Review Article

Proposed EU NGT legislation in light of plant genetic variation

Alan H. Schulman[1,2*], Frank Hartung[3†], Marinus J.M. Smulders[4†], Jens F. Sundström[5†], Ralf Wilhelm[3†], Odd Arne Rognli[6†], and Karin Metzlaff[7]

[1]HiLife Institute of Biotechnology and Viikki Plant Science Centre (ViPS), University of Helsinki, Finland; alan.schulman@helsinki.fi

[2]Production Systems, Natural Resources Institute Finland (LUKE), Helsinki, Finland; alan.schulman@luke.fi

[3]Julius Kuehn Institute (JKI) - Federal Research Centre for Cultivated Plants, Institute for Biosafety in Plant Biotechnology, Quedlinburg, Saxony-Anhalt, Germany; frank.hartung@julius-kuehn.de; ralf.wilhelm@julius-kuehn.de

[4]Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands; rene.smulders@wur.nl

[5]Department of Plant Biology, Swedish University of Agricultural Science, The Linnean Centre for Plant Biology, Box 7080, SE-75007 Uppsala, Sweden; jens.sundstrom@slu.se

[6]Faculty of Biosciences, Department of Plant Sciences, Norwegian University of Life Sciences (NMBU), Ås, Norway; odd-arne.rognli@nmbu.no

[7]European Plant Science Organisation (EPSO), Brussels, Belgium; karin.metzlaff@epsomail.org

*Correspondence: (Tel +358 407682242; email alan.schulman@helsinki.fi)

[†]These authors contributed equally to this work

Key words: gene editing; new genomic techniques (NGT); new breeding techniques (NBT); plant genetic diversity; mutagenesis; CRISPR/Cas9; plant breeding

Summary

The European Commission (EC) proposal for New Genomic Techniques (NGTs) of July 2023 specifies that Category 1 NGT (NGT1) plants, which are considered equivalent to conventional plants, i.e. those obtainable by conventional plant breeding or mutagenesis, may differ from the recipient or parental plant by no more than 20 insertions, which cannot be longer than 20 bp; deletions can be no more than 20 but of any size. Here, we examine the proposed 20/20 NGT1 limit against the background of the theoretical considerations and older data used to frame it and in light of recent data from highly contiguous long-read assemblies for reference genomes and pangenomes. We find that current genomic data indicate that natural variation in germplasm used by breeders is much greater than earlier understood and that both conventional breeding and mutagenesis can introduce genomic changes that are both more extensive in size and more frequent than the NGT Category 1 "20 insertions of maximum 20 bp" limit would allow. Furthermore, natural variation also scales with genome size and complexity, a factor not considered in the EC proposal. We conclude that the proposed cutoffs under which an NGT plant is considered equivalent to conventional plants do not align with what is observed in nature, conventional breeding, and mutagenesis. Updating the 20/20 rule to broader limits would facilitate breeding for climate resilience, farming sustainability, and nutritional security, while ensuring that NGT1 plants are equivalent to conventional ones.

Introduction

Annex 1 to the European Commission (EC) proposal (2023/0226) on New Genomic Techniques (NGTs) specifies the number and types of changes that would be regarded as equivalent to the variation found in conventional plants, i.e. those that "could have been produced through conventional plant breeding or classical mutagenesis" (as stated in COM(2023) 411 final). Plants within these bounds are regarded as NGT Category 1 (NGT1). With the rapid progress in genomic sequencing methods, our understanding of plant genomic variation is improving quickly in parallel. Here, we consider what is known about variation between plant genomes and to what extent the proposed NGT1 category reflects that. The EC also defines a Category 2 (NGT2), which includes all NGT modifications that exceed the limits specified in NGT1. The traceability and labelling regime currently in place for GMOs under Directive 2001/18/EC would be adapted to NGT2, likely reducing commercial interest in introducing NGT2 products into the market. As a counterbalance, the EC proposes regulatory incentives under Annex III for use of NGT2 for various sustainability traits, including disease resistance, abiotic stress tolerance, and nutrient and water use efficiency; herbicide tolerance is explicitly excepted. Given that the line between NGT1 and NGT2 will crucially affect the use and commercialization of NGTs, we will restrict our focus to the EC proposal and to the relevant research underlying it (Figure 1).

From Annex I:

"An NGT plant is considered equivalent to conventional plants when it differs from the recipient/parental plant by no more than [20] genetic modifications of the types referred to in points 1 to 5, in predictable DNA sequences. A predictable DNA sequence is any DNA sequence that shares sequence similarity with the targeted site."

(1) Substitution or insertion of no more than (20) nucleotides ;

(2) Deletion of any number of nucleotides;

(3) On the condition that the genetic modification does not result in an intragenic plant:

- (a) Targeted insertion of a contiguous DNA sequence existing in the breeder's gene pool;
- (b) Targeted substitution of an endogenous DNA sequence with a contiguous DNA sequence existing in the breeder's gene pool;

(4) Targeted inversion of a sequence of any number of nucleotides;

(5) Any other targeted modification of any size, on the condition that the resulting DNA sequences already occur (possibly with modifications as accepted under points 1 and/or 2) in a species from the breeders' gene pool.

Figure 1 Excerpt from Annex I, "Criteria of equivalence of NGT plants to conventional plants," to "Proposal for a Regulation of the European Parliament and of the Council on plants obtained by certain new genomic techniques and their food and feed, and amending Regulation (EU) 2017/625," of 5 July 2023.

Annex I sets a very specific standard for insertions within the NGT1 category. This can be interpreted as being consistent with the original 2001/18/EC legislation on the deliberate release into the environment of genetically modified organisms, where Article 2(2) specifies a GMO as one in which the genetic material has been altered in a way that *"does not occur naturally* by mating and/or natural recombination.*"* Likewise, the proposed restriction on the number of changes appears to respond to the 2018 ECJ Curia judgement (ECLI:EU:C:2018:583), which states according to the referring court that, "the new techniques of mutagenesis allows the production of modifications … *at a rate out of all proportion to the modifications likely to occur naturally or randomly…*", implying a resulting safety risk.

Standards of "naturalness" and "conventional" beg the question of what is found in nature. In the European Commission's document 14204/23 (Commission Services, 2023), "Regulation on new genomic techniques (NGT) – Technical paper on the rationale for the equivalence criteria in Annex I", the criteria are based on a literature analysis of 90 scientific, peer-reviewed original studies, which are cited in the annex of that document. The cited EFSA study on site-directed mutagenesis, however, is from 2012 (EFSA GMO Panel, 2012), which is well before the advent of the current state of the art. The EFSA risk assessment studies in 2020 (EFSA GMO Panel *et al*., 2020) and 2022 (EFSA GMO Panel *et al*., 2022) did not revisit the state of knowledge of genome structural variation, either natural or that induced by conventional mutagenesis. Virtually all of the 90 papers that support the proposed standards were based on research from before long-read sequencing. This recently available approach has greatly increased the contiguity and completeness of genome assemblies – akin to reproduction of manuscripts without missing punctuations or words, or misplaced sentences and paragraphs – and has thereby improved the detection of insertions and deletions ("indels", when taken together), chromosomal rearrangements, and both presence-absence and copy-number variations in gene families.

Critically, the true dynamic nature of the genome could not be resolved by the methods available before 2012, and in fact not before the advent of the PacBio HiFi long-read sequencing method in 2022, complemented today with e.g. Nanopore technology. Indeed, the 14204/23 document anticipates its own obsolescence, stating that "…improvement of detection methods (i.e. long-read sequencing) has started to unveil higher rates than previously estimated" for genomic changes larger than single-nucleotide polymorphisms. The PacBio HiFi method, for example, produces read lengths of 10 to 25 kb at accuracies of 99.5% or higher (Hon *et al.*, 2020). Long reads span retrotransposons, which can be 5 to 15 kb and even 30 kb in length, whose high copy number and conservation interfere with unique assignment and assembly of reads into genomes (Jayakodi *et al.,* 2023). To aid in correct orientation and assembly of long reads into chromosome-scale scaffolds that bridge the

remaining sequence gaps, the Chromosome Conformation Capture (Hi-C) sequencing method has come into use (Himmelbach *et al.,* 2018). The Hi-C method ligates together segments of DNA that are on proximal DNA strands in chromatin, thereby helping to establish the linear order of long sequence reads in scaffolds and also enabling phasing of haplotypes (Sun *et al.*, 2025). Hi-C can be supplemented as well by optical mapping (Paajanen *et al.,* 2019; Jayakodi *et al.,* 2023) to further improve long-read assemblies.

Insertion and deletion sizes in plant genomes vs the 20 bp limit

A key restriction in 14204/23 is the 20 bp insertion limit for NGT1. A likely rationale for the limit is to distinguish short, random repair-type insertions from long insertions that can be identified as unique or specific genomic constituents, i.e., equivalent to cisgenes. There are two components to this rationale: first, the assumption that natural "random" insertions are short; second, that insertions longer than 20 bp can be uniquely identified as pre-existing in the genome (i.e., in the "breeders' gene pool"). As is stated in the report, "*Insertions of more random sequences are typically of a length of less than ten nucleotides but have been observed to extend to approximately fifty nucleotides*" and that "…*a threshold of twenty nucleotides in criterion 1 for substitutions and insertions was set since it fits with the sizes observed in the scientific analysis*."

The first question one can raise is: What is the actual size distribution of spontaneous insertions and deletions in conventional plants, compared with the 20 bp limit of NGT1? The answer is that recent advances in genome sequencing show that natural variations extend from 1 to 1 million bp. Although the proposed regulations distinguish between insertions and deletions, in practice it is seldom possible to determine the initial state in accessions of cultivars or wild materials, i.e. whether it was a spontaneous insertion or deletion that occurred to distinguish the versions of a sequence. Hence, the term "indel" is used collectively for insertions and deletions; very often a particular "complex" indel will contain a combination of both. While the Commission draft proposal distinguishes between the legal status of insertions (of fixed number and size for NGT1) and deletions (of any size or number), the data from the actual natural world does not support this distinction (Figures 1—3). For indels found in sequenced genomes, even early (2013) data from long genome assemblies showed no bimodal distribution expected for short, random insertions and longer gene-like (or transposon-like) insertions within eukaryotes overall (Figure 2).

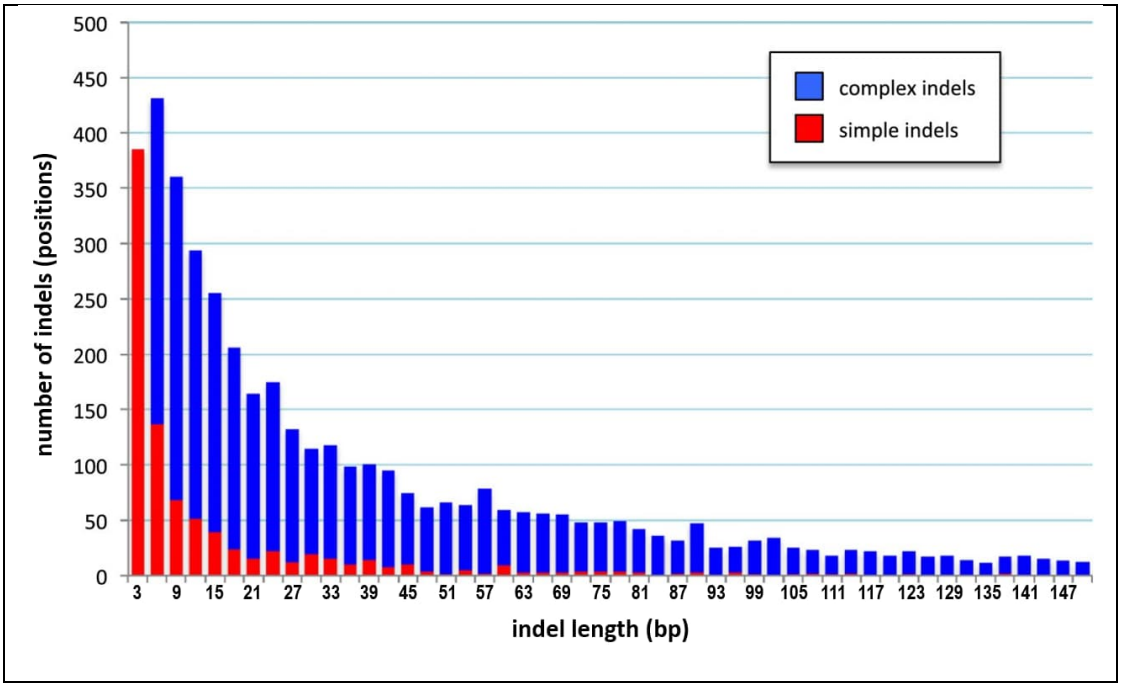
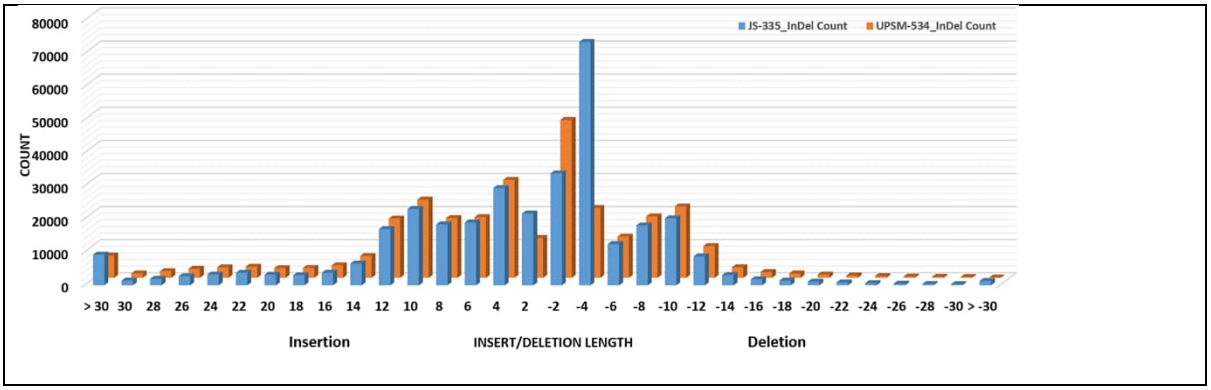


Figure 2 For each indel size class (x-axis), the number of simple (total = 901) and complex (total = 3,806) indels are indicated by the red and blue bars, respectively. 501 indels (10 simple indels and 491 complex indels) longer than 50 amino acid residues are not shown. Simple indels" occur in only two states, present or absent, and are potentially the result of a single indel event, while "complex indels" occur in two or more states and represent multiple compounded indel events. Modified from Ajawatanawong and Baldauf (2013).

"Complex" indels (Figure 2), which are the likely result of multiple, nested events over time, show no sharp decline and, as compound events, are *ipso facto* not structurally equivalent to cisgenes. In rice, indel markers varied from 3 to 39 bp, with 88.2% 6—25 bp, 6.2% ≤ 5 bp, and 5.6% were ≥ 26 bp (Zeng *et al*., 2013). Work from 2015 in soybean (Figure 3) with the older short-read technologies indicated a rapid drop-off in indel length, consistent with the earlier EFSA studies. Nevertheless, analyses of individual gene families (e.g., *RPB2* in barley; Sun *et al*. (2009)), where alignments were carefully constructed for the genes), indicated that indels of 20 – 100 bp are quite common; MITE transposons, which are abundant, are 90 – 100 bp.

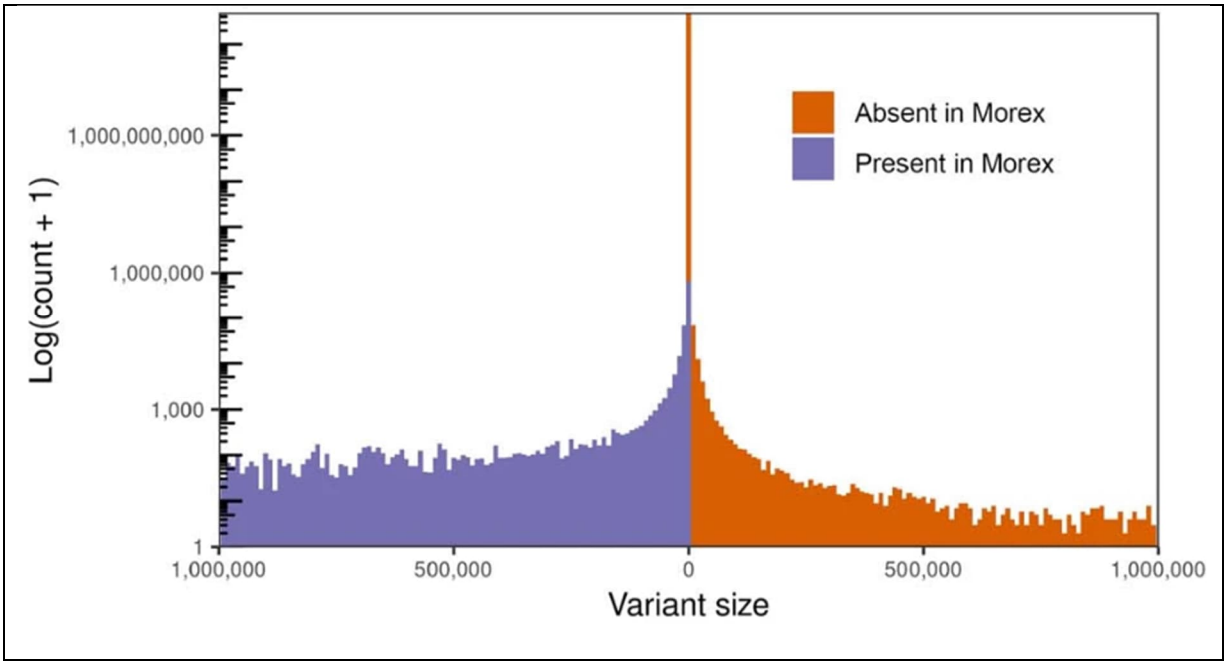


157 Figure 3 Frequency of length distribution of indels between soybean cultivars JS-335 and UPSM-534. From
158 Yadav *et al*. (2015).

159

160 Critically, it is has become clear that the apparent indel length distribution can be influenced by the
161 limits of alignment and assembly of short-read ("Illumina") sequences; a better picture is now
162 emerging from long-read (PacBio HiFi and Nanopore) sequencing approaches as anticipated but not
163 yet documented by research by EFSA in the 14204/23 technical paper. A striking example is the
164 distribution of indels and presence-absence variations (PAVs) between two well-assembled barley
165 cultivars (Figure 4). Another example was published in 2024 for lentil, *Lens culinaris* (Shivaprasad *et*
166 *al*., 2024). These researchers compared the genomes of a lentil parental line with recombinant
167 inbred bulks, finding almost 735 000 indels, of which almost 16 000 were longer than 20 bp, 3600
168 greater than 40 bp, and 1200 greater than 50 bp.

169



170 Figure 4 Size distribution of PAVs between Morex and Barke cultivars. Extended Data Fig. 6 in Jayakodi *et al*.
171 (2020).

Distribution of indels generated by break repair following intentional mutagenesis

The question of indel size distribution is relevant because the criteria of naturalness and pre-2001 methodology are used to set the outer limits for GE acceptable as "conventional-like," i.e. NGT1. Conventional breeding methods – including mutagenesis methods taken into use pre-2001 – are not subject to the 2001/18 regulatory regime and hence are worth comparing with the outcomes of NGT methods, which are subject to 2001/18 and considered in 14204/23. Ion-beam mutagenesis is one frequently used method (Guo *et al*., 2024). A recent study in Arabidopsis demonstrated that insertions generated by repair of the double-strand DNA breaks induced by ion-beam irradiation of seedlings ranged from less than 5 bp to over 100 bp, with an average of ~ 12 bp (Kitamura *et al*., 2024). In contrast, data shows that when the early CRISPR/*Cas9* method (SDN-1), which causes double-strand breaks, is used to knock-out gene function, the breaks are precisely repaired 36—41% of the time, the remainder not (Ben-Tov *et al*., 2024). In a study of 361 CRISPR/*Cas9*-mutated plants, the imprecise break repairs were predominantly short insertions or deletions; 87% of the induced indels were smaller than 10 bp (Zhang *et al*., 2020). Insertions comprised 30% of the total; 73% of the insertions were only 1 bp in length, 2% were 2—50 bp, and 6% > 50 bp.

In many cases, it will be necessary to introduce changes at the mutation site that preserve gene function rather than knocking it out by a deletion or by the repair process of the cell that can generate small random insertions. The currently most popular targeted (NGT) mutagenesis method, CRISPR/*Cas9*, generates distinctly smaller break-repair insertions (~1—10 bp) when used for knockouts than does either conventional mutagenesis or natural processes. Hence, while the Curia judgement of 2018 viewed the genetic changed wrought by new mutagenic techniques as far in excess of those occurring naturally or by earlier-established methods, the available data shows that the opposite is the case: NGT methods are therefore considerably gentler in their genomic impact than traditional breeding approaches, whether crossing or random mutagenesis.

Minimum length needed to specify a unique sequence in the genome

To address the need for practical monitoring under NGT regulatory regimes, an alternative approach for defining a maximum insertion length acceptable as NGT1 is that it should be below the minimum identifiable unique sequence in a genome, hence it should be one that could result from a random process. Report 14204/23 posits that, *"…when considering genome diversity, the JRC calculated that the theoretical probability that a random sequence is unique in the genome of various crops boils down to a consistent relatively narrow size range between 19 and 21 bases."* As justification for this

claim, it cites Broothaerts *et al.* (2021). However, this publication (section 4.4, p. 20), has no explanation given for the claim; only undescribed and unpublished results from rice are cited.

One conceivable explanation is that 20 bp length is based on a mathematical calculation. Assuming that the four nucleotides (A, C, G, T) occur at equal frequency, the likelihood of occurrence of any arbitrary nucleotide sequence of length bp in a genome of random nucleotides is $1/4^{bp}$. Hence, a 19mer would have a frequency of 3.6×10^{-12} bp; a 20mer, 0.9×10^{-12} bp; a 21mer, 2.3×10^{-13}. One of the largest crop genomes known is that of faba bean (*Vicia faba*), where the basic set of chromosomes (monoploid genome) comprises 13×10^{12} bp. An arbitrary 19mer would be expected, based on fully random sequence, to occur by chance three or four times in *V. faba* monoploid genome (seven times in the diploid, i.e, in all cells except for pollen and the egg cell), whereas a 20mer would be found only once in the monoploid and two or three times in the diploid; a 21mer, once or less.

Critically, this is an inaccurate estimate of the true frequency of oligomers, and therefore uniqueness, in a likely target crop for NGT1. First, the four nucleotides do not occur at the same frequency (usually, GC < AT) and plant genome sequences are far from random. This is due both to the functional importance of sequence information in genes (both regulatory and coding regions) and especially to the high percentage, even 80%, of large genomes represented by relatively few abundant retrotransposon families that comprise highly similar sequences. Analyses of the length required for true unique representation have been made (Figure 5); for single-occurrence frequencies, sequences must be ~400 bp even in compact crop genomes such as rice, although 100 bp (not 20 bp!) is sufficient for the small genome of Arabidopsis.

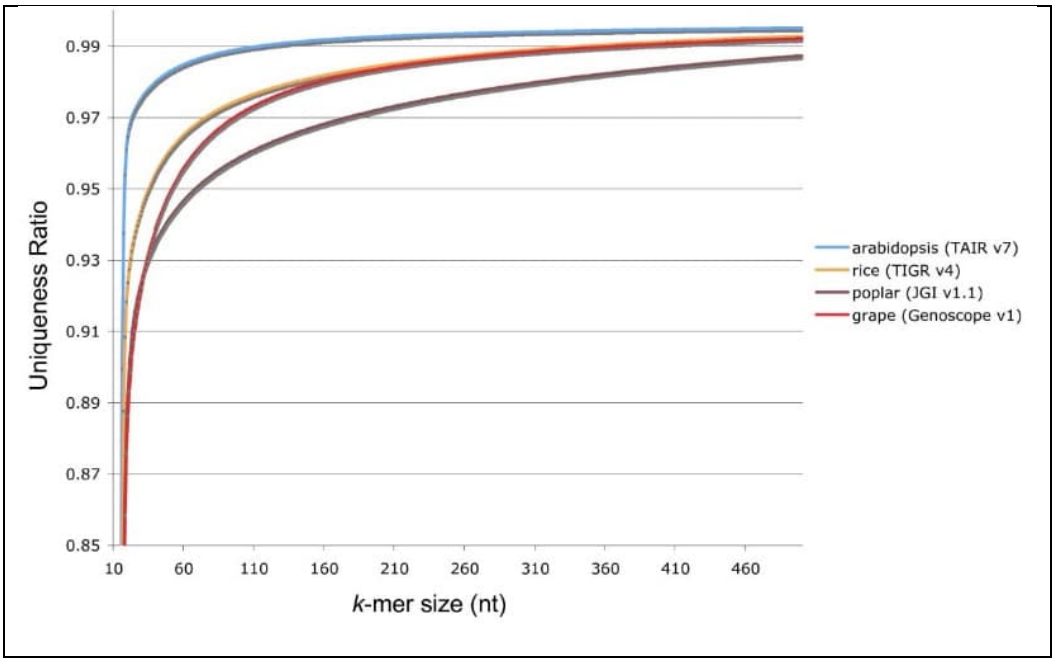


Figure 5 The k-mer uniqueness ratio for some assembled plant genomes as a function of k. The uniqueness ratio is the ratio of *k*-mers occurring exactly once relative to all *k*-mers in the set. It is computed for every *k* between 10 and 500. Extrapolating beyond the tested *k*-mer interval, it appears as though poplar, rice, and grape approach unity at a much slower rate than Arabidopsis. Source: Kurtz *et al*. (2008).

Moreover, at least currently, most targets for gene editing are protein-coding sequences, which are highly non-random. Eukaryotic proteins are encoded by only 64 specific triplets (codons), which form the genetic code for the amino acids, with some of these being much preferred for particular amino acids (De Amicis and Marchetti, 2000). Furthermore, some amino acids are over-represented in the encoded set of cellular proteins, the proteome. Hence the frequencies of 20mers, each representing ~7 amino acids and among the likely gene editing targets, are much higher than expected for random sequences of that length. Certainly, if we consider the contiguous protein-coding segments of a gene (exons), random insertions and deletions would equally likely destroy the protein's function unless, at least, they were precisely phased with the reading frame of the gene. Hence, 20 bp, the insertion limit under NGT1 is insufficient to specify a unique, non-random insertion in typical plant genomes; sequences at least 20-fold longer are still within the range of statistically random variation in plants.

Alternative approaches to uniqueness for insertions under NGT1

From the considerations above, random mutations found in nature provide no obvious limit to insertion size based on a naturalness criterion. If the uniqueness argument is used, the data (Figure 5) would indicate a limit of at least 400 bp would be needed under the NGT1 standard. An alternative approach would be to choose as the limit the largest insertion that would not contain the coding sequence of a full-length protein, in order to maintain a distinction between NGT1 and a gene insertion, i.e., achievable with transgenesis or cisgenesis.

Of the conventional cellular proteins, one-finger (Dof) proteins are plant-specific zinc finger proteins and typically contain 200 to 400 amino acids (Waschburger *et al*., 2024), equivalent to 600—1200 bp. Plant haemoglobins are still smaller, ~150 amino acids, encoded by 450 bp (Becana *et al*., 2020). "Miniproteins," recently discovered, are the smallest proteins to be found in plants (Gruber *et al.*, 2008). They are generally only 50 to 60 amino acids long, hence equivalent to 150 bp, but despite their small size, they can play important regulatory functions (Molesini *et al.*, 2012). For example, the cyclotides, a special class of miniproteins found in the family *Violaceae*, have antimicrobial and antifungal properties (Kim *et al.*, 2023; Lian *et al.*, 2024). Among mammalian proteins, insulin is exceptionally small, the mature form comprising two chains of 21 and 30 amino acids respectively; another example, the bioactive thymosin alpha 1 peptide, is 28 amino acids long (Tao *et al.*, 2023). Given that proteins of less than 50 or 60 amino acids in length are however unlikely to fold into an active form (Linsky *et al.*, 2022), 150 bp (given 3 bp per amino acid) is a reasonable insertion size for

distinguishing protein-coding sequences. This could serve as the maximum insertion size qualifying as NGT1.

It is worth noting that a limit of 50 amino acids or 150 bp generally distinguishes functional proteins, but that short peptides, even less than ten amino acids, if expressed, may have functionality, e.g. through their binding to enzymes in a cell. Coding sequences for short peptides may be generated naturally through point mutations or indels, such as those resulting from double-strand break repair processes, which are discussed above, and of course through proteolytic digestion. However, unless they are near an active promotor, are contained within an mRNA that will be translated, are produced in significant quantity, and have biological function, they are of no consequence. Possible formation of such peptides in GMO events, for example, is routinely checked against toxin databases.

Insertion and deletion numbers in plant genomes vs the 20-insertion limit

The idea in the proposed legislation is that what is achieved by NGT1 should be equivalent to, and no more than, what can be reached through conventional breeding (which includes radiation and chemical mutagenesis). The 14204/23 report states that, in the literature, "*the total number of genetic modifications in individual viable plants ranged from thirty to one hundred. The mutation frequency after using random mutagenesis was higher compared to natural mutation rates. It remained nevertheless below the total number of accumulated single nucleotide polymorphisms naturally occurring between different cultivars.*"

This concept raises the question of what current data show for the number of indels found between cultivars, landraces, and wild accessions. First, it is important to note that comparisons are based on sequence assemblies representing, for a given cultivar, landrace, or wild line, either a single individual or a consensus from several individuals. Heterozygosity within the accession or cultivar is filtered from the published sequence. However, very recent intra-varietal long-read sequencing has been made and phased into the two haplotypes of the clonally propagated "Fuji" apple (Cai *et al.*, 2024), allowing the discovery of 68,965 somatic SNPs across 74 individuals, or 932 per each. Intra-individual mutation rates vary greatly by tissue, by propagation method (clonal vs. sexual), and by life cycle (perennial vs. annual), ranging from 0.08—15.78 x 10^{-9} per bp per year, the highest rate being seen in wild strawberry (*Fragaria vesca*) stems (Wang *et al.*, 2019). This rate corresponds to 6 changes per diploid genome in each cell per year in strawberry plants clonally propagated by runners. In long-lived individuals, these changes accumulate; the same study found up to 19 inherited mutations (mean 11) per individual peach (*Prunus persica*) on one tree, which would be close to the limit permitted for NGT1 insertions under the proposed legislation.

Coming back to consensus sequences for plant lines, just as for indel size, current long-read assemblies provide a perspective on indel number that was generally unavailable before 2022. Taking barley as an example, the recent barley pan-genome (Jayakodi *et al.*, 2024), comprising long-read sequence assemblies of 76 wild and domesticated genomes and short-read sequence data of 1,315 genotypes, contains a total of 155 million SNPs and 9 million indels in 315 elite cultivars, or 493,837 SNPs and 28,983 indels per accession. Moreover, the extensive mutation breeding used for barley in the 1960s has left a legacy of abundant inversion polymorphisms in current germplasm that confer various selective advantages: among 69 barley genotypes (67 domesticated and 2 wild accessions) a total of 42 inversions were found that ranged from 4 to 141 Mb in size (mean 23.9 Mb). An independent, very complete survey of the barley gene pool (Weisweiler *et al.*, 2022) shows ~100,000 indels (lengths of 2—49 bp examined) in genic (exon + intron) regions among 23 inbred lines. Clusters of structural variants (SV) present per inbred ranged from less than 40,000 to more than 80,000.

The high level of SVs and indels is not unique to barley. Regarding rice, *Oryza sativa* ssp. *javanica* is a large-grain landrace. A recent study (Long *et al.*, 2022) found from 164,018 to 211,135 indels and 3,313 to 4,959 longer SVs in javanica compared to the commonly cultivated japonica or indica subspecies. In grapevine, Di Genova *et al.* (2014) identified 623,003 indels of 1 bp to 46 kb, of which 5981 were exon indels and 172,385 intron indels. In wheat, when the Chinese Spring reference genome was compared to other bread wheat accessions, some 36,904 frameshift indels where found that may impact protein function (Montenegro *et al.*, 2017).

The high level of variations found by genome sequencing of crop cultivars and landraces has direct practical implications. Conventional breeding involves crossing of elite cultivars with each other as well as introgression of genetic material from landraces and wild relatives. Crosses will introduce the full complement of variations, including SNPs, indels and other SVs, present on one haploid set of chromosomes, amounting to 40 to 80 thousand in the case of barley. The incorporation of massive numbers of genic and regulatory variations by crossing necessitates extensive back-crossing to the elite parental cultivar in most breeding programs, a process slowed by the “linkage drag” of unwanted variants flanking a desired introduced allele (Chitwood-Brown *et al.*, 2021; Deblieck *et al.*, 2022).

Not only conventional crossing, but also conventional random mutagenesis (not regulated under 2001/18), introduces large numbers of changes, the type and frequency depending on the mutagenesis agent and dosage. Mutation frequencies from the commonly used chemical mutagen ethyl methane sulphonate (EMS) can be 1.5—4.1 x 10^{-6}, corresponding to 7500 to 20,000 “off-target” mutations in the haploid barley genome of a mutagenized line (Jiang *et al.*, 2022). It is precisely the messiness of conventional mutagenesis compared with the clean introduction of edited alleles by

NBT that attracts breeders to gene editing (Yang *et al.*, 2023). Frequencies of off-target mutations induced by CRISPR-*Cas9* are very low, generally less than 5% in likely (i.e., almost identical) off-target sites (Slaman *et al.*, 2023), which would correspond to frequencies on the order 1×10^{-9} in the barley chemical mutagenesis example above. The few off-target mutations would be segregated away rapidly by onward breeding.

The studies described above, taken together, show that intra-plant, inter-individual, and inter-line indel numbers, both spontaneously occurring and obtained via conventional mutagenesis, are generally well in excess, even by a thousand-fold, over permissible insertion numbers under the proposed standard for NGT1. Even if we assume that half of the indels are insertions (restricted under NGT1) and the other half are deletions (unrestricted), targeted mutagenesis such as by CRISPR/*Cas9* or similar methods will not plausibly approach the amount of insertion-generated variations seen in the breeders' pool.

Practical consequences of the maximum 20 permitted NGT1 insertions

While any number and size of deletions is permitted for NGT1, in cases where insertions are used to edit multiple members of gene families, the question of gene family size versus natural variation within becomes relevant. Gene families in plants range from single-copy to hundreds of members. The many ongoing pan-genome projects in plants, in which high-quality genome assemblies for multiple accessions can be analysed, have revealed large variations in many gene family sizes both within and between species (Niu *et al.*, 2024). These together with structural variations, indels and SNPs and would thereby challenge the proposed 20/20 rule because the number of targets for editing under NGT1, as well as their initial state, may vary from cultivar to cultivar.

The NLR genes are a good example of an important NGT target limited by the 20-insertion rule.

Plant genomes typically contain hundreds of nucleotide-binding site leucine-rich repeat (NLR) genes, which are the largest family of plant disease resistance genes. The number of NLR genes per genome vary from 149 in Arabidopsis to ~3400 in bread wheat (Tong *et al.*, 2022). The NRL genes in Arabidopsis (Mondragon-Palomino *et al.*, 2017), wheat (Hao *et al.*, 2023), and soybean (Liu *et al.*, 2024), have been shown to have evolved and diversified through recombination and accumulation of SNPs and indels, with changes displaying association with disease resistance. Resistance genes are often "stacked", as described below, and modified rather than knocked out. Hence, the need to edit more than 20 by insertion approaches, especially to provide resistance against several pathogens, can likely easily arise.

The alpha-gliadin genes as an example of the impact of limitations arising from 20-insertion rule

The genes of alpha-gliadin family of storage proteins in wheat are part of the very dynamic *Gli-2* loci. The alpha gliadins are known for their importance in breadmaking as well as for their role in triggering celiac disease (CD). A combination of long-read sequencing and optical mapping was used to assemble the loci (Huo *et al.*, 2018). Three loci are found in each homoeologous set of chromosomes (A, B, D) of the hexaploidy bread wheat genome, in total nine loci, hence illustrating the importance of using the monoploid chromosome set as the standard for the number of permitted changes in plant genomes and increasing it by the ploidy level (see discussion below). Huo et al. (Huo *et al.*, 2018) identified a total of 47 α-gliadin genes in bread wheat, with only 26 encoding intact full-length protein products. Altogether 21 of the 47 were pseudogenes, 13 due to SNPs, 4 to deletions, others to rearrangements. Three contained TE insertions, premature stop codons, and frameshift indels. However, a 20mer associated with CD epitopes is present in 2161 copies at 93—100% identity in the alpha gliadin genes within the Chinese Spring genome (Schulman, unpublished). Others have attempted to analyse the relative abundance of CD types (Marin-Sanz *et al.*, 2023). An in-depth analysis of transcription and protein accumulation in the bread wheat Chinese cultivar Xiaoyan 81 (Wang *et al.*, 2017) found that 52 full-length gliadin genes were transcribed, 42 of these encoded proteins, 38 gliadins accumulated in mature grains, 10 did not carry any CD epitope, eight had one or two epitopes in their proteins, and 20 contained more than three epitopes in their proteins; of the 28 gliadins with CD epitopes, a total of 202 epitopes in the proteins were present at 100% match. Making the alpha-gliadins safe for CD patients by using NGT for all 28 CD-epitope-containing alpha-gliadin genes to alter all 202 CD epitopes would not be acceptable under the 20/20 rule within the current EC proposal. Removal through large deletions of the tandemly organised genes (Jouanin *et al.*, 2019), while permitted, is possible but not practical for all gliadin families if one wants to maintain baking quality (Jouanin *et al.*, 2020).

As a further example, receptor-like kinases (RLKs), which are critical for biotic and abiotic stress response, and therefore likely NGT targets, are found in 100s to 1000s copies depending on the plant species and have undergone a great degree of recombination and variation (Yan *et al.*, 2023). Another example of a large gene family in plants is that of cytochrome P450 (CYP450), which includes 100s of members in most plant genomes (Zhang *et al.*, 2023). A subgroup of CYP450, CYP71, which is connected to insect resistance, senescence, and yield-related traits, was studied in rice. In rice, 105 *OsCYP71* genes were found, of which 36 pairs were involved in gene duplication (in essence, large SVs); major indels of 20 bp affecting 20% of the varieties' promoter structures and thereby expression patterns and trait QTLs were found. In these sorts of cases, the natural variation would need to be confirmed in the edited and non-edited versions to confirm that the editing per se did not generate more than 20 changes for NGT1 status.

Impact on polyploid crops

Beyond variation in gene family number in the basic set of chromosomes, many plant species are not diploid (two sets of chromosomes) but rather tetraploid (four sets), hexaploid (six), octoploid (eight), or even higher. This means that gene family numbers likewise may double, triple, quadruple, or be of higher multiples, as described above for CD epitopes in wheat, complicating editing within a fixed, low limit of insertions under NGT1. For example, pasta (durum) wheat is tetraploid, as is potato, while bread (common) wheat is hexaploid, and cultivated strawberry is octoploid, as is sugar cane. Without adjustments for ploidy, the current 20/20 limits for NGT1 would therefore be far more restrictive for bread wheat than for pasta wheat, and both more than for einkorn wheat, which is a diploid as is barley. Cultivated roses (*Rosa hybrida*) can be either diploid, triploid, or tetraploid (Harmon *et al.*, 2023) but would be permitted the same maximum 20 insertions under NGT1. Clearly a more rational approach is needed.

Gene stacking versus NGT insertion number

An important goal in plant breeding is to "stack" or combine multiple beneficial traits into the same plant line, e.g. to improve an already commercially successful variety. This is to meet the widespread goal and urgent need for several classes of phenotypes: simultaneous resistance to multiple plant diseases; robust resistance to individual pathogens through combined use of different independently acting genes; both abiotic (e.g. drought) stress tolerance and disease resistance in a crop plant; both healthy crop plants and a harvest with human-health-promoting qualities (e.g. CD-safety). In some cases, even a single trait, for example CD-epitope-free gluten protein in wheat, requires the stepwise stacking of alleles. This is possible but slow by conventional breeding, requiring support by marker-assisted selection (MAS) and epitope immuno-assays. At each stage, the properties of the gliadins for baking quality would need to be preserved and tested. In fact, even use of GMO approaches to introduce multiple genes in parallel is technically highly challenging (Halpin, 2005). In practice, also transgenes therefore have been stacked through conventional crosses (Li *et al.*, 2023), with at most seven genes currently stacked (https://www.isaaa.org/gmapprovaldatabase/eventslist/), which is a maize line providing herbicide tolerance, multiple insect resistance, a modified alpha amylase, and altered mannose metabolism. The practical limitations to gene stacking raise several important questions: First, should the current technical limits of older conventional approaches serve as the basis for limiting NGT target numbers? Second, if so, can this limit be justified by some risk specific to gene stacking and not merely the sum of the individual risks? Conventional gene stacking is a moving target, as both biochemical phenotyping and marker-assisted selection improves. Moreover, no restrictions are imposed on stacked-gene conventional cultivars; rather, they command a premium price and are welcomed in the marketplace.

Conclusions and Future Prospects

We find that current genomic data indicate that natural variation in the germplasm used by breeders is much greater that earlier understood and that both conventional breeding and mutagenesis can introduce genomic changes that are more extensive in size and more frequent than the Category 1 (NGT1) “20 insertions of maximum 20 bp” rule would allow. Regarding genome size and polyploidy, the 20/20 rule for NGT1 does not take into account varying plant gene family sizes, the dynamic variation of gene family number and genome size in evolution, the effect of the limitation on improvement of the many polyploid crops in agriculture. Neither does it address the need for gene stacking to combine the traits needed for future-ready crops, which may lead to the limit being easily exceeded for many practical breeding goals. We conclude, moreover, that the criteria of “naturalness” and “uniqueness” which form the standards for the proposed rule are not met by the proposed NGT1 limits.

An approach based on the current state of knowledge, which would imply broadening the 20/20 rule, would better support the development of NGT1 plants while still ensuring they are equivalent to conventional plants. Such an approach would facilitate breeding for climate resilience, farming sustainability, and nutritional security. In March 2025, proposed amendments to Annex I introduced by the Polish Presidency of the Council of the EU, which serves through June 2025, appeared to achieve a qualified majority of the Council (Permanent Representatives Committee) to proceed to the trialogue (negotiations on the terms of the legislation and Annex I between the Council, the Commission, and the European Parliament). An important proposed amendment is that NGT1 limits would apply per monoploid genome. The 20/20 limit for NGT1, however, would remain in place per monoploid genome. The standards reached will greatly influence both the use of NGTs in Europe for research and applications and the introduction of NGT products into the marketplace.

Acknowledgements

This work was financially supported by Grant 210021 from the Jane and Aatos Erkko Foundation to A.H.S.

Author Contributions

All authors contributed to the review article. A.H.S. conceived of the study and drafted the manuscript. F.H., M.J.M.S., J.F.S., R.W., O.-A.R., and K.M. contributed data, text, and interpretation and made revisions. A.H.S. prepared the figures. All authors approved the final version of the manuscript.

Data availability statement

Only publicly available data was used and in this study; no new data were generated.

Conflict of interest disclosure

The authors declare no conflicts of interest.

Permission to reproduce material from other sources

References

Ajawatanawong, P. and Baldauf, S.L. (2013) Evolution of protein indels in plants, animals and fungi. *BMC Evol Biol* 13, 140.

Becana, M., Yruela, I., Sarath, G., Catalan, P. and Hargrove, M.S. (2020) Plant hemoglobins: a journey from unicellular green algae to vascular plants. *New Phytol.* 227, 1618-1635.

Ben-Tov, D., Mafessoni, F., Cucuy, A., Honig, A., Melamed-Bessudo, C. and Levy, A.A. (2024) Uncovering the dynamics of precise repair at CRISPR/Cas9-induced double-strand breaks. *Nature Comm.* 15, 5096.

Broothaerts, W., Jacchia, S., Angers, A., Petrillo, M., Querci, M., Savini, C., Van Den Eede, G. and Emons, H. (2021) *New Genomic Techniques: State-of-the-Art Review, EUR 30430 EN*. Luxembourg:Publications Office of the European Union.

Cai, Y., Gao, X., Mao, J., Liu, Y., Tong, L., Chen, X., Liu, Y., Kou, W., Chang, C., Foster, T., Yao, J., Cornille, A., Tahir, M.M., Liu, Z., Yan, Z., Lin, S., Ma, F., Ma, J., Xing, L., An, N., Zuo, X., Lv, Y., Zhao, Z., Li, W., Li, Q., Zhao, C., Hu, Y., Liu, H., Wang, C., Shi, X., Ma, D., Fei, Z., Jiang, Y. and Zhang, D. (2024) Genome sequencing of 'Fuji' apple clonal varieties reveals genetic mechanism of the spur-type morphology. *Nature Comm.* 15, 10082.

Chitwood-Brown, J., Vallad, G.E., Lee, T.G. and Hutton, S.F. (2021) Characterization and elimination of linkage-drag associated with Fusarium wilt race 3 resistance genes. *Theor. Appl. Genet.* 134, 2129-2140.

Commission Services (2023) Regulation on new genomic techniques (NGT) – Technical paper on the rationale for the equivalence criteria in Annex I. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_14204_2023_INIT

De Amicis, F. and Marchetti, S. (2000) Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Res.* 28, 3339-3345.

Deblieck, M., Szilagyi, G., Andrii, F., Saranga, Y., Lauterberg, M., Neumann, K., Krugman, T., Perovic, D., Pillen, K. and Ordon, F. (2022) Dissection of a grain yield QTL from wild emmer wheat reveals sub-intervals associated with culm length and kernel number. *Front. Genet.* 13, 955295.

Di Genova, A., Almeida, A.M., Munoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., Pinto, M., Hinrichsen, P., Orellana, A. and Maass, A. (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* 14, 7.

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms). (2012) Scientific opinion addressing the safety assessment of plants developed using Zinc Finger Nuclease 3 and other Site-Directed Nucleases with similar function. *EFSA J.* 10, 2943.

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), Mullins E, Bresson J-L, Dalmay T, Dewhurst IC, Epstein MM, Firbank LG, Guerche P, Hejatko J, Moreno FJ, Naegeli H, Nogué F, Rostoks N, Sánchez Serrano JJ, Savoini G, Veromann E, Veronesi F, Fernandez A, Gennaro A, Papadopoulou N, Raffaello T and Schoonjans R,, Mullins, E., Bresson, J.L., Dalmay, T., Dewhurst, I.C., Epstein, M.M., Firbank, L.G., Guerche, P., Hejatko, J., Moreno, F.J., Naegeli, H., Nogue, F., Rostoks, N., Sanchez Serrano, J.J., Savoini, G., Veromann, E., Veronesi, F., Fernandez, A., Gennaro, A., Papadopoulou, N., Raffaello, T. and Schoonjans, R. (2022) Criteria for risk assessment of plants produced by targeted mutagenesis, cisgenesis and intragenesis. *EFSA J.* 20, e07618.

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), Naegeli, H., Bresson, J.L., Dalmay, T., Dewhurst, I.C., Epstein, M.M., Firbank, L.G., Guerche, P., Hejatko, J., Moreno, F.J., Mullins, E., Nogue, F., Sanchez Serrano, J.J., Savoini, G., Veromann, E., Veronesi, F., Casacuberta, J., Gennaro, A., Paraskevopoulos, K., Raffaello, T. and Rostoks, N. (2020) Applicability of the EFSA Opinion on site-directed nucleases type 3 for the safety assessment of plants developed using site-directed nucleases type 1 and 2 and oligonucleotide-directed mutagenesis. *EFSA J.* 18, e06299.

517 Gruber, C.W., Elliott, A.G., Ireland, D.C., Delprete, P.G., Dessein, S., Goransson, U., Trabi, M., Wang,
518 C.K., Kinghorn, A.B., Robbrecht, E. and Craik, D.J. (2008) Distribution and evolution of circular
519 miniproteins in flowering plants. *Plant Cell* 20, 2471-2483.
520 Guo, X., Ren, J., Zhou, X., Zhang, M., Lei, C., Chai, R., Zhang, L. and Lu, D. (2024) Strategies to improve
521 the efficiency and quality of mutant breeding using heavy-ion beam irradiation. *Crit. Rev.*
522 *Biotechnol.* 44, 735-752.
523 Halpin, C. (2005) Gene stacking in transgenic plants--the challenge for 21st century plant
524 biotechnology. *Plant Biotechnol. J.* 3, 141-155.
525 Hao, Y., Pan, Y., Chen, W., Rashid, M.A.R., Li, M., Che, N., Duan, X. and Zhao, Y. (2023) Contribution of
526 duplicated nucleotide-binding leucine-rich repeat (NLR) genes to wheat disease resistance.
527 *Plants* 12, 2794.
528 Harmon, D.D., Chen, H., Byrne, D., Liu, W. and Ranney, T.G. (2023) Cytogenetics, ploidy, and genome
529 sizes of rose (Rosa spp.) cultivars and breeding lines. *Ornam. Plant Res.* 3, 10.
530 Himmelbach A, Walde I, Mascher M, Stein N. (2018) Tethered chromosome conformation capture
531 sequencing in Triticeae: A valuable tool for genome assembly. *Bio Protoc.* 8, e2955.
532 Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., *et al.* (2020) Highly accurate
533 long-read HiFi sequencing data for five complex genomes. *Sci Data* 7, 399.
534 Huo, N., Zhu, T., Altenbach, S., Dong, L., Wang, Y., Mohr, T., Liu, Z., Dvorak, J., Luo, M.C. and Gu, Y.Q.
535 (2018) Dynamic evolution of alpha-gliadin prolamin gene family in homeologous genomes of
536 hexaploid wheat. *Sci. Rep.* 8, 5181.
537 Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M.T., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A.,
538 Bekele, W., Brar, G.S., Braune, K., Bunk, B., Chalmers, K.J., Chapman, B., Jorgensen, M.E.,
539 Feng, J.W., Feser, M., Fiebig, A., Gundlach, H., Guo, W., Haberer, G., Hansson, M.,
540 Himmelbach, A., Hoffie, I., Hoffie, R.E., Hu, H., Isobe, S., Konig, P., Kale, S.M., Kamal, N.,
541 Keeble-Gagnere, G., Keller, B., Knauft, M., Koppolu, R., Krattinger, S.G., Kumlehn, J.,
542 Langridge, P., Li, C., Marone, M.P., Maurer, A., Mayer, K.F.X., Melzer, M., Muehlbauer, G.J.,
543 Murozuka, E., Padmarasu, S., Perovic, D., Pillen, K., Pin, P.A., Pozniak, C.J., Ramsay, L., Pedas,
544 P.R., Rutten, T., Sakuma, S., Sato, K., Schuler, D., Schmutzer, T., Scholz, U., Schreiber, M.,
545 Shirasawa, K., Simpson, C., Skadhauge, B., Spannagl, M., Steffenson, B.J., Thomsen, H.C.,
546 Tibbits, J.F., Nielsen, M.T.S., Trautewig, C., Vequaud, D., Voss, C., Wang, P., Waugh, R.,
547 Westcott, S., Rasmussen, M.W., Zhang, R., Zhang, X.Q., Wicker, T., Dockter, C., Mascher, M.
548 and Stein, N. (2024) Structural variation in the pangenome of wild and domesticated barley.
549 *Nature* 636, 654-662.
550 Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T., Kamal, N.,
551 Lang, D., Himmelbach, A., Ens, J., Zhang, X.Q., Angessa, T.T., Zhou, G., Tan, C., Hill, C., Wang,
552 P., Schreiber, M., Boston, L.B., Plott, C., Jenkins, J., Guo, Y., Fiebig, A., Budak, H., Xu, D., Zhang,
553 J., Wang, C., Grimwood, J., Schmutz, J., Guo, G., Zhang, G., Mochida, K., Hirayama, T., Sato, K.,
554 Chalmers, K.J., Langridge, P., Waugh, R., Pozniak, C.J., Scholz, U., Mayer, K.F.X., Spannagl, M.,
555 Li, C., Mascher, M. and Stein, N. (2020) The barley pan-genome reveals the hidden legacy of
556 mutation breeding. *Nature* 588, 284-289.
557 Jiang, C., Lei, M., Guo, Y., Gao, G., Shi, L., Jin, Y., Cai, Y., Himmelbach, A., Zhou, S., He, Q., Yao, X., Kan,
558 J., Haberer, G., Duan, F., Li, L., Liu, J., Zhang, J., Spannagl, M., Liu, C., Stein, N., Feng, Z.,
559 Mascher, M. and Yang, P. (2022) A reference-guided TILLING by amplicon-sequencing
560 platform supports forward and reverse genetics in barley. *Plant Commun.* 3, 100317.
561 Jouanin, A., Borm, T., Boyd, L., Cockram, L., Leigh, F., Santos, B., Visser, R. and Smulders, M. (2019)
562 Development of the GlutEnSeq capture system for sequencing gluten gene families in
563 hexaploid bread wheat with deletions or mutations induced by γ-irradiation or CRISPR/Cas9.
564 *J Cereal Sci* 88, 157-166.
565 Jouanin, A., Gilissen, L., Schaart, J.G., Leigh, F.J., Cockram, J., Wallington, E.J., Boyd, L.A., van den
566 Broeck, H.C., van der Meer, I.M., America, A.H.P., Visser, R.G.F. and Smulders, M.J.M. (2020)

CRISPR/Cas9 Gene Editing of Gluten in Wheat to Reduce Gluten Content and Exposure-Reviewing Methods to Screen for Coeliac Safety. *Front. Nutr.* 7, 51.
Kim, D.E., Jensen, D.R., Feldman, D., Tischer, D., Saleem, A., Chow, C.M., Li, X., Carter, L., Milles, L., Nguyen, H., Kang, A., Bera, A.K., Peterson, F.C., Volkman, B.F., Ovchinnikov, S. and Baker, D. (2023) De novo design of small beta barrel proteins. *Proc. Natl. Acad. Sci. USA* 120, e2207974120.
Kitamura, S., Satoh, K., Hase, Y., Yoshihara, R., Oono, Y. and Shikazono, N. (2024) Differential contributions of double-strand break repair pathways to DNA rearrangements following the irradiation of Arabidopsis seeds and seedlings with ion beams. *Plant J.* 120, 445-458.
Kurtz, S., Narechania, A., Stein, J.C. and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* 9, 517.
Li, B., Chen, Z., Chen, H., Wang, C., Song, L., Sun, Y., Cai, Y., Zhou, D., Ouyang, L., Zhu, C., He, H. and Peng, X. (2023) Stacking multiple genes improves resistance to *Chilo suppressalis*, *Magnaporthe oryzae*, and *Nilaparvata lugens* in transgenic rice. *Genes (Basel)* 14.
Lian, Y., Tang, X., Hu, G., Miao, C., Cui, Y., Zhangsun, D., Wu, Y. and Luo, S. (2024) Characterization and evaluation of cytotoxic and antimicrobial activities of cyclotides from *Viola japonica*. *Sci. Rep.* 14, 9733.
Linsky, T.W., Noble, K., Tobin, A.R., Crow, R., Carter, L., Urbauer, J.L., Baker, D. and Strauch, E.M. (2022) Sampling of structure and sequence space of small protein folds. *Nature Comm.* 13, 7151.
Liu, G., Fang, Y., Liu, X., Jiang, J., Ding, G., Wang, Y., Zhao, X., Xu, X., Liu, M., Wang, Y. and Yang, C. (2024) Genome-wide association study and haplotype analysis reveal novel candidate genes for resistance to powdery mildew in soybean. *Front. Plant Sci.* 15, 1369650.
Long, W., Luo, L., Luo, L., Xu, W., Li, Y., Cai, Y. and Xie, H. (2022) Whole genome resequencing of 20 accessions of rice landraces reveals Javanica genomic structure variation and allelic genotypes of a grain weight gene TGW2. *Front. Plant Sci.* 13, 857435.
Marin-Sanz, M., Barro, F. and Sanchez-Leon, S. (2023) Unraveling the celiac disease-related immunogenic complexes in a set of wheat and tritordeum genotypes: implications for low-gluten precision breeding in cereal crops. *Front. Plant Sci.* 14, 1171882.
Molesini, B., Pandolfini, T., Pii, Y., Korte, A. and Spena, A. (2012) Arabidopsis thaliana AUCSIA-1 regulates auxin biology and physically interacts with a kinesin-related protein. *PLoS One* 7, e41327.
Mondragon-Palomino, M., Stam, R., John-Arputharaj, A. and Dresselhaus, T. (2017) Diversification of defensins and NLRs in Arabidopsis species by different evolutionary mechanisms. *BMC Evol. Biol.* 17, 255.
Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.K., Visendi, P., Lai, K., Dolezel, J., Batley, J. and Edwards, D. (2017) The pangenome of hexaploid bread wheat. *Plant J.* 90, 1007-1013.
Niu, J., Wang, W., Wang, Z., Chen, Z., Zhang, X., Qin, Z., Miao, L., Yang, Z., Xie, C., Xin, M., Peng, H., Yao, Y., Liu, J., Ni, Z., Sun, Q. and Guo, W. (2024) Tagging large CNV blocks in wheat boosts digitalization of germplasm resources by ultra-low-coverage sequencing. *Genome Biol.* 25, 171.
Paajanen P, Kettleborough G, López-Girona E, Giolai M, Heavens D, Baker D, Lister A, Cugliandolo F, Wilde G, Hein I, Macaulay I, Bryan GJ, Clark MD. (2019) A critical comparison of technologies for a plant genome sequencing project. *Gigascience* 8, ngiy163.
Shivaprasad, K.M., Aski, M., Mishra, G.P., Sinha, S.K., Gupta, S., Mishra, D.C., Singh, A.K., Singh, A., Tripathi, K., Kumar, R.R., Kumar, A., Kumar, S. and Dikshit, H.K. (2024) Genome-wide discovery of InDels and validation of PCR-Based InDel markers for earliness in a RIL population and genotypes of lentil (*Lens culinaris* Medik.). *PLoS One* 19, e0302870.

Slaman, E., Lammers, M., Angenent, G.C. and de Maagd, R.A. (2023) High-throughput sgRNA testing reveals rules for Cas9 specificity and DNA repair in tomato cells. *Front. Genome Ed.* 5, 1196763.

Sun, G., Pourkheirandish, M. and Komatsuda, T. (2009) Molecular evolution and phylogeny of the RPB2 gene in the genus Hordeum. *Ann. Bot.* 103, 975-983.

Sun H, Tusso S, Dent CI, Goel M, Wijfjes RY, Baus LC, Dong X, Campoy JA, Kurdadze A, Walkemeier B, Sänger C, Huettel B, Hutten RCB, van Eck HJ, Dehmer KJ, Schneeberger K. (2025) The phased pan-genome of tetraploid European potato. *Nature* doi: 10.1038/s41586-025-08843-0.

Tao, N., Xu, X., Ying, Y., Hu, S., Sun, Q., Lv, G. and Gao, J. (2023) Thymosin alpha1 and Its Role in Viral Infectious Diseases: The Mechanism and Clinical Application. *Molecules* 28.

Tong, C., Zhang, Y. and Shi, F. (2022) Genome-wide identification and analysis of the NLR gene family in Medicago ruthenica. *Front. Genet.* 13, 1088763.

Wang, D.W., Li, D., Wang, J., Zhao, Y., Wang, Z., Yue, G., Liu, X., Qin, H., Zhang, K., Dong, L. and Wang, D. (2017) Genome-wide analysis of complex wheat gliadins, the dominant carriers of celiac disease epitopes. *Sci. Rep.* 7, 44609.

Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., Zhang, X., Zhao, L., Zhang, Y., Jia, Y., Qin, C., Yu, L., Huang, J., Yang, S., Hurst, L.D. and Tian, D. (2019) The architecture of intra-organism mutation rate variation in plants. *PLoS Biol.* 17, e3000191.

Waschburger, E.L., Filgueiras, J.P.C. and Turchetto-Zolet, A.C. (2024) DOF gene family expansion and diversification. *Genet. Mol. Biol.* 46, e20230109.

Weisweiler, M., Arlt, C., Wu, P.Y., Van Inghelandt, D., Hartwig, T. and Stich, B. (2022) Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation. *Theor. Appl. Genet.* 135, 3511-3529.

Yadav, C.B., Bhareti, P., Muthamilarasan, M., Mukherjee, M., Khan, Y., Rathi, P. and Prasad, M. (2015) Genome-wide SNP identification and characterization in two soybean cultivars with contrasting Mungbean Yellow Mosaic India Virus disease resistance traits. *PLoS One* 10, e0123897.

Yan, J., Su, P., Meng, X. and Liu, P. (2023) Phylogeny of the plant receptor-like kinase (RLK) gene family and expression analysis of wheat RLK genes in response to biotic and abiotic stresses. *BMC Genomics* 24, 224.

Yang, T., Ali, M., Lin, L., Li, P., He, H., Zhu, Q., Sun, C., Wu, N., Zhang, X., Huang, T., Li, C.B., Li, C. and Deng, L. (2023) Recoloring tomato fruit by CRISPR/Cas9-mediated multiplex gene editing. *Hortic. Res.* 10, uhac214.

Zeng, Y.X., Wen, Z.H., Ma, L.Y., Ji, Z.J., Li, X.M. and Yang, C.D. (2013) Development of 1047 insertion-deletion markers for rice genetic studies and breeding. *Genet. Mol. Res.* 12, 5226-5235.

Zhang, N., Roberts, H.M., Van Eck, J. and Martin, G.B. (2020) Generation and Molecular Characterization of CRISPR/Cas9-Induced Mutations in 63 Immunity-Associated Genes in Tomato Reveals Specificity and a Range of Gene Modifications. *Front. Plant Sci.* 11, 10.

Zhang, W., Li, H., Li, Q., Wang, Z., Zeng, W., Yin, H., Qi, K., Zou, Y., Hu, J., Huang, B., Gu, P., Qiao, X. and Zhang, S. (2023) Genome-wide identification, comparative analysis and functional roles in flavonoid biosynthesis of cytochrome P450 superfamily in pear (Pyrus spp.). *BMC Genom. Data* 24, 58.