser R, Bakker J, van Eck H. 2006. Construction of a 10,000-marker ultradense genetic recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics,* **173**: 1075-1087.

Ward J, Bhangoo J, Fernandez-Fernandez F, Moore P, Swanson J, Viola R, Velasco R, Bassil N, Weber C, Sargent D. 2013. Saturated linkage map construction in Rubus idaeus using genotyping by sequencing and genome-independent imputation. *BMC Genomics,* **14**.

Wilson A, Wickett N, Grabowski P, Fant J, Borevitz J, Mueller G. 2015. Examining the efficacy of a genotyping-by-sequencing technique for population genetic analysis of the mushroom Laccaria bicolor and evaluating whether a reference genome is necessary to assess homology. *Mycologia,* **107**: 217-226.

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

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *Plos One,* **7**.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti S, Patil V, Skryabin K, Kuznetsov B, Ravin N, Kolganova T, Beletsky A, Mardanov A, Di Genova A, Bolser D, Martin D, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop G, Sagredo B, Mejia N, Zagorski W, Gromadka R, Gawor J, Szczesny P, Huang S, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu D, Bonierbale M, Ghislain M, Herrera M, Giuliano G, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold S, Barreiro L, Massa G, Diambra L, Whitty B, Vaillancourt B, Lin H, Massa A, Geoffroy M, Lundback S, DellaPenna D, Buell C, Sharma S, Marshall D, Waugh R, Bryan G, Destefanis M, Nagy I, Milbourne D, Thomson S, Fiers M, Jacobs J, Nielsen K, Sonderkaer M, Iovene M, Torres G, Jiang J, Veilleux R, Bachem C, de Boer J,

Borm T, Kloosterman B, van Eck H, Datema E, Hekkert B, Goverse A, van Ham R, Visser R, Consortiu PGS. 2011. Genome sequence and analysis of the tuber crop potato. *Nature,* **475**: 189-U94.

Figure Legends

Figure 1. Distribution of contig and scaffold lengths

Scaffolds and contigs are grouped according to size range, from under 500 bp to over 280,000 bp, to indicate the proportion of sequence held by scaffolds/contigs within each size range group. Numbers of scaffolds/contigs in each size range group are plotted on the left hand vertical axis, with the corresponding base pair length for each group plotted on the right hand vertical axis. **(a)** Scaffold/contig counts and base pair lengths for each group are shown along with **(b)** cumulative running totals.

Figure 2 : Distribution of transcripts and gene models, and associated scaffolds

Scaffolds are grouped according to size range, from under 500 bp to over 280,000 bp, to indicate the proportion of transcripts from both Ruttink and Farrell sets and gene models from Byrne et al. contained by scaffolds within each size range group. Numbers of scaffolds in each size range group are plotted on the left hand vertical axis, with the number of associated transcripts for each group plotted on the right hand vertical axis. **(a)** Scaffold counts with numbers of transcripts and gene models for each group are shown along with **(b)** cumulative running totals.

Figure 3: Examples of graphical genotypes of chromosomes 2 and 3 a) before and b) after genotype error correction.

The x-axis consists of genotype calls of individuals and the y-axis consists of markers ordered by chromosomal map position. The blue colour represents the allele from the paternal grandparent, pink from the maternal grandparent and yellow for the heterozygous state.

Figure 4: Distribution of segregation distortion across the chromosomes. A line on the chromosome represents the framework marker on the map. Red indicates loci with segregation distortion (P.value <0.05) and green represents non-distorted loci. The highlighted loci with map position on the left and locus name on the right represents the marker location of previously published markers by Anhalt et al. (2008) on the current linkage map.

Figure 5: Distribution of marker density across the chromosome. The X-axis represents 5cM map interval and the Y-axis represents the number of GBS markers present in the interval.

Figure 6: Graphical genotypes of parents along with a subset of individuals from chromosome1. The first column represents the F1 parent, the second column the paternal grandparent. Blue represents alleles from the paternal grandparent, pink represents alleles from the maternal grandparent and yellow indicates a heterozygous allelic state.

Table 1: Summary statistics for *Lolium perenne* genome draft assembly

	Contigs	Scaffolds
No. sequences	624,485	424,750
Max. size	94,591	282,695
Mean size	1,338	2,618
N50	3,790	25,212
Total length	835,987,474	1,112,005,533
%GC	44.21	44.16
%N	3.08	27.19
No. sequences >= N50	54,586	10,877
% sequences >= N50	8.74	2.56
No. sequences < 500 bp	320,891	254,591
% sequences < 500 bp	51.38	59.94
No. bases in sequences < 500 bp	104,144,782	81,476,618
% bases in sequences < 500 bp	12.46	7.33

Table 2: Statistics of identified SNP markers (Number and proportion of transition versus transversion type markers)

Type	Type of variation	Number	Proportion of type
Transition	C/T	2832	31
Transition	A/G	2681	29
Transversion	C/G	1334	15
Transversion	A/C	875	10
Transversion	G/T	870	10
Transversion	A/T	535	6

Table 3: Summary of genetic map (linkage group (LG), total number of markers, number of SNP, PA markers, number of SNP bins, map length, number and size of scaffolds anchored for each linkage group).

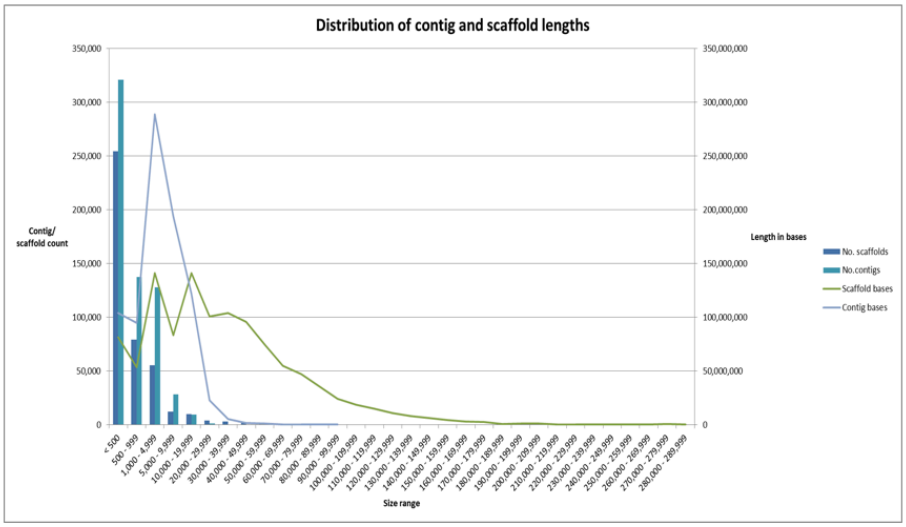
LG	Total no. of markers	No of SNP markers	No of PA markers	No of SNP bins	Map length (cM)	No of unique scaffolds anchored	No of bases anchored
1	1319	418	901	240	124.5	573	22,689,863
2	1502	466	1036	287	139.4	665	28,082,237
3	1986	563	1423	317	153.9	886	36,169,696
4	1839	557	1282	331	196	853	34,568,556
5	1421	370	1051	233	119.8	617	25,520,777
6	845	269	576	179	89.3	373	15,266,325
7	1440	449	991	278	129.7	624	25,795,316
Total	10352	3092	7260	1865	952.6	4591	188,092,770

Table 4: Summary of the size distribution of anchored scaffolds in the map of the F2 Biomass population

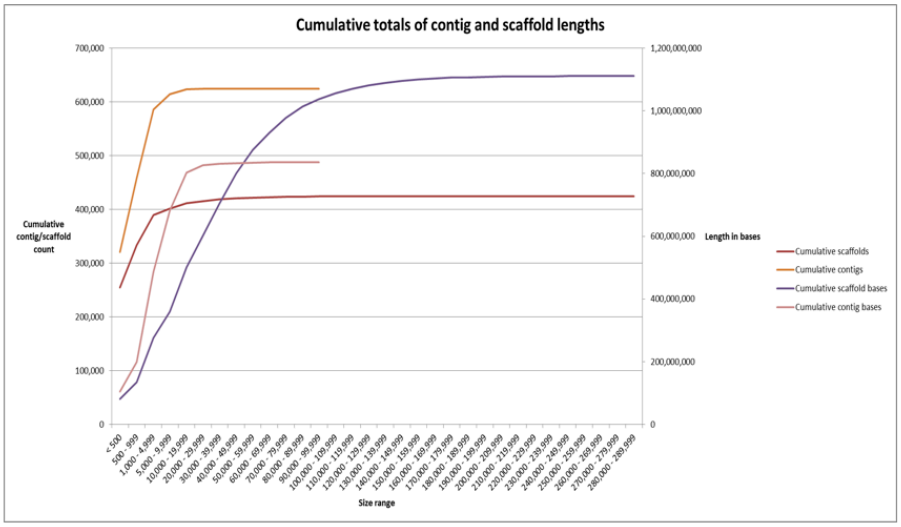
Scaffold size range	Number of anchored scaffolds
<=500	75
500 to 1000	63
1k to 5k	223
5k to 10k	296
10k to 50k	2556
50k to 100k	1241
100k to 500k	313
Total	4767

Figure 1

a)



b)

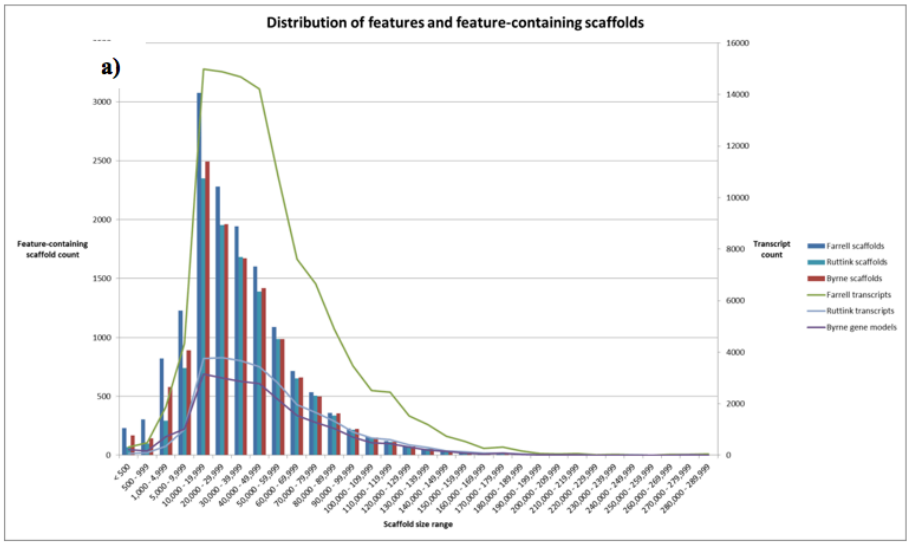


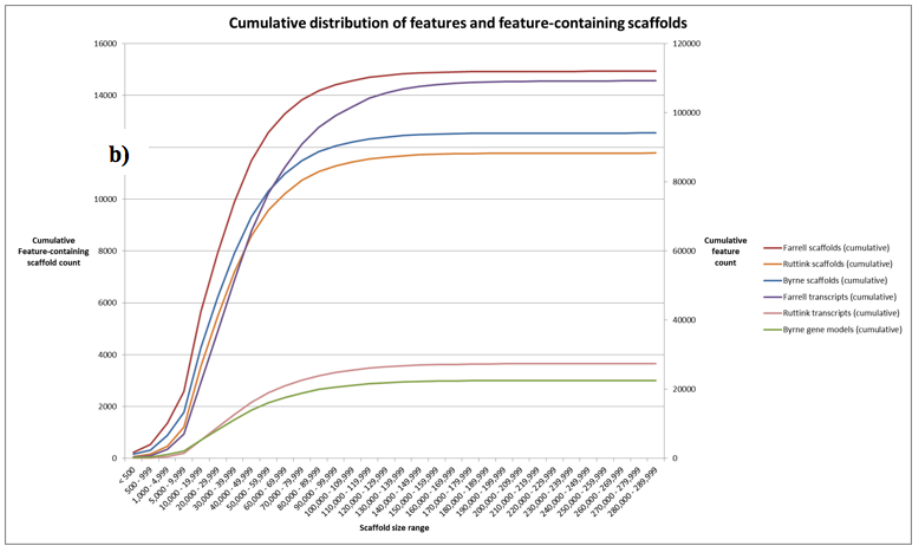
1158

1159

1160

Figure 2



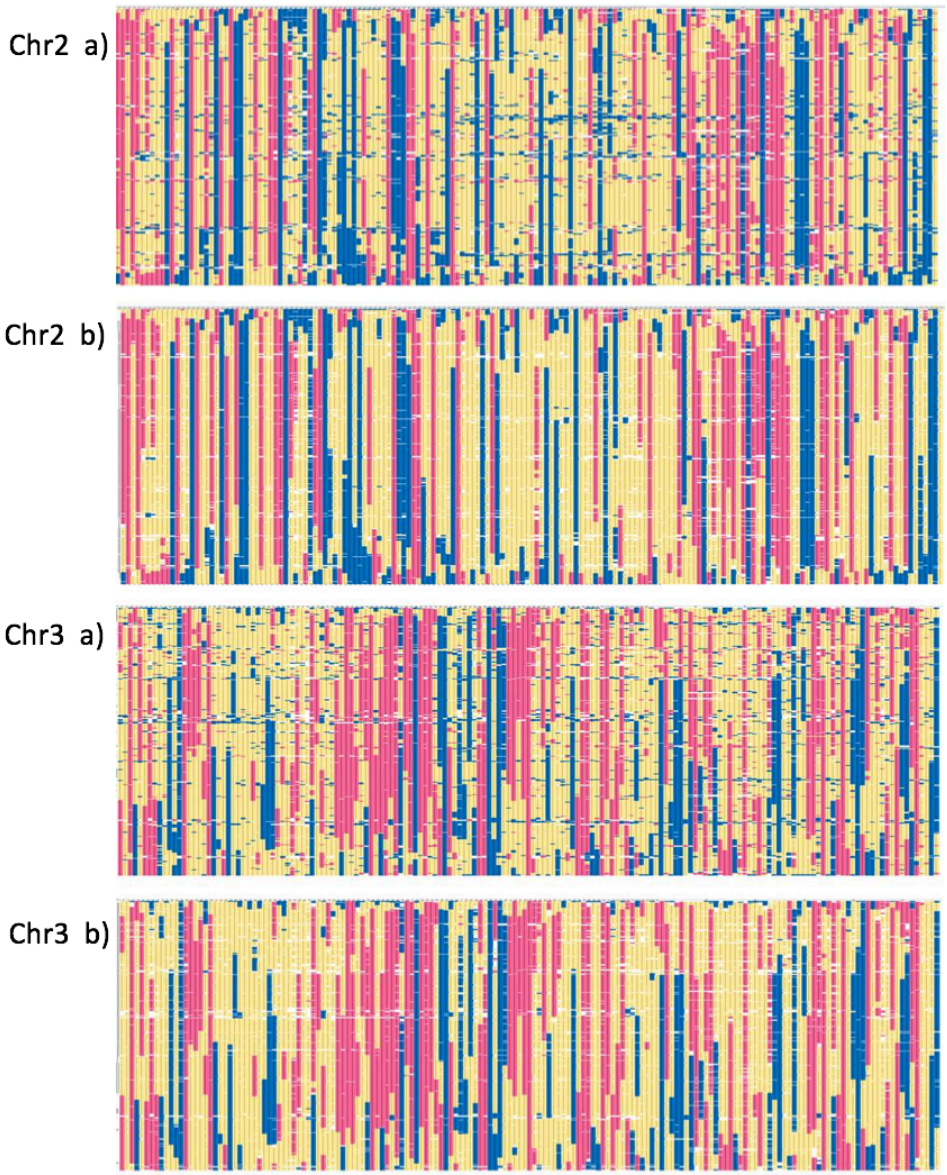


1161

1162

1163

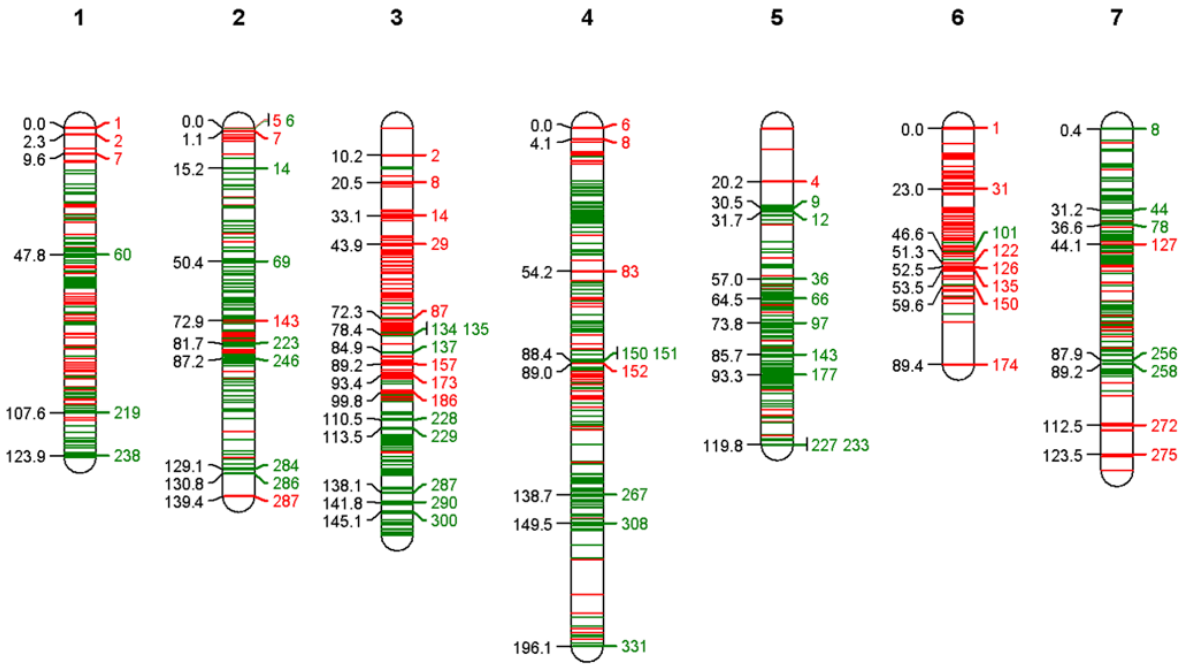
Figure 3



1164

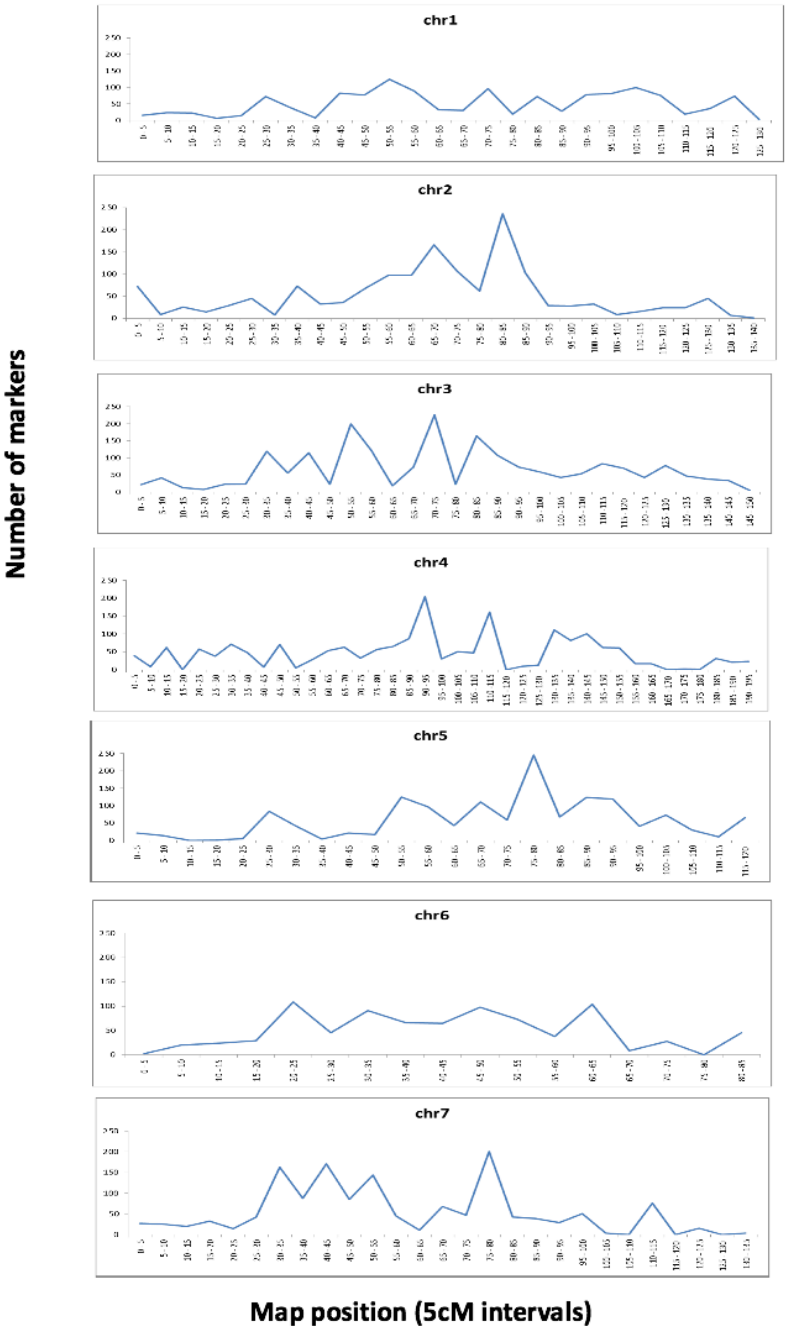
1165

Figure 4


1
2
3
4
5
6
7
0.0 1
2.3 2
9.6 7
47.8 60
107.6 219
123.9 238
0.0 5 6
1.1 7
15.2 14
50.4 69
72.9 143
81.7 223
87.2 246
129.1 284
130.8 286
139.4 287
10.2 2
20.5 8
33.1 14
43.9 29
72.3 87
78.4 134 135
84.9 137
89.2 157
93.4 173
99.8 186
110.5 228
113.5 229
138.1 287
141.8 290
145.1 300
0.0 6
4.1 8
54.2 83
88.4 150 151
89.0 152
138.7 267
149.5 308
196.1 331
20.2 4
30.5 9
31.7 12
57.0 36
64.5 66
73.8 97
85.7 143
93.3 177
119.8 227 233
0.0 1
23.0 31
46.6 101
51.3 122
52.5 126
53.5 135
59.6 150
89.4 174
0.4 8
31.2 44
36.6 78
44.1 127
87.9 256
89.2 258
112.5 272
123.5 275

1171

Figure 5



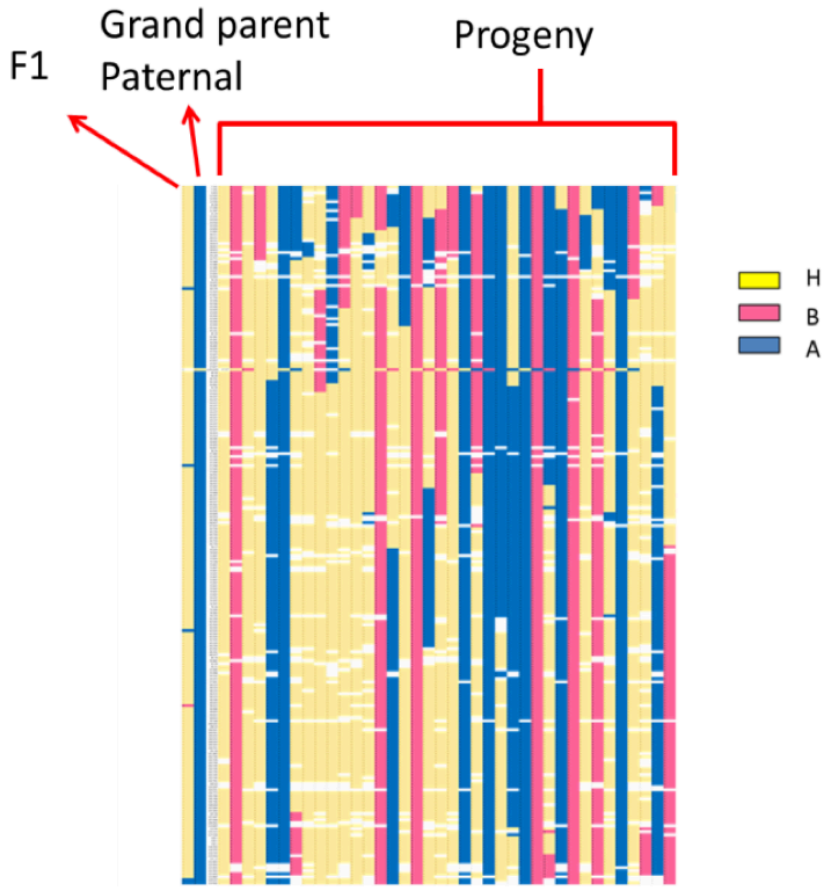
1172

1173

1174

1175

Figure 6



1176

1177

1178