Sodium azide mutagenesis induces a unique pattern of mutations

Chaochih Liu[1], Giulia Frascarelli[1], Adrian O. Stec[1], Shane Heinen[1], Li Lei[2], Skylar R. Wyant[3], Erik Legg[4], Monika Spiller[5], Gary J. Muehlbauer[1], Kevin P. Smith[1], Justin C. Fay[6], Peter L. Morrell[1]

[1] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108
[2] U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[3] Department of Ecology & Evolutionary Biology, University of California, Irvine, CA 92697
[4] Syngenta Crop Protection Inc., Greensboro, NC 27409
[5] KWS LOCHOW GmbH, Wetze 3, 37154 Northeim, Germany
[6] Department of Biology, University of Rochester, Rochester, NY 14627

Abstract

The nature and effect of mutations are of fundamental importance to the evolutionary process. The generation of mutations with mutagens has also played important roles in genetics. Applications of mutagens include dissecting the genetic basis of trait variation, inducing desirable traits in crops, and understanding the nature of genetic load. Previous studies of sodium azide-induced mutations have reported single nucleotide variants (SNVs) found in individual genes. To characterize the nature of mutations induced by sodium azide, we analyze whole-genome sequencing (WGS) of 11 barley lines derived from sodium azide mutagenesis, where all lines were selected for diminution of plant fitness owing to induced mutations. We contrast observed mutagen-induced variants with those found in standing variation in WGS of 13 barley landraces. Here, we report indels that are two orders of magnitude more abundant than expected based on nominal mutation rates. We found induced SNVs are very specific, with C→T changes occurring in a context followed by another C on the same strand (or the reverse complement). The codons most affected by the mutagen include the sodium azide-specific CC motif (or the reverse complement), resulting in a handful of amino acid changes and relatively few stop codons. The specific nature of induced mutations suggests that mutagens could be chosen based on experimental goals. Sodium azide would not be ideal for gene knockouts but will create many missense mutations with more subtle effects on protein function.

Introduction

Mutagens can quickly generate novel, heritable genetic variation for identifying gene function when naturally occurring variation will not suffice. Sodium azide (NaN_3), ethyl methanesulfonate (EMS), and fast neutron (FN) radiation have been the most commonly used mutagens in plants. Sodium azide and EMS are chemical mutagens expected to cause point mutations (Henry et al., 2014). FN is known for generating many large structural variants and small indels (especially deletions) that generate frameshifts that disrupt gene function (Bolon et al., 2014, Wyant et al., 2022). Mutagenizing agents have been utilized in many species to create knockouts or knockdowns of individual genes to understand gene function (Schneeberger, 2014). Despite decades of active use (Kleinhofs et al., 1978, Owais and Kleinhofs, 1988) and growing interest in the applications of mutagenesis (FAO/IAEA, 2018, Knudsen et al., 2022), the characterization of the nature of the mutations generated has been relatively limited (Zhu et al., 2017). More generally, the number of studies examining the effects of mutagens at the nucleotide sequence level is limited (Belfield et al., 2012, Henry et al., 2014, Li et al., 2017, Wyant et al., 2022).

The nature and context in which mutations occur are essential determinants of their likely functional impact (Morton et al., 2006, Zhu et al., 2020). Recently reported efforts to employ sodium azide mutagenesis on a massive scale (500,000 mutated lines) clarify the need for an improved understanding of the number and types of mutations likely to be generated (Knudsen et al., 2022).

We can characterize induced mutations by comparing newly generated (*de novo)* mutations to variants occurring naturally in untreated lines. However, there are essential considerations to minimize false positives when distinguishing between induced mutations and existing variants. One challenge with identifying induced *de novo* changes is that the mutations are always a mixture of induced *de novo* and spontaneously occurring mutations (Zhu et al., 2020). The experimental design necessary to study induced mutations in plant populations requires multiple generations of seed bulking before the mutagenesis treatment and several generations in which new mutations are made homozygous in inbred lines; thus, multiple generations in which new spontaneous variants can arise (Wyant et al., 2022). Mutagenized plants are identified as the M_0. At least two generations of selfing are required before it is possible to perform phenotypic screening and ensure all mutations identified are meiotically heritable is the M_2 (Nations, 2018). Mutagen dosage also affects the nature of observable mutations, with an effective dosage of mutagens is defined by the LD50, or lethal dose for 50% of the treated sample, which results in the death of half the individuals. Of course, this creates significant attrition due to lethal mutations, resulting in some mutations or combinations that cannot be observed directly. Observable mutations are necessarily less damaging, allowing the plant to survive and reproduce.

New mutations, whether induced or naturally occurring, are more likely to be harmful than segregating variants subject to generations of natural (purifying) selection. More potentially harmful changes are observed among newer mutations in very deep resequencing panels (Schaibley et al., 2013) and among induced mutations (Wyant et al., 2022). Newer mutations tend to include more changes of large effect, including nonsynonymous variants, changes in start and stop codons, intron splicing variants, and frameshifts (Eyre-Walker and Keightley, 2007, Kono et al., 2016). Comparative approaches using phylogenetic constraint have been used for decades to predict which mutations are more likely to harm the organism (Sunyaev et al., 2001, Ng and Henikoff, 2003). More recent studies of induced mutations in *Arabidopsis thaliana* suggest that these approaches accurately identify phenotype-changing variants and are more likely to have a major effect on organismal fitness (Kono et al., 2018). Thus, predicting the harmfulness of mutations has the potential to identify the nature of variants most likely to impact organismal fitness or yield in crops (Lu et al., 2006, Morrell et al., 2012, Moyers et al., 2018).

In barley, sodium azide primarily generates cytosine-to-thymine changes (Olsen et al., 1993). More recent studies have determined that mutagens create unique suites of mutations and that these mutations typically occur in a specific, immediate nucleotide context (Henry et al., 2014, Zhu et al., 2020). This observation is important because the nature of mutations, particularly when they occur predominantly in the presence of flanking motifs, can determine their relative effect. For example, C→ T transitions, particularly in specific local sequence contexts, may limit the potential for mutagens to generate premature stop codons, even when the reverse complement of a mutation motif is considered. Olsen et al. (1993) looked at sodium azide mutations in the barley *Ant18* gene and found A.T→G.C base-pair transitions were most frequently generated. However, resequencing this single barley gene does not capture the effects of sodium azide at the whole-genome scale.

In the present study, we examine sodium azide-induced mutations in 11 mutagenized lines of Morex, the barley variety used as the primary reference genome (Mascher et al., 2021). We also generated whole-genome sequence for a sample of the Morex seed stock used for mutagenesis. In addition to single nucleotide changes, we find evidence of sequence insertions and deletions (indels) private to each mutagenized line. Mutations identified in sodium azide-treated lines were compared to variants present in 13 barley landrace samples subject to whole-genome sequencing. To understand the nature of mutations induced by sodium azide, we address the following questions: 1) What is the nature of the variants induced by the mutagen? More specifically, does sodium azide tend to generate SNPs, indels, or other types of structural variants? 2) What is the observed mutation rate in sodium azide-treated samples in barley? 3) Do induced variants differ in type or predicted effect from variants occurring in untreated individuals? and 4) Is a greater number of harmful mutations associated with a reduction in yield in barley?

Results & Discussion

Identifying sodium azide-induced mutations

Data collected and analyzed in the present study includes three distinct datasets. First, we used multiple resequencing datasets to identify differences in our Morex samples and the Morex reference genome. This included new Illumina paired-end data with 10X Genomics linked reads, as well as Oxford Nanopore Technologies (ONT) and published Pacific Biosciences (PacBio) sequence (Mascher et al., 2021) (Table 1; Table S1). A second data set included lines treated with sodium azide that were resequenced with Illumina paired-end data; a subset of these lines also have linked reads and ONT DNA sequencing data (Table S1). We used this data to identify single nucleotide and structural variant differences between our Morex line subject to mutagenesis and the published Morex_v3 reference genome (Mascher et al., 2021). Finally, for contrast, we used Illumina paired-end data to examine the nature of spontaneously occurring variants in 13 barley landraces.

Variants in our Morex sample relative to the Morex reference genome

A major challenge in isolating *de novo* variants induced by the mutagen treatment is the need to distinguish between variants present in the mutagenized seed stock (i.e., Morex) and those that arise during mutagenesis. Heterogeneity, genetic variation within an inbred cultivar or variety, can contribute large numbers of variants (Haun et al., 2011) that are not due to the mutagen. Experimental contamination through unintended hybridization (Michno and Stupar, 2018, Wyant et al., 2022) can also contribute large numbers of variants (Figure S1). Mutagen-induced variants are expected to be relatively rare, requiring filtering of the variants that account for: (1) differences between the parental line used for mutagenesis and the reference genome, (2) uncallable regions (see Methods) that include genomic regions with unknown nucleotide state in the reference, or the regions where sequence reads do not align uniquely, and (3) heterogeneity among lines (Figures S2 & S3). With an estimated genome size of 4.2 Gb for Morex_v3 (Mascher et al., 2021), we identified callable portions of the genome that were 820 Mb and 817 Mb (the latter excludes low complexity sequence). Callable regions capture 88% of high-confidence (HC) genes (31,625 HC genes out of 35,827 total HC genes) annotated in the Morex_v3 reference genome. No individuals in our experiment show an excess of variants or runs of variants consistent with recent hybridization (Michno and Stupar, 2018).

Differences between the Morex_v3 reference genome and our Morex sample are mutations that have arisen in individual seed stocks, errors in the reference assembly, or errors in variant calling. We generated 10X Genomics-based resequencing data set with 46x average mapped read coverage and an ONT data set with 5x mapped coverage (Table S1) to address these issues. We identified 52,596 SNPs, 8,203 1-bp indels (insertions and deletions), 3,182 indels ranging in size from 2-204 bp, and 53 deletions ranging from 41 bp - 60 Kbp that survived rigorous filtering (Figures S2 & S3). Filtered ONT variant calls include 68 insertions and 42

deletions. Using published ONT (at 85x coverage) and PacBio data (at 27x coverage) sampled from 100 seedlings (Mascher et al., 2021) (Table 1) that were incorporated in the Morex_v3 assembly, we detected an additional 178 insertions and 93 deletions that were excluded from identification as mutagen-induced variants.

de novo variants in mutagenized barley lines

Reference-based read mapping, variant calling in GATK (DePristo et al., 2011, Kono et al., 2016, Van der Auwera and O'Connor, 2020), and filtering of the 11 M_5 mutagenized lines identified 23,339 SNVs (Single Nucleotide Variants) with an observed transition to transversion ratio (Ts/Tv) = 5.24. This compares to a ratio of Ts/Tv = 1.7 previously reported based on resequencing of naturally occurring variants in barley (Kono et al., 2016). Among mutagenized lines, we also identified 5,376 smaller indels ranging in size from 1 - 296 bp for a total of 28,715 variants potentially induced by the mutagen (Figure 1A). Here, we use the term SNV (Single Nucleotide Variants) to identify variants generated during mutagenesis and SNP (Single Nucleotide Polymorphisms) to identify variants that are segregating in the non-mutagenized population. Each variant is private to a mutagenized plant (i.e., occurs in only a single individual). There was an average of 2,122 SNVs and 489 indels per mutated sample, with more indels identified in 10X Genomics samples utilizing linked reads (Figure 1B, Table 2). Those numbers are much higher (17x and 139x) than estimates of the average number of mutations that would spontaneously arise in the absence of a single-generation of sodium azide treatment. Those estimates were calculated as below. Using the mean nucleotide substitution rate estimate of 6.5×10^{-9} base substitutions per site per generation from Ossowski et al. (2010) and accounting for our experimental design, we expect ~124 SNVs per individual in the 4.2 Gbp genome. For indels, we expect ~3.5 indels per individual to arise in the absence of the mutagen treatment based on an average indel mutation rate of 0.45×10^{-9} for 1-3 bp indels and 0.5×10^{-9} for >3 bp indels (Ossowski et al., 2010).

The 10X Genomics linked reads and ONT reads of the same three mutagenized lines improved the detection of larger structural variants (SVs). No inversions passed the filtering criteria, and a set of high-quality duplications could not be identified; thus, structural variant calling focused on insertions and deletions. The SVs detected in the 10X Genomics data set included 52 deletions ranging from 41 bp to 60 Kbp (Figure S4). ONT reads for three samples (Table S1) were intended to validate the larger structural variants identified in the 10X Genomics data. The average ONT mapped read coverage was 3.4x (Table S1). A total of 86 insertions (36 - 4,786 bp) and 26 deletions (36 - 300 bp) were called by Sniffles2; 8 insertions (21 - 172 bp) and 6 deletions (16 - 116 bp) were called by cuteSV (Figure S4 & S5). These ONT calls provided direct sequence read-based confirmation of two larger deletion calls that were also called in the 10X Genomics data set.

Variants in untreated barley landraces

For comparison, we generated whole-genome resequencing of 13 barley landraces (Sommer et al., 2020). Average coverage ranged from 41 - 93x (Table S1), and after variant calling and quality filtering, we identified a total of 6.7 million SNPs with Ts/Tv = 1.74 and 849,618 indels ranging in size from 1 - 388 bp. Out of the 6,746,637 SNPs, 2,277,853 SNPs were categorized as rare (i.e., non-reference allele count of two or less, the allele was identified in one or two genotypes) with Ts/Tv = 1.71 and 4,468,784 were common (i.e., non-reference allele count of three or higher) with Ts/Tv = 1.76 (Figure 1A). Rare variants were compared to *de novo* variants in the treated lines because they have experienced fewer generations of selection, and their mutational spectrum is likely more similar to that in treated lines.

Comparison of mutagenized versus untreated samples

SNPs in untreated samples are primarily transitions, particularly C→T* (Figure 2), where the notation C→T* includes the reverse complement G→A. Partitioning variants in the untreated samples into "rare" or "common" had a limited effect on the proportion of variants among each class, with a slight skew of rare variants to more C→ T* transitions and fewer A→G*. Variants in sodium azide-treated lines were dominated by C→T* transitions, which comprised 79.1% of all SNVs in the mutagenized lines (Figure 2). All other variants, including A→G* transitions, comprise 3-5% of all SNVs in sodium azide-treated lines.

Single base pair changes dominate insertion and deletion variants, particularly in sodium-azide treated lines, constituting 28.3% of insertions and 36.4% of deletions (Figure 3). The pattern of indels in treated lines is similar to that in rare and common variants in standing variation. Notable differences include more 1- and 2-bp deletions and 1-bp insertions in treated lines (Figure 3). The 53 larger deletions (41 bp - 60 Kbp) called in the 10X Genomics data set represent roughly 1.7% of all 3,135 deletions in the mutated lines. Linked and long read sequencing was not possible for the landrace lines, precluding a direct comparison.

In mutated samples, 9.7% of SNVs and 8.9% of indels occur in genic regions (Figure S6). The percentage of variants in genic regions is lower in untreated lines for rare and common variants. Rare variants had 6.5% SNPs and 1.7% indels occurring in genic regions and are more directly comparable with mutated sample SNVs and indels. Variant Effect Predictor (VeP) (McLaren et al., 2016) identified most sodium azide-induced variants as occurring in intergenic regions or genomic regions up or downstream of genes. A larger proportion of variants were found in intergenic regions among induced variants than in untreated lines. Fewer sodium azide-induced SNVs and indels were adjacent to genes (Figures S7 & S8). Sodium azide-treated lines have a slightly higher proportion of missense variants (3.86%) than untreated lines (2.19% Rare, 2.32% Common), but this effect is small. Slight increases in the proportion of start-stop-related changes (0.2% Mutated, 0.09% Rare, and 0.07% Common) and splice donor and acceptor sites

are also observed (0.09% Mutated, 0.01% Rare, and 0.01% Common). However, these variants are considered the most potentially damaging based on VeP categorization (Figures S7 & S8). Larger deletion variants (41 bp - 60 Kbp) detected among the three lines with linked reads disrupt a genic region in 11.3% of cases (6 out of 53 total).

Harmful mutations based on phylogenetic constraint

On average, sodium azide-treated lines include 78.6 nonsynonymous SNVs per sample, with 865 nonsynonymous SNVs identified among the 11 mutated lines (Figure S9). Estimates of putative variant effects based on phylogenetic constraints (Kono et al., 2018) were used to identify potentially damaging nonsynonymous variants among primary transcripts in the barley genome. This analysis includes missense variants (a change in amino acid), start lost, stop gained, and stop lost variants based on the Sequence Ontology definition of nonsynonymous changes (Eilbeck et al., 2005). For the 11 mutated lines, 611 nonsynonymous mutations in primary transcripts were tested for phylogenetic constraint relative to 72 other angiosperms to identify potentially universally harmful mutations. Among the successfully annotated mutations, 155 (35.9%) were annotated as “harmful” (i.e., deleterious), while the remaining 277 (64.1%) were identified as “tolerated.” This value compares to 9,716 “rare” nonsynonymous variants tested in 13 barley landraces, where 1,633 (13.2%) are identified as “harmful” and 10,693 (86.8%) as tolerated. For “common” nonsynonymous variants, 14,537 were tested, where 23,506 (7%) are “harmful,” and 311,121 (93%) are tolerated.

An average of 14.1 (±6.46) deleterious SNVs were identified per mutagenized sample (Figure S10). This suggests that there are roughly 14 phenotype-changing variants per individual treated line. The ratio of nonsynonymous SNVs (nSNV) to synonymous SNVs (sSNV) in mutated lines is 1.8:1. In comparison, the ratio of nSNVs:sSNVs in rare and common SNP categories is 1.36:1 and 1.22:1, respectively. The proportion of nSNVs inferred to be deleterious was 17.9% in treated lines versus nSNPs at 3.2% in rare and 4.1% in common categories. To standardize results among samples, we identified the number of harmful mutations per codon in 10 Mbp windows. The proportion of dSNPs per codon was lower near the centromeres for rare and common SNPs in the landraces (Figures S11-S13).

The context of variants induced by sodium azide

Biochemical interactions between mutagenic compounds and DNA produce SNVs in specific nucleotide contexts (Zhu et al., 2020). We used the program Mutation Motif (Zhu et al., 2017) for all SNVs to examine this effect in sodium azide-treated barley lines. The predominant mutation types in both treated and untreated lines are C→T* changes. In sodium azide-treated lines, the cytosine that changes to thymine is frequently followed by another cytosine, creating a CC context of mutation on the forward strand (Figure 4). To our knowledge, there have not been

previous studies on the preferential context of sodium azide mutations. There are highly significant differences between sodium azide-induced variants and variants spontaneously originating in the genome (Table S2). In untreated lines, the mutated cytosine is generally followed by guanine at the +1 or +2 site (downstream) from the C, thus resulting in a CG or potentially CGG context in which mutations occur. In the complete 4.2 Gb Morex_v3 genome assembly, the CC, CG (the two bp motif for CpG changes (Duncan and Miller, 1980), and CGG motifs occur ~228 million, ~154 million, and ~40 million times, respectively. In the 820 Mb region in which unique single nucleotide variants could be called, CC occurs ~48 million times, CG occurs ~35 million times, and CGG motifs occur ~9 million times. This suggests that, on average, a single generation of sodium azide treatment resulted in the mutation of 0.013% of CC sites at which unique mutations could be detected. The CC and CG motifs constitute 5.8% and 4.3% of all two nucleotide combinations in the 820 Mb callable regions, and the CGG motif constitutes 1.1% of three nucleotide combinations. In contrast, AA and TT motifs are the most frequent two nucleotide motifs, making up 8.1%.

Amino acid changes identified in sodium azide-treated lines are dominated by those that include the CC or (reverse complement) GG motif (Figure S14). Glycine to aspartic acid, proline to serine, and alanine to threonine are the three most abundant amino acid changes in SNVs identified as harmful (i.e., deleterious) (Table 3; Figure S15). The top four changes in tolerated SNVs (tSNVs) in mutagenized samples are similar to those annotated as harmful; tolerated amino acid changes include alanine to threonine, alanine to valine, glycine to aspartic acid, and proline to serine (Table 3). This contrasts with amino acid changes induced by rare and common variants in standing variation, where transitions associated with CpG are more abundant. For rare and common dSNPs, alanine-to-threonine and alanine-to-valine changes appear at the highest frequencies. The arginine-to-cysteine amino acid change had the third highest frequency in the common dSNPs class and frequently annotates as deleterious.

Putatively harmful SNVs and phenotypic variation

A total of 25 mutagenized barley lines self-fertilized for five to seven generations ($M_{5:7}$) were used for yield testing at one location in the first year (St. Paul, MN) and three locations in Minnesota (Crookston, Lamberton, and St. Paul) in years 2 and 3. Yield testing was performed in the presence of 5-8 check lines (see Methods) and the original Morex line untreated with the mutagen. Data for heading days after planting (DAP) and plant height were also collected. After spatial adjustment for variation across plots, the average yield for each line was calculated for all years combined. As expected, most mutagenized lines had lower grain yields than the Morex W2017 parental line, with six mutagenized lines yielding roughly the same or slightly higher than the parental line (Figure 5). M29 is among the three lowest-yielding lines and is the only mutagenized line with a visibly distinct phenotype, described as onion-like, short-stature, and very compact (Figure S16). Mutagenized lines tend to have diminished yield relative to Morex,

though some line-by-year combinations slightly exceeded the yield in Morex and some checks (Figure S17). The average diminution in yield across years and lines for the mutagenized lines was 32.8% relative to the Morex W2017 parent. The heading DAP in mutagenized lines increased by 6.2%, and height was reduced by 15% compared to the Morex W2017 parent line. To determine if observed damaging mutations impacted yield, we compared the relative order of yield to the number of damaging mutations per line. We found a slightly negative but nonsignificant correlation of -0.28 (P=0.4) between the number of harmful variants and yield (Table S3). Most mutagenized lines had lower variance across replicates than the check lines. This is likely due to the experimental design with seeds originating from plants that can be traced through single-seed descent following the mutagen treatment, whereas check lines derive from more heterogeneous seed stocks.

Conclusions

Sodium azide is widely used as a mutagen in experimental plant populations. It is frequently used for inducing variants in barley (Olsen et al., 1993, Talamè et al., 2008), including recent reports of extremely large-scale experiments involving the characterization of hundreds of thousands of individual plants (Knudsen et al., 2022). However, most studies have focused on the phenotypic effects of mutagenesis (Schneeberger, 2014) or changes induced at individual genes (Olsen et al., 1993). The genome-level effects of the mutagen have rarely been examined.

Consistent with a prior single gene resequencing study, we find that C→T* transitions dominate induced mutations (Olsen et al., 1993, Knudsen et al., 2022) (Figure 2). Sodium azide appears to generate primarily single nucleotide variants. We identify an average of 2,122 SNVs per mutagenized line. This is an ~88-fold increase in SNVs compared to expectations in the absence of a single-generation sodium azide treatment (see Methods for equations 1-6). Induced indels of all sizes are less abundant in the genome (Figures 1 and 3) but are found to occur at ~130-140 fold higher than nominal mutation rates.

The observation of higher indel rates derives from comparison of data from multiple sequencing platforms, including linked-reads and long-read sequencing, with an average of 489 indels per mutagenized M_5 line. The mutations present in M_5 lines are necessarily a mixture of induced mutations and mutations that arose spontaneously during line maintenance (Zhu et al., 2020). Mutation accumulation resequencing studies in *Arabidopsis thaliana*, found that the 1-3 bp mutation rate averaged 0.45×10^{-9} indels per site per generation, and large deletions (>3 bp) occurred at 0.5×10^{-9} per site per generation (Ossowski et al., 2010). In the 820 Mb portion of the barley genome, where variants can be called unambiguously, we expect ~3.5 indels per individual to arise naturally without mutagenesis treatment over the course of the experiment (see Methods for equations 1-6). In the 4.2 Gbp barley genome, we would expect ~18 indels per

individual (Figure 1B). This is a ~130-140 fold increase in the indel mutation rate to an average of 5.13×10^{-7}.

Most sodium azide-induced mutations occur in a specific nucleotide sequence context, as C→T* changes in a CC mutation motif (Figure 6) or the reverse complement. This results in a relatively small number of amino acid changes that predominate among induced mutations. In the most extreme comparison, sodium azide-induced amino acid changes are very distinct from the majority of amino acid changes segregating in barley. The amino acid changes that annotate as harmful and predominate in the mutagenized samples are glycine to aspartic acid, proline to serine, and alanine to threonine. In comparison, the top three amino acid changes annotated as harmful and segregating in the untreated barley landraces include alanine to threonine, alanine to valine, and arginine to cysteine. For projects seeking to induce novel changes, for example, in disease resistance genes or genes associated with stress tolerance, sodium azide will induce many coding changes that are rarely observed among standing variation.

Resequencing of individual genes identified many sodium azide-induced SNVs in barley (Olsen et al., 1993). Induced indels and SVs were not previously reported but would be difficult to identify with Sanger sequencing. Indeed, sequencing technology continues to present a limitation. Many of the SVs identified here were identified by linked reads (in two cases verified by ONT long reads) but could not be identified by Illumina paired-end reads alone. Regarding relative effect, indels and SVs identified with linked reads and verified with ONT result in six disruptive mutations that either induce a frameshift or eliminate a portion of a coding gene. This results in an average of two structural disruptions of genes per individual instead of an average of 33.4 per individual due to 1 - 3 bp nucleotide sequence-level changes.

Our mutated lines average a 37.7% reduction in yield relative to their non-mutated parental Morex line. This reduction in yield can be attributed to an average of 14.1 induced SNVs and 37.3 indels per line. The typical line has an average of 273.3 disruptions of coding variants, including SNVs and indels. The approach used in this study identified a finite number of deleterious (i.e., harmful) mutations induced by sodium azide. It was successful at creating lines that had lower yield than the untreated Morex parent line in the experiment. The reduction in fitness (using yield as a proxy for fitness) following the mutagen treatment was expected, given that most new amino-acid-changing mutations that impact fitness will be deleterious (Eyre-Walker and Keightley, 2007, Boyko et al., 2008).

In practical applications, deleterious variants are relatively easy to detect, which makes it possible to select against or eliminate them via targeted replacement of individual variants (Morrell et al., 2012, Moyers et al., 2018, Smith et al., 2018, Johnsson et al., 2019). However, it is still challenging to identify the effects of individual deleterious variants. With lines that have reduced yield and a better understanding of the nature of changes generated by sodium azide and the sequence contexts in which they occur, there is the possibility of training machine-learning

models to predict which variants contribute to harmful phenotypic change (Plekhanova et al., 2018, Benegas et al., 2023). Then, quantify the effect size of each harmful variant into a ranked list and combine the predictions with existing genomic prediction approaches to benefit plant and animal breeding programs (Wallace et al., 2018) and the study of complex human diseases.

Our results indicate that a portion of the new mutations induced by chemical mutagens may be predictable in the sense that they occur in a specific nucleotide context and thus result in forseeable effects on protein-coding changes. This is likely true for both mutagens in experimental treatments and those for mutagens that organisms encounter in natural environments. Ongoing characterization of the nature of induced mutations provides the potential for a better understanding of the nature of mutagen-induced variation and their potentially damaging or phenotype-changing effects (Zhu et al., 2020).

Materials and Methods

Plant materials and mutagenesis

Barley from a Morex seed stock was treated with sodium azide following the protocol in Döring et al. (1999). Morex is a 6-row malting barley variety used as the primary reference genome (Mascher et al., 2021). Morex was chosen for these experiments because new mutations are more directly detectable relative to the reference genome. The Morex line in this experiment traces back to the parent seed stock used to generate the Steptoe x Morex doubled haploid barley mapping population (Kleinhofs et al., 1993). To generate sufficient Morex seeds for sodium azide mutagenesis, 120 seeds were planted from a single Morex plant (Figure S1). Next, 200 seeds from the resulting bulk of seeds were planted; this was repeated one more time. A portion of the seeds was treated with sodium azide (1 mM NaN_3) following the Döring et al. (1999) protocol; the remaining portion of untreated Morex seeds was planted for another round of seed bulking and then planted to collect leaf tissue for sequencing (Figure S1). The M_1 individuals were grown to maturity and harvested after the mutagen treatment. These individuals then underwent single-seed descent until M_5.

Estimates for the expected number of spontaneous mutations occurring without mutagen treatment were calculated using the experimental design for seven generations of self-fertilization (Figure S1) and rate estimates from Ossowski et al. (2010). The mutation rates used for SNPs was 6.53×10^{-9} and for 1-3 bp indels was 0.45×10^{-9}. For the 820 Mbp (820,594,305 bp) callable regions, a diploid genome size of 1,641,188,610 bp was used. For estimates of the 4.2 Gbp (4,225,605,719 bp) genome size (Mascher et al., 2021), a diploid genome size of 8,451,211,438 bp was used. Each generation, new mutations appear in the heterozygous state, and the number of new heterozygous mutations (N_{het}) is given by

$$N_{het} = Diploid\ bp * Mutation\ rate \quad (1)$$

The experimental design involved multiple generations of selfing, meaning *de novo* mutations from previous generations are being lost or fixed over time. In each generation, heterozygous mutations are inherited (I_{het}) and are represented by

$$I_{het} = \begin{cases} 0 & for\ x = 0 \\ 0.5 \times N_{het}G_{x-1} & for\ x = 1 \\ 0.5 \times N_{het}G_{x-1} + 0.5 \times I_{het}G_{x-1} & for\ x > 2 \end{cases} \quad (2)$$

where x is the current generation. Similarly, each generation homozygous mutations are inherited (I_{hom}) and are represented by

$$I_{hom} = \begin{cases} 0 & for\ x < 2 \\ 0.25 \times I_{het}G_{x-1} & for\ x = 2 \\ I_{hom}G_{x-1} + 0.25 \times I_{het}G_{x-1} & for\ x > 2 \end{cases} \quad (3)$$

The number of heterozygous and homozygous spontaneous mutations accumulated by generation seven in our experimental design is given by

$$Total_{het}G_7 = I_{het}G_7 \quad (4)$$

$$Total_{hom}G_7 = I_{hom}G_7 \quad (5)$$

$$Total_{spont}G_7 = I_{het}G_7 + I_{hom}G_7 \quad (6)$$

and is used as our estimated number of spontaneous mutations that would have been present without mutagen treatment.

Phenotypic data collection

Twenty-five $M_{5:7}$ mutated lines and $F_{1:2}$ W2017 Morex parent line were evaluated in yield trials at one location (St. Paul, MN) in 2020 and three locations in Minnesota (Crookston, Lamberton, and St. Paul) in 2021 and 2022. Lines were grown in a randomized complete block design. Phenotypic data on grain yield, heading days after planting (DAP), height, and lodging were collected. Check varieties were used to adjust for spatial variation across trial plots for traits with a continuous scale (yield, heading DAP, and height). Spatial adjustments were performed using the R package mvngGrAd (Frank, 2015). All checks had 21 replicates and included Conlon, FEG141-20, Lacey, ND20448, ND26104, ND_Genesis, Pinnacle, and Rasmusson.

Whole-genome short-read and long-read sequencing

We generated whole-genome sequencing in 25 barley (*Hordeum vulgare* ssp. *vulgare*) accessions: the parent of the mutagenized lines (W2017 Morex), 11 mutagenized lines, and 13 barley landraces (Sommer et al., 2020) for comparative analyses (Table 1; Supplemental Table 1). High molecular weight genomic DNA was extracted from 4-6 week-old leaf tissue collected on ice using the Cytiva Nucleon PhytoPure kit for the mutagenized lines. We sequenced three of the 11 M_5 mutagenized lines (M01, M20, and M29) and the W2017 Morex line using the 10X Genomics linked read library preparation followed by Illumina NovaSeq 6000 sequencing with

150-bp paired-end technology to a target depth of 40x. For the remaining eight M_5 mutagenized lines (M02, M11, M14, M28, M35, M36, M39, and M41), libraries were prepared using the Illumina DNA Prep followed by Illumina NovaSeq 6000 sequencing with 150-bp paired-end technology to a target depth of 16x. Sequences for the 13 landraces were generated using the Illumina TruSeq DNA Nano Prep followed by Illumina NovaSeq 6000 sequencing with 150-bp paired-end technology to a target depth of 40x.

We used Oxford Nanopore Technologies (ONT) to sequence W2017 Morex and three of the 11 M_5 mutagenized lines (M01, M20, and M29, see Table 1; Table S1). This data was collected to provide read-level confirmation of SVs indicated by the Illumina short-read resequencing. For sample M01, following the ONT protocol, a high molecular weight gDNA extraction was performed using the Qiagen Genomic-tip kit (10262) with Carlson Lysis buffer (10450002-1). High molecular weight gDNA extractions for samples M20 and M29 were generated using the NucleoBond HMW DNA kit (740160.20) from Takara Bio USA. Size selection was performed using Circulomics SRE buffer, and DNA was quantified using the Qubit assay. The libraries were prepared with 400 ng of gDNA using the Rapid Sequencing Kit (SQK-RAD004) following the protocol version RSE_9046_v1_revT_14Aug2019. The library was primed using the flow cell priming kit (EXP-FLP002), then 400 ng of the library was loaded onto an R9.4.1 flow cell (FLO-MIN106D). M01 was run for 72 hours on a MinION Mk1C (MIN-101C). We found the flow cells were no longer collecting new data after 24 hours and modified the run for the remaining two samples. For M20 and M29, 400 ng of the library was loaded onto an R9.4.1 flow cell, run for 12 hours, and then paused. At this point, the pores were cleared using the flow cell wash kit (EXP-WSH004), then 400 ng of additional library was loaded and run for another 12 hours before the run was paused. Again, pores were cleared with the flow cell wash kit, then 400 ng of an additional library was loaded. The flow cell was run for an additional 24-30 hours. Three reactions were run for a single flow cell; a single sample with two washes ran for a total of 48-56 hours. This approach produced the highest data output for our samples. This process was repeated until each mutagenized line was sequenced to a target depth of 2-3x (Table S1). Basecalling was performed using Guppy v5.0.12+eb1a981 (for all runs except one) and Guppy v5.0.17+99baa5b (for one M01 run #3) using the default setting on the MinION Mk1C.

Read mapping and variant calling

Read alignment and variant calling for the eight mutagenized and 13 landrace WGS lines were processed using the sequence_handling workflow (https://github.com/MorrellLAB/sequence_handling), which integrates publicly available software into a series of bash scripts (Liu et al., 2022). The configuration files, which identify software versions and parameters, and scripts are available in the GitHub repositories https://github.com/MorrellLAB/hybrid_barley and

https://github.com/MorrellLAB/Barley_Mutated. Reads were aligned against the third version of the barley Morex reference genome (Morex_v3) (Mascher et al., 2021) with parameters adjusted to account for the level of nucleotide diversity in barley. Variants were called as part of a larger set of samples and followed the Genome Analysis Toolkit (GATK) best practices recommendations (DePristo et al., 2011, Van der Auwera and O'Connor, 2020). SNPs underwent GATK VariantRecalibrator with the following as input: filtered variants and SNPs from genotyping assays, which include 2,975 BOPA SNPs (Close et al., 2009), 7,541 9K SNPs (Comadran et al., 2011), and 41,813 50K SNPs (Bayer et al., 2017). Only polymorphic and biallelic SNPs were included. Additional SNP filtering criteria include allele balance deviation of 0.1, proportion heterozygous genotypes at a site > 0.1, per sample minimum DP < 5, per sample maximum DP > 158, proportion missing genotypes at a site > 0.30, QUAL < 30, and GQ < 9. SNPs identified in a barley Sanger resequencing dataset (Morrell et al., 2006, 2014) were used for validation. Indels were filtered following GATK's Best Practices Guidelines for hard filtering since we did not have enough truth and training datasets to run indels through VariantRecalibrator. All filtering criteria are detailed in the scripts available in the GitHub repository https://github.com/MorrellLAB/Barley_Mutated.

For the four 10X Genomics samples (W2017 Morex and three mutagenized lines, see Table S1), reads were aligned to the Morex v3 reference genome, and variants were called with the 10X Genomics software, Long Ranger v2.2.2. The Long Ranger pipeline processes the Chromium-prepared sequencing samples. Variants were filtered based on filters generated by Long Ranger, which include: 10X_QUAL_FILTER, 10X_ALLELE_FRACTION_FILTER, 10X_PHASING_INCONSISTENT, 10X_HOMOPOLYMER_UNPHASED_INSERTION, 10X_RESCUED_MOLECULE_HIGH_DIVERSITY, and LOWQ. SNPs and 1-bp indels were filtered to sites with per sample DP between 5 and 78 and an allele balance deviation of +/- 0.2 (from the expected 0.5) for heterozygous genotypes.

For the ONT data of W2017 Morex, M01, M20, and M29, read quality and summary statistics were generated with NanoPlot v1.38.1 and pycoQC (Leger and Leonardi, 2019). Adapters were trimmed with Porechop v0.2.4 (Wick et al., 2017). Reads were then aligned to the barley Morex v3 reference genome using Minimap2 v2.17 (Li, 2018) with parameters recommended for ONT sequence reads. The resulting SAM files were then realigned using a modified version of the Vulcan pipeline (Fu et al., 2021) (customized version, https://gitlab.com/ChaochihL/vulcan), which utilizes NGMLR (Sedlazeck et al., 2018) for read realignment, converted to BAM format, and sorted using Samtools v1.9 (Li et al., 2009). Structural variants were then called using Sniffles v2.0.3 (Sedlazeck et al., 2018, Smolka et al., 2022).

We also used a publicly available Morex individual sequenced with PacBio CCS reads (BioProject PRJEB40587, ERR numbers ERR4659245-ERR4659249) (Mascher et al., 2021). We used HiFiAdapterFilt (Sim et al., 2022) to filter adapters. Reads were aligned using

Minimap2 v2.17 (Li, 2018) using parameters recommended for PacBio sequence reads. Similar to the ONT data, the resulting SAM files were realigned using a modified version of the Vulcan pipeline (Fu et al., 2021) (customized version, https://gitlab.com/ChaochihL/vulcan), converted to BAM format, and sorted using Samtools v1.9 (Li et al., 2009). Structural variants were called used Sniffles v2.0.3 (Sedlazeck et al., 2018, Smolka et al., 2022).

For all datasets in this study, mapped coverage was estimated using Mosdepth v0.3.1 (Pedersen and Quinlan, 2018). All mapping parameters and filtering criteria are detailed in scripts available in the GitHub repository (https://github.com/MorrellLAB/Barley_Mutated).

Identifying *de novo* variants

Part 1: Finding differences between Morex samples and Morex reference genome

To identify *de novo* variants induced by the mutagen, we generated a list of regions where variants were called in Morex-sample2 (W2017 Morex); these are differences between the Morex parent in this study and the Morex reference genome. Variants called in the Mascher et al. (2021) Morex ONT and PacBio data were also counted as differences from the reference; these are potentially due to heterogeneity (variation among individuals in the Morex variety). To minimize spurious SV calls in difficult-to-call regions for the 10X Genomics, ONT, and PacBio data, we filtered out variants that overlap with uncallable regions, which includes annotated repeats, stretches of N's in the reference genome sequence, and "high copy" regions (i.e., regions where plastids, rDNA repeats, and centromere repeats align). For the ONT and 10X Genomics data, SVs that overlap low complexity regions (defined as regions containing low-copy sequence) were also filtered out because they can be non-biological artifactual sequences that result in unmapped sequences or sequences mapped to multiple locations. Low complexity regions were generated using BBMask from BBTools (BBMap – Bushnell B. – sourceforge.net/projects/bbmap/) with entropy set at 0.7, which was determined through data exploration to capture a majority of low complexity sequences (scripts available at https://github.com/MorrellLAB/morex_reference/tree/master/morex_v3). For the ONT and PacBio data, SVs were filtered out if they had less than five supporting reads. For Morex-sample2, SNPs in 100 bp windows with >2% diversity were filtered out. Such high diversity windows are unlikely when aligning Morex-sample2 to the Morex reference genome. SVs were then visually inspected in IGV or scored in SV-plaudit (in the case of the 10X Genomics larger deletions), and filtering criteria were tuned if necessary (summarized in Figure S2). The filtered SVs form a high confidence set of places where *de novo* variants shouldn't be called due to regions that are difficult to align or are heterogeneous among Morex individuals.

Part 2: de novo filtering mutated individuals

For the 10X Genomics and ONT sequenced mutagenized lines (M01, M20, M29), SVs that overlap uncallable regions or low complexity regions were filtered out (Figure S3). To benefit from using the strengths of distinct SV callers for ONT data, we utilize Sniffles2 (Sedlazeck et al., 2018) and cuteSV (Jiang et al., 2020). SVs called by cuteSV and Sniffles2 in the ONT data were filtered similarly, except in the Sniffles2 calls, we required at least five supporting reads. For all sequenced mutagenized lines (10X Genomics, ONT, and Illumina WGS), variants that overlap the "differences from reference" regions were filtered out (summarized in Figure S2). SNPs identified in the mutagenized samples that also appear in the BOPA (Barley Oligo Pooled Assay 1 and 2) on the Illumina Golden Gate genotyping platform (Close et al., 2009), Barley 9K Illumina Infinium iSelect Custom Genotyping BeadChip (Comadran et al., 2012), and Barley 50K iSelect SNP array (Bayer et al., 2017) panels were also excluded. Variants were filtered to those private to individual mutagenized samples, meaning the variant only exists in one of the mutagenized samples at a genomic position. This is based on the expectation that variants induced by the mutagen are new (arose after the mutagen treatment was applied) and unique to each individual. So, variants identified in the mutagenized lines that also exist in the Morex samples are likely due to heterogeneity in Morex and were not generated by the mutagen treatment. Again, variants were visually inspected in IGV (Thorvaldsdóttir et al., 2013) or igv-reports (https://github.com/igvteam/igv-reports), and filtering criteria were tuned if necessary.

An image scoring approach was conducted to verify the larger deletions in the 10X Genomics three mutagenized samples. Images of the SVs were created with the SamPlot software (Belyeu et al., 2021). Following the pipeline implemented in SV-Plaudit (Belyeu et al., 2018), the SV images were stored in the Amazon Web Services cloud storage and scored by multiple investigators through the PlotCritic website (https://github.com/jbelyeu/PlotCritic) based on the following criteria: coverage, insert size, and linked/split read evidence. This produced a set of scored deletions where a majority of scorers confirmed read evidence for the variant. This confirmed variant set was then verified using the ONT-sequenced samples (M01, M20, and M29).

Nucleotide composition of variants

Mutation Motif (Zhu et al., 2017) was used to identify the most frequent sequence motifs that affect SNPs in the mutated, rare, and common variant classes. This program performs comparative statistical analyses of neighborhoods (5 bp windows) centered around each SNP to identify the frequency of sequence motifs and the influence of neighboring bases for each SNP. The neighborhood of each SNP (e.g., C→T mutation) is compared to the neighborhood of a reference occurrence of the same nucleotide (e.g., C) randomly sampled from within +/- 100 bp of the mutation. Comparing the motifs associated with mutated, rare, and common classes of SNPs allowed us to identify specific motifs that sodium azide may preferentially target. The

number of 2- and 3- bp motifs in the 4.2 Gb reference genome and 820 Mb callable regions was calculated using the EMBOSS (Rice et al., 2000) compseq tool.

Deleterious predictions

Ensembl Variant Effect Predictor (VeP) (McLaren et al., 2016) was used to determine the predicted effect of each variant in the filtered VCF file, which includes SNPs, insertions, and deletions. Identification of nonsynonymous variants (includes missense, start lost, stop gained, and stop lost variants) used gene models provided by Mascher et al. (2021). Nonsynonymous variants for the mutated samples and landraces were extracted from the VeP reports and assessed using BAD_Mutations (Kono et al., 2016, 2018), which includes a likelihood ratio test (Chun and Fay, 2009) that compares codon conservation across Angiosperm species to determine if a base substitution is likely to be deleterious. We ran the BAD_Mutations pipeline with a set of 72 Angiosperm species genome sequences that are available through Phytozome v13 (https://phytozome-next.jgi.doe.gov/, last accessed November 29, 2021) and Ensembl Plants (http://plants.ensembl.org/, last accessed November 29, 2021). We ran BAD_Mutations using 35,827 primary transcripts. A SNP was annotated as deleterious if the *P*-value for the test was <0.05 with a multiple tests correction based on the number of tested codons, minimum of 10 sequences, maximum constraint of 1, and if the alternate or reference allele was not seen in any of the other species. Our thresholds for the three data groups were 8.1E-5 (611 codons tested) for the mutated samples, 5.1E-6 (9,716 codons tested) for the rare variants, and 3.4E-6 (14,537 codons tested) for the common variants. SNPs that failed this set of criteria were annotated as tolerated.

Data Availability

SRA numbers: WGS Morex and Morex treated with sodium azide BioProject PRJNA849997. WGS barley landraces BioProject PRJNA674330. ONT of Morex and Morex treated with sodium azide BioProject PRJNA967725.

Github: https://github.com/MorrellLAB/Barley_Mutated

DRUM: TBA

Acknowledgements

The authors thank Ron Okagaki for help with the initial screening of the mutant lines; Lucie Lu for submitting the raw sequencing data to NCBI SRA; Elaine Lee for processing the 10X Genomics datasets using Longranger; Nadia Janis for scoring images of the deletions; Malik Samuel, Emily Vonderharr, Samuel Hamann, and Mackenzie Linane for help with growing out the first couple of generations and processing the seed; Erica Sun for making digital sketches of

the barley plants, and Max Okagaki for help with Figure S14. This study was supported by a University of Minnesota Informatics Institute MnDRIVE Graduate Assistantship award to Chaochih Liu, the National Science Foundation (grant IOS-1339393), and the Minnesota Agricultural Experiment Station fund (MIN-13-122 in support of Peter L. Morrell). This research was carried out with software and hardware support provided by the Minnesota Supercomputing Institute (MSI) at the University of Minnesota.

Author Contributions

CL, PLM, and GF wrote the paper. CL, GF, and PLM performed the analyses. SH and GJM generated the mutant population. CL, PLM, JCF, and KPS decided which individuals to sequence and include in this study. AOS, GF, LL, PLM, and SRW generated Oxford Nanopore DNA sequencing data. EL and MS provided the landrace and cultivar samples.

Tables

Table 1. Brief overview of sequencing datasets used in this study.

Acronym	Number Samples	Description	Reference	BioProject
10X Genomics	4	Reads generated from 10X Genomics library preparation followed by 2x150bp PE Illumina whole genome sequencing for Morex and 3 mutated lines; 46X mapped coverage	This study	will add
WGS	21	2x150bp PE Illumina whole genome sequencing	This study	will add
ONT	4	Oxford Nanopore reads of Morex and 3 mutated lines; 5X mapped coverage	This study	will add
ONT	1 (100 seedlings)	Oxford Nanopore reads of Morex (sampled from 100 seedlings); 85X coverage	Mascher et al. 2021	PRJEB40588
PacBio	1 (100 seedlings)	PacBio circular consensus reads of Morex (sampled from 100 seedlings); 27X coverage	Mascher et al. 2021	PRJEB40587

Table 2. Number of variants in each of the the sodium azide treated barley samples and the parent of the mutant lines (Morex-sample2). Counts for indels are variants private (appears in only one sample) to each individual. Morex-sample2 is the same accession as the reference genome but a derived from a different seed lot.

					Larger SVs (41 bp-60 Kbp)		
Accession	**Type**	**Sequencing**	**SNV Count**	**Indel Count (1-296 bp)**	**Total larger deletions (10x Genomics)**	**Total indels (ONT-Sniffles2)**	**Total indels (ONT-cuteSV)**
Morex-sample2	elite/ parent	10x Genomics, ONT	52,596	11,385	53	110	N/A
M01	mutant	10x Genomics, ONT	1,134	441	17	40	1
M20	mutant	10x Genomics, ONT	3,225	464	14	41	9
M29	mutant	10x Genomics, ONT	3,669	445	21	31	4
M02	mutant	Illumina 2x150bp	1,585	56	N/A	N/A	N/A
M11	mutant	Illumina 2x150bp	1,222	54	N/A	N/A	N/A
M14	mutant	Illumina 2x150bp	1,446	62	N/A	N/A	N/A
M28	mutant	Illumina 2x150bp	2,357	80	N/A	N/A	N/A
M35	mutant	Illumina 2x150bp	2,431	75	N/A	N/A	N/A
M36	mutant	Illumina 2x150bp	1,562	64	N/A	N/A	N/A
M39	mutant	Illumina 2x150bp	1,613	705	N/A	N/A	N/A

M41	mutant	Illumina 2x150bp	3,095	63	N/A	N/A	N/A

Table 3. Most abundant amino acid changes annotated as deleterious (DEL) and tolerated (TOL) for mutated, rare, and common SNPs.

	Mutant			Rare			Common		
	AA Change	Number	%	AA Change	Number	%	AA Change	Number	%
DEL SNVs	G/D	22	14.19	A/T	62	3.97	A/T	140	3.18
	P/S	20	12.90	A/V	61	3.90	A/V	137	3.12
	A/T	19	12.25	G/D	38	2.43	R/C	91	2.07
	A/V	10	6.45	P/L	37	2.37	L/F	86	1.96
TOL SNVs	A/T	40	14.44	A/T	395	3.89	A/T	1632	3.29
	A/V	33	11.91	A/V	383	3.77	T/A	1590	3.21
	G/D	23	8.30	V/I	265	2.61	A/V	1552	3.13
	P/S	23	8.30	T/A	264	2.60	V/A	1548	3.12

Table S1. Detailed summary of samples sequenced in this study, which samples were treated with sodium azide, the library preparation and sequencing technology, and mapped average coverage.

Table S2. Log-linear analysis performed by Mutation Motif for C→T variants induced by sodium azide compared to those originating spontaneously in the reference sequence.
Position is relative to the mutating base. Deviance is a likelihood ratio from the log-linear model. Degrees-of-freedom (*df*) and *P*-values are from the chi-squared distribution.

Position(s)	Deviance	*df*	*P*-value
-2	37.6723914	3	3.32E-08
-1	45.9155839	3	5.91E-10
+1	4080.13219	3	0.0
+2	36.5891731	3	5.62E-08
(-2, -1)	265.618117	9	5.00E-52
(-2, +1)	59.5660276	9	1.63E-09
(-2, +2)	25.1941677	9	0.00276374
(-1, +1)	191.213936	9	2.29E-36
(-1, +2)	22.1115064	9	0.00853222
(+1, +2)	71.113564	9	9.21E-12
(-2, -1, +1)	37.6271604	27	0.08394335
(-2, -1, +2)	43.5490506	27	0.02300593
(-2, +1, +2)	42.1937658	27	0.03149572
(-1, +1, +2)	38.532008	27	0.06983028
(-2, -1, +2, +2)	102.703957	81	0.05214915

Table S3. The Pearson or Spearman correlation between the number of SNVs and the phenotypes across each functional class of variants.

Class	Test	Yield	Yield *P*-value	Heading DAP	Heading DAP *P*-value	Height	Height *P*-value
Noncoding	Pearson	-0.2	0.55	0.28	0.40	-0.47	0.14
Synonymous	Pearson	-0.052	0.88	0.09	0.80	-0.23	0.50
Nonsynonymous	Pearson	-0.16	0.65	0.04	0.91	-0.32	0.34
Tolerated	Pearson	-0.12	0.73	-0.03	0.93	-0.16	0.64
Deleterious	Spearman	-0.28	0.40	0.17	0.63	-0.39	0.23

Figures

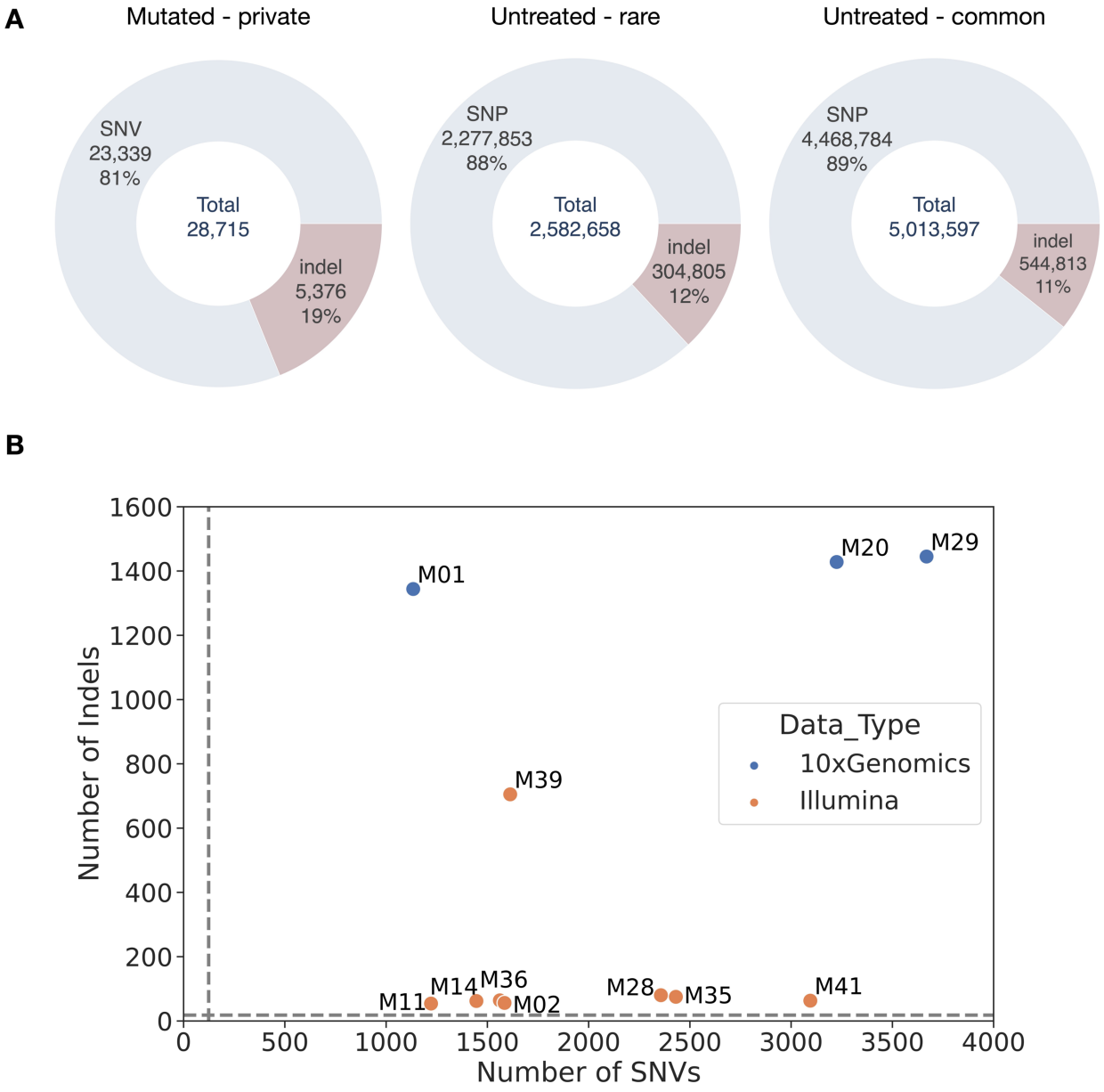

Figure 1. A) The number and percentage of SNVs and indels in mutagenized, untreated rare, and untreated common categories. Private is defined as variants identified in a single sample. Rare is defined as variants with a non-reference allele count of two or less, and common is defined as variants with a non-reference allele count of three or higher. B) The number of SNVs and indels identified in each mutagenized line. Colors indicate whether 10X Genomics linked read technology was utilized for the sample. Dashed lines indicate the expected number of SNPs and indels based on average mutation rates and accounting for experimental design.

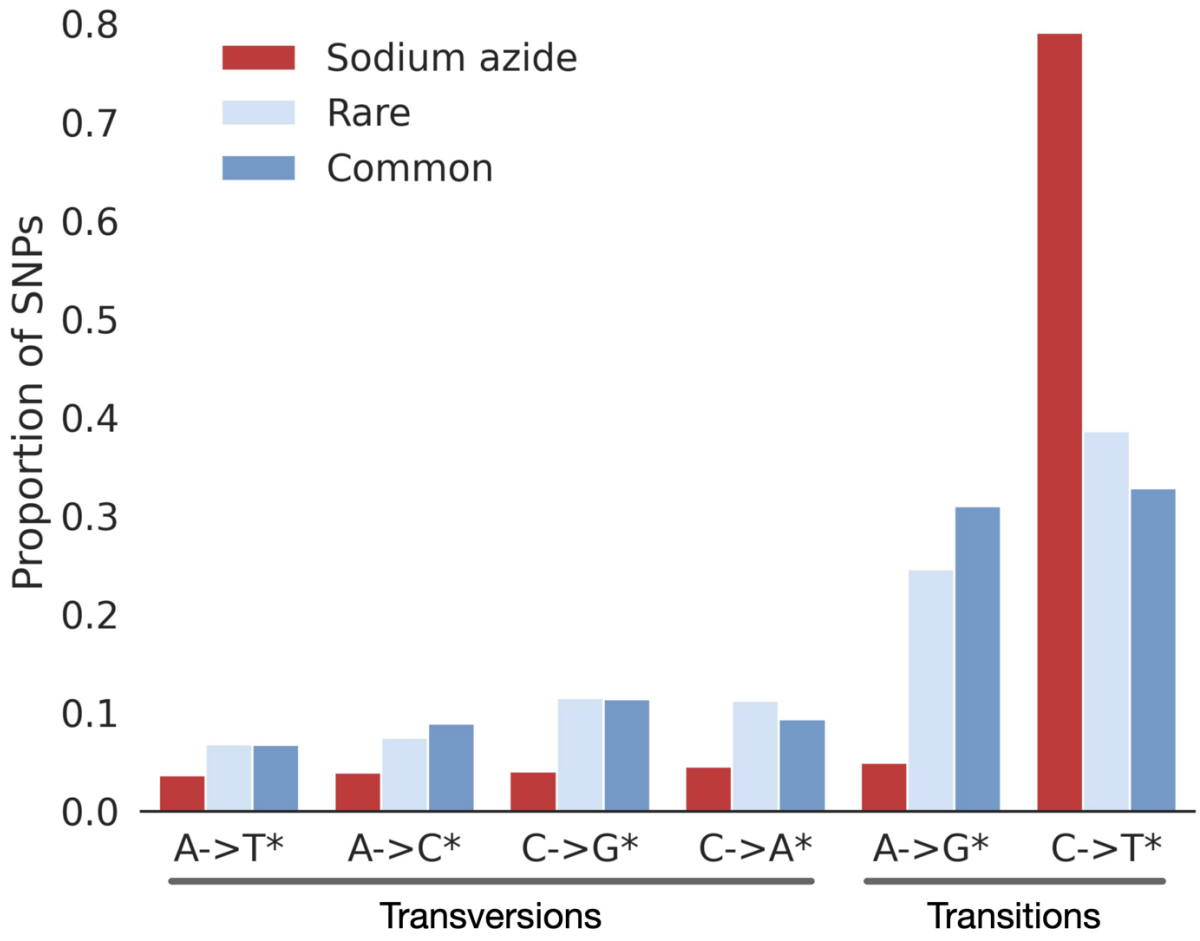

Figure 2. The mutational spectrum of mutagenized, untreated rare, and untreated common SNPs. Each bin includes the reverse complement. For example, the C→T* bin also includes G→A changes.

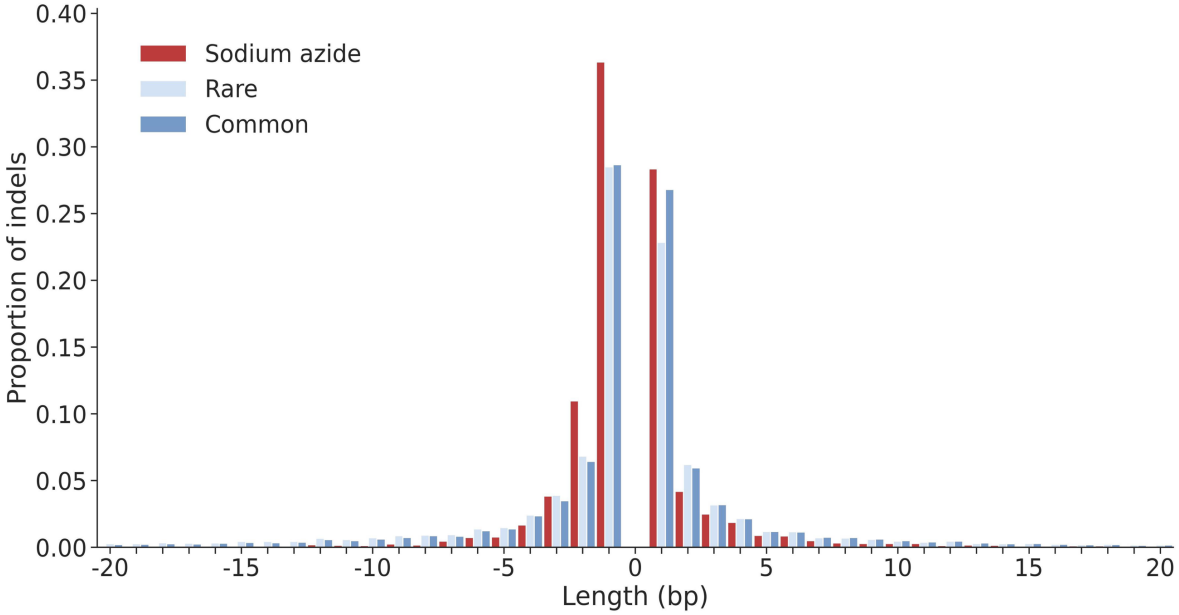

Figure 3. The distribution of insertion and deletion lengths for sodium azide treated lines versus rare and common categories. Insertions are shown as positive values and deletions are shown as negative values. Only variants with lengths <20 bp are shown here.

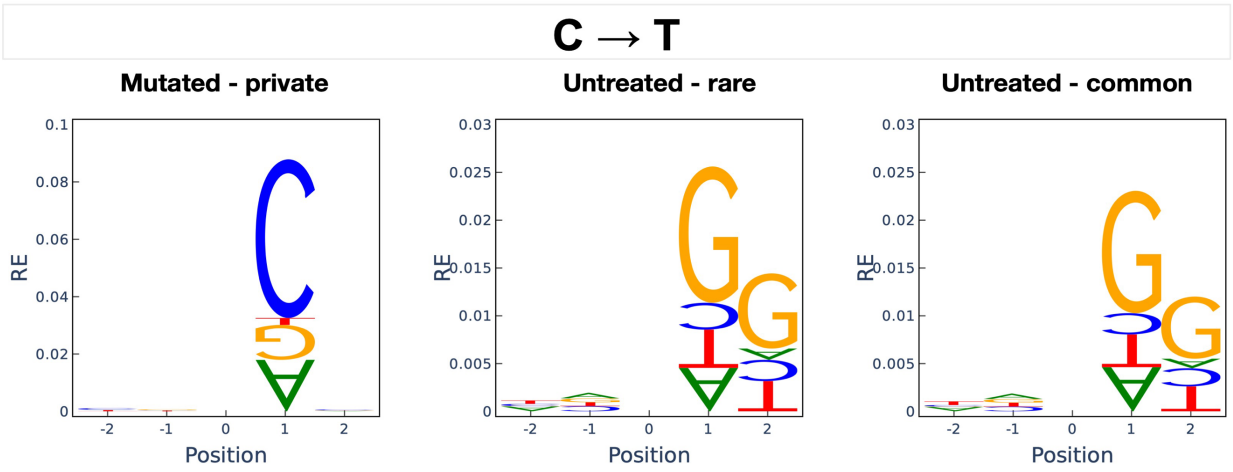

Figure 4. The nucleotide sequence context for C→T transitions relative to the reference genome in the mutated (*N*=10,048), untreated rare (*N*=439,563), and untreated common (*N*=732,565) SNVs. Position 0 indices where the C→T change occurred. The relative height of the letters indicates their relative entropy (RE), with a higher RE indicating a position has a greater influence on the mutation. Upright letters indicate overrepresented bases, whereas upside-down letters indicate underrepresented bases at positions neighboring position 0. The null expectation (RE of zero) is based on randomly sampling a nearby location with the same starting base (e.g., for a C→T mutation, a random choice of a position with a C is selected.

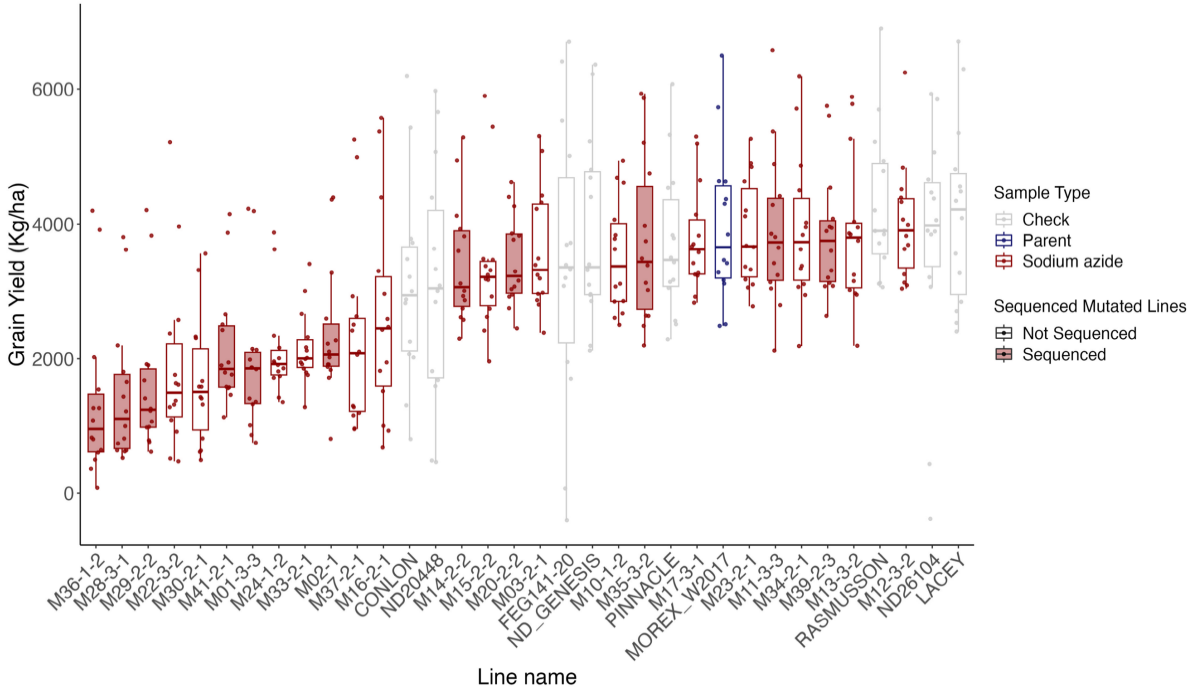

Figure 5. Grain yield for 25 mutated lines, the Morex W2017 parent, and 8 check lines. The box plots are sorted by the median for each line, and the bars in the box plot indicate the mean. Sodium azide-treated lines are represented by red outlines. Red shaded boxes indicate mutated lines that were sequenced in this study.

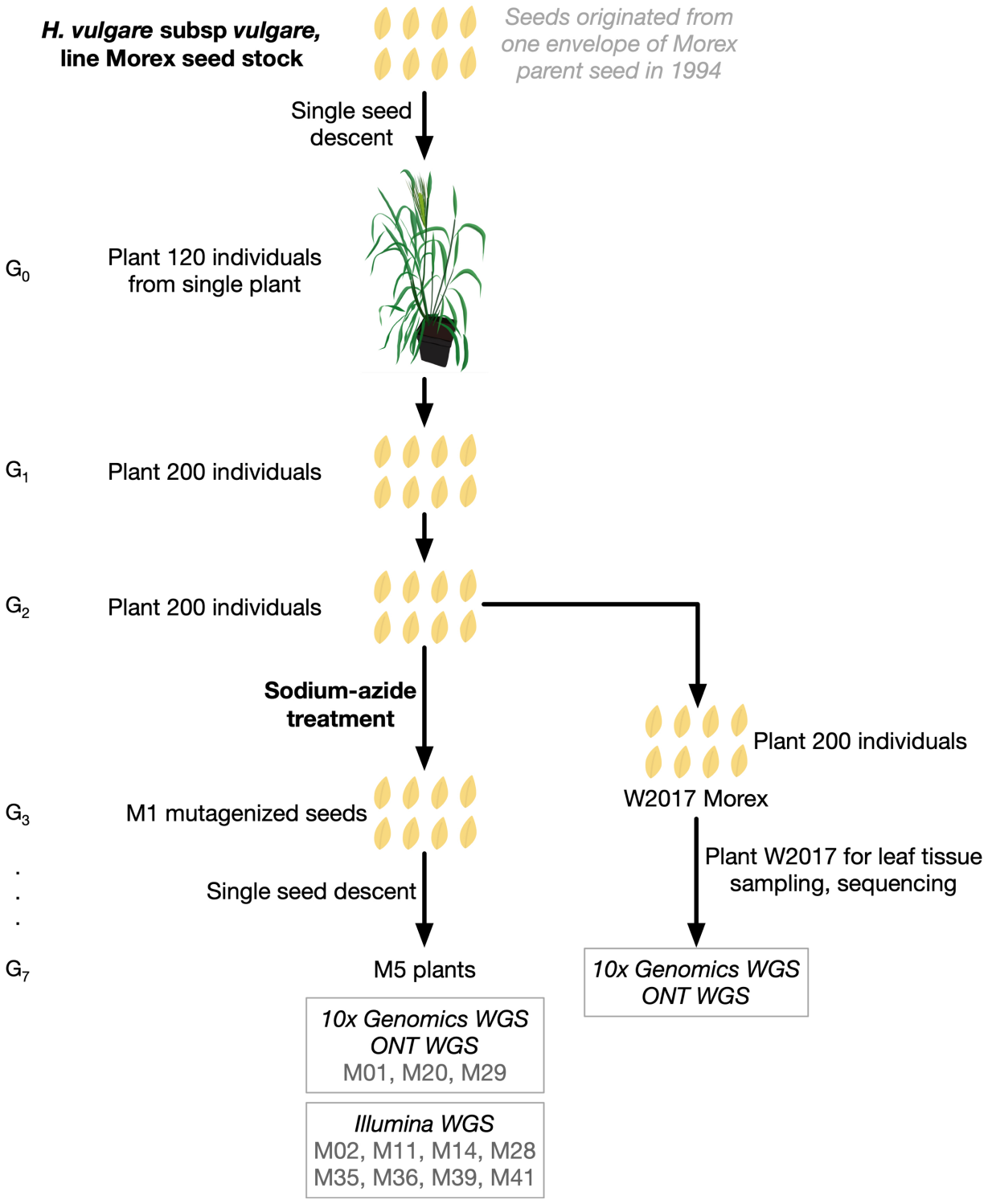

Figure S1. Experimental design for generating the sodium azide treated lines. G_x is the generation used for spontaneous mutation estimates (see Methods).

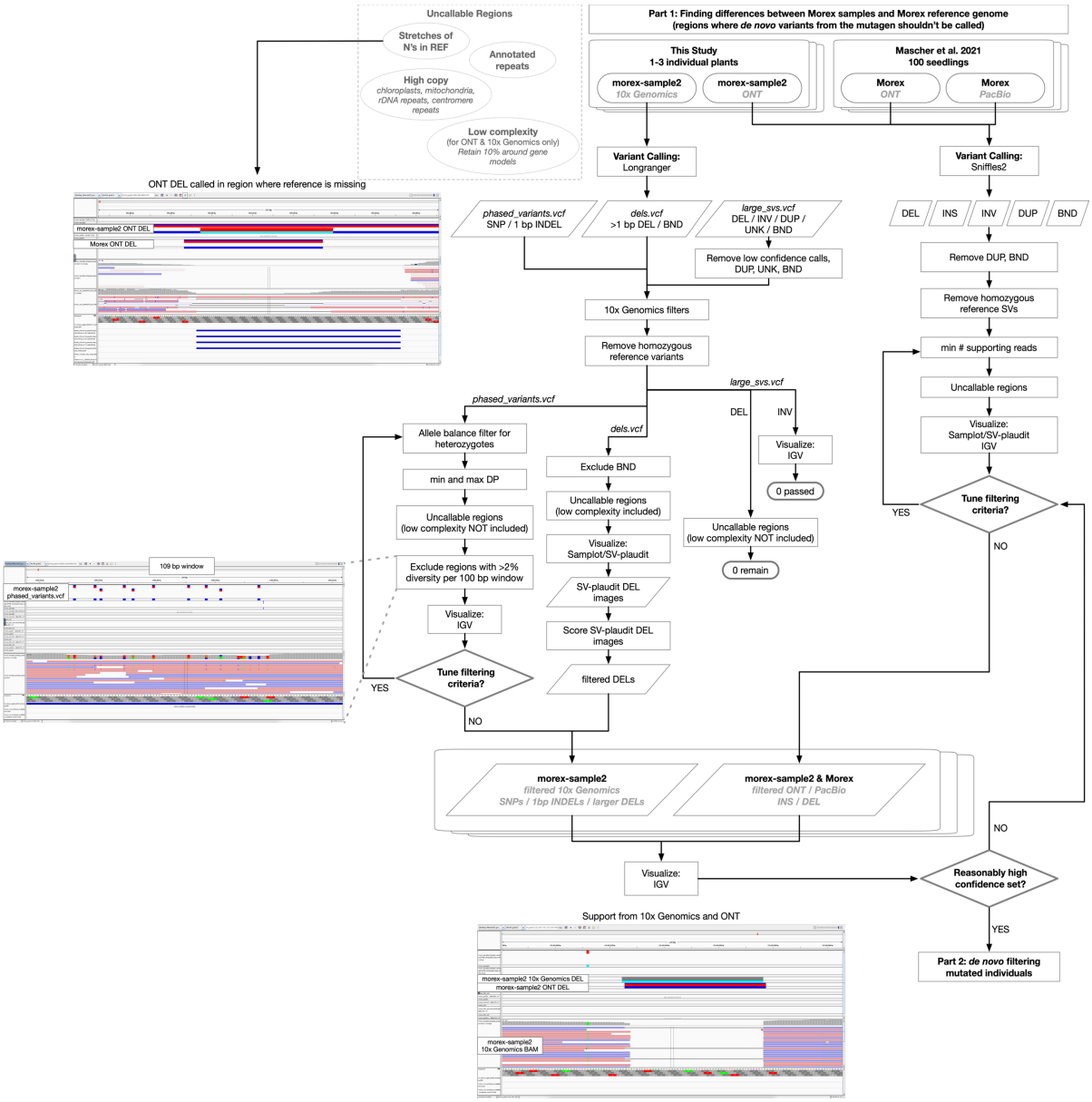

Figure S2. Flowchart of part 1 of variant filtering to identify differences between the Morex parent of the mutagenized lines and the Morex reference genome. These are regions where induced mutations should not be called as they are more likely to be variation among Morex individuals. Example screenshots from IGV are shown, from top to bottom: 1) A deletion called in ONT data in a region where the reference is missing - can't call a real deletion in these regions, 2) Regions with excessively high diversity based on the expectations in the most diverse barley genes are excluded, these are more likely alignment challenges than real variation, and 3) Read support for 10x Genomics called deletion with the same variant called in the ONT data providing additional support for the call.

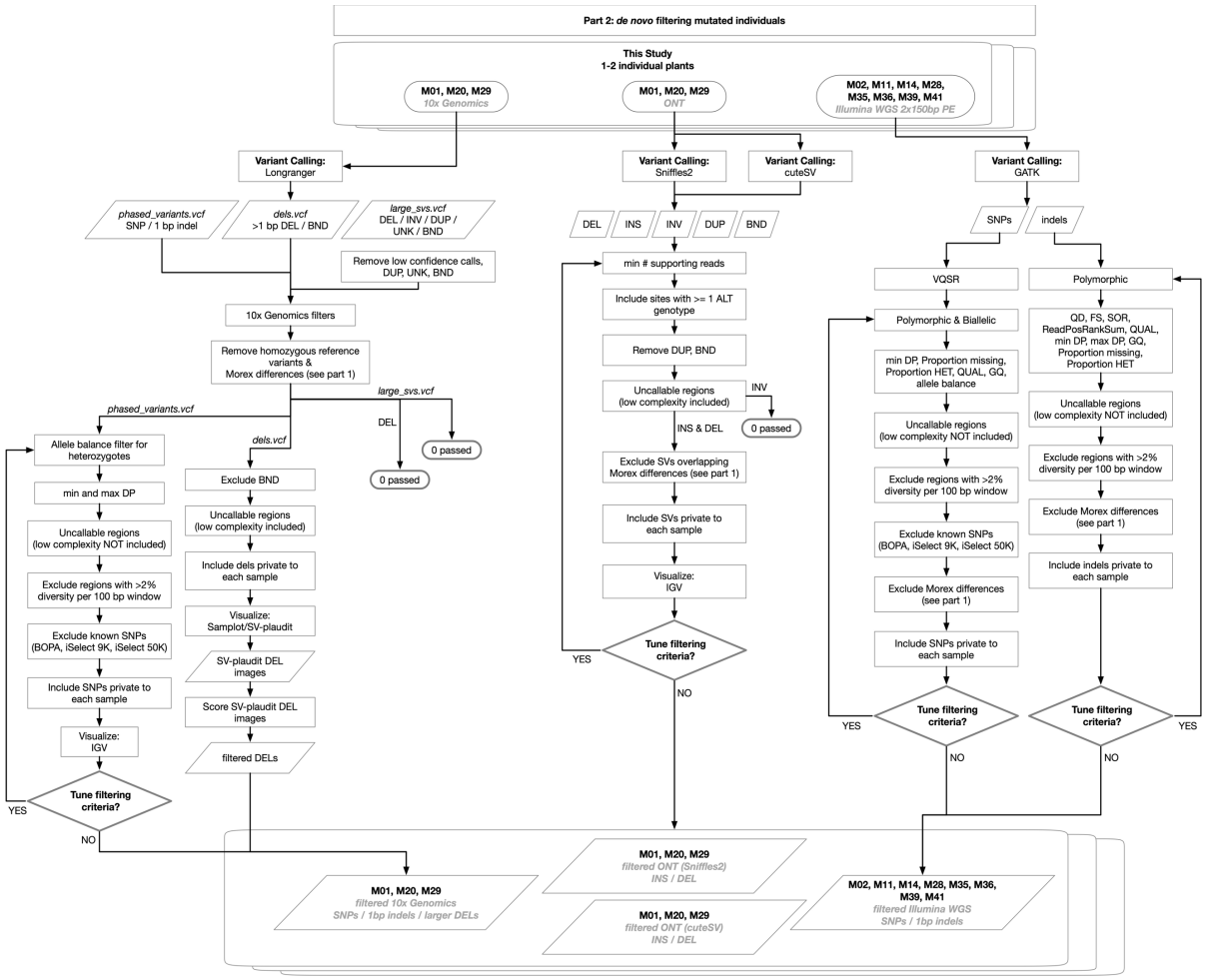

Figure S3. Flowchart of part 2 of variant filtering to identify *de novo* variants induced by sodium azide. Larger SVs were visually evaluated with a similar approach as in part 1 of the filtering (Figure S2).

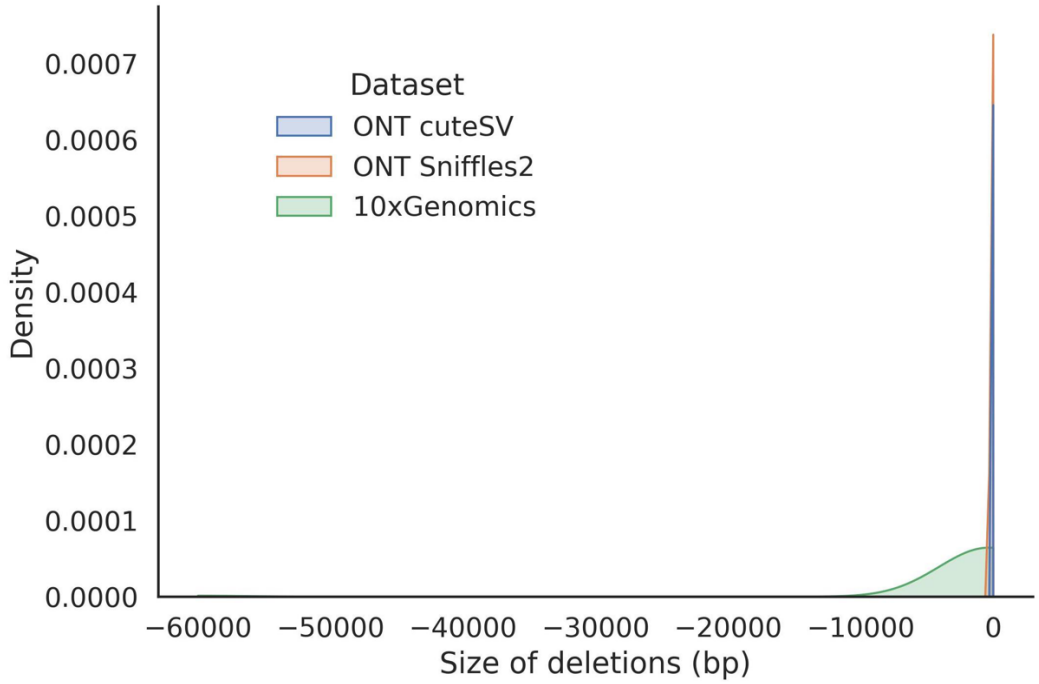

Figure S4. Distribution of larger deletion sizes in three mutagenized lines (M01, M20, and M29) called by 10x Genomics Longranger, Sniffles2 (ONT), and cuteSV (ONT).

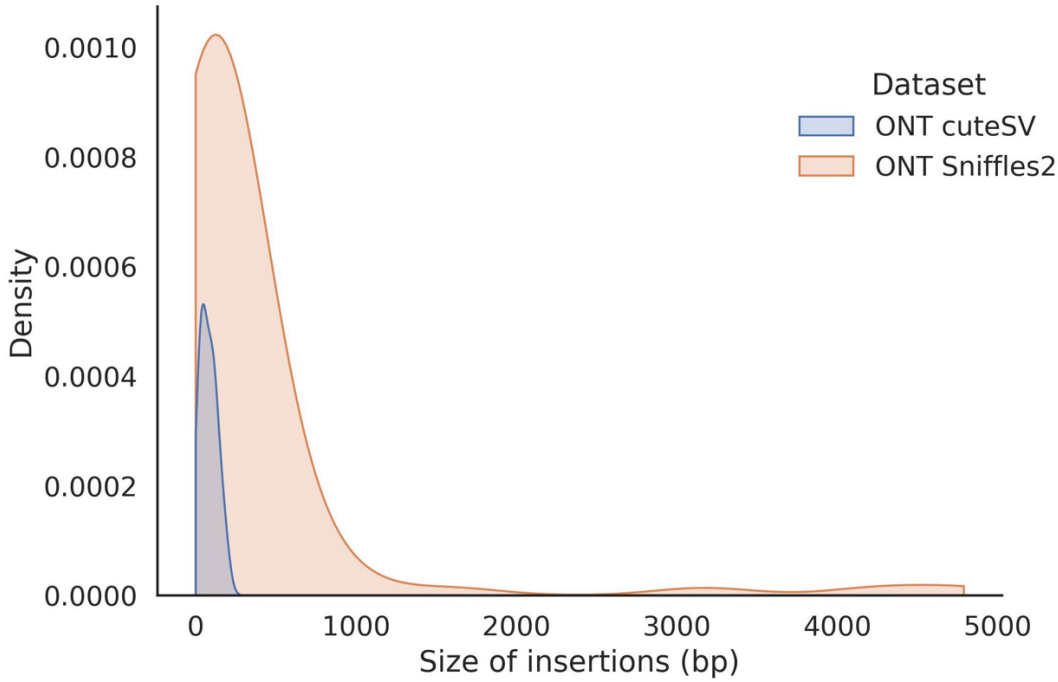

Figure S5. Distribution of larger insertion sizes in three mutagenized lines (M01, M20, and M29) called by Sniffles2 (ONT) and cuteSV (ONT). Larger insertions were not called in the 10x Genomics dataset.

Figure S6. A) Summary of the percentage and number of SNPs and indels in genic regions that can potentially disrupt genes for sodium azide treated samples, and rare vs common categories. B) Per sample breakdowns of the number and percentage of SNVs and indels that are in genic regions and can potentially disrupt genes.

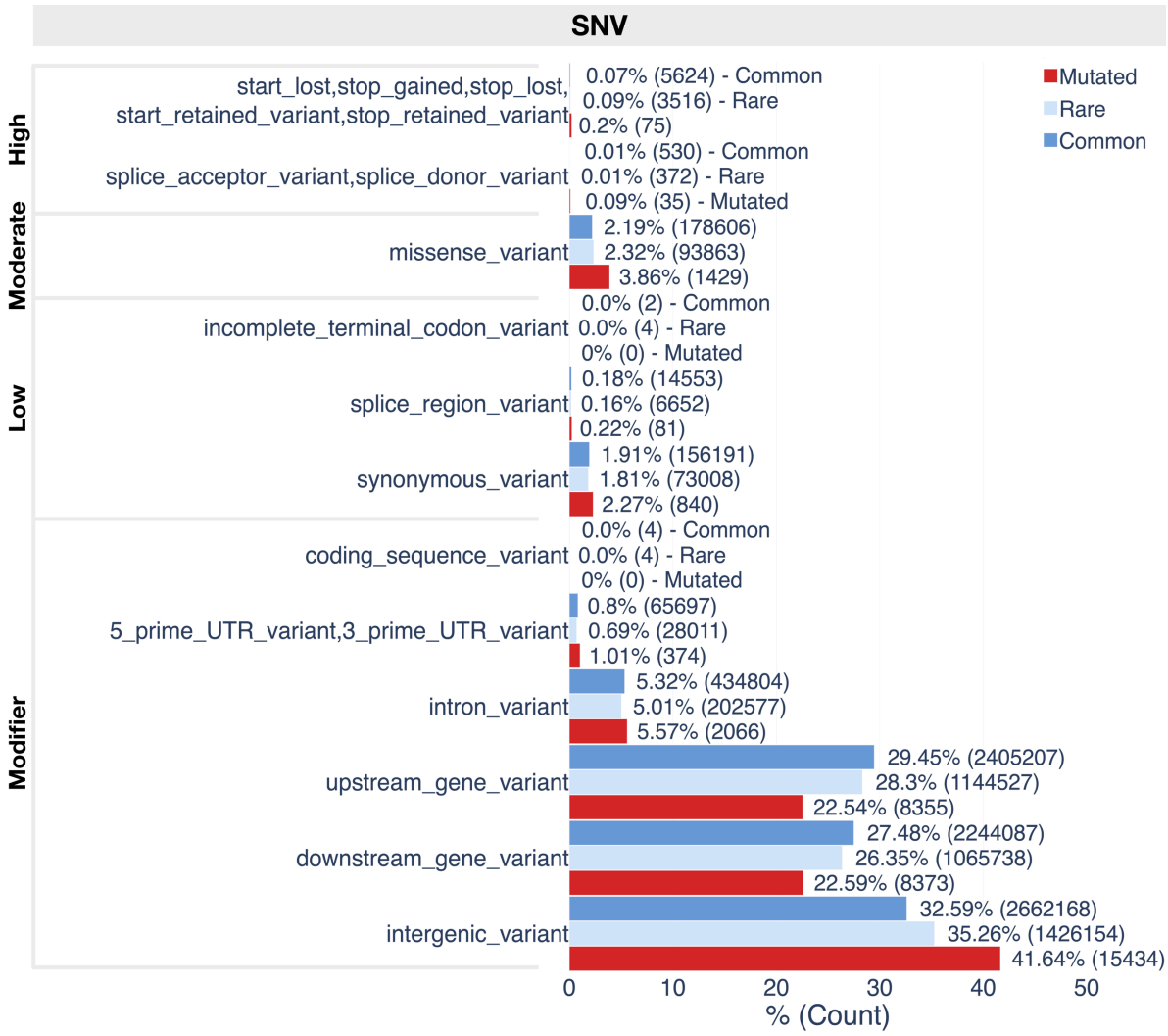

Figure S7. The functional effects of mutagenized, untreated rare, and untreated common SNVs/SNPs as annotated by VeP. Bars are labeled with the percentage and number of SNVs/SNPs in each consequence type. Boxes on the left side indicate the impact classification of the consequence type.

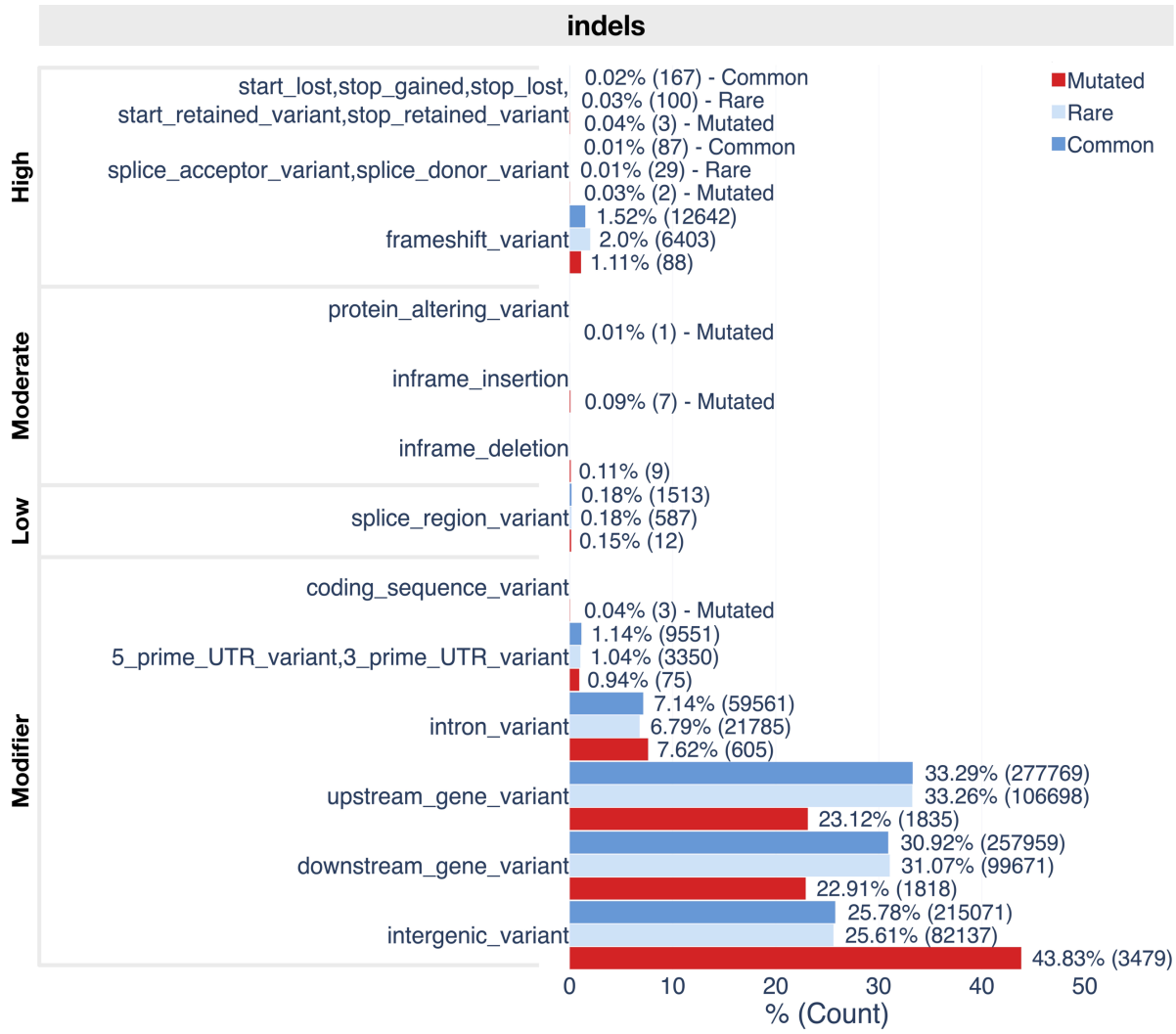

Figure S8. The functional effects of mutagenized, untreated rare, and untreated common indels as annotated by VeP. Bars are labeled with the percentage and number of indels in each consequence type. Boxes on the left side indicate the impact classification of the consequence type.

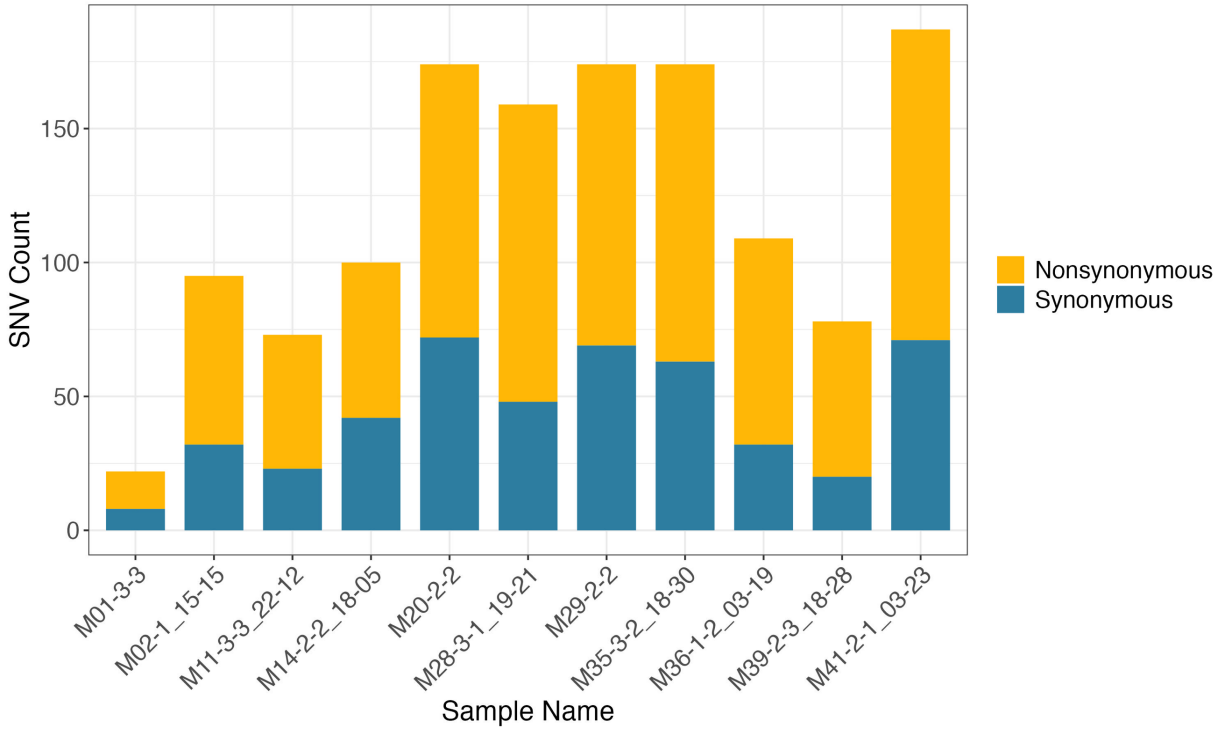

Figure S9. The number of nonsynonymous and synonymous SNVs in each mutated sample.

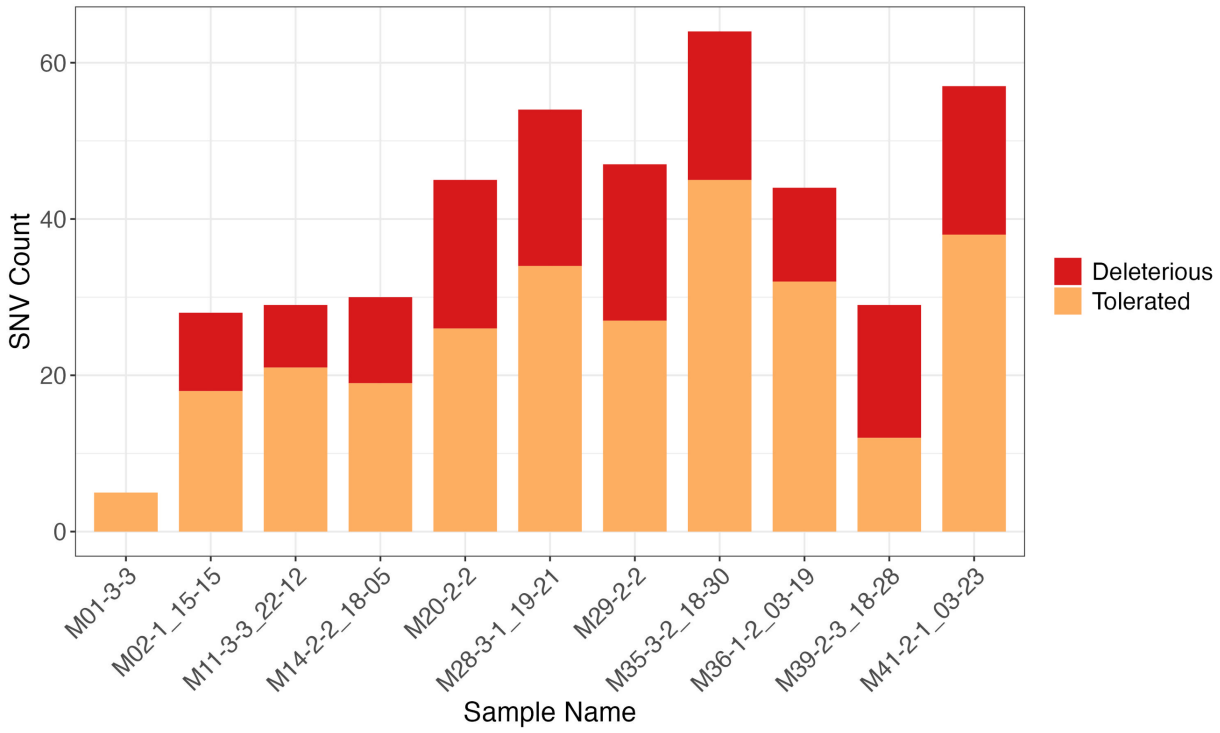

Figure S10. The number of nonsynonymous SNVs in each mutated sample partitioned into "Deleterious" and "Tolerated."

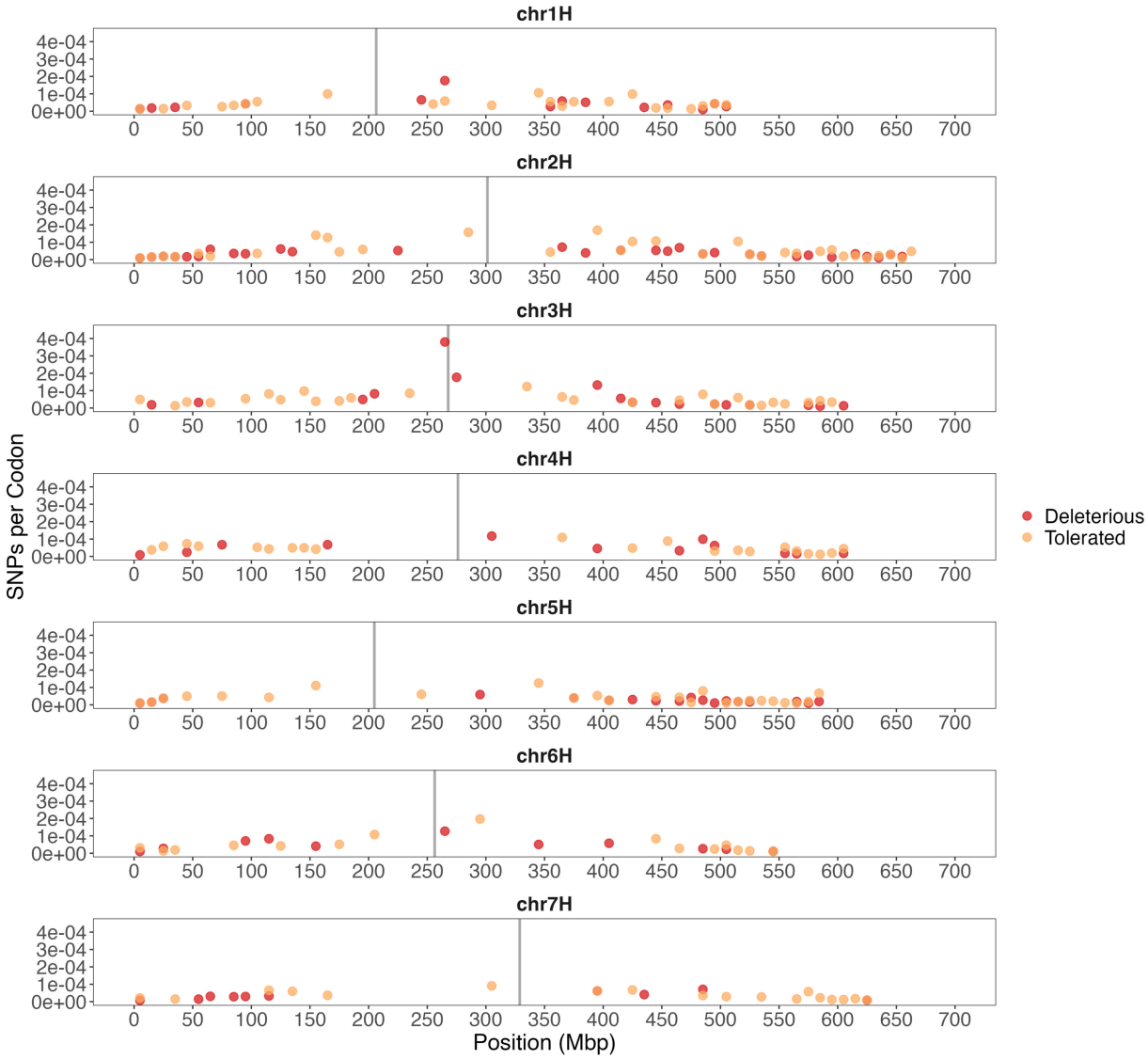

Figure S11. The number of nonsynonymous SNPs per covered codon in 10 Mb windows in mutated samples across the barley genome. Nonsynonymous SNPs are separated into "Deleterious" vs "Tolerated" and are plotted separately. Vertical grey line indicates the centromeric region.

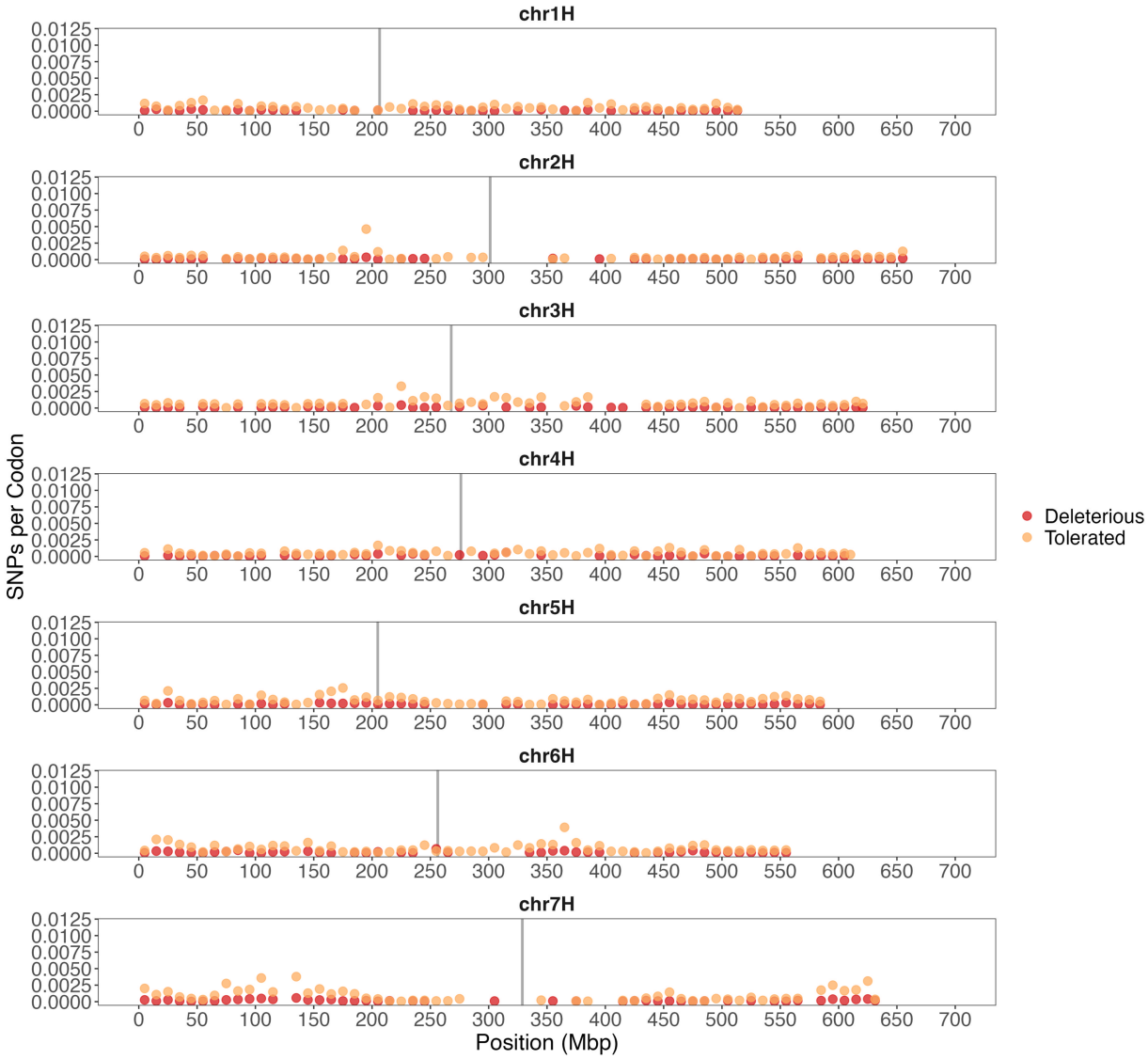

Figure S12. The number of nonsynonymous SNPs per covered codon in 10 Mb windows categorized as "rare" in the landrace samples across the barley genome. Nonsynonymous SNPs are separated into "Deleterious" vs "Tolerated" and are plotted separately. Vertical grey line indicates the centromeric region.

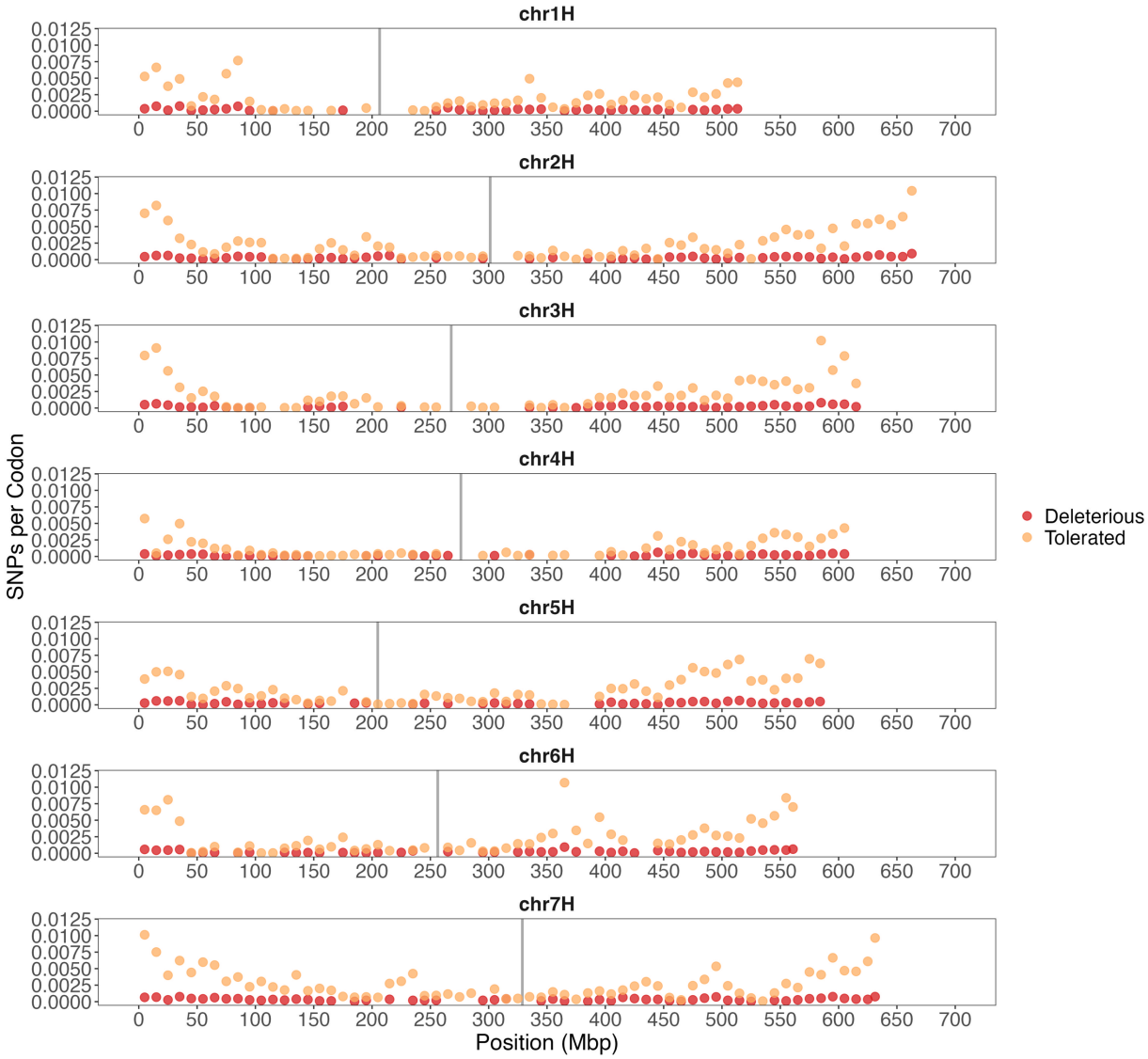

Figure S13. The number of nonsynonymous SNPs per covered codon in 10 Mb windows categorized as "common" in the landrace samples across the barley genome. Nonsynonymous SNPs are separated into "Deleterious" vs "Tolerated" and are plotted separately. Vertical grey line indicates the centromeric region.

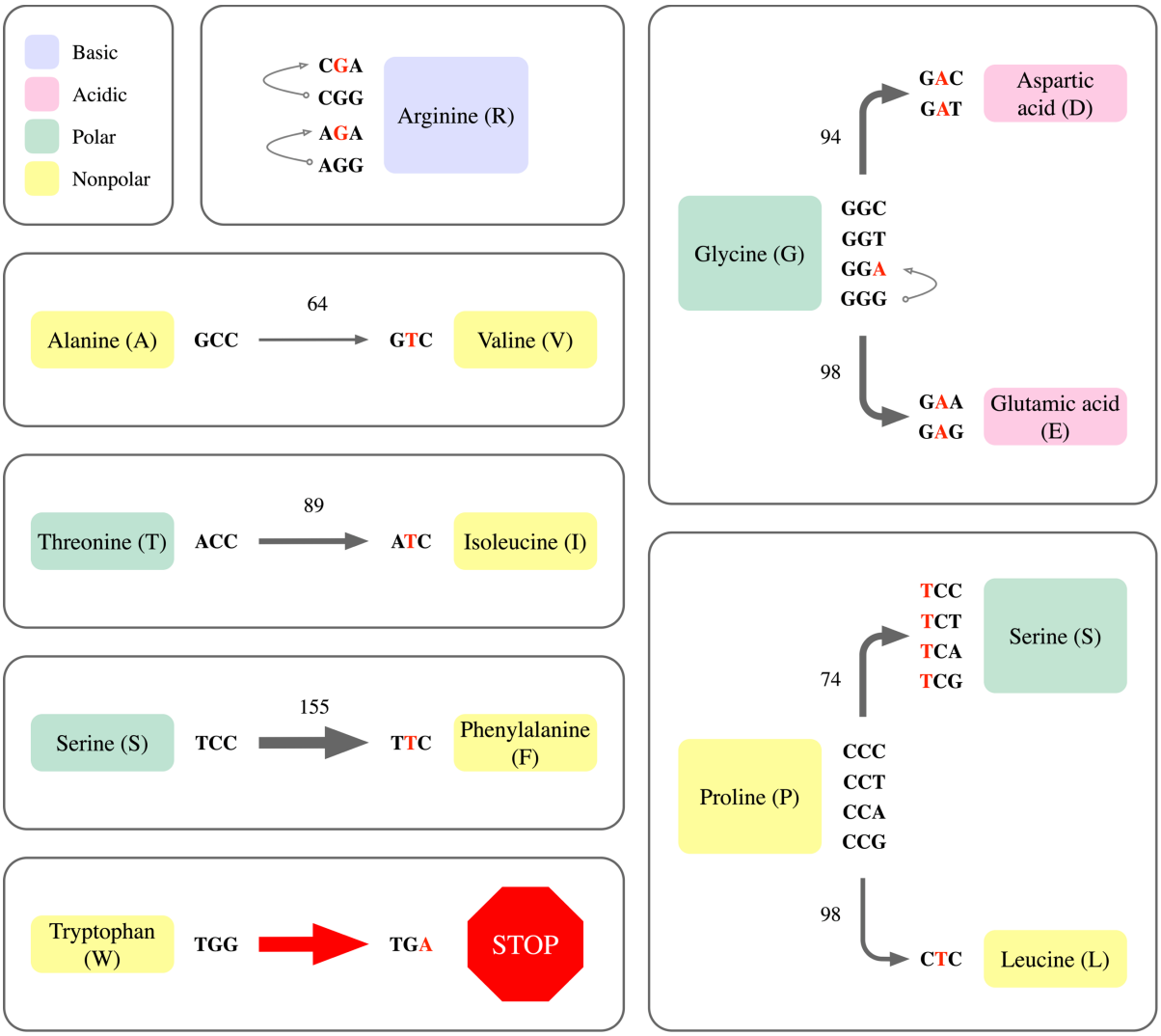

Figure S14. Representation of the mutations in the CC and GG motif that generate the most abundant amino acid changes identified as harmful in the SNVs. Colors represent the polarity of each amino acid. Arrows show the within codon changes, their line thickness and the side numbers indicate the correspondent Grantham score. A thicker line indicates a greater evolutionary distance between two amino acids. Red-letters represent the mutations induced by the sodium azide-associated motifs CC or reverse complement GG. Colors represent the primary properties of each amino acid including polarity and acidity.

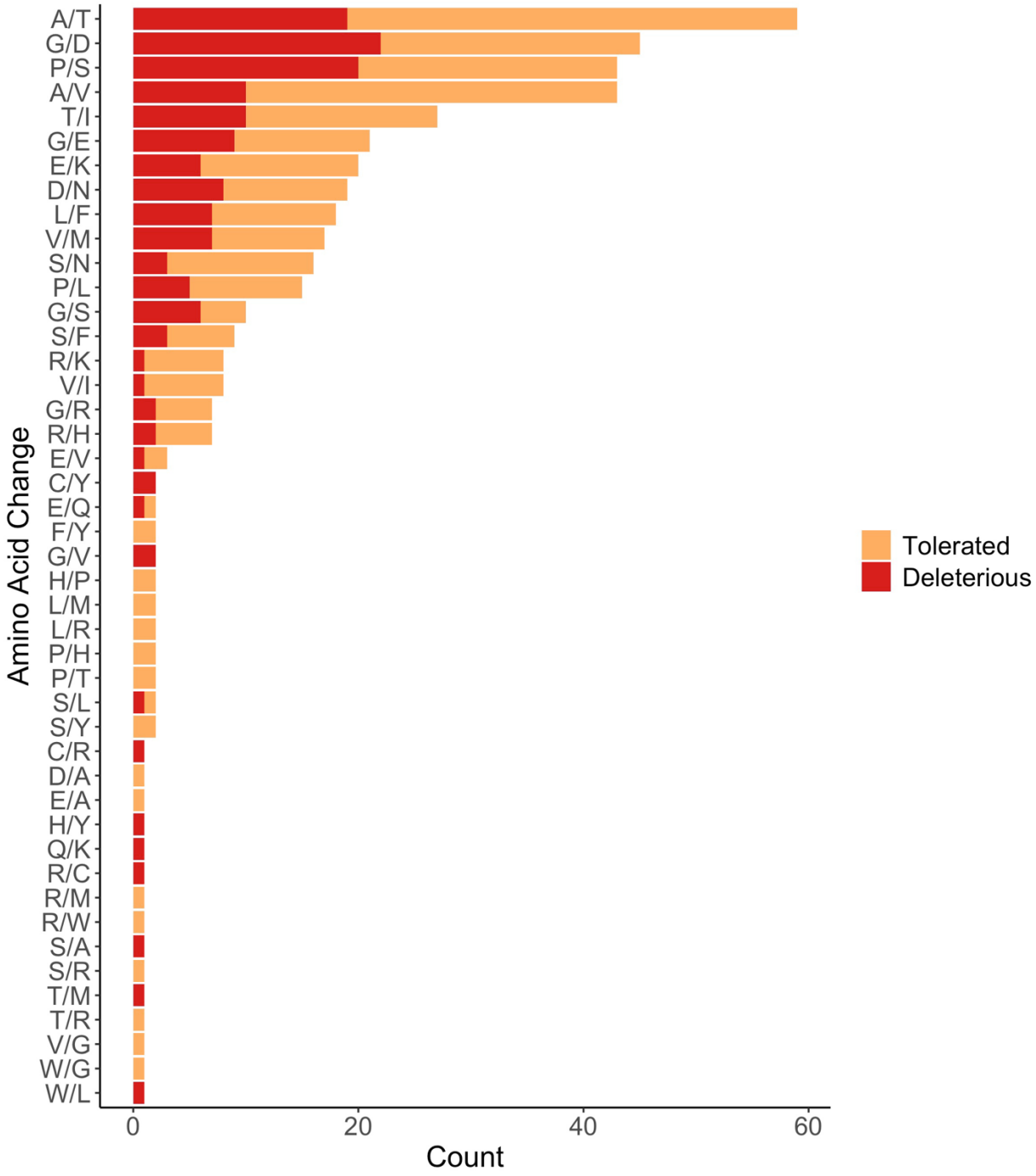

Figure S15. Frequency of amino acid changes for SNVs that annotate as tolerated versus deleterious in mutated lines.

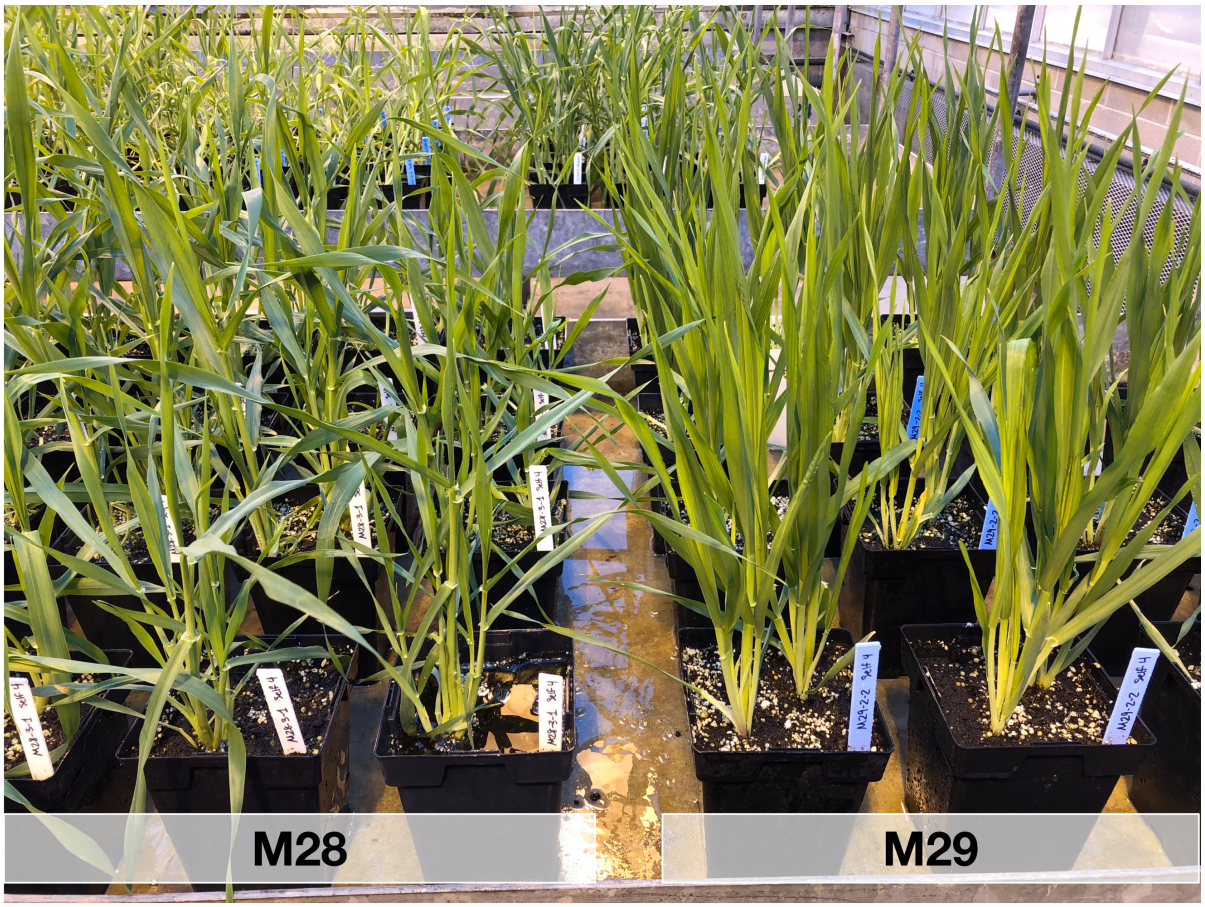

Figure S16. Picture of the mutagenized lines M28 next to M29, which has a distinct onion-like and compact phenotype that is atypical of barley. Both lines have been self-fertilized for four generations and photos were taken five weeks after planting.

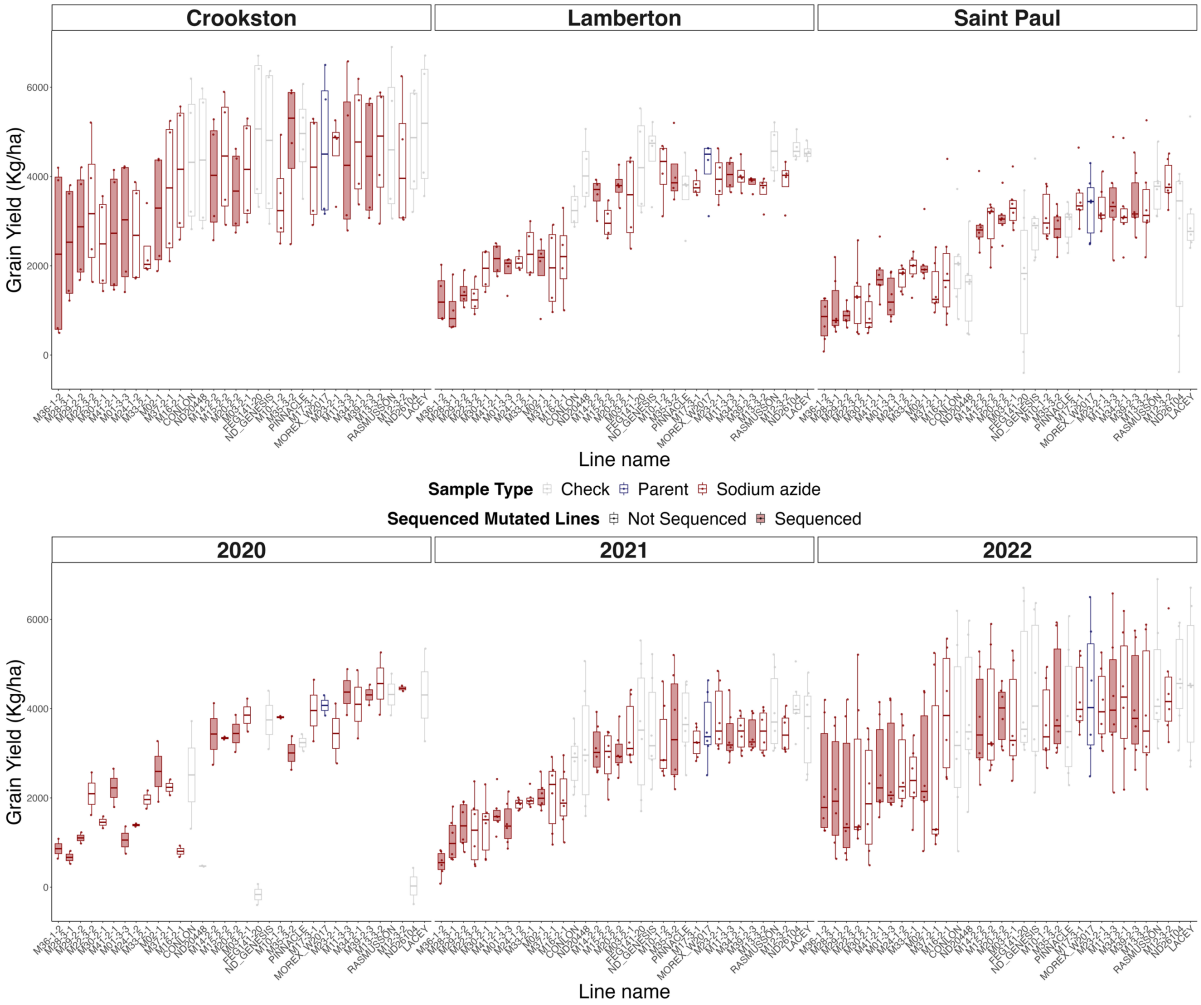

Figure S17. Grain yield for 25 mutated lines, the Morex W2017 parent, and eight check lines for three locations and three years. The box plots are sorted by the median for each line, and the bars in the box plot indicate the mean. Sodium azide-treated lines are represented by red outlines. Red shaded boxes indicate mutated lines that were sequenced in this study.

References

Bayer MM, Rapazote-Flores P, Ganal M et al. 2017. Development and evaluation of a barley 50k iSelect SNP array. Frontiers in Plant Science. 8:1792.

Belfield EJ, Gan X, Mithani A et al. 2012. Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of *Arabidopsis thaliana*. Genome Res. 22:1306-1315.

Belyeu JR, Nicholas TJ, Pedersen BS, Sasani TA, Havrilla JM, Kravitz SN, Conway ME, Lohman BK, Quinlan AR, Layer RM. 2018. SV-plaudit: a cloud-based framework for manually curating thousands of structural variants. GigaScience. 7:giy064.

Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, Layer RM. 2021. Samplot: a platform for structural variant visual validation and automated filtering. Genome Biol. 22:161.

Benegas G, Batra SS, Song YS. 2023. DNA language models are powerful predictors of genome-wide variant effects. Proceedings of the National Academy of Sciences. 120:e2311219120.

Bolon Y-T, Stec AO, Michno J-M et al. 2014. Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. Genetics. 198:967-981.

Boyko AR, Williamson SH, Indap AR et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLOS Genetics. 4:e1000083.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. Genome Res. 19:1553-1561.

Close TJ, Bhat PR, Lonardi S et al. 2009. Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics. 10:1-13.

Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, Shaw P, Bayer M, Thomas W, Marshall D. 2012. Natural variation in a homolog of Antirrhinum *CENTRORADIALIS* contributed to spring growth habit and environmental adaptation in cultivated barley. Nat Genet. 44:1388-1392.

Comadran J, Ramsay L, MacKenzie K, Hayes P, Close TJ, Muehlbauer G, Stein N, Waugh R. 2011. Patterns of polymorphism and linkage disequilibrium in cultivated barley. Theor Appl Genet. 122:523-531.

DePristo MA, Banks E, Poplin R et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491.

Döring HP, Lin J, Uhrig H, Salamini F. 1999. Clonal analysis of the development of the barley (*Hordeum vulgare* L.) leaf using periclinal chlorophyll chimeras. Planta. 207:335-342.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. Nature. 287:560-561.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 6:R44.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610-618.

FAO/IAEA. 2018. Manual on Mutation Breeding - Third Edition. Rome, Italy: Food and Agriculture Organization of the United Nations.

Nations FAAOOTU. 2018. Manual on Mutation Breeding Third Edition. Food & Agriculture Org.

Frank T. 2015. R package mvngGrAd: moving grid adjustment in plant breeding field trials.

Fu Y, Mahmoud M, Muraliraman VV, Sedlazeck FJ, Treangen TJ. 2021. Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment. GigaScience. 10:giab063.

Haun WJ, Hyten DL, Xu WW et al. 2011. The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol. 155:645-655.

Henry IM, Nagalakshmi U, Lieberman MC et al. 2014. Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. The Plant Cell. 26:1382-1397.

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 21:1-24.

Johnsson M, Gaynor RC, Jenko J, Gorjanc G, de Koning DJ, Hickey JM. 2019. Removal of alleles by genome editing (RAGE) against deleterious load. Genet Sel Evol. 51:14.

Kleinhofs A, Kilian A, Maroof S et al. 1993. A molecular, isozyme and morphological map of the barley (*Hordeum vulgare*) genome. Theor Appl Genet. 86:705-712.

Kleinhofs A, Owais WM, Nilan RA. 1978. Azide. Mutation Research/Reviews in Genetic Toxicology. 55:165-195.

Knudsen S, Wendt T, Dockter C et al. 2022. FIND-IT: Accelerated trait development for a green evolution. Sci Adv. 8:eabq2266.

Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. 2016. The role of deleterious substitutions in crop genomes. Mol Biol Evol. 33:2307-2317.

Kono TJY, Lei L, Shih C-H, Hoffman PJ, Morrell PL, Fay JC. 2018. Comparative genomics approaches accurately predict deleterious variants in plants. G3 (Bethesda). 8:3321-3329.

Leger A, Leonardi T. 2019. pycoQC, interactive quality control for Oxford Nanopore Sequencing. Journal of Open Source Software. 4:1236.

Li G, Jain R, Chern M et al. 2017. The sequences of 1504 mutants in the model rice variety Kitaake facilitate rapid functional genomic studies. The Plant Cell. 29:1218-1231.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094-3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078-2079.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 22:126-131.

Mascher M, Wicker T, Jenkins J et al. 2021. Long-read sequence assembly: a technical evaluation in barley. Plant Cell. 33:1888-1906.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. Genome Biol. 17:122.

Michno J-M, Stupar RM. 2018. The importance of genotype identity, genetic heterogeneity, and bioinformatic handling for properly assessing genomic variation in transgenic plants. BMC Biotechnol. 18:38.

Morrell PL, Buckler ES, Ross-Ibarra J. 2012. Crop genomics: advances and applications. Nat Rev Genet. 13:85.

Morrell PL, Gonzales AM, Meyer KKT, Clegg MT. 2014. Resequencing data indicate a modest effect of domestication on diversity in barley: a cultigen with multiple origins. J Hered. 105:253-264.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2006. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. Genetics. 173:1705-1723.

Morton BR, Bi IV, McMullen MD, Gaut BS. 2006. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics. 172:569-577.

Moyers BT, Morrell PL, McKay JK. 2018. Genetic Costs of Domestication and Improvement. J Hered. 109:103-116.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31:3812-3814.

Olsen O, Wang X, von Wettstein D. 1993. Sodium azide mutagenesis: preferential generation of AT--> GC transitions in the barley Ant18 gene. Proc Natl Acad Sci USA. 90:8043-8047.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 327:92-94.

Owais WM, Kleinhofs A. 1988. Metabolic activation of the mutagen azide in biological systems. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 197:313-323.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 34:867-868.

Plekhanova E, Nuzhdin SV, Utkin LV, Samsonova MG. 2018. Prediction of deleterious mutations in coding regions of mammals with transfer learning. Evolutionary Applications.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276-277.

Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, Jean PLS, Abecasis GR, Novembre J, Zöllner S, Li JZ. 2013. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res. 23:1974-1984.

Schneeberger K. 2014. Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet. 15:662-676.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods. 15:461-468.

Sim SB, Corpuz RL, Simmonds TJ, Geib SM. 2022. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics. 23:1-7.

Smith KP, Thomas W, Gutierrez L, Bull H. 2018. Genomics-Based Barley Breeding. The Barley Genome. Springer. p. 287-315.

Smolka M, Paulin LF, Grochowski CM, Mahmoud M, Behera S, Gandhi M, Hong K, Pehlivan D, Scholz SW, Carvalho CMB. 2022. Comprehensive structural variant detection: from mosaic to population-level. BioRxiv. 2022.04. 04.487055.

Sommer L, Spiller M, Stiewe G, Pillen K, Reif JC, Schulthess AW. 2020. Proof of concept to unmask the breeding value of genetic resources of barley (*Hordeum vulgare*) with a hybrid strategy. Plant Breed. 139:536-549.

Sunyaev S, Ramensky V, Koch I, Iii L, Warren, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. Hum Mol Genet. 10:591-597.

Talamè V, Bovina R, Sanguineti MC, Tuberosa R, Lundqvist U, Salvi S. 2008. TILLMore, a resource for the discovery of chemically induced mutants in barley. Plant Biotechnology Journal. 6:477-485.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 14:178-192.

Van der Auwera GA, O'Connor BD. 2020. Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media.

Wallace JG, Rodgers-Melnick E, Buckler ES. 2018. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. Annu Rev Genet. 52:421-444.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics. 3

Wyant SR, Rodriguez MF, Carter CK, Parrott WA, Jackson SA, Stupar RM, Morrell PL. 2022. Fast neutron mutagenesis in soybean enriches for small indels and creates frameshift mutations. G3 (Bethesda). 12:jkab431.

Zhu Y, Neeman T, Yap VB, Huttley GA. 2017. Statistical methods for identifying sequence motifs affecting point mutations. Genetics. 205:843-856.

Zhu Y, Ong CS, Huttley GA. 2020. Machine learning techniques for classifying the mutagenic origins of point mutations. Genetics. 215:25-40.