the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) **113**, 866–880 doi: 10.1111/tpj.16079

RESOURCE

Direct access to millions of mutations by whole genome sequencing of an oilseed rape mutant population

Srijan Jhingan[1], **Avneesh Kumar**[1,†], **Hans-Joachim Harloff**[1], **Felix Dreyer**[2], **Amine Abbadi**[2], **Katrin Beckmann**[2], **Christian Obermeier**[3] and **Christian Jung**[1,*]

[1] *Plant Breeding Institute, Christian-Albrechts-Kiel University, Olshausenstrasse 40, 24098, Kiel, Germany,*

[2] *NPZ Innovation GmbH, Hohenlieth-Hof, 24363, Holtsee, Germany, and*

[3] *Department of Plant Breeding, Justus Liebig University Giessen, Heinrich-Buff-Ring 26-32, 35392, Giessen, Germany*

Received 19 October 2022; revised 20 December 2022; accepted 21 December 2022; published online 27 December 2022.

*For correspondence (e-mail c.jung@plantbreeding.uni-kiel.de).

†Present address: Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO, 80521, USA

SUMMARY

Induced mutations are an essential source of genetic variation in plant breeding. Ethyl methanesulfonate (EMS) mutagenesis has been frequently applied, and mutants have been detected by phenotypic or genotypic screening of large populations. In the present study, a rapeseed M_2 population was derived from M_1 parent cultivar 'Express' treated with EMS. Whole genomes were sequenced from fourfold (4×) pools of 1988 M_2 plants representing 497 M_2 families. Detected mutations were not evenly distributed and displayed distinct patterns across the 19 chromosomes with lower mutation rates towards the ends. Mutation frequencies ranged from 32/Mb to 48/Mb. On average, 284 442 single nucleotide polymorphisms (SNPs) per M_2 DNA pool were found resulting from EMS mutagenesis. 55% of the SNPs were C → T and G → A transitions, characteristic for EMS induced ('canonical') mutations, whereas the remaining SNPs were 'non-canonical' transitions (15%) or transversions (30%). Additionally, we detected 88 725 high confidence insertions and deletions per pool. On average, each M_2 plant carried 39 120 canonical mutations, corresponding to a frequency of one mutation per 23.6 kb. Approximately 82% of such mutations were located either 5 kb upstream or downstream (56%) of gene coding regions or within intergenic regions (26%). The remaining 18% were located within regions coding for genes. All mutations detected by whole genome sequencing could be verified by comparison with known mutations. Furthermore, all sequences are accessible via the online tool 'EMSBrassica' (http://www.emsbrassica.plantbreeding.uni-kiel.de), which enables direct identification of mutations in any target sequence. The sequence resource described here will further add value for functional gene studies in rapeseed breeding.

Keywords: ***Brassica napus*, ethyl methanesulfonate, EMS mutagenesis, TILLING, rapeseed.**

INTRODUCTION

Oilseed rape (*Brassica napus* L.) is the primary oil crop in the world's temperate regions and the third-largest seed oil and second-largest protein meal source globally (Wang et al., 2018). In 2019, approximately 34 million hectares yielded 70 million tons of oilseed rape globally (http://www.fao.org/faostat). Although the seeds are a major source of edible oil (45–50%), byproducts after oil extraction are conventionally utilized as animal feed and as a substrate for biodiesel production in Europe. Increasing and stabilizing the yield potential in combination with improved seed quality (e.g. reduced glucosinolate content) are the primary aims of rapeseed breeding.

Oilseed rape belongs to the family of crucifers (Brassicaceae). It is an allotetraploid (AACC, 2*n* = 38) resulting from a spontaneous interspecific hybridization between the diploid AA (2*n* = 20) and CC (2*n* = 18) genomes of turnip rape (*Brassica rapa* L., syn. *campestris*) and cabbage (*Brassica oleracea* L.), respectively. Its origin was traced back to the Mediterranean region approximately 7500 years ago and the total genome size is estimated to be approximately 1.13 Gb (Chalhoub et al., 2014). Today, several high-quality rapeseed reference genomes are available. The 'Darmor-*bzh*' was the first reference genome assembled using a European winter-type oilseed rape. The genome encompassed 314.2 Mb of the A sub-genome and 525.8 Mb of the C sub-genome, predicted with 101 040 gene models (Chalhoub et al., 2014). More recently, an improved long-read assembly of the 'Darmor-*bzh*' reference genome (Rousseau-Gueutin et al., 2020) was published. Moreover, high-quality whole genome assemblies of eight oilseed accessions across three ecotypes (Song et al., 2020), a winter-type (Lee et al., 2020) and a semi-winter (Chen et al., 2021) oilseed rape are now publically available.

As a result of its short history of evolution and domestication, the genetic diversity within *B. napus* is low (Rahman, 2013). Ethyl methanesulfonate (EMS) induced mutagenesis has been used to create new allelic variation (Gilchrist et al., 2013; Harloff et al., 2012; Lee et al., 2018; Tang et al., 2020; Wang et al., 2008; Wells et al., 2014). In recent years, clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) technology to create targeted mutations has been successfully applied and numerous mutants have been published (Braatz et al., 2017; Karunarathna et al., 2020; Sashidhar et al., 2020; Zheng et al., 2020). Although targeted mutagenesis offers several advantages, random mutagenesis still has its importance in rapeseed breeding, mainly because CRISPR/Cas mutants are legally classified as genetically modified organisms in the European Union and therefore their usage in practical breeding is limited (Jung & Till, 2021).

EMS mutant discovery in oilseed rape has conventionally relied on the activity of DNA mismatch specific endonucleases and polyacrylamide gel-based detection assays classically termed TILLING (Targeting Induced Local Lesions in Genomes) (Braatz et al., 2018; Emrani et al., 2015; Karunarathna et al., 2020; Shah et al., 2018) using pooled genomic DNA from M_2 individuals of large mutant populations. Because this procedure is time-consuming and laborious, amplicon sequencing-based detection methods (Gilchrist et al., 2013; Sashidhar et al., 2019; Wells et al., 2014) have gained increasing popularity because of their efficiency and sensitivity. However, similar to conventional TILLING, detection of EMS mutations via amplicon sequencing approaches is restricted to single gene families with high sequence conservation.

Sequencing whole mutant populations, termed TILLING by sequencing (TbySeq), is the gold standard of mutant detection (Jung & Till, 2021). In a pioneering study, Krasileva et al. (2017) demonstrated a TILLING by exome sequencing approach to detect EMS-induced mutations in tetraploid and hexaploid wheat. The TbySeq approach has also been reported from several crops such as rice (Abe et al., 2012), maize (Nie et al., 2021), tomato (Garcia et al., 2016), soybean (Lakhssassi et al., 2021), sunflower (Fanelli et al., 2021) and cotton (Fang et al., 2020). In a first TILLING by whole genome sequencing (TbyWGS) approach for oilseed rape, a limited number of EMS mutants were whole genome sequenced to detect EMS-induced mutations (Tang et al., 2020).

The present study aimed to develop a bioinformatic resource for detecting EMS-induced mutations on a genome-wide scale (Harloff et al., 2012). We sequenced the whole genomes of 1988 M_2 plants from an EMS mutagenized winter oilseed rape population. A TbyWGS pipeline was established that allows the identification of mutations within any genomic region of interest. The sequences can be screened via the online resource 'EMSBrassica' (http://www.emsbrassica.plantbreeding.uni-kiel.de). Thus, our TbyWGS platform constitutes a long-lasting sequence repository of mutants.

RESULTS

Whole genome sequencing reveals high mutation density

To gather initial results on sequencing quality and mutant detection, we first performed pilot experiments with fifty 4× pools sequenced at 10× coverage by mapping the raw reads to the Express617 v1 reference genome. The effective genome size of this reference assembly was 925 095 059 bp (Lee et al., 2020). After single nucleotide polymorphism (SNP) detection and filtration, many SNPs were filtered into the 'high confidence' category (data not shown) but later could not be validated via Sanger sequencing. Therefore, we decided to double the intended coverage to 20× for sequencing the 4× pools and sequenced pooled DNA from 1988 M_2 plants on the NovaSeq 6000 platform (Illumina, San Diego, CA, USA) (Figure 1). The raw dataset encompassing a total of 497 4× pools represented an average of 35.8 Gb per pool. On average, approximately 116 million raw reads per pool were generated. In terms of read quality, Phred scores for all paired-end reads varied well above the optimum scores (Figure S1). Mapping reads to the Express617 reference genome resulted in a mapping rate between 96.1 and 99.1% and an average coverage depth of 34.2× with a coverage breadth of 97.7–99.2%.

We then calculated the number and frequency of SNPs. Variant calling revealed 843 000 unfiltered SNPs per pool on average. Using our SNP filtering criteria [read depth (DP) ≥ 10, allele depth (AD) = 12.5–60% and mapping quality (MQ) ≥ 30], on average, 156 479 SNPs per pool were $C \rightarrow T$ and $G \rightarrow A$ transitions, characteristic for EMS induced mutations (Table 1). This corresponds to an average of 39 120 EMS type mutations per single M_2 plant. Following the terminology used by Fanelli et al. (2021), we termed the EMS type $C \rightarrow T$ and $G \rightarrow A$ mutations as 'canonical' transitions and all others as 'non-canonical'

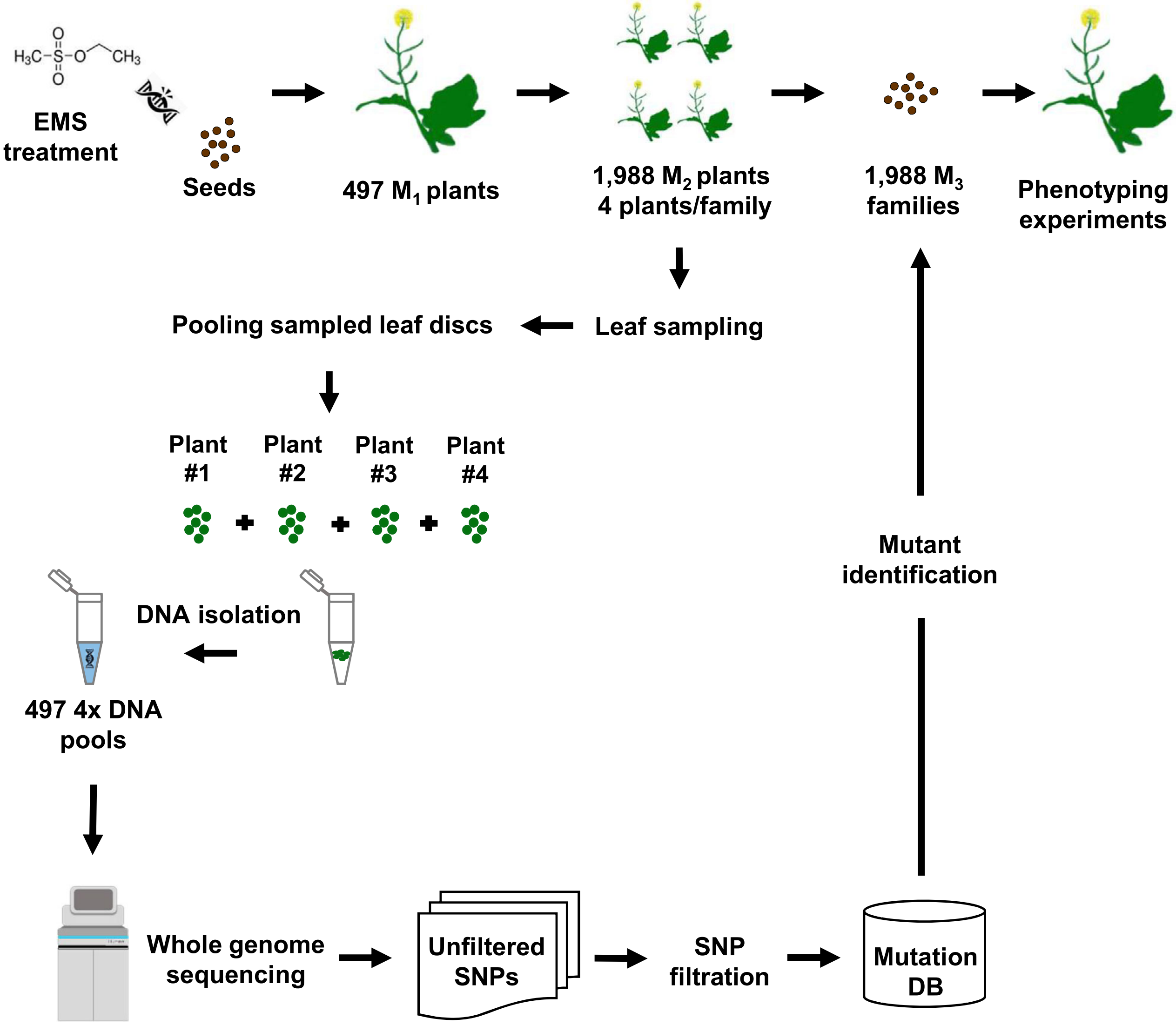

Figure 1. Workflow of the mutant detection pipeline.
M_2 seeds were harvested from 497 M_1 plants and four plants/M_2 family were grown. Leaf material was sampled from 1988 M_2 plants as leaf discs. Leaves from each family were pooled and DNA was isolated and sequenced on the NovaSeq 6000 platform. Raw reads were mapped to the Express617 reference genome. SNPs were called and filtered using GATK 4.1.4.1 and custom Unix scripts. Raw SNPs were filtered based on canonical EMS mutations (C → T and G → A transitions) possessing read depth (DP) $\geq$ 10 allele depth (AD) = 12.5–60% and quality controlled for mapping quality (MQ) $\geq$ 30. Mutation effects were predicted using ENSEMBL VARIANT EFFECT PREDICTOR (VEP) tool. High confidence mutations from each pool were merged into a single accessible database. After identification of desirable M_2 mutants, corresponding M_3 plants can be analyzed to verify for phenotypic effects. EMS, ethyl methanesulfonate; SNP, single nucleotide polymorphism.

transitions or transversions. We termed them collectively as single nucleotide variants (SNVs). We detected an average of one C → T and G → A mutation per 23.6 kb of the genomic sequence of the Express617 reference genome (925 Mb). Noteworthy, the 45% share of non-canonical transitions (15%) or transversions (30%) was almost equal to that of the canonical C → T and G → A mutations among the high confidence SNPs (Figure 2). Additionally, on average, we detected 88 725 high confidence insertions and deletions (InDels) per M_2 DNA pool. However, we restricted subsequent sequence analyses to SNVs only.

Detecting functional mutations

We considered how many high confidence EMS mutations could have a putative effect on gene function. First, we analyzed mutations within predicted gene models from the Express617 reference. We reasoned that an SNV might alter the function of the encoded protein if it (i) introduces a premature stop codon, (ii) results in a splice variant, (iii) confers a non-synonymous amino acid substitution or (iv) causes a change in the translation start site. On average, 12 129 mutations per M_2 DNA pool fulfilled these criteria

Table 1 Summary statistics of SNPs detected after whole genome sequencing (including regions without chromosomal annotations) of 1988 M_2 rapeseed plants assembled in 497 4× DNA pools

	Canonical transitions		Non-canonical transitions		Transversions							
SNP	C → T	G → A	T → C	A → G	C → A	G → T	A → C	T → G	T → A	A → T	G → C	C → G
Number of SNPs[a]	78 169	78 310	21 073	21 174	13 770	13 776	9942	9930	12 839	12 754	6290	6414
Sum	156 479		42 247		85 715							
%	27.0	28.0	7.0	7.0	5.0	5.0	3.0	3.0	5.0	4.0	2.0	2.0

AD, allelic depth; DP, read depth; MQ, mapping quality; SNP, single nucleotide polymorphism.

SNPs were filtered based on DP ≥ 10, AD = 12.5–60% and MQ ≥ 30. Characteristic C → T and G → A transitions were named as 'canonical EMS-type' transitions. All other substitutions were named as 'non-canonical' transitions or transversions.

[a]Average of 497 4× pools analyzed.

(Table 2). Some 75% of all characterized mutations were located within annotated regions of the Express617 reference genome (chromosomes A01–A10 and C01–C09), whereas 25% were located on non-annotated scaffolds (occupying approximately 17% of the reference genome) that could not be confidently anchored to any chromosome (Table S1). Out of all high confidence canonical mutations located within annotated chromosomes, 0.4, 7.0, 4.1 and 4.9% were predicted as nonsense, missense, synonymous and intronic, respectively (Table 2). Out of all canonical mutations located within coding regions, 2.3% were nonsense mutations, 40.4% were missense mutations, 24.1% were synonymous mutations and 28.7% were located within introns (Figure S2). Start site loss and splice site variants were observed as the rarest type of mutations contributing < 1% of the total EMS-type mutations (Figure 3). On average, we observed 20 757 genes to be mutated per fourfold pool.

We then analyzed the functional effects of mutations other than C → T and G → A transitions (Figure S3). Approximately 85% of such mutations were located either 5 kb upstream or downstream (approximately 27% each) of gene coding regions or within intergenic regions (31%). The remaining 15% were located within regions coding for genes (Figure S4). As predicted, 3.6% of these mutations could have drastic effects in the form of nonsense (0.05%), missense (3.5%) and splice site (0.03%), or even as START site (0.01%) mutations. By contrast, 10.4% do not confer functional effects because they were characterized as silent mutations such as synonymous (4%) or intronic (6.3%) and rare mutations where the stop codon is retained (0.003%) (Table 2).

Sequence analyses reveal patterns of mutation frequency and distribution

Our mutation detection approach operates on the whole genome scale. Therefore, we investigated the distribution of EMS mutations across all chromosomes, including intergenic regions and 5 kb upstream and downstream regions. To check for a possible bias in mutation frequencies for the A and C sub-genomes of oilseed rape, we calculated the number and density of high confidence canonical mutations across all annotated chromosomes (Table 3). As expected, the number of C sub-genome mutations (76 818) was significantly higher than A sub-genome mutations (41 412) because, in the assembled genome, the C genome exceeds the A genome by 173.25 Mb (Figure 4). Surprisingly, mutation frequencies on average were higher for the C sub-genome (41 mutations/Mb) than the A sub-genome (35 mutations/Mb) (Table S1). In the next step, we searched both sub-genomes for mutation hotspots by visualizing EMS mutations in 1 Mb non-overlapping windows (Figure 5). Although canonical mutations were evenly distributed across all chromosomes in general, we found regions with significantly increased or decreased mutation densities (Figure 5). As a general tendency, mutation frequencies were lower towards the end of chromosomes except for A06, C03 and C06, where mutation frequencies were higher in the telomeric regions.

Furthermore, we compared the distribution of canonical and non-canonical mutations for all chromosomes. For this, we used all C → T and G → A transitions and then all other mutation types that qualified as high confidence SNPs (DP ≥ 10, AD = 12.5–60% and MQ ≥ 30) as distinct sets. Interestingly, a similar distribution of high confidence canonical and non-canonical mutations per 1 Mb non-overlapping windows was observed across most chromosomes (Figure S5). This suggests that the distribution of the non-canonical mutations is not random across chromosomes but follows a defined distribution pattern like that observed for the C → T and G → A transitions.

Validation of functional mutations

The functional mutations detected by the TbyWGS approach were verified in two ways. We chose four gene families affecting agronomically important traits in oilseed rape. First, we verified mutations in the four gene families previously detected by a conventional gel-based TILLING

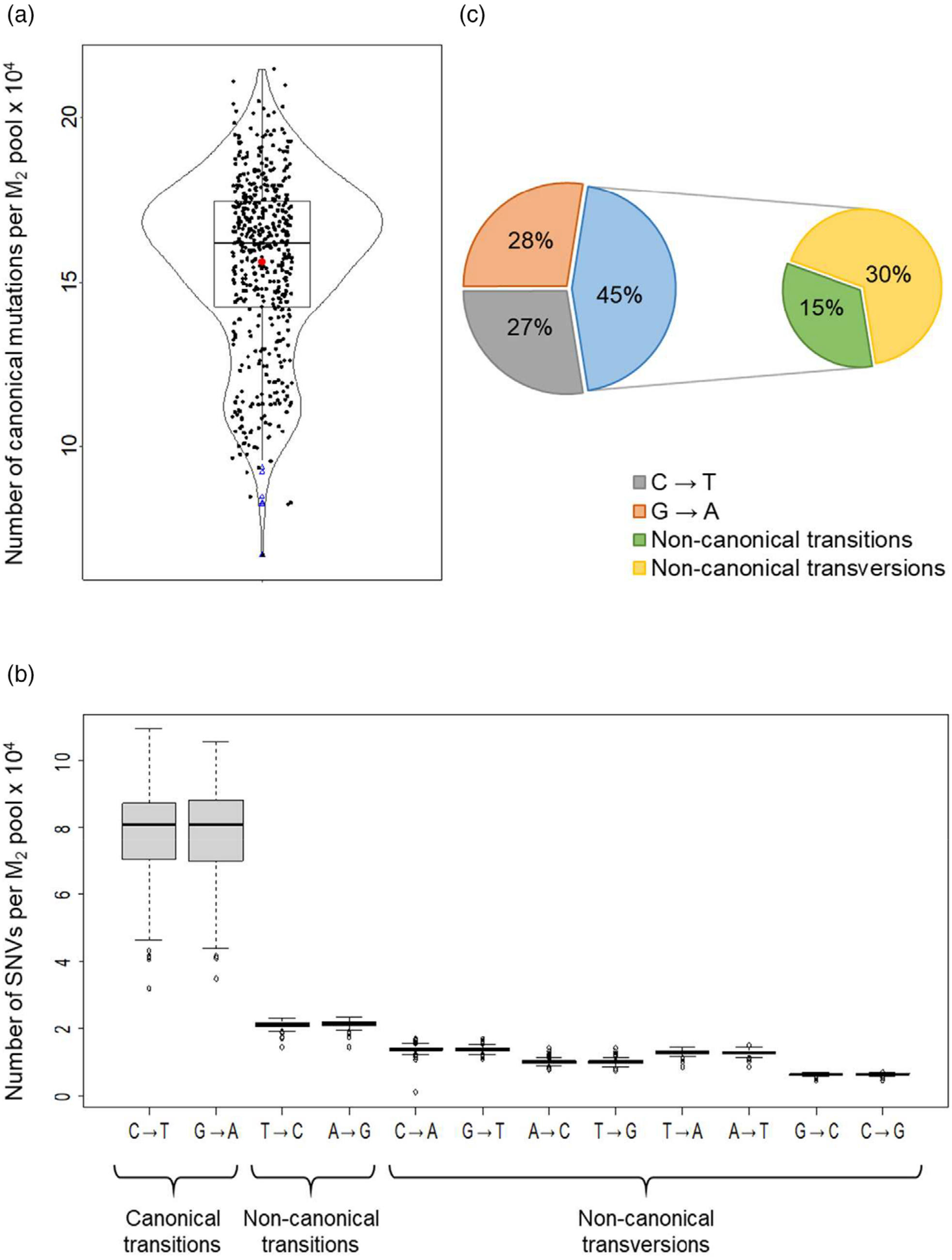

Figure 2. Analysis of SNP counts from 1988 sequenced M_2 plants constituting 497 4× pools.
(a) Violin plot showing the number of canonical C → T and G → A transitions. Black dots and blue triangles depict the number of filtered SNPs from each of the sequenced pools and outliers (pools with a very low number of called SNPs), respectively. The red dot represents the mean of SNP counts from all pools.
(b) Boxplots showing the number of individual SNP types observed and (c) percentage share of single nucleotide variants (SNVs) by type. C → T and G → A transitions accounting for 55% of total filtered SNVs were named as 'canonical' mutations. All other nucleotide substitutions (45%) have been named as 'non-canonical' transitions or transversions. All SNVs were filtered based on SNP type, minimum read depth (DP) ≥ 10, allele depth (AD) = 12.5–60% and mapping quality (MQ) ≥ 30 parameters. All boxplots show the upper and lower quartiles separated by the median (horizontal line). Whiskers represent maximum and minimum values. SNP, single nucleotide polymorphism.

of the Express617 TILLING population (Emrani et al., 2015; Harloff et al., 2012; Karunarathna et al., 2020). *SFAR* (*SEED FATTY ACID REDUCER*) genes encode GDSL lipases, which affect a wide range of primary and secondary functions in plants. Plants possessing EMS-induced functional mutations in *BnSFAR* genes had displayed an elevated seed oil content (Karunarathna et al., 2020). *REF1* and *SGT* genes encode for enzymes UDP-glucose:sinapic acid glucosyltransferase and sinapaldehyde dehydrogenase/coniferaldehyde dehydrogenase, respectively. Both genes play a

Table 2 Summary statistics of predicted mutation effects within the predicted gene models of the Express617 genome

	Number of mutations predicted[a]					
	Canonical transitions	Non-canonical transitions	Transversions			
Mutation type	C → T/G → A	T → C/A → G	C → A/G → T	A → C/T → G	T → A/A → T	G → C/C → G
Nonsense	637	0	37	3	11	1
Splice site[b]	247	8	7	3	10	4
Missense	11 228	956	622	685	534	546
START lost	17	4	0	3	2	0
STOP retained	16	3	0	0	0	0
Synonymous	6701	1857	551	521	564	357
Splice region[b]	959	195	104	98	137	45
Intronic	7969	1958	1134	997	1297	595
Downstream[c]	45 108	9074	4915	3907	4872	2215
Upstream[c]	45 310	9193	4943	4002	4937	2291
Intergenic	42 054	10 365	6275	4711	6140	2588
Total	160 246	33 613	18 588	14 930	18504	8642

AD, allelic depth; DP, read depth; MQ, mapping quality.

[a] Predicted mutation effects may overlap across mutation types depending on the structure of predicted gene models.

[b] Splice site variants include both acceptor and donor site mutations. Splice region variants include change(s) within the region of the splice site; within 1–3 bases of the exon or 3–8 bases of the intron.

[c] Variants located 5 kb upstream or downstream of the translation START and STOP sites, respectively.

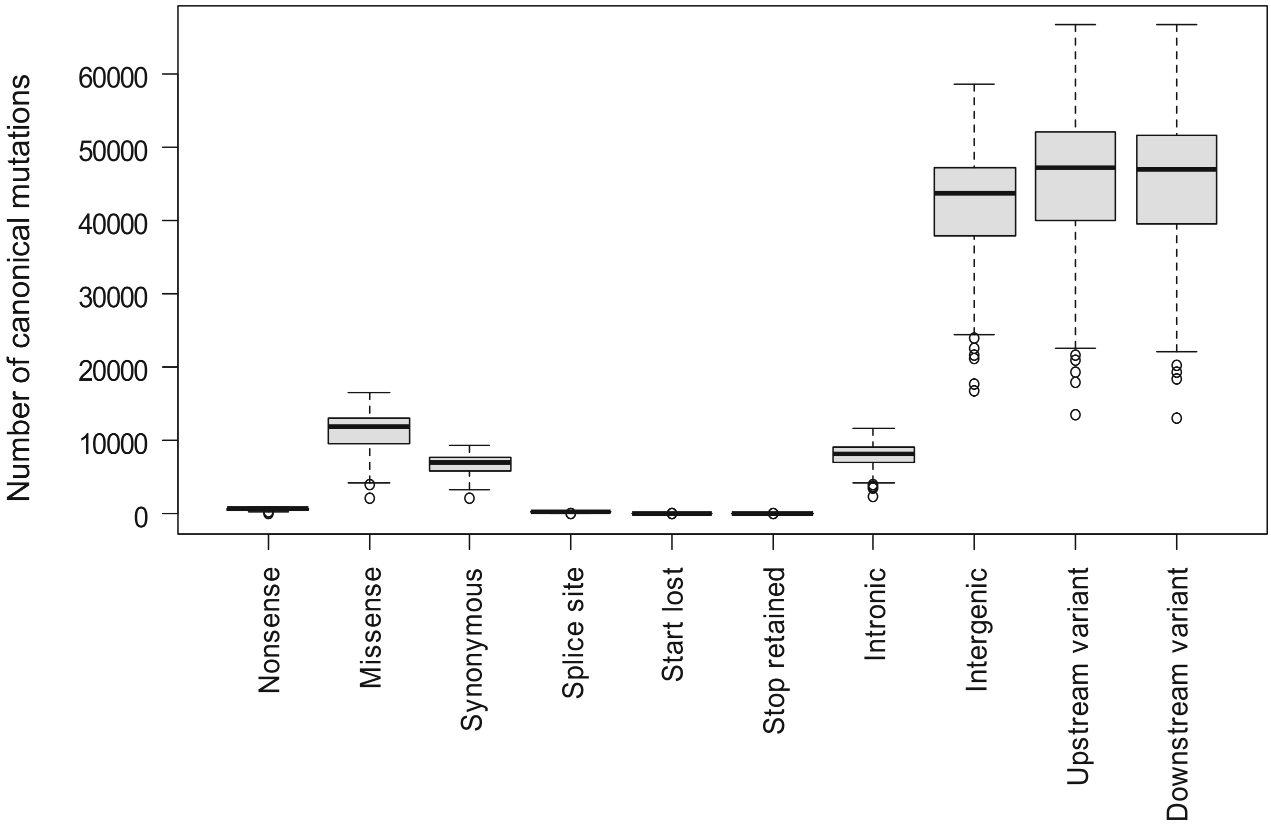

Figure 3. Characterization of C → T and G → A mutations with predicted effects on a genome-wide scale.

Boxplots show distribution of SNP effects from the sequenced 4× pools. The upper and lower quartiles are separated by the median (horizontal line). Whiskers represent maximum and minimum values. Circles depict outliers. Mutation effects are predicted using SNPs filtered with $DP > 10$, $AD = 12.5$–60% and $MQ \geq 30$ parameters. The ENSEMBL VARIANT EFFECT PREDICTOR (release 99) was used in offline mode. Mutation effects were characterized within all predicted gene models using the general feature format (GFF) file of the Express617 reference genome. Splice site variants include acceptor and donor site mutations. Stop retained mutations refer to unchanged STOP codons in spite of induced mutation(s). Upstream and downstream variants are located within a distance of 5 kb from the transcription START and STOP sites, respectively. AD, allele depth; DP, read depth; MQ, mapping quality; single nucleotide polymorphism.

Table 3 Summary statistics of EMS type transitions in the A and C sub-genomes of the Express617 mutant population

	Chromosome	Chromosome length (bp)[a]	Number of mutations/ chromosome[b]	Number of mutations/Mb per 4× pool[b]	Number of mutations/Mb per plant[b]
Sub-genome A	A01	29 969 127	4521	151	37.7
	A02	30 174 133	3975	132	32.9
	A03	38 328 780	5055	132	33.0
	A04	22 208 790	2852	128	32.1
	A05	29 330 807	4372	149	37.3
	A06	31 862 090	4787	150	37.6
	A07	27 714 229	3988	144	36.0
	A08	22 295 061	2963	133	33.2
	A09	43 308 710	6183	143	35.7
	A10	20 498 486	2716	133	33.1
Sub-genome C	C01	44 118 044	8606	195	48.8
	C02	61 556 739	9961	162	40.5
	C03	62 379 756	8626	138	34.6
	C04	56 192 105	9056	161	40.3
	C05	46 536 336	7457	160	40.1
	C06	46 870 576	8935	191	47.7
	C07	39 009 375	6106	157	39.1
	C08	52 066 946	8509	163	40.9
	C09	60 209 689	9562	159	39.7

Mutation counts and frequencies were calculated for individual chromosomes of the two sub-genomes (chrA01–A10 and chrC01–C09) for all sequenced 4× pools.

[a] Based on the assembled Express617 reference genome (Lee et al., 2020).

[b] Calculated as an average of 497 4× pools analyzed. For estimations per plant, frequencies of each pool were divided by 4, since each sequenced pool is a compilation of four single M_2 plants.

crucial role in the biosynthesis of sinapine, an important anti-nutritive compound in oilseed rape. EMS-induced loss of function mutations had been detected in both genes. Double and triple mutants showed a significant reduction of the seed sinapine content (Emrani et al., 2015; Harloff et al., 2012). To verify the mutations by our TbyWGS approach, we identified seven M_2 families with the respective M_2 mutants. We then screened whole genome sequences of the 4× pools harboring these M_2 families (Table S2). As a result, all seven mutations could be found, proving the reliability of our TbyWGS pipeline (Figure S6).

As a next step, we screened the WGS dataset for new mutations. We chose five genes involved in the glucosinolate biosynthesis and transportation pathway. *MYB28* and *CYP79F1* genes encode for an R2-R3-MYB transcription factor (Gigolashvili et al., 2008) and a cytochrome P450 enzyme (Reintanz et al., 2001), respectively. Both genes are involved in the core structure formation of aliphatic glucosinolates in oilseed rape. The *GTR* (*GLUCOSINOLATE TRANSPORTER*) genes have been characterized by seed-specific glucosinolate transportation activity in *Brassica* species (Nour-Eldin et al., 2017). We detected 12 M_2 DNA pools with nonsense mutations for all selected gene families (Figure S7). Then, we isolated genomic DNA from each of the 48 M_2 plants. DNA fragments were amplified by PCR using paralog-specific primers flanking the putative mutation sites (Table S3). Sanger sequencing revealed a success rate of 100% because all PCR fragments contained the expected mutations (Figure S8).

A web-based interface for screening the mutant population

Our TbyWGS resource 'EMSBrassica' provides access to 78 083 182 high confidence canonical EMS mutations. We developed a web-based resource enabling the user-friendly and convenient detection of high confidence C → T and G → A mutations with predicted effects for ease of screening. As a first step, the Express617 reference genome assembly and the supplemented GFF file assembled by Lee et al. (2020) can be downloaded from https://doi.org/10.5281/zenodo.3524259. Using BLAST queries, genomic regions of interest within this reference assembly concerning the chromosome and their exact locations can be identified. Then, the WGS database can be searched based on the 'Chromosome', 'Start' and 'End' information. By screening the database, mutations within genomic regions of interest can be identified from the respective 4× pools. The database 'EMSBrassica' is accessible at: http://www.emsbrassica.plantbreeding.uni-kiel.de.

DISCUSSION

The present study aimed to develop a database of oilseed rape mutant sequences for identifying putative functional mutations on a genome-wide scale. Based on the available Express617 mutant population, the reference genome

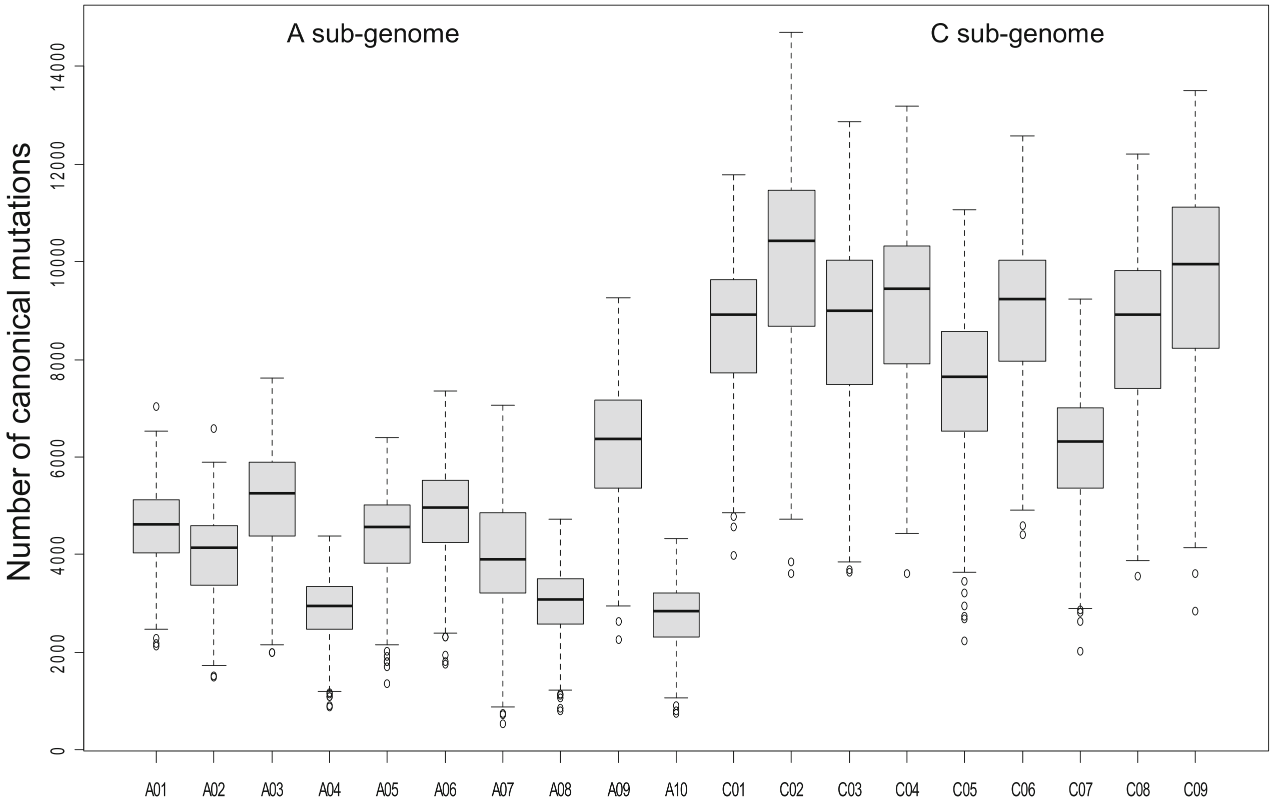

Figure 4. Number of EMS-induced transitions across the A- and C sub-genomes of 497 sequenced M_2 pools.
Boxplots represent the average number of SNPs originating from the regions of the Express617 genome annotated with chromosomes (chrA01–A10 and chrC01–C09). Black circles represent individual data points for SNP counts per sequenced 4× pool. The upper and lower quartiles are separated by the median (horizontal line). Whiskers represent maximum and minimum values. Circles depict outliers. EMS, ethyl methanesulfonate; SNP, single nucleotide polymorphism.

Express617 (Lee et al., 2020) and whole genome sequencing of M_2-Pools of mutants, we have established a TbyWGS protocol to detect mutations in any sequence of the respective rapeseed mutant genome. M_2 plants with the desired mutation can be directly identified to perform functional studies in the M_3 offspring. This procedure avoids the laborious gel-based mutant screening and the long-term storage of high-quality M_2 DNA needed for conventional TILLING.

In our TbyWGS approach, we used a 4× pooling strategy instead of the original 8× pooling previously used for gel-based screening of the Express617 mutant population (Harloff et al., 2012). We found that low sequence coverage results in the false identification of sequencing errors or artifacts as genuine mutations. Therefore, a 20× sequence coverage was chosen, which theoretically results in five raw reads from each M_2 plant. The degree of heterozygosity in the Express617 inbred line (F_{11}) used for EMS mutagenesis is expected to be below 0.0005%. Therefore, we expect that > 99.9995% of the SNPs detected within our M_2 families are caused by EMS mutagenesis and not by residual heterozygosity. We adjusted the SNP filtering criteria to remove low-quality SNPs with poor MQ and sequencing errors. According to classical Mendelian genetics, the frequency of mutant alleles in a segregating M_2 generation is expected to be 50% if we assume the M_1 to be hemizygous for the mutation. However, this value might be lower as a result of a mosaic of mutated and non-mutated cells, which is expected for almost all M_1 plants. Also, poor fitness of mutant gametophytes and low vitality of homozygous M_2 plants can account for decreased mutant frequencies in the M_2 offspring. In addition, we can expect a statistical bias caused by the low number of M_2 plants in each pool because only four M_2 plants per pool were selected. In the worst case, only one plant out of four might carry a mutant allele. In another case, the mutation might be more frequent because of an increased representation of mutated plants carrying one or more mutant alleles within an M_2 pool. Respecting the Mendelian segregation and biases as a result of mosaic M_1 and fitness of M_2 plants in all sequenced pools, it was reasonable in our screening to select a lower limit for the allelic depth of 12.5% to cover rare mutations in each pool. On the other hand, the upper limit from the expected Mendelian

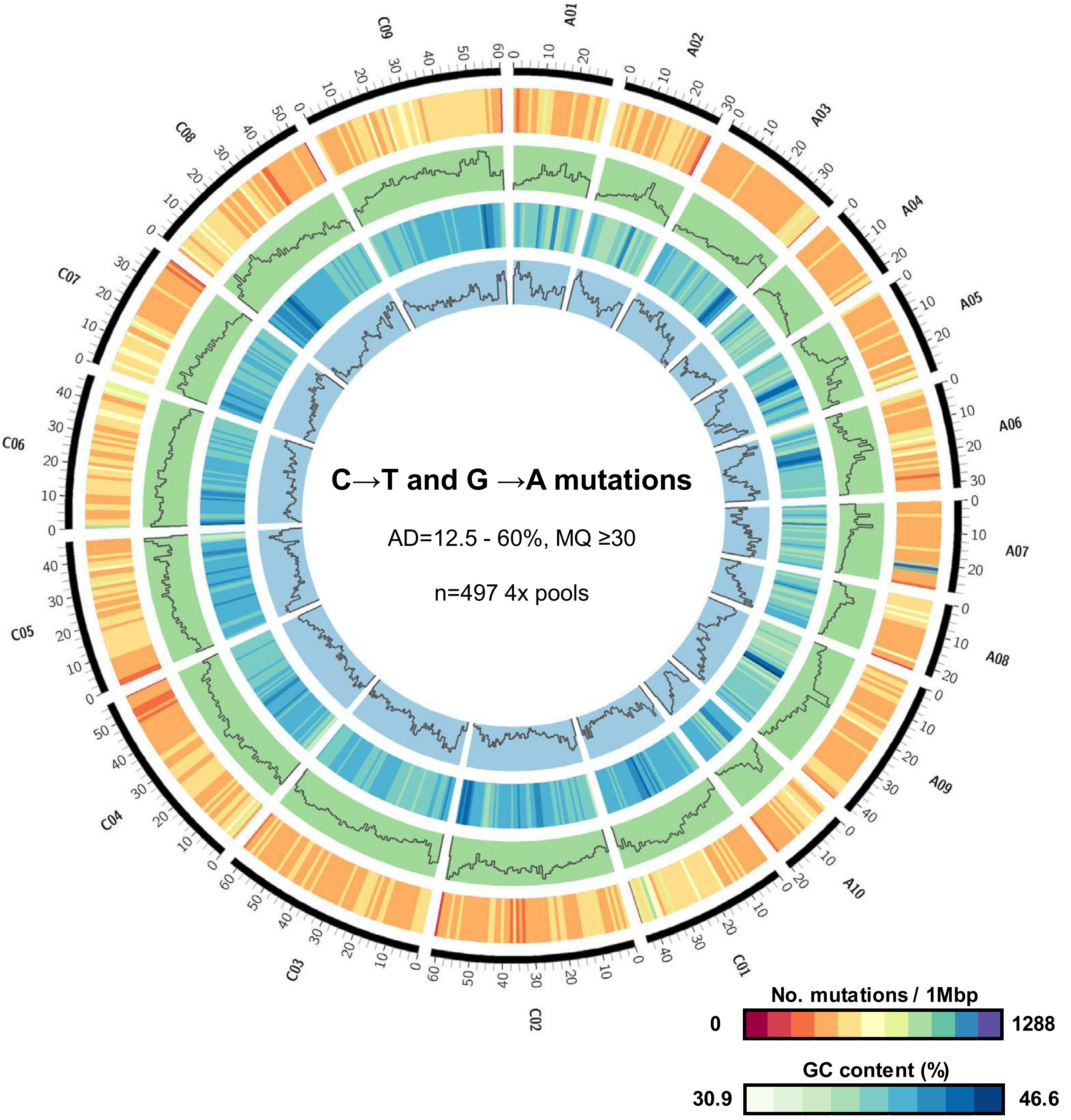

Figure 5. Visualization of C → T and G → A mutations at the whole genome-scale.

The CIRCOS plot shows five tracks (outside to inside) representing chromosomes A01–A10 and C01–C09 (black), number of mutations (dark red to violet), genomic regions annotated with repetitive sequences (light green), GC content (%) per 1 Mb windows (white to dark blue) and regions coding for genes (light blue). Chromosomal lengths are in Mb. Mutation densities were calculated per 1 Mb windows across all chromosomes from all sequenced 4× pools. Regions representing repetitive sequences and gene coding sequences were calculated as the number of base pairs per 1 Mb annotated across the whole genome and within predicted gene models, respectively. Mutation densities and GC content were plotted using ‘scale_log_base = 0.5’ values. 1 Mb non-overlapping windows were used for calculations.

average of 50–60% was increased to correct for an overrepresentation of mutations in a given M_2 pool. A recommended MQ $\geq$ 30 was consistently ensured for all called variants for quality control. Moreover, we only used SNPs from loci that had a total coverage depth of at least 10 mapped reads at that position. SNV calling was purposely based on these relaxed parameters to avoid false-negative classification of putative useful mutants in target genes based on too strict filtering parameters. This might also have resulted in the calling of a high percentage of non-canonical SNVs. However, validation by Sanger sequencing of known mutants revealed that the positive mutant calls were all correctly called and thus make this data set a useful resource for identification of genotypes with mutations in selected target genes.

Mutation frequencies can vary between different rapeseed EMS mutant populations. (Gilchrist et al., 2013; Harloff et al., 2012; Tang et al., 2020; Wang et al., 2008; Wells et al., 2014). This can be a result of several factors such as EMS concentration and permeability and the physiology and developmental stage of the treated tissue (Henry et al., 2014). Based on previous studies using conventional gel-based assays, the mutation frequency in our EMS population ranged between 1/12 and 1/72 kb (Braatz et al., 2018; Emrani et al., 2015; Guo et al., 2014; Harloff et al., 2012; Karunarathna et al., 2020; Shah et al., 2018). Using an amplicon sequencing approach, a mutation frequency of 1/27 kb was reported from the same TILLING platform by Sashidhar et al. (2019). These differences can be explained because mutation frequencies were estimated based on the length of the amplicons analyzed. Moreover, only protein-coding sequences were studied. In our TbyWGS approach, we estimated a mutation frequency of one high confidence EMS mutation per 23.6 kb of the Express617 reference genome, which is in line with previous estimations from the same EMS population. However, it is expected that mutation frequencies calculated exclusively for gene coding regions will vary compared to the distribution of mutations for the entire genome. Therefore, a more suitable and direct comparison can be drawn from the mutation frequencies of 1/23.64 kb to 1/303.86 kb observed by Tang et al. (2020) from 20 whole genome sequenced single oilseed rape EMS mutants. Mutation frequencies in EMS resources from other crops like rice (Till et al., 2007), tomato (Garcia et al., 2016) (Wells et al., 2014) and wheat (Krasileva et al., 2017) were estimated to be 1/530 kb, 1/125 kb and as high as 35–40 per kb, respectively.

The non-random distribution of mutation across the whole genome is an important result of our study. We report a lower density of EMS-induced mutations at the ends of the chromosomes (Figure 5). Using the representative group of 19 annotated chromosomes in the Express617 reference assembly (chromosomes A01–A10 and C01–C09), we calculated mutation frequencies from different regions of the genome (Figure S5; Table 3). Although the higher number of mutations in the C sub-genome is explained by its genome size, the striking difference in the mutation frequencies between the two sub-genomes deserves a further explanation. One reason could be different GC contents. In the assembled Express617 genome, the average GC contents per 1 Mb windows of the A and C sub-genomes are 35.2% and 36.2%, respectively. Although this difference is not high, we reason that a high GC content invariably affects mutation frequencies because mutations observed were primarily C → T and G → A transitions (55%). Therefore, regions with high GC content could be potential mutation hotspots. Furthermore, it can be speculated that both sub-genomes are not randomly targeted by EMS. At the time of EMS treatment, the vast majority of cells are in the interphase. It is known from studies with polyploid species e.g. wheat (Fussell, 1987; Martinez-Perez et al., 2001) that sub-genomes adopt the 'Rabl configuration' and are located separately in the interphase nucleus. This could have an impact on their accessibility to EMS resulting in varying mutation densities. Also, the frequency and distribution of repetitive elements alter the conformation of the chromatin fiber during the interphase. It is tempting to speculate that EMS targets DNA differentially in condensed and relaxed chromatin.

EMS is considered to confer mainly C → T and G → A transitions (Till et al., 2006). This is reflected in our TbyWGS resource since 55% of all high confidence SNPs were canonical mutations. Using *in vitro* DNA ethylation experiments, Sega (1984) demonstrated a range of possible modifications other than C → T and G → A transitions resulting from DNA depurination and depyrimidation. It was observed that even transversions could originate because a random nucleotide was inserted opposite apurinic or apyrimidic sites. Because we observed 45% non-canonical mutations in our TbyWGS approach, we questioned why such mutations were rarely observed in our previous studies with the Express617 mutant population. Traditionally, functional genomic studies utilizing EMS mutants have focused on identifying mutations within coding regions (Braatz et al., 2018; Emrani et al., 2015; Karunarathna et al., 2020). Moreover, these studies used conventional detection assays utilizing the CEL I endonuclease, an S1 single strand-specific nuclease that cleaves heteroduplexes of mutant and wild-type DNA. It has been reported that, although CEL I identifies all sequence variants, including InDels (Oleykowski et al., 1998), it has the highest preference for cleaving heteroduplexes arising from G → C, G → A, G → T and C → G mutations (Oleykowski et al., 1998; Triques et al., 2007; Yang et al., 2000). Using our TbyWGS approach, we observed that, for high confidence SNPs located in gene coding regions, 76.3% were within this group of mutations highly preferred by CEL I. Out of these, 66.7% were canonical mutations and only 9.6% were G → C, G → T or C → G mutations. We reason that such non-canonical mutations were rarely reported from previous conventional screenings for two main reasons. First, mutant detections were performed mainly within gene coding regions. Second, most non-canonical transitions and transversions within gene coding regions represented the least preferred mismatch sites for CEL I cleavage (Oleykowski et al., 1998). Within gene coding regions, a low frequency of non-canonical mismatch heteroduplexes favored by CEL I is possibly a reason for low detection rates in the past.

We detected 45% non-canonical transitions and transversions. This is in line with a recent report where > 50% non-canonical transitions, transversions and InDels were observed in three out of 10 whole genome sequenced M_2 EMS mutagenized rapeseed plants (Tang et al., 2020). Moreover, our observation is in line with several reports utilizing conventional and sequencing-based detection methods for EMS mutants from rice (Till et al., 2007), maize (Lu et al., 2018), barley (Caldwell et al., 2004), soybean (Lakhssassi et al., 2020) and sunflower (Fanelli et al., 2021). In these studies, the share of non-canonical transitions and transversions ranged from 12.3 to 31%.

Presently, our mutant database encompasses the canonical transitions. Adding a screening method for non-canonical transitions, transversions could be an interesting option and a possible expansion in the future, especially regarding the regulatory 5 kb up- and downstream regions of genes. Deletions and insertions are expected to be present in our M_2 population as a result of EMS treatment, an effect that has been reported before in a study on DNA ethylation experiments (Sega, 1984). Moreover, using an exome sequencing approach, Krasileva et al. (2017) have confirmed the presence of small InDels (< 20 bp) in EMS mutagenized populations of tetraploid and hexaploid wheat. Although small (< 10 bp) InDels can be detected in our Illumina sequenced population, screening for large InDels requires long-range sequencing technology.

In conclusion, our web-accessible whole genome sequencing mutant platform is an unprecedented resource that can serve as a strong foundation for molecular breeding and functional genomics in oilseed rape research. Because the resource represents whole genome sequencing data, it not only enables functional characterization of EMS mutations within genes but also offers an opportunity to analyze mutations within regulatory sequences. In this respect, investigation of mutations detected upstream and downstream of genes or even within intergenic regions is now a possibility and perhaps a compelling proposition for researchers and plant breeders alike. Although CRISPR/Cas mediated editing of multiple gene families has gained immense momentum in the past years, especially for the improvement of oilseed rape (Braatz et al., 2017; Karunarathna et al., 2020; Sashidhar et al., 2020; Zheng et al., 2020), legal restrictions within the European Union have deterred application in current breeding programs. In this regard, EMS mutagenized functional mutants hold great promise for successful integration in plant breeding and crop improvement (Jung & Till, 2021).

In summary, the present study reports a TbyWGS platform for detecting EMS-induced mutations on the whole genome level from a winter rapeseed EMS mutant population. The resource contains 78 083 182 high confidence EMS-induced C → T and. G → A mutations originating from whole genome sequencing data from 1988 M_2 plants. On average, each plant possessed approximately 39 000 mutations with a frequency of 1/23.6 kb on the genome level. Approximately 82% of the mutations were located in 5 kb upstream or downstream (28% each) of gene coding regions or within intergenic regions (26%). The remaining 18% were located within regions coding for genes. Of these, 0.4, 7.0, 4.1, and 4.9% were predicted as nonsense, missense, synonymous and intronic mutations, respectively. On average, 20 757 genes per 4× pool could be inferred to harbor canonical mutations. The present study also notes distinct distribution patterns of EMS induced mutations on a genome-wide scale. Our designed web-based resource incorporates whole genome sequencing data enabling the user-friendly detection of EMS mutations in any genomic region of interest. This effectively bypasses the need for cumbersome gel-based mutation screening techniques for the detection of EMS induced functional mutations in oilseed rape in a time and cost-effective manner.

EXPERIMENTAL PROCEDURES

Plant material

The offspring of the winter oilseed rape inbred line Express617 (F_{11}) had been treated with EMS and M_2 and M_3 generations had been produced previously (Harloff et al., 2012). 1988 M_2 plants (four plants/family) representing 497 M_2 families (four plants/family) were grown under greenhouse conditions (16:8 h light/dark photocycle at 22°C) at the NPZ Innovation GmbH (Holtsee, Germany). Plants were vernalized for 12 weeks at 4°C. All M_2 plants were self-pollinated and M_3 seeds were harvested.

DNA isolation and whole genome sequencing

Twenty 4-mm disks were taken from leaves of each of the M_2 plants before vernalization. DNA was isolated from bulked leaf samples of each M_2 family. The pooled leaf samples were lyophilized (ALPHA 1–4 LDplus; Martin Christ Gefriertrocknungsanlagen GmbH, Osterode am Harz, Germany) for 96 h. Genomic DNA of 497 4× DNA pools was isolated using the DNeasy Plant Mini Kit in accordance with the manufacturer's instructions (QIAGEN GmbH, Hilden, Germany) and sequenced on an NovaSeq 6000 platform (Illumina) using 150-bp paired ends reads and a 350-bp insert DNA library with 20× depth of coverage (Novogene, Co., Ltd, Beijing, China).

Data processing and SNP detection

Raw data were obtained in the FASTQ format. The quality of raw reads was checked using FASTQC, version 0.11.5 (Andrews, 2010) and MULTIQC (Ewels et al., 2016) programs. Based on the quality checks, samples representing low sequencing outputs or reads possessing poor Phred scores, deviating GC count per read and a high number of ambiguously called bases were removed from the analysis. A long-read genome assembly of Express617 was used as the reference genome (Lee et al., 2020). Raw reads from all 4× pools were mapped to the Express617 oilseed rape reference, using BWA-MEM (Li, 2013) with default parameters for local alignment. SAMTOOLS (http://www.htslib.org/) was used to convert, sort, and index the resulting files to binary alignment map files. To estimate the mapping rates, the percentage of mapped reads were calculated using the SAMTOOLS – flagstat function. The average depth and breadth of coverage per pool were calculated using the SAMTOOLS – depth function.

Using PICARD tools (https://broadinstitute.github.io/picard/) AddOrReplaceReadGroups and MarkDuplicates, mapped reads were pre-processed before variant calling. Variants were called using GATK 4.1.4.1 HaplotyeCaller (McKenna et al., 2010; Poplin et al., 2018). The resulting Variant Call Format (vcf) files were filtered for single nucleotide polymorphisms characteristic to the canonical EMS type C → T and G → A transitions. SNPs were retained based on DP ≥ 10, MQ ≥ 30 and AD between 12.5 and 60% parameters using BCFTOOLS (http://samtools.github.io/bcftools/bcftools.html) and custom Unix scripts. AD of mutations was calculated as the percentage share of reads possessing mutations from total reads (DP) mapped at that position. These SNPs are hereafter referred to as 'high confidence' SNPs. The same criteria were used to create a subset of variants representing all other nucleotide substitutions referred to as 'non-canonical' transitions and transversions. The VARIANT EFFECT PREDICTOR (VEP) tool (McLaren et al., 2016) was used in the offline mode to characterize and predict SNP effects on the polypeptide level. The General Feature Format (GFF) file for the Express617 reference assembly (Lee et al., 2020) representing predicted protein models was used to characterize all SNP effects on a genome-wide scale.

Estimation of mutation frequencies

To approximate the mutation frequency per kb of the reference genome, the number of high confidence SNPs along the chromosomal length for each of the annotated chromosomes (A01–A10 and C01–C09) and the non-annotated share of the reference genome was calculated. A genome-wide mutation density was calculated by dividing the number of mutations by the effective genome size (approximately 925 mb). Within 1 b (megabases) non-overlapping windows, the frequency of C → T and G → A transitions and all other transitions and transversions were then calculated from each of the 4× pools for all chromosomes. A moving average for the number of SNPs across all sequenced pools per 1 Mb non-overlapping windows was calculated for all chromosomes. The general additive model (GAM) in the ggplot2 package on R (Wickham, 2009) was used to smoothen the curves while plotting values. The number of putative functional mutations conferred by C → T and G → A transitions and all other mutations were then calculated. Mutations located within the open reading frames of all predicted gene models in the Express617 reference genome were identified to estimate the number of genes harboring at least one canonical mutation.

Developing the web-based mutant database

The web-accessible database EMSBrassica (http://emsbrassica.plantbreeding.uni-kiel.de) was built using DJANGO (https://www.djangoproject.com), a PYTHON-based free and open-source web framework. PostgreSQL (http://www.postgresql.org), an open-source object-relational database system, was used as the database engine. Webpages were built using HTML and BOOTSTRAP (https://getbootstrap.com).

ACKNOWLEDGEMENTS

We thank the Institute of Clinical Molecular Biology in Kiel for providing Sanger sequencing as supported in part by the DFG Clusters of Excellence 'Precision Medicine in Chronic Inflammation' and 'ROOTS'. This work was funded by the Federal Ministry of Education and Research (BMBF) within the framework of the project IRFFA: Improved Rapeseed as Fish Feed in Aquaculture (grant number 031B0357B). Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

SJ, H-JH, AA, FD, CO and CJ designed the research. AA, FD and KB generated the plant material and developed the fourfold pools. SJ conducted the experiments and analyzed the data. SJ and AK designed and developed the webpage. HJH and CJ supervised the research. SJ wrote the original draft. H-JH, AA, CO and CJ reviewed and edited the manuscript. All authors participated in the discussion and revision of the manuscript. The authors read and approved the final version of the manuscript submitted for publication.

CONFLICT OF INTEREST

AA, FD and KB are employed by NPZ Innovation GmbH, Germany. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships and declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets supporting the conclusions of this article are available in the NCBI Sequence Read Archive – BioProject: PRJNA758762. All individual files representing SNP effects from whole genome sequenced pools are accessible at https://doi.org/10.5281/zenodo.6617711. All codes used can be made available from the authors upon request.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Summary of EMS-type transitions originating from the annotated and non-annotated regions of the Express617 genome. Regions of the Express617 reference assembly without chromosomal annotations for the A- or C sub-genomes are denoted as non-annotated regions. The share of C → T and G → A transitions located within annotated and non-annotated regions was calculated as an average from 497 sequenced 4× pools.

Table S2. Summary of validation experiments to confirm mutations in seven M_2 DNA pools harboring previously detected and characterized EMS mutations within four candidate gene families, *BnREF*, *BnSGT*, *BnSFAR4* and *BnSFAR1*. For each of the selected mutants, corresponding M_2 families were identified. Individual read alignments from each of the 4× pools (named with prefix 'D' or 'IR') representing the selected mutant M_2 families were visualized for regions harboring expected mutations.

Table S3. Summary of validation experiments to confirm the presence of mutations in 12 pools harboring EMS mutations within candidate gene families *BnMYB28*, *BnCYP79F1* and *BnGTR2*. 4× pools (named with prefix 'D') with nonsense mutations for the candidate genes were first identified. Genomic DNA was isolated separately from the leaf samples of the four individuals (#1–4) bulked in each pool. Standard PCR with locus-specific primers

was used to amplify regions encompassing the detected mutations. PCR fragments were Sanger sequenced to validate the presence of a mutation.

Figure S1. Screenshot of the FastQC reports showing a graphical overview of quality checks for raw reads from all sequenced 4× pools. Mean quality scores represent Phred scores for individual bases called across the length 150-bp paired-end reads. Phred scores for all 4× pools are within the optimum levels of > 30 (blocks marked in green). Raw reads with Phred score = 20–26 are tolerated (yellow zones), Phred score ≤ 20 (red zones) are rejected. Analysis was performed via FASTQC, version 0.11.5, and MULTIQC.

Figure S2. The share of EMS-induced mutations predicted to have putative functional effects (a) on a genome-wide scale and (b) specifically within coding regions. SNP effects were first predicted on a genome-wide scale for all predicted gene models on the Express617 reference genome. The share of mutation effects within coding regions was calculated using SNPs that were present within and including the translation START and STOP sites. Splice site variants include the sum of splice acceptor and donor site mutations. Upstream and downstream variants encompass 5-kb genomic regions from the START and STOP sites, respectively. The Ensembl VARIANT EFFECT PREDICTOR tool was used for SNP effect prediction using the Express617 reference genome (Lee et al., 2020). *Reflects the mean SNP counts from 497 IRFFA 4× pools.

Figure S3. Characterization of 'Non-EMS' type (a) Transitions (T → C and A → G) and (b) Transversions (C → A, G → T, A → C, T → G, T → A, A → T, G → C and C → G) with predicted effects on a genome-wide scale. Boxplots show the distribution of SNP effects as a mean of 497 4× pools. Mutation effects are predicted using filtered SNPs with DP ≥ 10, AD = 12.5–60% and MQ ≥ 30. Ensembl VARIANT EFFECT PREDICTOR release 99 was used in offline mode. Mutation effects were predicted within the gene models of the Express617 genome. All gene IDs were extracted from the general feature format (GFF) of the Express617 reference genome. Splice site variants include acceptor and donor site mutations. Upstream and downstream variants are located within a distance of 5 kb from the transcription START and STOP sites, respectively. DP, read depth; AD, allele depth; MQ, mapping quality.

Figure S4. Characterization of 'Non-EMS' type (a) transitions (T → C and A → G) and (b) transversions (C → A, G → T, A → C, T → G, T → A, A → T, G → C and C → G) with predicted effects exclusively within coding regions. Boxplots show the distribution of SNP effects as a mean of 497 4× pools. Mutation effects are predicted using filtered SNPs with DP ≥ 10, AD = 12.5–60% and MQ ≥ 30. Ensembl VARIANT EFFECT PREDICTOR release 99 was used in offline mode. Mutation effects were predicted within the gene models of the Express617 genome. All gene IDs were extracted from the general feature format (GFF) of the Express617 reference genome. Splice site variants include acceptor and donor site mutations. Upstream and downstream variants are located within a distance of 5 kb from the transcription START and STOP sites, respectively. DP, read depth; AD, allele depth; MQ, mapping quality.

Figure S5. Distribution and density of filtered mutations per 1 Mb windows across all sequenced pools for chromosomes A01–A10 and C01–C09. The red and blue lines represent the running average of frequency distribution of filtered mutations for EMS type (C → T and G → A transitions) and non-EMS type mutations (all other nucleotide substitutions), respectively, across all pools. Black dots reflect the corresponding mutation densities per 1-Mb bins for each of the sequenced pools. The *x*-axis represents the length of chromosomes in Mb and each interval corresponds to a 1-Mb window. The *y*-axis shows the number of SNPs located within each window. To smoothen the curves, the General Additive Model (GAM) in the ggplot2 package of R was used.

Figure S6. Screenshot of the INTEGRATIVE GENOMICS VIEWER (IGV) showing read alignments from the seven selected fourfold pools harboring previously detected EMS mutations. Pool IDs are named with the prefix 'IR' or 'D'. (1) Tracks showing 41-bp regions with 20 bases up and downstream of the detected mutations. (2) Coverage tracks showing read alignments with horizontal bars representing individual reads. SNPs are marked in different colors. (3) Inset windows representing the count details for each of the called nucleotides (A, adenine; C, cytosine; G, guanine; T, thymine; N, unknown) and their share from all mapped reads at that position.

Figure S7. Screenshot of the INTEGRATIVE GENOMICS VIEWER (IGV) showing read alignments from 12 fourfold pools harboring EMS mutations detected within candidate genes: *BnGTR2*, *BnMYB28* and *BnCYP79F1*. Pool IDs are named with the prefix 'IR' or 'D'. (1) Tracks showing 41-bp regions with 20 bases up and downstream of the detected mutations. (2) Coverage tracks showing read alignments with horizontal bars representing individual reads. SNPs are marked in different colors. (3) Inset windows representing the count details for each of the called nucleotides (A, adenine; C, cytosine; G, guanine; T, thymine; N, unknown) and their individual share from all mapped reads at that position.

Figure S8. Sanger sequencing results from experiments validating the presence of EMS-induced mutations detected in 12 4× pools (Table S3). Using genomic DNA from the four representative individuals of the selected pools (Table S3), PCR amplicons encompassing detected mutants were Sanger sequenced. Genomic sequences of the candidate genes were extracted from the Express617 reference genome. Location and type of detected mutations were annotated for each of the candidate genes (marked in blue arrows). Sanger sequencing reads were mapped to the corresponding candidate gene sequences to check for the presence of the expected mutations. Validated mutations from M_2 individuals are marked with orange arrows. 4× pools are named with the 'D' prefix. The 4 M_2 individuals corresponding to each of the pools have been named with the pool IDs followed by the '_1', '_2', '_3' and '_4' suffixes. Sequence analysis was performed using CLC MAIN WORKBENCH 7.

REFERENCES

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H. *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnology*, **30**(2), 174–178.

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data*. Cambridge, UK: Babraham Bioinformatics, Babraham Institute.

Braatz, J., Harloff, H.-J., Emrani, N., Elisha, C., Heepe, L., Gorb, S.N. *et al.* (2018) The effect of *INDEHISCENT* point mutations on silique shatter resistance in oilseed rape (*Brassica napus*). *Theoretical and Applied Genetics*, **131**(4), 959–971.

Braatz, J., Harloff, H.-J., Mascher, M., Stein, N., Himmelbach, A. & Jung, C. (2017) CRISPR-Cas9 targeted mutagenesis leads to simultaneous modification of different homoeologous gene copies in polyploid oilseed rape (*Brassica napus*). *Plant Physiology*, **174**(2), 935–942.

Caldwell, D.G., McCallum, N., Shaw, P., Muehlbauer, G.J., Marshall, D.F. & Waugh, R. (2004) A structured mutant population for forward and reverse genetics in barley (*Hordeum vulgare* L.). *The Plant Journal*, **40**(1), 143–150.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X. *et al.* (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, **345**(6199), 950–953.

Chen, X., Tong, C., Zhang, X., Song, A., Hu, M., Dong, W. *et al.* (2021) A high-quality *Brassica napus* genome reveals expansion of transposable elements, subgenome evolution and disease resistance. *Plant Biotechnology Journal*, **19**(3), 615–630.

Emrani, N., Harloff, H.-J., Gudi, O., Kopisch-Obuch, F. & Jung, C. (2015) Reduction in sinapine content in rapeseed (*Brassica napus* L.) by induced mutations in sinapine biosynthesis genes. *Molecular Breeding*, **35**(1), 37. Available from: https://doi.org/10.1007/s11032-015-0236-2

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**(19), 3047–3048.

Fanelli, V., Ngo, K.J., Thompson, V.L., Silva, B.R., Tsai, H., Sabetta, W. *et al.* (2021) A TILLING by sequencing approach to identify induced mutations in sunflower genes. *Scientific Reports*, **11**(1), 9885. Available from: https://doi.org/10.1038/s41598-021-89237-w

Fang, D.D., Naoumkina, M., Thyssen, G.N., Bechere, E., Li, P. & Florane, C.B. (2020) An EMS-induced mutation in a tetratricopeptide repeat-like superfamily protein gene (*Ghir_A12G008870*) on chromosome A12 is responsible for the li_y short fiber phenotype in cotton. *Theoretical and Applied Genetics*, **133**(1), 271–282. Available from: https://doi.org/10.1007/s00122-019-03456-4

Fussell, C.P. (1987) The Rabl orientation: a prelude to synapsis. In: Moens, P.B. (Ed.) *Meiosis*. New York: Academic Press, pp. 275–299.

Garcia, V., Bres, C., Just, D., Fernandez, L., Tai, F.W.J., Mauxion, J.-P. *et al.* (2016) Rapid identification of causal mutations in tomato EMS populations via mapping-by-sequencing. *Nature Protocols*, **11**(12), 2401–2418. Available from: https://doi.org/10.1038/nprot.2016.143

Gigolashvili, T., Engqvist, M., Yatusevich, R., Müller, C. & Flügge, U.I. (2008) HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in *Arabidopsis thaliana*. *New Phytologist*, **177**(3), 627–642. Available from: https://doi.org/10.1111/j.1469-8137.2007.02295.x

Gilchrist, E.J., Sidebottom, C.H., Koh, C.S., MacInnes, T., Sharpe, A.G. & Haughn, G.W. (2013) A mutant *Brassica napus* (canola) population for the identification of new genetic diversity via TILLING and next generation sequencing. *PLoS One*, **8**(12), e84303.

Guo, Y., Harloff, H.-J., Jung, C. & Molina, C. (2014) Mutations in single *FT*- and *TFL1*-paralogs of rapeseed (*Brassica napus* L.) and their impact on flowering time and yield components. *Frontiers in Plant Science*, **5**, 282.

Harloff, H.-J., Lemcke, S., Mittasch, J., Frolov, A., Wu, J.G., Dreyer, F. *et al.* (2012) A mutation screening platform for rapeseed (*Brassica napus* L.) and the detection of sinapine biosynthesis mutants. *Theoretical and Applied Genetics*, **124**(5), 957–969.

Henry, I.M., Nagalakshmi, U., Lieberman, M.C., Ngo, K.J., Krasileva, K.V., Vasquez-Gross, H. *et al.* (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *The Plant Cell*, **26**(4), 1382–1397.

Jung, C. & Till, B. (2021) Mutagenesis and genome editing in crop improvement: perspectives for the global regulatory landscape. *Trends in Plant Science.*, **26**, 1258–1269.

Karunarathna, N.L., Wang, H., Harloff, H.J., Jiang, L. & Jung, C. (2020) Elevating SEED oil content in a polyploid crop by induced mutations in *SEED FATTY ACID REDUCER* genes. *Plant Biotechnology Journal*, **18**(11), 2251–2266.

Krasileva, K.V., Vasquez-Gross, H.A., Howell, T., Bailey, P., Paraiso, F., Clissold, L. *et al.* (2017) Uncovering hidden variation in polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America*, **114**(6), E913–E921.

Lakhssassi, N., Lopes-Caitar, V.S., Knizia, D., Cullen, M.A., Badad, O., El Baze, A. *et al.* (2021) TILLING-by-sequencing$^+$ reveals the role of novel fatty acid desaturases (GmFAD2-2 s) in increasing soybean seed oleic acid content. *Cell*, **10**(5), 1245.

Lakhssassi, N., Zhou, Z., Liu, S., Piya, S., Cullen, M.A., El Baze, A. *et al.* (2020) Soybean TILLING-by-sequencing$^+$ reveals the role of novel *GmSACPD* members in the unsaturated fatty acid biosynthesis while maintaining healthy nodules. *Journal of Experimental Botany.*, **71**, 6969–6987.

Lee, H., Chawla, H.S., Obermeier, C., Dreyer, F., Abbadi, A. & Snowdon, R. (2020) Chromosome-scale assembly of winter oilseed rape *Brassica napus*. *Frontiers in Plant Science*, **11**, 496. Available from: https://doi.org/10.3389/fpls.2020.00496

Lee, Y.-H., Park, W., Kim, K.-S., Jang, Y.-S., Lee, J.-E., Cha, Y.-L. *et al.* (2018) EMS-induced mutation of an endoplasmic reticulum oleate desaturase gene (*FAD2-2*) results in elevated oleic acid content in rapeseed (*Brassica napus* L.). *Euphytica*, **214**(2), 1–12.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Lu, X., Liu, J., Ren, W., Yang, Q., Chai, Z., Chen, R. *et al.* (2018) Gene-indexed mutations in maize. *Molecular Plant*, **11**(3), 496–504.

Martinez-Perez, E., Shaw, P. & Moore, G. (2001) The *Ph1* locus is needed to ensure specific somatic and meiotic centromere association. *Nature*, **411** (6834), 204–207.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20** (9), 1297–1303. Available from: https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A. *et al.* (2016) The Ensembl variant effect predictor. *Genome Biology*, **17**(1). https://doi.org/10.1186/s13059-016-0974-4

Nie, S., Wang, B., Ding, H., Lin, H., Zhang, L., Li, Q. *et al.* (2021) Genome assembly of the Chinese maize elite inbred line RP125 and its EMS mutant collection provide new resources for maize genetics research and crop improvement. *The Plant Journal*, **108**, 40–54.

Nour-Eldin, H.H., Madsen, S.R., Engelen, S., Jørgensen, M.E., Olsen, C.E., Andersen, J.S. *et al.* (2017) Reduction of antinutritional glucosinolates in *Brassica* oilseeds by mutation of genes encoding transporters. *Nature Biotechnology*, **35**, 377–382. Available from: https://doi.org/10.1038/nbt.3823

Oleykowski, C.A., Bronson Mullins, C.R., Godwin, A.K. & Yeung, A.T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Research*, **26**(20), 4597–4602.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A. *et al.* (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178.

Rahman, H. (2013) Review: breeding spring canola (*Brassica napus* L.) by the use of exotic germplasm. *Canadian Journal of Plant Science*, **93**(3), 363–373. Available from: https://doi.org/10.4141/cjps2012-074

Reintanz, B., Lehnen, M., Reichelt, M., Gershenzon, J., Kowalczyk, M., Sandberg, G. *et al.* (2001) Bus, a bushy *Arabidopsis CYP79F1* knockout mutant with abolished synthesis of short-chain aliphatic glucosinolates. *The Plant Cell*, **13**(2), 351–367.

Rousseau-Gueutin, M., Belser, C., Da Silva, C., Richard, G., Istace, B., Cruaud, C. *et al.* (2020) Long-read assembly of the *Brassica napus* reference genome Darmor-bzh. *GigaScience*, **9**(12), giaa137.

Sashidhar, N., Harloff, H.J. & Jung, C. (2019) Identification of phytic acid mutants in oilseed rape (*Brassica napus*) by large scale screening of mutant populations through amplicon sequencing. *New Phytologist*, **225** (5), 2022–2034.

Sashidhar, N., Harloff, H.J., Potgieter, L. & Jung, C. (2020) Gene editing of three *BnITPK* genes in tetraploid oilseed rape leads to significant reduction of phytic acid in seeds. *Plant Biotechnology Journal*, **18**(11), 2241–2250.

Sega, G.A. (1984) A review of the genetic effects of ethyl methanesulfonate. *Mutation Research/Reviews in Genetic Toxicology*, **134**(2–3), 113–142.

Shah, S., Karunarathna, N.L., Jung, C. & Emrani, N. (2018) An *APETALA1* ortholog affects plant architecture and seed yield component in oilseed rape (*Brassica napus* L.). *BMC Plant Biology*, **18**(1), 380. Available from: https://doi.org/10.1186/s12870-018-1606-9

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S. *et al.* (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants*, **6**(1), 34–45.

Tang, S., Liu, D.X., Lu, S., Yu, L., Li, Y., Lin, S. *et al.* (2020) Development and screening of EMS mutants with altered seed oil content or fatty acid composition in *Brassica napus*. *The Plant Journal.*, **104**, 1410–1422.

Till, B.J., Cooper, J., Tai, T.H., Colowit, P., Greene, E.A., Henikoff, S. *et al.* (2007) Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biology*, **7**(1), 1–12.

Till, B.J., Zerr, T., Comai, L. & Henikoff, S. (2006) A protocol for TILLING and Ecotilling in plants and animals. *Nature Protocols*, **1**(5), 2465–2477.

Triques, K., Sturbois, B., Gallais, S., Dalmais, M., Chauvin, S., Clepet, C. *et al.* (2007) Characterization of *Arabidopsis thaliana* mismatch specific endonucleases: application to mutation discovery by TILLING in pea. *The Plant Journal*, **51**(6), 1116–1125.

Wang, B., Wu, Z., Li, Z., Zhang, Q., Hu, J., Xiao, Y. *et al.* (2018) Dissection of the genetic architecture of three seed-quality traits and consequences for breeding in *Brassica napus*. *Plant Biotechnology Journal*, **16**(7), 1336–1348.

Wang, N., Wang, Y., Tian, F., King, G.J., Zhang, C., Long, Y. *et al.* (2008) A functional genomics resource for *Brassica napus*: development of an EMS mutagenized population and discovery of *FAE1* point mutations by TILLING. *New Phytologist*, **180**(4), 751–765.

Wells, R., Trick, M., Soumpourou, E., Clissold, L., Morgan, C., Werner, P. *et al.* (2014) The control of seed oil polyunsaturate content in the polyploid crop species *Brassica napus*. *Molecular Breeding*, **33**(2), 349–362.

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis (use R!)*. New York: Springer, pp. 970–978.

Yang, B., Wen, X., Kodali, N.S., Oleykowski, C.A., Miller, C.G., Kulinski, J. *et al.* (2000) Purification, cloning, and characterization of the CEL I nuclease. *Biochemistry*, **39**(13), 3533–3541.

Zheng, M., Zhang, L., Tang, M., Liu, J., Liu, H., Yang, H. *et al.* (2020) Knockout of two *BnaMAX1* homologs by CRISPR/Cas9-targeted mutagenesis improves plant architecture and increases yield in rapeseed (*Brassica napus* L.). *Plant Biotechnology Journal*, **18**(3), 644–654.