DNA Research, 2018, 0(0), 1–12
doi: 10.1093/dnares/dsy033
Full Paper

OXFORD

Full Paper

Overcoming challenges in variant calling: exploring sequence diversity in candidate genes for plant development in perennial ryegrass (*Lolium perenne*)

Elisabeth Veeckman [1,2,3], Sabine Van Glabeke[1], Annelies Haegeman[1], Hilde Muylle[1], Frederik R.D. van Parijs[1], Stephen L. Byrne[4], Torben Asp[5], Bruno Studer[6], Antje Rohde[1†], Isabel Roldán-Ruiz[1,3], Klaas Vandepoele[2,3,7], and Tom Ruttink[1,2]*

[1]ILVO, Plant Sciences Unit, B-9090 Melle, Belgium, [2]Bioinformatics Institute Ghent, Ghent University, B-9052 Ghent, Belgium, [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium, [4]Teagasc, Crop Science Department, Carlow R93 XE12, Ireland, [5]Department of Molecular Biology and Genetics, Faculty of Science and Technology, Research Center Flakkebjerg Aarhus University, DK-4200 Slagelse, Denmark, [6]Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, CH-8092 Zurich, Switzerland, and [7]Center for Plant Systems Biology, VIB, B-9052 Ghent, Belgium

*To whom correspondence should be addressed. Tel. +32 9 272 28 78. Fax. +32 9 272 29 01. Email: tom.ruttink@ilvo.vlaanderen.be

[†]Present address: BASF Agricultural Solutions, Technologiepark 38, B-9052 Ghent, Belgium.

Edited by Dr Satoshi Tabata

Received 26 June 2018; Editorial decision 4 September 2018; Accepted 6 September 2018

Abstract

Revealing DNA sequence variation within the *Lolium perenne* genepool is important for genetic analysis and development of breeding applications. We reviewed current literature on plant development to select candidate genes in pathways that control agronomic traits, and identified 503 orthologues in *L. perenne*. Using targeted resequencing, we constructed a comprehensive catalogue of genomic variation for a *L. perenne* germplasm collection of 736 genotypes derived from current cultivars, breeding material and wild accessions. To overcome challenges of variant calling in heterogeneous outbreeding species, we used two complementary strategies to explore sequence diversity. First, four variant calling pipelines were integrated with the VariantMetaCaller to reach maximal sensitivity. Additional multiplex amplicon sequencing was used to empirically estimate an appropriate precision threshold. Second, a *de novo* assembly strategy was used to reconstruct divergent alleles for each gene. The advantage of this approach was illustrated by discovery of 28 novel alleles of *LpSDUF247*, a polymorphic gene co-segregating with the S-locus of the grass self-incompatibility system. Our approach is applicable to other genetically diverse outbreeding species. The resulting collection of functionally annotated variants can be mined for variants causing phenotypic variation, either through genetic association studies, or by selecting carriers of rare defective alleles for physiological analyses.

Key words: genomic diversity, targeted resequencing, variant calling, allele reconstruction, natural variation

1. Introduction

Perennial ryegrass (*Lolium perenne* L.) is one of the most widely cultivated grass species in Europe. It is of interest for grazing, hay and silage production as it has a long growing season, and relatively high yield and nutritive value. Because of its outbreeding nature, individual plants are highly heterozygous and the diploid perennial ryegrass genome is highly heterogeneous both within and across breeding populations and wild accessions. As genomic variation forms the foundation of phenotypic variation, revealing DNA sequence variation within the genepool is important for genetic analysis and development of breeding applications.[1]

Several studies used a candidate gene-based approach to associate sequence polymorphisms with phenotypic variation. Examples include the association of *Late embryogenesis abundant 3* (*LEA3*) with drought tolerance,[2] *Brassinosteroid insensitive 1* (*BRI1*) with shoot morphology,[3] *Gibberellic acid insensitive* (*GAI*) with organ growth,[4] *Heading date 1* (*HD1*) with carbohydrate content[5] and *Flowering locus T* (*FT*) with flowering time.[5–7] While these studies show the power of testing gene–trait associations, the limited number of genes per study was mostly due to the high cost of genotyping at the time. However, this approach is not amenable to study complex traits related to plant development and phenology, which are typically regulated by the interaction of many genes. Therefore, we need versatile and cost-efficient methods to characterize the genetic variation in parallel for hundreds of candidate genes and hundreds of genotypes. This enables breeders to perform higher resolution screening of genetic diversity in their material and link genotypic and phenotypic variation.

Single-nucleotide polymorphisms (SNPs) are the most prevalent type of genomic variation and are convenient molecular markers. Two complementary SNP genotyping arrays are available for high-throughput screening in perennial ryegrass.[8,9] These arrays target SNPs in genic regions, but do not allow discovery of new sequence variants. In contrast, genotyping-by-sequencing (GBS) allows for simultaneous discovery of genome-wide SNPs and genotyping of a large number of individuals or pools, thereby avoiding ascertainment bias. Therefore, GBS has broad applications in plant breeding and genetics studies, including linkage maps, genome-wide association studies, genomic selection and genomic diversity studies.[10] Only short fragments (range about 100–300 bp) are sequenced and there is no *a priori* control over which genes are tagged. In combination with a local short linkage disequilibrium that is typical for an outbreeding species as *L. perenne*,[3,11] it can be very difficult to identify SNPs that are causal for the phenotype of interest.

The large size of the *L. perenne* genome (2 Gbp) and high repetitive sequence content (76%)[12] currently precludes whole genome sequencing at sufficient depth in hundreds of accessions as has been done in e.g. Arabidopsis, rice and soybean. To study trait genetics in forage and turf grasses, we identified hundreds of candidate genes in genetic pathways that control plant development and quality traits, and analysed their genome sequence using probe capture enrichment for targeted resequencing in a large germplasm collection of 736 genotypes. We specifically focussed on genes involved in pathways related to interesting agronomic traits, such as plant growth and architecture (important for biomass yield), development and transition to flowering (important for seasonal control of growth), cell wall biogenesis (important for digestibility) and phytohormone biosynthesis, signalling and response [including abscisic acid (ABA), auxin, brassinosteroids, cytokinins, ethylene, gibberellic acid and strigolactones]. Identifying sequence variants in these genes provides insights in the range of naturally occurring genomic diversity that can be expected in gene-rich regions of the genome. The variants can be used as markers for association genetics studies (as previously described for *LEA3*, *BRI1*, *GAI*, *HD1* and *FT3*), and/or to identify alleles with altered amino acid sequence, mRNA splicing or mRNA stability, hence altered gene function or regulation possibly resulting in an altered physiology and thereby affecting the phenotype.

Multiple bioinformatics methods are available to identify sequence variants using next-generation sequencing (NGS) data, but defining a complete and reliable variant set remains difficult. *De novo* discovery of genomic polymorphisms commonly relies on mapping reads to a single reference genome sequence. Although the GATK best practices are the most commonly used variant calling (VC) pipeline, there is no single best VC pipeline available with both good sensitivity and precision. Moreover, there is low concordance between VC pipelines, even with the same input data.[13–16] In addition, each VC pipeline returns different variant annotations that can be used for quality filtering. Choosing the appropriate filtering criteria and thresholds [for instance, minimum read depth (RD)] is not straightforward as NGS data typically has a non-uniform distribution of coverage[17] and estimated quality values may be dataset dependent so that optimal settings need to be calibrated for each dataset. The high density of sequence polymorphisms in the germplasm collection with respect to the *L. perenne* genome sequence could also hamper variant identification. More divergent alleles could contain the most interesting genomic variation, but are also the most difficult to detect as reads that are highly divergent from the reference genome may fail to map if the parameters for short read alignment are too stringent.[18] Hence, if capture and/or mapping efficiency of highly divergent sequences precludes their detection, it is to be expected that routine workflows of mapping and VC lead to an underestimation of the genetic diversity at highly divergent regions, a known problem in genome resequencing studies.[19]

Here, we present the identification and annotation of 503 *L. perenne* orthologues of known genes that regulate plant growth and development. These genes were resequenced in a germplasm collection of 736 genotypes to describe the genomic variation in *L. perenne*. Two complementary strategies were used to obtain a reliable and complete catalogue of genomic variation. First, four VC pipelines were compared and automatically integrated to reach maximal sensitivity. The influence of mapping algorithms was assessed and hard filtering was compared with precision-based filtering to reach sufficient specificity. Additionally, an alternative strategy consisting of *de novo* assembly followed by overlap-layout-consensus (OLC) clustering was used to circumvent read mapping bias and to construct alternative alleles for each gene. This reference independent allele reconstruction is particularly important for gene families with highly divergent alleles. We demonstrated the benefit of this approach for *LpSDUF247* and identified 28 novel alleles that were not detected using traditional VC pipelines. This approach is broadly applicable to other highly heterozygous outbreeding species. Finally, we used all this information to create a comprehensive catalogue of functionally annotated genetic variation across many pathways that control growth, development and agricultural traits.

2. Materials and methods

2.1 Candidate gene identification and manual curation

Gene families of *A. thaliana* candidate genes were identified using the comparative genomics platform PLAZA 3.0 Monocots.[20] *Brachypodium distachyon* family members were used to identify homologous loci in the draft genome sequence of *L. perenne*[12] using BLASTx analysis (*E*-value 10e-5). At each *L. perenne* locus, predicted protein sequences were added to the corresponding PLAZA 3.0 Monocots gene family. Using protein sequences of all gene family members of *A. thaliana*, *B. distachyon* and *L. perenne*, a phylogenetic tree was built with MUSCLE (v3.8.31)[21] and PhyML[22] using default settings, and for each *A. thaliana* candidate gene the closest orthologous *L. perenne* gene was selected for further manual curation. The *L. perenne* gene models were evaluated using multiple protein sequence alignments with MUSCLE using orthologous proteins from *B. distachyon*, *O. sativa*, *Z. mays* and *S. bicolor* according to PLAZA 3.0 Monocots. Additional RNA-seq data[23–25] mapped with TopHat (v2.0.13)[26] using default settings, was used to refine gene models and delineate untranslated regions, or to design a new gene model if required (Supplementary Data S1).

2.2 Probe design, library construction and sequencing

The coding strand of each of the 503 target regions (gene model and 1,000 bp upstream promoter region) was tiled with 120 bp probes, starting every 40 bp using OligoTiler.[27] Probes showing high sequence similarity to non-targets, other probes, repetitive sequences, mitochondrial or chloroplast sequences, or with extreme GC content (<25% or >65%) were removed. Finally, 57,693 SureSelect probes of 120 bp (Agilent) were retained, covering 2.3 Mb of the intended 2.8 Mb target region, at around 3× tiling.

Genomic DNA was extracted from freeze-dried leaf material from 736 *L. perenne* genotypes representing current cultivars, breeding material and wild accessions using the cetyltrimethylammonium bromide (CTAB) method.[28] DNA concentration was measured using the Quantus double-stranded DNA assay (Promega, Madison, WI, USA). For each genotype, an indexed shotgun sequencing library was prepared from 100 ng DNA by (i) Adaptive Focused Acoustic fragmentation on a Covaris S2 instrument (Covaris, Inc.), (ii) adapter ligation and (iii) magnetic bead purification using an adapted protocol of Uitdewilligen et al.[29] The libraries were pooled without normalization into eight pools, each containing 96 libraries of individual genotypes. Each pool was used for a probe capture hybridization reaction according to the SureSelect protocol (Agilent SureSelectXT2 Target Enrichment for Illumina Paired-End Sequencing Library Protocol, v. 1.0). After PCR amplification of purified enriched pooled libraries, each pool of 96 libraries was sequenced on one lane of a HiSeq2000 instrument using 2 × 91 PE sequencing (BGI, Shenzhen, China). The raw data is available in the NCBI Sequence Read Archive (BioProject PRJNA434356, Accessions SRR6812717–SRR6813075).

2.3 Read mapping and variant calling

Raw reads were trimmed and quality filtered by Trimmomatic (v0.32)[30] and mapped onto the draft perennial ryegrass genome sequence[12] with default settings of BWA-MEM (version 0.7.8-r455)[31] and GSNAP (version 2016-09-23).[32] Duplicate reads were marked using Picard-tools (release 1.113). Local realignment around indels was performed according to the best practices workflow of the Genome Analysis Toolkit (GATK) (v.3.7).[36,37] RD and coverage were calculated on the resulting BAM files using BEDTools (v2.25.0).[33]

Four different VC pipelines were used: SAMtools (version 1.2-115-gb8ff342),[34] Freebayes (v1.0.2-2-g7ceb532),[35] GATK Unified Genotyper (GATK UG) and GATK HaplotypeCaller (GATK HC).[36,37] Multi-allelic variants were removed using VCFtools (v0.1.14).[38] For hard filtering, a custom Python script was used to remove variant positions and genotype calls with a RD lower than 6 and a genotype quality (GQ) lower than 30. SNPs and indels were automatically integrated by the VMC (v1.0),[39] in 10 and 2 partitions, respectively. The estimated precision (EP) was calculated using a custom Python script based on the formulas given in Gézsi et al.[39] The concordance of SNP and indel sets identified by four VC pipelines was determined using information in the INFO field of the VCF file returned by VMC and visually represented using Upset,[40] before and after precision-based filtering (EP > 80%). Functional effects of sequence variants were predicted with SnpEff (version 4.3T).[41] To validate consistency of genotype calls in an F1 segregating population, Mendelian inheritance errors (MIE) were defined after precision-based filtering (EP > 80%) using PLINK (v1.90b2t), for two parents and their F1 progeny of 29 individuals. Variants with a missing genotype call in either one of the parents were excluded from analysis, as were MIEs derived from a missing genotype call in one of the 29 F1 progeny.

2.4 Hi-Plex amplicon sequencing

To generate an independent variant set, 78 genotypes were selected for resequencing of 171 amplicon regions of 80–140 bp. Of these, 147 amplicons overlap with 28 candidate genes. Primers were designed with Primer3[42] and divided into two highly multiplex (Hi-Plex) PCR-reactions according to their amplification efficiency (Supplementary Data S4). DNA was extracted using the CTAB method[28] and DNA concentration was measured using the Quantus double-stranded DNA assay (Promega, Madison, WI, USA). Per sample, the final DNA concentration was adjusted to 40 ng/µl and the amplicons were PCR-amplified while adding sample specific indices. Libraries were prepared using the KAPA Hyper Prep PCR-free Kit according to manufacturer directions (Kapa Biosystems, USA). Hi-Plex amplification reactions and library preparations were done by Floodlight Genomics LLC (Knoxville, TN, USA). The libraries were sequenced with 2 × 150 PE on a HiSeq3000 instrument (OMRF, Oklahoma City, OK, USA). Paired-end reads were merged with PEAR (v0.9.8)[43] and adapter sequences were removed. The read data is available in the NCBI Sequence Read Archive (BioProject PRJNA437219, Accessions SRR6813540–SRR6813585). BWA-MEM was used for read mapping, and VC was done by running the four VC pipelines. Bi-allelic variants were extracted using VCFtools and combined by VMC and the EP was calculated as described above.

2.5 Identification of divergent alleles of *LpSDUF247*

Per genotype, all reads were used for De Bruijn Graph assembly without scaffolding (CLC Genomics Workbench 9.5.3, https://www.qiagenbioinformatics.com (date last accessed 14 september 2018)). Contigs of at least 200 bp were retained and mapped onto the reference genome with BWA-MEM using default parameters, to group all contigs of the 736 genotypes per candidate gene. Per candidate gene, sequences of overlapping allelic fragments were extracted from the

BAM files using BEDtools and clustered with the OLC assembler CAP3 (version date 02/10/15).[44] Singlet sequences returned by CAP3 were removed from further analysis. All resulting alleles are assigned to their respective candidate gene and are available as (Supplementary Data S5), allowing the reader to repeat the analyses described below for any other candidate gene.

One of the candidate genes of the 503 gene set, *LpSDUF247*, is known to be highly polymorphic and was selected to demonstrate in-depth reconstruction of divergent alleles. The 34 contigs of *LpSDUF247* were aligned using MUSCLE and six highly similar sequences (>98% identity) were removed. The reference gene model of *LpSDUF247* was projected onto the contigs using GenomeThreader (v 1.6.6)[45] to identify CDS regions (Supplementary Data S6) and corresponding protein sequences (Supplementary Data S7).

All *B. distachyon* members of the DUF247 gene family (HOM03M000101) were used in a tBLASTn search against the perennial ryegrass genome sequence, and 25 *LpDUF247* genes were identified and manually annotated (Supplementary Data S8). After multiple sequence alignment of all 25 LpDUF247 protein sequences with *B. distachyon* and *H. vulgare* gene family members using MUSCLE, a phylogenetic tree was built with PhyML using 100 rounds of bootstrapping (Supplementary Fig. S5). Similarly, a phylogenetic tree was built using the reference protein sequences of LpDUF247-01, LpSDUF247, LpDUF247-03 and LpDUF247-04, the protein sequences of the LpSDUF247 alleles, and five LpSDUF247-02 alleles identified by Manzanares et al.[46] (Supplementary Fig. S6).

The 28 novel alleles were added to the reference genome sequence and read mapping was repeated for all 736 genotypes onto this multi-allelic reference genome. A matrix was created with the average RD per *LpSDUF247* allele per genotype using BEDtools. This matrix was normalized per genotype, by dividing the RD *per LpSDUF247* allele by the sum of RDs across *all LpSDUF247* alleles, to identify alleles with the highest relative RD for each genotype while correcting for differences in library size and capture efficiency across the set of 736 samples.

3. Results and discussion

3.1 Identification, classification and curation of target genes

To identify *L. perenne* genes putatively involved in the regulation of plant growth and development, plant architecture, induction of flowering, cell wall biogenesis and phytohormone biosynthesis, signalling and response, we first searched the literature for *Arabidopsis thaliana* genes with a well-defined molecular and physiological function (Supplementary Table S1). Next, the corresponding 174 gene families were identified with the comparative genomics platform PLAZA 3.0 Monocots.[20] For each of the *A. thaliana* candidate genes, a comprehensive list of orthologous loci in the draft genome sequence of *L. perenne*[12] was delineated. A phylogenetic tree was built for *A. thaliana* and *B. distachyon* gene family members of the 174 PLAZA gene families, to select the closest orthologous *L. perenne* sequences of the candidate genes. When no clear one-to-one orthologous pairs were found due to lineage-specific gene duplication or gene loss events, the best two or three *L. perenne* loci were selected from the respective clades. The final selection contained 503 *L. perenne* candidate genes (Supplementary Table S1). For 407 of these loci, an annotated gene model was available.[12] For the other 96 loci, a gene model needed to be annotated *ab initio*, in line with previous observations that the annotated gene space of *L. perenne* is 76% complete.[47] The available gene models were evaluated using multiple protein sequence alignments with all their monocot gene family members according to PLAZA 3.0 Monocots. In addition, mapped RNA-seq data[23–25] was used to refine gene models and delineate untranslated regions. Taken together, manual curation of 503 gene models (Supplementary Data S1), showed that previously available gene models[12] were correct for 272 loci (54%) and needed small adaptations for 135 loci (27%). A completely new gene model was annotated at 96 loci (19%) using RNA-seq data. The length of the protein sequences corresponds well to that of their closest *B. distachyon* orthologs (Supplementary Fig. S1), showing that the 503 manually curated *L. perenne* gene models are of high quality (HQ). This was required to delineate regions for probe design and to correctly position variants relative to the reading frame in the CDS to functionally interpret the consequences of sequence polymorphism in the genic regions. Finally, the 503 candidate genes were assigned to biological processes based on the known function of their *A. thaliana* orthologs (Table 1 and Supplementary Table S1). This HQ gene set can also be used to train and validate gene prediction algorithms to improve genome-wide gene annotation.

3.2 Design and efficacy of targeted resequencing by probe capture enrichment

For each candidate gene, a target region was delineated spanning the curated gene model and an additional 1,000 bp upstream promoter region, as described previously.[48] Probes were designed for a total length of 2.3 Mbp, corresponding to a coverage of 85% of each target region on average, as probes targeting repetitive regions were excluded (Supplementary Data S2). Targeted resequencing of 736 genotypes resulted in 3.2 million reads per genotype on average (range 20 thousand–31 million). After duplicate read removal, a mean of 1.9 million reads was retained per genotype, corresponding to a mean RD of 80× per position within the target regions. For VC analysis in heterozygous diploid species, a coverage of at least 6–10× is desirable to avoid false negative heterozygous calls.[49] Saturation curves show a non-linear relationship between number of reads per library and target region coverage at a given RD threshold, as expected for probe capture enriched shotgun sequencing libraries (Fig. 1). At least 550,000 uniquely mapped reads per genotype were required to reach the probe region coverage plateau at 95% for RD $\geq$ 1. Further increasing the number of reads per sample did not substantially increase probe region coverage (see Ruttink et al.[48]). The probe region coverage was slightly lower at higher RD thresholds (89% for RD $\geq$ 6 and 85% for RD $\geq$ 10 (boxplots in Fig. 1).

3.3 Optimization of variant calling pipelines to compile a reliable catalogue of sequence variation

To obtain a complete and reliable variant set, we selected two mapping algorithms and four frequently used multi-sample VC pipelines to reach maximal sensitivity. Read mapping algorithm BWA-MEM[31] was compared to GSNAP,[32] which is able to handle short and long insertions and deletions. Two alignment based VC pipelines were selected for their strength in SNP calling[50]: GATK Unified Genotyper (GATK UG)[36,37] and SAMtools.[34] Additionally, two haplotype-based VC pipelines were chosen for their strength in indel detection: GATK HaplotypeCaller (GATK HC) and Freebayes.[35] We compared the resulting variant sets and assessed the performance of hard filtering to improve the precision of variant sets. Finally, the four individual variant sets and corresponding variant quality annotations were merged by the VMC,[39] allowing for precision-based filtering as an alternative to hard filtering.

Table 1. Assignment of 503 candidate genes to pathways and distribution of high impact mutations per pathway

Pathway	Gene families	# candidate genes	Stop gain	Splice site	Frame shift
Plant development and architecture					
Development	BCH1, BRIZ, CBP80, DRM1, HB13, HYL1, ING2, RSM1, SAMDC4	14	2 (14%)	4 (29%)	–
Cell wall	4CL, ALDH, C3H, C4H, CAD, CAD2, CCoAOMT, CCR, CES, COMT, F5H, HCT, HPRGP, IRX, LAC, OFP, PAL, POX, SND, XylS, XylT	121	41 (34%)	16 (13%)	5 (4%)
Cell wall TF	ERF, WRKY	6	2 (33%)	–	1 (17%)
Cell wall TF MYB	MYB	21	3 (14%)	–	1 (5%)
Cell wall TF NAC	NAC	11	2 (18%)	4 (36%)	1 (9%)
Chromatin remodelling	MET1, SWI	4	3 (75%)	2 (50%)	–
Lateral organ initiation	ANT, SLOMO, TOP1A	6	1 (17%)	–	–
Lateral organ patterning morphogenesis	AS, CLF, DOT5, GRF, KAN, NOV, SE, TRN1, YABBY, ZPR1, ZPR3	30	7 (23%)	3 (10%)	2 (7%)
Lateral organ identity	AN3, BOP, HDZIPIII	10	4 (40%)	1 (10%)	–
Light signalling	bHLHABAI, CO1, COP9, CRY, DET1, HY5, LHY, PCI, PFT1, PHYB, PIF, SPA	29	4 (14%)	7 (24%)	–
Shoot apical meristem	BARD1, BLH, CLPS3, FTA, KNAT, OBE1, ULT1, USP1, VEF2, WOX14, WUS	25	8 (32%)	5 (20%)	3 (12%)
Self-incompatibility	DUF247, GK	4	2 (50%)	1 (25%)	1 (25%)
Transition to flowering	CCA, FCA, FIE, FKF1, FLD, FPA, FT, FVE, FWA, FY, GI, LHP1, MBD9, PHP, RAV, SDG8, SPL3, VIL3, VRN1, VRN1-like	45	19 (42%)	12 (27%)	1 (2%)
Flower development	ESD4, HAC3, LFY3, LUG, MADS, RGA, SEU, SUF4, SUP	31	2 (6%)	4 (13%)	–
Transcription factor	BIM2, TCP	8	2 (25%)	–	–
Phytohormone biosynthesis, signalling and response					
ABA biosynthesis	NCED1, PDS1, PDS3	4	1 (25%)	1 (25%)	–
ABA signalling	ABI1, ABI3, ABI5, ABI8, AIP3, DRIP, GBF, GPA, GTG2, HD2C, PSY, SAD1, SIR3, WIG, ZEP	29	10 (34%)	3 (10%)	–
Auxin biosynthesis	TAA1, TAR2, YUC	6	3 (50%)	–	1 (17%)
Auxin signalling	ADA2B, AMP1, ARF, AUXIAA, AXR, AXR1, AXR4, AXR6, CAND1, GH3, TIR1	20	8 (40%)	3 (15%)	–
Auxin transport	AUX1, ENP, PGP4, PID2, PIN1, PIN1like, SPS	12	1 (8%)	–	–
Brassinosteroid biosynthesis	DWF1, DWF3, DWF5, DWF7, SQS	8	2 (25%)	1 (13%)	–
Brassinosteroid signalling	BES1	2	–	–	–
Cytokinin signalling	ARR, CRE, GCR1, RR	11	2 (18%)	1 (9%)	–
Ethylene biosynthesis	ACS	2	1 (50%)	–	–
Ethylene signalling	EBF1, EBF2, EIL3, EIN2, ETO1, ETR1	13	7 (54%)	1 (8%)	1 (8%)
Gibberellin biosynthesis	GAOX	11	4 (36%)	–	2 (18%)
Gibberellin signalling	GID1A, SHI, SPY	5	–	1 (20%)	–
Strigolacton biosynthesis	D14, D27, MAX1, MAX3, MAX4	11	3 (27%)	2 (18%)	–
Strigolacton signalling	MAX2, TB1	4	–	–	–
Total	**180**	**503**	**144**	**72**	**19**

3.4 Influence of read mapping algorithms and variant calling pipelines

For each VC pipeline, the similarity of SNP and indel sets identified using BWA-MEM or GSNAP mappings was calculated using the Jaccard Index (Fig. 2). The similarity of SNP_{BWA} and $\text{SNP}_{\text{GSNAP}}$ sets was lowest for Freebayes (0.73) and highest for SAMtools (0.83). The similarity of $\text{indel}_{\text{BWA}}$ and $\text{indel}_{\text{GSNAP}}$ sets was lower than that of SNP_{BWA} and $\text{SNP}_{\text{GSNAP}}$, independent of the VC pipeline. Jaccard index between $\text{indel}_{\text{BWA}}$ and $\text{indel}_{\text{GSNAP}}$ sets ranged from 0.47 (Freebayes) to 0.70 (GATK HC). On average, 11% of the SNPs were uniquely identified in the SNP_{BWA} set and 9% of the SNPs were uniquely identified in the $\text{SNP}_{\text{GSNAP}}$ set. Likewise, on average, 17% of the indels were uniquely identified in the $\text{indel}_{\text{BWA}}$ set and 19% of the indels were uniquely identified in the $\text{indel}_{\text{GSNAP}}$ set. In summary, the choice of read mapper did not affect the SNP and indel sets as much as the choice of VC pipeline. For results presented below, only variants identified on BWA-MEM mappings are shown, as the same trend was observed for GSNAP mappings.

3.5 Concordance of variant sets produced by four variant calling pipelines

Next, the size and concordance of variant sets (bi-allelic SNPs and indels) identified by the four VC pipelines were compared (Fig. 3). The number of SNPs was highest for GATK UG and SAMtools and considerably lower for Freebayes. The number of indels was at least four times lower than the number of SNPs identified by the same VC

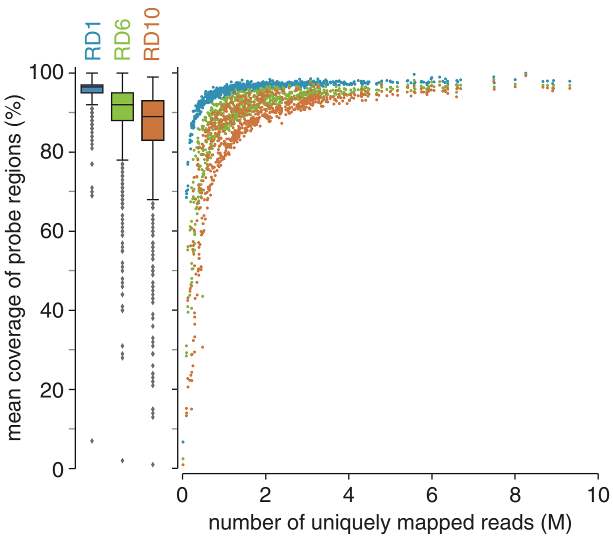

Figure 1. Target region coverage per genotype. For each of the 503 candidate genes, the target region was delineated as the gene model and an additional 1,000 bp upstream promoter region. The mean fraction of the target region covered per genotype is shown in function of the number of uniquely mapped reads using BWA-MEM after duplicate removal, using different RD thresholds [RD ≥ 1 (blue), RD ≥ 6 (green) and RD ≥ 10 (orange)].

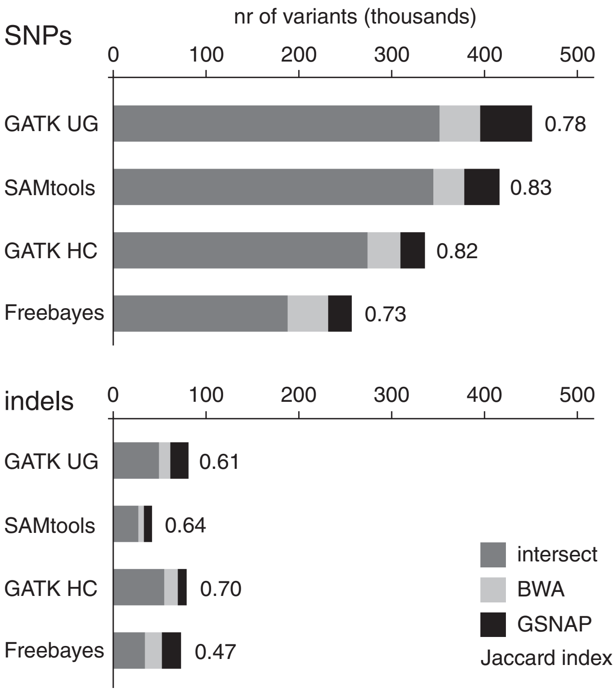

Figure 2. Overlap of variant sets generated using BWA-MEM and GSNAP mappings as input for four VC pipelines. SNPs and indels were determined for 503 candidate genes in 736 genotypes for BWA-MEM and GSNAP mappings using four VC pipelines. The intersect of variants sets was calculated to determine common variants (dark grey) and uniquely identified variants using BWA-MEM mappings (light grey) or GSNAP mappings (black) as input. The Jaccard index value indicates the corresponding similarity.

pipeline. GATK UG and GATK HC identified the highest number of indels, and SAMtools the least. The concordance of all four VC pipelines was low: only 150k SNPs (33% of the total number of SNPs identified) and 6.8k indels (5% of the total number of indels identified) were commonly identified, in line with previous reports.[14]

3.6 Precision-based filtering is more reliable than hard filtering

Hard filtering on e.g. minimal RD and GQ is a commonly used strategy to improve the precision of variant sets. As expected, both number of variant positions and call rate (number of genotype calls per position across 736 genotypes) decreased by filtering on minimal RD of six and minimal GQ score of 30 (Fig. 4a and b). Notably, hard filtering did not increase the concordance between VC pipelines (Supplementary Fig. S2), indicating that true variants were not necessarily identified by multiple VC pipelines. These results corroborate that it is difficult to build a reliable catalogue of sequence variation using a single VC pipeline and applying hard filtering.[51]

As an alternative for hard filtering on individual VC pipelines, the VMC[39] uses support vector machines to automatically combine multiple information sources (including RD and GQ values) generated by the four VC pipelines, and estimates the probability that a variant is a true genetic variant and not a sequencing artefact. The unfiltered, VMC integrated probe capture variant set contained 444,222 SNPs and 132,766 indels, determined in 736 genotypes and 503 candidate genes. By ordering the variants according to their probability, an EP was calculated for each variant, which can be used for precision-based filtering. In general, variants identified by multiple VC pipelines were assigned higher EP scores. As precision in this context refers to the number of true called variants, choosing an EP threshold is equivalent to finding a dataset- and aim-specific balance between sensitivity and precision of VC.[39]

3.7 Empirical determination of the EP threshold

Instead of using an arbitrary EP threshold, we reasoned that the EP threshold should be determined empirically, based on the distributions of EP values of HQ and low quality variants (LQ). In the absence of a published reference set of variants for the genotypes and genes used in this study, we generated an independent variant set for a subset of 78 genotypes using a Hi-Plex amplicon sequencing assay[52] of 171 amplicons, of which 147 overlap with 28 out of the 503 candidate genes. Hi-Plex amplicon sequencing resulted in 126,000 reads per genotype on average (range 11,000–418,000), corresponding to an average RD of 619 reads per amplicon (range 24–20,000).

Using the four VC pipelines integrated by the VMC resulted in a Hi-Plex variant set containing 813 SNPs and 184 indels, compared with 775 SNPs and 246 indels in the probe capture variant set that overlap with these amplicons. In total, 593 SNPs and 60 indels were commonly identified by the two independent sequencing-based genotyping methods. Together, these variants were defined as the HQ subset of variants. Conversely, SNPs and indels that were unique to either set (i.e. non-reproducible and more likely to be random artefacts), were defined as the LQ subset of variants per genotyping method.

To further validate HQ variants, we compared genotype calls of two methods (probe capture vs Hi-Plex) at the individual genotype level. The mean genotype call consistency, calculated as percentage of identical genotype calls on the total of 593 HQ SNPs and 60 HQ indels, over all 78 genotypes was 97% (range 93–100%). This high level of genotype call consistency confirmed the HQ of commonly identified variants. Inconsistent genotype calls are most likely the result of failed probe capture, low RD, the complexity of the region

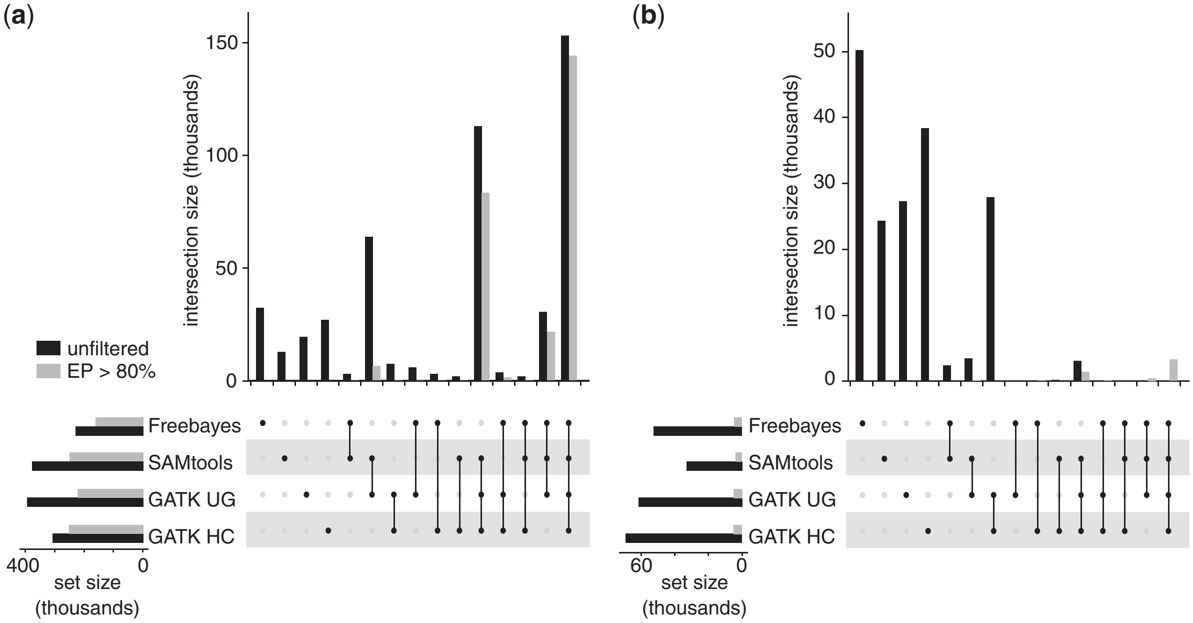

Figure 3. Size and concordance of bi-allelic SNP and indel sets of four VC pipelines, before and after precision-based filtering. SNPs and indels were identified for 503 candidate genes in 736 genotypes using four VC pipelines and concordance was calculated for bi-allelic SNPs (a) and indels (b). Per Upset plot, the lower left panel shows the total number of variants per VC pipeline; the lower right panel shows the overlap in call sets between the four VC pipelines. The bar graph shows the size per concordance group before (black) and after integration by VMC and precision-based filtering (EP > 80%) (light grey).

potentially hampering read mapping, allele specific amplification bias in the amplicon sequencing data, or combinations thereof.

Comparison of EP value distributions of HQ and LQ variant positions (Fig. 5) revealed that EP values of the Hi-Plex variant set were generally lower than those of the probe capture variant set, possibly because of higher RD and lower complexity of amplicon reads. Furthermore, EP values associated with HQ SNPs were higher than EP values of LQ SNPs for both Hi-Plex and probe capture SNP sets. Taken together, these data show that an EP threshold of 80% differentiates most HQ variants from LQ variants in the probe capture SNP set, whereas the EP threshold needs to be set at 70% to remove LQ variants from the Hi-Plex SNP set. This further indicates that different EP thresholds ought to be used depending on the genotyping method. In contrast to SNPs, there was no clear differentiation between EP values associated with HQ or LQ indels (Fig. 5). This shows that indel detection remains challenging because of mapping and/or realignment errors, and errors near repetitive regions,[53] thus leading to an incomplete indel set and underestimation of frameshift variants.

Using the empirically validated EP threshold of 80% for the probe capture variant set containing variants of 736 genotypes and 503 candidate genes resulted in 252,406 SNPs and 5,074 indels (Supplementary Data S3). The high genotype call consistency indicates that the VMC, at least for SNPs, was able to reliably integrate variant sets without losing genotype call quality. Moreover, using the VMC for precision-based filtering led to a higher genotype call rate compared with hard filtering of individual VC pipelines (Fig. 4c and d).

3.8 Validation of the resulting variant set in an F1 progeny

Mendelian inheritance in a segregating F1 progeny derived from a bi-parental cross was used as an accuracy measurement for the precision-based filtered variant set: MIEs are most likely the result of erroneous genotype calls. The set of 736 individuals contained 2 parents and their respective F1 progeny of 29 individuals. The genotype calls of these individuals were used to calculate the number of MIEs. Out of the 257,480 variants, 10,669 contained a missing genotype call in either one or both parents (4%) and could not be tested. For the 246,881 remaining variants and 29 individuals, 89,789 MIEs were identified, of which 57,326 (63%) were due to a missing genotype call. The other 32,463 MIEs represent a genotype call error in only a fraction of all genotype calls among the 246,881 variants in this F1 progeny (<0.5%). Moreover, these MIEs corresponded to 9,440 variant positions (4%) of which most had a MIE in a single individual (Supplementary Fig. S3).

3.9 Effects of sequence variation on gene function

We investigated the consequences of sequence variants on predicted gene function, using the manually curated HQ gene models to annotate the variants with SnpEff.[41] The complete annotation of functional effects for each of 252,406 SNPs and 5,074 indels is available as Supplementary Data S3. Out of the 257,480 variants, 65,225 resided in exon regions (25%) and 116,274 in intron regions (45%) corresponding to a density of 8.6 and 10.1 variants per 100 bp, respectively. Among the SNPs in coding regions, 38% were non-synonymous substitutions, which is consistent with previous observations in *L. perenne* transcriptomes.[23–25]

A general overview of the abundance of high impact effects on gene function, listed per functional category or pathway is presented in Table 1. These include gain of stop codons, frame shifts and alterations in splice sites, as they are most likely to disrupt protein function, possibly leading to loss of function (LOF), and causing phenotypic variation. For instance, the variant set contained 256 stop gain variants, affecting 144 out of the 503 candidate genes. The position of each stop gain relative to the total CDS length could indicate the degree

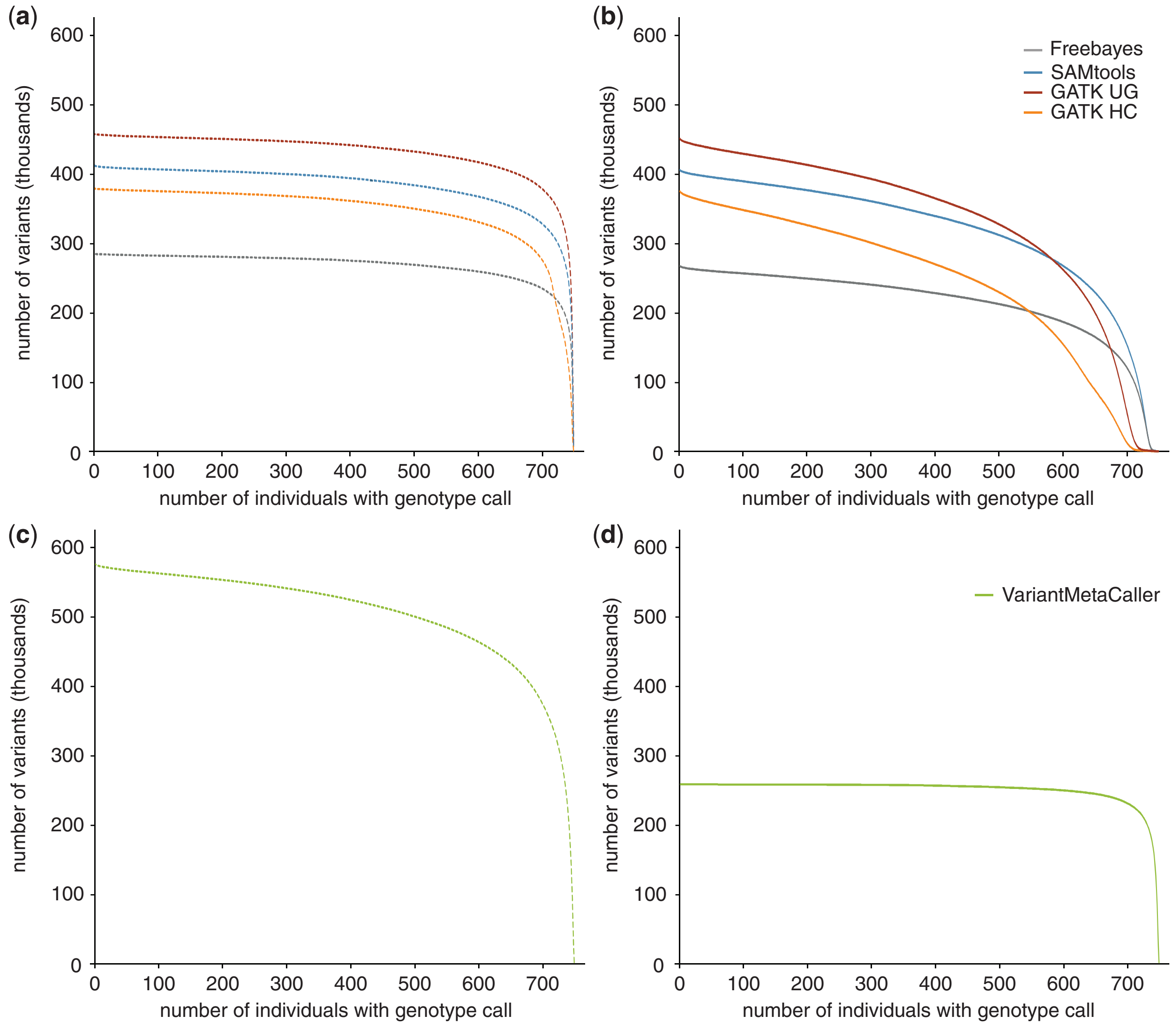

Figure 4. Effect of hard filtering and precision-based filtering on the saturation of genotype calls across the 736 genotypes. SNPs and indels were determined for 503 candidate genes in 736 genotypes using four VC pipelines and integrated using the VMC. The genotype call rate was calculated as the number of genotype calls present for each variant, over the total number of genotypes, and plotted cumulatively to estimate the genotype call saturation. This was done for bi-allelic variant sets: (a) before and (b) after hard filtering (RD > 6, GQ > 30) of the variant sets returned by the four VC pipelines and (c) before and (d) after precision-based filtering (EP > 80%) of the VMC output.

to which the protein is affected (Supplementary Fig. S4). Most of these stop gain variants occur at low allele frequency across the germplasm collection. Additionally, 72 candidate genes were affected by splice site variants: 40 variants affected donor splice sites and 47 variants affected acceptor splice sites. In line with the relatively low number of indels (5,074), only 20 frameshift variants were identified in 19 genes. In summary, naturally occurring LOF alleles could be readily identified in as much as one-third of the genes tested across various pathways that are important for plant growth and development.

This variant catalogue can be exploited in a dual fashion: (1) to associate genomic variation with phenotypic variation using an association mapping approach, which we are currently performing for architectural traits and cell wall digestibility, or (2) to mine for rare defective alleles, i.e. variants that disrupt gene function or regulation, and to subsequently select carriers of these variants for detailed phenotypic analysis. For example, we observed naturally occurring alleles for the single copy genes *GIGANTEA* (*LpGI-01*) and *ENHANCED RESPONSE TO ABSCISIC ACID 1* (*LpERA1-01*), in which a premature stop codon truncates translation at 5% and 23% of the protein length, respectively. Crosses with the carriers of these putative null alleles could help to clarify the function of *LpGI-01* in the regulation of flowering time, circadian clock, and/or hypocotyl elongation[54] and *LpERA1-01* in meristem organization and the ABA-mediated signal transduction pathway.[55]

Sequence variants were determined in a germplasm collection representing commercial cultivars, breeding populations and wild accessions, so as to ensure the downstream application in current breeding programs. For instance, the 170 amplicons used to estimate the EP threshold, and to validate the genotype calls of the probe capture set, were designed to cover the genetic diversity in 28 genes putatively involved in flowering time and other phenotypic traits of interest to breeders. Design and validation of these amplicons is a clear illustration of the application of the variant set. Since a comprehensive set of SNPs and indels are now known for our breeding materials, detailed and customised design of PCR primers targeting specific SNPs in candidate genes spread across the genome, while avoiding polymorphisms in the

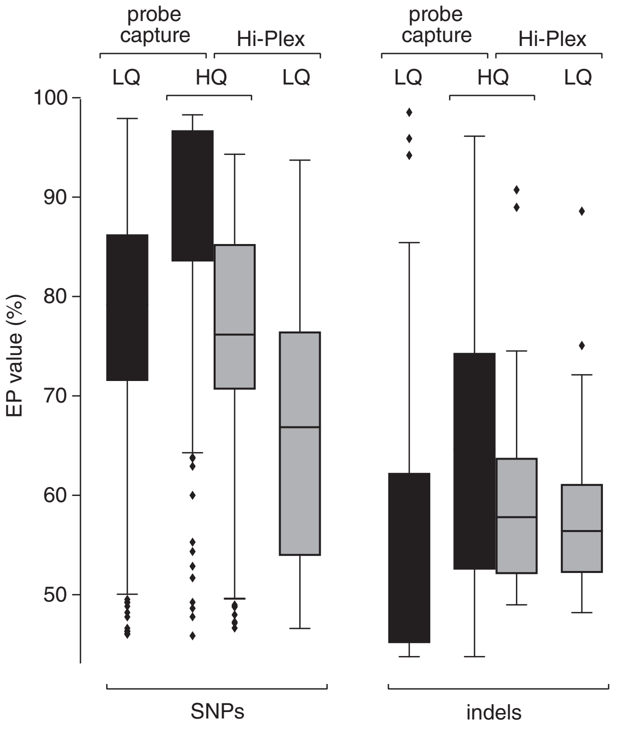

Figure 5. Distribution of EP values in HQ and LQ variant sets. For variants present in 78 genotypes and 147 amplicons, box plots show the distributions of EP values for commonly identified (HQ) and uniquely identified variants (LQ) in the probe capture and Hi-Plex variant sets.

flanking primer binding site, becomes feasible. Similar methods and criteria apply to the design of hybridization probes for high density SNP arrays. We are currently using Hi-Plex amplicon sequencing as a very cost- and time-efficient method to screen hundreds of variants simultaneously in a few thousand genotypes, a scale required to screen for putative associations with phenotypic traits in our current breeding populations.

3.10 Reconstruction of divergent alleles enables better characterization of genetic variation

The prime goal of targeted resequencing is to *de novo* discover alleles that are divergent from the reference genome sequence. However, the capture efficiency of a divergent sequence is reduced with increasing sequence dissimilarity to the reference sequence for which the probes were designed. Additionally, reads may fail to map to highly divergent regions if the parameters for short read alignment are too stringent.[18] To circumvent mapping short reads to a single reference sequence, on which classical VC pipelines rely, we devised a *de novo* assembly strategy to reconstruct full-length alleles. First, *de novo* assembly of the captured reads was performed per individual genotype to reconstruct alleles for each of the 503 candidate genes in parallel. Next, all *de novo* assembled contigs from all 736 genotypes were aligned to the reference genome to sort out and extract all corresponding allelic fragments per candidate gene. Per gene, all contigs were clustered using the OLC assembler CAP3[44] to collapse allelic redundancy and resolve fragmented gene sequences. This three-step approach results in a collection of alternative alleles assigned to each of the 503 candidate genes (on average 58 contigs per gene, range 4–203). The entire set of 29,320 CAP3 contigs is available as Supplementary Data S5. The value of this approach is that we can now characterize genetic variation in regions of high sequence diversity where traditional short read mapping-based VC pipelines fail.

We demonstrate the value of this approach for *LpSDUF247*, but all analyses described below may be repeated for any of the other 502 candidate genes, using the sequence data supplied as Supplementary Data S5. *LpSDUF247* is a highly polymorphic gene co-segregating with the *S*-locus that determines the grass self-incompatibility system.[46] CAP3 clustered the allelic diversity present in the 736 genotypes into 34 separate contigs. Alignment of CAP3 contigs identified a central region of the protein with high sequence divergence, while this region is virtually free from SNPs in the VC dataset, showing the limitations of read mapping-based VC. Six contigs displayed high sequence similarity (98%) to the reference sequence or to other contigs and were removed. Within the remaining allelic contigs, a single exon encoded for the DUF247 protein. The translated proteins showed only 73–84% global sequence identity with each other (Fig. 6a). These data are consistent with the previously reported identification of at least five unique alleles of *LpSDUF247* with 80–90% protein sequence identity.[46] Phylogenetic analysis confirmed that *de novo* assembled contigs were indeed novel alleles of *LpSDUF247* at the *S*-locus, and not of any of the 24 other *DUF247 paralogs* in the *L. perenne* genome (Supplementary Figs S5 and S6).

Next, we analysed the distribution of *LpSDUF247* alleles across the *L. perenne* germplasm collection. The 28 novel alleles were added to the reference genome sequence, thus complementing the reference *LpSDUF247* allele, and giving reads the opportunity for near-perfect mapping at their respective allele. Differential read mapping *across* the alleles in a multi-allelic context was then used to score which alleles are present in each genotype. Mapping reads in a multi-allelic context eliminates the need for VC, but only if alleles are sufficiently divergent so that differential RD can be used to identify which alleles are present per genotype. Near-perfect mapping of the raw reads onto the newly constructed *LpSDUF247* alternative alleles confirmed their existence in the *L. perenne* germplasm collection, except for allele 31 which had no read support (Fig. 6c). This also shows that the capture efficiency of 120 bp probes was sufficient to detect alleles with as little as 80% sequence identity to the reference genome sequence.

In the vast majority of genotypes (501 out of 736) the reads almost exclusively mapped onto a combination of two *LpSDUF247* alleles, often at similar RD, and only a minor fraction (<5%) of reads mapped to additional alleles (Fig. 6c). There was no bias for combinations of alleles across wild accessions, breeding populations and cultivars, and clear segregation of alleles was observed in the F1 progeny ($n = 29$) of a bi-parental cross that was included in the set of 736 individuals (Supplementary Fig. S7). Furthermore, 57 genotypes displayed reads mapping only to a single allele, suggesting either homozygosity or the failure to capture and sequence yet undiscovered alternative alleles with even stronger sequence divergence to the reference genome sequence used to design the probes. Finally, 177 genotypes displayed RD spread over three or more alternative alleles. In 65 of them, the higher allele count could be explained by a consistent co-segregation of *LpSDUF247*-04 with *LpSDUF247*-28 suggesting a gene duplication, in combination with an additional, variable third allele. In the remaining 112 genotypes, the observation that reads map to more than two alleles in a multi-allelic reference genome, could indicate ambiguity of read mapping between closely related alleles, or the presence of additional alleles derived from cross-over events at the *LpSDUF247* locus.

Although *LpSDUF247* was the most extreme case of sequence divergence in alternative alleles, Supplementary Figure S8 presents four

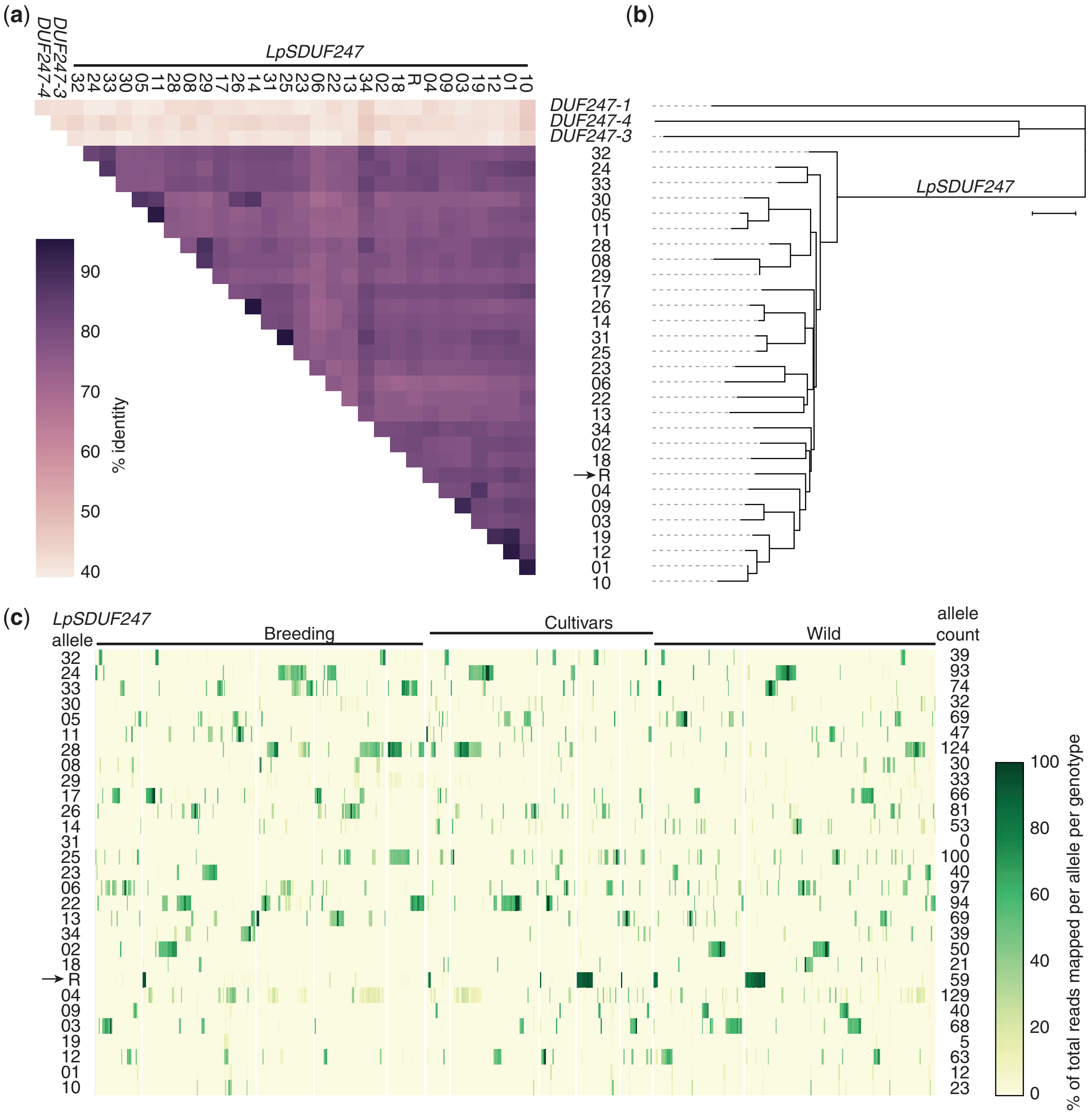

Figure 6. Sequence diversity and distribution of 28 newly identified alleles of *LpSDUF247* across breeding populations and wild accessions. A similarity matrix (a) and phylogenetic tree (b) were built using the protein sequences of 28 alleles and reference sequence (R) of *LpSDUF247*, together with three additional *DUF247* genes. Panel c gives an overview of the distribution of the *LpSDUF247* alleles across the gene pool. The alleles present per genotype were identified by mapping the reads to a multi-allelic reference genome, and calculating the ratio of average RD per allele over the total number of reads mapping to *LpSDUF247* alleles.

other candidate genes with different levels of divergence, global or local. The alternative alleles of *LpMAX3-01* and *LpETR1-01* showed only local sequence divergence, at the introns and 5′UTR regions, respectively. The sequence variation of *LpFT-04* was captured in only four contigs, explaining why the variant density across the gene region was low, especially when low frequent SNPs were filtered out. Taken together, the analysis of *LpSDUF247* demonstrates the rich sequence diversity that can be mined for in this catalogue of genomic diversity across 503 candidate genes.

Acknowledgements

We thank the breeders of ILVO, Barenbrug and Eurograss for providing plant material, András Gézsi for support and advice for running the VMC and Kurt Lamour (Floodlight Genomics) for his services provided for the Hi-Plex amplicon sequencing assay. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government—department EWI. This research was supported by the Agency for the promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium (project LO-080510), and by Aarhus University and the Godfred Birkedal Hartmann foundation in Denmark. Elisabeth Veeckman was supported by an ILVO PhD fellowship.

Funding

Agency for the promotion of Innovation by Science and Technology (IWT) project LO-080510, Aarhus University, Godfred Birkedal Hartmann foundation.

Accession numbers

SRR accession numbers: SRR6812717–SRR6813075, SRR6813540–SRR6813585.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Roldán-Ruiz, I. and Kölliker, R. 2010, Marker-assisted selection in forage crops and turf: a review. *In: Sustainable Use of Genetic Diversity in Forage and Turf Breeding* (pp. 383–390), ed. Huyghe, C. 383–390. Berlin, Germany: Springer.
2. Yu, G., Sauchyn, D. and Li, Y. F. 2013, Drought changes and the mechanism analysis for the North American Prairie, *J. Arid Land*, **5**, 1–14.
3. Brazauskas, G., Pašakinskienė, I., Asp, T. and Lübberstedt, T. 2010, Nucleotide diversity and linkage disequilibrium in five *Lolium perenne* genes with putative role in shoot morphology, *Plant Sci*, **179**, 194–201.
4. Auzanneau, J., Huyghe, C., Julier, B. and Barre, P. 2007, Linkage disequilibrium in synthetic varieties of perennial ryegrass, *Theor. Appl. Genet.*, **115**, 837–47.
5. Skot, L., Humphreys, J., Humphreys, M. O. et al. 2007, Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.), *Genetics*, **177**, 535–47.
6. Fiil, A., Lenk, I., Petersen, K. et al. 2011, Nucleotide diversity and linkage disequilibrium of nine genes with putative effects on flowering time in perennial ryegrass (*Lolium perenne* L.), *Plant Sci*, **180**, 228–37.
7. Skot, L., Sanderson, R., Thomas, A. et al. 2011, Allelic variation in the perennial ryegrass flowering locus T gene is associated with changes in flowering time across a range of populations, *Plant Physiol*, **155**, 1013–22.
8. Studer, B., Byrne, S., Nielsen, R. O. et al. 2012, A transcriptome map of perennial ryegrass (*Lolium perenne* L.), *BMC Genomics*, **13**, 140.
9. Blackmore, T., Thomas, I., McMahon, R., Powell, W. and Hegarty, M. 2015, Genetic-geographic correlation revealed across a broad European ecotypic sample of perennial ryegrass (*Lolium perenne*) using array-based SNP genotyping, *Theor. Appl. Genet.*, **128**, 1917–32.
10. Chung, Y. S., Choi, S. C., Jun, T. H. and Kim, C. 2017, Genotyping-by-sequencing: a promising tool for plant genetics research and breeding, *Hortic. Environ. Biotechnol.*, **58**, 425–31.
11. Xing, Y., Frei, U., Schejbel, B., Asp, T. and Lubberstedt, T. 2007, Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*, *BMC Plant Biol.*, **7**, 43.
12. Byrne, S. L., Nagy, I., Pfeifer, M. et al. 2015, A synteny-based draft genome sequence of the forage grass *Lolium perenne*, *Plant J.*, **84**, 816–26.
13. Liu, X., Han, S., Wang, Z., Gelernter, J. and Yang, B.-Z. 2013, Variant callers for next-generation sequencing data: a comparison study, *PLoS One*, **8**, e75619.
14. O'Rawe, J., Jiang, T., Sun, G. et al. 2013, Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing, *Genome Med.*, **5**, 28.
15. Pirooznia, M., Kramer, M., Parla, J. et al. 2014, Validation and assessment of variant calling pipelines for next-generation sequencing, *Hum. Genomics*, **8**, 14.
16. Yu, X. Q. and Sun, S. Y. 2013, Comparing a few SNP calling algorithms using low-coverage sequencing data, *BMC Bioinformatics*, **14**, 274.
17. Quail, M. A., Smith, M., Coupland, P. et al. 2012, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, *BMC Genomics*, **13**, 341.
18. Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B. and van Nimwegen, E. 2014, Automated reconstruction of whole-genome phylogenies from short-sequence reads, *Mol Biol Evol*, **31**, 1077–88.
19. Gan, X., Stegle, O., Behr, J. et al. 2011, Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*, *Nature*, **477**, 419.
20. Proost, S., Van Bel, M., Vaneechoutte, D. et al. 2015, PLAZA 3.0: an access point for plant comparative genomics, *Nucleic Acids Res*, **43**, D974–81.
21. Edgar, R. C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*, **32**, 1792–7.
22. Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. 2010, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, *Syst Biol*, **59**, 307–21.
23. Farrell, J. D., Byrne, S., Paina, C. and Asp, T. 2014, De novo assembly of the perennial ryegrass transcriptome using an RNA-Seq strategy, *PLoS One*, **9**, e103567.
24. Paina, C., Byrne, S. L., Domnisoru, C. and Asp, T. 2014, Vernalization mediated changes in the *Lolium perenne* transcriptome, *PLoS One*, **9**, e107365.
25. Ruttink, T., Sterck, L., Rohde, A. et al. 2013, Orthology guided assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in *Lolium perenne*, *Plant Biotechnol. J.*, **11**, 605–17.
26. Trapnell, C., Pachter, L. and Salzberg, S. L. 2009, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105–11.
27. Bertone, P., Trifonov, V., Rozowsky, J. S. et al. 2006, Design optimization methods for genomic DNA tiling arrays, *Genome Res.*, **16**, 271–81.
28. Murray, M. G. and Thompson, W. F. 1980, Rapid isolation of high molecular-weight plant DNA, *Nucleic Acids Res.*, **8**, 4321–5.
29. Uitdewilligen, J. G. A. M. L., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J. A., Visser, R. G. F. and van Eck, H. J. 2013, A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato, *PLoS One*, **8**, e62355.
30. Bolger, A. M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
31. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754–60.
32. Wu, T. D. and Nacu, S. 2010, Fast and SNP-tolerant detection of complex variants and splicing in short reads, *Bioinformatics*, **26**, 873–81.
33. Quinlan, A. R. and Hall, I. M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
34. Li, H., Handsaker, B., Wysoker, A. et al. 2009, The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
35. Garrison, E. and Marth, G. 2012, Haplotype-based variant detection from short-read sequencing, *arXiv Preprint arXiv*, **1207**, 3907.
36. McKenna, A., Hanna, M., Banks, E. et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**, 1297–303.
37. Van der Auwera, G. A., Carneiro, M. O., Hartl, C. et al. 2013, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline, *Curr Protoc Bioinformatics*, **43**, 11 10–33.
38. Danecek, P., Auton, A., Abecasis, G. et al. 2011, The variant call format and VCFtools, *Bioinformatics*, **27**, 2156–8.
39. Gézsi, A., Bolgár, B., Marx, P., Sarkozy, P., Szalai, C. and Antal, P. 2015, VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering, *BMC Genomics*, **16**, 875.
40. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. and Pfister, H. 2014, UpSet: visualization of Intersecting Sets, *IEEE Trans. Vis. Comput. Graph.*, **20**, 1983–92.

41. Cingolani, P., Platts, A., Wang le, L. et al. 2012, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3, *Fly (Austin)*, **6**, 80–92.

42. Untergasser, A., Cutcutache, I., Koressaar, T. et al. 2012, Primer3-new capabilities and interfaces, *Nucleic Acids Res.*, **40**, e115.

43. Zhang, J. J., Kobert, K., Flouri, T. and Stamatakis, A. 2014, PEAR: a fast and accurate Illumina Paired-End reAd mergeR, *Bioinformatics*, **30**, 614–20.

44. Huang, X. Q. and Madan, A. 1999, CAP3: a DNA sequence assembly program, *Genome Res.*, **9**, 868–77.

45. Gremme, G., Brendel, V., Sparks, M. E. and Kurtz, S. 2005, Engineering a software tool for gene structure prediction in higher organisms, *Inform Software Tech*, **47**, 965–78.

46. Manzanares, C., Barth, S., Thorogood, D. et al. 2016, A gene encoding a DUF247 domain protein cosegregates with the S self-incompatibility locus in perennial ryegrass, *Mol. Biol. Evol.*, **33**, 870–84.

47. Veeckman, E., Ruttink, T. and Vandepoele, K. 2016, Are we there yet? Reliably estimating the completeness of plant genome sequences, *Plant Cell.*, **28**, 1759–68.

48. Ruttink, T., Haegeman, A., van Parijs, F.R.D. et al. 2015, Genetic diversity in candidate genes for developmental traits and cell wall characteristics in perennial ryegrass (*Lolium perenne*). In: Budak, H. and Spangenberg, G. (eds), *Molecular Breeding of Forage and Turf: The Proceedings of the 8th International Symposium on the Molecular Breeding of Forage and Turf*, pp. 93–109. Springer International Publishing, Cham.

49. Song, K., Li, L. and Zhang, G. 2016, Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology, *Sci. Rep.*, **6**, 35736.

50. Tian, S. L., Yan, H. H., Neuhauser, C. and Slager, S. L. 2016, An analytical workflow for accurate variant discovery in highly divergent regions, *BMC Genomics*, **17**, 703.

51. Park, M.-H., Rhee, H., Park, J. H. et al. 2014, Comprehensive analysis to improve the validation rate for single nucleotide variants detected by next-generation sequencing, *PLoS One*, **9**, e86664.

52. Nguyen-Dumont, T., Pope, B. J., Hammet, F. et al. 2013, Cross-platform compatibility of Hi-Plex, a streamlined approach for targeted massively parallel sequencing, *Anal Biochem*, **442**, 127–9.

53. Li, H. 2014, Toward better understanding of artifacts in variant calling from high-coverage samples, *Bioinformatics*, **30**, 2843–51.

54. Mishra, P. and Panigrahi, K. C. 2015, GIGANTEA – an emerging story, *Front. Plant Sci.*, **6**, 8.

55. Cutler, S., Ghassemian, M., Bonetta, D., Cooney, S. and McCourt, P. 1996, A protein farnesyl transferase involved in abscisic acid signal transduction in Arabidopsis, *Science*, **273**, 1239–41.