Received Date : 03-Aug-2016
Revised Date : 08-Nov-2016
Accepted Date : 21-Nov-2016
Article type : Resource

Towards a whole-genome sequence for rye (*Secale cereale* L.)

Eva Bauer[1*], Thomas Schmutzer[2*], Ivan Barilar[3], Martin Mascher[2], Heidrun Gundlach[4], Mihaela M. Martis[4,$], Sven O. Twardziok[4], Bernd Hackauf[5], Andres Gordillo[6], Peer Wilde[6], Malthe Schmidt[6], Viktor Korzun[6], Klaus F.X. Mayer[4], Karl Schmid[3], Chris-Carolin Schön[1], Uwe Scholz[2†]

* Co-first authors

†Corresponding authors: Eva Bauer, Uwe Scholz

[1] Technical University of Munich, Plant Breeding, Liesel-Beckmann-Str. 2, 85354 Freising, Germany

[2] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3, 06466 Stadt Seeland, Germany

[3] Universität Hohenheim, Crop Biodiversity and Breeding Informatics, Fruwirthstr. 21, 70599 Stuttgart, Germany

[4] Helmholtz Zentrum München, Plant Genome and Systems Biology, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

[5] Julius Kühn-Institute, Institute for Breeding Research on Agricultural Crops, Rudolf-Schick-Platz 3a, 18190 Sanitz, Germany

[6] KWS LOCHOW GMBH, Ferdinand-von-Lochow-Str. 5, 29303 Bergen, Germany

[$] Present address: NBIS (National Bioinformatics Infrastructure Sweden), Division of Cell Biology, Department of Clinical and Experimental Medicine, Linköping University, SE-558185 Linköping, Sweden

EB: e.bauer@tum.de (corresponding author 1)

TS: schmutzr@ipk-gatersleben.de

IB: ivan_barilar@uni-hohenheim.de

MM: mascher@ipk-gatersleben.de

HG: h.gundlach@helmholtz-muenchen.de

MMM: mihaela.martis@bils.se

SOT: sven.twardziok@helmholtz-muenchen.de

BH: bernd.hackauf@julius-kuehn.de

AG: andres.gordillo@kws.com

PW: peer.wilde@kws.com

MS: malthe.schmidt@kws.com

VK: viktor.korzun@kws.com

KFXM: k.mayer@helmholtz-muenchen.de

KS: karl.schmid@uni-hohenheim.de

CCS: chris.schoen@tum.de

US: scholz@ipk-gatersleben.de (corresponding author 2)

Running title: A whole-genome sequence for rye

Keywords: *Secale cereale* L., rye, whole-genome shotgun sequencing, *de novo* genome assembly, single nucleotide variants (SNVs), Rye600k genotyping array, high-density genetic map, rye genome zipper, diversity, selection signals

Summary

We report on a whole-genome draft sequence of rye (*Secale cereale* L.). Rye is a diploid Triticeae species closely related to wheat and barley and an important crop for food and feed in Central and Eastern Europe. Through whole-genome shotgun (WGS) sequencing of the 7.9 Gbp genome of the winter rye inbred line Lo7 we obtained a *de novo* assembly represented by 1.29 million scaffolds covering a total length of 2.8 Gbp. Our reference sequence represents nearly the entire low-copy portion of the rye genome. This genome assembly was used to predict 27,784 rye gene models based on homology to sequenced grass genomes. Through resequencing of 10 rye inbred lines and one accession of the wild relative *S. vavilovii*, we discovered more than 90 million single nucleotide variants (SNVs) and short insertions/deletions (indels) in the rye genome. From these variants, we developed the high-density Rye600k genotyping array with 600,843 markers which enabled anchoring the sequence contigs along a high-density genetic map and establishing a synteny-based virtual gene order. Genotyping data were used to characterize the diversity of rye breeding pools and genetic resources and to obtain a genome-wide map of selection signals differentiating the divergent gene pools. This rye whole-genome sequence closes a gap in Triticeae genome research and will be highly valuable for comparative genomics, functional studies and genome-based breeding in rye.

Introduction

Rye (*Secale cereale* L.) is a member of the Triticeae tribe of the grass family Poaceae and related to bread wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.). It has the largest genome (~7.9 Gbp; Bartoš *et al.*, 2008) among all diploid Triticeae with more than 90% repetitive sequences (Flavell *et al.*, 1974). As a first comprehensive sequence resource for rye, an expressed sequence tag (EST) library was established which allowed the

development of the Rye5k genotyping array (Haseneyer *et al.*, 2011), synteny-based comparisons with other cereal genomes which provided insights in the reticulate evolution of rye (Martis *et al.*, 2013), and mapping of quantitative trait loci (QTL) influencing agronomic traits (Miedaner *et al.*, 2012). As expected for an outcrossing species, previous studies in rye indicated higher levels of nucleotide diversity and a faster decay of linkage disequilibrium (LD) (Li *et al.*, 2011; Auinger *et al.*, 2016) compared to self-pollinating crops such as barley and wheat (Chao *et al.*, 2010; Zhou *et al.*, 2012). The low level of LD in rye promises high resolution in association genetic approaches for the identification of candidate genes for traits of interest, but at the same time requires large marker numbers which emphasizes the need for high-density genotyping platforms as they have been established recently for many major crop species (Voss-Fels and Snowdon, 2015).

Comprehensive whole-genome sequence information of the allogamous rye has been missing so far, whereas draft genome sequences of the related autogamous Triticeae species barley, bread wheat, *Aegilops tauschii* and *T. urartu* became available recently (Mayer *et al.*, 2012; Jia *et al.*, 2013; Ling *et al.*, 2013; Mayer *et al.*, 2014). These genomic resources are indispensable tools for understanding the biology and evolution of major Triticeae species through comparative genomic approaches (Spannagl *et al.*, 2016a) and for relating this knowledge to phenotypic traits (Esch *et al.*, 2015). As yet, rye is not well represented in public sequence databases, which prohibits large-scale functional analyses in rye and genomics-assisted genetic improvement of rye for sustainable crop production. Rye is an important model to elucidate the genetic and functional basis of traits which are also relevant for the genetic improvement of wheat and barley. It excels by an exceptional frost tolerance (Fowler and Carles, 1979) and outyields wheat and barley on poor and medium soils and under drought stress conditions (Schittenhelm *et al.*, 2014). Rye translocations are present in many wheat varieties grown worldwide and contribute to abiotic and biotic stress tolerance (Rabinovich, 1998), and in addition, rye is one of the parents of the man-made cereal triticale (× *Triticosecale*) (Oettler, 2005). Thus, the availability of rye whole-genome

sequences would facilitate the elucidation of genes and molecular mechanisms underlying important agronomic traits which are useful for the improvement of related Triticeae species. Here we report comprehensive rye genome resources consisting of a draft assembly, resequencing data from 10 rye inbred lines and *S. vavilovii*, a high-confidence gene set and a high-density genotyping array. We demonstrate their utility for comparative genomics, for investigating the genomic diversity in rye breeding pools and genetic resources (GR) and for detection of selection signals. These genomic resources will facilitate map-based cloning and functional characterization of genes underlying agronomic traits and fill a gap in current Triticeae genomics.

Results and discussion

Whole-genome shotgun sequencing, assembly and structural analysis

Genome sequencing and de novo assembly

The genome of the winter rye inbred line Lo7 was sequenced and *de novo* assembled using a WGS sequencing strategy. Several paired-end (PE) and mate-pair (MP) libraries were constructed and sequenced on the Illumina Hiseq2000 platform, resulting in approximately 72.4-fold total sequence coverage (Table S1, Table S2). Deep sequencing revealed an average GC content of 46.1% in PE350 and PE450 reads (Table S2), only slightly higher than the estimate of 45.9% reported from BAC-end survey sequencing of rye chromosome arm 1RS (Bartoš *et al.*, 2008). An elevated GC content of 46.6% observed in the assembled contigs indicates that genic regions, which tend to have a higher GC content in plants (Glémin *et al.*, 2014), are well represented in the rye WGS assembly. The *de novo* assembled rye genome consisted of 1.58 million contigs totalling 1.68 Gbp of gap-free sequence which most likely covers the low-copy portion of the rye genome. Through subsequent scaffolding we obtained 1.29 million scaffolds with a length of 2.80 Gbp (Table 1) which corresponds to around 35% of the rye genome. This value might

underestimate genome coverage since typically in WGS assemblies of large plant genomes single- or low-copy sequences are enriched whereas highly repetitive sequences are difficult to assemble or may collapse (Treangen and Salzberg, 2012).

The genome assembly was used to predict 27,784 rye high-confidence (HC) gene models through a reference-based approach (Table S3), which is similar to the 26,159 HC genes reported in barley (Mayer *et al.*, 2012). We validated the genome assembly using the 'Benchmarking Universal Single-Copy Orthologs' (BUSCO; (Simão *et al.*, 2015)) gene set and found 89% of all BUSCO plant genes being represented in the assembly (Table S4). With this proportion the genome assembly and annotation completeness in rye is comparable to other plant species (Visser *et al.*, 2015; Xu *et al.*, 2015). Previously published draft genomes as for instance the close relative barley (Mayer *et al.*, 2012) accelerated forward genetic approaches and enabled novel strategies for genome research such as exome-capture sequencing (Mascher *et al.*, 2013b; Mascher *et al.*, 2014). Therefore, we expect that the rye draft genome will promote genome analysis and gene discovery to a new level.

Repetitive elements

Transposable elements (TE), constituting the major portion of genomic repeat elements, were annotated and classified by a homology-based approach using a comprehensive *Poaceae* repeat library. The overall TE content of rye as estimated from 800 Mbp random Illumina sequence reads amounted to at least 72% (Table S5). Long terminal repeat (LTR) retrotransposons were prevalent with a content of 60%, followed by a much lower amount (7%) for DNA transposons. Although short sequence reads of highly repetitive genomes are typically difficult to assemble, the 1.68 Gbp assembly still contained 60% (~1 Gbp) TEs, but with a moderately different distribution of transposon classes: LTR retrotransposons, in particular the generally younger and still more repetitive *copia* subgroup, were depleted in the assembly, whereas DNA transposons and non-LTR retrotransposons were enriched (Table S5). Both of these increased TE types, especially MITEs (short DNA transposon

derivatives), are known to reside in the vicinity of genes and thus confirm the high gene content of the assembled sequences. The TE compositions for the rye Lo7 WGS assembly and raw Illumina sequence reads, especially the enrichment of the gene-associated MITEs, were in line with previous findings for barley (Mayer *et al.*, 2012). The rye genome is one example for the massive mobile element accumulation within the Triticeae (Middleton *et al.*, 2013). The availability of this whole-genome sequence in combination with resequencing data from 10 inbred lines and *S. vavilovii* will enable insights in TE dynamics in rye and will be a useful resource for gaining more insights in Triticeae genome evolution (Wicker *et al.*, 2009).

Variant calling and diversity patterns in rye

To assess the sequence diversity in contemporary rye breeding lines on a genome-wide level and to discover sequence variants for the development of a genotyping array suitable for rye research and breeding, we sequenced five representative lines from each of the two heterotic pools used in hybrid breeding – the seed and pollen parent pool. In addition, one accession of the wild species *S. vavilovii* was sequenced as a putative ancestor of cultivated rye. We obtained 14.0- to 15.4-fold genome coverage in the 11 samples which together yielded 1.27 Tbp of sequence data (Table S6). On average, we found for each genotype ~245 million reads properly paired, covering 80.2% of the reference genome with more than four reads. As expected due to higher sequence divergence, coverage of the reference genome was reduced to 72.4% in *S. vavilovii* (Table S6). In total, 90,012,964 SNVs and short indels were discovered in the rye genome. For convenience, we refer to both types of variants in the following as SNVs unless stated otherwise. After stringent quality filtering and removal of heterozygous sites, 8,626,622 variants including 220,766 indels remained. Around 24.3% of these SNVs were unique to *S. vavilovii*, 15.2% and 22.0% were specific for the seed and pollen parent pools, respectively, and 11.2% were shared between the two breeding pools and *S. vavilovii* (Figure S1). The remaining 27.2% were shared by only two of the three groups.

Based on this filtered set of SNVs we determined nucleotide diversity (π per site) in the two breeding pools. The average values in the five seed and five pollen parent pool lines (0.260 and 0.254, respectively) were significantly different ($p < 2.2e^{-16}$). Taking the 10 lines from both pools together to represent the overall diversity in rye breeding germplasm, a higher level of nucleotide diversity of 0.295 was observed. As shown above a considerable portion of SNVs was specific for each of the breeding pools, which was also reflected by an F_{ST} estimate of 0.108 indicating a moderate differentiation between the two populations. Despite a relatively small sample size, this level of differentiation between the two heterotic pools in rye is similar as observed in a large and diverse panel of temperate maize lines from two heterotic pools (Unterseer *et al.*, 2016) and consistent with expectations for divergent breeding pools in outcrossing species.

A condensed overview of SNV distribution along the rye chromosomes is shown in Figure 1. Our analysis revealed patterns of divergence between the seed and pollen pool elite lines with obvious differences mainly in (peri-)centromeric regions. Diversity hot spots were visible in *S. vavilovii* in regions which showed reduced variation in rye elite lines, e.g. on chromosome arms 1RS and 6RS. Such regions could be targets for mining the diversity present in genetic resources and wild ancestors of rye. The centromeric and extensive peri-centromeric regions of rye chromosomes contain a large number of genes that are enclosed in recombinationally inactive genomic regions which is similar to findings in barley and wheat (Mayer *et al.*, 2011; Mayer *et al.*, 2014). Approximately 17,196 (32.5%) of the genetically anchored WGS contigs were assigned to these regions, which are challenging to access in positional cloning and in plant breeding since it is difficult to break up the large linkage blocks. Especially in these (peri-)centromeric regions of the genetic map which cover large physical distances, residual heterozygosity was observed in the inbred reference line Lo7 (Figure 1, outmost track). The residual heterozygosity in the highly inbred line Lo7 corroborates the assumption that due to genetic load rye inbred lines retain a certain rate of heterozygosity to avoid lethal effects of homozygous deleterious genes (Thompson and

Rees, 1956). A similar phenomenon was reported in maize, where excess heterozygosity was observed in peri-centromeric regions (McMullen *et al.*, 2009). It cannot be excluded however, that to some extent the apparent heterozygosity in the assembly of rye line Lo7 arises from mapping of reads to duplicated sequences that collapsed in the assembly.

The majority of SNVs (97.6%) in rye was located in intronic/intergenic regions. In 9,675 out of the 27,784 predicted genes we found a considerable proportion (1.01%) of non-synonymous SNVs (nsSNVs) in at least one of the resequenced lines (Table 2, Table S7, Figure S2). These nsSNVs may encode functional polymorphisms and thus may influence gene integrity. Loss-of-function by gain or loss of a stop codon is the main large-effect polymorphism in coding sequences. We compared the resequenced lines from the seed and pollen parent pools and found more nsSNV mutations in the pollen (8,483) than in the seed (7,907) parent pool (Figure S2). This was expected, since lines from the pollen parent pool are genetically more distant from the reference line Lo7, which belongs to the seed parent pool, and thus exhibit a larger total number of SNVs than lines from the seed parent pool. However, with 1.01% the proportion of nsSNVs was the same in seed and pollen parent pools (Table 2, Table S7). In a single sequenced plant of *S. vavilovii*, our study revealed a high number of nsSNVs (6,341) which was almost twice than in each of the other ten lines, but with 0.88% the proportion relative to the total number of SNVs found in *S. vavilovii* was at a slightly lower level than in the two rye breeding pools (Table 2, Table S7). Due to filtering of heterozygous SNV calls in our dataset potentially deleterious alleles in heterozygous stage in *S. vavilovii* might have been removed which may explain the lower proportion compared to the inbred lines.

We further investigated the gene content among the studied rye inbred lines and *S. vavilovii* and found substantial presence/absence variation (PAV). Nine percent (9,007) of all exons were missing in at least one of the 11 resequenced genotypes. When focusing on the gene-space of rye we found that 2,934 gene models were missing in at least one genotype. Cluster analysis of these gene models with PAV patterns showed a cluster split in the two

breeding pools and separated *S. vavilovii* from the *S. cereale* lines, indicating pool- or species-specific PAV patterns (Figure S3). A considerable proportion (39.0%) of these gene losses was detected only in single inbred lines but we also found 1,251 (42.6%) gene models missing in at least three genotypes. Among these more frequently missed genes we detected pool-specificity for 80 and 132 candidate genes of the seed and pollen parent pool, respectively.

Design of a Rye600k genotyping array

The complete set of SNVs discovered by resequencing of 10 elite lines and one accession of *S. vavilovii* and their functional annotation was subsequently used to design a Rye600k genotyping array which aimed for a uniform representation of markers across all rye chromosomes and optimal coverage of exonic SNVs (for details see Experimental procedures). Given the genetic composition of the resequencing panel with mainly lines from two divergent elite breeding pools, we included a substantial number of SNVs from the wild species *S. vavilovii* on the array to additionally cover polymorphisms representative for rye GR (Figure S4). Phylogenetic trees constructed from the WGS data set and from the SNVs represented on the Rye600k array showed the same topology and reflected the breeding history of the lines (Figure S5).

The 600,843 SNVs on the Rye600k array were experimentally validated in a broad germplasm panel comprising 84 elite inbred lines from the seed and pollen parent breeding pools, 46 diverse accessions from GR, and 133 recombinant inbred lines (RILs) from a mapping population. This genotyping panel included the reference line Lo7, seven of the 10 resequenced inbred lines and the *S. vavilovii* accession. Average call rates in samples from the seed and pollen parent pools and GR were high with 96.4%, 96.2% and 94.4%, respectively. The fact that probe sequences on the array were derived from the inbred line Lo7 likely contributed to the higher call rate in the seed parent pool, whereas the lower call

rates in GR may be explained by increased sequence difference between GR and elite material. The lowest call rates were observed in samples from the three wild relatives *S. strictum* (89.6%), *S. vavilovii* (87.5%) and *S. sylvestre* (84.3%), consistent with their evolutionary distance from cultivated rye (Jones and Flavell, 1982). After genotype clustering, more than half (52.7%) of the SNVs fell in one of the three Affymetrix SNV categories "Poly High Resolution" (PHR), "Off-Target Variant" or "No Minor Homozygote" which are useful for genotyping and can be regarded as technically validated (Table S8). PHR SNVs had the highest proportion in this group (39.3% of all SNVs) and can be considered as producing the most reliable genotype calls. The remaining SNVs (47.3%) were classified as "Other", "Call Rate Below Threshold" or "Mono High Resolution", indicating difficulties with genotype clustering in the first two cases or a lack of polymorphism in the latter case. Overall, the validation rate was similar to genotyping arrays of other species (Unterseer *et al.*, 2014; Winfield *et al.*, 2016). The experimentally validated SNVs are a comprehensive and very valuable resource for genome analysis and marker-assisted breeding in rye. They may be converted into other highly flexible SNV assay formats such as Kompetitive Allele-Specific PCR (KASP) to target specific genomic regions, since flanking sequence information is available and conversion rates among platforms are generally high (Semagn *et al.*, 2014).

High-density genetic map and cross-species comparison

Using the Rye600k array, we generated a high-density genetic map as a backbone for anchoring the WGS contigs along the rye genome. In a RIL population derived from an inter-pool cross between the genome reference line Lo7 from the seed parent pool and line Lo225 from the pollen parent pool, 87,820 SNVs were genetically mapped (Table S9). They represented 44,371 Lo7 WGS contigs (covering 158.2 Mbp of sequence) and 3,022 contigs from previous projects (e.g. Haseneyer *et al.*, 2011). The linkage map had a total length of 1,245 cM which is in the same order as other genetic maps in rye (Martis *et al.*, 2013).

Owing to the high marker number on average 42.3 markers cosegregated per locus. The average distance of 0.6 cM between loci indicated a nearly saturated genetic map. Large areas around the centromeres exhibited low recombination rates, leading to extensive linkage blocks which are generally observed in the large cereal genomes (Mayer *et al.*, 2012) (Figure 1). The high-density genetic map served to generate an updated and extended version of the Rye Genome Zipper (Martis *et al.*, 2013) which provides a virtual linear order of rye sequence contigs based on genes in syntenic blocks of the model grass genomes *Brachypodium distachyon*, *Oryza sativa* and *Sorghum bicolor.* The new Rye Genome Zipper links the ordered rye genome sequence with these grass species at high resolution and thus makes a wealth of sequence resources directly accessible for genomic and cross-species analyses.

The comparison of the gene order established in rye with barley (Figure 2) and wheat (Figure S6) indicated a well-conserved genome collinearity between the three species as is evident from large syntenic blocks that are interrupted by breakpoints corresponding to the previously described chromosomal rearrangements (Naranjo and Fernández-Rueda, 1991; Devos *et al.*, 1993; Martis *et al.*, 2013). The large pericentromeric regions with reduced recombination frequency (Figure S7), were conserved between the (sub-)genomes of all three species. In addition to the phylogenetically conserved genetic centromere of 5R, we found a second region on the long arm of this chromosome with reduced recombination frequency relative to wheat and barley (Figure 2, Figure S6, Figure S7), which might be related to a previously described neocentric activity on 5RL (Schlegel, 1987; Manzanero *et al.*, 2000).

Genomic diversity in rye breeding pools and genetic resources

To investigate the genomic diversity within and differentiation between elite breeding pools and GR, we investigated a diverse panel of inbred lines from each of the two heterotic pools and a broad panel of GR with the Rye600k array. The seed parent pool was represented by

38 and the pollen parent pool by 46 inbred lines. The GR comprised 46 individuals from open-pollinated populations mainly from Eastern Europe but also included accessions from Portugal, Canada, U.S.A., a primitive rye from Iran, and three wild *Secale* species (*S. strictum*, *S. sylvestre*, *S. vavilovii*). The seed and the pollen parent pool for hybrid rye breeding were initially developed from the two genetically distant pools Petkus and Carsten through recurrent selection programs after introgression of dominant self-fertility genes which overcame the natural self-incompatibility of rye (Geiger and Miedaner, 2009). Major differences in agronomic traits are observed between these two pools: whereas lines from the seed parent pool typically exhibit high yield performance, good kernel development, tolerance to abiotic stress and high pre-harvest sprouting resistance, lines from the pollen parent pool are characterized by large spikes and good seed setting, but low lodging and pre-harvest sprouting resistance and low yield performance. For broadening the genetic basis of the elite breeding pools a rich reservoir of diversity preserved in GR with favourable alleles for different agronomic traits is available, but the utilization of GR is hampered by strong inbreeding depression due to genetic load and linkage drag with undesired alleles. With the Rye600k array a novel tool is available for a genome-wide detailed characterization of the diversity in different elite and GR gene pools.

To compare the diversity estimates obtained based on WGS data from 10 lines from the two elite pools with estimates from the genotyping array we calculated nucleotide diversity π per site for each pool and for the GR based on 235,460 SNVs of class PHR from the array. Confirming the results from WGS data, nucleotide diversity calculated from the genotyping array was significantly higher in the seed than in the pollen parent pool (0.327 and 0.311; $p < 2.2e^{-16}$), respectively. Both values were higher than the corresponding estimates from WGS data reported above which may be due to a more representative sample size in the genotyping panels and/or due to the filtering process during array construction. The GR exhibited significantly higher nucleotide diversity (0.371; $p < 2.2e^{-16}$) than the two elite breeding pools, which indicates that these accessions harbour allelic diversity not

represented in the elite lines and thus might be of interest for broadening the genetic basis of the elite germplasm. With array data a strong differentiation between the seed and pollen parent pool was observed with an F_{ST} estimate of 0.229. Differentiation between the GR accessions and each of the breeding pools was significantly lower (GR vs. seed parent pool: $F_{ST} = 0.109$; GR vs. pollen parent pool: $F_{ST} = 0.116$; both: $p < 2.2e^{-16}$), suggesting an intermediate position of the GR between the two breeding pools. This was also reflected by a Principal Coordinate Analysis which showed a clear separation between the two breeding pools and a more central position of the GR between the two divergent breeding pools (Figure 3). No clear population structure was observed within each of the three groups. Consistent with results based on simple sequence repeat markers (Fischer *et al.*, 2010; Parat *et al.*, 2015) and with the breeding history of many Eastern European open-pollinated populations, the GR group which comprised many accessions of Eastern European origin had a significantly lower F_{ST} with the seed than with the pollen parent pool.

Genome-wide screens for selection signals

To detect selection signals along the genome differentiating the two rye elite breeding pools and the diverse GR accessions we analysed the 78,731 genetically mapped SNVs. Using Lositan (Antao *et al.*, 2008) we identified F_{ST} outliers in the three pairwise group comparisons, which resulted in 592 SNVs (from 480 contigs) between the seed and pollen parent pools, 1,187 SNVs (from 996 contigs) between the GR accessions and the seed parent pool and 3,420 SNVs (from 2,815 contigs) between the GR accessions and the pollen parent pool. About 40% (237) of the SNV outliers between the two breeding pools were highly differentiated ($F_{ST} > 0.8$) which points towards fixation of different alleles that may contribute to heterosis between the breeding pools. We further calculated the Bayenv2.0 X^TX statistic which is analogous to F_{ST} but accounts for genome-wide covariance of allele frequencies (Günther and Coop, 2013). Since this measure is closely related to F_{ST} we compared the highest 1% X^TX values with the F_{ST} outliers for each of the three pairwise

group comparisons and found a strong overlap of highly differentiated SNVs with both methods. Genome-wide maps of selection signals revealed the genetic differentiation between the rye breeding pools (Figure 4) and each of the breeding pools with the GR group (Figure S8, Figure S9). In all three comparisons outlier SNVs clustered in few distinct genomic regions that may harbour targets of divergent selection.

To demonstrate the usefulness of the new rye genomic tools for assigning putative functions to selection targets, WGS contigs which contained rye gene models and harboured SNVs that were identified as selection candidates from the overlap of the highest 1% X^TX values with the F_{ST} outliers were used for a tBLASTX analysis against the Q-TARO database (Yonemaru *et al.*, 2010). Q-TARO comprises 1,949 cloned and functionally characterized rice genes. The functional characterization of these genes in rice allows hypotheses on the possible roles of their orthologs in rye and gives first insights which genes contribute to the differentiation between rye germplasm pools. In total, 27 rice orthologs could be detected on 22 Lo7 contigs (Table S10). Ten rice orthologs were found for potential selection candidates in the comparison between seed and pollen parent pool, eight between seed parent pool and GR, and 14 between pollen parent pool and GR. Five of them were found in two comparisons. For the 22 Lo7 contigs for which rice orthologs were identified, we calculated nucleotide diversity π (Table S11). As expected, in most cases the highest nucleotide diversity was observed in the GR group. Of the 27 rice orthologs, six affect plant height (*Dwarf 1 gene*, *gid1*, *OsDOG*, *OsGAE1*, *OsGSK2/BIN2*, *OsPH1*). Plant height is a quantitative, highly heritable trait and a major selection target to improve lodging resistance of rye (Geiger and Miedaner, 2009). Strong differentiation in five of the six orthologs governing plant height was observed in the comparison of the seed or pollen parent pools with GR, which are typically much taller than elite inbred lines. This finding is consistent with the large number of genomic regions affecting plant height that were previously detected in a rye introgression library constructed from a seed parent pool inbred line and an Iranian GR (Miedaner *et al.*, 2011; Mahone *et al.*, 2015). Two of the six genes (orthologs of *Dwarf 1*

gene and *OsPH1*) were strongly differentiated between the seed and pollen parent pools (Table S10). The pollen parent pool was mainly derived from the population Carsten's Kurzstroh ("Carsten's short-strawed") and has reduced plant height compared to the taller 'Petkus' population from which the seed parent pool was developed and thus these two genes may contribute to the phenotypic differences between the two pools. The rye ortholog of *OsPH1*, a gibberellic acid (GA)-responsive protein, maps to the long arm of chromosome 5R and may serve as a candidate for the GA_3-insensitive rye dwarfing gene *ct2*, which has been mapped on 5RL (Plaschke *et al.*, 1993). Three of the rice orthologs (*d11*, *OsGPX1*, *SP1)* affect grain size, number of seeds per spikelet or panicle length in rice, respectively (Tanabe *et al.*, 2005; Li *et al.*, 2009b; Passaia *et al.*, 2014). Interestingly, the ortholog controlling the number of seeds per spikelet (*OsGPX1*) was detected in the comparison of seed and pollen parent pools, which strongly differ in their ear morphology, with generally longer ears with many smaller kernels in the pollen parent pool, as opposed to the shorter and more compact ears with larger kernels in the seed parent pool (Figure 5A). Large differences in ear morphology are also observed between the seed parent pool and GR. In this comparison, we identified the rye ortholog of *SP1* which affects panicle elongation and grain size in rice. The rye ortholog of *SP1* is located in the centromeric region of chromosome 4R and might be a candidate for a thousand-kernel weight QTL identified in a rye introgression library (Falke *et al.*, 2009b). As typical examples for differences of candidate selection loci between the germplasm pools haplotypes of two Lo7 contigs which carry the rye orthologs of *SP1* and *OsGPX1/OsGPX3*, respectively, are shown in Figure 5B. A clear difference in haplotype frequencies between the three pools was observed in both contigs. Especially in GR, rare haplotypes are prevalent which also fits with the high nucleotide diversity observed for the selection candidates in GR (Table S11). The rye ortholog of rice gene *DCW11* on chromosome 1R, encoding for a mitochondrial protein phosphatase 2C protein (Fujii and Toriyama, 2008), was differentiated between the seed parent pool and GR. The knockdown of *DCW11* in rice leads to defects in pollen germination ability and reduced seed set which are interesting phenotypes in the context of hybrid

breeding in rye. Hybrid production in rye is based on cytoplasmic-genic male sterility (CMS) in the seed parent pool combined with effective nuclear male fertility restorer genes on the pollen parent side (Geiger and Miedaner, 2009). Intensive selection in the seed parent pool ensures full male sterility in CMS cytoplasm, which may explain the strong differentiation between the seed parent pool and GR for this gene. The *DCW11* ortholog on 1R may correspond to QTL for partial pollen-fertility restoration genes identified in an early inbred line from seed parent pool (Wricke *et al.*, 1993; Miedaner *et al.*, 2000) and in the rye introgression library with the Iranian rye donor accession (Falke *et al.*, 2009a). The other rice orthologs of rye selection candidates have functions in plant development and morphology, abiotic and biotic stress and regulation of multiple physiological processes and thus are interesting targets for follow-up studies to investigate gene classes affected by intensive selection and differentiation of the rye breeding pools.

All selection candidates described above affect phenotypic traits known to differ between the three pools investigated in our study. Our analyses can be seen as a first step towards detecting selection signals in the rye genome and demonstrate the utility of the Rye600k array for population genetic analyses. The genome-wide screens have enabled us to identify candidates with selection signals in the pairwise comparisons of two breeding pools and GR which warrant further research. Since sequence information is available from WGS sequencing, results from such studies can now easily be linked to genomic resources from other grass species, either directly or via the Rye Genome Zipper.

Summary and outlook

We present a whole-genome draft sequence assembly of rye, a diploid crop species of the Triticeae with a 7.9 Gbp genome, which covers most of the non-repetitive portion of the rye genome. Different strategies were employed to anchor the WGS contigs in the rye genome which resulted in the assignment of around half of the 1.58 million contigs to one of the

seven rye chromosomes. For specific regions of interest, even more contigs may be anchored through synteny-based approaches using the wheat and barley draft genome sequences which have become available recently. The rye gene set represented by 27,784 predicted high-confidence gene models will greatly promote transcriptomic approaches and genome-wide functional analysis in rye. The functional genomic analysis of agronomic traits such as abiotic stress tolerance may have implications on other cereals as well and contribute to breeding better varieties for challenging environmental or climatic conditions. Our population genetic analyses revealed insights in the structure and diversity of elite breeding pools and GR in rye. A genome-wide scan for selection signals between the breeding pools and/or GR revealed candidate genes with rice orthologs affecting agronomic traits differing between pools. Candidates detected in the comparison of the seed and pollen parent pools are targets for further investigating the differentiation between these two pools in the context of heterosis. Altogether, the rye whole-genome sequence, the gene models and the high-density genotyping array will enable comparative genomic analyses at unprecedented resolution and open new avenues for genome-based breeding, genome mapping and gene cloning in rye. Making use of sequencing platforms which provide longer reads combined with physical genome mapping in nanochannel arrays (Hastie *et al.*, 2013) may help to further improve the assembly and to explore structural variation in rye in more depth.

Experimental procedures

Plant material, nucleic acid preparation and whole-genome sequencing

A diverse panel of cultivated rye (*Secale cereale* L.) inbred lines from two heterotic groups from a hybrid breeding program (KWS LOCHOW GMBH) was selected for WGS, assembly and variant detection. Six inbred lines (Lo7, Lo90, Lo115, Lo117, Lo176, Lo191) represent the rye seed parent pool, whereas five lines (Lo282, Lo298, Lo310, Lo348, Lo351) originate

from the pollen parent pool. Lines from the seed and pollen parent pools were selfed for 5-6 and 2-3 generations, respectively. A *S. vavilovii* accession was provided by Thomas Miedaner (University of Hohenheim). Genomic DNA of each inbred line was prepared from bulked young leaf tissue from five plants which were pre-tested for homogeneity and from a single plant of *S. vavilovii* using a modified CTAB protocol (Saghai-Maroof *et al.*, 1984). The preparation of sequencing libraries and Illumina (http://www.illumina.com/) sequencing was done by Eurofins Genomics GmbH (http://www.eurofinsgenomics.eu/). For deep sequencing of inbred line Lo7 different sequencing libraries were prepared using standard protocols recommended by the manufacturer or developed by the service provider. Two PE shotgun libraries with insert sizes of 150-350 bp (PE350) and 250-500 bp (PE450), two methylation-filtered paired-end shotgun libraries (100-600 bp, METH), and three long jumping distance (LJD; similar to the MP library protocol from Illumina, but with adaptor-guided ligation of genomic fragments) libraries with insert sizes of ~3 kbp (LJD3), ~8 kbp (LJD8) and ~20 kbp (LJD20) were prepared. For resequencing of the other 10 rye inbred lines and *S. vavilovii*, PE shotgun libraries with insert sizes of 250-500 bp were prepared. All libraries (Table S1) were sequenced on an Illumina HiSeq 2000 sequencer with chemistry v3.0 and the 2 x 100 bp paired-end read module.

Sequence assembly and hierarchical scaffolding

FASTQ sequence data from the Lo7 reference line, 10 rye inbred lines and one *S. vavilovii* accession are archived at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the project numbers PRJEB6214 for the reference inbred line Lo7 and PRJEB6215 for the 11 resequencing data sets. LJD MP sequences of Lo7 were provided with trimmed adapters and removed barcodes and are decoded as PE sequences (read pairs point to each other 'fr'). Prior to data analysis all raw sequence data sets were quality trimmed using 'clc_quality_trim', a subroutine of the software suite CLC Assembly Cell v4.1 (http://www.clcbio.com). As cut-off we used a minimal base quality of 20 and required that at

least 50% of a read is exceeding this threshold. On average, 83.9% of all sequence base pairs and 87.6% of all sequence reads passed the quality trimming. Sequencing quality was checked with FastQC version v0.10.1 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) and the ea-utils command line tool (https://code.google.com/p/ea-utils/). We observed on average a PHRED base quality of ~38.

Sequences were *de novo* assembled using CLC Assembly Cell v4.1. The quality trimmed reads of short Lo7 libraries (PE350, PE450 and METH) were integrated in the assembly process. The singleton reads of all libraries were merged to a joint dataset and integrated as one FASTQ file. The assembly was calculated in PE mode using calculated minimum and maximum distances of paired read libraries. We used a base pair distance ranging from 150 to 350, 250 to 500 and 100 to 600, respectively, for the short libraries PE350, PE450 and METH. These distances were based on the estimated fragment length peak for each PE library. A minimum contig length of 300 bp was used. Subsequently, the constructed WGS assembly was screened for contaminations. Therefore, we downloaded the NCBI nt nucleotide collection (date: 10th January 2013) and performed a BLASTN analysis to screen for non-plant species using a sequence identity threshold of 90% (Zhang *et al.*, 2000). In total, 1,782 WGS contigs were removed from the WGS assembly with a cumulative size of 828.2 kbp. In these discarded contigs hits to 783 different non-plant species were detected.

To gain higher specificity of the constructed WGS assembly the Lo7 contigs were analysed by chromosome arm assignment (CarmA). CarmA is a bioinformatics tool that uses the resource of flow-sorted chromosome (arm) sequence sets to determine the chromosomal origin of unassigned query sequences, like WGS contigs. It was used before to aid the gene based assemblies of barley and wheat (Mayer *et al.*, 2012; Mayer *et al.*, 2014). For the CarmA approach about 1x coverage of 454 chromosome survey sequences (CSS) from each of the seven flow-sorted rye chromosomes was available (Martis *et al.*, 2013) which allowed the assignment of Lo7 WGS contigs and LJD MP reads from the LJD3, LJD8 and

LJD20 libraries to chromosomes as input for the virtual gene order (Genome Zipper) approach and thus reduced scaffolding ambiguities. The principle of CarmA is based on a homology search for each query sequence separately against the CSS bins from each of the seven chromosomes. The homology search was done via vmatch (http://www.vmatch.de) under high stringency conditions: -d-p -l 100 or 90 (forward and reverse strand search, perfect match, minimum hit length 100 bp for contigs, 90 bp for LJD reads). Queries with hits to more than one bin were assigned to the highest coverage bin if the signal noise ratio to the second lower bin was ≥ 1.5. Under these parameter settings 67% of all Lo7 contig basepairs, corresponding to 50.9% of all contigs, could be allocated to one of the seven chromosomes (Table S12). CarmA was also used for the assignment of 38.4 mio LJD mate-pairs to their chromosomal origin.

Ordering and continuation of WGS contigs into scaffolds is feasible if large distance MP libraries are available. As described above, all Lo7 WGS contigs and LJD MP reads were processed by CarmA to establish chromosome assignment. To construct a scaffolded genome reference sequence we implemented a hierarchical scaffolding approach (Figure S10) as is widely used especially in large and complex plant genomes (International Brachypodium Initiative, 2010; Schatz *et al.*, 2010; Chapman *et al.*, 2015). Among various scaffolding tools we decided to use SSPACE (Boetzer *et al.*, 2011), which had good performance results in a broad evaluation study of scaffolding tools (Hunt *et al.*, 2014) and application in other plant genome studies (Beier *et al.*, 2015). To reduce the risk of chimeric scaffolding where contigs from distant genome regions (e.g. different chromosomes) are erroneously linked, we integrated the information of genetic map positions obtained from ~15.000 rye DArT-seq markers that were anchored to the Lo7 WGS assembly by BLASTN. BLASTN was performed using an e-value of 1e-5, perc_identity of 98% and requiring a minimal overlap of the contig with the DArT marker sequence length of 90%. If a WGS contig was assigned by CarmA to the same rye chromosome as a genetically mapped DArT-seq marker the required minimal overlap was reduced to 85%. In total, 3,443 WGS contigs were

assigned to genetic map positions using DArT-seq markers. If the genetic distance between WGS contigs exceeded 1 cM, a putative physical link was discarded. For hierarchical scaffolding we first used the shortest MP libraries (LJD3) that were characterized by a more accurate estimation of MP distances. Subsequently, we included the longer MP libraries with increasing distances (LJD8 and LJD20). Each chromosome was assessed separately. A prerequisite of the scaffolding is an alignment of the MPs to the WGS contigs that was performed using BWA version 0.7.8 (Li and Durbin, 2009). To integrate an external alignment in the scaffolding process of SSPACE we used an adapted version of the sam2TAB.pl scripts that were published with the barley BAC pipeline (Beier *et al.*, 2015). All WGS contigs represented in the genetic map and in the updated Rye Genome Zipper (see below) were used to assign a genetic map position to scaffolds. The assignment was performed using MegaBLAST (Zhang *et al.*, 2000) with stringent parameter settings of 'word_size' of 150, 'perc_identity' of 99.5% and 'evalue' of 1e-60. In total, 42,197 scaffolds could be anchored. From these, only a minor proportion of <0.3% (122) were classified as chimeric scaffolds originating from different rye chromosomes. The WGS scaffolds of Lo7 are deposited as digital object identifier *DOI 1* (see Data Availability).

Transposon composition in the rye genome

Transposons were detected and classified by a homology search against the REdat_9.0_Poaceae section of the PGSB transposon library (Spannagl *et al.*, 2016b). The program vmatch (http://www.vmatch.de) was used as a fast and efficient matching tool suited for large and highly repetitive genomes with the following parameters: identity ≥ 70%, minimal hit length 75 bp, seedlength 12 bp (exact command line: -d -p -l 75 -identity 70 -seedlength 12 -exdrop 5). The vmatch output was filtered for redundant hits via a priority based approach, which assigned higher scoring matches first and either shortened (<90% coverage and ≥ 50 bp rest length) or removed lower scoring overlaps, leading to an overlap-

free annotation. The analyses were performed both with the WGS assembly and with 800 Mbp of randomly selected Illumina reads.

High-confidence gene set

S. cereale gene structures were predicted using a reference-based approach. Available annotations of closely related plant genomes, namely *Brachypodium distachyon* v1.2 (International Brachypodium Initiative, 2010), *Sorghum bicolor* v1.4 (Paterson *et al.*, 2009), *Oryza sativa* MSU7 (Kawahara *et al.*, 2013), *H. vulgare* (Mayer *et al.*, 2012), *Triticum urartu* (Ling *et al.*, 2013), and *Aegilops tauschii* (Jia *et al.*, 2013), as well as three data sets of rye ESTs and transcripts (Haseneyer *et al.*, 2011; Banaei-Moghaddam *et al.*, 2015), were used as a template to model the genes by aligning the rye contigs against the protein and transcript sequences and identifying potential gene structures using the GenomeThreader gene prediction software (Gremme *et al.*, 2005). Gene candidates were filtered to remove models that translate for peptide sequences with internal stop codons. Redundant predicted genes from the different sources of information were clustered based on their genomic coordinates by using Cuffcompare, which is part of the Cufflinks package (Roberts *et al.*, 2011). For all non-redundant coding regions the open reading frames (ORFs) and peptide sequences were predicted using Transdecoder (http://transdecoder.sf.net) and the structure with the longest continuous amino acid sequence was selected for further analysis. The accuracy of the predicted genes was further increased by filtering known repeats from the predicted gene set using BLASTX searches against the Triticeae repeat database TREP (Wicker *et al.*, 2002) and against all known repetitive sequences in the reference genomes used as templates. In addition, we removed gene models with overlapping coordinates from the final data set.

To identify high-confidence protein-coding genes the predicted gene models were compared against the protein data sets of the reference genomes mentioned above and the best-matching reference protein selected as representative template sequences. Based on the similarity to the respective representative template sequence and the maximum coverage of the template sequences, the gene models were classified in five confidence classes: three high-confidence classes (HC1 to HC3), one low confidence class (LC), and one class containing pseudogenes and gene fragments (PGGF). The high-confidence classes contain only gene models with matches to one of the following reference genomes: *Brachypodium*, *Sorghum*, rice, barley, *T. urartu*, or *Ae. tauschii*, while the low-confidence class contains gene models with matches to rye ESTs/transcripts only. In the HC1 class the genes show a protein coverage greater than 70% of the representative template sequence, while in the HC2 class the coverage lies between 50% and 70%, and in the HC3 class between 30% and 50%. The genes of the LC class cover over 70% of the tagged rye ESTs/transcripts with a minimal length of 150 amino acids. The PGGF class contains genes covering less than 30% of the representative template sequence and genes without sequence homology to any of the reference species proteins. The 27,784 genes grouped in the first three confidence classes (HC1 - HC3) were referenced here as the rye gene set. The assignment of these rye gene models to Lo7 WGS contigs is available under *DOI 2* (see Data Availability).

Gene annotation

To provide insights into the putative function of genes, the rye gene set was processed using the 'Automatic assignment of human readable descriptions' (AHRD) pipeline (version 2.0; https://github.com/groupschoof/AHRD/) which integrates three types of database evidences to describe putative gene functions using standard nomenclature. A human readable protein description was inferred for 23,274 (83.24%) of the rye genes using BLASTP hits to TAIR10, Swiss-Prot and TrEMBL databases. AHRD annotation results are available under *DOI 2 (see Data Availability)*. Gene ontology (GO) terms were assigned to the rye gene set using the

Blast2GO version 2.8 (Götz *et al.*, 2008) pipeline in standard settings (except the e-value threshold for the annotation step which was set to 1e-3) identifying sequence similarity to other sequenced species using NRPEP, the NCBI non-redundant protein sequences database. In total, 16,275 rye genes were mapped, 12,716 were annotated and 3,247 were assigned to an EC number. Blast2GO annotation results are available under *DOI 2 (see Data Availability)*. A BUSCO analysis was performed according to Simão *et al.* (2015) using the BUSCO plant gene set.

Read alignment and variant calling

The genome-wide diversity of rye was investigated using WGS resequencing data from 10 rye inbred lines and one *S. vavilovii* accession. FASTQ sequences of quality trimmed paired end data were aligned to the Lo7 genome reference WGS assembly using BWA version 0.7.0 (Li and Durbin 2009). For read alignment a minimal base quality of 20 and a minimal mapping quality of 13 was required. The constructed SAM read alignment files were converted into the sorted binary BAM format using SAMtools version 0.1.18 (Li *et al.*, 2009a). Duplicated reads were removed using the 'rmdup' function of SAMtools. On average, ~9% of all PE read pairs were detected as duplicated reads and removed from further processing steps. Subsequently, this data resource was used for variant discovery using the MPILEUP format constructed by SAMtools (Li, 2011) that was further processed by VCFtools version 0.1.11 (Danecek *et al.*, 2011) resulting in the raw variant data file in the VCF format. Filtering of detected variant positions was performed with the following criteria: First, we discarded ambiguous positions with multiple alleles in the Lo7 reference line (e.g. heterozygous positions). Second, we filtered for homozygous positions that passed the additional requirements of minimal read coverage of five reads per genotype and a minimal quality score of 100. From the filtered set of variants, we calculated nucleotide diversity π and F_{ST} (Weir and Cockerham, 1984) on a per-site basis using VCFtools version 0.1.12a. A two-sided Wilcoxon rank sum test (Wilcoxon, 1945) was performed to test for differences in

nucleotide diversity between pools. The effects of SNVs and InDels were annotated with CooVar version 0.07 (Vergara *et al.*, 2012). The SNVs detected in the 10 resequenced lines and *S. vavilovii* were used to construct a phylogenetic tree using SNPhylo (Lee *et al.*, 2014). As filter criteria we used a minor allele frequency of 0.1, a maximal missing rate of 0.1 and a minimal depth of coverage of 3.

Genome positions and allele calls for the variants derived from the resequencing data can be downloaded as VCF files under *DOI 3* (see Data Availability). The functional annotations of variant positions are provided in separate GFF files for each genotype (*DOI 3;* see Data Availability).

Gene loss

Gene loss was investigated by the coverage breadth criterion using read alignments of the 10 resequenced rye inbred lines and *S. vavilovii*. A gene model was defined as absent if less than 5% of the gene length was covered with reads. Subsequently, a gene model was considered as strong candidate for divergent presence/absence variation (PAV) patterns, if more than 25% of the studied genotypes (>3) were flagged as absent. Genes that fulfilled this stringent criterion were regarded as candidates that most likely underwent gene loss. A list of 2,934 gene models which were missing in at least one genotype is deposited under the digital object identifier *DOI 2* (see Data Availability).

Development of the Rye600k Affymetrix genotyping array

For the development of the Rye600k Affymetrix Axiom HD genotyping array (http://www.affymetrix.com) we used 8.6 million informative SNVs derived from 10 resequenced cultivated rye lines and one accession of the wild progenitor *S. vavilovii* with unique positions in the Lo7 genome reference sequence. For selection of the final SNV dataset we followed the Axiom myDesign Custom Array recommendation giving priority to

SNVs that were observed in multiple lines, had sufficient sequence depth, high quality, no additional polymorphism within the flanking 30 bp marker sequences, were not classified as repetitive sequence and had Minor Allele Frequency (MAF) larger 0.05. In addition, 10,250 Infinium iSelect markers from a custom Rye16k Illumina genotyping array (Auinger *et al.*, 2016) were integrated in the Rye600k array. Initially, the complete set of 8.6 million SNVs was processed by the quality control of Affymetrix to classify if a SNV position is 'recommended', 'neutral' or 'not possible' for marker design. It was ensured that no adjacent SNVs occurred within recommended probe sequences. For an optimal selection of SNVs within the design process in a first phase additional information sets were pre-calculated and subsequently integrated in the evaluation pipeline using a custom PERL script. Figure S11A shows the utilized information sets. The information provided by the CarmA assignment as well as by DArT markers for Lo7 contigs were used to achieve a uniform representation of SNVs across chromosomes. *K*-mer frequencies were pre-calculated with the tool Kmasker (Schmutzer *et al.*, 2014) using a 21-mer index that was constructed based on a 10-fold representation of the Lo7 PE reads (PE350). SNVs with *k*-mer frequencies >10 were excluded. With this we intended to avoid selecting ambiguous SNVs from repetitive regions. In the second phase, information gained in phase I was assessed in four steps using the following filtering criteria: 1) SNVs from coding regions; 2) high stringency; 3) medium stringency, and 4) low stringency. Details of the stringency settings are given in Figure S11B. Finally, 600,843 SNVs were selected for marker design on the Rye600k array which represent 242,349 different Lo7 contigs and 21,479 rye gene models that have a variant site within their coding sequence (88,487 SNV markers are linked with gene models). SNVs used for marker design were used to construct a phylogenetic tree using SNPhylo (Lee *et al.*, 2014) with identical parameter settings as described above.

Variant validation by genotyping

For the experimental validation of SNVs, a total of 38 elite inbred lines from the seed parent pool, 46 elite inbred lines from the pollen parent pool, 43 diverse accessions derived from rye genetic resources (GR) including three accessions from wild species were genotyped with the Rye600k array (Table S13). For construction of the genetic linkage map 131 recombinant inbred lines (RIL) from a cross between rye lines Lo7 and Lo225 were genotyped. From GR we extracted DNA from a single plant per accession and from elite lines and RILs, DNA was extracted from a bulked sample of 8-10 plants using a protocol modified after Doyle and Doyle (1987). Per DNA sample 200 ng were processed on an Affymetrix Gene Titan platform by the Animal Breeding group at Technical University of Munich (Germany) according to manufacturer's protocols (http://www.affymetrix.com). Raw intensity data were analysed in one batch with the Affymetrix Genotyping Console (v. 4.2.0.26) following the manufacturer's best practice guidelines to obtain genotype calls. During genotype calling the level of inbreeding of the samples was taken into account by setting appropriate values for inbred penalties (ranging from a value of 2 for highly heterozygous single plants from rye GR to 8 for mostly homozygous inbred lines at higher selfing generations). Categorization of variants into one of the six Affymetrix SNV categories "Poly High Resolution", "No Minor Homozygote", "Mono High Resolution", "Off-Target Variant", "Call Rate Below Threshold" and "Other" was performed with the R package SNPolisher (v. 1.3.6.6) (Gao *et al.*, 2012) according to the Axiom Genotyping Solution Data Analysis Guide.

Genetic mapping

For construction of a high-density genetic map, 131 RILs in selfing generation F7 were used. Genotype calls from the two SNV categories "Poly High Resolution" and "Off-Target Variant" were filtered for markers polymorphic between the parents of the mapping population (Lo7 × Lo225), allowing a maximum of 5% missing values and a maximum of 5% heterozygous

calls per SNV, which resulted in 115,241 SNVs. This dataset was processed through a script from the POPSEQ pipeline to identify groups of SNVs with identical segregation patterns based on the Hamming distance (Mascher *et al.*, 2013a). All remaining heterozygous genotype calls were set to missing values for constructing the linkage map. A first genetic map was calculated with the R package ASMap v. 0.3-3 (Taylor and Butler, 2014), by using the function mstmap with the following parameters: pop.type="RIL6", dist.fun="kosambi", objective.fun="COUNT", p.value=1e-20, noMap.dist=20, noMap.size=2, miss.thresh=0.05. The resulting linkage groups were assigned to the seven rye chromosomes based on previously mapped markers (Martis *et al.*, 2013) and unlinked small groups with only a few markers were discarded. Two RILs exhibited very high numbers of crossovers and were excluded from further analyses. In two consecutive rounds of mapping using the same parameters as stated above, markers which led to double-crossovers were identified using the function statMark in the ASMap R package and discarded before final map construction. The final genetic linkage map contained 10,196 loci. All SNVs from the initial dataset which had a Hamming distance of 0 with only one of the mapped SNVs and which could be assigned to a unique map position were inserted into the map, resulting in a genetic map with 87,820 SNVs representing 44,371 Lo7 WGS contigs and 3,022 contigs from previous studies (e.g. Haseneyer *et al.*, 2011). The genetic map is available under *DOI 2* (see Data Availability).

Genome colinearity and synteny across rye, barley and wheat

The rye WGS assembly was aligned to coding sequences (CDS) of barley and wheat gene models with GMAP (Wu and Nacu, 2010). We used the high-confidence gene models annotated on the WGS assembly of barley cv. Morex (Mayer *et al.*, 2012) and on the chromosome survey sequence assembly of wheat cv. Chinese Spring (Mayer *et al.*, 2014). Only the best alignment of each CDS was considered, requiring at least 90% alignment identity and 50% coverage of the CDS. Genetic positions of CDS were taken from the

POPSEQ genetic maps of barley (Mascher *et al.*, 2013a) and wheat (Chapman *et al.*, 2015) and plotted against the map locations of the rye WGS contigs the CDS were aligned to. The positions of genetic centromeres in rye chromosomes were determined by tabulating the number of anchored sequence contigs in 1 and 5 cM bins. Centromere positions are given in Figure S7. Aggregation and plotting of positional information was performed with standard functions of the R statistical environment (R Core Team, 2015) and the R package "data.table" (https://cran.r-project.org/web/packages/data.table/index.html).

Updated version of the Rye Genome Zipper

The previously described framework of genetic map data, chromosomal gene content of rye, conserved synteny information to model grass genomes, rye EST assembly information and barley full-length cDNAs (Haseneyer *et al.*, 2011; Matsumoto *et al.*, 2011; Martis *et al.*, 2013) was extended to ~87k markers as described earlier (Mayer *et al.*, 2011). The complete Rye Genome Zipper v2 data sets for the seven rye chromosomes are available under *DOI 4* (see Data Availability).

Circos plot of rye genome structure and diversity

Genome structure and diversity distribution within the 10 resequenced rye inbred lines and *S. vavilovii* was visualized using a concatenated version of the Lo7 rye assembly (Figure 1). The assembly was ordered using the genetic map and the updated Rye Genome Zipper. Heterozygosity and gene density were analysed per 1 cM. To estimate the heterozygosity we performed a read alignment of Lo7 reads against the Lo7 WGS assembly and calculated the number of sites with multiple alleles using a minimal read coverage of ten and a minimal quality score of 100. Gene density, marker density and SNV counts were plotted in 500 kbp windows. For greater clarity, we calculated the 98 quantile within each single track and applied it as maximum per track. Calculated maximum values were 558 (heterozygosity), 78

(gene density per 1cM) and 472 (marker density per 500 kbp). The resulting figure was constructed using Circos (v.0.67) (Krzywinski *et al.*, 2009).

Population genetic analyses

The population genetic analyses were performed with biallelic SNVs of class PHR from the Rye600k array. A common dataset of 235,460 SNVs was available for the 130 samples from the seed and pollen parent pool and GR. From this dataset nucleotide diversity π and F_{ST} (Weir and Cockerham, 1984) on a per-site basis were calculated using VCFtools version 0.1.12a. Differences in nucleotide diversity and F_{ST} between pools were tested with the two-sided Wilcoxon rank sum test (Wilcoxon, 1945). For Principal Coordinate Analysis based on Rogers' distances (Rogers, 1972) in the R package ade4 (Dray and Dufour, 2007), data were further filtered for maximum 5% missing data and minor allele frequency > 0.01, resulting in 179,660 SNVs. Missing values were imputed by sampling from the marginal allele distribution of a marker using the codeGeno function of the R package Synbreed (Wimmer *et al.*, 2012).

In order to detect genomic regions under selection in rye we scanned 78,731 genetically mapped SNVs of class PHR of the Rye600k array for selection signals using Lositan (Antao et al., 2008) and Bayenv2.0 (Günther and Coop, 2013). By searching for SNVs that are strongly differentiated between two populations, candidate regions and genes for positive selection can be detected. Lositan uses a FST outlier detection method (Beaumont and Nichols, 1996). To find candidate SNVs with Lositan we analysed three population pairs: seed vs. pollen parent pool, seed parent pool vs. GR and pollen parent pool vs. GR material. For each population comparison three runs with 200,000 simulations were performed and only SNVs that were identified as outliers in all three runs were considered as selection candidates. We also used Bayenv2.0 to detect SNVs that are strongly differentiated between the three population pairs. Bayenv2.0 calculates standardized allele frequencies at each

SNP by removing unequal SNP sampling variances and covariance among populations. The resulting X^TX statistic therefore accounts for a shared population history and sampling noise and allows identifying loci with an unusually high allele frequency variance among populations which may reflect differential selection. We estimated the covariance matrix that reflects the population and kinship structure based on 100,000 MCMC iterations. For each population comparison we conducted 100,000 iterations and compared the top 1% X^TX values with the Lositan results to validate candidate SNVs.

To characterize possible selection candidates the WGS contigs which contained rye gene models and harboured SNVs identified as selection candidates from the overlap of the highest 1% X^TX values with the F_{ST} outliers were used for a tBLASTX analysis against the Q-TARO database with 1,949 cloned and functionally characterized rice genes (Yonemaru *et al.*, 2010). Hits with an e-value <$1e^{-40}$ and a minimum overlap of 40 amino acid residues (120 bp) are reported in Table S10. For 22 contigs which gave a hit with the Q-TARO database we calculated population-specific nucleotide diversity π using all PHR SNVs available for these contigs (Table S11). For two Lo7 contigs containing genes that may contribute to phenotypic differences between germplasm pools we show graphical representations of population-specific haplotypes. Only the more frequent haplotypes which occurred in at least four individuals (~10% frequency) in one of the three groups are shown.

Data availability

For comparative analysis we established a stand-alone BLAST web server similar to the IPK Barley Blast Server (Spannagl *et al.*, 2016a) using ViroBLAST (Deng *et al.*, 2007). For direct application of analyses based on sequence homology it contains the WGS assembly and the rye gene models. The IPK Rye Blast Server is accessible under http://webblast.ipk-gatersleben.de/ryeselect/.

The WGS scaffolds of Lo7 (version 2) are deposited as digital object identifier *DOI 1* (see below). A list of 2,934 gene models with divergent PAV patterns, the assignment of WGS contigs and scaffolds to rye chromosomes (based on CarmA, 88k genetic map and Rye Genome Zipper), the assignment of 27,784 rye gene models to Lo7 WGS contigs, and the 88k genetic map of RIL population Lo7 × Lo225 are deposited as *DOI 2* (see below). VCF files and related functional annotations of variant positions are accessible for download under *DOI 3* (see below). The updated version of the Rye Genome Zipper (v2) is deposited as digital object identifier *DOI 4* (see below). All DOIs were constructed using the tool e!DAL (Arend *et al.*, 2014) and stored in the Plant Genomics and Phenomics Research Data Repository – PGP (Arend *et al.*, 2016).

List of DOIs:

DOI 1: http://dx.doi.org/10.5447/IPK/2016/56

DOI 2: http://dx.doi.org/10.5447/IPK/2016/57

DOI 3: http://dx.doi.org/10.5447/IPK/2016/13

DOI 4: http://dx.doi.org/10.5447/IPK/2016/58

Accession numbers

All sequence data exploited in this study is deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena/). WGS data of the rye reference inbred line Lo7 is deposited under study accession PRJEB6214 with the sample accession number ERS446995. Resequencing data of the 10 rye inbred lines and *S. vavilovii* are stored under study accession PRJEB6215 with sample accession numbers ERS455621 to ERS455631. The Lo7 WGS assembly is stored under project study number PRJEB13501 with the sample accession number ERS1115868. The WGS contigs of this study (FKKI010000001-FKKI011581707) have the assigned analysis id ERZ291745. Information on the Affymetrix

Rye600k array including SNV IDs, probe sets, and alleles can be retrieved from NCBI GEO, platform GPL22066 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL22066).

Competing interests

The authors declare no competing financial interests.

Acknowledgements

We are grateful to Thomas Miedaner (University of Hohenheim) and the IPK Gatersleben Genebank for providing seeds of rye genetic resources; to Sylwia Schepella and Stefan Schwertfirm (Technical University of Munich, Plant Breeding) for excellent technical assistance; to Ruedi Fries, Hubert Pausch and Bettina Hayn (Technical University of Munich, Animal Breeding) for processing of the Rye600k arrays; and to Nils Stein and Andreas Houben (IPK Gatersleben) for sharing sequence data. We thank Chris Ulpinnis for help in the construction of Circos graphics, Sebastian Beier for initial quality assessment of sequencing data, Anne Fiebig and Doreen Stengel (IPK Gatersleben) for assistance in data handling (submission of sequence data sets to EBI/ENA). This study was funded by the German Federal Ministry of Education and Research (BMBF) within project RYE-SELECT (Grant IDs 0315946A-E).

Short supporting information legends

Supplementary figures are available in file RyeGenome_SupplementaryFigures.pdf.

Supplementary tables are available in file RyeGenome_SupplementaryTables.pdf or as separate Excel files for large tables (Tables S6, S7, S10).

Figure S1. Comparisons of diversity between two different breeding pools and *S. vavilovii.*

Figure S2. Comparison of rye genes with non-synonymous SNVs (nsSNVs) in the seed and pollen parent pools.

Figure S3. Comparison of 2,934 gene model candidates showing presence/absence variation.

Figure S4. Genotype proportion in the complete variant data set and the selected set of the Rye600k array.

Figure S5. Phylogenetic tree constructed (A) from complete set of chromosome assigned SNVs and (B) from the Rye600k high density array.

Figure S6. Collinearity between the genetic maps of rye and wheat.

Figure S7. Position of genetic centromeres in the rye genetic map.

Figure S8. Genome-wide map of selection signals between the seed parent pool and genetic resources.

Figure S9. Genome-wide map of selection signals between the pollen parent pool and genetic resources.

Figure S10. Hierarchical scaffolding scheme.

Figure S11. Process and parameter settings of the Rye600k array design.

Table S1. EBI/ENA sequence information.

Table S2. Statistics of the WGS sequencing and read quality processing of reference line Lo7.

Table S3. High-confidence gene set of rye.

Table S4. BUSCO analysis of genome (WGS assembly), protein and transcript data sets.

Table S5. Transposable element composition of the rye genome in the Lo7 WGS assembly and in 800 Mbp of random Illumina reads.

Table S6. Statistics for resequencing results and read alignment statistics of the 10 rye inbred lines and *S. vavilovii.*

Table S7. Functional annotation of SNVs discovered in 10 resequenced rye inbred lines and *S. vavilovii.*

Table S8. Classification of SNVs on the Rye600k array according to Affymetrix SNV categories.

Table S9. Overview statistics for the Lo7 × Lo225 high-density genetic map

Table S10. Rice orthologs for selection candidates.

Table S11. Nucleotide diversity π for 22 contigs harbouring SNVs which were under selection.

Table S12. Assignment of WGS contigs and scaffolds to rye chromosomes based on CarmA, 88k genetic map and Rye Genome Zipper.

Table S13. List of rye inbred lines and accessions genotyped with the Rye600k array for genome-wide selection screens.

References

Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A. and Luikart, G. (2008) LOSITAN: A workbench to detect molecular adaptation based on a Fst -outlier method. *BMC Bioinformatics*, **9**, 1-5.

Arend, D., Junker, A., Scholz, U., Schüler, D., Wylie, J. and Lange, M. (2016) PGP repository: a plant phenomics and genomics data publication infrastructure. *Database*, **2016**.

Arend, D., Lange, M., Chen, J., Colmsee, C., Flemming, S., Hecht, D. and Scholz, U. (2014) e!DAL - a framework to store, share and publish research data. *BMC Bioinformatics*, **15**, 214.

Auinger, H.-J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., et al. (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.) *Theor. Appl. Genet.*, **accepted**.

Banaei-Moghaddam, A.M., Martis, M.M., Macas, J., Gundlach, H., Himmelbach, A., Altschmied, L., Mayer, K.F.X. and Houben, A. (2015) Genes on B chromosomes: Old questions revisited with new tools. *BBA-Gene Regul. Mech.*, **1849**, 64-70.

Bartoš, J., Paux, E., Kofler, R., Havránková, M., Kopecký, D., et al. (2008) A first survey of the rye (*Secale cereale*) genome composition through BAC end sequencing of the short arm of chromosome 1R. *BMC Plant Biology*, **8**, 1-12.

Beaumont, M. and Nichols, R. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological Sciences*, **263**, 1619-1626.

Beier, S., Himmelbach, A., Schmutzer, T., Felder, M., Taudien, S., et al. (2015) Multiplex sequencing of bacterial artificial chromosomes for assembling complex plant genomes. *Plant Biotechnol. J.*, **14**, 1511-1522.

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

Chao, S., Dubcovsky, J., Dvorak, J., Luo, M.-C., Baenziger, S.P., et al. (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (Triticum aestivum L.). *BMC Genomics*, **11**, 1-17.

Chapman, J.A., Mascher, M., Buluç, A., Barry, K., Georganas, E., et al. (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.*, **16**, 26.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., et al. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.

Deng, W., Nickle, D.C., Learn, G.H., Maust, B. and Mullins, J.I. (2007) ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics*, **23**, 2334-2336.

Devos, K.M., Atkinson, M.D., Chinoy, C.N., Francis, H.A., Hartcourt, R.L., et al. (1993) Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor. Appl. Genet.*, **85**, 673-680.

Doyle, J.J. and Doyle, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11-15.

Dray, S. and Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**, 1-20.

Esch, M., Chen, J., Colmsee, C., Klapperstück, M., Grafahrend-Belau, E., Scholz, U. and Lange, M. (2015) LAILAPS: the plant science search engine. *Plant Cell Physiol.*, **56**, e8-e8.

Falke, K.C., Wilde, P. and Miedaner, T. (2009a) Rye introgression lines as source of alleles for pollen-fertility restoration in Pampa CMS. *Plant Breeding*, **128**, 528-531.

Falke, K.C., Wilde, P., Wortmann, H., Geiger, H.H. and Miedaner, T. (2009b) Identification of genomic regions carrying QTL for agronomic and quality traits in rye *Secale cereale* introgression libraries. *Plant Breeding*, **128**, 615-623.

Fischer, S., Melchinger, A.E., Korzun, V., Wilde, P., Schmiedchen, B., et al. (2010) Molecular marker assisted broadening of the Central European heterotic groups in rye with Eastern European germplasm. *Theor. Appl. Genet.*, **120**, 291-299.

Flavell, R.B., Bennett, M.D., Smith, J.B. and Smith, D.B. (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochemical Genetics*, **12**, 257-269.

Fowler, D.B. and Carles, R.J. (1979) Growth, development, and cold tolerance of fall-acclimated cereal-grains. *Crop Sci.*, **19**, 915-922.

Fujii, S. and Toriyama, K. (2008) *DCW11*, *down-regulated gene 11* in CW-type cytoplasmic male sterile rice, encoding mitochondrial protein phosphatase 2C is related to cytoplasmic male sterility. *Plant Cell Physiol.*, **49**, 633-640.

Gao, H., Webster, T., Pirani, A., Zhan, Y., Lu, Y. and Shen, M.-M. (2012) SNPolisher: tools for SNP classification, visualization and OTV genotyping (R package version 1.3.6.6).

Geiger, H.H. and Miedaner, T. (2009) Rye (*Secale cereale* L.). In *Cereals* (Carena, M.J. ed. New York, U.S.A.: Springer, pp. 157-181.

Glémin, S., Clément, Y., David, J. and Ressayre, A. (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet.*, **30**, 263-270.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420-3435.

Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, **47**, 965-978.

Günther, T. and Coop, G. (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205-220.

Haseneyer, G., Schmutzer, T., Seidel, M., Zhou, R., Mascher, M., et al. (2011) From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol.*, **11**, 131.

Hastie, A.R., Dong, L.L., Smith, A., Finklestein, J., Lam, E.T., et al. (2013) Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. *PLOS ONE*, **8**, e55864.

Hunt, M., Newbold, C., Berriman, M. and Otto, T.D. (2014) A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.*, **15**, 1-15.

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

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., et al. (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91-95.

Jones, J.D.G. and Flavell, R.B. (1982) The structure, amount and chromosomal localisation of defined repeated DNA sequences in species of the genus *Secale*. *Chromosoma*, **86**, 613-641.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., et al. (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639-1645.

Lee, T.-H., Guo, H., Wang, X., Kim, C. and Paterson, A.H. (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, **15**, 1-6.

Li, H. (2011) Improving SNP discovery by base alignment quality. *Bioinformatics*, **27**, 1157-1158.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., et al. (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Li, S., Qian, Q., Fu, Z., Zeng, D., Meng, X., et al. (2009b) *Short panicle1* encodes a putative PTR family transporter and determines rice panicle size. *Plant J.*, **58**, 592-605.

Li, Y., Haseneyer, G., Schön, C.-C., Ankerst, D., Korzun, V., Wilde, P. and Bauer, E. (2011) High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale* L.) genes involved in frost response. *BMC Plant Biol.*, **11**, 6.

Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., et al. (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, **496**, 87-90.

Mahone, G.S., Frisch, M., Bauer, E., Haseneyer, G., Miedaner, T. and Falke, K.C. (2015) Detection of donor effects in a rye introgression population with genome-wide prediction. *Plant Breeding*, **134**, 406-415.

Manzanero, S., Puertas, M.J., Jiménez, G. and Vega, J.M. (2000) Neocentric activity of rye 5RL chromosome in wheat. *Chromosome Research*, **8**, 543-554.

Martis, M.M., Zhou, R., Haseneyer, G., Schmutzer, T., Vrana, J., et al. (2013) Reticulate evolution of the rye genome. *Plant Cell*, **25**, 3685-3698.

Mascher, M., Jost, M., Kuon, J.-E., Himmelbach, A., Aßfalg, A., Beier, S., Scholz, U., Graner, A. and Stein, N. (2014) Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biol.*, **15**, R78.

Mascher, M., Muehlbauer, G.J., Rokhsar, D.S., Chapman, J., Schmutz, J., et al. (2013a) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.*, **76**, 718-727.

Mascher, M., Richmond, T.A., Gerhardt, D.J., Himmelbach, A., Clissold, L., et al. (2013b) Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.*, **76**, 494-505.

Matsumoto, T., Tanaka, T., Sakai, H., Amano, N., Kanamori, H., et al. (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.*, **156**, 20-28.

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

Mayer, K.F.X., Rogers, J., Dolezel, J., Pozniak, C., Eversole, K., et al. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, **345**, 1251788.

Mayer, K.F.X., Waugh, R., Brown, J.W.S., Schulman, A., Langridge, P., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711-716.

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., et al. (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737-740.

Middleton, C.P., Stein, N., Keller, B., Kilian, B. and Wicker, T. (2013) Comparative analysis of genome composition in *Triticeae* reveals strong variation in transposable element dynamics and nucleotide diversity. *Plant J.*, **73**, 347-356.

Miedaner, T., Glass, C., Dreyer, F., Wilde, P., Wortmann, H. and Geiger, H.H. (2000) Mapping of genes for male-fertility restoration in 'Pampa' CMS winter rye (*Secale cereale* L.). *Theor. Appl. Genet.*, **101**, 1226-1233.

Miedaner, T., Hübner, M., Korzun, V., Schmiedchen, B., Bauer, E., Haseneyer, G., Wilde, P. and Reif, J. (2012) Genetic architecture of complex agronomic traits examined in two testcross populations of rye (*Secale cereale* L.). *BMC Genomics*, **13**, 706.

Miedaner, T., Müller, B.U., Piepho, H.P. and Falke, K.C. (2011) Genetic architecture of plant height in winter rye introgression libraries. *Plant Breeding*, **130**, 209-216.

Naranjo, T. and Fernández-Rueda, P. (1991) Homoeology of rye chromosome arms to wheat. *Theor. Appl. Genet.*, **82**, 577-586.

Oettler, G. (2005) The fortune of a botanical curiosity - Triticale: past, present and future. *J. Agric. Sci.*, **143**, 329-346.

Parat, F., Schwertfirm, G., Rudolph, U., Miedaner, T., Korzun, V., Bauer, E., Tellier, A. and Schön, C.-C. (2015) Geography and end use drive the diversification of worldwide winter rye populations. *Mol. Ecol.*, **25**, 500-514.

Passaia, G., Caverzan, A., Fonini, L.S., Carvalho, F.E.L., Silveira, J.A.G. and Margis-Pinheiro, M. (2014) Chloroplastic and mitochondrial *GPX* genes play a critical role in rice development. *Biologia Plantarum*, **58**, 375-378.

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

Plaschke, J., Börner, A., Xie, D.X., Koebner, R.M.D., Schlegel, R. and Gale, M.D. (1993) RFLP mapping of genes affecting plant height and growth habit in rye. *Theor. Appl. Genet.*, **85**, 1049-1054.

R Core Team (2015) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabinovich, S.V. (1998) Importance of wheat-rye translocations for breeding modern cultivar of *Triticum aestivum* L. *Euphytica*, **100**, 323-340.

Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325-2329.

Rogers, J.S. (1972) Measures of genetic similarity and genetic distance. In *Stud. Genet.* (Wheeler, M.R. ed. Austin, Texas, U.S.A.: University of Texas, pp. 145-153.

Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A. and Allard, R.W. (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 8014-8018.

Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome Res.*, **20**, 1165-1173.

Schittenhelm, S., Kraft, M. and Wittich, K.-P. (2014) Performance of winter cereals grown on field-stored soil moisture only. *Eur. J. Agron.*, **52, Part B**, 247-258.

Schlegel, R. (1987) Neocentric activity in chromosome 5R of rye revealed by haploidy. *Hereditas*, **107**, 1-6.

Schmutzer, T., Ma, L., Pousarebani, N., Bull, F., Stein, N., Houben, A. and Scholz, U. (2014) Kmasker - A tool for in silico prediction of single-copy FISH probes for the large-genome species Hordeum vulgare. *Cytogenet. Genome Res.*, **142**, 66-78.

Semagn, K., Babu, R., Hearne, S. and Olsen, M. (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol. Breed.*, **33**, 1-14.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210-3212.

Spannagl, M., Alaux, M., Lange, M., Bolser, D.M., Bader, K.C., et al. (2016a) transPLANT resources for *Triticeae* genomic data. *The Plant Genome*, **9**, 1-13.

Spannagl, M., Nussbaumer, T., Bader, K.C., Martis, M.M., Seidel, M., Kugler, K.G., Gundlach, H. and Mayer, K.F.X. (2016b) PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.*, **44**, D1141-D1147.

Tanabe, S., Ashikari, M., Fujioka, S., Takatsuto, S., Yoshida, S., et al. (2005) A novel cytochrome P450 is implicated in brassinosteroid biosynthesis via the characterization of a rice dwarf mutant, *dwarf11*, with reduced seed length. *Plant Cell*, **17**, 776-790.

Taylor, J. and Butler, D. (2014) ASMap: An (A)ccurate and (S)peedy linkage map construction package for inbred populatons that uses the extremely efficient MSTmap algorithm. (R package version 0.3-3).

The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711-716.

Thompson, J.B. and Rees, H. (1956) Selection for heterozygotes during inbreeding. *Nature*, **177**, 385-386.

Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36-46.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., et al. (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics*, **15**, 823.

Unterseer, S., Pophaly, S.D., Peis, R., Westermeier, P., Mayer, M., et al. (2016) A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. *Genome Biol. Evol.*, **17**, 137.

Vergara, I.A., Frech, C. and Chen, N. (2012) CooVar: Co-occurring variant analyzer. *BMC Research Notes*, **5**, 1-7.
Visser, E.A., Wegrzyn, J.L., Steenkmap, E.T., Myburg, A.A. and Naidoo, S. (2015) Combined de novo and genome guided assembly and annotation of the *Pinus patula* juvenile shoot transcriptome. *BMC Genomics*, **16**, 1-13.
Voss-Fels, K. and Snowdon, R.J. (2015) Understanding and utilizing crop genome diversity via high-resolution genotyping. *Plant Biotechnol. J.*, **14**, 1086-1094.
Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.
Wicker, T., Matthews, D.E. and Keller, B. (2002) TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.*, **7**, 561-562.
Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M. and Stein, N. (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.*, **59**, 712-722.
Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**, 80-83.
Wimmer, V., Albrecht, T., Auinger, H.-J. and Schön, C.-C. (2012) synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, **28**, 2086-2087.
Winfield, M.O., Allen, A.M., Burridge, A.J., Barker, G.L.A., Benbow, H.R., et al. (2016) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.*, **14**, 1195-1206.
Wricke, G., Wilde, P., Wehling, P. and Gieselmann, C. (1993) An isozyme marker for pollen fertility restoration in the Pampa cms system of rye (*Secale cereale* L.). *Plant Breeding*, **111**, 290-294.
Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873-881.
Xu, C., Jiao, C., Zheng, Y., Sun, H., Liu, W., et al. (2015) *De novo* and comparative transcriptome analysis of cultivated and wild spinach. *Scientific Reports*, **5**, 17706.
Yonemaru, J.-i., Yamamoto, T., Fukuoka, S., Uga, Y., Hori, K. and Yano, M. (2010) Q-TARO: QTL Annotation Rice Online Database. *Rice*, **3**, 194-203.
Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203-214.
Zhou, H., Muehlbauer, G. and Steffenson, B. (2012) Population structure and linkage disequilibrium in elite barley breeding germplasm from the United States. *Journal of Zhejiang University. Science. B*, **13**, 438-451.

Tables

Table 1. Summary of WGS assembly and scaffolding.

	Contigs*	Scaffolds**	Scaffolds (>1kb)
Sum (Mb)	1,685	2,804	2,334
Total number	1,581,707	1,286,927	335,608
N50 contig length (bp)	1,708	9,448	12,472
N75 contig length (bp)	705	2,935	7,545
Maximum (bp)	35,334	148,970	148,970

* Minimal length of 200 bp.

** Minimal length of 300 bp.

Table 2. Genome annotation and variant calling statistics.

A) Total number of high-confidence gene models and number of gene models with AHRD or Blast2GO annotation and their representation by orthologous clusters.
B) Functional annotation of SNVs according to subclasses defined by the CooVar annotation (Vergara *et al.*, 2012). The class 'non-synonymous' includes frameshift and in-frame variants, non-conservative missense, conservative missense, variants in the splice acceptor or splice donor site and gain or loss of stop codon.

A) Total number of rye gene models:		**27,784**	**(%)**
Assigned functional description (AHRD or Blast2GO)		26,571	95.6
OrthoMCL link to grass species*		17,470	62.9
B) Total number of SNVs:		**8,626,622**	
Occurrence in elite breeding pool or *S. vavilovii*	Seed	Pollen	*S. vavilovii*
Total number	4,010,067	4,812,751	3,797,250
Silent (%)	97.59	97.59	97.92
Synonymous (%)	1.40	1.40	1.20
Non-synonymous (nsSNV, %)	1.01	1.01	0.88
Total number of gene models with nsSNV	7,907	8,483	8,041

* brachypodium, barley

Figure legends

Figure 1. Distribution of SNVs, genes and markers along the rye nuclear genome.

The figure depicts WGS contigs anchored to the rye genetic framework as a combined representation of the high-density genetic map and the rye genome zipper. This framework was constructed using 93,157 markers, assigning 52,901 different sequence contigs to 2,813 unique cM positions. The plot is separated into several circular tracks showing the seven chromosomes 1R to 7R. The outer track depicts the heterozygosity of the sequenced reference line Lo7 as histogram with the number of heterozygous sites per 1 cM. The second track shows the gene density per 1 cM along the genetic map plotted as heatmap. The subsequent track shows the short (grey) and the long (dark green) chromosome arm, where the breakpoint indicates the position of the centromere. The following connector track illustrates the anchoring of the genetic map to the physical sequence contigs. Subsequently, gene density is plotted per 500 kbp for the physical sequence level. The next track shows

the marker density (brown histogram). The eleven inner tracks present the SNV density as heatmap along the anchored WGS contigs for each of the resequenced rye genotypes. From outside to inside, the tracks represent the *S. vavilovii* accession, the five inbred lines from the pollen parent pool (Lo351, Lo348, Lo310, Lo298, Lo282) and the five inbred lines from the seed parent pool (Lo191, Lo176, Lo117, Lo115, Lo90).

Figure 2. Collinearity between the genetic maps of rye and barley.

A) The order of gene-bearing sequence contigs in the high-density genetic map of rye (x-axis) was compared to the order of their orthologous contigs in the assemblies of the barley genome (The International Barley Genome Sequencing Consortium, 2012) (y-axis). Chromosomes are separated by blue lines. Positions of genetic centromeres are marked with dotted grey lines.
B) Schematic representation of genomic rearrangements between barley chromosomes 1H to 7H (left) and rye chromosomes 1R to 7R (right).

Figure 3. Principal Coordinate Analysis (PCoA) of seed and pollen parent pool and genetic resources.

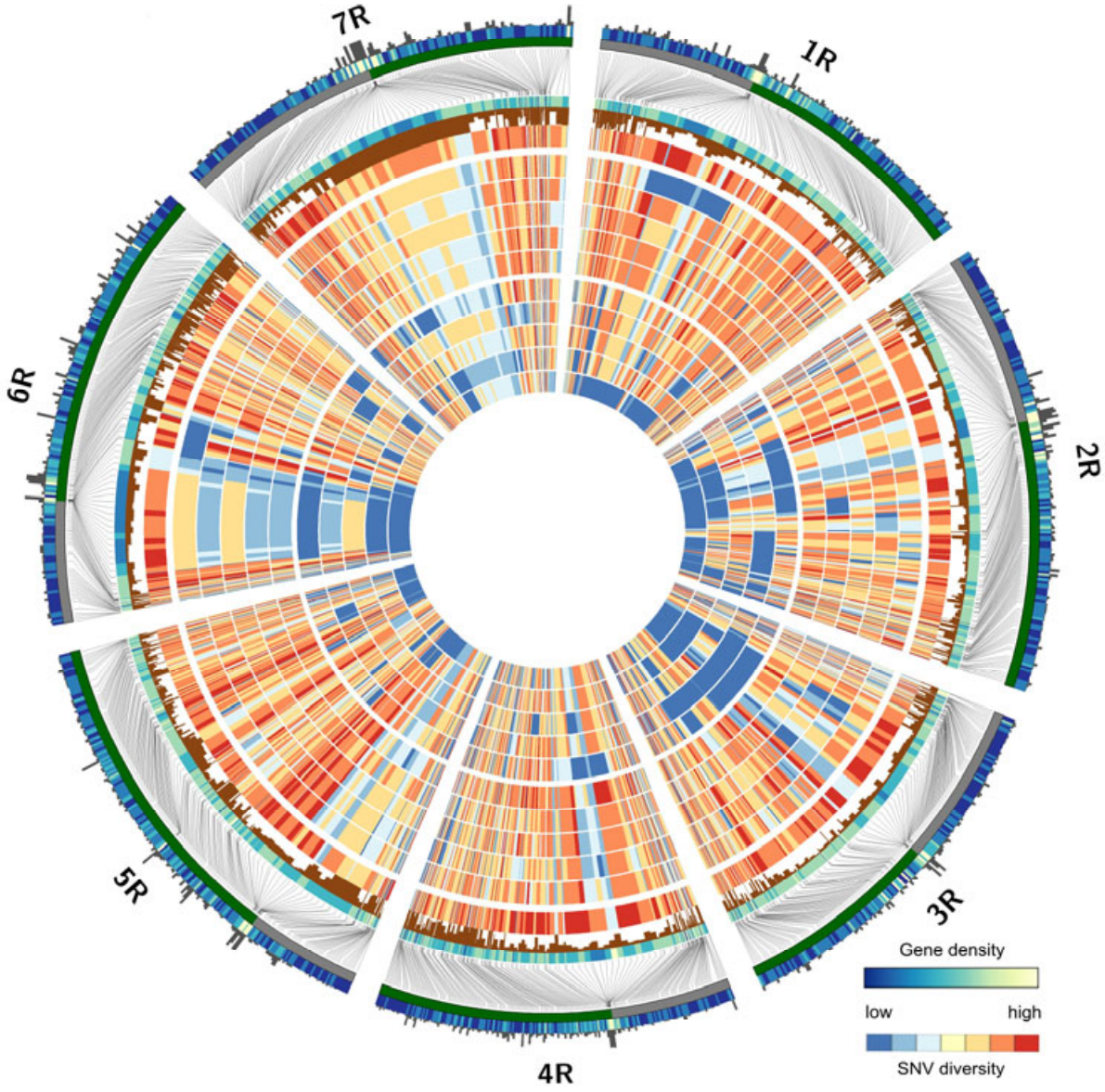
The PCoA was based on Rogers' distances calculated from 179,660 SNVs from the Rye600k array. Only SNVs with < 5% missing values and minor allele frequency > 0.01 were used. PCo1 and PCo2 are the first two principal coordinates. The percent of the variance explained by the respective PCo is indicated. GR: genetic resources.

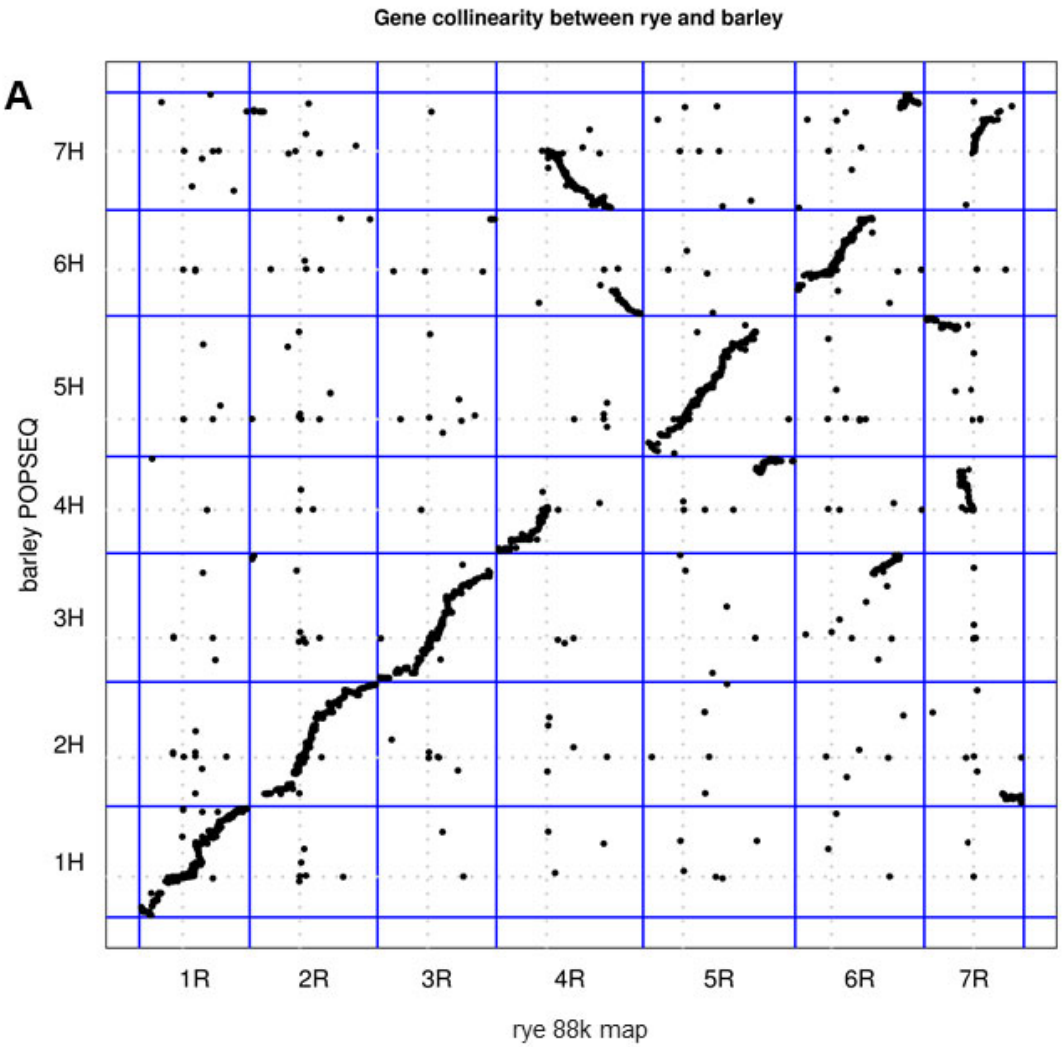
Figure 4. Genome-wide map of selection signals between the seed and pollen parent pools.

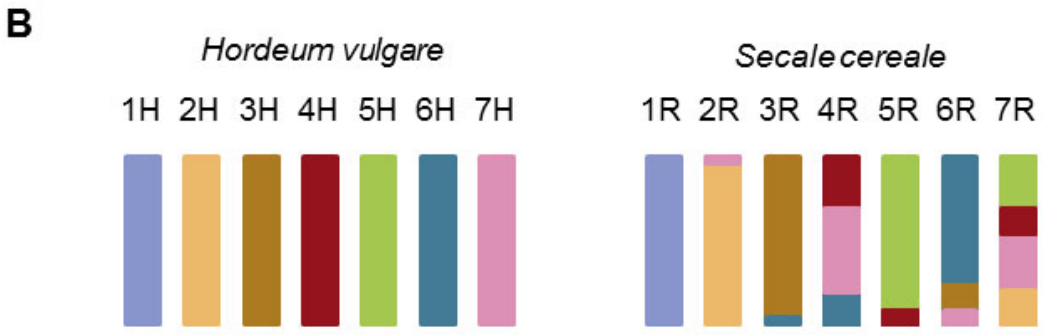
The plots for the seven rye chromosomes are based on 78,731 genetically mapped SNVs from the Rye600k array. The blue and red SNVs are the top 1% X^TX values. The red SNVs are shared Lositan (F_{ST}) outliers and top 1% X^TX values. Centromere positions are indicated by a triangle on the x-axis.

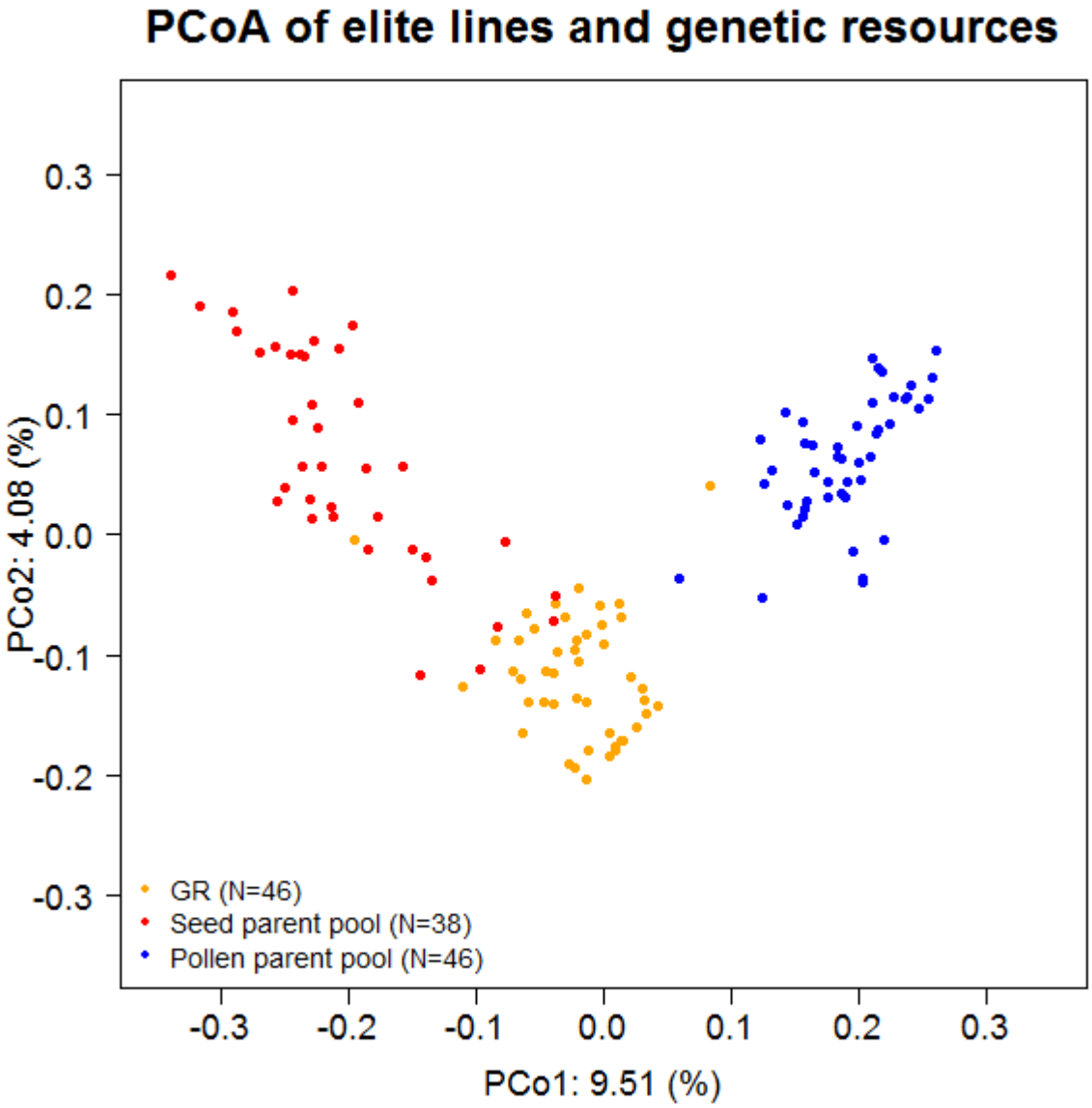
Figure 5. Morphological differences in ears from different rye gene pools and haplotype frequencies in two contigs with selection candidates.

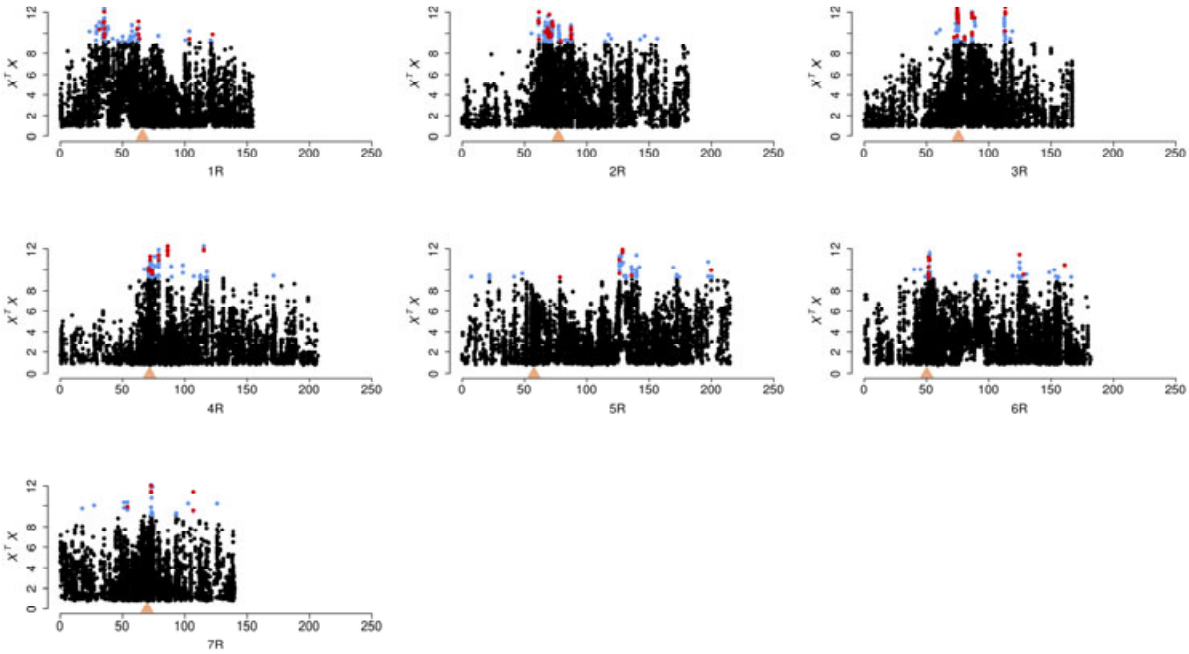
A) From left to right, three representative ears are shown for inbred lines from the seed and pollen parent pool and from genetic resources (Eastern European open-pollinated populations), respectively. The ruler on the right side indicates the size of the ears in cm.
B) Graphical representation of haplotype patterns for contigs Lo7_v2_contig_1355272 (top, 15 SNVs) and Lo7_v2_contig_63401 (bottom, 9 SNVs) which harbour the rye orthologs of rice genes *SP1* and *OsGPX1/OsGPX3*, respectively. Only the more frequent haplotypes present in at least four individuals (~10%) in one of the three populations are shown. Each row represents a haplotype with SNVs indicated by boxes. Reference allele (Lo7) homozygote: light grey, heterozygote: middle grey, alternative allele homozygote: dark grey. The frequencies of the haplotypes in seed and pollen parent pool and genetic resources (GR) are shown on the right side. Numbers in the columns do not add up to 1 since rare haplotypes are not displayed.











B Haplotypes in contigs carrying rye orthologs of *SP1* and *OsGPX1/OsGPX3*

	Seed	Pollen	GR
SP1	0.05	0.09	0.04
	0.29	-	0.02
	0.39	-	0.07
	-	0.13	-
	-	0.35	0.09
OsGPX1/ OsGPX3	0.87	0.02	0.41
	0.03	0.61	0.04
	-	0.13	0.02
	-	0.07	0.09