Uncovering hidden variation in polyploid wheat

Ksenia V. Krasileva[a,b,c], Hans A. Vasquez-Gross[a], Tyson Howell[a], Paul Bailey[c], Francine Paraiso[a], Leah Clissold[c], James Simmonds[d], Ricardo H. Ramirez-Gonzalez[c,d], Xiaodong Wang[a], Philippa Borrill[d], Christine Fosker[c], Sarah Ayling[c], Andrew L. Phillips[e], Cristobal Uauy[d,1,2], and Jorge Dubcovsky[a,f,1,2]

[a]Department of Plant Sciences, University of California, Davis, CA 95616; [b]The Sainsbury Laboratory, Norwich NR4 7UH, United Kingdom; [c]The Earlham Institute, Norwich NR4 7UG, United Kingdom; [d]John Innes Centre, Norwich NR4 7UH, United Kingdom; [e]Rothamsted Research, Harpenden AL5 2JQ, United Kingdom; and [f]Howard Hughes Medical Institute, Chevy Chase, MD 20815

Contributed by Jorge Dubcovsky, December 20, 2016 (sent for review November 22, 2016; reviewed by Beat Keller and Joachim Messing)

Comprehensive reverse genetic resources, which have been key to understanding gene function in diploid model organisms, are missing in many polyploid crops. Young polyploid species such as wheat, which was domesticated less than 10,000 y ago, have high levels of sequence identity among subgenomes that mask the effects of recessive alleles. Such redundancy reduces the probability of selection of favorable mutations during natural or human selection, but also allows wheat to tolerate high densities of induced mutations. Here we exploited this property to sequence and catalog more than 10 million mutations in the protein-coding regions of 2,735 mutant lines of tetraploid and hexaploid wheat. We detected, on average, 2,705 and 5,351 mutations per tetraploid and hexaploid line, respectively, which resulted in 35–40 mutations per kb in each population. With these mutation densities, we identified an average of 23–24 missense and truncation alleles per gene, with at least one truncation or deleterious missense mutation in more than 90% of the captured wheat genes per population. This public collection of mutant seed stocks and sequence data enables rapid identification of mutations in the different copies of the wheat genes, which can be combined to uncover previously hidden variation. Polyploidy is a central phenomenon in plant evolution, and many crop species have undergone recent genome duplication events. Therefore, the general strategy and methods developed herein can benefit other polyploid crops.

wheat | polyploidy | mutations | reverse genetics | exome capture

Since the dawn of agriculture, wheat has been a major dietary source of calories and protein for humans. The cultivated wheat species *Triticum turgidum* (tetraploid, AABB genome) and *Triticum aestivum* (hexaploid, AABBDD genome) originated via recent polyploidization events followed by domestication. *T. turgidum* originated less than 500,000 y ago from the hybridization of *Triticum urartu* (diploid, AA genome) and a now-extinct species related to *Aegilops speltoides* (diploid, SS similar to BB genome), whereas *T. aestivum* originated less than 10,000 y ago from the hybridization of tetraploid wheat with *Aegilops tauschii* (diploid, DD genome) (1).

As a result of the recent polyploidization, most genes in tetraploid and hexaploid wheat species are present in multiple functional copies, referred to as homeologs. These duplicated genes buffer the rapid natural changes occurring in the large and dynamic wheat genomes (1). As loss-of-function mutations in any single wheat homeolog are frequently masked by redundancy in other homeologs, this variation remains hidden from natural and human selection. This drawback becomes an advantage for the development of mutant populations, as redundancy confers tolerance to high densities of induced mutations (2). On average, mutation densities of ethyl methanesulfonate (EMS) mutant populations of hexaploid wheat (3–5) are as much as 10-fold higher than those of diploid barley (6). When mutations in individual homeologs have been identified, they can be combined to generate loss-of-function mutants and to overcome the masking effect of redundant homeologs.

Extensive utilization of the current wheat mutant populations has been limited by the need to physically access the DNAs of the mutant lines and by the time required for the mutant screens, which entail the development and optimization of genome-specific primers for each target gene. A pilot study using three Cadenza lines with known mutations in the *GA20ox* gene and a small capture array including 1,846 genes demonstrated that exome capture (7) followed by sequencing was a viable strategy to identify mutations in wheat (8). Whole-genome resequencing of mutant lines also has been used for species with small genomes (9), but is a very expensive alternative for the large genomes of tetraploid (12 Gb) and hexaploid (17 Gb) wheat (10).

In this study, we describe the development of a wheat exome capture platform and its use to sequence the coding regions of 2,735 mutant lines. We characterized the obtained mutations, organized them in a public database including more than 10 million mutations, identified deleterious alleles for ∼90% of the captured wheat genes, and discuss potential applications.

Results

Development of a Wheat Exome Capture Design. In collaboration with NimbleGen, we developed an 84-Mb exome capture assay including overlapping probes covering 82,511 transcripts (*SI Appendix*, Method S1 and Table S1). We aligned these transcripts to

Significance

Pasta and bread wheat are polyploid species that carry multiple copies of each gene. Therefore, loss-of-function mutations in one gene copy are frequently masked by functional copies on other genomes. We sequenced the protein coding regions of 2,735 mutant lines and developed a public database including more than 10 million mutations. Researchers and breeders can search this database online, identify mutations in the different copies of their target gene, and request seeds to study gene function or improve wheat varieties. Mutations are being used to improve the nutritional value of wheat, increase the size of the wheat grains, and generate additional variability in flowering genes to improve wheat adaptation to new and changing environments.

Author contributions: C.U. and J.D. designed research; K.V.K., T.H., L.C., J.S., X.W., P. Borrill, and C.F. performed research; K.V.K., H.A.V.-G., T.H., P. Bailey, F.P., R.H.R.-G., S.A., A.L.P., C.U., and J.D. contributed new reagents/analytic tools; K.V.K., H.A.V.-G., T.H., P. Bailey, R.H.R.-G., C.U., and J.D. analyzed data; K.V.K., H.A.V.-G., T.H., C.U., and J.D. wrote the paper; A.L.P. contributed to the original idea and provided the Cadenza TILLING population; C.U. proposed the original idea, and codirected the project; and J.D. proposed the original idea, codirected the project, and coordinated international collaboration.

Reviewers: B.K., University of Zürich; and J.M., Rutgers University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in NCBI BioProject (accession no. PRJNA258539) and European Read Archive (ENA study PRJEB11524). The variant calls are available at Plant Ensembl, plants.ensembl.org/Triticum_aestivum/Info/Index.

[1]C.U. and J.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: cristobal.uauy@jic.ac.uk or jdubcovsky@ucdavis.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619268114/-/DCSupplemental.

the draft wheat genome (11) and identified 286,799 exons, which we padded with 30 bp of intronic sequence, when possible, to maintain coverage at splice sites. Given that the assay probes cross-hybridize efficiently at high levels of sequence identity (12), we did not include the most similar homeologs in the capture design. The coding regions of wheat homeologs average 97.2% identity [SD 1.8% (13)], which is similar to the genome divergence found in peanuts (14) and slightly greater than the divergence found in soybeans (15).

We multiplexed captures from eight tetraploid or four hexaploid wheat lines per Illumina lane and obtained >20 million 100-bp paired-end reads per sample by using methods described in *SI Appendix*, Method S2 and Table S2. To improve the proportion of mapped reads, we supplemented the reference wheat genome sequence (11) with de novo assemblies of unmapped reads from both wheat species (*SI Appendix*, Method S3 and Table S3). The de novo assemblies (*SI Appendix*, Table S3) resulted in an additional 40,975 contigs for Kronos (33.4 Mb) and 67,632 contigs for Cadenza (41.3 Mb). The expanded reference improved the proportion of mapped reads from 93% to 98% in Kronos and from 96% to 99% in Cadenza (*SI Appendix*, Table S3). The initial capture design (alpha) was tested in 42 "Kronos" and 49 "Cadenza" mutants, and probes with extreme capture efficiencies were rebalanced at NimbleGen, resulting in an adjusted design (beta) with more uniform coverage (*SI Appendix*, Fig. S1).

Identification and Characterization of Induced Mutations. We used the adjusted design to capture and sequence the coding regions of 1,535 EMS mutants from the tetraploid variety "Kronos" (5) and 1,200 EMS mutants from the hexaploid variety "Cadenza" (3). The development of these populations and the sequencing strategy are summarized in Fig. 1. Mutations were then identified using the Mutation and Polymorphism Survey (MAPS) bioinformatics pipeline (16), which was modified by additional filters described in *SI Appendix*, Figs. S2–S5. We modified several MAPS parameters to optimize the detection of mutations in polyploid wheat (*SI Appendix*, Method S4). Using these parameters, the MAPS pipeline identified 119.2 Mb and 162.4 Mb positions with

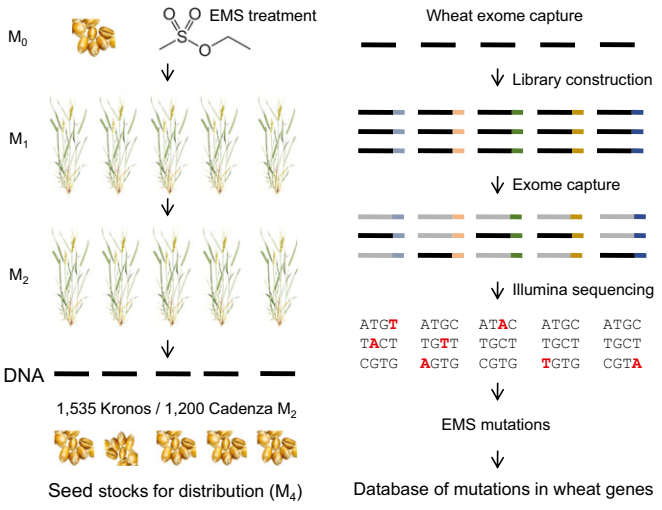

Fig. 1. Overview of the development of the sequenced mutant populations. M_0 seeds were mutagenized with EMS (resulting in M_1 plants), and a single M_2 plant was grown from each M_1 plant. Genomic DNAs were extracted from the M_2 plants, and M_3 seeds were obtained from the same plant. M_3 seeds were planted in the field to produce M_4 seed for distribution. Barcoded sequencing libraries were constructed, used for exome capture, and sequenced by using Illumina. Mutations were identified by using the MAPS pipeline and were deposited in public databases that can be searched online.

adequate coverage and quality for mutation identification in tetraploid and hexaploid wheat, respectively. These positions, designated hereafter as "valid positions," were larger than the 84 Mb covered by the original design. This was an expected result, as the captured sequences included homeologs that were not in the original design. The 1.4:1 ratio between valid positions in hexaploid and tetraploid wheat is close to the 1.5:1 ratio expected between lines containing three and two genomes.

Real mutations are expected to be present in multiple reads, whereas sequencing errors tend to be independently distributed. Therefore, the minimum number of reads including a mutation that are required to call a mutation [minimum coverage (*MC*)] is a critical parameter to differentiate real mutations from sequencing errors. To select an *MC* threshold that maximizes the number of detected mutations while keeping a low error rate, we compared the coverage, the number of mutations, and the associated errors at different *MC* stringency levels (*SI Appendix*, Method S4 and Tables S4–S6). Based on these data, we selected a minimum coverage of five mutant reads for heterozygous (*HetMC5*) and three for homozygous (*HomMC3*) mutations as the optimal thresholds for mutant identification. The median coverage at mutation sites for Kronos and Cadenza was 21× (*SI Appendix*, Table S4).

By using the *HetMC5*/*HomMC3* threshold, we identified 4.15 million uniquely mapped EMS-type mutations (G to A and C to T) in tetraploid wheat (2,705 mutations per line) and 6.42 million in hexaploid wheat (5,351 mutations per line; Table 1). Dividing these numbers by the number of valid positions identified by MAPS, we estimated an average mutation density of 23 mutations per Mb per individual in Kronos (34.8 mutations per kb for 1,535 lines) and 33 mutations per Mb per individual in Cadenza (39.5 mutations per kb for 1,200 lines). The distribution of mutations along the wheat pseudomolecules paralleled the distribution of gene densities (Fig. 2*A*), suggesting a uniform distribution of mutations along the wheat coding regions.

At the *HetMC5*/*HomMC3* threshold, the nonmutagenized control lines showed a much lower number of polymorphisms than the mutant lines (0.5% in Kronos and 1.2% in Cadenza). On average, only 14 SNPs per plant were detected in the nonmutagenized Kronos, along with 63 in the nonmutagenized Cadenza controls (Fig. 2*B*), suggesting a low error rate. This was also supported by the low percentage of non–EMS-type mutations (all mutation types except G to A and C to T) observed in the mutagenized lines of Kronos (0.9%) and Cadenza (0.8%; Fig. 2*C* and Table 1). A third estimate of the error rate was obtained by calculating the number of non–EMS-type transitions (A to G and T to C) within the non-EMS SNPs. This number is very similar to the reciprocal EMS-type mutations (G to A and C to T; *SI Appendix*, Method S5), and can be used to estimate the number of potentially erroneous EMS-type mutations. By using this method, the predicted error rate among uniquely mapped EMS-type mutations was less than 0.2% in both Kronos and Cadenza (Table 1). The low error rate of our mutant identification pipeline was also validated experimentally by resequencing PCR products and dedicated SNP assays. Among 280 EMS-type mutations, 278 (99.3%) were confirmed across both populations (*Materials and Methods* and *SI Appendix*, Text S1 and Tables S7–S10).

At lower *MC* stringency levels, we observed higher numbers of mutations, but also an increase in the associated errors (Fig. 2*D*). At *HetMC3*/*HomMC2*, for example, we detected an additional 0.9 million EMS-type mutations in tetraploid wheat (*SI Appendix*, Table S5) and 1.7 million in hexaploid wheat (*SI Appendix*, Table S6) that were not previously detected at *HetMC5*/*HomMC3*. However, the estimated errors for heterozygous mutations at exact coverage of 3 (*HetC3*) increased to 5.6% for Kronos (*SI Appendix*, Table S5) and 10.0% for Cadenza (*SI Appendix*, Table S6). These additional mutations still have an acceptable probability of being correct and can be accessed by selecting the desired *MC* in the public database

Table 1. Characterization of mutations in tetraploid and hexaploid wheat (*HetMC5/HomMC3*)

Mutations and SNPs characteristics	Tetraploid Kronos	Hexaploid Cadenza
Uniquely mapped SNPs*	4,189,561	6,470,733
Heterozygous/homozygous ratio at M_2*	1.87	2.21
Uniquely mapped EMS-type mutations*	4,152,707	6,421,522
Average EMS-type mutations/line*	2,705	5,351
Average EMS-type mutations per kilobase (population)	34.8	39.5
EMS-type, %*	99.1	99.2
Non–EMS-type transitions*	7,323	10,569
Maximum error in uniquely mapped EMS-type, %*,†	0.18	0.16
RH SNPs	69,651	38,626
Heterozygous/homozygous ratio in RH	0.33	0.30
Average SNPs per megabase per individual in RH	592	441
EMS-type SNPs in RH	16,412	6,023
EMS-type in RH, %	23.6	15.6
Non–EMS-type transitions in RH	20,358	8,669
Multimap SNPs	321,511	955,074
Heterozygous/homozygous ratio in multimap	2.85	6.16
Multimap EMS-type mutations	315,537	933,515
EMS-type mutations in multimap SNPs, %	98.14	97.74
Non–EMS-type transitions in multimap	1,166	5,968
Maximum error in multimapped EMS-type, %	0.37	0.64
Gene models with at least one mutation (GM_1)‡	48,172	73,895
GM_1 with at least one truncation	28,604 (59%)	45,311 (61%)
GM_1 with at least one missense mutation	46,198 (96%)	69,543 (94%)
Average number of missense mutations per GM_1	21.4	22.6
GM_1 with truncation and/or deleterious missense§	43,787 (91%)	67,830 (92%)
No. of unique genes eliminated in large deletions	832	6,657

*Excluding RH and deletion regions.
†Estimated from the number of reciprocal A>G and T>C transitions among non–EMS-type mutations.
‡GM in Ensembl (a more detailed analysis of variant effect predictions is provided in *SI Appendix*, Text S3).
§Predicted deleterious missense mutations by SIFT (<0.05).

search tools from the project (dubcovskylab.ucdavis.edu/wheat_blast and www.wheat-tilling.com).

Residual Genetic Heterogeneity. The original breeder's seed stocks of Kronos and Cadenza that were mutagenized had small regions of residual genetic heterogeneity (RH) that were identified after sequencing based on their lower proportion of EMS-type mutations, higher mutation density, higher proportion of homozygous mutations, and presence in multiple individuals (Table 1, Fig. 3 *A* and *B*, and *SI Appendix*, Fig. S6). Using an index that combined those criteria (*SI Appendix*, Method S5 and Table S11), we identified 69,651 RH-SNPs in Kronos (1.7% of the total SNPs) and 38,626 RH-SNPs (0.6%) in Cadenza at the *HetMC5*/*HomMC3* threshold (Table 1 and *SI Appendix*, Table S12). These RH levels are consistent with seed obtained after pooling multiple F_7 plants for Kronos and multiple F_8 plants for Cadenza, which is normal breeding practice.

Mutations Present in Multiple Individuals. Even after removal of the RH regions, approximately 1.4 million EMS-type mutations shared by more than one individual were detected in Kronos and Cadenza. The frequency of these mutations decayed rapidly from two to six individuals (Fig. 3 *C* and *D*, red bars) and was very different from the frequency distribution in the RH regions (Fig. 3 *A* and *B*). The distribution of these EMS-type mutations approached a Poisson distribution (Fig. 3 *C* and *D*, light blue bars). However, the closest theoretical Poisson distribution was obtained by using only one fifth of the available G/C sites (*SI Appendix*, Method S6 and Tables S13 and S14). This result suggests that some G/C positions have a lower probability of being affected by the EMS mutagen.

We hypothesize that the EMS preference for certain sequences flanking the mutated sites (Fig. 3 *E* and *F* and *SI Appendix*, Method S7) can affect the probability of mutations in some G/C positions, as previously observed in rice (16). This hypothesis is also supported by the observation that sequence preferences in the region flanking EMS-type mutations were stronger in non-RH mutations shared by multiple individuals than in those present in only one individual (*SI Appendix*, Fig. S7). We do not rule out the possibility that differences in chromatin structure and DNA methylation may have also affected the probability of mutations at some G/C sites.

Reads Mapping to Multiple Locations. For some genes, we detected very few or no mutations. Characterization of these genes revealed that this was mainly caused by duplicated regions in the reference genome (e.g., highly similar homeologs or incorrect duplicated assemblies). Reads associated with multiple mapping (MM) locations were assigned low mapping quality values and were eliminated in the MAPS pipeline. To recover mutations in these locations, we developed a custom bioinformatics pipeline that assigned the MM reads to a single location, recorded alternative locations, modified the mapping quality score, and redirected the reads to MAPS (*SI Appendix*, Method S8 and Figs. S2 and S3). By using this pipeline, we recovered 16.6 million reads and identified an additional 1.25 million EMS-type mutations (Table 1). More MM high-quality mutations were observed in hexaploid (933,515) than in tetraploid wheat (315,537), which was expected based on the presence of an additional genome. We validated 22 of the 25 MM mutations tested by PCR and resequencing (*SI Appendix*, Method S8 and Table S15).

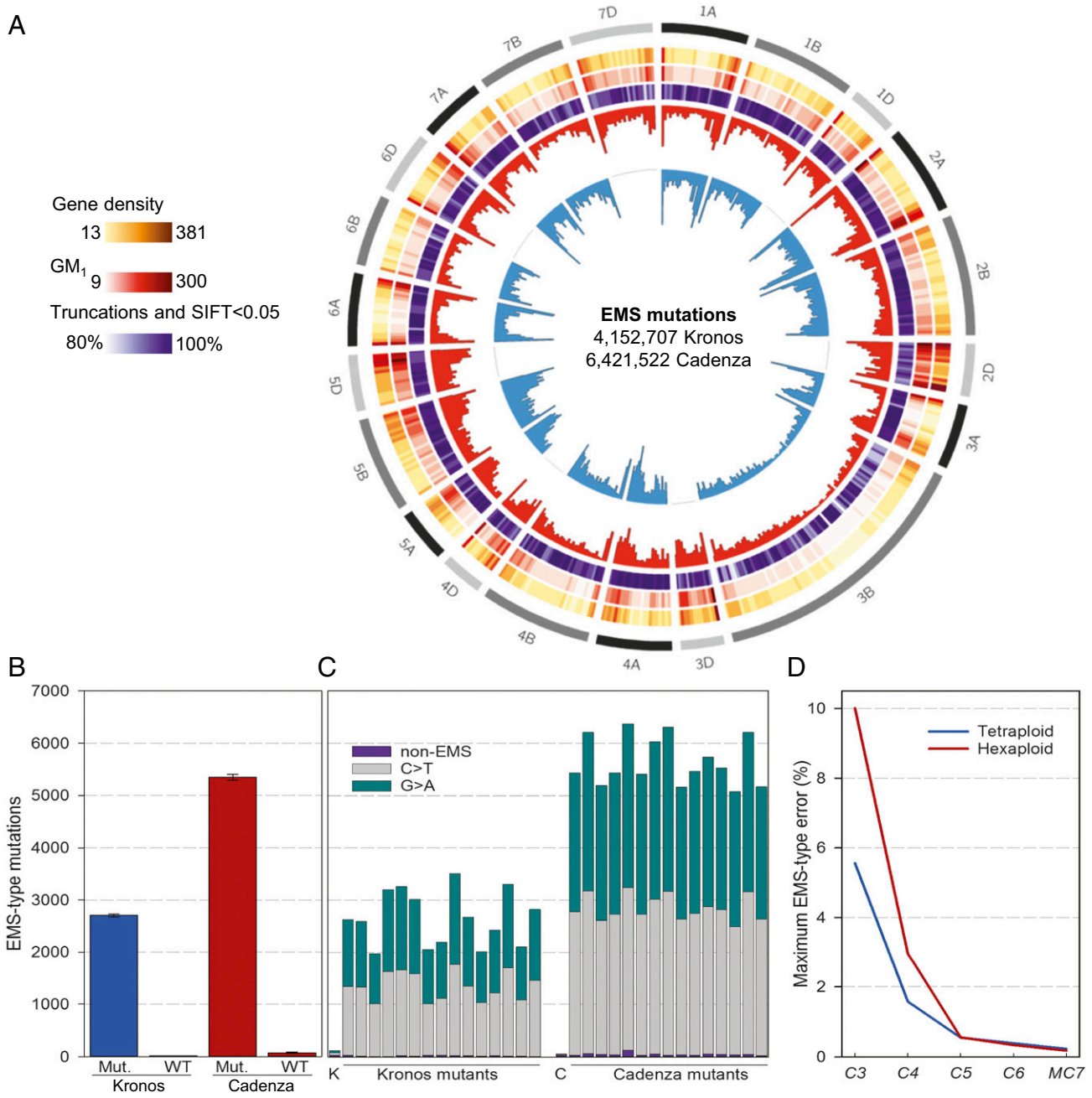

Fig. 2. Characterization of mutations in tetraploid Kronos and hexaploid Cadenza. (*A*) Genome-wide positions of identified mutations and their effects. The tracks from outside to inside represent gene density along wheat chromosomes (yellow-brown), number of gene models with at least one mutation (i.e., GM_1; red), percentage of GM_1 genes with at least one deleterious allele (truncation and/or missense mutation with SIFT score <0.05; purple), and total mutation densities for Cadenza (red) and Kronos (light blue). Each bin corresponds to a 10-Mb window. (*B*) Average number of SNPs in mutants and WT controls. (*C*) Distribution of mutations types in fifteen representative Kronos and Cadenza mutants ("K," nonmutagenized Kronos; "C," nonmutagenized Cadenza). Gray/teal indicates EMS-type mutations (C>T/G>A); violet indicates non–EMS-type mutations. (*D*) Estimated EMS-type error at different *HetMC* cutoffs (*C*, coverage; *MC*, minimum coverage).

Small and Large Deletions. In addition to EMS-type mutations, MAPS identified 616 and 1,268 small (<20 bp) deletions in Kronos and Cadenza, respectively (at *HetMC5*/*HomMC3*, excluding RH regions; *SI Appendix*, Table S16). We selected 15 of these predicted deletions for validation by Sanger resequencing (*SI Appendix*, Table S17 and Fig. S8) and were able to validate 14 of them (93.3%; *SI Appendix*, Table S17).

By using a dedicated bioinformatics pipeline (*SI Appendix*, Method S9 and Fig. S4), we detected a total of 870 large homozygous deletions covering five or more exons in Kronos and 7,971 in Cadenza (*SI Appendix*, Tables S18 and Figs. S9–S11). Overall, the large deletion events were mainly independent (94% of deletions were restricted to one or two individuals; *SI Appendix*, Text S2 and Fig. S11), and were confined to relatively small physical intervals comprising a median size of four to eight scaffolds in Kronos and Cadenza, respectively (*SI Appendix*, Fig. S10). The tetraploid Kronos population had fewer numbers of lines with homozygous deletions compared with the hexaploid Cadenza population, and those deletions were, on average, smaller and confined to fewer individuals. This is most likely a consequence of the higher EMS dosage used in the mutagenesis of the hexaploid

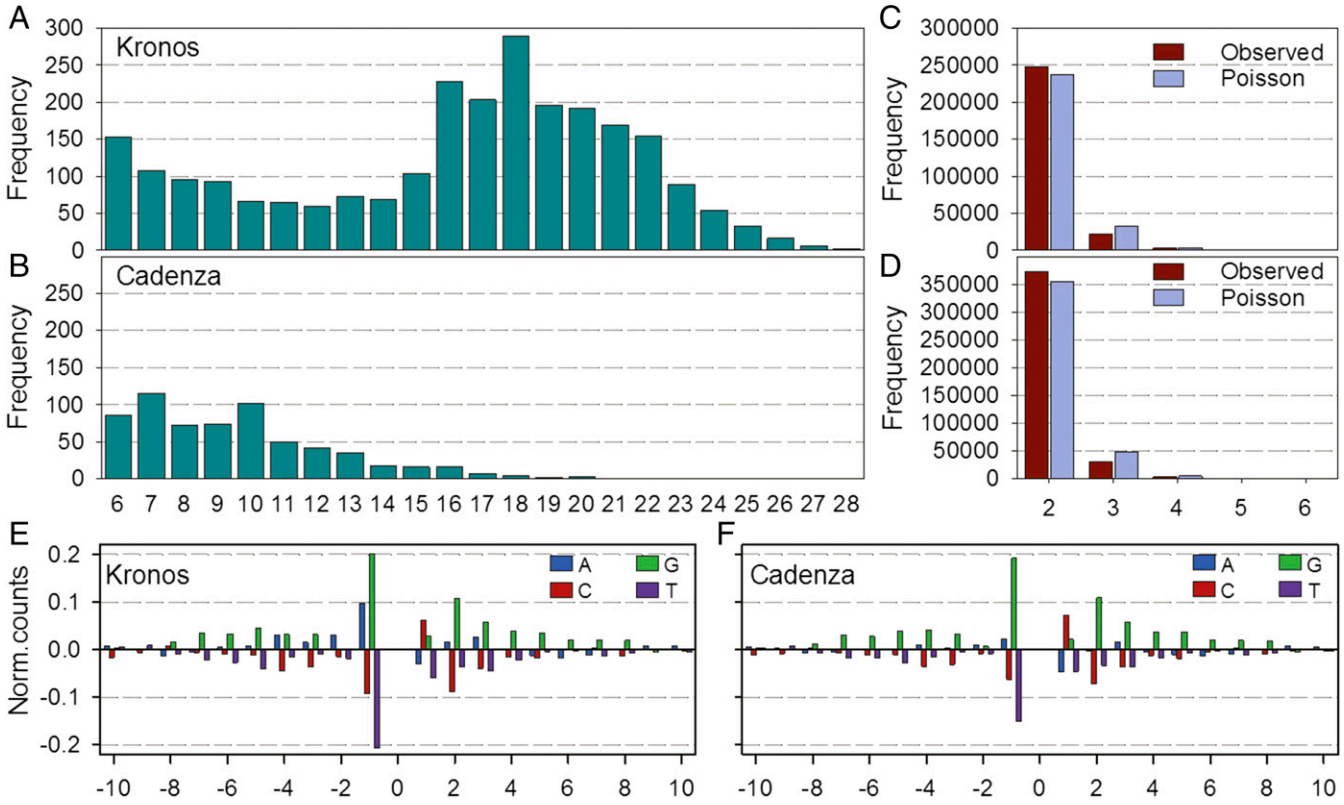

Fig. 3. EMS mutations present in multiple individuals. EMS sequence preference and RH: (*A*, *C*, and *E*) Tetraploid Kronos and (*B*, *D*, and *F*) hexaploid Cadenza. (*A* and *B*) Mutations shared by multiple individuals in RH regions. (*C* and *D*) Observed (red) and closest Poisson distribution (light blue) of mutations present in non-RH regions of multiple individuals. (*A*–*D*) The *x*-axis indicates the number of individuals sharing the same mutation. (*E* and *F*) Sequence preference in regions flanking EMS-type mutations (*SI Appendix*, Fig. S7). The *x*-axis indicates the number of nucleotides upstream (negative) and downstream (positive) from the mutated site.

lines, which is also reflected in the higher average mutation density in Cadenza relative to Kronos. For validation, we selected 11 homozygous large deletions and were able to confirm all of them (*SI Appendix*, Method S10 and Table S19).

Effect of Induced Mutations on Gene Models. We analyzed the effect of EMS-type mutations on gene models with at least one mutation (GM_1; Table 1 and *SI Appendix*, Method S11, Text S3, and Tables S20 and S21). In tetraploid wheat, 59% of GM_1 genes contained at least one truncation (premature stop or splice site) mutation and 96% at least one missense mutation (average, 21.4). In hexaploid wheat, 61% of GM_1 genes contained at least one truncation and 94% contained at least one missense mutation (average, 22.6; Table 1 and *SI Appendix*, Table S21). By using the "sorting intolerant from tolerant" (SIFT) algorithm (17), we found that more than 85% of GM_1 genes across both populations had at least one deleterious missense mutation (SIFT < 0.05). Results combining the SIFT and truncation analyses suggest that our database includes high-quality uniquely mapped mutations that eliminate or reduce function for more than 90% of the captured wheat genes (Table 1 and Fig. 2*A*). As an example of the high frequency of mutations in these populations, we show the presence of truncations or deleterious missense mutations in most of the genes from the wheat starch biosynthesis (Fig. 4*A* and *SI Appendix*, Method S12, Table S22, and Fig. S12) and flowering pathways (Fig. 4*B* and *SI Appendix*, Method S12, Table S23, and Fig. S12).

Mutations in the *Starch Branching Enzyme* genes have been already used to develop pasta and wheat germplasm with increased levels of resistant starch (18, 19), a dietary fiber associated with beneficial effects on human health (20–22). Mutations in wheat flowering genes have been used to dissect the wheat flowering pathway and to modulate wheat flowering time (23–28). For four of these genes, the effects of mutations in individual homeologs were negligible compared with those of the null mutations affecting all homeologs (Fig. 4 *C*–*F*). These results illustrate the limited effects of individual recessive mutations in polyploid wheat.

Access to Mutations, Seed Stocks, and SNP Markers. The EMS-type mutations detected in the Kronos and Cadenza populations at different stringency levels are accessible in public databases and can be visualized using a JBrowse graphic interface (*SI Appendix*, Text S4). Once the desired mutations are identified, the corresponding M_4 seeds can be requested from the University of California, Davis (dubcovskylab.ucdavis.edu/wheat-tilling), and the UK Germplasm Resources Unit (https://www.seedstor.ac.uk/shopping-cart-tilling.php).

In addition, predesigned "Kompetitive Allele Specific PCR" (KASP) primers are available to validate the mutations and to select them for downstream research and breeding applications. In total, we designed 2,771,688 KASP assays for the Kronos population and 3,872,892 assays for Cadenza, the majority of which are genome-specific or -semispecific (72.9% Kronos and 82.8% Cadenza). These primers are provided as part of the output from the public databases.

Discussion

Advantages and Limitations of Sequenced Mutant Populations. The exome-sequenced tetraploid and hexaploid mutant populations can be used for complementary purposes. The tetraploid mutant population is best suited for basic research projects because it allows quicker generation of complete null mutants. This can be achieved through a single cross between A and B genome mutant

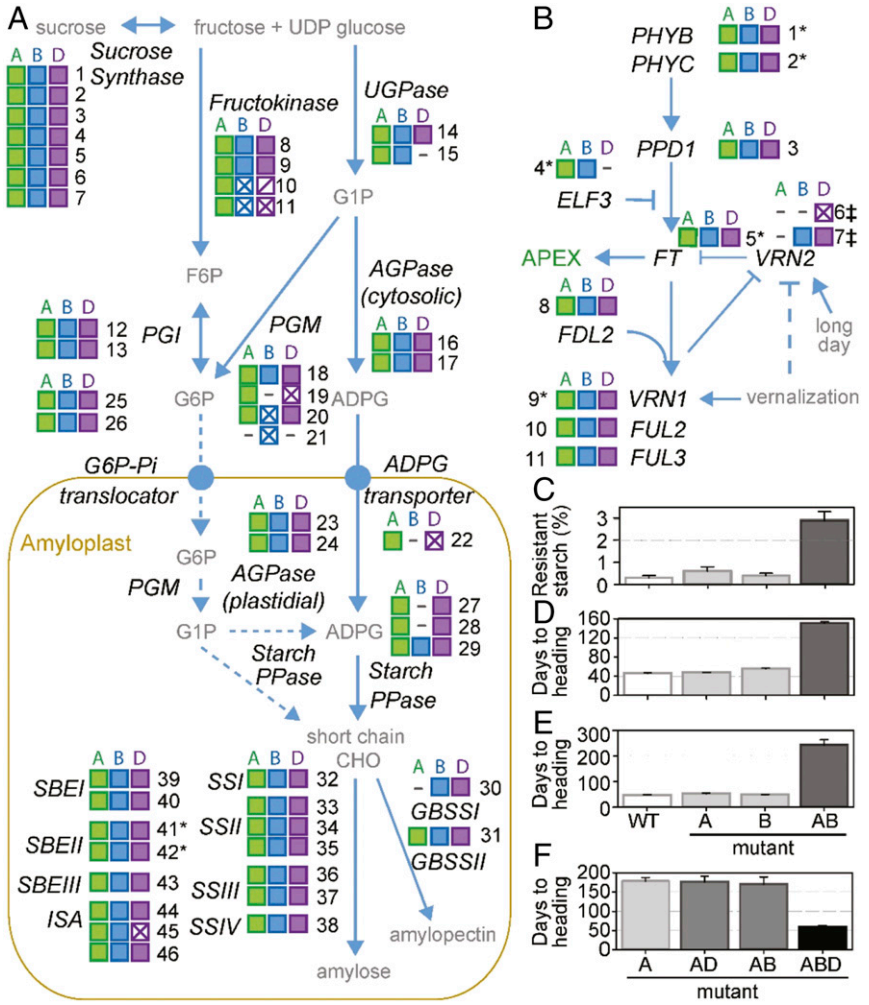

Fig. 4. Predicted effect of mutant alleles on wheat starch biosynthesis and flowering pathways genes. (*A*) Wheat starch biosynthesis and (*B*) flowering pathways. Squares represent individual genes and are colored according to genomes ("A," green; "B," blue; "D," violet). Filled squares represent genes with at least one truncation mutation from the mutant database, squares with crossed diagonals indicate deleterious missense mutations, and single diagonals denote tolerated missense mutations as predicted by SIFT (17). Dashed lines indicate homeologs absent or nonfunctional in the reference. Details for each gene are presented in *SI Appendix*, Tables S22 and S23. (*C–F*) Effect of loss-of-function mutations for (*C*) *STARCH BRANCHING ENZYME IIa/b* (18), (*D*) *PHYTOCHROME C* (25), (*E*) *PHYTOCHROME B* (27), and (*F*) *VERNALIZATION 2* (26). Asterisk indicates published characterization of mutant phenotype. Double dagger indicates published characterization of mutant phenotype based on natural mutations.

lines, followed by self-pollination of the F_1 plant, to generate a segregating F_2 population. Homozygous F_2 double mutants can then be selected by using predesigned KASP assays to generate complete null mutants in the third generation. For breeding applications, however, the hexaploid population is most relevant, as bread wheat represents more than 95% of the wheat grown globally (29). In hexaploid wheat, an additional cross is required to combine mutations across all three genomes. Both populations are in genetic backgrounds with a spring growth habit and have relatively short generation times (10–16 wk depending on growing conditions).

The high mutation rates of these populations are valuable for the identification of novel alleles, including truncation mutations. However, they can also have a negative impact in the characterization of a specific mutation. Strategies to account for this high level of background mutation include the use of multiple independent mutants, backcrossing to reduce mutation load, and selecting for isogenic sibling lines that share background mutations. For mutations with subtle phenotypic effects or that require field phenotyping, it is advisable to backcross the mutant lines to the nonmutagenized parent for at least two generations before combining homeologous mutations (30, 31). Sibling lines can then be selected for homozygous WT or null-mutant alleles. Given that isogenic sibling lines share many of the same background mutations, comparisons between them provide a better assessment of the mutation effect than comparisons vs. the nonmutagenized control. Conversely, for highly penetrant traits, the comparison of complete null and WT F_2 sibling lines without backcrossing may suffice for the analysis.

Recent developments in genome-editing technologies offer the potential to induce mutant alleles in targeted regions of the wheat genome (32–34). Simultaneous homozygous deletions in all three homeologs have been achieved by using transient expression of the Cas9 endonuclease and short guide RNAs complementary to the three homeologs (34). However, homozygous triple mutants in the first T_0 generation were rare (68 plants from 7,680 bombarded embryos) and only ~0.5% (41 of 7,680) corresponded to transgene-free triple homozygous mutants (34). This suggests that, for most genome editing experiments, crossing will still be required to combine single or double mutant lines or to segregate out the DNA construct. Genome editing has the advantage of lower levels of background mutations, but the disadvantage of the high upfront

expense in construct design and transformation costs. By contrast, the sequenced mutant populations provide researchers with instant access to the mutant alleles with a simple online search and inexpensive seed request. Efficient wheat transformation is still limited to a small number of large research institutions, so the public sequenced mutant populations have the potential to democratize access to reverse genetic resources in wheat.

It is still not clear if genome-edited crops will be regulated as nontransgenic in all countries, which may impose constrains for globally traded crops such as wheat. There are still no commercially available transgenic wheat varieties, whereas EMS mutations have been used in agriculture for almost a century and are not under any regulation. We therefore predict that the mutant populations developed in this work will be very valuable and highly complementary to editing approaches in the future.

Uncovering Hidden Recessive Variation. The results for the flowering repressor *VRN2* mutants (Fig. 4*F*) are particularly illustrative of the limited phenotypic effect of recessive mutations in polyploid wheat. To date, no polyploid wheat variety has been described with a spring growth habit associated with recessive *vrn2* alleles (26). This is not caused by a lack of effect in polyploid wheat, as loss-of-function mutations at all three *VRN2* homeologs in hexaploid wheat results in a spring growth habit (26). It is also not caused by limited selection pressure, as more than 10 independent dominant mutations for spring growth habit have been described for the meristem identity gene *VRN1* (26). Finally, it is unlikely that the lack of spring types associated with *vrn-2* is caused by a low adaptive value, as most accessions of cultivated diploid wheat *Triticum monococcum* (35) and a large number of diploid barley varieties (36, 37) have a spring growth habit associated with recessive *vrn-2* mutations. Based on the previous evidence, we hypothesize that recessive mutations at the *VRN2* locus in polyploid wheat have remained hidden from selection for more than 8,000 y by the redundancy conferred by multiple homeologs.

Given the recent origin of the polyploid wheat species, many potentially useful induced and natural mutations are likely masked by functional redundancy among homeologs. The >10,000,000 sequenced mutations identified in wheat coding regions in the present study facilitate the identification of loss-of-function mutations in different homeologs and generates a large number of alleles. These mutations can be combined to study gene function and to reveal previously hidden phenotypic variation. Likewise, the effects of candidate genes from diploid grass species can now be studied directly in wheat, as recently shown for the wheat *TaGW2-A1* mutants with increased grain size identified in the tetraploid population (31). The strategy and methods developed herein can be also applied to other young polyploid crops with closely related genomes.

In summary, the mutant populations sequenced in the present study represent an invaluable resource for wheat functional genetics and provide a powerful tool to uncover variation previously hidden to human and natural selection in a central crop species for global food security.

Materials and Methods

Exome Capture Design. The wheat exome capture designs used in the present study were developed in collaboration with NimbleGen (Roche) and are publicly available to order from Roche catalog numbers 140228_Wheat_Dubcovsky_D18_REZ_HX1 (tetraploid wheat) and 140430_Wheat_TGAC_D14_REZ_HX1(hexaploid wheat). The sequences in this design comprise protein-coding transcript data from *T. turgidum* and *T. aestivum* transcriptome studies, wheat ESTs, wheat sequences homologous to barley gene models (38) not present in the previous wheat datasets, and hand-annotated sequences that were crowd-sourced from the wheat research community (*SI Appendix*, Table S1). A detailed description of the methods used for the development of the exome capture is presented in *SI Appendix*, Method S1.

A total of 82,511 protein coding sequences passed through all filters into the final design (*SI Appendix*, Table S1). Exon prediction was performed by aligning transcripts to the Chinese Spring Survey (CSS) sequences using the exonerate program (39) as described previously (40). Individual exons were padded with 30 bp from the introns on each side of the exons to increase capture efficiency at exon/intron borders. For *T. aestivum* sequences, all exon and padded sequences were derived from the genomic assembly. For *T. turgidum* sequences, original sequences were retained for the exons and only padding was supplemented from the *T. aestivum* genome. In total, we included 219,383 padded and 67,416 unpadded exons in the design, covering a total of 84 Mb (*SI Appendix*, Table S1). The exome capture design is available for BLAST and can be downloaded at dubcovskylab.ucdavis.edu/wheat-tilling and www.wheat-tilling.com.

Sample Preparation. A detailed description of the methods used for genomic DNA extraction and shearing, library construction and barcoding, capture hybridization and DNA recovery, amplification. and sequencing are presented in *SI Appendix*, Method S2 and Table S2.

Data Processing and Mapping Rates. Illumina 100-bp paired-end reads were preprocessed to trim 3′ adapter sequences and low quality. Trimmed reads were aligned to genome scaffolds of the CSS sequence (AB genomes for Kronos and ABD genomes for Cadenza) using bwa (41). For Cadenza, CSS scaffolds for chromosome 3B were replaced with the 3B pseudomolecule assembly (42). Alignments were sorted by using samtools (41), and duplicate reads were removed with Picard tools. Additional information is provided in *SI Appendix*, Method S3.

De Novo Assembly of Unmapped Reads. To increase the proportion of mapped reads, we supplemented the CSS reference with a de novo assembly of unmapped reads from Kronos and Cadenza. Supplementary de novo assemblies for tetraploid and hexaploid wheat were constructed separately by using 43,073,616 unmapped reads from 14 Kronos samples and 56,988,370 unmapped reads from 10 Cadenza samples (*SI Appendix*, Method S3 and Table S3).

MAPS Parameter Optimization. To identify EMS-induced mutations, we used the MAPS software that was previously tested on rice and eight wheat mutant lines (16) (*SI Appendix*, Method S4 and Fig. S2). In each MAPS run, we processed batches of 24 samples for tetraploid wheat and 24 or 32 samples for hexaploid wheat. Only SNPs detected in a single sample of the batch are reported by the MAPS pipeline. This removes varietal SNPs between the CS reference and Kronos/Cadenza and is also critical in polyploid species to eliminate polymorphisms among homeologs, which are present across all samples.

In wheat, EMS generates almost exclusively G to A and C to T changes, which are referred in the present study as EMS-type mutations. Therefore, the proportion of non–EMS-type SNPs can be used as a first approximation to the error rate. We also estimated the error rate by comparing the number of SNPs detected in the nonmutagenized WT line (no mutations expected) with the average number of mutations detected in the EMS-treated plants. By using these error estimates, we empirically adjusted several parameters in the MAPS pipeline to minimize the detection of false mutations without losing too much sensitivity (*SI Appendix*, Method S4).

Validation of EMS Mutations. We validated EMS mutations by examining their status in M_4 plants by using genome-specific KASP assays designed with PolyMarker (43) and through direct Sanger sequencing (*SI Appendix*, Tables S8–S10). The main objectives were to confirm the presence of the mutation in the M_4 progeny seed deposited in the public repositories and classify the M_2 as homozygous or heterozygous in the M_4 progeny seed.

Correction for Heterozygous/Homozygous Mutation Classification. In a single M_2 individual, the frequency of WT and mutant alleles is expected to be close to 50% for a heterozygous mutation. However, MAPS classifies mutations as heterozygous even when a single WT read is present at low frequency. A single mismapped read can result in a homozygous mutation being misclassified as heterozygous and in the overestimation of the ratio between heterozygous to homozygous mutations. This problem is exacerbated in polyploid species with similar genomes. To correct these errors, we introduced a filter in the bioinformatics pipeline that reclassified heterozygous mutations as homozygous when the frequency of the WT allele was less than 15% of reads (*SI Appendix*, Fig. S5).

Calculation of Mutation Density and Coverage. To estimate the mutation density (number of mutations per kb of captured sequence) across the population, we divided the total number of uniquely mapped mutations identified at

HetMC5/*HomMC3* (excluding RH) by the average number of positions used by MAPS to identify mutations (119.2 Mb for Kronos and 162.4 Mb for Cadenza). To calculate these last two numbers, we first obtained the number of bases covered by at least one read at quality higher than 20 in at least $N - 4$ samples from the MAPS batch from the intermediate "MAPS-assay" file (*SI Appendix*, Fig. S2). We then determined the number of bases covered by at least four reads in each individual, and used the average across all mutants in the population to estimate the average number of positions used by MAPS to identify mutations.

The coverage values presented in *SI Appendix*, Table S4, are based on mutant positions. These positions are selected by MAPS to have a minimum coverage of three reads and to be present in a high proportion of the samples in the same run, and therefore their coverage can be higher than the average from the complete population. However, the coverage values obtained for Kronos by using the previous method (26.6) was almost identical to the value obtained for 89.2 million positions independently of mutations in the comparison of the α-design and β-design in Kronos (28.8; *SI Appendix*, Fig. S1). This result suggests that coverage values estimated from mutant positions are not very different from the ones in the overall population.

EMS Sequence Preference. Sequence preference in sites adjacent to EMS mutation sites were calculated by using the method described previously (16). Briefly, we measured nucleotide frequencies in 20-bp regions flanking the mutated G nucleotides and compared it to regions flanking a nonmutagenized G located 40–50 bp upstream and downstream of each mutation site. Sequence preference at each position was expressed as the difference between the percent frequency in the base flanking the mutated G and the corresponding frequency in the control sites (Fig. 3 *E* and *F* and *SI Appendix*, Fig. S7).

Reads Mapped to Multiple Locations (i.e., Multimapped Reads). The bioinformatics pipeline used to detect MM is described in *SI Appendix*, Fig. S3, and the methods used to select the primary location, visualize the alternative locations, and validate the MM mutations are described in *SI Appendix*, Method S8.

Identification and Validation of Large Deletions. To identify and characterize homozygous deletions in our mutant populations, we developed a custom bioinformatics pipeline that examines relative coverage of exons within and across mutant lines (*SI Appendix*, Method S9 and Fig. S4). The methods used to validate these large mutations are described in *SI Appendix*, Method S10).

Variant Effect Prediction. Mutation effects on gene function were predicted on the final SNP files (*HetMC5*/*HomMC3*) without RH using the Variant Effect Predictor program (44) from Ensembl tools release 78 in offline mode (*SI Appendix*, Method S11).

Access to and Visualization of Mutations. The raw reads for the tetraploid and hexaploid projects are available from National Center for Biotechnology Information BioProject PRJNA258539 and European Read Archive ENA (European Nucleotide Archive) study PRJEB11524, respectively. In addition, we deposited the uniquely mapped EMS-type mutations at *HetMC5*/*HomMC3* (excluding RH) in EnsemblPlants. These platforms are searchable through string searches or BLAST queries.

Access to Mutant Seed Stocks and SNP Markers. For the Kronos and Cadenza Targeting Induced Local Lesions in Genomes (TILLING) populations, M_3 seed was collected from the individual M_2 plants used for DNA extraction and exome sequencing. For initial seed bulking, ~30 M_3 siblings were grown in the field as single rows and all M_4 seed was harvested and bulked for each mutant line. For lines with low yields (<60 g), an additional set of M_3 siblings was grown in the glasshouse or field to increase seed quantity. Additional backups of the complete tetraploid mutant population have been deposited in Centro Internacional de Mejoramiento de Maiz y Trigo (Mexico), Shandong University, the University of Saskatchewan, the quarantine repository in Australia, the Cereal Disease Laboratory, and Washington State University. Likewise, backups of the complete hexaploid mutant population have been deposited at Rothamsted Research, National Institute of Agricultural Botany, the French National Institute for Agricultural Research, and University College Dublin.

To generate KASP assays for the *HetMC5*/*HomMC3* EMS-type mutations in the database, we ran the PolyMarker pipeline (43). For the allele-specific primers, fluorophore-compatible tails need to be added to the 5′ end before oligo synthesis (45).

Code Availability. All code is available through the wheat TILLING project GitHub page: https://github.com/DubcovskyLab/wheat_tilling_pub.

ACKNOWLEDGMENTS. The authors thank Dr. Marcelo Soria for help with the EST analysis; Dr. Wenjun Zhang for mutation validation; Dr. Luca Comai for valuable suggestions for the analysis of TILLING datasets; Dr. Robert King for assistance with the deletion analyses; Meric Lieberman for Mutation and Polymorphism Survey (MAPS) advice; Dr. Martin Trick for developing initial stages of the www.wheat-tilling.com database; Drs. Mike Ambrose, Adrian Turner, and Richard Horler for developing the TILLING seed database at SeedStor; the John Innes Centre (JIC) field trials and horticultural team for plant husbandry; members of the Platforms and Pipelines Group for the next-generation sequencing and library construction; and the Norwich BioScience Institutes Computing infrastructure for Science group through the JIC and Earlham Institute clusters. This work was supported by the Howard Hughes Medical Institute (J.D.); Gordon and Betty Moore Foundation Grant GBMF3031 (to J.D.); NRI Competitive Grants 2011-68002-30029 and 2017-67007-25939 from the US Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA; to J.D.); UK Biotechnology and Biological Sciences Research Council (BBSRC) Grants BB/J003557/1, BB/J003913/1, and BB/J003743/1 (to C.U., K.V.K., A.P., S.A., and L.C.); BBSRC and Institute Strategic Programme Grant at The Earlham Institute BB/J004669/1; BBSRC National Capability in Genomics at The Earlham Institute Grant BB/J010375/1; BBSRC Future Leader Fellowship BB/M014045/1 (to P. Borrill); USDA NIFA postdoctoral fellowship 2012-67012-19811 (to K.V.K.), a Norwich Research Park PhD Studentship (to R.H.R.-G.), and an The Earlham Institute Funding and Maintenance Grant (to R.H.R.-G.). The following institutions contributed funding to sequence 100 tetraploid mutant lines each: Shandong University, University of Saskatchewan, Commonwealth Scientific and Industrial Research Organisation, USDA–Agricultural Research Service Cereal Disease Laboratory, and Washington State University.

1. Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316(5833):1862–1866.
2. Tsai H, et al. (2013) Production of a high-efficiency TILLING population through polyploidization. *Plant Physiol* 161(4):1604–1614.
3. Rakszegi M, et al. (2010) Diversity of agronomic and morphological traits in a mutant population of bread wheat studied in the Healthgrain program. *Euphytica* 174: 409–421.
4. Slade AJ, Knauf VC (2005) TILLING moves beyond functional genomics into crop improvement. *Transgenic Res* 14(2):109–115.
5. Uauy C, et al. (2009) A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biol* 9:115.
6. Wang TL, Uauy C, Robson F, Till B (2012) TILLING *in extremis*. *Plant Biotechnol J* 10(7):761–772.
7. Mamanova L, et al. (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7(2):111–118.
8. King R, et al. (2015) Mutation scanning in wheat by exon capture and next-generation sequencing. *PLoS One* 10(9):e0137549.
9. Jiao Y, et al. (2016) A sorghum mutant resource as an efficient platform for gene discovery in grasses. *Plant Cell* 28(7):1551–1562.
10. Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci* 274(933):227–274.
11. Mayer KFX, et al. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345(6194):1251788.
12. Jordan KW, et al. (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol* 16:48.
13. Schreiber AW, et al. (2012) Transcriptome-scale homoeolog-specific transcript assemblies of bread wheat. *BMC Genomics* 13:492.
14. Guo Y, Abernathy B, Zeng Y, Ozias-Akins P (2015) TILLING by sequencing to identify induced mutations in stress resistance genes of peanut (*Arachis hypogaea*). *BMC Genomics* 16:157.
15. Cannon SB, Shoemaker RC (2012) Evolutionary and comparative analyses of the soybean genome. *Breed Sci* 61(5):437–444.
16. Henry IM, et al. (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell* 26(4):1382–1397.
17. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31(13):3812–3814.
18. Hazard B, Zhang X, Naemeh M, Dubcovsky J (2014) Registration of durum wheat germplasm lines with combined mutations in *SBEIIa* and *SBEIIb* genes conferring increased amylose and resistant starch. *J Plant Regist* 8(3):334–338.
19. Schönhofen A, Hazard B, Zhang X, Dubcovsky J (2016) Registration of common wheat germplasm with mutations in *SBEII* genes conferring increased grain amylose and resistant starch content. *J Plant Regist* 10(2):200–205.
20. Amini A, Khalili L, Keshtiban AK, Homayouni A (2016) Resistant starch as a bioactive compound in colorectal cancer prevention. *Bioactive Foods in Health Promotion*, eds Watson RR, Preedy VR (Academic, Cambridge, UK), pp 773–780.
21. Robertson MD, Bickerton AS, Dennis AL, Vidal H, Frayn KN (2005) Insulin-sensitizing effects of dietary resistant starch and effects on skeletal muscle and adipose tissue metabolism. *Am J Clin Nutr* 82(3):559–567.
22. Wong THT, Louie JCY (2016) The relationship between resistant starch and glycemic control: A review on current evidence and possible mechanisms. *Starke* 68:1–9.
23. Alvarez MA, Tranquilli G, Lewis S, Kippes N, Dubcovsky J (2016) Genetic and physical mapping of the earliness per se locus *Eps-A^{m}1* in *Triticum monococcum* identifies *EARLY FLOWERING 3* (*ELF3*) as a candidate gene. *Funct Integr Genomics* 16(4):365–382.

24. Chen A, Dubcovsky J (2012) Wheat TILLING mutants show that the vernalization gene *VRN1* down-regulates the flowering repressor *VRN2* in leaves but is not essential for flowering. *PLoS Genet* 8(12):e1003134.
25. Chen A, et al. (2014) PHYTOCHROME C plays a major role in the acceleration of wheat flowering under long-day photoperiod. *Proc Natl Acad Sci USA* 111(28):10037–10044.
26. Kippes N, Chen A, Zhang X, Lukaszewski AJ, Dubcovsky J (2016) Development and characterization of a spring hexaploid wheat line with no functional *VRN2* genes. *Theor Appl Genet* 129(7):1417–1428.
27. Pearce S, Kippes N, Chen A, Debernardi JM, Dubcovsky J (2016) RNA-seq studies using wheat *PHYTOCHROME B* and *PHYTOCHROME C* mutants reveal shared and specific functions in the regulation of flowering and shade-avoidance pathways. *BMC Plant Biol* 16(1):141.
28. Lv B, et al. (2014) Characterization of *FLOWERING LOCUS T1* (*FT1*) gene in *Brachypodium* and wheat. *PLoS One* 9(4):e94171.
29. Taylor RD, Koo WW (2015) 2015 Outlook of the U.S. and World Wheat Industries, 2015-2024. Agribusiness & Applied Economics. Agribusiness & Applied Economics Report 738 (North Dakota State Univ, Fargo, ND), p 23.
30. Hazard B, et al. (2015) Mutations in durum wheat *SBEII* genes affect grain yield components, quality, and fermentation responses in rats. *Crop Sci* 55(6):2813–2825.
31. Simmonds J, et al. (2016) A splice acceptor site mutation in *TaGW2-A1* increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. *Theor Appl Genet* 129(6):1099–1112.
32. Shan Q, Wang Y, Li J, Gao C (2014) Genome editing in rice and wheat using the CRISPR/Cas system. *Nat Protoc* 9(10):2395–2410.
33. Wang W, Akhunova A, Chao S, Akhunov E (2016) Optimizing multiplex CRISPR/Cas9-based genome editing for wheat. *bioRxiv*, 10.1101/051342.
34. Zhang Y, et al. (2016) Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nat Commun* 7:12617.
35. Yan L, et al. (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* 303(5664):1640–1644.
36. Karsai I, et al. (2005) The *Vrn-H2* locus is a major determinant of flowering time in a facultative x winter growth habit barley (*Hordeum vulgare* L.) mapping population. *Theor Appl Genet* 110(8):1458–1466.
37. Szűcs P, et al. (2007) Validation of the *VRN-H2*/*VRN-H1* epistatic model in barley reveals that intron length variation in *VRN-H1* may account for a continuum of vernalization sensitivity. *Mol Genet Genomics* 277(3):249–261.
38. Mayer KFX, et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491(7426):711–716.
39. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
40. Krasileva KV, et al. (2013) Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol* 14(6):R66.
41. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
42. Choulet F, et al. (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science* 345(6194):1249721.
43. Ramirez-Gonzalez RH, Uauy C, Caccamo M (2015) PolyMarker: A fast polyploid primer design pipeline. *Bioinformatics* 31(12):2038–2039.
44. McLaren W, et al. (2016) TheEnsembl Variant Effect Predictor. *Genome Biol* 17(1):122.
45. Ramirez-Gonzalez RH, et al. (2015) RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnol J* 13(5):613–624.