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

Anna Hawliczek1, Leszek Bolibok2, Katarzyna Tofil1, Ewa Borzęcka1, Joanna Jankowicz-Cieślak3, Piotr Gawroński1, Adam Kral1, Bradley J. Till3,4† and Hanna Bolibok-Brągoszewska1,*†

Abstract

Background: 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 L.).

Results: Using the amplicon sequencing approach 95 rye accessions of different improvement status and worldwide origin, each represented by a pooled sample comprising DNA of 96 individual plants, were evaluated for sequence variation in six candidate genes with significant functions on biotic and abiotic stress resistance, and seed quality. Seventy-four predicted deleterious variants were identified using multiple algorithms. Rare variants were recovered including those found only in a low percentage of seed.

Conclusions: We conclude that this approach provides a rapid and flexible method for evaluating stock heterogeneity, probing allele diversity, and recovering previously hidden variation.

A large extent of within-population heterogeneity revealed in the study provides an important point for consideration during rye germplasm conservation and utilization efforts.

Keywords: Secale cereale, Natural variation, Allele frequency, Variant calling, MATE1, FBA, TLP, GSP-1, Sinb, PBF
Background

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 [1–3]. 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 [4–7].

Luckily, the value of PGR as a reservoir of gene variants was recognised over a hundred years ago [8] and nowadays there are over 1700 ex situ germplasm collections worldwide, maintaining about 7.4 million accessions. Approximately 62% of these accessions are landraces and wild species [9]. 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 [5, 9, 10].

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 within gene sequences such as whole genome resequencing (WGS) [11–13] and exome capture [14–16]. 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 [17]. 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 [18]. 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. Compared to WGS or exome capture, amplicon approaches allow acquisition of a much higher coverage of the selected target bases pairs at a lower sequencing cost. This is because the total yield of the sequencing reaction, in terms of raw bases, is distributed to fewer unique bases of each sample in the pool (e.g. [19]). One application of amplicon sequencing is the simultaneous genotyping of hundreds of unique samples independently by employing strategies to barcode, or index, each sample uniquely [20]. In addition to this approach, the high sensitivity of current sequencing technologies enables “ultra deep” methods whereby nucleotide variants can be identified in samples containing pools of mixed genotypes. One example is the detection of rare somatic mutations in human samples [21]. Another example is the use of amplicon sequencing to measure intrahost virus diversity. Researchers showed that a rare Zika virus variant could be detected if present at >3% in a mixed sample when sequencing coverage was at least 400x [22]. In plants, experiments can be designed to discover rare nucleotide variants present at very low frequencies by screening large populations where genomic DNA has been pooled prior to PCR amplification and sequencing. Screening throughputs are increased and assay costs are reduced, making screening thousands of samples practical. This has been used for recovery of induced point mutations in TILLING by Sequencing assays [23]. 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 [23–26]. The approach has been adapted for recovery of natural variation in Populus nigra, Manihot esculenta Crantz (cassava), and Oryza sativa L., 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 [27]. In cassava, DNA from up to 281 accessions were pooled prior to sequencing for variants in starch synthesis pathway-related genes and herbicide tolerance genes in 1667 accessions [28]. In rice, pooling of DNAs prepared from 233 breeding lines was followed by sequencing for variants in starch synthesis genes [29]. Pooling of multiple samples from the same species has also been used in studies where WGS has been applied. There are many variations to this methodology that has been termed Pool-sequencing [30]. This includes cases where, contrary to TILLING assays, multiple individuals with similar genotypes are pooled together to estimate population allele frequencies. In such applications, sequencing coverage can be reduced to save costs, but are insufficient to find rare alleles in one or few individuals in the pool. Sequencing intra-accession pools has also been described such as in metagenomics studies [31].

Rye (Secale 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 [32]. Genetically rye is a diploid (n = 7), with a large (ca. 8 Gbp) and complex genome [33, 34]. There are over 21 thousand rye accessions in genebanks worldwide, approximately 35% of them are landraces and wild species [9]. Several studies on genome-wide diversity in rye were published to date [35–38]. 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 [35, 36, 38, 39]. To date neither NGS-based targeted amplicon sequencing, nor any other method of gene variant discovery was applied to rye genetic resources.

Abiotic and biotic stress resistance, and yield constitute the key targets in rye breeding [40, 41]. Although the number of well characterized rye genes is very limited [42], there are important candidate genes related to biotic and abiotic stress resistance and grain quality to consider. MATE1 (multidrug and toxic compound extrusion, also known as AACT1 - aluminum activated citrate transporter), 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 [43]. Rye is one of the most Al-tolerant cereals, with the degree of tolerance depending on the allelic variant of MATE1 [44]. TLPs (taumatin-like proteins) are a family of pathogenesis-related (PR) proteins, involved in fungal pathogen response in many plant species [45]. FBA (fructose-bisphosphate aldolase) is one of the key metabolic enzymes involved in CO₂ 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 [46, 47]. GSP-1 (grain softness protein) genes, belonging to the prolamine 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 [48]. Secaloindolines, products of genes Sin(a) and Sin(b), are main components of friabilin - a starch-associated protein fraction of cereal grains [49]. The wheat orthologues of Sin(a) and Sin(b), called Pina and Pinb, are key determinants of grain texture, an important breeding trait directly influencing the end-use [50]. PBF (prolambox binding factor) is an endosperm specific transcription factor involved in the regulation of protein and starch synthesis [51]. It binds to the prolambox 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 [52], while in wheat, mutating the homologous PBFs using TILLING resulted in a markedly decreased gluten content and high content of lysine [53].

Exploration of genic variation in outcrossing, generically propagated crops, such as rye, maize (Zea mays), sugar beet (Beta vulgaris), broccoli (Brassica oleracea var. italica), or carrot (Daucus carota), is a particularly demanding task. Natural, random-mating populations of such species are heterozygous and heterogeneous, with multiple alleles of a locus being present [54]. 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 heterogeneous 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 96 plants per accession. These were subjected to pooled amplicon sequencing in six target genes implicated in biotic and abiotic stress resistance and seed quality: MATE1, TLP, FBA, PBF, Sinb, and GSP-1. Three variant calling algorithms (GATK HaplotypeCaller [55],
SNVer [56] and CRISP [57]) 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 predicted to affect protein function that are 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 representing different germplasm groups: wild species, landraces, historical and modern cultivars.

Results
DNA sequencing, mapping and coverage
Pooled amplicon sequencing using Illumina sequencing by synthesis 2 × 300 paired end reads on 95 accessions and six genes produced a mean coverage of 13,948x and mean mapping quality of 58.65. Mean coverage per accession pool varied approximately 10 fold, between 2924x and 30,275x. Analysis of sequencing coverage at each nucleotide revealed that 94.2% of the experiment produced 20 or more reads to support a rare variant present at 5% in the DNA pool (Additional file 1: Table S1).

Evaluation of variant calling algorithms and predicted effects of nucleotide changes
Variant calling was first performed on each pool using HaplotypeCaller in GATK (v.4.0) with ploidy set to 192 in order to recover rare alleles. This resulted in 4115 called variants, of which 3682 were single nucleotide polymorphisms, 192 insertions, and 241 deletions. Evaluation of the Variant Call Format (VCF) file, allowed calculation of the frequency of a specific allele within the DNA pool created from the 96 seeds that were sampled to represent an accession. This is referred to as VAF (Variant Allele Frequency), to distinguish the measurement from AF (Allele Frequency) - the frequency of the allele within the set of 95 accessions analyzed in the present study. Data was plotted to evaluate the distribution of the mean VAF for each variant and the number of accessions harboring each discovered allele (Fig. 1).

Deleterious alleles with a high maximum VAF (the highest VAF reported in an accession) and present in only one accession were recovered along with alleles with a maximum VAF 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 VAF between 0.047 and 0.391 and are found in 1 to 21 accessions (Additional file 4: Table S2, Additional file 5: Figure S3).

Within target genes, 18 to 443 polymorphic positions were detected consistently by the three algorithms, corresponding to one SNP or InDel every 8-10 bp of sequence for five of the analyzed genes (Additional file 7: Table S3). 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 22th June 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 first assayed using Sanger sequencing of Sinb amplicons in a single individual plant from each of eight accessions, with the aim of evaluating variant prediction while keeping Sanger sequencing costs low. Twelve variants were predicted in this set. Only variants reported by all three algorithms for the tested accession, and where the lowest VAF was greater than 0.295 were validated (Additional file 8: Table S4). 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. 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, 13 out of 16 tested variants were recovered when sampling between six and 27 plants (Table 2). Observed allele frequencies calculated from the number of plants harboring the

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             |

Fig. 2 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 (track 1). 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.
tested sequence difference varied from the frequencies predicted from the amplicon sequencing data. Seven variants had observed VAF closest to GATK predictions, two variants were closest with SNVer and four with CRISP. The three 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.

Phylogenetic relationships between populations and comparison of VAF distributions

The relationship between accessions was evaluated by creating a Neighbor Joining (NJ) tree based on Nei’s genetic distance (Fig. 3). This resulted in accessions divided into six clusters (I–VI), with cluster I containing mostly cultivars, including the majority of modern cultivars analyzed. Nevertheless, a coincidence of the clustering with improvement status could not be observed. The accessions: *S. sylvestre* (abbreviation B6 in Fig. 3), *S. strictum* subsp. *kuprijanovii* (F8) and *S. strictum* subsp. *africanum* (B7), were indicated as the most divergent of the analyzed set, which is in agreement with the outcome of NJ clustering. The accessions: *S. sylvestre* (B6), *S. strictum* subsp. *kuprijanovii* (F8), *S. strictum* subsp. *africanum* (B7), and the sample of historical variety Imperial (B4) are very distant from the rest, while within the group of the remaining accessions several subgroups can be observed, which correspond to clusters indicated in the NJ tree (Fig. 3).

Private variants occurred in all germplasm groups included in the study: modern cultivars, historic cultivars, landraces and wild accessions. In the group of commonly identified variants, 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 VAF values (mean 0.4–0.45, median 0.22–0.28). Ten of the private variants detected in wild accessions were putatively deleterious: five in the *PBF* gene, four in *MATE1* and one in *GSP-1*. In the remaining germplasm groups mean VAF and median VAF did not exceed 0.12 and 0.007, respectively. The

| Table 2 | CAPS and Sanger validation of variants in multiple single plants of an accession |
|---------|------------------------------------------|
| Gene    | Pos | Ref | Alt | Method | RE used | Acc | GATK | SNVer | CRISP | VAFobs | No. plants |
|---------|-----|-----|-----|--------|---------|-----|------|-------|-------|--------|------------|
| MATE1   | 170 | A   | G   | CAPS   | NotI    | D2  | 0.880 | 0.587 | 0.819 | 0.79   | 26[11]     |
| MATE1   | 170 | A   | G   | CAPS   | NotI    | E12 | 0.875 | 0.592 | 0.783 | 0.70   | 27[8]      |
| MATE1   | 210 | A   | G   | CAPS   | TaqI    | H5  | 0.172 | 0.079 | 0.276 | 0.38   | 25[15]     |
| MATE1   | 364 | G   | C   | CAPS   | Mbol    | H5  | 0.307 | 0.137 | 0.393 | 0.24   | 25[8]      |
| PBF     | 310 | C   | T   | CAPS   | MnII    | D2  | 0.292 | 0.206 | 0.165 | 0.44   | 25[14]     |
| PBF     | 310 | C   | T   | CAPS   | MnII    | E12 | 0.120 | 0.059 | 0.059 | 0.10   | 25[5]      |
| PBF     | 517 | G   | A   | CAPS   | Mbol    | D2  | 0.286 | 0.262 | 0.180 | 0.44   | 27[12]     |
| PBF     | 517 | G   | A   | CAPS   | Mbol    | E12 | 0.016 | 0.104 | 0.065 | 0.09   | 27[5]      |
| PBF     | 532 | C   | T   | CAPS   | FokI    | D2  | 0.104 | 0.096 | 0.138 | 0.00   | 26[0]      |
| PBF     | 532 | C   | T   | CAPS   | FokI    | E12 | 0.536 | 0.405 | 0.472 | 0.43   | 25[14]     |
| PBF     | 666 | C   | T   | Sanger | na      | F8  | 0.401 | 0.371 | 0.359 | 0.58   | 25[11]     |
| PBF     | 810 | C   | T   | Sanger | na      | F10 | 0.042 | 0.022 | 0.068 | 0.00   | 6[0]       |
| PBF     | 810 | C   | T   | Sanger | na      | F11 | 0.214 | 0.074 | 0.216 | 0.16   | 16[5]      |
| PBF     | 846 | G   | C   | Sanger | na      | F8  | 0.094 | 0.053 | 0.104 | 0.10   | 25[5]      |
| PBF     | 847 | G   | A   | Sanger | na      | F8  | 0.094 | 0.064 | 0.100 | 0.08   | 25[4]      |
| Sinb    | 211 | A   | G   | CAPS   | FokI    | H5  | 0.026 | 0.183 | 0.111 | 0.00   | 25[0]      |

* = nucleotide position
| = reference sequence
| = variant sequence
d = accession code
a = algorithm predicted allele frequency (VAF)
| = observed allele frequency
| = numbers in brackets indicate the number of heterozygous individuals
| = not applicable
number of private variants varied from 0 (32 accessions) to 25 in *S. sylvestre* (B6) (Fig. 3). Among the cultivated rye (*S. cereale* subsp. *cereale*) accessions, the highest number of private variants (7) was observed in a landrace from Bosnia and Herzegovina (C2) and also in a *S. cereale* subsp. *cereale* accession of unknown improvement status from Israel (E4).

VAF values were used to prepare combined strip/violin plots in order to qualitatively compare accessions. Several distinct patterns of allele frequency distribution were observed (representative examples shown in Fig. 5). Based on the results of two-part Wilcoxon test of pairwise comparisons of VAF distributions, rye accessions were grouped into 20 clusters ranging in size from 1 to 16 (Additional file 10: Figure S6). Five accessions, characterized by a high proportion of variants with high VAF values, were consistently recognized as markedly different from the rest: *S. sylvestre* (B6), *S. strictum* subsp. *kuprijanovii* (F8), historic cultivars Imperial (B4) and Otello (G12), and landrace R1040 (F6) (Additional file 11: Table S5).

**Discussion**

To evaluate the distribution of frequencies of alleles within a landrace or cultivar, we chose to sample 96 plants from each accession of rye selected for our study. This allows the recovery of i) sequence differences compared to the reference sequence used that are fully
Fig. 4 PCoA plot based on Nei’s genetic distance matrix derived from VAF values, showing genetic relationships between 95 rye accessions. VAF values reported by GATK for 895 variants detected in common by tree algorithms were used. Colors indicate clusters in NJ tree (Fig. 3). To simplify output accession codes are shown only for the most distant accessions.

Fig. 5 Combined strip/violin plots of GATK VAF values for variants predicted in common by GATK HaplotypeCaller, SNVer and CRISP. Horizontal bars indicate median values. Representative accessions are shown. Combined strip/violin plots for all 95 accessions included in the study are presented in Additional file 10: Figure S6.
were at an allele frequency below 40% [59]. When se-
with settings for diploid samples, Li et al. reported that
ducing the sensitivity in order to lower false positive er-
variations in diploid samples. Even in optimized diploid
ings were found in all DNA pools suggesting that each pool
was suitable for PCR amplification and sequencing. De-
viations in coverage values, therefore, likely resulted
from differences associated with the quantification,
normalization and pooling of PCR products. Such vari-
ations were recently reported in a study comparing to-
mato, cassava and barley amplicon sequencing data sets
[26]. The study revealed that minor coverage improve-
ments could be achieved through the addition of extra
quantification methods. Alternative approaches to in-
crease 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 sequen-
cing 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 re-
ducing the sensitivity in order to lower false positive er-
ors. 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% [59]. When se-
quencing non-pooled samples, setting an allele fre-
cuency threshold of >40% for heterozygous variants
therefore reduced false positive errors. In non-pooled
samples, the choice of mapping software and variant
calling software can also affect predicted SNPs. Yao and
colleagues used whole exome capture wheat data sets
and seven variant calling tools to define putative true
variants that were identified by all tools [16]. Using this
set, the authors concluded that mapping with BWA
mem outperformed Bowtie2. Variant callers showed
variable performance with GATK Haplotype Caller out-
performing SNVer and Samtools/mpileup performing
best. Independent validation of SNPs by Sanger sequen-
cing was not carried out. Optimizing variant calling may
be more challenging in highly pooled samples. Algo-
rithms 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 [25]. That work described technical differ-
ences between different algorithms and concluded that
accuracy is improved when a variant call is predicted by
at least two algorithms. In cassava, up to 281 different
accessions were pooled together prior to sequencing in
an approach designed to quickly identify putative dele-
terious alleles [28]. 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. Similar to previous stud-
ies with pooled samples and wheat exome capture data,
multiple variant callers were used to find concordant
SNPs. The rye 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 ac-
cession (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 rese-
quencing as alleles can be recovered through assay of a
single seed [60, 61]. 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 mul-
tiple 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.
Importantly, the presence of variants that segregate at a
low frequency within an accession, and are never found
at high frequency in any tested accession, suggest that a
broader genetic diversity can exist in germplasm collec-
tions than previously known. We expect this to be most
common in outcrossing species like rye where admix-
tures of alleles are frequent.

Variants with mean VAF between 0.7 and 1 repre-
sented between 1.87 and 3.06% of all predicted alleles,
depending on the algorithm used. In this set of variants,
between 41 and 49% are private alleles found in only
one accession (Fig. 1, Additional file 2: Figure S1, Addi-
tional file 3: Figure S2). The highest number of variants
were found in the lowest VAFs. It is expected that false
positive errors will increase as the number and percent-
age of reads supporting the alternative allele decreases.
Studies have been carried out on errors associated with
MiSeq paired end sequencing, but a thorough investiga-
tion into errors in pooled samples has not been reported
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 VAF of 0.038 or lower, a reduction of more than 50% from the data from any single algorithm. Further experiments are required to determine what, if any, percentage of the sub 0.038 VAF 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 a VAF 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 VAF false positive errors. Sixty-five percent of sub 0.038 VAF variants commonly predicted by all algorithms were found in more than one accession. However, 96% had a maximum predicted VAF of less than 10%, and the highest maximum VAF 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 algorithms may reduce errors and serves as a useful method to prioritize alleles for further study.

We found qualitative evaluation of VAF values using strip and 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 S. sylvestre (B6). Molecular marker-based analyses of genetic diversity indicated this self-pollinating taxon as the most divergent in genus Secale [35, 58, 63]. Its large proportion of high VAF variants (Fig. 5) 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 (F8), is a perennial outbreeder, also genetically divergent from S. cereale. However, its violin plot differs markedly from plots obtained for the other two S. strictum samples included in the study, S. strictum subsp. africanum (B7) and S. strictum subsp. strictum (B9), which might indicate a sample tracking mistake during genebank preservation or laboratory handling, or a bottleneck during preservation. A sample of Imperial cultivar (B4), widely used in cytological studies, originating from the collection of A. J. Lukaszewski (UCLA, Riverside), showed an approximately equal abundance of variants with all possible VAF values and differed clearly from another sample of Imperial (C1), obtained from IPK Gatersleben genebank. Less pronounced (although also statistically significant) differences were also observed between the two samples of cultivar Dankowskie Zlote (C6 and F3), obtained from different sources. Samples of hybrid cultivars from KWS, such as Ballistic (D11) and KWS Florano (G10), exhibited a higher percentage of VAF 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 (F3 and C6), Petkus (G8), or Carstens Kortstra (E8), which is consistent with the use of the three line system in the development of hybrid rye cultivars. Statistical analysis also showed that wild accessions differed from modern varieties in terms of VAF value distributions, with wild accessions (accession codes F8, B7, B6, B10, B9, A9, F4, F7, E12, full names in Additional file 11: Table S5) always located in different clusters than modern varieties (accession codes F12, G11, F2, F1, G10, G9, D11, B12) in the dendrogram in Additional file 10: Figure S6. However a trend in median values differentiating wild accessions from modern varieties could not be observed.

In this study we analyzed six genes linked to biotic and abiotic stress resistance and seed quality. Using deep sampling and pooled amplicon sequencing numerous new variants were identified, (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/52 bp [64], 1 SNP/58 bp [65] or 1 SNP or InDel/31 bp [66]. In agreement with the results of previous genome-wide, DArT-marker based characterization of genetic diversity in rye [35], 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. On the other hand, the clustering of the accessions in the NJ tree generated based on the VAF of 895 variants detected in common did not agree with the improvement status of the accessions, suggesting that selective pressures other than breeding practices have influenced the diversity of the genes analyzed.

This study also points out that, in case of open pollinated populations (due to the high within-accession variability), the sampling of a single individual or a small number of individuals from an accession most likely results in an inaccurate and perhaps even misleading representation of genetic relationships between the accessions. This can be seen in NJ trees produced based on conversion of VAF values into genotyping-like scores, where a different clustering of accessions was observed at each range of VAF used to define heterozygous variants (Additional file 9: Figure S5).

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 described above). Homozygous variants present in more than 90 accessions were also recovered. Interestingly, putative deleterious variants were also identified where the maximum VAF 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). This suggests that alleles are segregating within rye accessions at low fractions that may affect gene function and potentially plant phenotype. Such variants would go undiscovered in conventional GBS or WGS assays where only one or two seed per accession are sampled, and may be useful for functional genomic characterizations and breeding. 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 aluminum in the soil). Possible mechanisms for the maintenance of rare low frequency alleles in populations, including meiotic effects, can also be investigated.

The rye amplicons used in this study were generated before the release of the rye genome [34]. It is expected that the recent release of the rye reference genome will enable improvements in gene target selection and primer design. The broadening of the genetic basis has been identified as one of the most important goals in rye hybrid breeding [40, 67], however, introduction of PGR into a breeding program is often challenging [68]. The experimental protocol validated in this study provides a means to rapidly and effectively screen numerous accession samples for the genes of interest and identify desired variants. Therefore it has the potential to advance the use of exotic and primitive germplasm for targeted broadening of variation in breeding schemes.

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.

Methods

Plant material

Ninety-five accessions of rye, each represented by a pooled sample comprising DNA of 96 individual plants, were analyzed in the study. This set included 90 accessions of S. cereale, among them 8 modern cultivars, 34 historic cultivars, 35 landraces, and 5 accessions of other Secale taxa, representing various geographic regions. In total 10 accessions from this set were described as wild/weedy. Seeds were obtained from several sources including genebanks and breeding companies (Additional file 11: Table S5).

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 18 h at –60 °C, 0.011 mbar, followed by 1 h 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: multidrug and toxic compound extrusion (MATE1, also known as AACT1 - aluminium activated citrate transporter), taumatin-like protein (TLP), fructose-biphosphate aldolase (FBA), prolamin-box binding factor (PBF), secaloindoline-b (Sinb) and grain softness protein (GSP-1) were retrieved from GenBank (Additional file 12: Table S6, Additional file 13: Figure S7). The entire sequences of Sinb and GSP-1 genes (456 and 506 bp, respectively) were amplified using primers described, respectively, by Simeone and Lafiandra [49] and Massa et al. [69]. 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 [70]. 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. Primer design and all other assays described in this work were carried out before the public release of the rye genome. PCR set up was as follows: 200 ng of template DNA, 2.5 mM MgCl₂, 0.2 mM 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 gel 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 1 kb 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. 20,015,965) 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 2 × 300 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 (Additional file 14: Table S7).

**Sequence evaluation**

FASTQ files were aligned to target amplicons using BWA mem (Version: 0.7.17-r1188) with commands -M -t 16 [71]. Amplicon fragment sequences were derived from public databases prior to the release of the rye reference genome. These were given target names that were used throughout the NGS analysis and the sequences are referred to as homozygous reference sequence throughout the manuscript (Additional file 12: Table S6). Samtools view (Version 1.7) was used to convert from SAM to BAM format [72]. BAM files are available in NCBI BioProject PRJNA593253. Coverage statistics were prepared using qualimap (v.2.2.1-dev) [73]. Variant calling was performed using three algorithms CRISP (Version 0.1), GATK (Version 4.0.1.2) and SNVer (Version 0.5.3). Parameters used for CRISP were −q offset 0, −poolsize 192 and −q offset 33 [57]. The GUI of SNVer was used with the following parameters: -bq 20, -mq 17, -s 0, -f 0, -pbonferroni = 0.1, -a 0, -u 30, -n 192, -t 0 [56]. HaplotypeCaller (GATK) was used following best practices with default settings with the exception that ploidy was set to 192 [55]. 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 allele frequency from the VCF files (called VAF in this manuscript) was extracted for each variant and each accession using R libraries vcfR [74] and VariantAnnotation [75] and used to produce AF tables. The potential effect of nucleotide variation on gene function was evaluated with SNPeff [76]. 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 [77]. Venn diagrams were produced using the R package eulerr (https://github.com/jolars/eulerr).

**Evaluation of VAF distributions**

For the variants detected in common by three algorithms, distributions of VAF values reported by GATK were compared pairwise using the two-part Wilcoxon test [78] resulting in a pairwise matrix of 0 s and 1 s, with 1 indicating that for the given pair of populations the distributions of VAF values are different at a = 0.05. This matrix was then used for hierarchical clustering analysis with the hclust function of the R package stats. Combined strip/violin plots were drawn using R libraries ggplot2 [79], ggbeeswarm and ggdendro.

**Evaluation of phylogenetic relationships between accessions**

For the purpose of illustrating the relationships between rye populations analyzed, a Nei's genetic distance [80] matrix was calculated using POPTREEW [81] using VAF values reported by GATK for the variants detected in common by three algorithms and imported into MEGA 5.2 [82] to produce a Neighbor Joining dendrogram. The Nei's genetic distance matrix was also used as input to perform a principal coordinates analysis with the 

tan function of the R package stats. This matrix was then used for hierarchical clustering analysis with the hclust function of the R package stats. Combined strip/violin plots were drawn using R libraries ggplot2 [79], ggbeeswarm and ggdendro.

**Validation of nucleotide variants**

For validation of nucleotide variants CAPS assays were developed based on output of PARSESNP [86], 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 several accessions, where the predicted variant resulted in gain or loss of a restriction enzyme recognition site in amplicons of genes MATE1, PBF and Sinh. Tissue harvest, DNA isolation and PCR reaction were done separately for each plant, using the procedures described above. The number of individual plants ranged from 10 to 27, depending on the availability of seeds after the initial issue collection for NGS amplicon sequencing experiments. Restriction digestion was done for 20 min 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 Tar-gońska et al. [36], or in 1.5% agarose gels containing ethidium bromide. For Sanger sequencing-based validation of variants, amplicons of PBF gene from six to 27 plants per accession, obtained as described above, were sent to an external service provider. Sequencing was done on an automated sequencer using fluorescent dye terminator chemistry. 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.

**Supplementary Information**

The online version contains supplementary material available at https://doi.org/10.1186/s12864-020-07240-3.

---

**Additional file 1: Table S1.** Sequencing coverage for each nucleotide position in the experiment.

**Additional file 2: Figure S1.** Percentage of private alleles (found in only one of the tested accessions) plotted by variant allele frequency (VAF). Data from GATK is plotted in light blue, CRISP in green and SNVer in orange.

**Additional file 3: Figure S2.** Scatter plots of variant allele frequency (VAF) data. VAF is plotted on the x-axis. Data is plotted on the z-axis to separate different variants that share the same VAF and number of accessions. The percentage of the total data from VAF 0 to a specific frequency is overlaid in red. Variants predicted by CRISP are plotted in panel A, and by SNVer in panel B.

**Additional file 4: Table S2.** Allele frequencies and number of accessions harboring alleles of predicted deleterious variants common to GATK, SNVer and CRISP.

**Additional file 5: Figure S3.** Venn diagram of variants called by GATK, SNVer and CRISP predicted to be deleterious using SIFT.

**Additional file 6: Figure S4.** Lollipop chart of allele frequencies of GATK, SNVer and CRISP. Each variant is assigned an arbitrary number (x axis) with maximum allele frequency values calculated from GATK VCF data is plotted on the y axis. Data is sorted into 5 distinct groups based on the number of accessions harboring the variant. This sorting is indicated by the colored ball at the end of the bar. Allele frequencies below 0.039 are not plotted.
Additional file 7: Table S3. Number of polymorphic positions detected in common by three algorithms per gene.

Additional file 8: Table S4. Sanger sequencing validation of variants in single plants from an accession.

Additional file 9: Figure S5. NJ dendrograms based on conversion of VAF values reported by GATK for variants identified in common into genotype scores. VAF values were converted into genotype scores: “0” = reference allele homozygote; “1” = variant allele homozygote; “2” = heterozygote using the following settings: A) VAF < 0.3 = 0; VAF ≥ 0.7 = 1; and values in between (greater than 0.3 and less than 0.7) = 2; ii) VAF < 0.4 = 0; VAF ≥ 0.6 = 1; and values in between = 2; and iii) VAF < 0.2 = 0; VAF ≥ 0.8 = 1; in between = 2. Colors of the nodes correspond to colors of the clusters in the NJ dendrogram derived from VAF data (Fig. 3, main manuscript) and indicate membership of the respective accessions in the clusters of the NJ dendrogram derived from VAF data.

Additional file 10: Figure S6. Dendrogram showing relationships between distributions of VAF values (shown as strip/violin plots) for 95 accessions (based on results of the two-part Wilcoxon test). Horizontal bars indicate median values.

Additional file 11: Table S5. Accessions used in this study.

Additional file 12: Table S6. Primer sequences used in this study.

Additional file 13: Figure S7. 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.

Additional file 14: Table S7. Correspondence between Illumina indexes and accession pools used in this study.

Abbreviations

MATE1: Multidrug and toxic compound extrusion; AACT1: Aluminum activated citrate transporter; TLP: Taumaatin-like protein; FBA: Fructose-biphosphate aldolase; PBF: Prolamine-box binding factor; Sintb: Secaloindoline-b; GSP-1: Grain softness protein; VAF: Variant allele frequency

Acknowledgements

Not applicable.

Authors’ contributions

Conceptualization, BJT and HB-B; Methodology, AH, JJ-C, BJT and HB-B; Investigation, AH, LB, KT, EB, JJ-C, PG, AK, BJT, and HB-B; Writing – Original Draft, AH, BJT and HB-B; Writing – Review & Editing, AH, JJ-C, BJT, and HB-B; Funding Acquisition, BJT and HB-B; Resources, BJT and HB-B; Supervision, BJT and HB-B. The author(s) read and approved the final manuscript.

Funding

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 funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

BAM files generated and analyzed in this study are available in NCBI BioProject PRJNA593253. The remaining data used and/or analyzed during this study is included in the supplementary information files or is available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

1Department of Plant Genetics, Breeding and Biotechnology, Institute of Biology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland.
2Department of Silviculture, Institute of Forest Sciences, Warsaw University of Life Sciences – SGGW, Warsaw, Poland. 3Plant 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. 4Veterinary Genetics Laboratory, University of California, Davis, Davis, California, USA.

Received: 20 August 2020 Accepted: 18 November 2020

References

1. Punyagunta MD, Fuller DQ. The nature of selection during plant domestication. Nature. 2009;457:843–8.
2. Mondal S, Rutkoski JE, Velu G, Singh PK, Crespo-Herrera LA, Guzman CG, et al. 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. 2016;7:991.
3. Joukhadar R, Daetwyler HD, Bansal UK, Gendall AR, Hayden MJ. Genetic diversity, population structure and ancestral origin of Australian wheat. Front Plant Sci. 2017;8:1–15.
4. Hoisington D, KharraZe M, Reeves T, Ribaut J-M, Skovmand B, Tabata S, et al. Plant genetic resources: what can they contribute to increased crop productivity? Proc Natl Acad Sci. 2002;96:5937–43.
5. McCouch S. Feeding the future. Nature. 2013;493:3–4.
6. Giri A, Zamar D. Unused natural variation can lift yield barriers in plant breeding. PLoS Biol. 2004;2:e245.
7. GanuvaZe R, Chin JH, Paraiso-Tanaka J, PatersaZe P, Catausan S, Dalei C, et al. The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. Nature. 2012;488:535–9.
8. McCouch S. Diversifying selection in plant breeding. PLoS Biol. 2004;2:e347.
9. FAO. The Second Report on the State of the World’s Plant Genetic Resources for Food and Agriculture. Rome: FAO 2010.
10. KelwaZe J, Kilian B, O’kazan H, Babben S, Perovic D, Mayer KFX, et al. Separating the wheat from the chaff - a strategy to utilize plant genetic resources from ex situ genebanks. Sci Rep. 2014;4:14–8.
11. Li Y, Zhao S, Ma J, Li D, Yan L, Li J, et al. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics. 2013;14:579.
12. Mehra P, Pandey BK, Giri J. Genome-wide DNA polymorphisms in low phosphate tolerant and sensitive rice genotypes. Sci Rep. 2015;5:13090.
13. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accesses identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33:408–14.
14. Hodges E, Kuan Z, Balija V, Kamer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007;39:1522–7.
15. Hussain M, Iqbal MA, Till BJ, Rahman M. Identification of induced mutations in hexaploid wheat genome using exome capture assay. PLoS One. 2018;13:e0201918.
16. Yao Z, You FM, N’Diaye A, Knox RE, McCartney C, Hiebert CW, et al. Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics. 2020;21:1–16.
17. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016;17:1–14.
18. Le Nguyen K, Gondzin A, Courtois B, Gantet P. Next-generation sequencing accelerates crop gene discovery. Trends Plant Sci. 2019;24:263–74.
19. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.
20. Campbell NR, Harmon SA, Narum SR. Genotyping-in-thousands by sequencing (gt-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. Mol Ecol Resour. 2015;15:855–67.
21. Dou Y, Gold HD, Luquette LJ, Park PJ. Detecting somatic mutations in normal cells. Trends Genet. 2018;34:545–57.
22. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimaSeq and iVar. Genome Biol. 2019;20:1–19.
an Iranian primitive accession as a new tool for rye breeding and genomics. Theor Appl Genet. 2008;117:641–52.

69. Massa AN, Morris CF, Gill BS. Sequence diversity of Puroindoline-a, Puroindoline-b, and the grain softness protein genes in Aegilops tauschii Coss. Crop Sci. 2004;44:1808–16.

70. Ye J, Coulotouris G, Zaretkaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.

71. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754–60.

72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Hormer N, et al. The sequence alignment / map format and SAMtools. Bioinformatics. 2009;25:2078–9.

73. Garcia-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Gótz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28:2678–9.

74. Knaus BJ, Grünwald NJ. Vcfr: a package to manipulate and visualize variant call format data in R. Mol Ecol Resour. 2017;17:44–53.

75. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014;30:2076–8.

76. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6:80–92.

77. Vasar R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11:1–9.

78. Gleiss A, Dakna M, Mischak H, Heinez G. Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters. Bioinformatics. 2015;31:2310–7.

79. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer New York; 2016.

80. Tateno Y, Nei M, Tajima F. Accuracy of estimated phylogenetic trees from molecular data – I. Distantly Related Species. J Mol Evol. 1982;18:387–404.

81. Takezaki N, Nei M, Tamura K. POPTREEW: web version of POPTREE for constructing population trees from allele frequency data and computing some other quantities. Mol Biol Evol. 2014;31:1622–4.

82. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9.

83. Rohlf FJ. NTsys-pc: Numerical Taxonomy and Multivariate Analysis System, Version 2.2. Exeter Software, Setauket, NY (2005).

84. ROD PEAKALL, PETER E. SMOUSE, genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 2006;6(1):288–95.

85. R. Peakall, P. E. Smouse, GenALEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. Bioinformatics. 2012; 28(19):2537–9.

86. Taylor NE, Greene EA. PARSESNP : a tool for the analysis of nucleotide polymorphisms. Nucleic Acids Res. 2003;31:3808–11.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions