Short title: Relevance of off-target changes in genome editing

Co-Corresponding Authors: Peter L. Morrell (pmorrell@umn.edu); Robert M. Stupar (stup0004@umn.edu), Department of Agronomy and Plant Genetics, 1991 Upper Buford Circle, 411 Borlaug Hall, St. Paul, MN 55108-6026 USA

Full title: Plant genome editing and the relevance of off-target changes

Author Names and Affiliations: Nathaniel Grahama,b, Gunvant B. Patilc, David M. Bubeckd, Raymond C. Dobb, Kevin C. Glenné, Annie T. Gutscheđ, Sandeep Kumarđ, John A. Lindbof, Luis Maa, Gregory D. Mayđ, Miguel E. Vega-Sancheze, Robert M. Stuparc, Peter L. Morrellc

a Department of Genetics, Cell Biology and Development, University of Minnesota, St. Paul, MN 55108 USA
b Pairwise, Durham, NC 27709 USA
c Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108 USA
d Corteva Agriscience™, Johnston, IA 50131, USA
e Bayer Crop Science, Chesterfield, MO 63017 USA
f HM Clause, Davis, CA 95618 USA
g Enza Zaden Research USA, San Juan Bautista, CA 95045 USA

Present address: Department of Plant and Soil Science, Texas Tech University, Lubbock, TX 79409.
**One Sentence Summary:** With well-designed protocols and guide RNAs, off-target changes induced by site-directed nucleases are negligible with fewer genetic differences than from standing variation or induced mutagenesis.

**Conflict of interest statement:** Some of the authors are employees of Corteva Agriscience, Bayer Crop Science, HM Clause, Enza Zaden, or Pairwise, which are companies involved in targeted genetic modification of plants. R.M.S. is a co-inventor of a patent in the plant gene-editing space. All other authors report no conflicts of interest relevant to this article.

**List of author contributions:** The authors contributed equally in conceiving, drafting, editing, and critical review of the text, tables, and figures of this review.

**Abstract**

Site-directed nucleases (SDNs) used for targeted genome editing are powerful new tools to introduce precise genetic changes into plants. Like traditional approaches, such as conventional crossing and induced mutagenesis, genome editing aims to improve crop yield and nutrition. Next-generation sequencing studies demonstrate that across their genomes, populations of crop species typically carry millions of single nucleotide polymorphisms and many copy number and structural variants. Spontaneous mutations occur at rates of approximately $10^{-8}$ to $10^{-9}$ per site per generation, while variation induced by chemical treatment or ionizing radiation results in higher mutation rates. In the context of SDNs, an off-target change or edit is an unintended, non-specific mutation occurring at a site with sequence similarity to the targeted edit region. SDN-mediated off-target changes can contribute to a small number of additional genetic variations compared to those that occur naturally in breeding populations or are introduced by induced-mutagenesis methods. Recent studies show that using computational algorithms to design genome editing reagents can mitigate off-target edits in plants. Finally, crops are subject
to strong selection to eliminate off-type plants through well-established multigenerational
breeding, selection, and commercial variety development practices. Within this context, off-
target edits in crops present no new safety concerns compared to other breeding practices. The
current generation of genome editing technologies is already proving useful to develop new
plant varieties with consumer and farmer benefits. Genome editing will likely undergo improved
editing specificity along with new developments in SDN delivery and increasing genomic
characterization, further improving reagent design and application.

Introduction: Plant Genetic Variability

Genetic differences between individuals are the basis of adaptation and evolution. Plant
breeding, as a form of directed evolution, has a long history of using genetic diversity for crop
improvement. During the process of crop domestication, humans selected individual plants with
favorable traits that resulted from novel mutations or standing variation in the ancestral species.
The process of selecting plant varieties with favorable characteristics for cultivation and
consumption continues to the present day. Modern plant breeding is a more directed process
than the crop improvement that occurred through the history and pre-history of most cultivated
species, but it continues to involve lengthy development cycles. Progress in our understanding
of the genetic basis of traits and refined approaches to select superior progeny have contributed
to increased yield for most major crops. Globally, crop yields have increased by greater than
one percent each year for much of the past century (Xu et al., 2017). Technological innovations,
including those that increased our understanding of nucleic acid sequence and function, have
contributed to this progress. Advances have primarily been incremental, with few step changes
(Bernardo, 2016). The relatively recent introduction of genome editing, which permits targeted
genetic changes, is part of a continuum of development of plant breeding techniques (Lusser et
al., 2012). Its primary appeal is that, in contrast to established mutagenesis techniques,
changes can be targeted to desired genes.
Over the past several decades, there has been rapid development of genome editing techniques (also referred to as gene editing or genome engineering) (Gaj et al., 2013; Jaganathan et al., 2018; Zhang et al., 2018). These technologies make use of site-directed nucleases (SDNs) and have broad applicability, including applications in many crop plants (Podevin et al., 2013; Voytas, 2013; Weeks et al., 2016). As described in greater detail below, SDNs are enzymes that create DNA double-strand breaks (DSB) at specific, pre-determined nucleotide sequences sites, permitting targeted genetic modifications (Voytas, 2013; Weeks et al., 2016). The most robust and best-characterized SDNs are the Zinc-Finger Nucleases (ZFN), Transcription Activator-Like Effector Nucleases (TALEN), and the type II Clustered Regularly-Interspaced Short Palindromic Repeats-CRISPR-associated proteins (e.g., Cas9 or Cas12a) systems, hereafter referred to as “CRISPR” (Gaj et al., 2013; Jaganathan et al., 2018; Zhang et al., 2018). In all cases, the targeted DSBs generated by these precision nucleases are repaired by the cell’s native DNA break repair mechanisms: the non-homologous end-joining (NHEJ) or the homologous recombination (HR) pathways (Danner et al., 2017).

One area of consideration for genome editing of plants, animals, and in human therapeutic applications is the potential for “off-target” edits that could lead to unintended mutations at untargeted loci in the genome. An off-target edit is an unintended, non-specific mutation that occurs at a site with sequence similarity to the targeted edit region. Off-target edits are distinct from untargeted mutations that occur at sites without similarity to the target region and that are usually associated with spontaneous or other type of background variation. The fidelity of SDNs to edit only the targeted sequence has been questioned, particularly in mammalian genome editing applications (Carroll, 2011; Pattanayak et al., 2011; Carroll, 2013; Wu et al., 2014; Zhang et al., 2015; Tsai and Joung, 2016; Kosicki et al., 2018). The potential to introduce DNA sequence changes at non-targeted loci in patients undergoing genome editing procedures has become an essential consideration for clinical and disease treatment applications (Araki and...
Furthermore, unintended mutations in treated human somatic cells could pose health risks such as unregulated cell proliferation (cancer) (Cox et al., 2015). Off-target edits were detected in the early stages of CRISPR-Cas9 research involving human cancer cell lines (Fu et al., 2013). The off-target edit frequency in human cells ranged from 0 to 150 mutations per genome, and this frequency was closely linked to the single-guide RNA (sgRNA) specificity (Frock et al., 2015). The potential for off-target edits and rate at which they could occur in plants are discussed below.

Plants differ from animals in substantive ways that alter the impact of induced changes. First, somatic cell changes in plants differ in consequences from those in mammals. Somatic changes in a portion of the plant are less likely to affect critical (irreplaceable) tissues. Unlike many animals, genetic changes in juvenile plants can be transmitted to reproductive tissues (Schmid-Siegert et al., 2017). However, plants frequently develop multiple independent reproductive structures, only a portion of which may be affected by any new mutations. Also, breeding to develop new lines for commercial release involves an intensive process of selection of individual plants with useful phenotypes while eliminating individuals with undesirable mutations or phenotypes (commonly known as “off-types”) (ASTA, 2016; Glenn et al., 2017; Kaiser et al., 2020). For these reasons, off-target edits in crops present fewer safety concerns than those that could arise with therapeutic applications of genome editing.

The goal of this review is a quantitative comparison of genetic variation that could arise through genome editing with the variation that exists in crop species or is generated by current breeding practices. The review examines our current understanding of the nature and abundance of standing and induced genetic variation in crops and how they compare to the potential for off-target changes during genome editing.

Sources and Types of Inherent (Standing) Genomic Variation in Plants
Before the identification of DNA as the molecular basis for the transmission of heritable genetic variation, crop domestication and plant breeding involved visible inspection and selection of desired phenotypes. There was limited knowledge of the types and extent of underlying genomic variation in crop plants and their wild relatives (Zohary, 1999). The development of genomic sequencing methodologies has provided an unprecedented understanding of the molecular basis of diversity in plant species, ranging from single nucleotide variants to macro-scale events such as genomic rearrangements or whole genome duplication leading to polyploidy and transposon-directed DNA transposition (Wendel et al., 2016; Gabur et al., 2019).

Our understanding of the primary sources of genetic diversity, namely mutations and recombination, and the cellular processes involved in generating this variation, have increased as both DNA sequencing technologies and genome assemblies have improved (Bevan et al., 2017). Mutation creates new variants while recombination arranges variants into new combinations (Morrell et al., 2006); both processes rely upon endogenous DNA repair mechanisms (see Box 1 on double-strand DNA break repair). Genetic drift and both positive and purifying selection remove variants from populations, maintaining an equilibrium that permits variation to persist at relatively constant levels in stable populations. Much of the process of plant breeding involves the creation of varieties that combine favorable variants from their parental lines.

**Reference genomes and assessment of genomic variation**

Standing variation in a species can be estimated at the whole genome level by comparing a sample of individuals using Next Generation Sequencing (NGS). High-throughput DNA sequencing, whole-genome reference sequences and assembly, and improved bioinformatic tools have provided an unprecedented understanding of the variation in populations of the
world’s major field and vegetable crop species (Morrell et al., 2011; Huang and Han, 2014; Bevan et al., 2017). NGS and reference-based read mapping have also been used to identify de novo sequence variation in lines accumulating induced and spontaneous mutations over several generations (Ossowski et al., 2010; Belfield et al., 2012; Henry et al., 2014). Many of the field crop and vegetable species that constitute a major part of the global diet now have a high-quality reference genome sequence (Michael and VanBuren, 2015). Reference genomes have several limitations, the most apparent being that all genes or gene variants are not present in any single accession (Hirsch et al., 2016; Jiao et al., 2017). Means of accounting for structural variation and gene content are rapidly evolving, but generally involve representing multiple genomes of a single species in a graph-like structure (Church et al., 2015; Paten et al., 2017).

A list of the classes of variation detected in cultivated plants is presented in Table 1. Single nucleotide variations (SNVs) (referred to as single nucleotide polymorphisms (SNPs) when segregating in populations), are the most abundant class (Innan et al., 1996). Because of their abundance, SNPs are the most commonly genotyped polymorphism for breeding applications and comparative studies in crops (Morrell et al., 2011). Insertion and deletion (indels) changes are also common forms of variation, but because indel events are superimposed over one another, over evolutionary time it becomes increasingly difficult to distinguish the boundaries of individual insertion or deletion events (Clegg and Morrell, 2004). These naturally occurring changes are similar to those induced by SDNs, with SNVs being equivalent to base edits (see below) and indels resembling the gene knockouts induces by CRISPR edits.

It is important to note that most whole genome sequencing studies to date have used short-read sequencing technologies (Gilad et al., 2009). As a result, the diversity in breeding populations due to structural variation such as variation in transposable element location and abundance,
presence-absence variation (PAV) and gene copy number variants (CNV) has been difficult to measure (Muñoz-Diez et al., 2012). PAV and CNV typically refer to changes that include genes and thus are functionally equivalent to SDN-mediated gene insertion or deletion. Although structural variants are less common than SNPs, they are an important source of variation (Gabur et al., 2019; Schiessl et al., 2019). There are many examples of gene presence/absence or copy number differences impacting agronomic traits including submergence tolerance in rice (Oryza sativa) (Xu et al., 2006), high oleic acid content in sunflower oil (Schuppert et al., 2006), soybean cyst nematode resistance (Cook et al., 2012), and seed coat pigmentation in soybean (Glycine max) (Todd and Vodkin, 1996). Because of the density of naturally occurring variation, intra- and inter-specific crosses of plants that occur during plant breeding can introduce millions of SNPs (see Figure 1), in addition to thousands of PAV or CNV sequence variations into resulting progeny. As more long-read sequencing technologies are deployed, more accurate measurements of the CNV and PAV in breeding populations will be possible.

**Sources of spontaneous de novo genome variation**

Table 2 reports current estimates of standing genomic variation in select crop species. Environmental conditions and breeding and genetic practices such as tissue culture propagation (described below) can further impact the rate at which mutations occur. The sources, types, and rates of de novo genomic variation are reviewed below.

[INSERT TABLE 2 HERE]

- **Spontaneous Nucleotide Variation:**

  *De novo* mutations arise primarily from errors in the DNA replication process. DNA repair processes (NHEJ and HR) can also result in new mutations, including variants that arise due to environmentally-triggered DNA damage (Jiang et al., 2014; Nisa et al., 2019). Whole-genome sequencing of mutation accumulation lines of Arabidopsis (Arabidopsis thaliana) estimated a
mutation rate of $7 \times 10^{-9}$ base substitutions per site per sexual generation (Ossowsky et al., 2010), a rate confirmed in a more recent report (Weng et al., 2019). This rate estimate in Arabidopsis is consistent with earlier estimates of a mutation rate of $5.9 \times 10^{-9}$ substitutions per site per year based on phylogenetic divergence in monocots (Gaut et al., 1996). Most of the mutations reported by Ossowsky et al. (2010) were SNPs, though 1-3 bp indel mutations were also detected at a rate of $4.0 \pm 1.6 \times 10^{-10}$ indels per site per generation. Deletions larger than 3 bp occurred at a rate of $0.5 \times 10^{-9}$ per site per generation, and the average larger deletion size was hundreds of bp per event (Ossowsky et al., 2010). In a similar study in rice, the mutation rate was estimated at $5.4 \times 10^{-8}$ per site per diploid genome (Tang et al., 2018). In maize (Zea mays), the spontaneous mutation rate was reported to be $2.2-3.9 \times 10^{-8}$ per site per generation (Yang et al., 2017). The general trend for single nucleotide mutation rates in plants is on the order of $10^{-8}$ to $10^{-9}$ per site per generation, similar to estimates for humans (Nachman and Crowell, 2000; Roach et al., 2010).

- **Transposons as a source of variation:**

In many plant species, transposable elements (TE) and the remnants of past TE insertion events make up the majority of the genome (Sahebi et al., 2018). For example, TEs make up approximately 85% of the maize genome, 76% of the pepper genome, and between 20-40% of the rice genome, depending upon the rice cultivar (Dubin et al., 2018; Anderson et al., 2019). Transposons are DNA elements that can auto-excise and re-insert in the genome. They have the potential to disrupt genes or create rearrangements. They can also result in complete copies of genetic elements being placed in other areas of the genome. Transposons can have a profound impact on genome structure and gene function (Lisch, 2012). Some types of transposons are known to be activated by environmental conditions (e.g., temperature) and by breeding practices, (e.g., tissue culture, described below) (Vitte et al., 2014; Rey et al., 2016).
Accurate measurements of rates of heritable transposition in a single generation in plants have not been determined.

**Summary of inherent genomic variation**

NGS has dramatically improved the estimation of genome-wide standing variation and made it possible to estimate rates of *de novo* mutations, especially SNVs. Numerous studies have examined the genomic variation between individuals in breeding populations for well-studied crops like maize (Hufford et al., 2012; Lu et al., 2015; Bukowski et al., 2017), soybean (Li et al., 2014; Maldonado dos Santos et al., 2016; Valliyodan et al., 2016), and rice (Zhao et al., 2018).

To date, short-read sequencing technology has been the principal tool used for resequencing plant genomes. However, short-read sequence technologies are of more limited utility for detection of large structural variants (such as PAV or CNV) (Gilad et al., 2009; Morrell et al., 2011; Fuentes et al., 2019). As long-read sequencing becomes a more standard component of comparative studies and additional reference-quality genomes from individual species are reported, diversity due to large structural changes will become more readily accessible (Gabur et al., 2019; Schiessl et al., 2019).

**Sources of Induced Genomic Variation in Plants**

Although naturally-occurring genetic variants have been the primary source of plant genetic diversity used for crop domestication and breeding, geneticists have developed approaches to augment natural plant genomic diversity and accelerate the development of improved cultivars. Genetic variation induced by chemical mutagens and ionizing radiation has been used in plant breeding since early in the 20th century (Friedberg et al., 2006). Somaclonal variation induced by tissue culture has also been a source of variation used in plant breeding. Mutation breeding is considered cost-effective, ubiquitously applicable, non-hazardous, environmentally friendly, and continues to benefit plant breeders and consumers globally (Ahloowalia et al., 2004). Many
types of mutagenic and ionizing agents have been evaluated to maximize genomic variation while minimizing detrimental or lethal effects in plants (Sikora et al., 2011). As described below, each type of mutagenic agent tends to have a predominant type of mutation induced across the genome.

The mutations induced by chemical and ionizing radiation occur throughout the genome, but at an increased rate compared to spontaneous mutations (Belfield et al., 2012; Henry et al., 2014). It can be difficult to select individuals with mutations for the desired phenotype, yet lacking mutations that produce undesirable phenotypes (Bolon et al., 2011; Saika et al., 2011). Mutant lines that carry numerous new mutations have been developed for genetic studies and breeding improved varieties in many crops (Tables 3 and 4). Over 3,200 mutant varieties, including cereals (1,584; 49.5%), flowers/ornamentals (700; 21.9%), legumes (480; 15%) and others (435; 13.6%), have been released for commercial use in more than 210 plant species in over 70 countries (Figure 2).

[INSERT TABLES 3 and 4 HERE]

Mutant population screening is commonly used in rice breeding, resulting in 828 registered lines in the Mutant Variety Database (Kharkwal et al., 2004; Shu et al., 2012; Wang and Jia, 2014). Most of the wheat (Triticum aestivum) varieties used for making Italian pasta were derived in part through mutation breeding (Micke et al., 1990; Mba, 2013). For some crops, mutation breeding is more widely utilized for crop improvement within specific geographic regions. For example, nearly 50% of the improved soybean cultivars grown in Vietnam are derived from induced mutations (Vinh et al., 2009). Mutagenesis has also been used to introduce desirable traits in horticultural crops. For example, researchers have used induced mutations to develop variation for tomato (Solanum lycopersicum) fruit size (Just et al., 2013; Park et al., 2014), starch content in potato (Solanum tuberosum) (Muth et al., 2008), and virus resistance in
peppers (*Capsicum annuum*) (Ibiza et al., 2010). Several grapefruit (*Citrus x paradisi* Macfad.)
varieties, including popular red flesh varieties, were developed using induced mutation (Mba,
2013). Although induced mutant plant varieties have enjoyed considerable global commercial
success and public acceptance, the molecular and physiological basis for the improved
characteristics in these varieties is generally unknown (Shu et al., 2012).

The genomic changes induced by mutagenesis or tissue culture derived plants have been
evaluated in numerous studies using NGS techniques (Tables 3 and 4). In addition to WGRS
methods (Table 3), TILLING studies have been used to estimate induced mutation density by
collecting sequencing data on a limited number of genes from a large population of individual
mutagenized plants (Table 4). Exome capture and whole exome resequencing studies also rely
on targeted sequencing of a portion of the genome, namely the protein coding regions (exons),
of mutagenized plant populations.

For example, mutations induced via fast neutron irradiation generate a wide variety of mutations
that differ in size and copy number, including single nucleotide variants, deletions, insertions,
inversions, translocations, and duplications (Bolon et al., 2014; Li et al., 2017; Jo and Kim,
2019). Li et al. (2017) generated and deep-sequenced a fast neutron mutant population (1,504
lines) in the model rice cultivar, Kitaake. The estimated mutation frequency analyzed was 61
mutations/plant (1.6 x 10^{-7} per bp, as shown in Table 3) with varied types of mutations observed
in 58% of all known rice genes. Induced mutations can differ from spontaneous mutations in
subtle ways. Transversions (a purine base substitutes for a pyrimidine base, or vice versa) were
found to be more frequent in fast neutron populations (Belfield et al., 2012). Transitions (purine
to purine or pyrimidine to pyrimidine substitutions) are concentrated at pyrimidine dinucleotide
sites, suggesting covalent linkages as a molecular mechanism for these mutations (Belfield et
al., 2012). Other mutagens may generate a narrower range of mutation types (Tables 3 and 4;
Figure 2). For example, EMS primarily creates transitions at guanine sites (Henry et al., 2014).
These changes have a slightly elevated probability of being flanked at the 5’ neighboring nucleotide by adenine or guanine and the 3’ side by cytosine (Henry et al., 2014). There is also evidence that some plant genomes can tolerate elevated mutation rates better than others. The hexaploid wheat and oat polyploid genomes, possessing multiple redundant alleles, can tolerate a larger number of accumulated mutations without any notable impact on survival or agronomic performance (Stadler, 1929; Krasileva et al., 2017; Hussain et al., 2018).

**Genome variation induced by conventional tissue culture practices**

Many modern breeding programs use tissue culture or *in vitro* culture systems to propagate plants, preserve elite genotypes, or for production of pure lines using double haploidy. Similar to chemical and irradiation-induced mutations, *in vitro* culture methods can also generate *de novo* genetic variability (referred to as somaclonal variation) (Neelakandan and Wang, 2012). Several studies have examined the rate of *de novo* sequence variation of endogenous genes in plants regenerated through tissue culture. Jiang et al. (2011) regenerated 28 Arabidopsis plants from root explants (living cells transferred to culture medium) from a single Columbia (Col-0) laboratory strain (Jiang et al., 2011). Five regenerated plants were sequenced and analyzed, and on average, each plant had ~30 *de novo* sequence changes (totaling 21 indels of 2 bp or fewer, and 131 single nucleotide variants substitutions across the five plants) (Jiang et al., 2011). The calculated somaclonal mutation frequencies of spontaneous mutations in this experiment was between $4 \times 10^{-7}$ and $24 \times 10^{-7}$ mutations per nucleotide. Tang et al. (2018) reported a mutation rate of $1.86 \times 10^{-7}$ in rice plants regenerated through tissue culture. These values are nearly two orders of magnitude higher than the rate Ossowski et al. (2010) estimated from greenhouse-grown plants derived from single seed descent (Shaw et al., 2000).

Somaclonal variation provides another source of genomic variability in plants (Krishna et al., 2016). Nearly two hundred commercial varieties from approximately 55 plant species (e.g., horticultural crops, cereals, legumes, medicinal, and aromatic plants) were derived through
somaclonal variation (reviewed by Neelakandan and Wang, 2012; Krishna et al., 2016). During the in vitro culture process, many regenerated plants develop differences in appearance relative to the parental genotype. Changes induced can include heritable genetic and epigenetic alterations (Bednarek et al., 2007; Machczyńska et al., 2015; Krishna et al., 2016). As an example, in the popular cultivar of banana (Musa acuminata) ‘Cavendish,’ the occurrence of off-types (different from the parental clone) from in vitro cultured plantlets ranges from 6 to 38% (Sahijram et al., 2003). The genomic variation induced during in vitro regeneration is influenced by media composition, length of incubation and subculture, explant origin, and genotype. Genome-wide DNA polymorphisms or structural variations resulting from tissue culture have been observed in soybean (Anderson et al., 2016), barley (Hordeum vulgare) (Bednarek et al., 2007), rice (Miyao et al., 2012; Zhang et al., 2014), potato (Fossi et al., 2019), papaya (Carica papaya) (Kaity et al., 2009), banana (Ray et al., 2006), and several other plant species. Bednarek et al. (2007) assessed somaclonal variation in barley and identified 6% gene expression variation (1.7% of this variation was due to nucleotide changes and the balance was due to altered methylation). Recently, Li et al. (2019) performed whole-genome sequencing of cotton plants derived from in vitro culture and CRISPR-Cas9 edited plants. They determined that most of the mutations were induced either by somaclonal variation or were attributable to heterogeneity present in parental plants (Li et al., 2019). Tissue culture and somaclonal variation have been successfully used to generate valuable heritable genomic changes in crops, and have been widely accepted by breeders for germplasm development and release of new varieties.

In mutation breeding (i.e., irradiation, chemical, or somaclonal), efforts are focused on identifying beneficial and novel phenotypes. However, in addition to the beneficial mutation(s), other “coincident” mutations or rearrangements can be generated elsewhere in the genome (Figure 3). While most of these mutations will have negligible or no phenotypic effect, it can be
necessary to remove undesirable mutations with observable deleterious effects (off-types)
through backcrossing and selection. Each generation of backcrossing is expected to reduce the
presence of coincident mutations by 50%. However, given their number and genomic
distribution, some mutations may persist through this process. This process of backcrossing and
selection is one reason why mutation breeding has been successfully used for decades with a
recognized history of safe use (Food and Drug Administration, 1992).

**Genome Editing as a Source of Variation**

The advantage of classical mutation breeding is that it allows for the rapid generation of
genomic variation that can be used as a source of traits not currently identified in a crop’s
germplasm. The drawback of these methods is that variation is induced randomly both at
desired locations and across the genome. Because of these factors, screening, isolating, and
introgressing desired mutations into elite germplasm requires large populations and lengthy
breeding development timelines. Thus, classical mutation breeding as a tool to generate
variation is of great value but is relatively inefficient.

Genome editing is a targeted and more exact form of mutation breeding. These tools, based on
the use of SDNs, provide researchers the ability to modify sequences at specific locations within
the genome. SDNs generate targeted DSBs that, regardless of the SDN used, are repaired by
the same native DNA break repair mechanisms, the NHEJ or HR pathways described
previously. As described below and depicted in Figure 1, these methods generally result in
many fewer mutations elsewhere in the genome relative to earlier mutagenic techniques. SDNs
continue to be improved to further reduce the likelihood of non-targeted mutations (Zhao and
Wolt, 2017; Hahn and Nekrasov, 2018).

The genome editing toolbox and its utility
Though still a relatively new technology, genome editing using SDNs has demonstrated great promise as a tool for crop improvement. Recent reviews catalog the increasing number of phenotypes and characteristics in crop plants that have been generated using SDNs (Jaganathan et al., 2018; Zhang et al., 2018; Chen et al., 2019). Some of these traits were derived from loss-of-function mutations within coding regions or gene regulatory regions that resulted from NHEJ-mediated repair leading to mutations (insertions/deletions) at the break site. Other traits were derived from specific nucleotide insertion or substitution, and gene/allele replacement edits, which can be generated by repairing SDN-generated DSBs with a DNA donor template (Townsend et al., 2009; Li et al., 2015; Petolino et al., 2016; Filler Hayut et al., 2017; Shi et al., 2017; Yu et al., 2017). Several types of SDNs that have been used in plants rely on either protein/DNA or RNA/DNA binding to provide editing specificity. As shown in Figure 4 (Panel A-B), ZFN and TALEN are chimeric proteins consisting of DNA recognition/binding domains fused to the catalytic domain of the FokI nuclease (Christian et al., 2010; Carroll, 2011; Weeks et al., 2016). TALENs are derived from transcription activator-like (TAL) effectors found in the plant pathogen *Xanthomonas*. TAL effectors bind to specific DNA sequences by using tandem amino acid repeats. Variations in these repeats, known as the repeat variable di-residues (RVDs), allow for specificity to nucleotide sequences. For both ZFNs and TALENs, specificity for a chosen target sequence is improved by expanding the length of DNA targets (e.g., 18 to 32 bp for the paired nuclease) and the spacer (i.e., the distance between paired nucleases). The designed proteins interact as heteromeric pairs, with DSBs occurring only when both protein halves are bound at the same genomic location, further increasing the specificity of these reagents for creating DSBs only at the desired site.

RNA-Guided Endonucleases (RGENs) function as a protein complex that is directed to a genomic target site through a short non-coding guide RNA (gRNA). The nuclease protein and the gRNA are transcribed separately and then form a complex in the cell nucleus. A portion of
the gRNA contains a sequence that is directly complementary to a genomic target site that
guides the induction of DSB formation at that location. In the case of the CRISPR system, Cas9
or another Cas or Cpf nuclease recognizes the gRNA-DNA pair, thus targeting this site for a
DSB (Figure 4, Panel C) (Jinek et al., 2012; Cong et al., 2013; Weeks et al., 2016). The gRNA
of the most widely used Cas9, *Streptococcus pyogenes* Cas9 (SpCas9), binds to a 20 bp
genomic site, termed the protospacer, that is adjacent to a triplet DNA sequence known as the
protospacer adjacent motif (PAM). Unlike complex protein engineering needed for DNA binding
using ZFNs and TALENs, Cas9 can be programmed to new target sites by changing the spacer
sequence of the guide RNA, which simplifies vector design and construction while substantially
reducing the cost. Due to this ease of design and greater efficiency, CRISPR has become the
SDN of choice for genome editing (Globus and Qimron, 2018).

Recently developed CRISPR editing systems do not rely on DSB formation to induce targeted
changes, but instead involve chimeric protein units consisting of a disabled nuclease and an
additional protein domain. In these systems, the disabled nuclease is used to target the
additional protein units to specific genomic locations for different applications. Applications
include targeted base editing with deaminase domains (Komor et al., 2016; Gaudelli et al.,
2017; Adli, 2018), transcriptional knock down using repressors (Qi et al., 2013; Thakore et al.,
2015), targeted DNA methylation and histone modification (Kearns et al., 2015; Stepper et al.,
2016), gene activation using strong transcriptional activators (Cheng et al., 2013), activation of
developmentally silent endogenous loci through chromatin looping (Deng et al., 2014; Morgan et
al., 2017), DNA transposition (Strecker et al., 2019), site-specific recombination (Standage-
Beier et al., 2019), and prime editing (Anzalone et al., 2019). The targeting components of the
nucleases are still intact, allowing for nucleotide changes in a site-directed manner. For
example, cytosine and adenine base editors (converts C to T and A to G, respectively) fuse a
nickase-type Cas9 with a deaminase domain and thus do not induce DSBs. Both are useful
genome editing tools, however the cytosine base editor has been reported to induce off-target editing both in animal cells (Zuo et al., 2019) and in plants ((Jin et al., 2019); discussed in more detail below).

**Designing genome editing protocols and methods to optimize editing specificity and reduce the potential for off-target edits**

Optimizing the specificity of genome editing methods can be achieved in multiple ways, often through altering SDNs delivery into plants. At present, the most common practice involves stably integrating a transgene into the plant genome that expresses the SDN "reagents" (protein or protein/guide RNA complex that is encoded by the transgene) and template DNA molecule for HR edits. The transgene(s) encoding the SDN components are typically integrated at a locus that is unlinked to the gene(s) being edited. A subsequent outcross or selfing generation results in the independent assortment of unlinked genes to segregate the SDN-encoding transgene away from the edited target sequence. Resulting progeny plants can be screened to select non-transgenic plants carrying the targeted edit and lacking the SDN-encoding insert (Gao et al., 2016; Char et al., 2017). The resulting plant variety is often molecularly indistinguishable from varieties with mutations that either occur naturally or are produced by induced mutagenic techniques. It should be noted that chromosomal re-arrangements due to T-DNA insertions have been observed (Clark and Krysan, 2010; Hu et al., 2017), but these can generally be detected by next-generation sequencing-based methods and removed by segregation and selection. Transient delivery methods, which seek to limit the overall exposure of the genome to the SDN and avoid the stable integration of a transgene, have also been developed in plants (Woo et al., 2015; Kelliher et al., 2019). The most common transient delivery method uses a ribonucleoprotein (RNP) complex that consists of a purified RGEN protein bound to a guide RNA (Kim et al., 2017; Andersson et al., 2018). The complex is delivered directly to plant cells,
through bombardment or other means (Ran et al., 2017), so that the active protein can cause a DSB, but the SDN-encoding DNA is not integrated into the genome (Chen et al., 2018).

Off-target edits, and the means to reduce such edits, have received a great deal of attention in the human therapeutic and medical fields (Cho et al., 2014; Shen et al., 2014; Lee et al., 2016; Akcakaya et al., 2018; Lee et al., 2018; Moon et al., 2019) and are also of interest in the plant science community (Feng et al., 2014; Wolt et al., 2016; Tang et al., 2018; Jin et al., 2019; Li et al., 2019; Young et al., 2019). As described in the Introduction, changes observed following genome editing protocols can be split into those that occur at sites without similarity to the targeted region, referred to as untargeted mutations, and “off-target edits” that occur at sites with sequences similar to the targeted region. Numerous studies discussed below have found that the potential for and level of off-target activity in plants can be predicted by comparing the sequence of the target site to the rest of the genome. SDNs bind to specific nucleotide sequences, with the most likely location of potential off-target editing to occur at sites with nucleotide-sequence level similarity to the intended target (Jiang and Doudna, 2017).

Bioinformatics prediction tools have been designed to predict sites that may result in an off-target edit. Programs have been developed for all of the major SDNs [PROGNOS (Fine et al., 2014); TALE-NT (Doyle et al., 2012); CAS-OFF Finder (Bae et al., 2014)], and are routinely included in the design of SDN-based editing protocols (Liu et al., 2017; Listgarten et al., 2018; Minkenberg et al., 2019). Prediction algorithms are useful for identifying potential off-target sites in well-sequenced genomes. Plant species with limited reference genome sequence data may be poorly assessed by these tools. For plant species without a comprehensive genome sequence, it may be difficult to predict potential off-target SDN activity. However, the amount of variation introduced would likely be less than traditional mutagenesis techniques classically utilized in plant breeding strategies (Figure 1). As described in more detail below, it can also be
difficult to assess and discriminate between bona fide off-target SDN activity versus variation attributable to transformation or standing variation.

**Off-target Edit Evaluation and Frequency**

The approaches reported for the identification of off-target edits vary widely and can impact the level of off-target activity observed. As noted above, many studies rely on algorithms to predict the optimal on-target activity and potential off-target edit sites based on similarity with the gRNA sequence. The most common method used to assess whether the off-target activity has occurred at these site(s) is PCR amplification, followed by amplicon sequencing. PCR approaches provide a rapid method for determining if SDN-mediated off-target edits have occurred at a predicted or suspected site. However, it is reliant on accurate predictions of potential off-target sites, which in turn relies on a well-sequenced source or reference genome for thorough analysis. Several alternative, “non-biased” methods to assess and detect DSB and edits have been developed for off-target edit analysis (reviewed in (Zischewski et al., 2017)), although few have been evaluated in plants.

Recently, high-throughput whole genome resequencing (WGRS) has been used to evaluate off-target edits and untargeted mutations in plants. While WGRS is the most comprehensive strategy currently available, it requires that researchers carefully control for differences between the reference genome and the experimental line, heterogeneity in experimental lines, and sequencing errors (Michno and Stupar, 2018). WGRS cannot always distinguish between standing variants and de novo untargeted mutations generated by tissue culture or SDN treatments or attributable to sequencing miscalls. Several recent studies have used WGRS to examine genome-wide mutational profiles in genome-edited rice (Tang et al., 2018), maize (Lee et al., 2019), and cotton (Li et al., 2019). A large-scale study examined the amount of variation that occurs due to various genome editing techniques (i.e., RGENs, CRISPR-Cas9 and Cpf1) (Tang et al., 2018). The study reported WGRS of 69 individual plants that had been regenerated
via tissue culture (with and without *Agrobacterium* treatment), edited with various SDNs reagents, or received no treatment (field grown). WGRS analysis revealed the presence of variation (e.g., SNPs, indels) throughout the genome in all plants, though more so in plants that had gone through the tissue culture process. Relative to plants that went through tissue culture, there was no increase in the variation observed from the plants that had been exposed to SDN treatment. The exception was in plants edited with a gRNA purposefully designed to be more likely to elicit off-target edits (Tang et al., 2018). The detected off-target edits in these plants were predicted by the program Cas-OFFinder (Bae et al., 2014), confirming that intentionally inadequate guide design led to the observed off-target activity. A recent study in maize determined that regardless of the SDN delivery method (*Agrobacterium*, particle bombardment, or RNP), proper design of gRNAs leads to undetectable levels of off-target edits (Young et al., 2019). Studies examining rice (Tang et al., 2018), cotton (Li et al., 2019), and maize (Lee et al., 2019; Young et al., 2019) report that nearly all of the observed variation from RGEN experiments is attributable to tissue culture or natural variation, and not the delivered nucleases. Tang et al. (2018) found that plants regenerated via tissue culture contained ~100-150 SNVs and ~40-80 indels. The number of *de novo* variants attributable to tissue-culture and SDN delivery is small relative to standing variation that differentiates individual lines (Table 2).

A recent study has applied the WGRS approach to evaluate the off-target edit frequency of base editors in rice (Jin et al., 2019). For potential off-target edits, the authors found that SNVs were not significantly different between control plants (exposed to tissue culture with no editing machinery) and adenine base edited plants. The rate of off-target edits was significantly higher only in plants generated with cytosine (C to T) base editors, irrespective of whether a gRNA was supplied. The percentage of changes that were C to T transitions was higher in plants containing cytosine base editing constructs, suggesting these reagents were inducing background variation. A larger number of off-target mutations suggests that the source of off-
target variation was not the SDN, but the enzymatic domain of the cytosine base editing complex. Previous studies had found that the cytosine deaminase domain, and the additional uracil glycosylase inhibitor domain used in cytosine base editors, generate higher background levels of C to T changes (Radany et al., 2000; Harris et al., 2002). The studies suggest that increased optimization of the cytosine base editor would be necessary to minimize off-target activity. However, it is worth noting that when compared to the frequencies of mutagenesis reported in rice for somaclonal variation (Miyao et al., 2011, Tang et al, 2018) or other forms of induced mutagenesis (see Tables 3 and 4), mutation rates with the cytosine base editor are comparable to methods used in conventional and mutation breeding. This study suggests that, as new protein motifs and editing strategies are developed, the intended activity and potential for off-target activity should be evaluated.

WGRS studies (Tang et al., 2018; Jin et al., 2019; Li et al., 2019) show that the amount of induced background variation as a result of introducing editing enzymes into plants is generally low, especially for editing applications with SDNs. The identification of high background mutation rates for C-to-T base editor enzymes (Jin et al., 2019) suggests the need for careful screening of the effects of new approaches. Taken together, these results suggest that WGRS is best applied to new approaches and may not be necessary for routine application of SDNs, as unintended activity will generally be predictable by computational algorithms that allow for the optimal design of gRNAs (Tang et al., 2018; Lee et al., 2019; Young et al., 2019). Indeed, conventional mutagenesis methods have been found to be safe using existing agronomic evaluations, generally in the absence of WGRS evaluation.

Conclusions

Whole-genome examinations of variability in plants and improvements in understanding of gene function are occurring concurrently with advances in genome editing technology. Targeted genome editing utilizing SDN techniques are the most recent option for the introduction of
precise genetic variation for crop improvement (Figure 5). SDNs are designed to introduce
mutations with a high degree of sequence specificity, resulting in fewer unintended genomic
changes than earlier mutagenesis techniques. As the understanding of nuclease activity has
increased, SDN protocols and associated prediction algorithms to more specifically target
intended sequences have improved and been aided by the increasing number of well-annotated
and sequenced genomes. NGS has facilitated a more thorough understanding of plant genomes
and the genetic changes that accompany the development of new plant varieties. These
methods allow more accurate estimations of de novo genomic variation. Each plant generation
has new genetic variants that are the result of a range of naturally-occurring processes. The
potential for, and possible effects of, SDN-mediated off-target changes must be put in the
context of naturally occurring standing variation in crops and mutations induced or introduced
during plant breeding practices (Table 5).

[INSERT TABLE 5 HERE]

The history of safe consumption of foods from plants has been built on the fact that many
mutations, regardless of origin, have no phenotypic effect such that breeders cannot remove
these “neutral” genetic changes from plant populations. By comparison, plants displaying off-
type phenotypes due to unintended mutations are eliminated or ameliorated through well-
established, multigenerational breeding, selection, and commercial variety development
practices (Figure 5). Varieties developed through genome editing will be subjected to the same
screening and selection practices as is used for improved varieties developed using other
sources of genetic variation (Figure 5). While off-target edits are deemed to be unacceptable for
therapeutic applications of genome editing, off-target edits in the context of crops are smaller in
magnitude than that of current crop improvement methods, such as conventional or mutation
breeding (see Outstanding Questions). The current generation of genome editing technologies
has facilitated the efficient generation of desirable genomic variation and new plant varieties that
would have been more challenging to achieve through other breeding or molecular approaches. Genome editing technology continues to be refined, with improved editing specificity, developments in the delivery of editing enzymes, and ever-increasing genomic characterization.

Acknowledgments

The authors wish to acknowledge Wayne A. Parrott for comments on an earlier version of the manuscript and Emily E. Vonderharr for copy editing of the text. We also thank Allison Devitre and Emily Turnbough for their help with figure art. Funding provided by USDA / National Institute of Food and Agriculture to Robert M. Stupar and Peter L. Morrell (2015-33522-24096 and 2019-33522-30200).

Box 1: The importance of double-strand break repairs

Box 1 cited articles: Pacher and Puchta, 2017; Puchta, 2005; Schuermann et al., 2005; Waterworth et al., 2010; Waterworth et al., 2011; Waterworth et al., 2016; Friedberg et al., 2006; Voytas, 2013; Pacher et al., 2017; Brunet and Jasin, 2018; Čermák et al., 2017; Knoll et al., 2014; Gaut et al., 2007; Hajjar and Hodgkin, 2007
Table 1. Common types of genomic variation in the breeding populations of field crops and vegetables.
<table>
<thead>
<tr>
<th>Variant type</th>
<th>Variation</th>
<th>Description/comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide polymorphisms</td>
<td>SNP</td>
<td>A variant at a single position where one base replaces another base. SNPs arise from errors during DNA replication or repair. Also referred to as single nucleotide variant (SNV) or single base substitution (SBS).</td>
</tr>
<tr>
<td></td>
<td>Single Nucleotide Polymorphism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MNP</td>
<td>Changes at 2 or more adjacent nucleotides MNPs are relatively rare.</td>
</tr>
<tr>
<td></td>
<td>Multi-nucleotide polymorphisms</td>
<td></td>
</tr>
<tr>
<td>Structural variants</td>
<td>Indel</td>
<td>Sites where DNA is lost (deletion) or gained (insertion) typically describing relatively small (1-20 bp but up to 1000 bp) changes. Usually classified as sequence differences smaller than PAV.</td>
</tr>
<tr>
<td></td>
<td>Insertion/deletions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSR</td>
<td>SSRs are repeats of 2 to 6 nucleotides that vary in the number of repeat motifs present. Also referred to as short tandem repeats (STRs) or microsatellites. Can be considered a type of CNV</td>
</tr>
<tr>
<td></td>
<td>Simple sequence repeats</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CNV</td>
<td>Sequences with variable copy number between individuals. The term CNV generally applies genic regions. Can be due to duplications, deletions or insertions. Simple sequence repeats (SSR) can be considered as very small CNVs. Typically,</td>
</tr>
<tr>
<td><strong>CNVs</strong></td>
<td>Refer to larger sequence repeats (&gt;1kb for example) and may encompass one or more genes or parts of genes.</td>
<td></td>
</tr>
<tr>
<td><strong>PAV</strong></td>
<td>Presence/Absence Variation. Large sequence block (&gt;1 kb) present in one line/individual but completely missing in another line. Like CNV, PAV typically refers to a region containing genes and can be thought of as the extreme example of CNV.</td>
<td></td>
</tr>
<tr>
<td><strong>TE</strong></td>
<td>Transposable elements. Mobile genetic elements. The number and location of each element frequently differ from the reference genome.</td>
<td></td>
</tr>
<tr>
<td><strong>Large Structural Variation (SV) and ploidy (whole and partial genome duplication)</strong></td>
<td>Includes DNA rearrangements such as chromosomal inversions, translocations, partial genome duplications, duplications, etc.</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Estimated genomic diversity in the breeding populations of select field and vegetable crops. Not all forms of genomic variation were assessed for all species listed. This is denoted by ND (Not Determined). In addition, because of the short-read sequencing technologies used, the number of PAV/CNV structural variants, representing large deletions, inversions and duplications, is likely underestimated.
<table>
<thead>
<tr>
<th>Plant Species</th>
<th>Evaluated sample size and description</th>
<th>Estimate d Genome Sizea (Mbp)</th>
<th>Estimate d SNPs present (million)</th>
<th>Average Pairwise Diversityb</th>
<th>Estimate d Indels (million)</th>
<th>Structural Variants (CNV &amp; PAV)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bean (Phaseolus vulgaris)</td>
<td>37</td>
<td>521</td>
<td>45.0</td>
<td>~5.0 x 10^3</td>
<td>2.1</td>
<td>18.5 k</td>
<td>(Schmutz et al., 2014; Lobaton et al., 2018)</td>
</tr>
<tr>
<td>Wild Cabbage (Brassica oleracea)</td>
<td>10 (1 wild species)</td>
<td>489</td>
<td>4.8</td>
<td>ND</td>
<td>ND</td>
<td>11.5 k</td>
<td>(Golicz et al., 2016)</td>
</tr>
<tr>
<td>Chickpea (Cicer arietinum)</td>
<td>35</td>
<td>738</td>
<td>2.0</td>
<td>ND</td>
<td>0.3</td>
<td>9.9 k</td>
<td>(Thudi et al., 2016)</td>
</tr>
<tr>
<td>Cucumber (Cucumis sativus)</td>
<td>115</td>
<td>194</td>
<td>3.3</td>
<td>ND</td>
<td>0.03</td>
<td>0.8 k</td>
<td>(Zhang et al., 2015)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Qi et al., 2013)</td>
</tr>
<tr>
<td>Crop</td>
<td>Accessions</td>
<td>Lines</td>
<td>Genomesize</td>
<td>Pairwise diversity</td>
<td>Pairwise Diversity Ref.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------------</td>
<td>-------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maize (Zea mays)</td>
<td>2815</td>
<td>13</td>
<td>2,400</td>
<td>~6.0 x 10^3</td>
<td>Gore et al., 2009; Romay et al., 2013; Wang et al., 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rice (Oryza sativa ssp. Japonica)</td>
<td>3010</td>
<td>375</td>
<td>29.0</td>
<td>~1.0 x 10^3</td>
<td>Liu et al., 2017; Wang et al., 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soybean (Glycine max)</td>
<td>106 (Landraces, US cultivars and wild)</td>
<td>978</td>
<td>10.4</td>
<td>1.8 x 10^3</td>
<td>Valliyodan et al., 2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tomato (Solanum lycopersicum)</td>
<td>69 (incl 12 wild species)</td>
<td>950</td>
<td>17.0</td>
<td>~2 x 10^-4</td>
<td>Causse et al., 2013; Sauvage et al., 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pepper (Capsicum annuum)</td>
<td>20 (3 wild species)</td>
<td>3,500</td>
<td>9.8</td>
<td>~2.6 x 10^-3</td>
<td>Aguilar-Meléndez et al., 2009; Qin et al., 2014</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Notes:**

- Estimated genome sizes from references or [https://plants.ensemble.org](https://plants.ensemble.org) April 2019.
- Pairwise diversity (also referred to as nucleotide diversity) is the average number of nucleotide differences per site between any two DNA sequences in all possible pairs in the sample population.
(Tajima, 1983). It is a measure that allows for diversity/variation to be compared across organisms that vary widely in genome size.
Table 3: Examples of induced mutation types and frequency in food crops and plant model species evaluated via whole genome resequencing (WGRS) analyses

<table>
<thead>
<tr>
<th>Plant</th>
<th>Mutagen</th>
<th>No. Plants analyzed</th>
<th>Mean mutations per plant $^b$ (per diploid genome)</th>
<th>Total mutation frequency $^c$ (mutations per nucleotide site per experiment)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice ($Oryza sativa$)</td>
<td>FN</td>
<td>1,504</td>
<td>29 SNPs, 26 Indels, 3 translocations, 3 inversion/duplications, 61 total</td>
<td>[1.6 x 10$^{-7}$]</td>
<td>(Li et al., 2017)</td>
</tr>
<tr>
<td>[375 Mb]</td>
<td>GR</td>
<td>6</td>
<td>2,419 SNPs, 452 Indels, 69 SV, 383 CNV, 3,323 total</td>
<td>[8.86 x 10$^{-6}$]</td>
<td>(Li et al., 2016)</td>
</tr>
<tr>
<td>Tissue culture</td>
<td>2</td>
<td>114 SNPs, 36 Indels, 150 total</td>
<td>1.86 x 10$^{-7}$</td>
<td></td>
<td>(Tang et al., 2018)</td>
</tr>
<tr>
<td>Plant</td>
<td>Method</td>
<td>Mutagen</td>
<td>sSNPs</td>
<td>Indels</td>
<td>Total</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Sorghum (Sorghum bicolor) [708 Mb]</td>
<td>EMS</td>
<td>256</td>
<td>7,660 SNPs</td>
<td></td>
<td>[1.1 x 10⁻⁵]</td>
</tr>
<tr>
<td>Soybean (Glycine max) [978 Mb]</td>
<td>EMS</td>
<td>12</td>
<td>12,796 SNPs</td>
<td>12,854 SNPs (Md)</td>
<td>[1.3 x 10⁻⁵]</td>
</tr>
<tr>
<td>Tomato (Solanum lycopersicum) [950 Mb]</td>
<td>EMS</td>
<td>4</td>
<td>1348 SNPs</td>
<td>17 Indels</td>
<td>[1.4 x 10⁶]</td>
</tr>
<tr>
<td></td>
<td>GR</td>
<td>3</td>
<td>137 SNPs</td>
<td>40 Indels</td>
<td>[1.9 x 10⁻⁷]</td>
</tr>
<tr>
<td>Lotus japonica [472 Mb]</td>
<td>EMS</td>
<td>2</td>
<td>2,201 SNPs</td>
<td></td>
<td>[4.7 x 10⁻⁶]</td>
</tr>
<tr>
<td>Arabidopsis thaliana [135 Mb]</td>
<td>GR</td>
<td>6</td>
<td>11 SNPs (8.5 Md)</td>
<td>7 Indels (6.5 Md)</td>
<td>3.6 x 10⁻⁷</td>
</tr>
<tr>
<td></td>
<td>Tissue culture</td>
<td>5</td>
<td>26 SNP (18 Md)</td>
<td>4 Indels (5 Md)</td>
<td>4.2 x 10⁻⁷ - 24.2 x 10⁻⁷</td>
</tr>
</tbody>
</table>
Abbreviations: EMS: Ethyl Methanesulphonate; FN: Fast Neutron; Na$_3$-MNU: sodium azide-methyl-nitro urea; GR: Gamma Ray; Indel: Insertion/deletion; SNP: single nucleotide polymorphism; SV: structural variant; CNV: copy number variation; WGRS: whole genome resequencing; kb: kilobase; Mb: Megabase.

a Assembled genome size based on EnsemblPlants database (http://plants.ensembl.org) unless otherwise noted.

b Mean number of mutations per analyzed plant are either values as published or calculated based on total mutations observed across all plants divided by the number of plants analyzed. All plants listed are diploids and studies confirmed full (>90%) genomic coverage unless otherwise noted in footnote c. For some studies, median values (Md) for number of mutations per plant could also be calculated.

c Total mutation frequency represents total mutations per nucleotide site per experiment as published or numbers in brackets were calculated by dividing total mutations per plant by the assembled genome size. As noted, total mutation frequency was not limited to a single type of mutation.

d Average genome sequence coverage was <90% for tomato (86% coverage; 815 Mb) and *Lotus japonica* (67% coverage; 315 Mb); partial genomic coverage values were used in normalizing number of mutations per diploid genome and in calculating total mutation frequency.
<table>
<thead>
<tr>
<th>Plant</th>
<th>Mutagen</th>
<th>No. Plants analyzed</th>
<th>Reported mutation density [calculated mutation frequency]</th>
<th>Number of genes/genomic region analyzed</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice</td>
<td>NaN-MNU</td>
<td>768</td>
<td>1 SNP/265 kb [3.8×10^-6]</td>
<td>10 genes [7.9 Mb total]</td>
<td>(Till et al., 2007)</td>
</tr>
<tr>
<td>(Oryza sativa)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[375 Mb]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maize</td>
<td>EMS</td>
<td>1,086</td>
<td>1 SNP/774 kb [1.3 x 10^-6]</td>
<td>Whole exome resequence [139 Mb/plant]</td>
<td>(Lu et al., 2018)</td>
</tr>
<tr>
<td>(Zea mays)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2,400 Mb]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wheat</td>
<td>EMS</td>
<td>1,535 tetraploid 1,200 hexaploid</td>
<td>23 SNPs/Mb (4n) [2.3 x 10^-5] 33 SNPs/Mb (6n) [3.3 x 10^-5]</td>
<td>Whole exome resequence [119 Mb/plant (4n) 162 Mb/plant (6n)]</td>
<td>(Krasileva et al., 2017)</td>
</tr>
<tr>
<td>(Triticum aestivum)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[14,540 Mb]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barley</td>
<td>NaN3</td>
<td>3,148</td>
<td>1 SNP/374 kb [2.7 x 10^-6]</td>
<td>4 genes [8.2 Mb total]</td>
<td>(Talamè et al., 2008)</td>
</tr>
<tr>
<td>(Hordeum)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Examples of induced mutation types and frequency in food crops using TILLING or exome resequencing genomic analyses.
<table>
<thead>
<tr>
<th>Species</th>
<th>Genotype</th>
<th>Mutagen</th>
<th>Mutations</th>
<th>Mutation Density</th>
<th>Genes</th>
<th>Genomic Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cotton</td>
<td>(Gossypium</td>
<td>EMS</td>
<td>1 SNP/153 kb</td>
<td>8,000 M_2 lines (2 varieties)</td>
<td>1 SNP/326 kb</td>
<td>8 genes</td>
</tr>
<tr>
<td></td>
<td>hirsutum)</td>
<td></td>
<td>6.5 x 10^{-6}</td>
<td>6.5 x 10^{-6}</td>
<td>3.1 x 10^{-6}</td>
<td>58.1 Mb/variety</td>
</tr>
<tr>
<td></td>
<td>[2,100 Mb]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tomato</td>
<td>(Solanum</td>
<td>EMS</td>
<td>1 SNP/322 kb</td>
<td>1,352 M_3 lines</td>
<td>1 SNP/322 kb</td>
<td>7 genes</td>
</tr>
<tr>
<td></td>
<td>lycopersicum)</td>
<td></td>
<td>3.1 x 10^{-6}</td>
<td>3.1 x 10^{-6}</td>
<td>3.1 x 10^{-6}</td>
<td>10.9 Mb total</td>
</tr>
<tr>
<td></td>
<td>[950 Mb]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** EMS: Ethyl Methanesulphonate; NaN₃-MNU: sodium azide-methyl-nitro urea; TILLING: Targeting Induced Local Lesions In Genomes; SNP: single nucleotide polymorphism; kb: kilobase; Mb: Megabase.

*a* Assembled genome size based on EnsemblPlants database (http://plants.ensembl.org).

*b* Mutation density (mutations/kb or Mb) is as reported and was determined by dividing the total genomic region evaluated (kb) by the number of nucleotide changes observed. Calculated mutation frequency is the arithmetic calculation of the mutations per nucleotide site per experiment.

*c* Number of gene or genomic regions analyzed in mutagenized plants are as reported. The total DNA length analyzed across all plants and genes (for TILLING) or per plant (for exome sequencing) is also included.
Table 5. Examples of genetic variation introduced by different breeding methods, as reported in recent studies in rice.

<table>
<thead>
<tr>
<th>Variation Type</th>
<th>Breeding Method</th>
<th>Re-mixing standing variants</th>
<th>de novo variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cross pollination&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2 × 10⁶</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self pollination&lt;sup&gt;b&lt;/sup&gt;</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Classical mutagenesis (Fast neutron)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>43</td>
<td>114</td>
</tr>
<tr>
<td>Number of single nucleotide variants introduced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tissue culture&lt;sup&gt;b&lt;/sup&gt;</td>
<td>&gt; 5 × 10³</td>
<td>18</td>
</tr>
<tr>
<td>Number of insertions/deletions introduced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genome editing&lt;sup&gt;b&lt;/sup&gt;</td>
<td>&gt; 2 × 10⁶</td>
<td>41</td>
</tr>
<tr>
<td>Total number of mutations introduced</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Wang et al., 2018
<sup>b</sup> Tang et al., 2018
<sup>c</sup> Li et al., 2017
Figure Legends

Figure 1: Comparison of the average number of SNPs and indels per genome (individual) introduced into tomato by different breeding strategies. The data represents approximate number of SNPs between the genome sequence of *S. lycopersicum* (Hetz 1706 reference genome) and other cultivars or wild relatives which have been used in breeding of modern tomato cultivars. After selfing (X Self), each individual plant in the next generation will have (on average) approximately 6 random single nucleotide polymorphisms (SNPs) (assuming the *de novo* rate of spontaneous heritable SNP formation in tomato is similar to that of lab-grown Arabidopsis (Ossowski et al., 2010). Most modern elite tomato lines commercially grown typically have four or more disease resistance genes that have been introgressed by crossing with wild tomato species such as (*Solanum pennellii* or *S. pimpinellifolium*) (Foolad, 2007). These initial elite x wild species hybridization events introduced millions to tens of millions of SNPs in addition to indels, copy number (CNV) and presence/absence variation (PAV) (Aflitos et al., 2014). Crosses with more closely related domestic tomato lines or landraces (i.e., cultivar [cv.] San Marzano) will introduce (on average) hundreds of thousands of SNPs/indels (Ercolano et al., 2014). Creating random variation by treating seeds with the chemical mutagen ethyl methanesulphonate (EMS) typically introduces thousands of SNPs per individual (Minoia et al., 2010). Treating cells with a well-designed gene editing reagent (including a guide RNA homologous to target sequence and adjacent protospacer adjacent motif [PAM]) can create a single SNP or indel at a precise, pre-determined location. Crossing with wild or closely related species can also introduce additional indels, CNV and PAV into the genome, which is not considered in this figure.

Figure 2. Number of officially released mutant varieties (top 20 countries) showing direct release of improved varieties (orange bars) and mutants used as breeding material (blue bars).

Figure 3: Overview of sources genomic variation in crop plants. The simplified gene models depict observed variation among individual crop species. The comparison of inherent/standing (genotype X and Y) and induced variation (chemical and irradiation) with the reference genome shows a range of intended and unintended mutations, including single nucleotide polymorphisms (SNPs), small indels, transposon insertions/movement, or a large segmental deletion. However, the intended edit induced via a site-directed nuclease (SDN) occurs at the target site. Note: not drawn to the scale; intergenic variation not shown.

Figure 4. Most widely used SDNs. A. TALENs are composed of a DNA binding domain and the nuclease Fok1; the DNA binding domain contains an array of nearly identical protein subdomains each varying at two specific amino acids, known as the repeat variable di-residues (RVDs); specific RVDs recognize unique bases on the target DNA molecule. B. ZFNs are also composed of DNA binding domains and the Fok1 nuclease; each ZFN is composed of three zinc-finger domains that are custom designed to recognize a triplet of DNA bases on the target sequence. For TALENs and ZFNs, two subunits are needed per target region, each binding closely spaced DNA sequences. C. The Cas9 nuclease binds to the target sequence via an RNA molecule, known as the sgRNA; the 5' region of the sgRNA, the proto-spacer, is typically 20 nucleotides long and is complementary to the target DNA; a PAM sequence is also required for recognition (shown here in bold and underlined).

Figure 5: An overview of conventional and modern mutation breeding approaches for crop improvement. The figure illustrates the systematic approach for introducing intended and unintended genomic variation in crops for trait discovery, germplasm development and commercial release of new crop varieties. After the introduction of genomic variation, the
mutant populations go through a series of selection and testing to remove undesirable mutations and phenotypic variations. The best performing line is selected for subsequent commercial release. This breeding process is common to all methods used for crop product development (Kaiser et al., 2020).

Literature Cited


Downloaded on May 3, 2021. - Published by https://plantphysiol.org
Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.


Clark KA, Krysan PJ (2010) Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines. The Plant Journal 64: 990-1001


Food and Drug Administration (1992) Food for human consumption and animal drugs, feeds, and related products: foods derived from new plant varieties; policy statement, 22984. In FDA Federal Register, Department of Health and Human Services, Vol 57, p 22984


Gabur I, Chawla HS, Snowdon RJ, Parkin IAP (2019) Connecting genome structural variation with complex traits in crop plants. Theoretical and Applied Genetics 132: 733-750


Li S, Zheng Y-c, Cui H-r, Fu H-w, Shu Q-y, Huang J-z (2016) Frequency and type of inheritable mutations induced by γ rays in rice as revealed by whole genome sequencing. Journal of Zhejiang University-SCIENCE B 17: 905-915


Mba C (2013) Induced mutations unleash the potentials of plant genetic resources for food and agriculture. Agronomy 3: 200-231


Neelakandan AK, Wang K (2012) Recent progress in the understanding of tissue culture-induced genome level changes in plants and potential applications. Plant cell reports 31: 597-620


Shirasawa K, Hirakawa H, Nunome T, Tabata S, Isobe S (2016) Genome - wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato. Plant biotechnology journal 14: 51-60

Shu Q-Y, Forster BP, Nakagawa H (2012) Plant mutation breeding and biotechnology. CABI


Zohary D (1999) Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. Genetic Resources and Crop Evolution 46: 133-142

ADVANCES

- The importance and potential impact of “off-target” genetic changes during genome editing are subjects of ongoing debate.
- Recent studies have demonstrated that with appropriate design of genome editing reagents, off-target edits in plants are nearly undetectable.
- Off-target edits with base editing proteins can be minimized through optimization of the base-editing enzyme domain.
OUTSTANDING QUESTIONS

- What additional technical and non-technical roadblocks will need to be overcome to facilitate further research and commercial use of plant genome editing tools?
- As new genome editing tools and more facile reagent delivery methods are developed, what is an appropriate level of evaluation for potential off-target activity in plants?
- How can the plant science community effectively convey to society at large the relevance of potential off-target genomic changes when placed in the context of highly variable and dynamic plant genomes?
Double-strand breaks (DSB) in DNA occur as a part of routine biological functions in all organisms (Pacher and Puchta, 2017). DNA DSB repair is crucial to many ordinary cellular functions such as DNA replication, homologous recombination, and in response to transposon replication and movement (Puchta, 2005; Schuermann et al., 2005; Waterworth et al., 2011). DSBs also occur during the seed dormancy process and subsequent seed germination is dependent on the effective repair of DSBs (Waterworth et al., 2010; Waterworth et al., 2011; Waterworth et al., 2016). Mutagenesis (including radiation-based techniques such as fast-neutron mutagenesis) and genome editing with SDNs generate DSBs (Friedberg et al., 2006; Voytas, 2013) which are repaired by endogenous repair mechanisms.

Regardless of their origin, DSB in a somatic plant cell typically initiate break repair by one of two primary mechanisms; NHEJ or homology-dependent repair (HDR) (Pacher and Puchta, 2017). NHEJ is the predominant repair mechanism in plant somatic cells, accounting for nearly all repair events (Puchta, 2005). Repair typically results in restoration of the original sequence, but small (typically 1-20 base pair (bp)) insertions or deletions (indels) also occur (Waterworth et al., 2011; Brunet and Jasin, 2018). It has been recently shown in plants that only 15% of DSBs generated by CRISPR-Cas9 are repaired imperfectly, leading to mutations (Čermák et al., 2017). HDR in plant somatic cells is rare (Knoll et al., 2014) but is of critical importance during meiosis leading to recombination between intact homologous regions. Sequences used as templates in HDR can be allelic (from homologous chromosomes), ectopic (from non-homologous chromosomes), or intrachromosomal (from the same chromosome) (Puchta, 2005). Homologous and non-homologous recombination following DSB repair contributes to genomic rearrangements and has allowed the introgression of many beneficial traits from crop wild relatives into modern cultivars by conventional cross-breeding (Gaut et al., 2007; Hajjar and Hodgkin, 2007).
Approximate Number of SNPs

- ~6
- X Self
- X S. pennelli
- X S. pimpinellifolium
- X cv. San Marzano
- 170,000
- 4,000,000
- 7,000,000
- ~3,000
- 1

Genomic DNA

Cas9

Target Sequence

PAM

Guide RNA

Crossing Plants

Random, Unknown and Uncharacterized Genetic Changes

Treat in Lab

Precise Change

EMS Genome Editing

https://plantphysiol.org

Downloaded on May 3, 2021. - Published by https://plantphysiol.org

Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.
A. TAL effector nucleases (TALENs)

DNA-binding domain

Repeat Variable Di-residues

Target DNA

B. Zinc-finger nucleases (ZFN)

5' -CGCCACTAGCGTATGATCGTACTGTGCAGTTGTGGTTTGTCTACCGTATGATCGTACTGTGCAGTTGTGGTTTGTCTACCGTA

3' -CGCGTGATCGCATACTAGCATCAGACGTCAACACCAAACAGATGGCAT

C. Clustered Regularly-Interspaced Short Palindromic Repeats (CRISPR) Associated Protein 9 (Cas9)

5' -ACGTAGCGT

3' -TGCATCGCA

5' -ACGTAGCGT CATCGCCACTAGCGTATGATCGTACTGTGCAGTTGTGGTTTGTCTACCGTA

3' -TGCATCGCA GTAGCGTGATCGCATACTAGCATCAGACGTCAACACCAAACAGATGGCAT

Proto-spacer Adjacent Motif (PAM)

sgRNA proto-spacer sequence
Introduction of genetic variation for crop improvement

Conventional breeding  |  Chemical/radiation mutagenesis  |  Genome editing

Nucleotide polymorphisms
- SNPs
- Indels
- CNV/PAV

Gene content polymorphisms
- Deletions
- Duplications

Chromosomal changes
- Translocations
- Inversions

Selection & testing
- Removal of off-types
- Phenotypic characteristics
- Performance attributes
- Processing features
- End-user attributes

Genotyping & Inbreeding  ↔  Small-scale field trials

Large-scale field trials

Crosses & multi-year, multi-location field trials

Commercial release

Downloaded on May 3, 2021 - Published by https://plantphysiol.org
Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.


Clark KA, Krysan PJ (2010) Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines. The Plant Journal 64: 990-1001


Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.

Food and Drug Administration (1992) Food for human consumption and animal drugs, feeds, and related products: foods derived from new plant varieties; policy statement, 22984. In FDA Federal Register, Department of Health and Human Services, Vol 57, p 22984


Gabur I, Chawla HS, Snowdon RJ, Parkin IAP (2019) Connecting genome structural variation with complex traits in crop plants. Theoretical and Applied Genetics 132: 733-750


Li S, Zheng Y-c, Cui H-r, Fu H-w, Shu Q-y, Huang J-z (2016) Frequency and type of inheritable mutations induced by γ rays in rice as revealed by whole genome sequencing. Journal of Zhejiang University-SCIENCE B 17: 905-915


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only
Title Only
Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mba C (2013) Induced mutations unleash the potentials of plant genetic resources for food and agriculture. Agronomy 3: 200-231

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title


Neelakandan AK, Wang K (2012) Recent progress in the understanding of tissue culture-induced genome level changes in plants and potential applications. Plant cell reports 31: 597-620


Shu Q-Y, Forster BP, Nakagawa H (2012) Plant mutation breeding and biotechnology. CABI


Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zohary D (1999) Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. Genetic Resources and Crop Evolution 46: 133-142
Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title

Pubmed: Author and Title
Google Scholar: Author Only Title Only Author and Title