Title

Deep sampling and pooled amplicon sequencing reveals hidden genic variation in heterogeneous rye accessions

Authors

Anna Hawliczek[1], Leszek Bolibok[2], Katarzyna Tofil[1], Ewa Borzęcka[1], Joanna Jankowicz-Cieślak[3], Piotr Gawroński[1], Adam Kral[1], Bradley J Till[3,4]*, Hanna Bolibok-Brągoszewska[1]*

Affiliations

[1] Department of Plant Genetics, Breeding and Biotechnology, Institute of Biology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland

[2] Department of Silviculture, Institute of Forest Sciences, Warsaw University of Life Sciences – SGGW, Warsaw, Poland

[3] Plant Breeding and Genetics Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, IAEA Laboratories Seibersdorf, International Atomic Energy Agency, Vienna International Centre, Vienna, Austria

[4] Plant Genomics Unit, Centro de Genòmica Nutricional Agroacuìcola, Temuco, Chile

Contact Information

Bradley J Till: till.brad@gmail.com

Hanna Bolibok-Brągoszewska: hanna_bolibok_bragoszewska@sggw.pl

Additional Footnotes

* equally contributing, corresponding authors

Running Title

Recovery of hidden genic variation in rye

Abstract

Loss of genetic variation negatively impacts breeding efforts and food security. Genebanks house over 7 million accessions representing vast allelic diversity that is a resource for sustainable breeding. Discovery of DNA variations is an important step in the efficient use of these resources. While technologies have improved and costs dropped, it remains impractical to consider resequencing millions of accessions. Candidate genes are known for most agronomic traits, providing a list of high priority targets. Heterogeneity in seed stocks means that multiple samples from an accession need to be evaluated to recover available alleles. To address this we developed a pooled amplicon sequencing approach and applied it to the out-crossing cereal rye *(Secale cereale)*. Ninety-six plants from an accession were pooled together, and 95 unique accessions evaluated for sequence variation in six target genes involved in seed quality, biotic and abiotic stress resistance, resulting in 74 predicted deleterious variants using multiple algorithms. Rare variants were recovered including those found only in a low percentage of seed. We conclude that this approach provides a rapid and flexible method for evaluating stock heterogeneity, probing allele diversity, and recovering previously hidden variation in large germplasm collections.

Keywords: *Secale cereale*, natural variation, allele frequency, variant calling, *MATE 1*, *FBA*, *TLP*, *GSP-1*, *Sinb*, *PBF*

INTRODUCTION

Development of crop varieties that are more productive and tolerant to biotic and abiotic stresses is needed, in view of the predicted climatic and demographic changes, such as a rise of average temperature by up to 4.0 °C by the end of the century, a higher frequency of extreme weather events, and an increase of human population by 26% until 2050 (Wheeler and von Braun, 2013; UN, 2019), Yet, yearly yield increases of major crops were found to be in the range of 0.9 to 1.6 % and need to be increased by ca. 90 to 170 %, depending on the crop, to meet the projected demand by 2050 (Ray et al., 2013).

Plants can be made more resilient, yields stabilized, and nutritional components enhanced through selection and combination of gene variants that control these traits. Crop improvement is therefore dependent on the existence of genetic variability for the trait in question. For the past 10,000 years humans have been selecting and combining genetic variants to improve crops. However, most of the history of crop development was carried out without a knowledge of genetics or DNA, and thus modern cultivars have a relatively narrow genetic base, resulting from bottleneck-like effects of domestication and breeding practices (Purugganan and Fuller, 2009; Mondal et al., 2016; Joukhadar et al., 2017). Therefore, the allelic variability existing within contemporary cultivars or breeding programs may be insufficient for successful identification of gene variants for satisfactory productivity and resilience of the crop.

Useful alleles conferring important traits that have been lost in modern cultivars may still exist in nature. Plant genetic resources (PGR), such as landraces and wild relatives of crop plants, possess a much higher genetic diversity. While not high yielding and having often undesirable agronomic characteristics, they were shown to contain gene variants that can improve performance of successful modern cultivars (Hoisington et al., 2002; McCouch, 2013)**.** Gur and Zamir achieved yield increases in tomato exceeding 50% by introgressing three independent genome regions from wild species *Solanum pennellii* (Gur and Zamir, 2004). An example directly relevant to plants' resilience is that of the gene PSTOL1, conferring tolerance to phosphorus starvation. It was discovered in a traditional rice cultivar Kasalath, and found to improve root growth and phosphorus uptake when transferred via marker-assisted backcrossing to Asian lowland rice cultivars naturally lacking this gene (Gamuyao et al., 2012; Pariasca-Tanaka et al., 2014).

Luckily, the value of PGR as a reservoir of gene variants was recognised over a hundred years ago (McCouch, 2004), and nowadays there are over 1,700 *ex situ* germplasm collections worldwide, maintaining about 7.4 million accessions. Approximately 62% of these accessions are landraces and wild species (FAO, 2010). Unfortunately, in most cases little is known about the

extent and structure of genetic diversity within a given collection. The available data is often limited to passport information, and some phenotypic measurements or DNA marker-based genetic diversity assessment for a subset of accessions. Such information is not sufficient to make an informed choice of PGR for inclusion into a breeding program. Therefore the utilisation of primitive, exotic germplasm in crop improvement is limited (FAO, 2010; McCouch, 2013; Keilwagen et al., 2014).

To fully profit from the allelic variation of PGR, methods for efficient and reliable screening of hundreds of accessions to discover useful gene variants are needed. Rapid development of next generation sequencing (NGS) technologies resulted in the establishment of various approaches, which can be used for high-throughput assessment of genic variation, such as whole genome resequencing (WGS) (Li et al., 2013; Mehra et al., 2015; Zhou et al., 2015) or exome capture (Hodges et al., 2007; Hussain et al., 2018). Unfortunately, these approaches are not yet applied in many species owing to factors including genome size, polyploidy, and associated costs of sequencing and capture probe development. While a future can be envisioned where comprehensive genomic data is available for every accession of every important crop, the current state of technology and funding means that material is prioritized, and compromises made. Insofar as evaluation of WGS data provides information useful for understanding population genetics and evolution, it is expected that only a small fraction of base pairs of a genome are controlling key agronomic traits (Wendel et al., 2016). Targeting candidate genes and their regulatory elements provides a tremendous reduction in data collected. Indeed, many studies have revealed quantitative trait loci and associated candidate genes that can be used to identify orthologous sequences in other plants (Nguyen et al., 2019). An alternative to whole genome or exome capture sequencing is amplicon sequencing. In this approach, selected genomic regions are first amplified by PCR and then subjected to massively parallel sequencing. Far fewer nucleotide bases are sequenced compared to exome capture or WGS. This results in much higher sequencing coverage and greater sensitivity (e.g. (Sims et al., 2014)). The high sensitivity of current sequencing technologies further enables “ultra deep” methods whereby PCR products from many different genomic DNA samples are pooled together prior to sequencing. Experiments can be designed to discover rare nucleotide variants present at very low frequencies, while simultaneously increasing screening throughput and reducing assay costs. This has been used for recovery of induced point mutations in TILLING by Sequencing assays (Tsai et al., 2011). Here, genomic DNAs from different lines harboring induced mutations are pooled, subjected to target-specific PCR and the PCR products are then pooled and sequenced. The method has been used to recover rare mutations in genomic DNA samples pooled from 64 to 256 fold. These studies suggest that variant calling accuracy is improved when using multiple variant calling algorithms (Tsai et al., 2011; Pan et al., 2015; Gupta et al., 2017;

Tramontano et al., 2019). The approach has been adapted for recovery of natural variation in *Populus nigra* and *Manihot esculenta* Crantz (cassava), whereby DNAs from different accessions were pooled together prior to PCR and variant discovery. In *P. nigra*, PCR products were prepared from pooled genomic DNA from 64 accessions to identify variants in lignin biosynthesis genes in 768 accessions (Marroni et al., 2011). In cassava, DNA from up to 281 accessions were pooled prior to sequencing for variants in starch biosynthesis pathway-related genes and herbicide tolerance genes in 1667 accessions (Duitama et al., 2017).

Rye (Se*cale cereale L.*) is an outcrossing cereal, popular in Europe and North America, and an important source of variation for wheat breeding due to its high tolerance to biotic and abiotic stresses (Crespo-Herrera LA et al., 2017). Genetically rye is a diploid (n=7), with a large (ca. 8 Gbp) and complex genome (Bartos et al., 2008). There are over 21 thousand rye accessions in genebanks worldwide, approximately 35% of them are landraces and wild species (FAO, 2010). Several studies on genome-wide diversity in rye were published to date (Bolibok-Bragoszewska et al., 2014; Targońska et al., 2016; Maraci et al., 2018; Sidhu et al., 2019). It was shown that accessions from genebanks are genetically distinct from modern varieties, which highlighted the potential of PGR in extending the variability in current rye breeding programs (Bolibok-Bragoszewska et al., 2014; Targońska et al., 2016). To date neither NGS-based targeted amplicon sequencing, nor any other method of gene variant discovery was applied to rye genetic resources.

Exploration of genic variation in outcrossing, generatively propagated crops, such as rye, maize, sugar beet, broccoli, or carrot, is a particularly demanding task. Natural, random-mating populations of such species are heterozygous and heterogenous, with multiple alleles of a locus being present (Souza Jr., 2012). Such population structure has important implications for the design of NGS-allele mining experiments. Firstly, due to high levels of heterozygosity, a higher sequencing coverage is needed even when sequencing non-pooled samples to ensure reliable nucleotide variant calling. Secondly, due to the heterogeneity of accessions, a large enough number of individuals of a given accession needs to be included in the screen to obtain a faithful representation of within-accession variability and to successfully recover rare variants. Many potentially useful and interesting alleles may go undiscovered with current experimental designs. To address this, a low-cost, high-throughput, and reliable amplicon sequencing approach suitable for assessment of genic variation in heterozygous and heterogenous rye accessions was developed. Rather than pool DNA from different accessions, ultra deep amplicon sequencing was used to evaluate intra-accession heterogeneity while also providing information on novel genetic variation. DNA pools were created that contain ninety-six plants per accession. These were subjected to pooled amplicon sequencing in six target genes implicated in biotic and abiotic stress resistance and seed quality: aluminium activated citrate transporter (*MATE1)*, taumatin-like protein (*TLP*),

fructose-biphosphate aldolase (*FBA*), prolamine-box binding factor (*PBF*), secaloindoline-b (*Sinb*) and grain softness protein *(GSP-1*). *MATE1* is a gene involved in aluminum (Al) tolerance of rye. Al-toxicity is one of the main constraints to agricultural production on acidic soils, which constitute ca. 50% of the arable land on Earth (Maron et al., 2013). Rye is one of the most Al-tolerant cereals, with the degree of tolerance depending on the allelic variant of *MATE1* (Santos et al., 2018). TLPs are a family of pathogenesis-related (PR) proteins, involved in fungal pathogen response in many plant species (Zhang et al., 2018). FBA is one of the key metabolic enzymes involved in CO_2 fixation and sucrose metabolism. *FBA* genes were found to have an important role in regulation of growth and development, and responses to biotic and abiotic stresses, such as chilling, drought and heat (Lv et al., 2017; Cai et al., 2018). *GSP*-1 genes, belonging to the prolamin superfamily of seed storage proteins, encode precursor proteins, which after post-translational processing give rise to arabinogalactan peptide AGP and the grain softness protein GSP-1 (Wilkinson et al., 2017). Secaloindolines, products of genes *Sina* (not analyzed in this study) and *Sinb*, are main components of friabilin - a starch-associated protein fraction of cereal grains (Simeone and Lafiandra, 2005). The wheat orthologues of *Sina* and *Sinb*, called *Pina* and *Pinb*, are key determinants of grain texture, an important breeding trait directly influencing the end-use, (Liu et al., 2017). *PBF* is an endosperm specific transcription factor involved in the regulation of protein and starch synthesis (Zhang et al., 2016). It binds to the prolamine-box motif occurring in promoter regions of multiple cereal seed storage proteins. In barley, SNPs located in *PBF* were associated with crude protein and starch content (Haseneyer et al., 2010), while in wheat, mutating the homeologous *PBF*s using TILLING resulted in a markedly decreased gluten content and high content of lysine (Moehs et al., 2019).

In the present study, three variant calling algorithms (GATK HaplotypeCaller (Poplin et al., 2017), SNVer (Wei et al., 2011) and CRISP (Bansal, 2010)) were used to identify putative variants at frequencies as low as one heterozygous event per 96 plants assayed in each pool. A subset of variants was independently validated and the functional effect of each variant was evaluated *in silico*. Common and rare variants were recovered, including variants present in only a small fraction of seed representing an accession. This data provides preliminary knowledge on the levels of variant allele frequencies in accessions of different types: wild species, landraces, historical and modern cultivars.

RESULTS

DNA sequencing, mapping and coverage

Pooled amplicon sequencing using Illumina sequencing by synthesis 2x300 paired end reads on 95 accessions and six genes produced a mean coverage of 13948x and mean mapping quality of 58.65 . Mean coverage per accession pool varied approximately 10 fold, between 2924x and 30275x. Analysis of sequencing coverage at each nucleotide revealed that 88.5% of the experiment produced five or more reads to support the rarest possible nucleotide variant (Figure 1, Supplemental Table 1.

Evaluation of variant calling algorithms and predicted effects of nucleotide changes

Variant calling was first performed using HaplotypeCaller in GATK (v.4.0) with ploidy set to 192 in order to recover rare alleles. This resulted in 4,115 called variants, of which 3,682 were single nucleotide polymorphisms, 192 insertions, and 241 deletions. To evaluate the distribution of allele frequencies (AF), the data was placed into six classes based on the mean AF of each variant. Frequency boundaries used were 1 to 0.7 (class A), 0.69 to 0.4 (class B), 0.39 to 0.1 (class C), 0.099 to 0.069 (class D), 0.068 to 0.038 (class E) and those with values below 0.038 (class F). Private variants occurring in only one accession were recovered in each allele frequency class (Figure 2, Supplemental Table 2). The percentage of private alleles in classes A through E was 41.0, 19.8, 21.7, 30.6, and 47.9%, respectively. These data represent 1,103 of the 4,115 variants called using HaplotypeCaller. The remaining 3,012 (73.2%) are found in class F. In this set, 1,389 (46.1%) were private variants and 93.7% of variants were found in 10 or fewer accessions.

Variant calling was next carried out using SNVer and CRISP producing 1,570 and 3,261 variant calls, respectively. Similar to data produced with GATK, class F contained the highest percentage of variants (43.4 and 75.6%, respectively). Private variants in all allele frequency classes ranged from 17.02 to 48.98 % with SNVer and from 7.69 to 42.62% with CRISP (Supplemental Table 2). In total, 895 variants were common between the three methods (Figure 3). Within these common variants, the mean allele frequency and the number of accessions carrying the variant differed between the three algorithms used.

Private variants occurred in all germplasm groups included in the study: modern cultivars, historic cultivars, landraces and wild accessions. In the group of variants commonly identified by three algorithms the number of private variants per accession coincided with the domestication status: private variants were most frequent in wild accessions (five to seven per accession), followed by landraces with approximately two variants per accession, and historic and modern cultivars with less than one private variant per accession. Private variants in wild accessions also

had the highest AF values (mean 0.4 -0,45, median 0.22-0.28). In the remaining germplasm groups mean AF and median AF did not exceed 0.12 and 0.007, respectively. Allele frequencies from individual pools were used to prepare violin plots in order to qualitatively compare different accessions and identify potential outliers. Several distinct patterns of allele frequency distribution were observed (Figure 4, and not shown). Five accessions, characterized by a high proportion of variants with high AF values were consistently recognized as markedly different from the rest: *S. sylvestre, S. strictum subsp. kuprijanovii*, historic cultivar Imperial (from A Lukaszewski's collection, UCLA, Riverside), and landraces R 3 and R1040 (Supplemental Table 1).

The effect on gene function of putative variants was evaluated with SNPeff and SIFT . This resulted in 695 putative deleterious variants from GATK, 171 from SNVer and 578 from CRISP, with 74 putative deleterious variants common to all three algorithms (Table 1, Supplemental Table 3). Deleterious alleles with a high maximum AF and present in only one accession were recovered along with alleles with a high maximum AF that were present in 90 or more accessions (Supplemental Table 3). Alleles with a maximum AF less than 0.4 were also identified, suggesting the presence of rare alleles segregating within an accession. In the GATK data set, for example, 29 of the 74 predicted deleterious common variants have a maximum AF between 0.047 and 0.391 and are found in 1 to 21 accessions.

Within target genes, 18 to 443 polymorphic positions were detected consistently by the three algorithms, corresponding to one SNP or InDel every eight to ten bp of sequence for five of the analyzed genes. For the sixth gene, *Sinb*, this frequency was markedly lower, with one SNP per 25 bp. The number of putatively deleterious variants per gene ranged from 11 (*GSP-1*) to 21 (*MATE1*), corresponding to one deleterious variant every 40 to 80 bp, with exception of *Sinb*, where only three deleterious variants were identified in 447 bp of coding sequence. Previous data on genic variation was available solely for *MATE1* (a total of 112 unique variants from 26 sequences deposited in GenBank as of 7th January 2020) and Sinb with seven unique variants reported (Liu et al., 2017). The present study identified 62 new variants in *MATE1* coding sequence, including seven putatively deleterious, and 15 new variants in *Sinb*, including all three putatively deleterious variants. Most new variants identified in *MATE1* and *Sinb* were private or rare (median of the number of accessions with a given variant equaled two in *MATE1* and one in *Sinb)*.

The presence of predicted variants was assayed using Sanger sequencing of *Sinb* amplicons in a single individual plant from each of eight accessions. Twelve variants were predicted in this set. Only variants reported by all three algorithms for the tested accession, and where the lowest AF was greater than 0.295 were validated (Supplemental Table 4). Because allele frequencies were calculated from a pooled DNA sample, it was concluded that lower frequency alleles likely represent alleles that are not present in every seed of an accession. Therefore, failure to validate

variants may have resulted from sampling DNA from only one seed. To test this, subsequent validation assays were carried out whereby multiple plants from each accession were assayed independently. In CAPS and Sanger sequencing assays on *MATE1*, *PBF*, and *Sinb* amplicons eleven out of 16 tested variants were recovered when sampling between six and 16 plants (Table 2). Observed allele frequencies calculated from the number of plants harboring the tested sequence difference varied from the frequencies predicted from the amplicon sequencing data. Five variants had observed AF closest to GATK predictions, three variants were closest with SNVer and three with CRISP. The four variants not recovered by CAPS or Sanger assays had frequencies reported by GATK below 0.15. Failure to recover low frequency alleles may have resulted from testing an insufficient number of individuals.

DISCUSSION

To evaluate the distribution of frequencies of alleles within a landrace or cultivar, we chose to sample 96 plants from each seed stock of rye selected for our study. This allows the recovery of i) sequence differences compared to the reference sequence used that are fully homozygous (those with an allele frequency of 1), ii) heterozygous variants present in all pooled plants and iii) variants of lower frequencies that are not present in every seed in the seed stock used to represent an accession. To streamline the approach, tissue from each plant was collected and pooled prior to DNA extraction. The experiment was designed such that an allele found in a single plant could be identified. Investigation of sequencing coverage revealed that 88.5% of the experiment produced five or more reads to support the rarest possible allele. High coverage values were found in all DNA pools suggesting that each pool was suitable for PCR amplification and sequencing. Deviations in coverage values, therefore, likely resulted from differences associated with the quantification, normalization and pooling of PCR products. Such variations were recently reported in a study comparing tomato, cassava and barley amplicon sequencing data sets (Tramontano et al., 2019). The study revealed that minor coverage improvements could be achieved through the addition of extra quantification methods. Alternative approaches to increase read coverage at all nucleotide positions include increasing sequencing yields by adjusting the number of samples in an experiment and/or the number of target genes (amplicons) used in a single sequencing run. As sequencing costs drop, it may prove more cost effective and faster to simply produce more base pairs of data per experiment than to fine-tune the other experimental parameters.

Many applications employing next generation sequencing of genomic DNA involve the evaluation of sequence variations in diploid samples. Even in optimized diploid conditions, a balance is struck between maximizing allele calling sensitivity to reduce false negative errors and reducing the sensitivity in order to lower false positive errors. For example, when using GATK HaplotypeCaller with settings for diploid samples, Li et al reported that more than 80% of false positive errors in diploid rice were at an allele frequency below 40% (Li et al., 2019). When sequencing non-pooled samples, setting an allele frequency threshold of >40% for heterozygous variants therefore reduced false positive errors. Such optimizations are more challenging in pooled samples. Algorithms such as GATK HaplotypeCaller, SNVer and CRISP provide parameter settings to call low frequency variants. Yet, optimal parameters still need to be determined. For example, evaluation of six SNP calling algorithms in tomato TILLING samples pooled either 64 or 96 fold revealed that accuracy ranged between 89.33 and 25.33 % when comparing to Sanger validated SNP mutations (Gupta et al., 2017). That work described technical differences between different algorithms and concluded that accuracy is improved when a variant call is predicted by at least two software. In cassava, up to 281 different accessions were pooled together prior to sequencing in an approach designed to quickly identify putative deleterious alleles (Duitama et al., 2017). In that study 24% (79/325) of called variants were predicted by four algorithms tested.

The experimental design for rye differed from previous studies in order to allow the discovery and analysis of intra-accession allele variation. The assay was designed to recover two types of what can be considered “rare” variation. The first type of rare variants are alleles that are found in only one accession (known as private alleles) or very few accessions in the tested set, and occur with a high frequency within the respective accessions. This type of rare variation is easily recovered using conventional genotyping and resequencing as alleles can be recovered through assay of a single seed (Wang et al., 2018; Balfourier et al., 2019). The second type of rare variants are more difficult to discover. These variants segregate at a low frequency within an accession and are never found at high frequency in any tested accession. To recover this type of rare variant requires the sampling of multiple individual seed per accession. As such, these alleles are hidden from discovery when using traditional methods that sample one or few seed per accession. Using pooled amplicon sequencing we have recovered both types of rare alleles in the tested rye accessions. Common variation was also detected.

Variants with mean AF between 0.7 and 1 represented between 1.87 and 3.06% of all predicted alleles, depending on the software used. In this set of variants, between 41 and 49% are private alleles found in only one accession. The highest number of variants were found in the lowest

allele frequency class (0.038 and below) ranging from between 43 and 76% of the total. It is expected that false positive errors will increase as the number and percentage of reads supporting the alternative allele decreases. Studies have been carried out on errors associated with MiSeq paired end sequencing, but a thorough investigation into errors in pooled samples has not been reported (Schirmer et al., 2015). False positive errors are expected to be random and therefore infrequently independently predicted when applying multiple variant calling algorithms. Indeed, of the 895 variants common to GATK, SNVer and CRISP, only 20% had a predicted mean AF of 0.038 or lower, a reduction of more than 50% from any single algorithm. Further experiments are required to determine what, if any, percentage of the sub 0.038 AF variants predicted by all three algorithms are false positive errors. This requires extensive genotyping, as many individual seed need to be tested to ensure true variants are recovered. In the present study, genotyping assays using approximately 10 seed per accession were sufficient to validate alleles with an AF of 0.15 or higher that were predicted in the same accession by all three algorithms. We expect it is necessary to test more than 100 seeds per accession to validate the lowest frequency alleles in the data. Some very low frequency false positive errors are expected and may result from biological contamination, for example, from pollen contamination on the leaf tissue collected. This can be ruled out in the present study because seedlings were grown, and tissue collected in growth room conditions where there were no rye plants flowering. Sample to sample cross contamination of DNA or PCR product may also be a source of low AF false positive errors. Sixty-five percent of sub 0.038 AF variants commonly predicted by all algorithms were found in more than one accession. However, 96 % had a maximum predicted AF of less than 10%, and the highest maximum AF was 24.5%. This means that a large volume of accidental liquid transfer between samples would be needed to create a detectable false positive. With the caveat of possible very low frequency false positive errors, we conclude that selecting variants commonly called by multiple software may reduce errors and serves as a useful method to prioritize alleles for further study.

We found qualitative evaluation of AF values using violin plots to be useful to estimate the influence of a taxon's reproductive biology, preservation history and breeding on the genetic composition of an accession. For example, one of the outlier accessions identified is this study is a representative of *S. sylvestre*. Molecular marker-based analyses of genetic diversity indicated this allogamous taxon as the most divergent in genus *Secale* (Bolibok-Bragoszewska et al., 2014; Al-Beyroutiova et al., 2016; Schreiber et al., 2018). Its large proportion of high AF variants (Figure 4) likely corresponds to homozygosity for alternative alleles, since reference sequences used during variant calling originated from cultivated rye accessions. Another outlier, *S. strictum* subsp. *kuprijanovii*, is a perennial outbreeder, also genetically divergent from *S. cereale*. However, its violin plot differs from plots obtained for the other two *S. strictum* samples included in the study, *S.*

strictum subsp. *africanum* and *S. strictum* subsp. *strictum*, which might indicate a sample tracking mistake during genebank preservation or laboratory handling, or a bottleneck during preservation. A sample of Imperial cultivar, widely used in cytological studies, originating from the collection of A. J. Lukaszewski (UCLA, Riverside) showed an approximately equal abundancy of variants with all possible AF values and differed clearly from another sample of Imperial, R1150, obtained from IPK Gatersleben genebank. Differences (although less pronounced) were also observed between the two samples of cultivar Dankowskie Zlote obtained from different sources. Samples of hybrid cultivars from KWS, such as Ballistic and KWS Florano, exhibited a higher percentage of AF values in the range 0.3-0.5, with median ca. 0.3, and also higher percentage of AF values close to 1.0, in comparison to population cultivars included in the study, such as Dankowskie Zlote, Petkus, or Carsten Kortstra, which is consistent with the use of the three line system in the development of hybrid rye cultivars.

The approach of deep sampling and pooled amplicon sequencing allows discovery of variants in candidate genes and also an evaluation of the effect of variants on gene function. This provides an additional filter to prioritize variants. The SIFT program was used to identify 73 putative deleterious alleles commonly identified by the three variant calling algorithms. This data set contained different classes of alleles for example, homozygous variants found in one or few individual accessions (private deleterious alleles, the first category of rare variants). Homozygous variants present in more than 90 accessions were also recovered. Interestingly, putative deleterious variants were also identified where the maximum AF was between 0.15 and 0.3 and the variant was found in only one or two individual accessions (the second category of rare variant described above). Further studies are being designed to evaluate the different classes of putative deleterious alleles. For example, homozygous private alleles may represent alleles where a fitness penalty results in the allele having been expunged from most populations. Homozygous putative deleterious alleles present in most tested accessions may represent alleles with no fitness penalty, or may represent alleles that have no negative effect on fitness under their natural growing conditions (e.g. low aluminium in the soil). Possible mechanisms for the maintenance of rare low frequency alleles in populations, including meiotic effects, can also be investigated.

In this study we analyzed six genes linked to biotic and abiotic stress resistance and seed quality Using deep sampling and pooled amplicon sequencing we identified numerous new variants, (including putatively deleterious ones), in each of the analyzed genes, providing potential targets for future functional studies and, eventually, inclusion in breeding schemes in rye and related species (wheat, triticale). Consistent with a high diversity of the germplasm set used (with respect to

domestication status and origin) we obtained a several fold higher estimate of SNP frequency in rye (on average one SNP or InDel every 12 bp), than those reported in the past: 1 SNP/52bp (Li et al., 2011), 1 SNP/58bp (Varshney et al., 2007) or 1 SNP or InDel/31 bp (Bauer et al., 2017). In agreement with the results of previous genome-wide, DArT-marker based characterization of genetic diversity in rye (Bolibok-Bragoszewska et al., 2014), data obtained in the present study on distribution of private alleles among germplasm groups indicates that the genetic diversity in modern rye cultivars is relatively narrow, with less than one private allele identified per modern cultivar tested, and provides further evidence for the value of rye PGR in genetic research and crop improvement, with more than five private alleles identified per accession, stressing the importance of conservation and characterization efforts.

The rye amplicons used in this study were generated before the release of the rye genome (Rabanus-Wallace et al., 2019). It is expected that the recent release of the rye reference genome will enable improvements in gene target selection and primer design. Reference genomes have been produced for few of the hundreds of thousands of plant species existing on the planet. Because pooled amplicon sequencing does not require complete genome sequence, we expect that the approach described for rye can be adapted for many plant species and can facilitate better characterization of existing rich germplasm collections. We predict that flexible and low-cost methods for recovery of rare genetic variation will support future efforts to promote sustainable food security.

MATERIALS AND METHODS

Plant material

Eighty-three accessions of cultivated rye (*S. cereale* spp. *cereale)* and 12 accessions of other *Secale* taxa were selected to represent maximum genetic diversity from a larger set, based on DArTseq genotyping results (Hawliczek at al., unpublished). Seeds were obtained from several sources including genebanks and breeding companies (Supplemental Table 5).

Genomic DNA extraction, quantification and pooling

Seeds were placed in a growth room in containers lined with moist paper towels. Ten days after germination a 20 mm long leaf segment was harvested from each plant. For each accession 96 plants were sampled. Leaf segments from 16 plants of the same accession were collected into one 2

mL centrifuge tube, with six tubes from 16 individual plants obtained for each accession. Collected leaves were freeze-dried in an Alpha 2-4 LDplus lyophilizator (Christ), for 18h at -60 °C, 0.011 mbar, followed by 1h at -64°C, 0,006 mbar and ground to fine powder using a laboratory mill MM 301 (Retsch) for 2.5-5 min at frequency 30.0 1/s. Genomic DNA was extracted using Mag-Bind Plant DNA DS Kit (OMEGA Bio-Tek) following manufacturer's protocol. Quality and quantity of DNA was assessed using spectrophotometry (NanoDrop2000, Thermo) and electrophoresis in 1% agarose gels stained with ethidium bromide. DNA concentration of each sample tube was adjusted to 100 ng and an equal volume of all samples from an accession were pooled together.

Primer design and PCR amplification of target genes.

Sequences of six target genes: aluminium activated citrate transporter (*MATE1*, also known as *AACT1*), taumatin-like protein (*TLP*), fructose-biphosphate aldolase (*FBA*), prolamine-box binding factor (*PBF*), secaloindoline-b (*Sinb*) and grain softness protein *(GSP-1*) were retrieved from GenBank (Supplemental Table 6). The entire sequences of *Secb* and *GSP-1* genes (456 and 506 bp, respectively) were amplified using primers described, respectively, by Simeone and Lafiandra (2005) and Massa et al. (2004). For genes *FBA*, *MATE1*, *PBF* and *TLP* primer pairs for generation of overlapping, ca. 600 bp long amplicons, covering the entire gene sequence were designed using Primer-BLAST (Ye et al., 2012). Primer pairs were tested using the DNA of rye inbred line L318 and those producing single product of expected length were used for amplification of gene fragments from pooled DNAs. PCR set up was as follows: 200 ng of template DNA, 2.5 mM $MgCl_2$, 0.2µM of each primer, 0.2 mM of each dNTP, 1x Dream *Taq* Green buffer, 0.5 U Dream *Taq* DNA polymerase (Thermo Scientific). The reactions were carried out in 25 µL in Mastercycler epgradient S (Eppendorf) thermal cyclers. For all primer pairs the thermal profile of initial denaturation for 60s at 95°C, 30 cycles of 30s at 95°C, 30s at 56°C and 60s at 72°C, followed by final extension for 5 min at 72°C was used. A volume of 5 µL from each reaction was used to check the amplification success using electrophoretic separation in 1.5% agarose gels stained with ethidium bromide. PCR products were shipped to Plant Breeding and Genetics Laboratory, Joint FAO/IAEA Division, International Atomic Energy Agency (Seibersdorf, Austria) for further processing.

PCR product quantification and pooling

PCR products were quantified using egel 96well gels (Thermo Fisher Scientific) and quantitative lambda DNA standards as previously described (Huynh et al., 2016). PCR product concentration was adjusted to 10 ng/ul in TE. All PCR products from a single gDNA pool were then pooled together. Pooled PCR products from each of the 95 accessions were then quantified using the Advanced Analytical® Fragment Analyzer™ with the low sensitivity 1kb separation matrix with 30 cm capillaries (Advanced Analytical®#DNF935). All sample pools were normalized to 30 nM concentration in TE prior to library preparation.

Library preparation and sequencing

Indexed DNA library for NGS was prepared using the TruSeq® Nano DNA HT Library Preparation Kit (Illumina, cat. 20015965) according to manufacturer's recommendation. Indexed libraries were then quantified using a Q-bit fluorometer (Thermo Fisher Scientific) and pooled together at an equal concentration. The pooled library was diluted to 18 pM concentration. Sequencing was performed on an Illumina MiSeq® using 2x300 PE chemistry according to manufacturer's protocol. The reads were de-multiplexed with the MiSeq Reporter software and were stored as FASTQ files for downstream analysis.

Sequence evaluation

FASTQ files were aligned to target amplicons using BWA mem with commands -M -t 16 (Li and Durbin, 2009). Amplicon fragments were given target names that were used throughout the NGS analysis (Supplemental Table 6). Samtools view was used to convert from SAM to BAM format (Li et al., 2009). Bam files are available in NCBI BioProject PRJNA593253

Coverage statistics were prepared using qualimap v.2.2.1-dev (García-Alcalde et al., 2012). Variant calling was performed using three algorithms CRISP, GATK and SNVer. Parameters used for CRISP were –OPE 0, --poolsize 192 and –qvoffset 33 (Bansal, 2010). The GUI of SNVer was used with the following parameters: -bq20,-mq17,-s0,-f0,-pbonferroni=0.1,-a0,-u30, -n192,-t0 (Wei et al., 2011). HaplotypeCaller (GATK 4.0) was used with default settings with the exception that ploidy was set to 192 (Poplin et al., 2017). For each method, VCF files from individual pools were merged using bcftools. Following this, read group information was unified between the three files using picard tools AddOrReplaceReadGroups function (http://broadinstitute.github.io/picard/index.html). Data for calculation of AF value for each variant and each accession was extracted using R libraries vcfR (Knaus and Grünwald, 2017) and

VariantAnnotation (Obenchain et al., 2014) and used to produce AF tables. The potential effect of nucleotide variation on gene function was evaluated with SNPeff (Cingolani et al., 2012). For this, a genome database was prepared using the build -genbank function. The effect of reported nucleotide variation was also evaluated with SIFT4G using a self-prepared genomic database with the fasta file of amplicon sequences used for mapping with BWA mem, a self-prepared gtf file and the uniref90 protein database (Vaser et al., 2016). Venn diagrams were produced using the R package eulerr (https://github.com/jolars/eulerr).

Validation of nucleotide variants

For validation of nucleotide variants CAPS assays were developed based on output of PARSESNP (Taylor and Greene, 2003), which provides a list of restriction endonuclease sites that are gained or lost due to the predicted SNV or indel. Serial Cloner 2.6.1. (http://serialbasics.free.fr/Serial_Cloner.html) software was used to digest *in silico* the gene fragment of interest and predict restriction patterns for reference and mutant alleles.
New batches of seeds were sown for selected accessions. Tissue harvest, DNA isolation and PCR reaction were done separately for each plant, using the procedures described above. Restriction digestion was done for 20 minutes using 10 µL of PCR reaction as template and 1 µL of the restriction enzyme in the total volume of 20 µL. FastDigest restriction enzymes (ThermoFisher) with dedicated buffers were used. The digestion products were separated in 6% denaturing polyacrylamide gels (if the predicted products were shorter than 200 bp or differed in length by less than 50 bp) and visualized by silver staining as described by Targońska et al. (Targońska et al., 2016), or in 1.5% agarose gels containing ethidium bromide. For Sanger sequencing-based validation of variants PCR reactions were sent to an external service provider. The analyzed plants were classified based on electrophoretic separation patterns/chromatograms as homozygous reference (RefRef), heterozygous (RefAlt), or homozygous variant (AltAlt). The variant frequency was calculated using the formula (RefAlt x 1 + Alt/Alt x 2)/ n x 2, where n is the total number of individuals analyzed.

Author Contributions:

Conceptualization, B.J.T. and H.B-B. ; Methodology, A.H., J.J-C., B.J.T. and H.B-B.; Investigation, A.H., L.B., K.T., E.B., J.J-C., P.G., A.K., B.J.T., and H.B-B.; Writing – Original Draft, A.H., B.J.T. and H.B-B.; Writing – Review & Editing, A.H., J.J-C., B.J.T., and H.B-B.; Funding Acquisition, B.J.T. and H.B-B.; Resources, B.J.T. and H.B-B.; Supervision, B.J.T. and H.B-B.

Acknowledgments:

This research was funded by the Polish National Science Center grant No. DEC-2014/14/E/NZ9/00285. Funding for DNA sequencing was provided by the Food and Agriculture Organization of the United Nations and the International Atomic Energy Agency through their Joint FAO/IAEA Program of Nuclear Techniques in Food and Agriculture. The authors declare that there is no conflict of interest.

Supporting Information:

Supplemental Table 1. Sequencing coverage for each nucleotide position in the experiment.

Supplemental Table 2. Total and private alleles by allele frequency class.

Supplemental Table 3. Allele frequencies and number of accessions harboring alleles of predicted deleterious variants common to GATK, SNVer and CRISP.

Supplemental Table 4. Sanger sequencing validation of variants in single plants from an accession.

Supplemental Table 5. Accessions used in this study.

Supplemental Table 6. Primer sequences used in this study.

References:

Al-Beyroutiova, M., Sabo, M., Sleziak, P., Dusinsky, R., Bircak, E., Hauptvogel, P., Kilian, A., and Svec, M. (2016). Evolutionary relationships in the genus Secale revealed by DArTseq DNA polymorphism. *Plant Syst. Evol.* **302**:1083–1091.

Balfourier, F., Bouchet, S., Robert, S., Oliveira, R. De, Rimbert, H., Kitt, J., Choulet, F., and Paux, E. (2019). Worldwide phylogeography and history of wheat genetic diversity. *Sci. Adv.* **5**:eaav0536.

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* **1**:318–324.

Bartos, J., Paux, E., Kofler, R., Havrankova, M., Kopecky, D., Suchankova, P., Safar, J., Simkova, H., Town, C. D., Lelley, T., 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 Biol.* **8**:95.

Bauer, E., Schmutzer, T., Barilar, I., Mascher, M., Gundlach, H., Martis, M. M., Twardziok, S. O., Hackauf, B., Gordillo, A., Wilde, P., et al. (2017). Towards a whole-genome sequence for rye (*Secale cereale* L.). *Plant J.* **89**:853–869.

Bolibok-Bragoszewska, H., Targonska, M., Bolibok, L., Kilian, A., and Rakoczy-Trojanowska, M. (2014). Genome-wide characterization of genetic diversity and population structure in Secale. *BMC Plant Biol.* **14**:184.

Cai, B., Li, Q., Liu, F., and Bi, H. (2018). Decreasing fructose-1 , 6-bisphosphate aldolase activity reduces plant growth and tolerance to chilling stress in tomato seedlings. *Physiol. Plant.* **163**:247–258.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., Ruden, D. M., Cingolani, P., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**:80–92.

Crespo-Herrera LA, Garkava-Gustavsson L, and Åhman Inger (2017). A systematic review of rye (Secale cereale L.) as a source of resistance to pathogens and pests in wheat (Triticum aestivum L.). *Hereditas* **154**:1–9.

Duitama, J., Kafuri, L., Tello, D., Leiva, A. M., Hofinger, B., Datta, S., Lentini, Z., Aranzales, E., Till, B., and Ceballos, H. (2017). Deep assessment of genomic diversity in cassava for herbicide tolerance and starch biosynthesis. *Comput. Struct. Biotechnol. J.* **15**:185–194.

FAO (2010). *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome.

Gamuyao, R., Chin, J. H., Pariasca-Tanaka, J., Pesaresi, P., Catausan, S., Dalid, C., Slamet-Loedin, I., Tecson-Mendoza, E. M., Wissuwa, M., and Heuer, S. (2012). The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* **488**:535–539.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., and Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics* **28**:2678–2679.

Gupta, P., Reddaiah, B., Salava, H., Upadhyaya, P., Tyagi, K., Datta, S., Malhotra, B., Thomas, S., Sunkum, A., Devulapalli, S., et al. (2017). Next-generation sequencing (NGS)-based identification of induced mutations in a doubly mutagenized tomato (Solanum lycopersicum) population. *Plant J.* **92**:495–508.

Gur, A., and Zamir, D. (2004). Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol.* **2**:e245.

Haseneyer, G., Stracke, S., Piepho, H., Sauer, S., Geiger, H. H., and Graner, A. (2010). DNA polymorphisms and haplotype patterns of transcription factors involved in barley endosperm development are associated with key agronomic traits. *BMC Plant Biol.* **10**:5.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J., et al. (2007). Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**:1522–1527.

Hoisington, D., Khairallah, M., Reeves, T., Ribaut, J.-M., Skovmand, B., Taba, S., and Warburton, M. (2002). Plant genetic resources: What can they contribute toward increased crop productivity? *Proc. Natl. Acad. Sci.* **96**:5937–5943.

Hussain, M., Iqbal, M. A., Till, B. J., and Rahman, M. (2018). Identification of induced mutations in hexaploid wheat genome using exome capture assay. *PLoS One* **13**:e0201918.

Huynh, O. A., Jankowicz-Cieslak, J., Saraye, B., Hofinger, B. J., and Till, B. J. (2016). Low-cost methods for DNA extraction and quantification. In *Biotechnologies for Plant Mutation Breeding* (ed. Jankowicz-Cieslak, J.), Tai, T.), Kumlehn, J.), and Till, B. J.), pp. 227–239. Cham: Springer.

Joukhadar, R., Daetwyler, H. D., Bansal, U. K., Gendall, A. R., and Hayden, M. J. (2017). Genetic diversity, population structure and ancestral origin of Australian wheat. *Front. Plant Sci.* **8**:1–15.

Keilwagen, J., Kilian, B., Özkan, H., Babben, S., Perovic, D., Mayer, K. F. X., Walther, A., Poskar, C. H., Ordon, F., Eversole, K., et al. (2014). Separating the wheat from the chaff - A strategy to utilize plant genetic resources from ex situ genebanks. *Sci. Rep.* **4**:14–18.

Knaus, B. J., and Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**:44–53.

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., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment / Map format and SAMtools. *Bioinformatics* **25**:2078–2079.

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.

Li, Y., Zhao, S., Ma, J., Li, D., Yan, L., Li, J., Qi, X., Guo, X., Zhang, L., He, W., et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**:579.

Li, F., Shimizu, A., Nishio, T., Tsutsumi, N., and Kato, H. (2019). Comparison and characterization of mutations induced by gamma-ray and carbon-ion irradiation in rice (Oryza sativa L .) using whole-genome resequencing. *G3 (Bethesda).* **9**:3743–3751.

Liu, H., Zhou, X., Li, X., Chen, J., Cui, D., and Chen, F. (2017). Molecular characterization of

secaloindoline genes in introduced CIMMYT primary hexaploid triticale. *Crop J.* **5**:430–437.

Lv, G.-Y., Guo, X.-G., Xie, L.-P., Xie, C.-G., Zhang, X.-H., Yang, Y., Xiao, L., Tang, Y.-Y., Pan, X.-L., Guo, A.-G., et al. (2017). Molecular Characterization, Gene Evolution, and Expression Analysis of the Fructose-1, 6-bisphosphate Aldolase (FBA) Gene Family in Wheat (Triticum aestivum L.). *Front. Plant Sci.* **8**:1–18.

Maraci, O., Ozkan, H., and Bilgin, R. (2018). Phylogeny and genetic structure in the genus Secale. *PLoS One* **13**:1–21.

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., and Bradbury, P. J. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci.* **110**:5241–5246.

Marroni, F., Pinosio, S., Di Centa, E., Jurman, I., Boerjan, W., Felice, N., Cattonaro, F., and Morgante, M. (2011). Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: Towards next-generation Ecotilling. *Plant J.* **67**:736–745.

Massa, A. N., Morris, C. F., Gill, B. S. (2004) Sequence diversity of puroindoline-a, puroindoline-b, and the grain softness protein in Aegilops tauschii Coss. *Crop Sci*. **44**:1808-1816

McCouch, S. (2004). Diversifying selection in plant breeding. *PLoS Biol.* **2**:e347.

McCouch, S. (2013). Feeding the future. *Nature* **499**:3–4.

Mehra, P., Pandey, B. K., and Giri, J. (2015). Genome-wide DNA polymorphisms in low Phosphate tolerant and sensitive rice genotypes. *Sci. Rep.* **5**:1–14.

Moehs, C. P., Austill, W. J., Holm, A., Large, T. A. G., Loeffler, D., Mullenberg, J., Schnable, P. S., Skinner, W., van Boxtel, J., Wu, L., et al. (2019). Development of decreased-gluten wheat enabled by determination of the genetic basis of lys3a barley. *Plant Physiol.* **179**:1692–1703.

Mondal, S., Rutkoski, J. E., Velu, G., Singh, P. K., Crespo-Herrera, L. A., Guzman, C. G., Bhavani, S., Lan, C., He, X., and Singh, R. P. (2016). Harnessing diversity in wheat to enhance grain yield, climate resilience, disease and insect pest resistance and nutrition through conventional and modern breeding approaches. *Front. Plant Sci.* **7**:991.

Nguyen, K. Le, Grondin, A., Courtois, B., and Gantet, P. (2019). Next-generation sequencing accelerates crop gene discovery. *Trends Plant Sci.* **24**:263–274.

Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**:2076–2078.

Pan, L., Shah, A. N., Phelps, I. G., Doherty, D., Johnson, E. A., and Moens, C. B. (2015). Rapid identification and recovery of ENU-induced mutations with next-generation sequencing and Paired-End Low-Error analysis. *BMC Genomics* **16**:83.

Pariasca-Tanaka, J., Chin, J. H., Dramé, K. N., Dalid, C., Heuer, S., and Wissuwa, M. (2014). A novel allele of the P-starvation tolerance gene OsPSTOL1 from African rice (Oryza glaberrima Steud) and its distribution in the genus Oryza. *Theor. Appl. Genet.* **127**:1387–1398.

Poplin, R., Ruano-Rubio, V., Depristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van Der, Kling, D. E., Gauthier, D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* **1**:http://dx.doi.org/10.1101/201178.

Purugganan, M. D., and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature* **457**:843–848.

Rabanus-Wallace, M. T., Hackauf, B., Mascher, M., Lux, T., Wicker, T., Gundlach, H., Houben, A., Mayer, K. F. X., Guo, L., Poland, J., et al. (2019). Chromosome-scale genome assembly provides insights into rye biology , evolution , and agronomic potential. *bioRxiv* Advance Access published 2019.

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLoS One* **8**:e66428.

Santos, E., Benito, C., Gallego, F. J., and Figueiras, A. M. (2018). Characterization, genetic diversity, phylogenetic relationships, and expression of the aluminum tolerance MATE1 gene in Secale species. *Biol. Plant.* **62**:109–120.

Schirmer, M., Ijaz, U. Z., Amore, R. D., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**:e37.

Schreiber, M., Himmelbach, A., Börner, A., and Mascher, M. (2018). Genetic diversity and relationship between domesticated rye and its wild relatives as revealed through genotyping-by-sequencing. *Evol. Appl.* Advance Access published 2018, doi:10.1111/eva.12624.

Sidhu, J. S., Ramakrishnan, S. M., Ali, S., Bernardo, A., Bai, G., Abdullah, S., Ayana, G., and Sehgal, S. K. (2019). Assessing the genetic diversity and characterizing genomic regions conferring Tan Spot resistance in cultivated rye. *PLoS One* **14**:1–22.

Simeone, M. C., and Lafiandra, D. (2005). Isolation and characterisation of friabilin genes in rye. *J. Cereal Sci.* **41**:115–122.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**:121–132.

Souza Jr., C. L. de (2012). Cultivar development of allogamous crops. *Crop Breed. Appl. Biotechnol.* **11**:8–15.

Targońska, M., Bolibok-Brągoszewska, H., and Rakoczy-Trojanowska, M. (2016). Assessment of genetic diversity in Secale cereale based on SSR markers. *Plant Mol. Biol. Report.* **34**:37–

51.

Taylor, N. E., and Greene, E. A. (2003). PARSESNP : a tool for the analysis of nucleotide polymorphisms. *Nucleic Acids Res.* **31**:3808–3811.

Tramontano, A., Jarc, L., Jankowicz-Cieslak, J., Hofinger, B. J., Gajek, K., Szurman-Zubrzycka, M., Szarejko, I., Ingelbrecht, I., and Till, B. J. (2019). Fragmentation of pooled PCR products for highly multiplexed TILLING. *G3 (Bethesda).* **9**:2657–2666.

Tsai, H., Howell, T., Nitcher, R., Missirian, V., Watson, B., Ngo, K. J., Lieberman, M., Fass, J., Uauy, C., Tran, R. K., et al. (2011). Discovery of rare mutations in populations: TILLING by sequencing. *Plant Physiol.* **156**:1257–1268.

UN (2019). *World Population Prospects 2019: Data Booklet*.

Varshney, R. K., Beier, U., Khlestkina, E. K., Kota, R., Korzun, V., Graner, A., and Börner, A. (2007). Single nucleotide polymorphisms in rye (Secale cereale L.): discovery, frequency, and applications for genome mapping and diversity studies. *Theor. Appl. Genet.* **114**:1105–16.

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* **11**:1–9.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**:43–49.

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer : a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* **39**:1–13.

Wendel, J. F., Jackson, S. A., Meyers, B. C., and Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biol.* **17**:1–14.

Wheeler, T., and von Braun, J. (2013). Climate change impacts on global food security. *Science* **341**:479–485.

Wilkinson, M. D., Tosi, P., Lovegrove, A., Corol, D. I., Ward, J. L., Palmer, R., Powers, S., Passmore, D., Webster, G., Marcus, S. E., et al. (2017). The Gsp-1 genes encode the wheat arabinogalactan peptide. *J. Cereal Sci.* **74**:155–164.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**:134.

Zhang, Z., Zheng, X., Yang, J., Messing, J., and Wu, Y. (2016). Maize endosperm-specific transcription factors O2 and PBF network the regulation of protein and starch synthesis. *Proc. Natl. Acad. Sci.* **113**:10842–10847.

Zhang, J., Wang, F., Liang, F., Zhang, Y., Ma, L., Wang, H., and Liu, D. (2018). Functional

analysis of a pathogenesis- related thaumatin-like protein gene TaLr35PR5 from wheat induced by leaf rust fungus. *BMC Plant Biol.* **18**:76.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**:408–414.

Figure Legends:

Figure 1. Target regions used in this study. Introns are colored blue, non-coding sequence grey and exons green. Relative nucleotide positions in base pairs are listed.

Figure 2. Scatter plot of allele frequency data from GATK HaplotypeCaller. Variants are grouped into six classes on the x-axis based on mean allele frequency (AF). Class A (blue dots) represents data with a mean AF from 1 to 0.7, green (B) from 0.69 to 0.4, orange (C) from 0.39 to 0.1, purple (D) from 0.099 to 0.069, red (E) from 0.068 to 0.038, and black (F) with a mean AF below 0.038. The y-axis shows the number of accessions harboring a specific variant. Each variant is plotted on a unique position on the z-axis.

Figure 3. Common and unique variants called by GATK, SNVer and CRISP. The Venn diagram shows the overlap of variant calls for the three algorithms (interior image). Eight hundred and ninety-five variants were commonly identified. The outer image is a circos plot of the common variants. Only the PCR amplified regions of gene targets are displayed (track1). Gene models are shown on track 2 with exons and introns represented by thick and thin black lines, respectively. Tracks 3, 4, and 5 show the position and frequency (indicated by bar height) of variants predicted by GATK, SNVer and CRISP, respectively.

Figures and Tables:

FIGURE 1

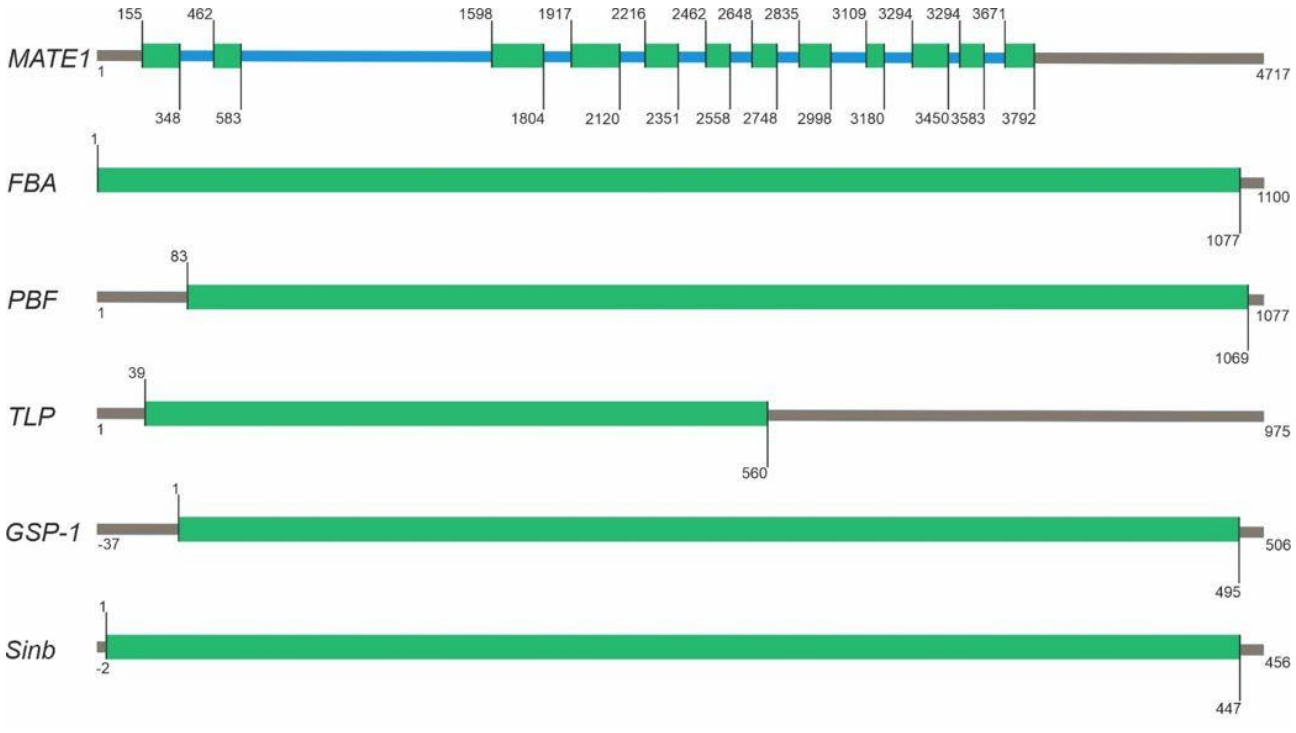

Figure 1. Target regions used in this study. Introns are colored blue, non-coding sequence grey and exons green. Relative nucleotide positions in base pairs are listed.

FIGURE 2

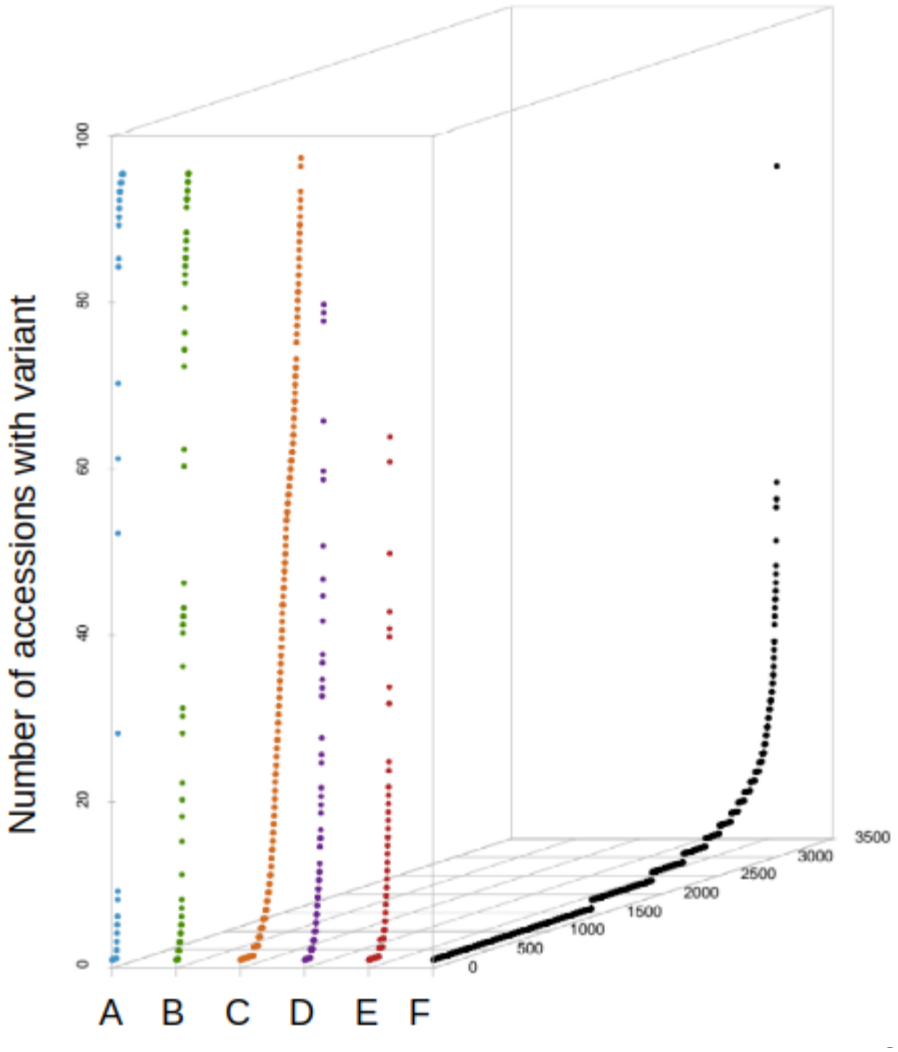

Figure 2. Scatter plot of allele frequency data from GATK HaplotypeCaller. Variants are grouped into six classes on the x-axis based on mean allele frequency (AF). Class A (blue dots) represents data with a mean AF from 1 to 0.7, green (B) from 0.69 to 0.4, orange (C) from 0.39 to 0.1, purple (D) from 0.099 to 0.069, red (E) from 0.068 to 0.038, and black (F) with a mean AF below 0.038. The y-axis shows the number of accessions harboring a specific variant. Each variant is plotted on a unique position on the z-axis.

FIGURE 3

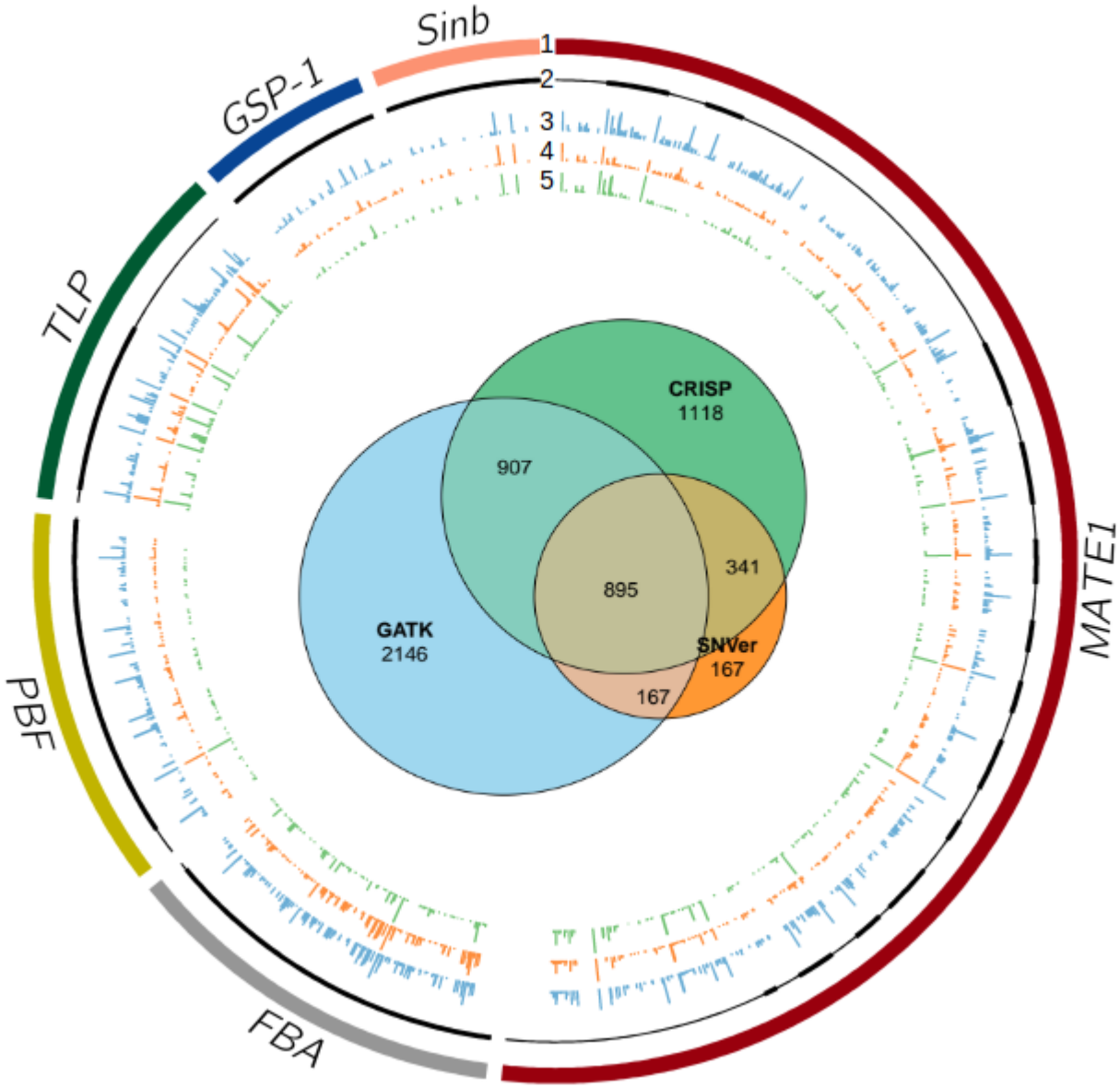

Figure 3. Common and unique variants called by GATK, SNVer and CRISP. The Venn diagram shows the overlap of variant calls for the three algorithms (interior image). Eight hundred and ninety-five variants were commonly identified. The outer image is a circos plot of the common variants. Only the PCR amplified regions of gene targets are displayed (track1). Gene models are shown on track 2 with exons and introns represented by thick and thin black lines, respectively. Tracks 3, 4, and 5 show the position and frequency (indicated by bar height) of variants predicted by GATK, SNVer and CRISP, respectively.

FIGURE 4

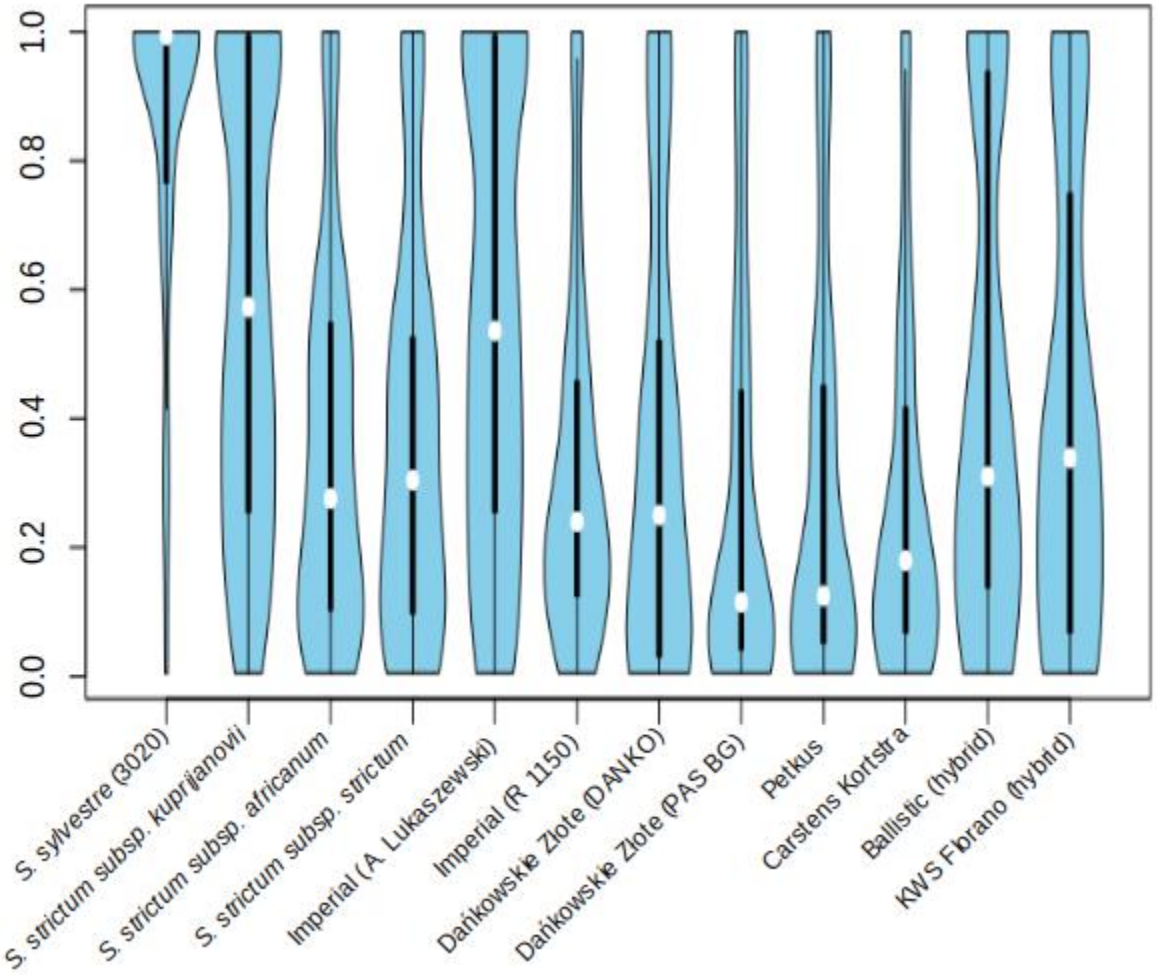

Figure 4. Violin plots of allele frequencies for variants predicted in common by GATK HaplotypeCaller, SNVer and CRISP. Data from HaplotypeCaller are shown for selected accessions.

Table 1. Missense, nonsense and silent changes with different variant calling methods

	GATK	SNVer	CRISP	Common variants
Missense	1183	336	868	164
Nonsense	14	7	9	2
Total	1770	602	1322	348

Table 2. CAPS and Sanger validation of variants in multiple plants of an accession

Gene	Pos[a]	Ref[b]	Alt[c]	Method	RE used	Acc. [d]	GATK[e]	SNVer[e]	CRISP[e]	AFobs[f]	No. plants
MATE 1	170	A	G	CAPS	*NotI*	D2	0.880	0.587	0.819	0.85	10
MATE 1	170	A	G	CAPS	*NotI*	E12	0.875	0.592	0.783	0.86	11
MATE 1	210	A	G	CAPS	*TaqI*	H5	0.172	0.079	0.276	0.38	13
MATE 1	364	G	C	CAPS	*MboI*	H5	0.307	0.137	0.393	0.19	13
PBF	310	C	T	CAPS	*MnlI*	D2	0.292	0.206	0.165	0.47	12
PBF	310	C	T	CAPS	*MnlI*	E12	0.120	0.059	0.059	0.00	11
PBF	517	G	A	CAPS	*MboI*	D2	0.286	0.262	0.180	0.42	12
PBF	517	G	A	CAPS	*MboI*	E12	0.016	0.104	0.065	0.00	11
PBF	532	C	T	CAPS	*FokI*	D2	0.104	0.096	0.138	0.00	12
PBF	532	C	T	CAPS	*FokI*	E12	0.536	0.405	0.472	0.41	11
PBF	666	C	T	*Sanger*	na[g]	F8	0.401	0.371	0.359	0.38	12
PBF	810	C	T	*Sanger*	na	F10	0.042	0.022	0.068	0.00	6
PBF	810	C	T	*Sanger*	na	F11	0.214	0.074	0.216	0.16	16
PBF	846	G	C	*Sanger*	na	F8	0.094	0.053	0.104	0.17	12
PBF	847	G	A	*Sanger*	na	F8	0.094	0.064	0.100	0.17	12
SecB	211	A	G	CAPS	*FokI*	H5	0.026	0.183	0.111	0.00	13

[a] nucleotide position, [b] reference sequence, [c] variant sequence, [d] accession code, algorithm predicted allele frequency, [f] observed allele frequency, [g] not applicable

Supplemental Information