Title: Ultra-deep sequencing of somatic mutations induced by a maize transposon

Authors: Justin Scherer[1,2], Michael Hinczewski[3], and Brad Nelms[1,2,4]*

Affiliations:

[1]Department of Genetics, University of Georgia, Athens, GA 30602

[2]The Plant Center, University of Georgia, Athens, GA 30602

[3]Department of Physics, Case Western Reserve University, Cleveland, OH 44106

[4]Department of Plant Biology, University of Georgia, Athens, GA 30602

*Correspondence: nelms@uga.edu

Abstract: Cells accumulate mutations throughout development, contributing to cancer, aging, and evolution. Quantitative data on the abundance of *de novo* mutations within plants or animals are limited, as new mutations are often rare within a tissue and fall below the limits of current sequencing depths and error rates. Here, we show that mutations induced by the maize Mutator (Mu) transposon can be reliably quantified down to a detection limit of 1 part in 12,000. We measured the abundance of millions of *de novo* Mu insertions across four tissue types. Within a tissue, the distribution of *de novo* Mu allele frequencies was highly reproducible between plants, showing that, despite the stochastic nature of mutation, repeated statistical patterns of mutation abundance emerge. In contrast, there were significant differences in the allele frequency distribution between tissues. At the extremes, root was dominated by a small number of highly abundant *de novo* insertions, while endosperm was characterized by thousands of insertions at low allele frequencies. Finally, we used the measured pollen allele frequencies to reinterpret a classic genetic experiment, showing that evidence for late Mu activity in pollen are better explained by cell division statistics. These results provide insight into the complexity of mutation accumulation in multicellular organisms and a system to interrogate the factors that shape mutation abundance.

INTRODUCTION

Multicellular organisms accumulate mutations throughout development, producing genetic heterogeneity within and between tissues. With the increased sensitivity to detect *de novo* mutations through sequencing, it has become clear that genetic mosaicism is ubiquitous even in healthy individuals[1–8]: over 1,000 single-base substitutions are present per adult human fibroblast[1] and megabase-sized structural variants can be observed in 30% of healthy human neurons[2]. In plants, low frequency mutations can be transmitted to the next generation[3], and preexisting (somatic) mutations contribute to variation between plants regenerated in tissue culture[4].

To interpret and predict the effect of *de novo* mutations, it is critical to understand what influences their abundance and spread within the organism. This is challenging for both biological and technical reasons. Biologically, mutation accumulation is complex and depends on processes that impact the initial mutation rate (e.g. mutagenic exposure, DNA repair) as well as the spread of mutations once they arise (e.g. tissue development, selection, cell death). While there have been theoretical advances in understanding how these factors interact to shape mutation abundance[9–15], there is need for quantitative, empirical data to constrain and inform the theory.

This is where the technical challenge comes in: new mutations can span many orders of magnitude in their abundance, down to 1 per cell, pushing the limits of current sequencing depths and error rates. To date, genome-wide studies on *de novo* mutations in plants and animals have reported detection limits around 1-5%[2–8] (**Fig. S1A**), which cover only the most abundant mutations. Targeted sequencing has helped bridge this gap[16–20], with an inverse relationship between genomic coverage and sensitivity to detect rare mutations (**Fig. S1A**). However, targeted sequencing still suffers from a limited dynamic range, as more abundant mutations are unlikely to occur within a narrow genomic window.

Mutations caused by transposable elements (TEs) play important roles in evolution, contributing to genome-size evolution[21], alleles selected during crop domestication[22], and the origin of new genes through exon shuffling[23]. Unlike other classes of mutation, TE insertions introduce defined sequences into the genome that can be targeted by PCR[24]. The potential of this is significant: by selectively amplifying only genome sequences containing the TE, *de novo* insertions can be identified without sequencing through an overwhelming number of wild-type copies at the same location (**Fig. S1B**). This shares advantages of targeted mutation sequencing without needing to focus on a predefined genomic region.

Here, we evaluate the maize Mutator (Mu) transposon as a quantitative model of *de novo* mutation accumulation in multicellular tissues. Mu has long been a valuable model in maize genetics[25–28] because of its high forward mutation rate, availability of both Mu-active and inactive genetic stocks, and ease of identifying Mu insertion locations by sequencing[24,29]. Mu is a class I (DNA) transposon that predominately transposes duplicatively, i.e. transposing to new locations without loss of the donor element[30]; this apparent ‘copy-and-paste’ behavior is thought to be caused by DNA repair pathways that restore the original sequence after transposition[25,26]. Mu transposes into unlinked sites, with no preference to insert near its site of origin[30].

We find that Mu sequencing can accurately measure the absolute allele frequencies of *de novo* Mu insertions within complex tissues, with a sensitivity, dynamic range, and error rate that are orders of magnitude better than currently possible for single-base substitutions. We then measured the allele frequency distribution for *de novo* Mu insertions in leaf, root, pollen and endosperm. Mu had broad

activity in all four tissues, with no evidence for a preference of late insertion in pollen. These results provide a rich dataset with which to test and refine theoretical models of mutation accumulation in multicellular organisms, and highlight the importance of tissue organization in shaping the abundance of *de novo* mutations during development.

RESULTS

Sensitive identification of *de novo* TE insertion sites. To identify Mu TE insertions, we established a sequencing assay based on MuSeq[24], which has been widely used to map Mu insertions in maize genetic stocks[24,31,32]. MuSeq applies nested PCR to specifically amplify and sequence DNA fragments that span the TE-genome boundary (**Fig. S2**). We optimized MuSeq for quantifying the abundance of rare, *de novo* TE insertions within heterogeneous tissue samples. Two key changes were implemented in 'MuSeq2'; first, we introduced molecule counting by incorporating unique molecular identifiers[33] (UMIs) during an initial adapter ligation step (**Fig. S2**). This makes it possible to identify and remove PCR duplicates, improving quantitative accuracy. Second, we limited the amplification of non-Mu products using suppression PCR, providing more specific transposon amplification with fewer PCR cycles (**Fig. S3**).

To test MuSeq2, we first applied it to seedling leaves from Mu-active and inactive maize lines. Samples were sequenced to a mean of 1.7 and 2.9 million TE-spanning molecules per Mu-active and inactive plant, respectively. For the inactive plants, all Mu elements are expected to map to a fixed set of genomic locations, representing historical TE insertions. Indeed, 99.8% of molecules from Mu-inactive samples mapped to only 29 locations (**Table S1**). In contrast, Mu-active plants had Mu elements mapping to a wide range of new genomic sites (**Fig. S4A,B**), with a mean of 184,410 insertion sites detected per leaf. These can be confirmed as *bona fide* Mu insertions because (i) the TE border was consistently sequenced along with the genomic region (**Fig. S4C**) and (ii) 123,312 sites were sequenced out of both directions of the TE, including the 9 bp target site duplication that is characteristic of Mu insertions[26]. To estimate the error rate of MuSeq2, we leveraged the fact that Mu-inactive lines have a negligible rate of new Mu transposition, providing a genetic control for no transposon activity. Assuming that all molecules mapping outside the 29 historical locations were false positives, the error rate of MuSeq2 is 0.1 falsely identified insertions per diploid cell (2.6×10^{-11} false positive insertions per bp), two orders of magnitude lower than the most accurate duplex methods to measure single-base substitutions[34].

***De novo* and inherited Mu insertions across matched tissues.** We next applied MuSeq2 to leaf, pollen, endosperm, and root from Mu-active plants. Our experimental design used a combination of sequential tissue isolations and controlled genetic crosses to unambiguously separate *de novo* insertions from inherited ones, and further divide the inherited insertions by parent-of-origin (**Fig. 1A**). First, the plants were generated from a cross between a Mu-inactive female and Mu-active male; the female parent contributes a defined set of historical insertions (**Fig. 1B**; **Table S1**), and so all other insertion sites in the offspring are either *de novo* or paternally inherited. To distinguish *de novo* from paternally inherited insertions, we used matched tissues with early and well-defined divergence times. The endosperm, which comprises the bulk of maize seed mass, inherits its paternal DNA from a sister sperm cell during double fertilization (**Fig. 1A**, left). Thus, insertions present at high abundance in both

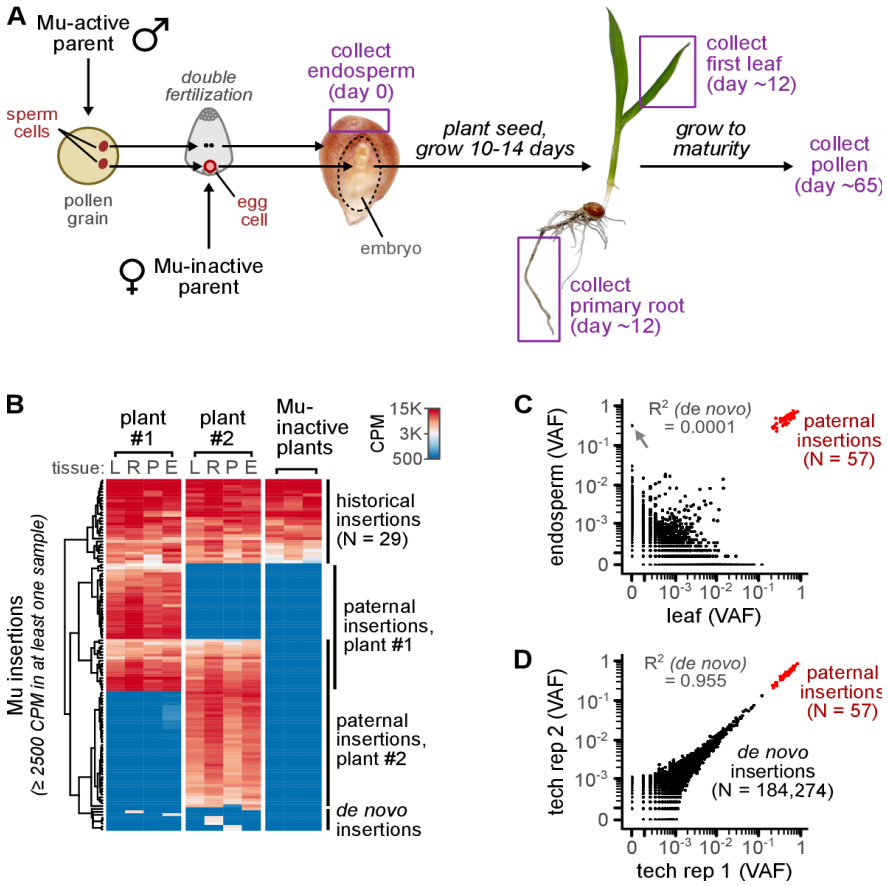

Figure 1. Sensitive and quantitative assessment of *de novo* mutation abundance for an active maize transposon.

(**A**) Cartoon of experimental design and tissue collection. Sequential, matched isolations from endosperm and other tissues make it possible to distinguish inherited from *de novo* insertions, because endosperm is derived from a sister sperm cell during double fertilization.

(**B**) Heatmap showing the abundance of Mu insertions in matched tissue samples from two siblings as well as control Mu-inactive plants. The Mu-inactive samples were from the family used as the female parent, and represent historical insertions that were maternally inherited. All insertion sites with >= 2500 CPM in at least one of the samples are shown. CPM, counts per million (number of TE-spanning molecules at a given genomic site).

(**C**) Allele frequencies of Mu insertions for matched endosperm and leaf from a single plant. Paternal insertions were abundant in both samples, while *de novo* insertions were only abundant in one (e.g. gray arrow). Black dot, *de novo* insertion; red dot, paternal insertion; VAF, variant allele frequency.

(**D**) Technical replicates for a representative leaf sample.

endosperm and embryo-derived tissues must be paternally inherited (hereafter: 'paternal insertions'). Indeed, paternal insertions were well-separated from *de novo* insertions based on their abundance in both endosperm and other tissues from the same plant (**Fig. 1B,C**).

To determine the Mu insertion rate in our line, we compared the inherited insertions between parents and their offspring. On average, there were 1.1 x 10^{-8} new Mu insertions per bp per generation; for context, this is comparable to the per generation single-base substitution rate in *Arabidopsis*[35] (0.7 x 10^{-8}), maize[36] (1.6 x 10^{-8}), and human[37] (1.3 x 10^{-8}). Thus, while the Mu-active line has an unusually high mutation rate for a change the size of a TE insertion (>1 kb for Mu elements), the number of events is similar to the background rate of single-base substitutions.

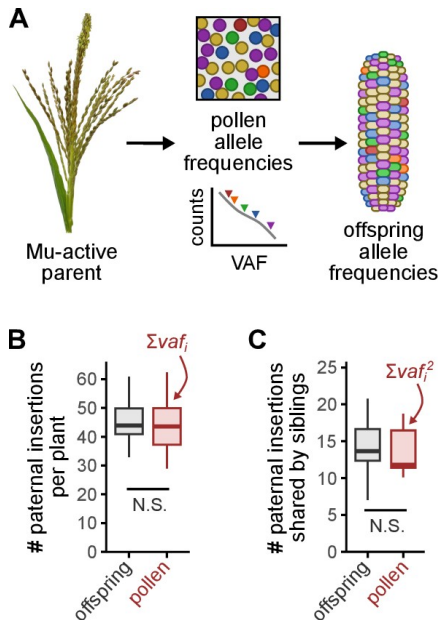

Figure 2. Pollen allele frequencies match paternal inheritance patterns.

(**A**) Allele frequencies in pollen should predict allele frequencies in the offspring.

(**B**) Number of paternal insertions per plant, measured in offspring (gray) or estimated from the pollen allele frequency distribution (red). *N.S.* not significant (p = 0.67, Mann-Whitney U test).

(**C**) Number of paternal insertions shared by two siblings, measured in offspring (gray) or estimated from the pollen allele frequency distribution (red). *N.S.* not significant (p = 0.42, Mann-Whitney U test). For panels B and C, N = 30 offspring, 9 pollen samples. Insertions on chromosome 9 were excluded because a reporter gene on this chromosome (*bz1*) was actively selected to maintain Mu activity, skewing allele frequencies for linked TEs.

Quantifying absolute allele frequencies of *de novo* Mu insertions. The initial output of MuSeq2 is the relative abundance of different Mu insertion sites within a sample. To convert relative abundances (UMI counts) to absolute allele frequencies (variant allele frequency; VAF), we normalized the data using the paternal insertions, which have a known allele frequency within the sample: 0.5 (heterozygous) in leaf, root and pollen and 0.33 in triploid endosperm (this normalization is insensitive to realistic Mu excision rates; **Fig. S5**). The measured allele frequencies were reproducible between technical replicates (independent libraries prepared from the same DNA; **Fig. 1D**), with strong quantitative agreement across 5 orders of magnitude (R^2 = 0.997 for all insertions; R^2 = 0.955 for *de novo* insertions). In contrast, there was no correlation in the abundance of *de novo* insertions between matched tissues from the same plant (**Fig. 1C**), reflecting their independent and recent origin. In total, MuSeq2 measured TE allele frequencies down to a detection limit of 8.3 $\times 10^{-5}$ for the median sample (1 part in 12,029).

To validate the measured allele frequencies, we leveraged the fact that allele frequencies in pollen should match those in the next generation (**Fig. 2A**). First, we asked whether the pollen data could accurately predict the number of inherited insertions per plant. The expected number of paternal insertions can be calculated as the sum of allele frequencies (Σvaf_i). From pollen, we predict an average of 46.5 paternal insertions per plant, in close agreement with the empirical value of 48.7 (**Fig. 2B**). Similarly, the expected number of paternal insertions shared by any two siblings (e.g. 17 insertions were shared by the siblings in **Fig. 1B**) can be estimated as the sum of allele frequencies squared (Σvaf_i^2). The pollen data predicts that 13.4 inherited insertions would be shared by siblings, again in agreement with the empirical value of 13.8 (**Fig. 2C**). Thus, allele frequencies measured in pollen accurately match paternal inheritance patterns in the offspring.

***De novo* Mu insertions occur at a wide range of allele frequencies.** A histogram of Mu allele frequencies for a representative leaf sample is shown in **Fig 3A**. There were 211,097 *de novo* insertions detected in this single leaf, with allele frequencies ranging from 0.28 down to $<10^{-4}$ (the detection limit of the assay). These data suggest that Mu is active throughout development, including

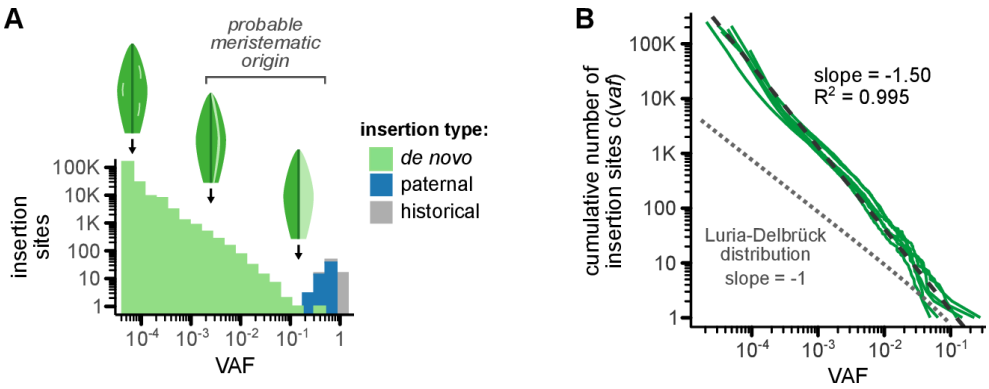

Figure 3. *De novo* Mu insertions occur across a wide range of allele frequencies.

(**A**) Histogram of Mu allele frequencies in a representative leaf sample. Colors indicate whether the insertion sites were historical, paternally inherited, or *de novo*. Cartoons on top show the potential spatial distribution of mutations at selected VAFs, estimated from sector sizes in ref. 39. Top bracket, insertions that likely originated in the meristem, based on estimate that 250 meristematic cells form a leaf primordia in maize[38]. VAF, variant allele frequency.

(**B**) Cumulative number of Mu insertion sites in individual leaf samples (N = 6). Dashed line, best linear fit to the log-log transformed data; gray dotted line, theoretical expectation for random mutation in an exponentially dividing cell population (Luria-Delbrück distribution).

insertions that likely arose in the meristem (based on their high abundance[38]; **Fig. 3A**) down to lower frequency insertions that more likely arose in the leaf itself[39]. Mu has a strong preference to insert into and immediately upstream of genes[29], targeting a reduced portion of the genome. Despite this, we did not observe saturation of the available Mu target sites (**Fig. S6**). The most abundant *de novo* Mu insertions occurred at uncorrelated, independent sites between samples (**Fig. 1C**). Thus, the specific mutations induced by Mutator were stochastic and infrequently repeated between plants. In contrast, the allele frequency distribution was highly reproducible across its entire range (**Fig. 3B**). Essentially, while the specific set of transposon insertions varied widely, a predictable number of insertions were present at any given abundance.

Mu insertion activity is much broader than for Mu excisions. The majority of prior data on somatic Mu activity is based on the excision of Mu elements in the endosperm[25,26,40], which can be observed by the appearance of revertant purple sectors after Mu excises from an anthocyanin reporter gene (**Fig. S7**). Endosperm excisions produce almost entirely small sectors[40], suggesting that Mu excision activity is highest later in development[25,26]. The excision rate also varies 1000-fold between tissues, ranging from ~10% excisions per element in endosperm[40] down to <10^{-4} excisions per element transmitted through pollen[26]. Compared to excisions, *de novo* Mu insertions were much more broadly distributed across space and time (**Fig. 4A**). There was substantial new insertion activity in every tissue type, despite large divergence in the developmental origins and biology of the selected tissues. Furthermore, *de novo* Mu insertions were observed at a wide range of allele frequencies. While Mu excisions almost never occur above an allele frequency of 0.002[40], Mu insertions were often observed beyond this limit, even within endosperm (**Fig. 4A** and **S7**). Thus, Mu insertions and excisions behave differently, with Mu insertions being more widely distributed.

Tissues show distinct allele frequency distributions for *de novo* Mu insertions. While not as dramatic as the divergence between excisions and insertions, there were significant differences in the

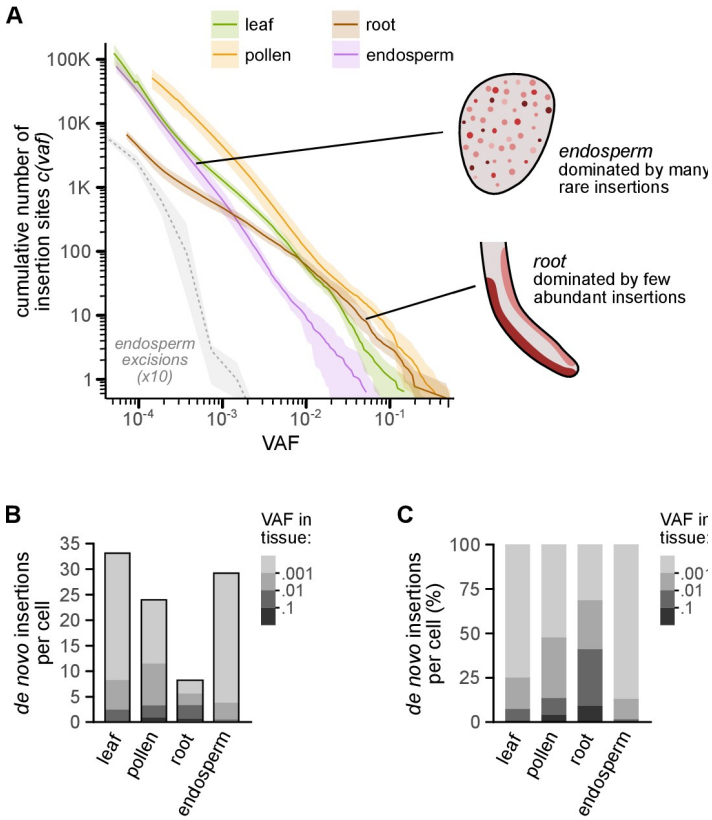

Figure 4. Allele frequency distribution of *de novo* Mu insertions across maize tissue types.

(**A**) Cumulative number of *de novo* Mu insertion sites at different allele frequencies. For insertion data (solid lines), curves show the mean and 95% confidence interval (bootstrap test). For endosperm excision data (dotted line), curve shows the reported values; shaded area, 95% confidence interval assuming Poisson counting error. Endosperm excision data is from ref. 40; the reported number of excision events was multiplied by 10 to make the insertion and excision distributions easier to compare. VAF, variant allele frequency.

(**B**,**C**) Number and % of *de novo* Mu insertions per cell, calculated from the sum of allele frequencies (Σvaf_i). Colors indicate the contribution of insertions at different allele frequencies.

behavior of *de novo* Mu insertions between tissues (**Fig. 4**). The total number of *de novo* insertions per cell varied by up to 4-fold (**Fig. 4B**), ranging from 8.2 in root to 32.8 in leaf. Moreover, each tissue had a reproducible but distinct allele frequency distribution (**Fig. 4A**). Root was most dominated by insertions at high allele frequencies (**Fig. 4B,C**), suggesting that new Mu insertions often formed relatively large sectors in this tissue. At the other extreme, endosperm had a much higher proportion of rare (low VAF) insertions. Thus, there was variation not only in the total number of Mu insertions, but also in how widespread the individual insertions were throughout the tissue.

What might contribute to the observed allele frequencies? Since early studies on bacterial mutation by Luria and Delbrück, many theoretical models of mutation accumulation have been developed. As a starting place, we compared the empirical allele frequency distributions to established theory. Leaf, pollen, and endosperm all closely followed a linear relationship on a log-log plot (a power law distribution; **Fig. 3B** and **S8**). Power-law relationships are well-known in mutation accumulation[14], as this distribution occurs in an exponentially dividing cell population subjected to a constant rate of neutral mutations (a Luria-Delbrück process). However, the empirical data was a bad fit to the Luria-Delbrück model, because the slopes were far steeper than the theoretical expectation[14] of -1 (**Fig. 3B** and **S8**). In animals, a common model for mutation accumulation is based on an exponential growth phase early in development, followed by a later, stable-population phase[9,41–44]; however, this model predicts a strong deviation from power law behavior and a shallow slope for much of the range, again a poor fit to the data (**Fig. S9**). Other models of mutation accumulation, including boundary-driven growth[13,14] (where cells divide preferentially at the edge of an expanding population), linear growth[9] (such as occurs during asymmetric stem-cell divisions), and the glandular fission model[12] (developed for solid cancer tumors) also predict sharp deviations from power law behavior. Given the complexity of multicellular development and transposon regulation, it is perhaps

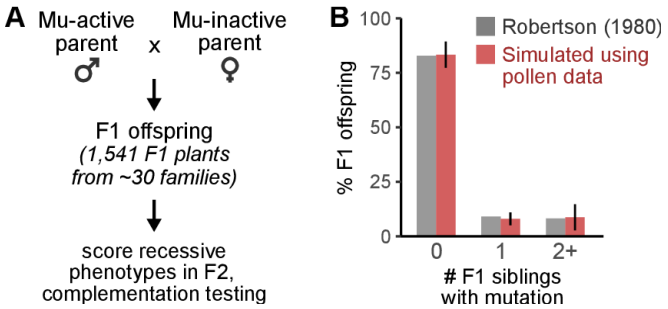

Figure 5. Pollen allele frequencies are consistent with outcross data from classical genetics.

(**A**) Experimental design from Robertson (1980). F1 offspring were generated by outcrossing a Mu-active male parent, then the offspring were assessed for appearance of visible mutant phenotypes after self-fertilization. F1 siblings segregating similar mutant phenotypes were subjected to complementation testing to determine if they shared the same (allelic) mutation.

(**B**) The experimental design in A was simulated using mutant alleles randomly drawn with probabilities matching the measured pollen allele frequencies. The % of F1 offspring that share mutations with 0, 1, or 2+ siblings were then calculated and compared to Robertson (1980). Error bars, standard error of the mean.

unsurprising that established theory cannot explain the data. The availability of quantitative allele frequency data across several orders of magnitude can inform and constrain future theoretical developments to understand mutation accumulation in plants.

Mu outcross experiments can be explained by cell division statistics. The classic view has been that Mu is most active late in germinal development[25–28], with activity peaking around the time of meiosis or during pollen maturation. In contrast, our data suggests that Mu insertions occur rather continuously throughout development (**Fig. 4A**). The most direct evidence for Mu activity late in germinal development comes from a study by Robertson[27], in which he outcrossed Mu-active plants and characterized new mutations in the F1 offspring (**Fig. 5A**). He identified 177 mutant F1 plants that segregated recessive seedling phenotypes; then, through extensive complementation testing, determined that 82.8% of the F1 offspring had unique mutations. The frequent occurrence of unique mutations among the offspring led to the idea that Mu must be most active late in development[25–28].

To reconcile these results, we directly compared our data to Robertson (1980). We previously showed that pollen allele frequency data can predict inheritance patterns in the offspring (**Fig. 2**); this approach can also be used to predict more complex experimental designs, such as Robertson's. We simulated Robertson's experiment 1,000 times, randomly drawing new mutations at probabilities defined by the bulk pollen data (see *Methods*). On average, the simulations predict that 83.3% of F1 offspring would have unique mutations (**Fig. 5B**), in close agreement with the reported value ($p = .83$, two-tailed bootstrap test). An advantage of the simulated experiments is that it is possible to computationally go 'back in time' and see how abundant any given mutation was in the Mu-active parent (**Fig. S10**). For offspring that shared mutations with 2+ siblings, the source mutations had an average allele frequency of 0.13 in pollen (consistent with a mutation at the time the seed was planted[45]); thus, the fact that Robertson observed any such offspring (8.2% of the total) suggests that early Mu activity occurred at an appreciable rate in this experiment.

Here, we can provide an alternative explanation for Robertson's data: in a dividing cell population, most mutations will be rare simply because there are more cells later in development and therefore more opportunities for a mutation to occur. While there is one chance for a mutation in the zygote,

there are two in the following division, then four, and so forth. This effect will lead to increasing numbers of mutations at decreasing allele frequencies, as was observed in all tissues for Mu insertions (**Fig. 3A**) and is even predicted by the Luria-Delbrück distribution (**Fig. 2B**). Rather than evidence for tissue-specific activity, the preponderance of unique mutations in Mu outcross families is better explained by the statistics of cell division.

DISCUSSION

De novo mutations are difficult to identify because they can be extremely rare within a tissue. This has led to an acute depth-vs-breath trade-off[46], where mutations can either be sequenced to lower depth across the genome or at higher depth for targeted loci. Here, we overcame this technical barrier with a strategic model system – the maize Mu transposon. We show that Mu sequencing can accurately measure the absolute allele frequency of *de novo* TE insertions genome-wide, while achieving a detection limit rivaling the most sensitive of targeted mutation studies[18].

As a model of mutation, a limitation of our approach is that it is only applicable to TEs; however, several findings are likely to be representative of other mutation classes. First, there were a large number (>100,000) of *de novo* Mu insertions per sample. If single-base substitutions could be sequenced to the same depth, a similar number of events might be expected. It has been estimated that every gene is mutated multiple times in an organism the size of maize or humans[9], a prediction consistent with data from deep sequencing single-genes[18]. The number of *de novo* Mu insertions per cell was similar to the germline single-base substitution rate in *Arabidopsis*[35] and far below the number of single-base substations per somatic cell in animals[34]. Thus, Mu simply provides a glimpse into the scope of genetic mosaicism for an organism with a cell population measured in the trillions.

Second, most *de novo* mutations were present at low allele frequencies. A strong trend towards low frequency mutation is expected from the statistics of cell division, as there are exponentially more cells later in development and thus more chances for mutations to occur. While individually rare, these mutations can collectively add up to important effects and may contribute to aging, cancer, and evolution. Finally, tissues varied not only in the number of mutations per cell, but also in how widespread the mutations were. When considering the rise and spread of *de novo* mutations, it will be important to recognize that multicellular organisms are large, complex populations with extensive heterogeneity.

Our results provide greater resolution into Mu activity across maize tissues. Mu insertions have been observed in somatic tissues such as leaf[29], but quantitative data on their number and abundance were not available. We found that Mu insertions occurred continuously throughout development in both somatic and germinal tissue. This is in contrast to Mu excisions, as there is a clear bias against early excision activity[40] and a >1,000-fold range in excision rates between endosperm[40] and pollen[26] (vs. 4-fold maximum range for *de novo* insertions). This provides further evidence that Mu insertions can be decoupled from excision outcomes, perhaps due to tissue-specific differences in the use of DNA repair to restore a Mu element after a transient excision event[26].

What might drive the tissue-specific variation in Mu allele frequencies? Differences in transposon activity may contribute, but are not the only explanation. For instance, spatial biases in cell division rates have a profound impact on mutant allele frequencies[9,11–13], and so differences in tissue

development may contribute to the patterns we observed. Selection for and against specific mutations has been observed in healthy human tissues[16], and might similarly impact the persistence or spread of *de novo* Mu insertions. Future work can dissect the relative contribution of tissue-specific transposon activity, cell division patterns, selection, and other processes on the ultimate abundance of *de novo* mutations within and across the plant.

ACKNOWLEDGMENTS

We thank Jonathan Gent, Robert Erdmann, Bob Schmitz, and Chris McFarland for invaluable discussions and critical reading of the manuscript. We thank Grant Freeman for initial testing of the MuSeq2 protocol. We thank the Duke University Sequencing and Genomics Technologies Core for sequencing services. Funding was provided by NIH grant R35GM151237 to B.N.

AUTHOR CONTRIBUTIONS

Conceptualization, B.N.; Methodology, J.S. and B.N.; Visualization, B.N.; Software, J.S. and B.N.; Formal Analysis, M.H. and B.N.; Investigation, J.S.; Writing – Original Draft, B.N.; Writing – Review & Editing, J.S., M.H., and B.N.

REFERENCES

1. Abyzov, A., Tomasini, L., Zhou, B., Vasmatzis, N., Coppola, G., Amenduni, M., Pattni, R., Wilson, M., Gerstein, M., Weissman, S., et al. (2017). One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. *27*, 512–523. 10.1101/gr.215517.116.

2. Schmid, M., Lee, J.S., Ahn, J.H., Amasino, R.M., Johnson, K.A., Chowdhury, N.I., Braam, J., Fitter, R.S., Wolkovich, E.M., Parmesan, C., et al. (2013). Mosaic copy number variation in human neurons. Science *342*, 632–638.

3. Schmitt, S., Heuret, P., Troispoux, V., Beraud, M., Cazal, J., Chancerel, É., Cravero, C., Guichoux, E., Lepais, O., Loureiro, J., et al. (2024). Low- frequency somatic mutations are heritable in tropical trees Dicorynia guianensis and Sextonia rubra. Proc. Natl. Acad. Sci. *121*, e2313312121. 10.1073/pnas.

4. Amundson, K.R., Prem, M., Marimuthu, A., Nguyen, O., Sarika, K., Demarco, I.J., Phan, A., Henry, I.M., and Comai, L. (2023). Differential mutation accumulation in plant meristematic layers. bioRxiv, 2023.09.25.559363.

5. Grimes, K., Jeong, H., Amoah, A., Xu, N., Niemann, J., Raeder, B., Hasenfeld, P., Stober, C., Rausch, T., Benito, E., et al. (2024). Cell type-specific consequences of mosaic structural variants in hematopoietic stem and progenitor cells. Nat. Genet. *56*. 10.1038/s41588-024-01754-2.

6. Siudeja, K., Beek, M., Riddiford, N., Boumard, B., Wurmser, A., Stefanutti, M., Lameiras, S., and Bardin, A.J. (2021). Unraveling the features of somatic transposition in the Drosophila intestine. EMBO J. *40*, 1–19. 10.15252/embj.2020106388.

7. Moore, L., Cagan, A., Coorens, T.H.H., Neville, M.D.C., Sanghvi, R., Sanders, M.A., Oliver, T.R.W., Leongamornlert, D., Ellis, P., Noorani, A., et al. (2021). The mutational landscape of human somatic and germline cells. Nature *597*, 381–386. 10.1038/s41586-021-03822-7.

8. Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B.J., Venturini, E., et al. (2018). Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science *359*, 550–555. 10.1126/science.aan8690.

9. Frank, S.A. (2010). Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. Proc. Natl. Acad. Sci. U. S. A. *107*, 1725–1730. 10.1073/pnas.0909343106.

10. Shahriyari, L., and Komarova, N.L. (2015). The role of the bi-compartmental stem cell niche in delaying cancer. Phys. Biol. *12*. 10.1088/1478-3975/12/5/055001.

11. Chkhaidze, K., Heide, T., Werner, B., Williams, M.J., Huang, W., Caravagna, G., Graham, T.A., and Sottoriva, A. (2019). Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLoS Comput. Biol. *15*. 10.1371/journal.pcbi.1007243.

12. Noble, R., Burri, D., Le Sueur, C., Lemant, J., Viossat, Y., Kather, J.N., and Beerenwinkel, N. (2022). Spatial structure governs the mode of tumour evolution. Nat. Ecol. Evol. *6*, 207–217. 10.1038/s41559-021-01615-9.

13. Lewinsohn, M.A., Bedford, T., Müller, N.F., and Feder, A.F. (2023). State-dependent evolutionary models reveal modes of solid tumour growth. Nat. Ecol. Evol. *7*, 581–596. 10.1038/s41559-023-02000-4.

14. Fusco, D., Gralka, M., Kayser, J., Anderson, A., and Hallatschek, O. (2016). Excess of mutational jackpot events in expanding populations revealed by spatial Luria-Delbrück experiments. Nat. Commun. *7*, 12760. 10.1038/ncomms12760.

15. Postek, W., Staskiewicz, K., Lilja, E., and Waclaw, B. (2024). Substrate geometry affects population dynamics in a bacterial biofilm. Proc. Natl. Acad. Sci. *121*, e2315361121. 10.1073/pnas.2315361121.

16. Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R., et al. (2018). Somatic mutant clones colonize the human esophagus with age. Science *362*, 911–917. 10.1126/science.aau3879.

17. Young, A.L., Spencer Tong, R., Birmann, B.M., and Druley, T.E. (2019). Clonal hematopoiesis and risk of acute myeloid leukemia. Haematologica *104*, 2410–2417. 10.3324/haematol.2018.215269.

18. Salazar, R., Arbeithuber, B., Ivankovic, M., Heinzl, M., Moura, S., Hartl, I., Mair, T., Lahnsteiner, A., Ebner, T., Shebl, O., et al. (2022). Discovery of an unusually high number of de novo mutations in sperm of older men using duplex sequencing. Genome Res. *32*, 499–511. 10.1101/gr.275695.121.

19. Bae, J.H., Liu, R., Roberts, E., Nguyen, E., Tabrizi, S., Rhoades, J., Blewett, T., Xiong, K., Gydush, G., Shea, D., et al. (2023). Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. Nat. Genet. *55*, 871–879. 10.1038/s41588-023-01376-0.

20. Waneka, G., Pate, B., Monroe, J.G., and Sloan, D.B. (2024). Investigating low frequency somatic mutations in Arabidopsis with Duplex Sequencing. bioRxiv, 2024.01.31.578196. 10.1101/2024.01.31.578196.

21. Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. Science *338*, 758–767. 10.1126/science.338.6108.758.

22. Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. *43*, 1160–1163. 10.1038/ng.942.

23. Cosby, R.L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E.J., and Feschotte, C. (2021). Recurrent evolution of vertebrate transcription factors by transposase capture. Science *371*, eabc6405. 10.1126/science.abc6405.

24. McCarty, D.R., Latshaw, S., Wu, S., Suzuki, M., Hunter, C.T., Avigne, W.T., and Koch, K.E. (2013). Mu-seq: Sequence-based mapping and identification of transposon induced mutations. PLoS One *8*, e77172. 10.1371/journal.pone.0077172.

25. Lisch, D. (2015). Mutator and MULE transposons. Mob. DNA III, 801–826. 10.1128/9781555819217.ch36.

26. Lisch, D. (2002). Mutator transposons. Trends Plant Sci. *7*, 498–504. 10.1016/S1360-1385(02)02347-6.

27. Robertson, D.S. (1980). The timing of Mu activity in maize. Genetics *94*, 969–978.

28. Robertson, D.S. (1981). Mutator activity in maize: Timing of its activation in ontogeny. Science *213*, 1515–1517.

29. Zhang, X., Zhao, M., McCarty, D.R., and Lisch, D. (2020). Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. Nucleic Acids Res. *48*, 6685–6698. 10.1093/nar/gkaa370.

30. Lisch, D., Chomet, P., and Freeling, M. (1995). Genetic characterization of the mutator system in maize: Behavior and regulation of Mu transposons in a minimal line. Genetics *139*, 1777–1796.

31. Liu, P., McCarty, D.R., and Koch, K.E. (2016). Transposon mutagenesis and analysis of mutants in UniformMu maize (Zea mays). Curr. Protoc. Plant Biol. *1*, 451–465. 10.1002/cppb.20029.

32. Marcon, C., Altrogge, L., Win, Y.N., Stöcker, T., Gardiner, J.M., Portwood, J.L., Opitz, N., Kortz, A., Baldauf, J.A., Hunter, C.T., et al. (2020). BonnMu: A sequence-indexed resource of transposon-induced maize mutations for functional genomics studies. Plant Physiol. *184*, 620–631. 10.1104/pp.20.00478.

33. Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74. 10.1038/nmeth.1778.

34. Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S. V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y., et al. (2021). Somatic mutation landscapes at single-molecule resolution. Nature *593*, 405–410. 10.1038/s41586-021-03477-4.

35. Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science *327*, 92–94. 10.1126/science.1180677.

36. Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., Wang, B., Liu, Z., Chen, J., Li, W., et al. (2012). Genome-wide genetic changes during modern breeding of maize. Nat. Genet. *44*, 812–815. 10.1038/ng.2312.

37. Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., et al. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature *549*, 519–522. 10.1038/nature24018.

38. Poethig, S. (1984). Cellular parameters of leaf morphogenesis in maize and tobacco. In Contemporary Problems in Plant Anatomy, pp. 235–259.

39. Langdale, J.A., Lane, B., Freeling, M., and Nelson, T. (1989). Cell lineage analysis of maize bundle sheath and mesophyll cells. Dev. Biol. *133*, 128–139.

40. Levy, A., and Walbot, V. (1990). Regulation of the timing of transposable element excision during maize development. Science *248*, 1534–1537.

41. Frank, S.A., and Nowakt, M.A. (2003). Developmental predisposition to cancer. Nature *422*, 494. 10.1038/422494a.

42. Meza, R., Jeon, J., Moolgavkar, S.H., and Georg Luebeck, E. (2008). Age-specific incidence of cancer: Phases, transitions, and biological implications. Proc. Natl. Acad. Sci. U. S. A. *105*, 16284–16289. 10.1073/pnas.0801151105.

43. Moeller, M.E., Mon Père, N. V., Werner, B., and Huang, W. (2024). Measures of genetic diversification in somatic tissues at bulk and single-cell resolution. Elife *12*, RP89780. 10.7554/eLife.89780.

44. Gunnarsson, E.B., Leder, K., and Foo, J. (2021). Exact site frequency spectra of neutrally evolving tumors: A transition between power laws reveals a signature of cell viability. Theor. Popul. Biol. *142*, 67–90. 10.1016/j.tpb.2021.09.004.

45. Poethig, R.S., Coe, E.H., and Johri, M.M. (1986). Cell lineage patterns in maize embryogenesis: A clonal analysis. Dev. Biol. *117*, 392–404. 10.1016/0012-1606(86)90308-8.

46. Gydush, G., Nguyen, E., Bae, J.H., Blewett, T., Rhoades, J., Reed, S.C., Shea, D., Xiong, K., Liu, R., Yu, F., et al. (2022). Massively parallel enrichment of low-frequency alleles enables duplex sequencing at low depth. Nat. Biomed. Eng. *6*, 257–266. 10.1038/s41551-022-00855-9.

47. Springer, N.M., Anderson, S.N., Andorf, C.M., Ahern, K.R., Bai, F., Barad, O., Barbazuk, W.B., Bass, H.W., Baruch, K., Ben-Zvi, G., et al. (2018). The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat. Genet. *50*, 1282–1288. 10.1038/s41588-018-0158-0.

48. deepTools: Effective genome size https://deeptools.readthedocs.io/en/develop/content/feature/effectiveGenomeSize.html.

49. Rockweiler, N.B., Ramu, A., Nagirnaja, L., Wong, W.H., Noordam, M.J., Drubin, C.W., Huang, N., Miller, B., Todres, E.Z., Vigh-Conrad, K.A., et al. (2023). The origins and functional effects of postzygotic mutations throughout the human life span. Science *380*, eabn7113. 10.1126/science.abn7113.

50. Muyas, F., Sauer, C.M., Valle-Inclán, J.E., Li, R., Rahbari, R., Mitchell, T.J., Hormoz, S., and Cortés-Ciriano, I. (2024). De novo detection of somatic mutations in high-throughput single-cell profiling data sets. Nat. Biotechnol. *42*, 758–767. 10.1038/s41587-023-01863-z.

51. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. 10.1093/bioinformatics/bty560.

52. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. 10.1038/nmeth.1923.

53. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

54. Eden, M. (1961). A two-dimensional growth process. Dyn. Fractal Surfaces *4*, 598.

METHODS

Literature survey on VAF detection limit vs. genomic coverage

For **Fig. S1A**, only papers that reported VAFs in the main text or figures were considered. The selected studies were not meant to be exhaustive, but rather representative of a range of techniques and mutation types. VAF detection limits were as given in each paper; when multiple limits were provided for different sample types (e.g. tissues), the lowest reported value was used. For targeted sequencing studies, genomic coverage was calculated from the reported target size divided by the mappable genome size[48]: 2,864,785,220 bp for human and 119,482,012 bp for *Arabidopsis*. For the purposes of the figure, whole-genome sequencing papers were considered to have 100% genomic coverage.

Two classes of study were not included: First, studies that identified *de novo* mutations from transcript data were excluded (e.g. ref [49]). RNA-seq produces very uneven read depths across the genome, and so there is not a well-defined relationship between VAF detection limit and genome coverage – highly expressed genes, which represent a small portion of the transcriptome, dominate the minimum observed VAFs. Single-cell mutation studies were also not included[50]. This was not due to active exclusion, but rather because available single-cell mutation papers largely do not report VAF detection limits. While single-cell methods provide additional cell-type resolution and information about cell-to-cell variation, they do not fundamentally overcome the limitations on sequencing depth and error rates that are also present in bulk experiments.

Plant growth and tissue collection

Mu-active plants were maintained by continual outcrossing of Mu-active pollen onto Mu inactive ears, using the bz1-Mum9 anthocyanin reporter to confirm Mu activity. The Mu-inactive maintainer (female) parents were descended from maize Co-op stock 910I, and carry the sh1-bb1981 and bz1-m4::Ds alleles. Mu-active seeds were descended from maize Co-op stock 919J, which carries a mutable bz1-Mum9 allele. Both stocks were originally ordered from the Stock Center in January 2010 by Jonathan Gent. Continual outcrossing of Mu active lines onto Mu inactive ears was required to maintain Mu activity. Mu active kernels were phenotypically identified by the speckling pattern that occurs when Mu somatically excises from the bz1-Mum9 allele.

For tissue collection, kernels were chipped using a razor blade and the resulting endosperm samples were stored in 2 mL tubes and frozen. To minimize sample cross-contamination, the surface that kernels were chipped on was wiped with a 10% bleach solution and razor blades were only used once. Chipped kernels were then planted in vermiculite (Therm-O-Rock Vermiculite 3A-HORT Medium). After the second seedling leaf was fully emerged (V2 stage; 10-13 days after planting), plants were removed from vermiculite. Roots were rinsed thoroughly in water to remove any vermiculite, and ~1 inch of the primary root was collected into 2 mL tubes and frozen. The topmost half of the first leaf (~3/4 inch) was also harvested and collected into 2 mL tubes and frozen. Seedlings were then transplanted to soil in the Botany Greenhouses in Athens, Georgia, where they were grown in sunlight supplemented with LED fluorescent lights (Medic Grow 550W Slim Power 2) until maturity. At maturity, pollen was collected into 2 mL tubes in the morning from the plants at first pollen shed (9-10 am) and frozen.

DNA isolation

Leaf DNA isolation: Leaf tissue was disrupted in a 2 mL tube using liquid nitrogen and a pestle (Agilent cat. no. PES-15-B-SI). Once disrupted, DNA was extracted with the Qiagen DNeasy Plant Mini Kit (Qiagen cat. no. 69104) and eluted in two steps, first with 30 ul of elution buffer followed by 25 uL elution buffer. DNA size distributions were evaluated using 5 ul of sample on a 0.8% agarose gel.

Root DNA isolation: Root tissue was disrupted with a Qiagen Tissue Lyser II and three to six 3 mm glass beads per sample. Prior to disruption, the sample box was chilled overnight at -80 °C with root samples and 3 mm glass beads inside. Pre-chilled root tissue was then shaken in the Tissue Lyser at max frequency for 5 minutes. Samples were removed from the shaker and agitated using a pestle to dislodge root debris from the tube walls. Shaking was repeated at max frequency for 5 minutes. After this process, root DNA was extracted as described for leaf isolation using the Qiagen DNeasy Plant Mini Kit. DNA size distributions were evaluated on a 0.8% agarose gel.

Endosperm DNA isolation: Genomic DNA was isolated from endosperm using a modified CTAB DNA extraction protocol. To prepare CTAB buffer, CTAB stock was made with 3% Cetyltrimethyl ammonium bromide, 1.4 M NaCl, 20 mM EDTA (pH 8.0), 100 mM Tris-Cl (pH 8.0). The day of DNA extractions, 2% w/v polyvinylpyrrolidone (PVP, MW 40 kDa) was dissolved into CTAB stock by heating the solution to 65 °C, and then 800 ul preheated lysis buffer was aliquoted into tubes (one tube per sample to be processed). Then 8 ul proteinase k (ThermoFisher cat. no. EO0491) and 1 ul beta-mercaptoethanol were added to each tube.

Endosperm tissue was disrupted using liquid nitrogen, mortar, and pestle. Disrupted tissue was then incubated at 65º C for 1 h in the preheated lysis buffer. Samples were inverted to mix every 10 minutes during incubation. Following this, samples were spun down at 5000 rcf for 8 mins to pellet tissue debris. Lysate was transferred to a new tube using a metal spatula and combined with 1 volume of a 24:1 chloroform isoamyl alcohol solution. Samples were mixed by inversion for 5 minutes and then centrifuged at 8,000 rcf for 10 minutes. The upper aqueous phase was carefully transferred to a new tube following centrifugation. To precipitate DNA, 0.7X volumes of cold isopropanol was added to each sample and inverted to mix. Samples were incubated at -20 °C for 1 hour. Samples were then centrifuged at 10,000 rcf for 15 minutes. The supernatant was removed and the DNA pellet was washed using 1000 ul of freshly prepared 70% ethanol. Samples were inverted to mix and incubated at room temperature for 5 minutes. Samples were then centrifuged for 5 minutes at 10,000 rcf. The ethanol wash was repeated one more time and DNA pellets were dried until the pellet became translucent. The DNA pellet was resuspended using 55 ul of ultra-pure H2O (ThermoFisher cat. no. 10977015) and incubated overnight at 4 °C. Size distributions were visualized using 5 ul of purified DNA on a 0.8% agarose gel. The first batch of endosperm samples showed signs of cross-contamination; these data were used to identify paternal insertions but excluded from all other analyses (see ‘Sample assessment and quality control’). Prior to processing subsequent endosperm samples, the mortar, pestle, and metal spatula were incubated for 5 min in 10% bleach solution and then thoroughly rinsed with water; this additional washing step removed the cross-contamination.

Pollen DNA isolation: Pollen was disrupted using a Qiagen Tissue Lyser II as described for root. During disruption, pollen debris would stick to the lid of the tubes and so a pestle was used to scrape off the debris into the tube. After disruption, 800 ul of preheated CTAB lysis buffer (prepared as

described for endosperm) was added to each sample. A pestle was again used to scrape off any material from the tube lid back into the tube as well as break up any pellet that had formed in the bottom of the tube. This step ensured a homogenous mixture during lysis, which greatly increased DNA quality and quantity. DNA extractions were then performed as described for endosperm, with the additional of third ethanol wash after DNA precipitation. The first batch of pollen samples had much lower sequencing depth compared to the other tissues (lower UMIs / sample). For subsequent samples, pollen DNA was purified an additional time with a Monarch DNA and PCR cleanup kit (New England Biolabs cat. no. T1030S); the DNA cleanup kit was performed after CTAB extraction and DNA shearing, prior to the end repair step in MuSeq2.

MuSeq2 adapter preparation

MuSeq2 adapter oligos (**Table S2**) were ordered from Integrated DNA Technologies, suspended to 100 uM in TE. The general adapter structure is as follows:

5'-[phos]rrrrrrrrrrUGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNbbbbbbbbbbT-3'

The 10 nt sequences labeled as strings of 'r' and 'b' are reverse complements of each other, allowing the adapter to form a hairpin with a 3' T overhang. These sequences vary by adapter (**Table S2**), providing a sample-specific barcode during adapter ligation. The series of 'N's is the 10 nt Unique Molecular Identifier (UMI). A uracil (U) near the 5' end makes it possible to cut the hairpin after adapter ligation. Read 2 begins at the UMI and continues through the sample barcode and subsequent genome sequence.

To prepare MuSeq2 adapters and anneal the hairpin, 7.5 ul of adapter oligo (100 uM) was diluted with 25 ul Duplex Buffer (Integrated DNA Technologies cat. no. 11-05-01-03) and 17.5 ul H2O. Diluted adapters were then placed in a thermocycler and incubated at 95ºC for 2 minutes followed by a .1º C ramp down in 1-second intervals for 700 cycles, reaching a final temperature of 10ºC. Adapters were then stored at -80ºC.

MuSeq2 library preparation

DNA samples were sheared using a Covaris E220 Evolution instrument in 50 uL of water. Shearing settings were optimized for each tissue to shear to a mean of 1000 bp. All tissues used settings of 2% Duty Factor with 200 cycles per burst. For pollen, the peak incident power was 140 and time was 50 seconds; Endosperm: 100 Peak Incident Power and 30 seconds; Leaf: 70 Peak Incident Power and 30 seconds. Root: 100 Peak Incident Power and 20 seconds. Concentrations and size distributions for sheared DNA was measured using an Agilent 4200 TapeStation with a D5000 screentape (Agilent cat. no. 5067-5589).

Sheared DNA (200-1000 ng / sample) was end-repaired using the NEBNext Ultra II DNA Library Preparation Kit (New England Biolabs cat. no. E7370L) according to manufacturer instructions, except that all reaction volumes were cut in half. MuSeq2 adapters were then ligated to the DNA using the same kit (NEBNext Ultra II) with half reaction volumes; a separate adapter was used for each sample, providing up to 48 sample-specific barcodes during the initial ligation step. After ligation, 1.5 uL USER enzyme (New England Biolabs cat. no. M5505S) and 2 uL Exonuclease 1 (New England Biolabs cat. no. M0293S) were added to each sample and the reaction was incubated at 37 for 15 min then 80 °C

for 15 min. This step linearizes the hairpin adapters by cleaving at a uracil base, and the addition of Exonuclease 1 degrades residual unligated adapter to minimize carryover in subsequent PCR. Samples were then purified with Ampure XP Beads (Beckman Coulter cat. No A63880) using a bead:sample ratio of 0.8X. After bead purification, libraries were resuspended in 5 ul ultra-pure H2O.

Adpter ligated libraries were processed through 3 rounds of PCR to selectively amplify Mu-containing fragments and complete the Illumina adapter sequences (PCR primer sequences in **Table S3**). For the first PCR, 5 ul sample was mixed with 6 ul NEBNext Ultra II Q5 Master Mix (New England Biolabs cat. no. M0544S), 0.5 ul TIR6 primer (4.8 uM stock concentration; 0.2 uM final), and 0.5 ul UDz_i7 primer (4.8 uM stock concentration; 0.2 uM final). Reactions were pipetted to mix and incubated at 98 °C for 30 s, then 14 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72°C for 30 s, followed by 72 °C for 2 min. To remove excess primers, 0.5 ul Exonuclease I was added and the tube was incubated at 37 °C for 15 min then 80 °C for 15 min.

For PCR2, an additional 4 ul of Q5 master mix was added along with 0.4 ul of Museq2_NestedTIR primer (10 uM stock), 0.4 ul of P7 primer (10 uM stock), and 2.7 ul of ultra-pure H2O. Reactions were pipetted to mix and incubated at 98 °C for 30 s, then 6 cycles of 98 °C for 10 s, 59 °C for 30 s, and 72°C for 30 s, followed by 72 °C for 2 min. To remove excess primers, 0.5 ul Exonuclease I was added and the tube was incubated at 37 °C for 15 min then 80 °C for 15 min.

For PCR3, 5 ul of PCR2 product was mixed with 25 ul Q5 master mix, 19.5 ul ultra-pure H2O, and 5.5 ul xGen indexed primer pairs (Integrated DNA Technologies xGen UDI Primers Plate 1, cat. no. 10005922). A distinct primer pair was added to each sample to allow for multiplexing. Reactions were pipetted to mix and incubated at 98 °C for 30 s, then 5-15 cycles of 98 °C for 10 s, 59 °C for 30 s, and 72°C for 30 s, followed by 72 °C for 2 min. The number of PCR cycles varied by sample and was determined using qPCR as follows: prior to PCR3, 5 ul of the prepared PCR reaction was withdrawn and mixed with 0.5 ul of a 1:1000 dilution of SYBR Green I DNA Gel Stain (Thermo Fisher cat. no. S7563) in water. The reaction aliquot with SYBR was then run on a BioRad CFX96 Real Time PCR Thermal Cycler for 25 cycles using the reaction conditions listed above. The cycle number for each PCR reaction was chosen to be within 30-80% of the plateau height, and the remaining PCR mix was run using the selected cycle number.

After PCR, libraries were cleaned up and size selected using magnetic beads. For cleanup, 50 ul Ampure XP beads were added to each PCR sample (1X ratio) and then the DNA was purified according to manufacturer instructions and eluted in 40 ul ultra-pure H2O. Size selection was then performed using SPRIselect beads (Beckman Coultier cat. no. B23317) with 0.6/0.8 bead ratios according to manufacturer instructions. DNA was eluted in a final volume of 20 ul and the size distribution and concentration was measured using an Agilent 4200 TapeStation with a D5000 screentape. Libraries were pooled to 15 nM such that each individual library was equally represented. Paired-end 150 bp sequencing was performed at the Duke University Sequencing and Genomics Technologies Core on an Illumina NovaSeq X Plus instrument with 20% PhiX spike-in.

<u>Mu insertion mapping and quantification</u>

For MuSeq2 libraries, read 1 contains the last 29 bp of the Mu transposon followed by genomic DNA sequence. Read 1 was first pre-processed by removing the first 23 bp, which contain transposon sequence matching the PCR primer, and moving bp 24-29 (the ‘validation sequence’) to the read

header using Fastp v0.23.4[51] with all filters disabled. Read 2 contains the adapter ligated fragment with an 8 bp UMI, 11 bp sample-specific barcode, and genomic sequence. Read 2 was pre-processed using Fastp to move both the 8 nt UMI and 11 nt sample-specific barcode to the read header. During this step, adapter sequences were also trimmed and fragments with under 40 bp in read 2 were removed. Paired-end reads were then mapped to the W22 V2 genome[47] using Bowtie2 v2.5.4[52], with both mixed and discordant mapping disabled (--no-mixed --no-discordant). The UMI-tools v1.1.6[53] 'group' function was used to identify reads sharing the same UMI while accounting for sequence errors.

Next, fragments were filtered using a custom R script to remove low quality mapping and PCR duplicates. Fragments were excluded if they had a mapping quality score <10 or a validation sequence that did not match 'TRTCTC' (the sequence at the edge of the Mu transposon). They were further excluded if they did not have an exact match to the sample-specific barcode added during adapter ligation. To remove PCR duplicates, fragments mapping to the same position (both read 1 and 2) with the same UMI were merged. The libraries in this study were intentionally over-sequenced to increase the amount of error-correcting from molecular counting, with a mean of 5.2 sequenced fragments per molecule (UMI). Each molecule was required to have a minimum support of at least 1/5 the average number of reads for a given sample; for the median library, this means that each molecule (UMI) was sequenced with a minimum of two reads.

Most Mu elements contain an intact Terminal Inverted Repeat (TIR) at both ends of the transposon, and so can be sequenced out of each direction. To connect molecules mapping to the left or right border of the same Mu element, the following steps were taken: First, for Mu elements present in the reference genome (N = 20; all historical insertions), the left and right borders were defined based on the genome sequence. For other Mu elements, the left and right borders were connected by expecting a 9 bp target site duplication (TSD) to be generated during Mu insertion; this would result in both ends of the transposon mapping 8 bp apart in reverse orientation (for a 9 bp TSD, there is an 8 bp distance between positions 1 and 9). To allow for discrepancies in TSD length, we searched for cases where a left and right border were within 15 bp of the expected TSD, but where both borders had over 50-fold more molecule counts than the corresponding border 8 bp away. In such cases, the two borders with higher counts were considered to come from the same element. Deviations from the 9 bp TSD were rare, with only 273 such instances identified compared to more than 3 million with the 'ideal' 9 bp TSD.

The total number of molecule counts mapping to either the left or right transposon border were then added to provide a single estimate for each element. A subset of elements were not sequenced effectively out of both directions, which could result in under-counting as there was only one border available for sequencing instead of the usual two. To adjust for this effect, we identified any elements where there was a greater than 2-fold difference in molecule counts between the left and right border after adding a pseudocount of 500. For these elements, the number of molecules was estimated as 2 times the greater of the left or right border counts. This process affected 422 elements (0.00013%). Finally, 19 elements were 'blacklisted' and removed from analysis (**Table S4**) because they were identified at moderate abundance (between 10-1000 counts per million) in over half of all samples or half of the Mu-inactive controls; many of the blacklisted sites were ancestral Mu elements with diverged sequences and would not be expected to amplify efficiently during MuSeq.

Estimating variant allele frequencies from Mu count data

To convert Mu insertion counts to variant allele frequencies (VAF), the data for each sample was first scaled to counts per million (CPM). Paternal insertions were identified as insertion sites with ≥1000 CPM in both endosperm and at least one matched sporophytic tissue (leaf, root, or pollen), excluding the 29 historical insertions (**Table S1**). CPM data were then normalized to VAF by dividing each sample by the mean CPM of the paternal insertions and multiplying by 1/2 (for leaf, root, and pollen) or 1/3 (for endosperm); the difference in normalization factor for endosperm is because the endosperm is triploid with a 2:1 maternal:paternal ratio, and so paternal DNA makes up 1/3 of the DNA in this tissue. The random error for normalization is estimated to be 6.2% (standard error of the mean for the paternal insertion sites).

Sample assessment and quality control

In total, 46 Mu-active samples were collected and sequenced for this study. All samples were used when identifying inherited insertions (paternal and historical insertion sites), but 17 were excluded from further analysis because of concerns with library quality: 6 samples were excluded because they did not meet a minimum VAF detection threshold of 10^{-3}. This set included the first 5 pollen samples, which had consistently low molecule counts, and one endosperm sample. Subsequent pollen libraries incorporated an additional round of DNA purification (see 'DNA isolation' section, above) and resulted in much better sequencing depth. Second, the first 11 endosperm libraries were excluded because they showed evidence of cross-contamination; in these libraries, paternal insertions from one library consistently showed unusually high abundance in the others. For subsequent endosperm libraries, all non-disposable items used for tissue disruption (mortar, pestle, metal spatula) were subjected to a more stringent washing protocol that included soaking in 10% bleach for 5 min (see 'DNA isolation' section, above); this additional cleaning step resolved the cross-contamination issue. These sample exclusion criteria were set prior to analyzing the data further.

Interpreting the allele frequency distribution of *de novo* Mu insertions

To estimate the mean allele frequency distribution for each tissue (e.g. **Fig. 4A**), the cumulative number of *de novo* Mu insertion sites at or above a given VAF was first calculated for the individual samples. The single-sample allele frequency distributions were then log-transformed and interpolated at 200 evenly spaced points between $\log_{10}(10^{-5})$ and $\log_{10}(1)$ using the R function *approx* (R version 4.3.0). The mean and 95% confidence interval (CI95) for each tissue was then calculated by bootstrapping with 2000 bootstrap replicates. As the sequencing depth varied between samples, the mean was reported down the minimum VAF covered by at least 75% of samples in a tissue. Power-law fits to the allele frequency distributions were performed using the R *lm* function after $\log_{10}$ transformation.

Simulating the Robertson (1980) experiment from pollen allele frequency data

Robertson (1980) performed a series of outcrosses between Mu-active plants (F0) and Mu-inactive donors. The F1 progeny were then evaluated to determine if they segregated new mutations and

whether any of the mutations were shared with siblings. In total, 1541 F1 offspring were tested, of which there were 171 mutant plants (11.1%) carrying an estimated 154 distinct mutations.

To simulate the results of one Mu outcross from Robertson (1980), a MuSeq pollen sample was first randomly selected. This sample represents a single Mu-active F0 plant from Robertson's study. Mutations (Mu insertion sites) were then randomly drawn based on the measured pollen allele frequencies. For instance, a mutation with a VAF of 0.1 was drawn with a 10% chance of occurring in each F1 offspring. After simulating 50 such F1 offspring (roughly the average number of offspring evaluated per outcross in ref. [27]), the number of times a mutation occurred 1, 2, or >3 times among the offspring was recorded. Robertson's entire study had ~30 such outcrosses, for a total of 1541 F1 plants. Thus, to simulate a full iteration of Robertson's study, 30 simulated outcross experiments were performed using 30 different pollen samples (randomly sampled with replacement) and the totals were added together.

To estimate confidence intervals, it is important that the simulated study reflects the variation expected under the conditions of Robertson (1980). From the pollen data, an average of 32,038 mutations were recovered for each simulation, far more than the 154 mutations recovered by Robertson (1980). This discrepancy is explained because Robertson tracked mutations with visible seedling phenotypes, which would represent only a small portion of the total. To better match the counting noise during Robertson (1980), the simulated mutations were downsampled so that an average of 154 were recovered per simulation. This downsampling makes the simulation-to-simulation variation better matched to Robertson (1980), but does not affect the mean estimates: 83.1% of mutations were found to be unique prior to downsampling, compared to 83.3% after downsampling.

Details of theoretical model comparisons

In **Fig. 3B** and **S9** we show the experimental cumulative distribution for leaf compared to the predictions of a variety of theoretical growth models. The model details are as follows:

Exponential growth: Starting with a single cell in generation $k = 0$, we consider a simple doubling process where in the k th generation there are $d(k) = 2^k$ cell divisions. If the probability of a mutation (i.e. a transposon insertion) per cell division is μ , then on average $\mu d(k)$ novel mutations occur in the k th generation (assuming each mutation is distinct). The cumulative number of mutations that occur from generation $k = 0$ through n is $c(n) = \mu \sum_{k=0}^{n} d(k) = \mu(2^{n+1} - 1)$. Let N be the total number of generations of growth, so the total cell population is $N_{cell} = 2^N$. An insertion in the n th generation will ultimately spread to 2^{N-n} copies by the final generation, so the associated frequency is $f = 2^{N-n}/N_{cell} = 2^{-n}$. Inverting this relation, $n = -\log_2 f$, and plugging it into the expression for $c(n)$, we find our final form for the cumulative distribution as a function of frequency:

$$c(f) = \mu\left(\frac{2}{f} - 1\right),$$

consistent with the scaling $c(f) \sim f^{-1}$ at small f we expect for the case of neutral mutations under exponential growth.

Linear growth: Here we consider a process where after each cell division only one daughter cell is capable of dividing further. Hence $d(k) = 1$ and $c(n) = \mu \sum_{k=0}^{n} d(k) = \mu(n + 1)$. After N generations there are $N_{cell} = N + 1$ total cells, and the frequency achieved by an insertion in the n th generation is $f = \frac{N-n+1}{N_{cell}} = \frac{N-n+1}{N+1}$. Solving for n in terms of f and plugging into $c(n)$ we find:

$$c(f) = \mu\big((N + 1)(1 - f) + 1\big).$$

Exponential + linear growth: In this case there are M generations of exponential growth, followed by $N - M$ generations of linear growth. Following an analogous argument as the previous two cases, we find:

$$d(k) = \begin{cases} 2^k & k \le M \\ 2^M & k > M \end{cases},$$

$$c(n) = \begin{cases} \mu(2^{n+1} - 1) & n \le M \\ \mu(2^{M+1} - 1 + 2^M(n - M)) & n > M \end{cases}.$$

The total number of cells is $N_{cell} = 2^M(N - M + 1)$, and the frequency associated with an insertion in the n th generation is:

$$f = \begin{cases} 2^{-n} & n \le M \\ 2^{-M}\dfrac{N - n + 1}{N - M + 1} & n > M \end{cases}.$$

Inverting this expression to find n as a function of f and plugging into $c(n)$ we find:

$$c(f) = \begin{cases} \mu\left(\frac{2}{f} - 1\right) & f \geq 2^{-M} \\ \mu(2^M(N - M + 3) - 4^M f(N - M + 1) - 1) & f < 2^{-M} \end{cases}.$$

This expression recovers the purely exponential result when $N = M$ and the purely linear case when $M = 0$.

Boundary-driven growth (BDG) simulations: As a final type of growth model, we consider a more complex scenario that requires numerical simulations to determine $c(f)$. We adapted the spatial model of bacterial range expansion in ref. 14, which in turn was derived from the Eden model[54]. The system consists of a d-dimensional lattice (square lattice in $d = 2$, cubic lattice in $d = 3$) with each grid point either empty or containing a single cell. A cell can only divide if there is at least one empty neighboring grid point which can be occupied by a daughter cell. At each time step one cell that is not totally surrounded is chosen at random to divide, replaced by two daughter cells (one at the current grid point, one at a random empty neighbor). As in the above models, at each cell division a novel mutation can occur with probability μ and gets inherited by both daughter cells (and all subsequent descendants). We keep track of the mutations carried by each cell in the population. The simulation is initiated by a single cell in the center of the lattice, and the cell colony expands in a roughly radially symmetric pattern, with the total lattice size at least four times the desired final population size N_{cell} (so no cells reach the edge). After reaching the population N_{cell}, we tally the mutations to obtain the cumulative distribution $c(f)$. Since there is stochastic variation between replicates of the same simulation, in Fig. S9 we show $c(f)$ averaged over 300 replicates in 2D, and 50 replicates in 3D.

Comparison to experimental data: The largest possible frequency in the theoretical models is $f = 1$, for a mutation that occurs in the first cell division. The corresponding value of the cumulative distribution $c(1) = \mu$, since this mutation will occur with probability μ. To align the models with the experimental data, we extrapolated the large frequency limit of the leaf distribution to find $\mu = 0.076$. Note that the largest frequency that appears in the experimental data is $f = 1/2$, since maize is diploid. To compare against the implicitly haploid growth models described above, we divided all frequencies in the theory plots in Fig. S9 by a factor of 2. We also chose a total population value of $N_{cell} = 10^6$ for all the models, as a typical experimentally plausible sampling size. Different value of N_{cell} would shift the plateau value of $c(f)$ at small frequencies, but the qualitative features of the theory predictions (and the discrepancies with the experimental curve) remain similar. Fixing μ and N_{cell} completely determines the theory curves, except in the case of the exponential + linear model, where there is still a free parameter M (the number of exponential generations). Here we did a best-fit to the leaf distribution (in log-log space), to find $M = 10$.

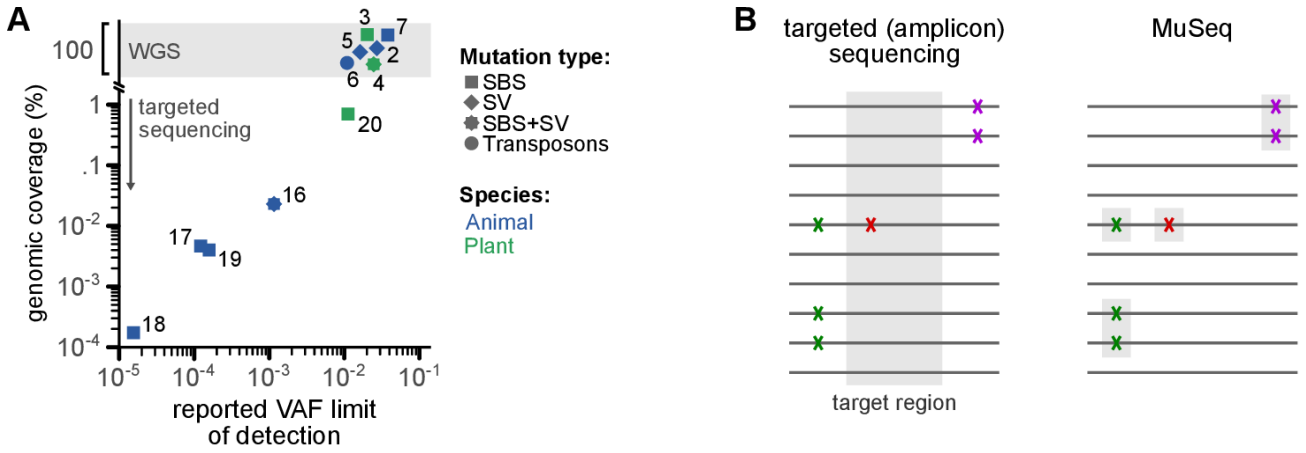

Figure S1. Sequencing depth limitations make it difficult to assess rare *de novo* mutations

(**A**) Relationship between VAF detection limit vs genomic coverage for selected studies. Numbers reflect the citation number in the main text references. Here, '100% genomic coverage' implies there was not intentional selection for a subset of the genome; in practice, this means the 'mappable genome' and excludes regions that are repetitive or otherwise difficult to amplify or sequence. VAF, variant allele frequency; SBS, single-base substitution; SV, structural variant.

(**B**) Comparison between targeted (amplicon) sequencing and MuSeq. While both approaches limit the sequencing to a portion of the genome, they do so in different ways. In these cartoons, 10 example DNA sequences are illustrated as dark gray lines; there are three mutations at different abundances, colored as green, red, and purple 'X's. For targeted sequencing, the 'X's could represent any class of mutation (SBS, SV, transposon); for Museq, these must be Mu transposon insertions. The region targeted by each technique is highlighted in gray.

Targeted sequencing selects a predefined set of genome loci to sequence deeply. Both wild type and mutant alleles are sequenced and any mutations outside of the target region are missed. MuSeq, in contrast, sequences transposon insertion sites throughout the genome, and reduces sequencing depth by avoiding the wild-type (transposon-free) alleles. In this hypothetical example, targeted sequencing would require at least 10 reads but only capture a single mutation; MuSeq would require fewer reads yet would capture all 3 mutations. While MuSeq is limited to transposons (Mu in this case), the opportunity is that it enables orders of magnitude greater sensitivity and dynamic range than is possible for other classes of mutation.

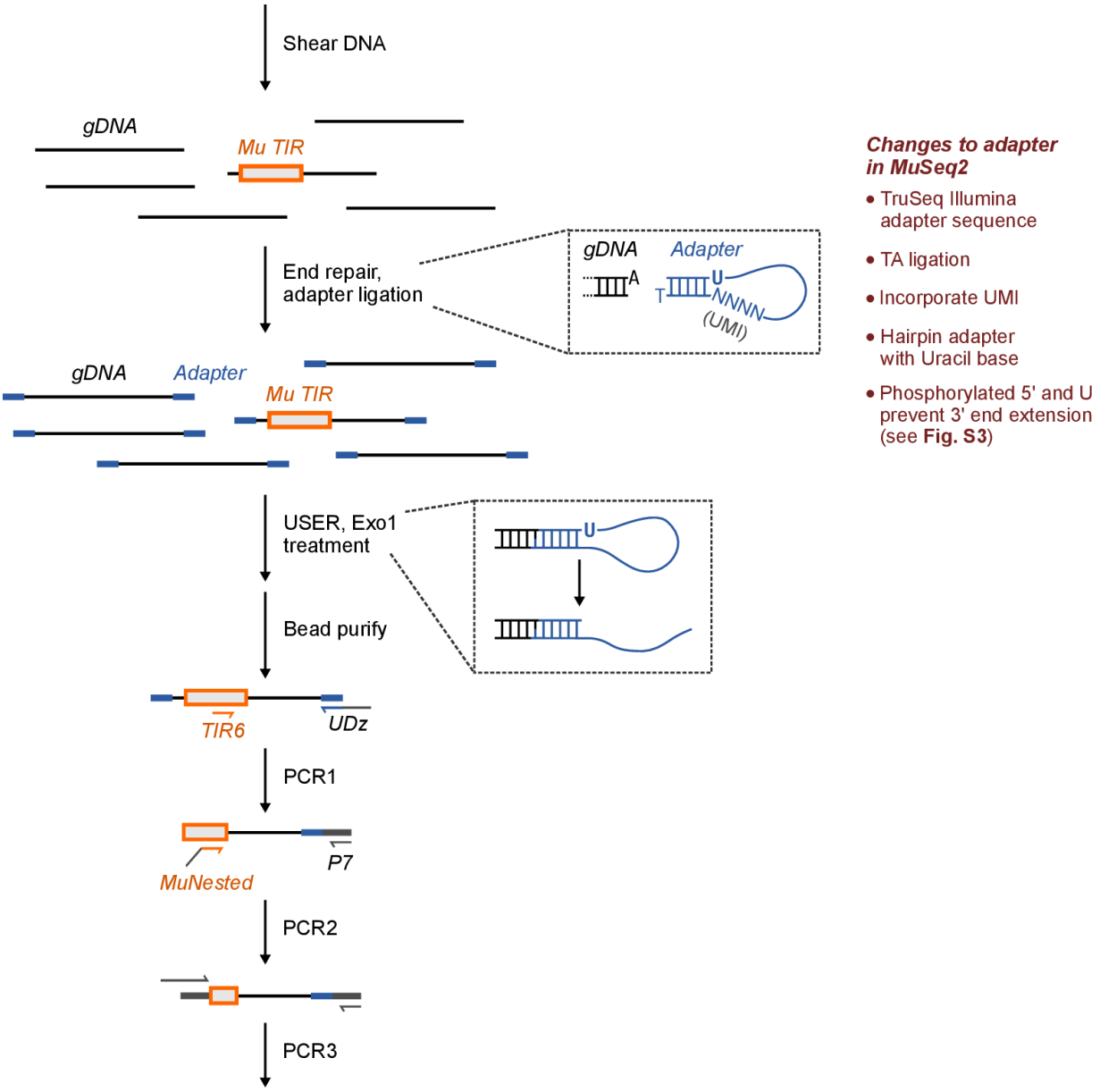

Figure S2. Overview of the MuSeq2 protocol

MuSeq2, similar to MuSeq, uses an adapter ligation followed by a series of nested PCR reactions to specifically amplify fragments spanning the transposon genome junction. Several changes to the adapter were made in MuSeq2, including incorporating a Unique Molecular Identifier (UMI) for molecular counting and updating to modern Illumina adapter sequences. USER treatment cleaves the adapter at the Uracil, making the ends compatible with the downstream PCR reactions.

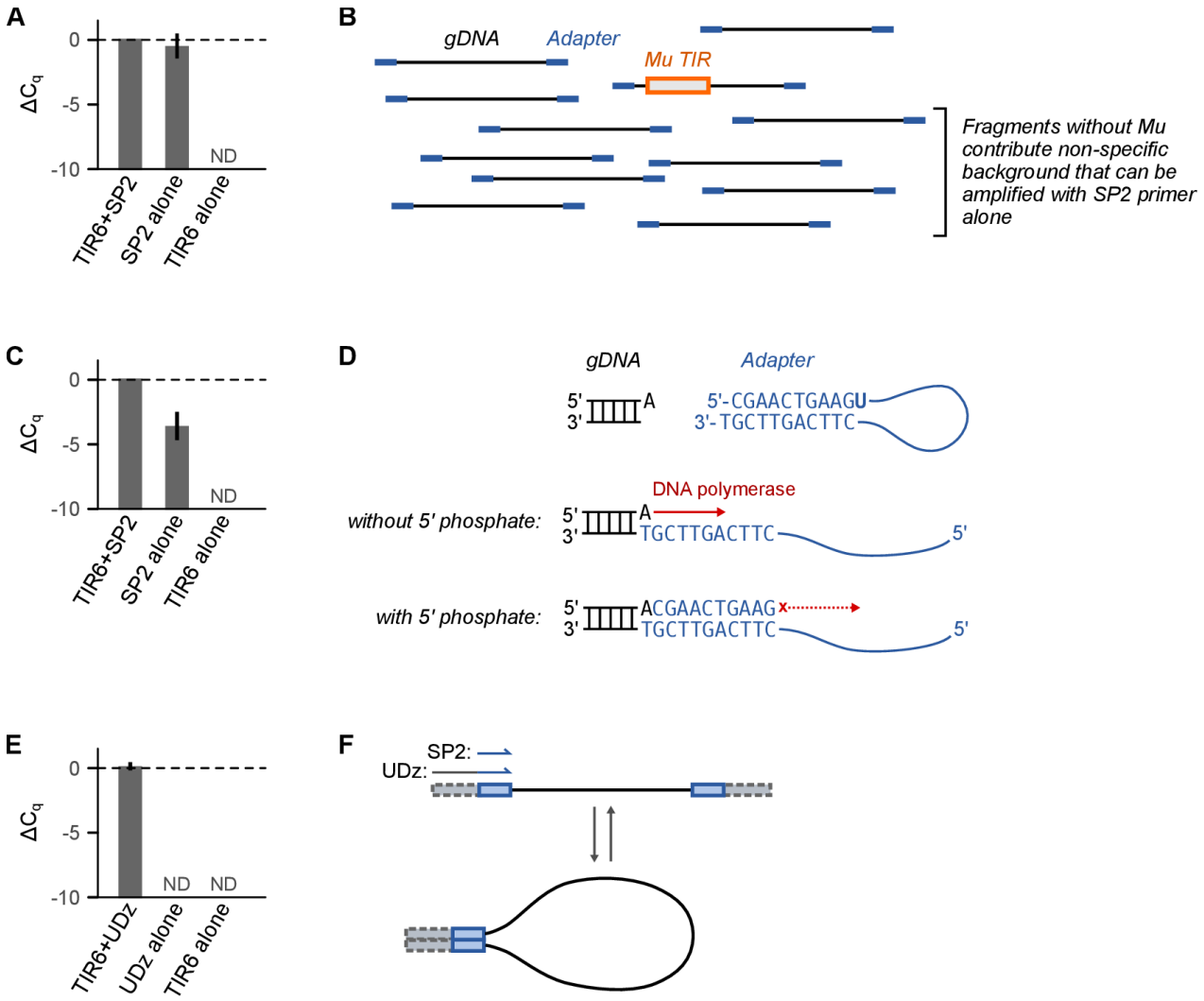

Figure S3. Reducing non-specific amplification through changes to adapter structure and suppression PCR

(**A**) Quantitative PCR for libraries prepared using an adapter design similar to the original MuSeq: the adapter was unphosphorylated and the reverse primer (SP2) matches the adapter-ligated sequence only. TIR6 is a Mu-specific primer; the SP2 primer matches the Illumina sequence ligated onto sheared genomic DNA. Three independent DNA samples were sheared and ligated, then the ligated DNA was split for qPCR with different primer combinations. The amplification is not specific, as SP2 alone amplifies similarly to when the Mu-specific primer was included. Cq was normalized to TIR6+SP2; in this experiment, the average number of cycles at ΔCq = 0 was 16.4.

(**B**) The reason for non-specific amplification with the SP2 primer is that the majority of DNA fragments from sheared genomic DNA do not contain a Mu element. Fragments without Mu are estimated to outnumber Mu-containing fragments by more than 10,000 fold. The background fragments will have adapter DNA on both sides, forming a potential priming site for PCR with the SP2 primer. The adapter structure limits background amplification in part because it has a 5' overhang and does not initially have the sequence needed for primer binding (the primer binds to the reverse complement of the overhang; this was also true in the original MuSeq); however, the 5' overhang can be copied by DNA polymerase at the start of PCR. This is likely an inefficient process, but given the excess of fragments without Mu it still contributed meaningful background (panel A).

(**C**,**D**) One modification to reduce non-specific background in MuSeq2 was to replace the unphosphorylated oligo with a phosphorylated one. With a 5' phosphate on the adapter, both adapter strands can be ligated to the sheared genomic DNA. After a Uracil on the adapter is cleaved to release the hairpin, it leaves a 3' phosphate overhang. By ligating the adapter on both strands, there is no free 3' hydroxyl available – blocking extension by DNA polymerase. Panel **C** shows the same experiment as panel **A**, except using a phosphorylated adapter. Cq was normalized to TIR6+SP2; in this experiment, the average number of cycles at ΔCq = 0 was 16.2.

(**E**,**F**) A second modification in MuSeq2 was to use a longer primer, UDz, in place of SP2 during the first PCR. Fragments with adapter sequence on both sides do not amplify as efficiently because they have self-complementary ends and can form a hairpin (suppression PCR). The UDz primer adds the entire Illumina adapter sequence during PCR. Because this primer makes the self-complementary region longer, it favors hairpin formation and increases the amount of suppression PCR for non-specific fragments. Panel **E** shows qPCR using the same adapter-ligated samples as in panel **C**, except that PCR was performed with the UDz primer. Cq was normalized to TIR6+SP2 from panel **C** and so the ΔCq values are directly comparable between these panels. There was no decrease in specific amplification when switching to the UDz primer, but non-specific amplification was completely suppressed.

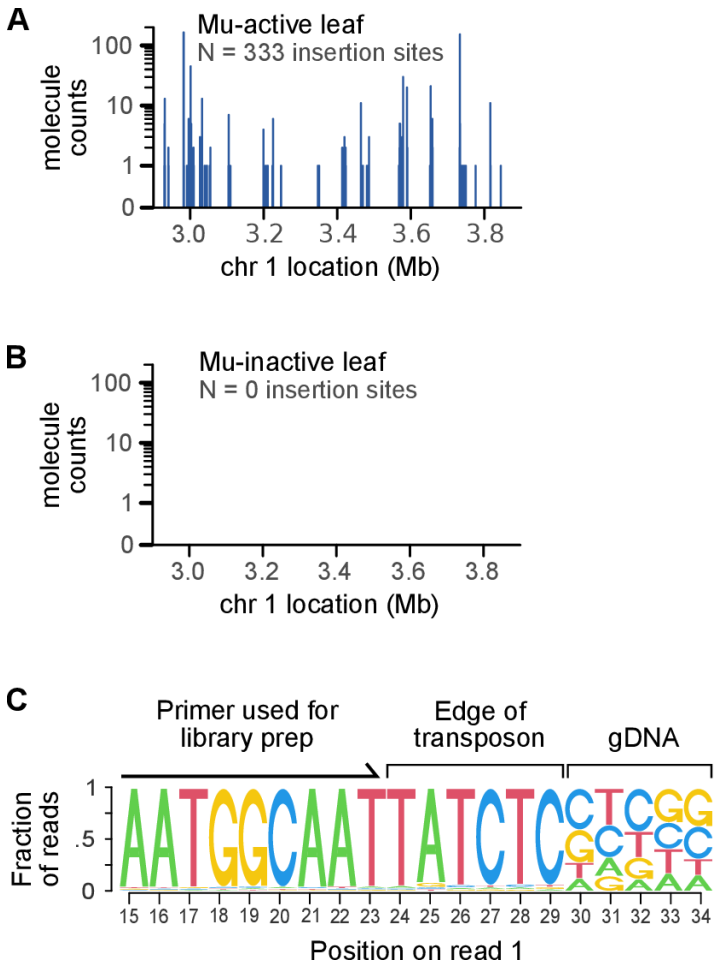

Figure S4. Specificity of MuSeq2

(**A**) A representative 1 Mb region showing all insertion sites identified in a single Mu-active leaf sample. In total, 333 insertion sites were found in this region, covering a range of molecule abundances (UMI counts).

(**B**) The same region as in **A**, but for a Mu-inactive leaf sample. No insertion sites were observed in this region.

(**C**) Sequence composition for a portion of read 1, which covers the transposon-genome junction. The last 6 bp of the transposon were not included in any primer used during library prep, and provides independent validation that the sequencing is specific to Mutator. The 'validation sequence' matches the known transposon sequence TATCTC for the vast majority of reads.

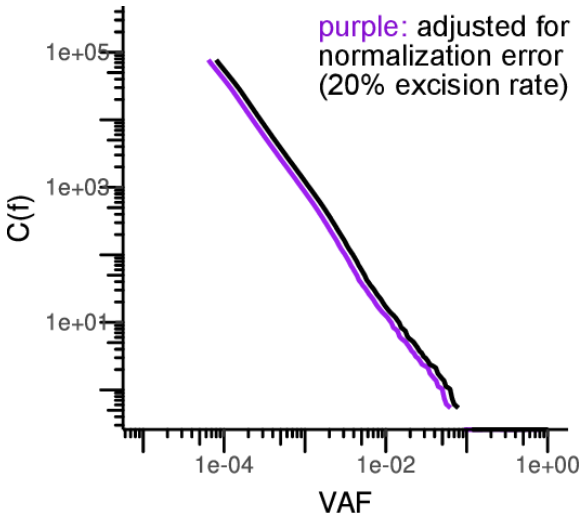

Figure S5. Mu excision rates have a negligible impact on normalization

Allele frequencies were normalized using the paternal insertions, which were assumed to be at their original abundance (VAF = ½ in most tissues, 1/3 in endosperm). In the presence of excisions, the true allele frequency of paternal insertions would be less, resulting in systematic error during normalization. We estimate the endosperm excision rate in our line is ~10%, based on the proportion of endosperm surface that has reverted to purple (this line carries a mutable bz1-Mum9 reporter allele that allows for purple pigment expression after excision). To be conservative, we used double this rate – 20% – and calculated the effect this would have on the measured allele frequencies. This figure shows the allele frequency distribution for a representative endosperm sample before (black line) and after (purple line) adjusting for normalization error due to a 20% excision rate. Even with double the observed excision rate, normalization error has a minimal impact on the allele frequency spectrum. This is because a change on the order of 20% is small when the measured frequencies vary by many orders of magnitude (log-scale). As a result, we did not consider excision further in our analyses; all main text results were not adjusted for excision (e.g. the black line above).

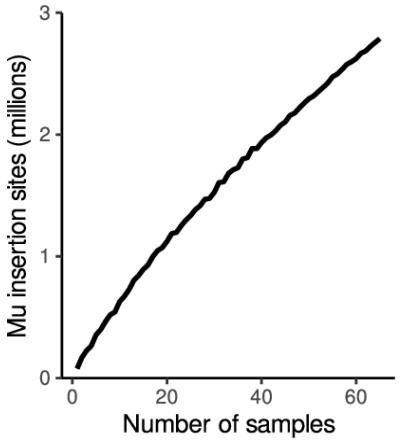

Figure S6. The number of Mu insertion sites has not reached saturation

Random subsets of samples were drawn and then the total number of genomic insertion sites was calculated. The total number of insertion sites has not reached saturation under the conditions of this study.

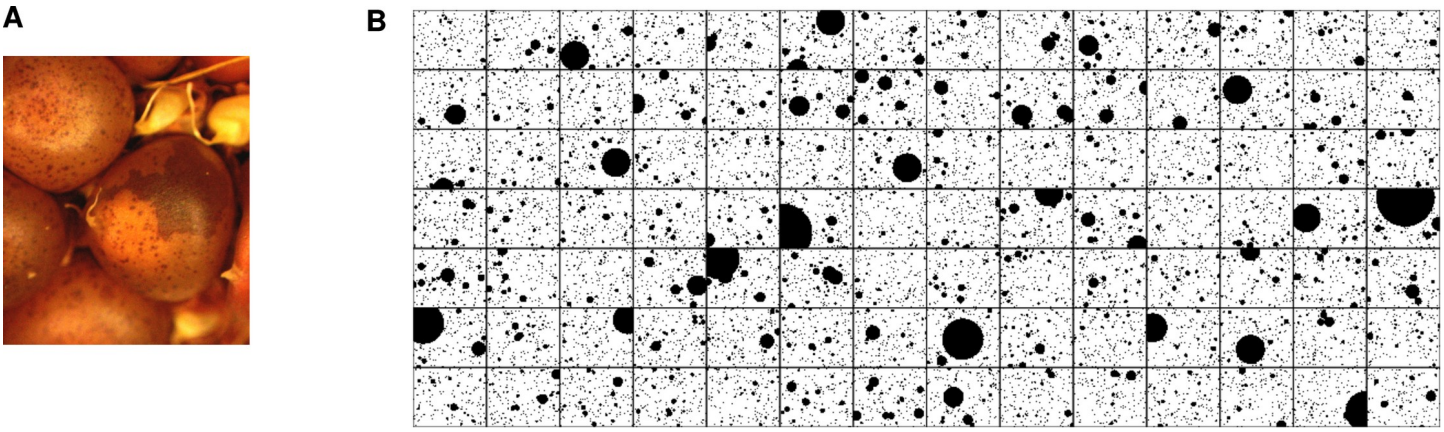

Figure S7. Mu insertions and excisions behave differently in endosperm

(**A**) Example sector sizes in Mu-active kernels with the bz1-Mum9 reporter. The kernel on the top left with many small spots is representative. The very large sector on the kernel in the middle is exceedingly rare; we observed 7 out of 1844 kernels with sectors making up at least 5% of the kernel area. Prior quantitative data on excision spot size found even fewer large sectors, with 0 sectors at a frequency under 2^8 (VAF ~ 10^{-3}) out of 2000 kernels (Levy and Walbot, 1990).

(**B**) A simulated ear with sectors drawn to represent Mu insertions. Spot size was defined by randomly drawing *de novo* endosperm insertions based on the measured allele frequency spectrum. This is intended as a simple visual representation to highlight the qualitative difference between excisions and insertions. The frequency of large spots in this diagram is dramatically higher than what is seen for endosperm excision sectors. In this simulation, we assumed all divisions happen within a 2D plane, which may be approximately true for aleurone (the outer cell layer of endosperm where the visible pigment is produced).

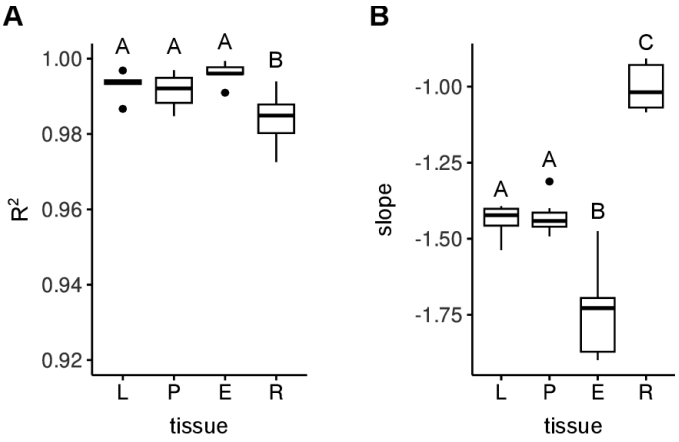

Figure S8. Power-law fits to the allele frequency spectra from various tissues

Best linear fit parameters to log-transformed data. Fitting was performed on individual samples and the results plotted as a boxplot separated by tissue. Letters indicate statistical significance: groups not sharing a letter have a significantly different mean ($p \leq 0.05$; Tukey's honest significant difference test). N = 6 samples for leaf and root; N = 9 samples for endosperm and pollen.

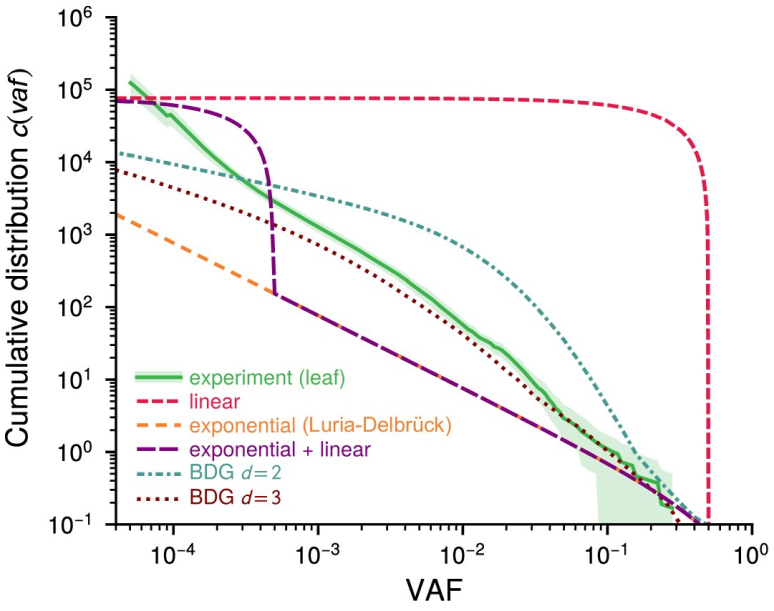

Figure S9. Fit of experimental leaf data to various models of mutation accumulation

Details of the theoretical models are described in the SI Text. All models assume no cell death, a constant mutation probability $\mu = 0.076$ over time (chosen to agree with the experimental curve at the largest frequency), and a final cell population size of $N_{cell} = 10^6$. The experimental data for leaf is shown in dark green, with the 95% confidence interval shaded in light green. Linear = linear growth, where after each cell division only one daughter cell is capable of further cell division. Exponential = exponential growth, where both daughters are capable of division (Luria-Delbrück model). Exponential + linear = 10 generations of exponential growth, with the remaining generations linear. BDG = boundary-driven growth simulations based on the Eden model, which were carried out on both 2D and 3D square lattices.

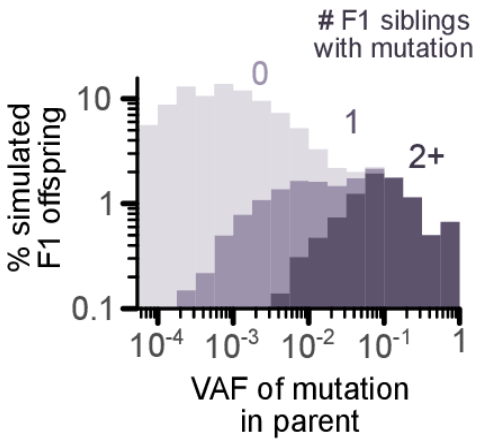

Figure S10. During simulations of Robertson (1980), F1 offspring that share the same mutation as their sibling are most often derived from pollen insertions at high allele frequencies.

For the simulations of Robertson (1980) in **Fig. 5B**, we recorded the VAF for any Mu insertion transmitted to the simulated F1 offspring. In this histogram, the bars are color coded based on whether the pollen mutation was inherited by 0, 1, or 2+ F1 siblings. Simulated F1 offspring were derived from Mu insertions at a wide range of allele frequencies, suggesting these occurred throughout development. For mutations present in larger clusters of F1 offspring (2+ siblings with the same mutation), the average VAF in the parent was 0.13; this corresponds 1 insertion in every 3.8 diploid pollen progenitors, roughly the number of meristematic cells in the seed that ultimately form the maize tassel (the male flower; ref: Poethig, Coe, and Johri, 1986). VAF, variant allele frequency.