Title: The *Setaria viridis* genome and diversity panel enables discovery of a novel domestication gene

Authors: Pu Huang[1,5], Sujan Mamidi[2], Adam Healey[2], Jane Grimwood[2], Jerry Jenkins[2], Kerrie Barry[3], Avinash Sreedasyam[2], Shengqiang Shu[3], Maximilian Feldman[1,6], Jinxia Wu[1,7], Yunqing Yu[1], Cindy Chen[3], Jenifer Johnson[3], Hitoshi Sakakibara[4,8], Takatoshi Kiba[4,9], Tetsuya Sakurai[4,9], Daniel Rokhsar[3], Ivan Baxter[1], Jeremy Schmutz[2,3], Thomas P. Brutnell[1,7], Elizabeth A. Kellogg[1],*

[1] Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132, USA

[2] HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA

[3] Department of Energy Joint Genome Institute, Walnut Creek, California, USA

[4] RIKEN Center for Sustainable Resource Science, Tsurumi, Yokohama 230-0045, Japan

[5] present address: BASF Corporation, 26 Davis Dr., Durham, NC 27709, USA

[6] present address: USDA-ARS Temperate Tree Fruit and Vegetable Research Unit, 24106 N. Bunn Rd., Prosser, WA 99350, USA

[7] Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[8] present address: Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya 464-8601, Japan

[9] present address: Multidisciplinary Science Cluster, Kochi University, Nankoku, Kochi, 783-8502, Japan

*author for correspondence

Abstract

We present a platinum-quality genome assembly for the model grass *Setaria viridis*, and high quality genomic sequences of 600+ wild accessions (average 42.6x coverage). Presence-absence variation (PAV) and single-nucleotide polymorphisms (SNPs) identify several subpopulations in North America. Using genome-wide association mapping plus CRISPR-Cas9 technology, we identified and validated *Less Shattering1* (*SvLES1*), a gene for seed shattering with a retrotransposon insertion in the domesticated *S. italica* (foxtail millet) allele. We also identified a candidate gene for erect leaves, orthologous to the maize gene *liguleless2*. These results demonstrate the utility of the model plant *S. viridis* for complex trait dissection in panicoid crops.

Introduction

Maize, sorghum and most species of millet are mainstays of industrial and small-holder agriculture. All are warm-season (C_4) grasses in the subfamily Panicoideae, a group of ca. 3300 species that includes not only essential grain, forage and biofuel crops, but also ecological dominants of tropical and warm temperate environments [1-3]. Improving the crops, exploring the domestication process, and unraveling the ecological and evolutionary history of the group all depend on development of genetic and genomic resources.

Setaria viridis (green foxtail) has been proposed as a practical experimental model for understanding the phenotypic consequences of genetic variation in panicoid grasses [4-6]. *Setaria viridis* plants are generally small (Figure 1a), with a short life cycle (seed to seed in 8-10 weeks) and self-compatible, with a single inflorescence that often produces hundreds of seeds. Transformation is efficient, and as we demonstrate here, is amenable to high efficiency CRisPR-Cas9 mediated mutagenesis.

As in most wild species, seeds of *S. viridis* fall off the plant at maturity, a process known as shattering [7]. A zone of specialized cells, the abscission zone (AZ), forms below the spikelet during seed maturation, and loss of connections between the cells of the AZ result in cell separation and seed drop. *Setaria italica* (foxtail millet) is the domesticated form of *S. viridis*; like most domesticates it fails to form an AZ and thus is non-shattering [8]. While loss of shattering has been an important early step in the domestication of all cultivated grasses, the genes involved generally differ among species [7,9-12]. In particular, genes and QTL identified in rice domestication are, for the most part, not conserved among other species [7,11,12]. However, most of these studies

have compared wild species to domesticates and none have considered the standing variation for reduced shattering in wild accessions.

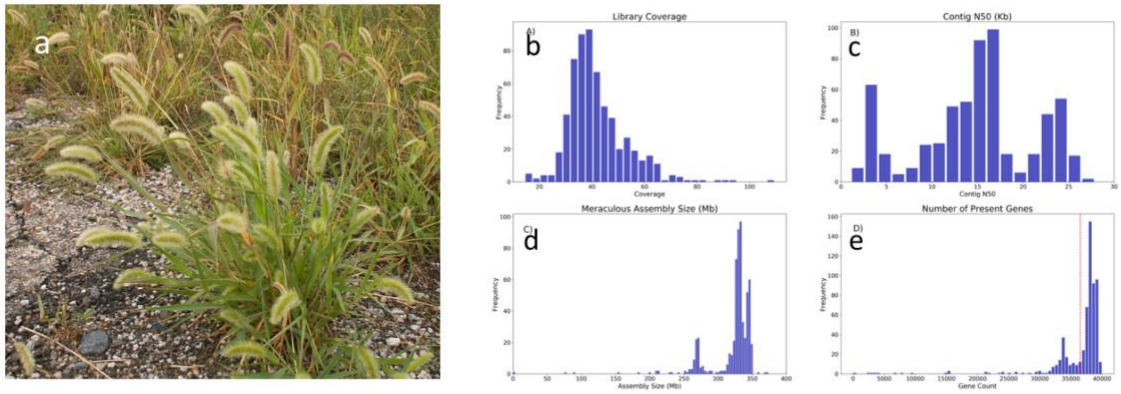

Fig. 1 a *Setaria viridis*, in its common highly disturbed habitat next to a road. **b-e** Assembly and gene presence/absence variation statistics for the 605 accessions of the *S. viridis* diversity panel. **b** Average library coverage. **c** Contig N50 (Kb). **d** Meraculous assembled bases. **e** Number of genes per library; red vertical line represents the number of genes necessary for a library to be included for PAV analysis (n=36,500).

In this work we identify a novel locus associated with reduced seed shattering by surveying the genetic diversity in natural populations of *S. viridis*. To achieve this we have deployed a large sequence-based resource for *S. viridis*, and analyzed its population structure using single-nucleotide polymorphisms (SNPs) and presence-absence variation (PAV) of individual genes. Subsequently we verified the function of the shattering locus using genome editing and, by investigating sequences of the orthologous gene in *S. italica*, showed that the locus also contributed to domestication. Data presented here show that genomics and biotechnological resources developed to support this model panicoid species can be used to accelerate our mechanistic understanding of evolutionary genetic processes.

Diversity panel and an updated genome

We have generated significant new resources for the *Setaria* community. A new assembly (v. 2.0) was generated for the *S. viridis* reference line A10.1, using a combination of Pacbio and Illumina sequencing technologies. The final version 2.0 release (https://phytozome.jgi.doe.gov/pz/portal.html) contains 395.1 Mb of sequence, consisting of 75 contigs with a contig N50 of 11.2 Mb and a total of 99.95% of assembled bases in chromosomes. The structural annotation of the genome was completed using the Joint Genome Institute plant annotation pipeline (Shu, unpublished) to identify 38,334 gene models with an additional 14,125 alternative transcripts.

Setaria viridis diversity samples (n=605) collected throughout North America were sequenced using Illumina 2x150 paired end (PE) libraries to an average genome coverage of 42.6X (Figure 1b). Individual de novo assemblies for each line were also constructed using Meraculous (v2.2.5) [13] with a kmer size of 51, selected to maximize the contig N50 in the resultant assemblies, and to ensure that alternative haplotypes would have the best chance of being split apart. The average results per assembly were: Number of contigs: 75,001; Contig N50: 16.2 Kb (Figure 1c); Assembly bases: 322.5 Mb (Figure 1d).

To construct chromosomes for each library, exons from the *S. viridis* annotated reference (v2.1; number of genes = 38,334; number of exons = 289,357) were aligned to each Meraculous assembly (blastn, word_size = 32), and exon alignments with identity >=90% and coverage >=85% were retained. Chromosomes were constructed by first joining scaffolds based on exon alignments into gene-based scaffolds; synteny

and exon alignments were then used to order and orient the sequences into chromosomes (Figure S1).

SNP calling and PAV identification

SNP variation was assessed for 598 of the 605 strains of *S. viridis* spread over North America (seven excluded because of low sequence coverage)(Table S1). An average of 56 million (M) high quality paired end reads (range 19M-177M) per line corresponded to an average of 42.6x coverage (range 14x – 134x), with ca. 80% of the reads of each accession mapped to the reference genome. On average, 88% of the genome had a coverage of 8x to 500x, and was used for SNP calling. 8.58 million SNPs, or about one for every 21.6Kbp, were identified, of which 612 thousand (K) mapped to exons; of these 335K SNP are missense, 6K are nonsense, and 275K are silent.

To determine gene presence/absence variation (PAV) across the diversity panel, *S. viridis* (v2.1) proteins and non-orthologous *Setaria italica* (v2.2) proteins (based on InParanoid comparisons [14]) were aligned using blat (-noHead -extendThroughN -q=prot -t=dnax) [15] to each chromosome integrated assembly. *S. italica* genes were included in the PAV analysis to capture genes that may vary in wild accessions of *S. viridis*, but are not annotated in the v2.1 reference. Genes were considered present if they aligned with greater than 85% coverage and identity, or at least 90% if the exons were broken up and located on no more than three scaffolds. Libraries with fewer than 36,500 genes present were excluded to avoid introduction of artifacts attributable to poor assembly and integration (Figure 1e). The resultant PAV matrix (Table S2) was then converted

into vcf format, using each protein's best hit genomic coordinates within the *S. viridis* v2.1 reference.

To empirically determine and cluster genes into their pan-genome designation (core, "soft"-core, shell, cloud) [16-18], discriminant function analysis of principal components (DAPC) [19] was used. DAPC found four distinct clusters, designated core (average observation: 98% across all libraries), "soft"-core (average observation: 70%), shell (average observation: 41%), and cloud (average observation: 14%).

Population genetic analyses

Population structure was assessed independently for SNPs and PAV data (shell genes) using fastStructure [20] and Admixture [21]. Both approaches identified three main sub-populations along with a heavily admixed population (Figure 2a,b). After removing admixed individuals from the PAV analysis ($Q<0.7$; n=217), there were three dominant subpopulations (n = 34, 72, 96, respectively), which ranged from the west-coast (green symbols), central-north (purple), and central-east (blue), plus a smaller central population (red; n=15; Figures 2c, S2). Comparing the PAV STRUCTURE classification to that of the SNP data, we found 98% agreement between the two (216 of 220 non-admixed samples). 3,355 genes varied significantly ($p<0.05$) among their expected observations within each subpopulation (Table S3). Results of a GO analysis for genes that were over- and under-represented in each population are shown in Figure S3. The Central-north population has the highest SNP diversity of the three large populations (Table S4). However, a neighbor-joining (NJ) tree (Figure 2d) shows that it is not

monophyletic, indicating a rapid radiation or complex history of gene exchange, possibly explaining the high diversity estimate.

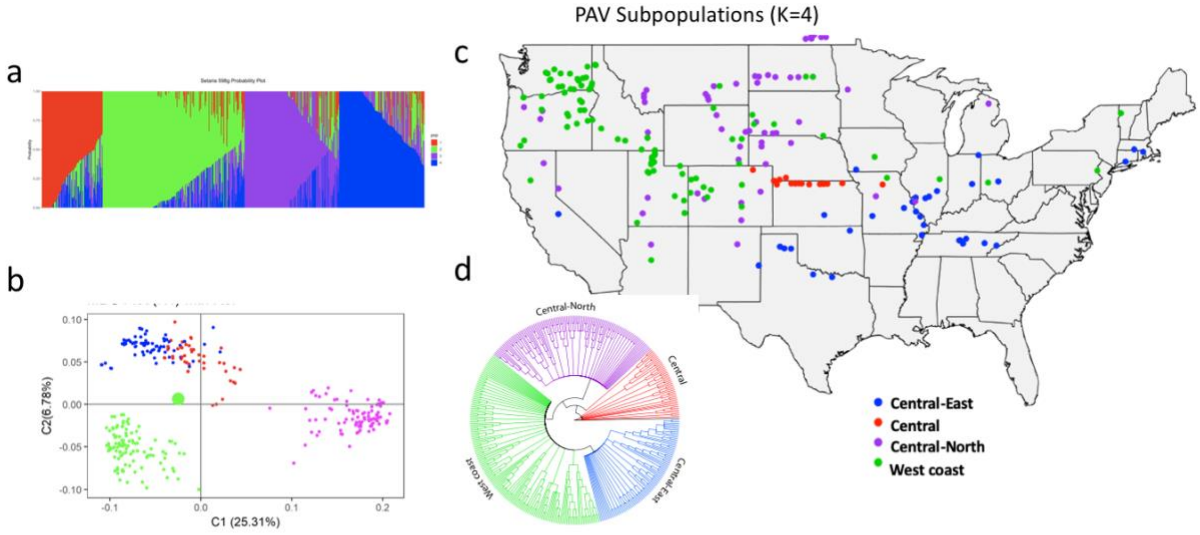

Fig. 2 Population differentiation of *Setaria viridis* in North America. **a, b, d** data from SNPs. **c** data from PAV. Subpopulations color coded as: Central-north, purple, Central, red, Central-east, blue, West coast, green. **a** STRUCTURE analysis with K=4; Central (red) population is extensively admixed. **b** multidimensional scaling of the same data as in **a**, showing distinctiveness of the Central-north (purple) population. **c** distribution of subpopulations with admixed (Q<0.7) individuals removed; population assignment and distribution are highly similar to that shown by SNP data (**Fig. S2**). **d** Neighbor-joining tree of SNP data, rooted with *S. adhaerens* (outgroup).

The root of the NJ tree falls within the small Central population 1 (red dots in Figure 2c, S2), corresponding approximately to that identified by Huang et al [22] as including material from China. The history of *S. viridis* in North America is unknown, although other phylogenetic studies place it within a clade of Asian species [23,24]. It is presumed to have made its way across both the Atlantic and Pacific Oceans following the voyages of exploration, although we cannot rule out the possibility that it is a North American native that has spread worldwide. Huang et al. [22] showed that each North American population has a counterpart elsewhere in the world, consistent with a history of repeated introductions.

Selection

All CDS in the Setaria genomes were tested for evidence of positive (diversifying) selection using four distinct tests, each of which has a different set of assumptions and underlying models. Loci inferred to be under selection by combinations of selection tests are listed in Table S5, with those identified by all four tests in Table S6.

A novel gene, *SvLes1*, controls seed shattering in *S. viridis*

We deployed the new high quality genome, pan-genome and population genetic analyses to link genotype to agronomically important phenotypes. A genome-wide association study (GWAS) identified a single strong QTL (peak $-\log_{10} p>30$) for seed shattering on Chr_05 (Figure 3a) covering a large region (2Mb) above the significance threshold as defined by experiment-wise p=0.01 (Bonferroni correction). We isolated 119 mutations in this region that alter protein sequences relative to the reference A10, primarily missense mutations. To narrow down the causal mutation, we used a computational approach [29] to predict deleterious mutations that are more likely to alter the biological function of protein products. Combining this prediction and the association score of SNPs (Figure 3b), we prioritized a single C to T mutation (Chr_05:6849363) in this region as our primary candidate. This mutation occurs in a gene encoding a MYB transcription factor, *SvLes1* (Sevir.5G085400), with two MYB DNA-binding domains. The mutation leads to a R84S substitution in *SvLes1*, located in a conserved position in the second MYB DNA-binding domain (Figure 3c). We name these two alleles *SvLes1-WT* and *SvLes1-mutant*, associated with high seed shattering and reduced shattering, respectively. The *SvLes1-mutant* allele appears in more than 5% of the ~200

accessions of the GWAS panel, and these accessions show clear association with reduced shattering scores. Interestingly, the reference line A10.1 is one of them, possibly favored by researchers due to its low shattering phenotype.

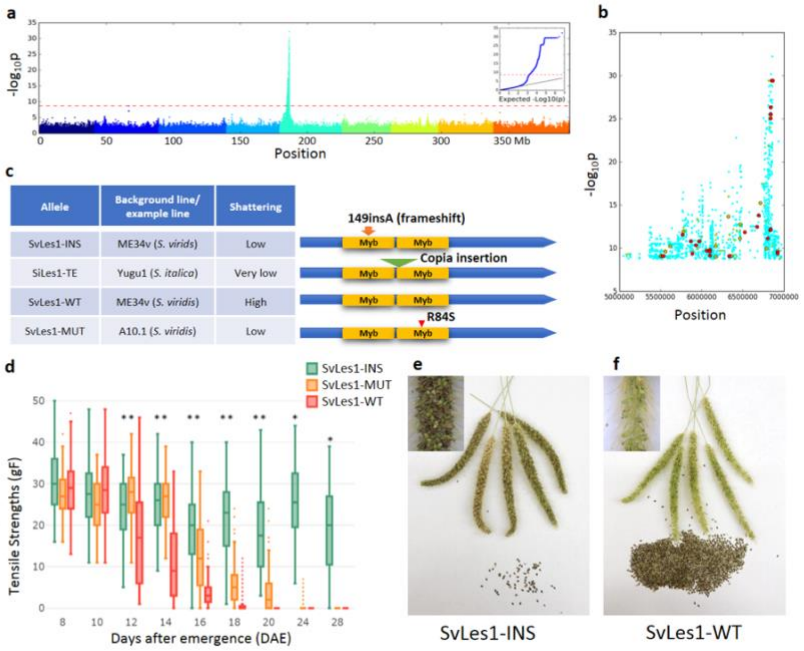

Fig. 3 Phenotyping and GWAS of reduced shattering. **a** Manhattan plot of GWAS result, red line showing p=0.01 after correction. **b** Zoom in to peak on Chromosome 5; larger dots represent missense SNPs identified by snpEff. Different colors of missense SNPs indicate provean score range (blue for >-2.5, green for <-2.5 and >-4.1, red for <-4.1; -2.5 and -4.1 represent 80% and 90% specificity). Lower scores indicate higher likelihood of deleterious effects of the mutation. **c** Allele obtained, background, shattering phenotype and allele effect on protein product. **d** Tensile strength measurements across panicle developmental stages with three alleles. **e,f** End point shattering phenotype of *SvLes1-WT* and *SvLes1-INS* in isogenic (ME034v) background.

To validate *SvLes1* as the causal gene, we used CRisPR-Cas9 to create a new allele. We assumed that *SvLes1-WT* was more likely to be the functional allele while *SvLes1-mutant* was only partially functional, because *SvLes1-WT* has an intact conserved MYB domain; additionally high shattering is likely preferred in natural populations for seed dispersal. We hypothesized that disruption of the *SvLes1-WT* allele would lead to reduced shattering. Therefore, we disrupted the wild type allele in a high-shattering accession (ME034v, corresponding to accession TB0147) to create a novel allele, *SvLes1-INS* (insertion), which has an “A” inserted at transcript position 149 (Figure 4c). This insertion led to a frameshift mutation predicted to completely abolish gene function.

After segregating out the transgenes encoding Cas9 and guide RNAs, homozygotes of this allele were phenotypically examined in the T3 generation.

To quantify seed shattering, we measured tensile strength of the abscission zone (AZ)[8,30] in homozygous lines of the three alleles during inflorescence development. Among the three alleles, *SvLes1-WT* and *SvLes1-INS* were in a completely isogenic ME34v background, with this SNP being the only genetic difference between the two lines. *SvLes1-mutant* is in an A10.1 (reduced shattering) background. We clearly observed high tensile strength (reduced seed shattering) in *SvLes1-INS* compared to the control *SvLes1-WT*, while A10.1 with the *SvLes1-mutant* allele showed moderate shattering compared to the other two lines (Figure 4d,e,f).

Like other known shattering genes in the grass family (*Sh4* [30], *qSh1* [31], *Shat1* [32] in rice, and *Sh1* [33] in sorghum), *SvLes1* is a transcription factor. It is a member of the very large MYB family, a set of genes also implicated in shattering by the comparative study of Yu et al. [12]. The precise mechanism of shattering is not known in any cereal, and recent data suggest that each species may have a unique mechanism [12]. Unlike other grasses, the AZ in *Setaria viridis* is not histologically distinct, with only subtle differences between wild type and domesticated plants [7,8]. As expected, we found that the anatomy and histology of *SvLes1-INS* and *SvLes1-WT* spikelets were indistinguishable (Figure S4).

Recent transposable element insertion in *SvLes1* contributed to domestication of foxtail millet

Reduced seed shattering is an early step in domestication of almost all cereals. To examine if *SvLes1* played a role in foxtail millet domestication, we compared the *SvLes1* locus in the genomes of *S. viridis* and *S. italica* (Yugu1). We discovered a ~6.5kb *copia* transposable element (*copia38*) inserted between the two Myb domains of *SiLes1* in Yugu1. We call this the *SiLes1-TE* (transposable element) allele. The structure strongly suggests *SiLes1-TE* is also a loss of function allele similar to *SvLes1-INS*, which should also produce a low shattering phenotype, thus potentially contributing to the domestication of foxtail millet. The *copia38* transposable element was aligned to each of the pan-genome assemblies (n=605) to investigate whether other diversity lines showed insertion into the *SvLes1* locus. Only two samples aligned to the *copia38* TE within the CDS sequence of *SvLes1*, but the nucleotide identity and coverage of the alignments were poor (32% and 6% respectively). In contrast, *Copia38* is almost fixed in foxtail millet lines (78 out of 79).

The low-shattering QTL co-localizes with a selective sweep. Genome-wide, *S. italica* has about 22% of the SNP variation of *S viridis*, based on coalescent simulations (using sequences generated in [34]), which is expected given its domestication history. In the *SiLes1* region, however, this number is significantly reduced to 4.1-8.2% of the diversity in *S. viridis* depending on the size of the region compared (10-100 kb, Table S7)(p=0.0066 based on 100,000 coalescent simulations). Diversity is relatively high within the gene itself, which could be caused by the comparatively small size of the gene (ca. 2 kb) plus its high level of conservation, or may mean that selection is on a regulatory region or additional locus under the QTL. Estimates of linkage disequilibrium (LD) support this interpretation (Table S8). Across the same intervals of up to 100 kb

surrounding the *SvLes1* region, LD is generally similar and relatively low in *S. viridis*. In *S. italica* on the other hand, LD is nearly complete for the 20 kb region surrounding *SiLes1*. When that region is expanded to 40 kb LD drops to levels approximating that in *S*. *viridis*. Thus selection appears to be keeping a block that includes *SiLes1* and *copia38* in LD.

On the other hand, *SiLes1* is probably not the only gene to account for low shattering in modern elite foxtail millet lines. The *SiLes1* region does not colocalize with the selective sweeps in foxtail millet elite lines compared to landraces, whereas the ortholog of sorghum shattering gene *Sh1* does [34]. In addition, using A10.1 as a reference, tensile strengths are higher in two elite foxtail millet lines, Yugu1 and B100, than *SvLes1-INS* homozygotes when comparing our results to a previous study [8], also indicating mechanisms in addition to *SiLes1-TE* are involved in determining low seed shattering in foxtail millet.

We have dated the *copia38* insertion in *SiLes1*. *Copia38* is a long terminal repeat (LTR) retroelement with two 451 bp LTR sequences. The two LTRs have identical sequences, suggesting recent insertion. Phylogenetic analysis showed *Copia38* tightly clusters with a few homologous copies on a long branch, indicating a shared burst event for copies this cluster (Figure S5). Pairwise distance among the copies suggests this burst was recent, on average about 45k (minimum and maximum 23k – 81k) years ago, assuming a neutral mutation rate of 6.5×10^{-9}/bp/year [35]. More interestingly, the copies from this burst only occur in Yugu1 but not in A10.1, suggesting a specific expansion in the line leading to the domesticate.

Combining the results and evidence above, a plausible scenario is that the *SiLes1-TE* allele arose about 50k years ago from the LTR retroelement insertion. This rare event/low frequency allele was selected during domestication due to its favored low shattering phenotype, and nearly fixed in foxtail millets. Later, the low shattering phenotype was further strengthened by additional loci with stronger effects (i.e. *SvSh1* [34,36]) during more recent crop improvement. Genome editing technology today allows us to recreate a low-shattering phenotype from ancestral *S. viridis* alleles, mimicking the initial phase of foxtail millet domestication through a different novel allele *SvLes1-INS.*

Parallel genetic control of leaf angle in *S. viridis* and corn by *liguleless2* orthologs

In addition to being a powerful tool for GWAS, the diversity panel also is useful for uncovering the basis of rare traits. We discovered a single accession in the panel (TB0159) with reduced auricle development and marked upright leaves (small leaf angle) (Figure 4a,b). As a GWAS approach is not suitable for mapping traits with low frequency and strong effects, bulked segregant analysis (BSA)[37,38] was used to identify the underlying causal gene. TB0159 was crossed to A10.1, and the F1 plants showed wild type leaf angle, showing that small leaf angle is a recessive allele. The WT and small leaf angle trait in the F_2 population segregated at 264:153 which differs significantly from a 3:1 ratio (p=0.000238). However, segregation distortion is known in *Setaria* crosses [39], so we proceeded assuming a single recessive causal gene for small leaf angle.

With BSA, we mapped the reduced leaf angle phenotype to a homozygous region of ~800 kb on Chr_05 (Figure 4c). 104 disruptive SNPs and 687 indels (393

single bp) were isolated from this QTL region. Because the small leaf angle phenotype is recessive and unique to TB0159 in the panel, the causal allele should be homozygous and occur only in TB0159 as well. This gives us the resolution to a single primary candidate mutation at Chr_5:41489494, a single G insertion inside the coding region of gene *SvLiguless2* (*SvLg2*) (Sevir.5G394700) that is predicted to cause a frameshift. Importantly, *SvLg2* is the syntenic ortholog to *liguless2* in maize (Figure 4d), mutations in which are known to reduce auricle development and leaf angle. Accordingly, we conclude this indel mutation is the mostly likely causal mutation for the observed small leaf angle in TB0159. This result gives one example of parallel phenotype-genotype relationships between *S. viridis* and maize, and provides a tool for deep exploration of *lg2* function in a tractable genetic system.

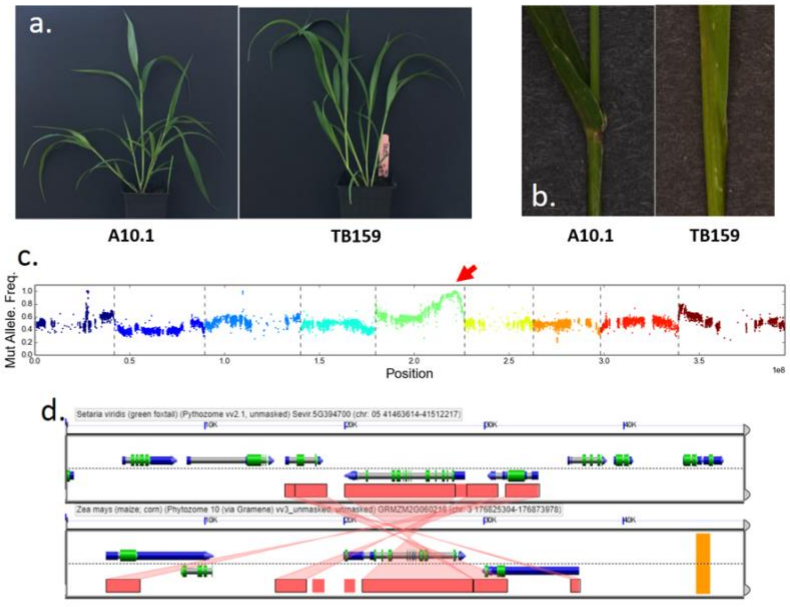

Fig. 4. Phenotype and mapping of small leaf angle. **a, b** Small leaf angle phenotype in TB159. **c** BSA mapping result, red arrow indicating QTL. **d** Synteny analysis around *SvLg2* and maize *lg2* locus.

Data presented here demonstrate the power of the *S. viridis* genome and associated diversity lines for gene discovery. Together these tools permit discovery of novel

genes, such as *SvLes1*, and additional insight into genes such as *SvLg2* that have been identified from other related systems. We show that *S. viridis* resources can uncover the genetic basis of a trait when the genes are unknown and likely differ from those in other related species. We also show that *S. viridis* can be a tool for identifying and investigating the function of genes in which an orthologue is known in a crop. Specifically, we have used the *S. viridis* genome and diversity panel to study the genetic basis of seed shattering and leaf angle (*SvLes1* and *SvLg2*, respectively), traits that influence crop yield by affecting harvestability, productivity and density of planting.

Methods

Plant materials

The reference line A10.1 is a descendant of the line used by Wang et al. [40] used in early RFLP maps. The original line was found to be heterozygous and thus A10.1 was propagated via single-seed descent by Andrew Doust (Oklahoma State University, Stillwater, OK, pers. comm.). It is thought to have originated in Canada. The other reference, ME034 (also known as ME034v), was collected by Matt Estep (Appalachian State University, Boone, NC) in southern Canada as part of a diversity panel [24] included among the diversity lines sequenced here. Transformation is more efficient for ME034 than for A10.1 (Joyce van Eck, Boyce Thompson Institute, Ithaca, NY, pers. comm.) and thus the former is being used widely for functional genetic studies.

The 605 individuals of the diversity panel were collected over a period of several years. About 200 lines have been described in previous studies [22,24] whereas others

were new collections added for this project. Individuals were propagated by single seed descent, although the number of generations varies by accession. *S. viridis* is inbreeding by nature (ca. 1% [22]), so we assume that initial heterozygosity was generally low and then further reduced in propagation.

Library creation and sequencing

To prepare DNA for sequencing of the reference line, 100 ng of DNA was sheared to 500 bp using the Covaris LE220 (Covaris) and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). The prepared library was then quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then multiplexed with other libraries, and the pool of libraries was prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2000 sequencer using a TruSeq SBS sequencing kit, v3, following a 2x150 indexed run recipe.

Plate-based DNA library preparation for Illumina sequencing was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Kapa Biosystems library preparation kit. 200 ng of sample DNA was sheared to 600 bp using a Covaris LE220 focused-ultrasonicator. Sheared DNA fragments were size selected by double-SPRI and then the selected fragments were end-repaired, A-tailed, and ligated with

Illumina compatible sequencing adaptors from IDT containing a unique molecular index barcode for each sample library.

The prepared library was quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then then multiplexed with other libraries, and the pool of libraries was then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3 or v4, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2000 or HiSeq2500 sequencer using HiSeq TruSeq SBS sequencing kits, v3 or v4, following a 2x150 indexed run recipe.

Sequencing of reference genome

We sequenced *Setaria viridis* A10.1 using a whole genome shotgun sequencing strategy and standard sequencing protocols. Sequencing reads were collected using Illumina HISeq and PACBIO SEQUEL platforms at the Department of Energy (DOE) Joint Genome Institute (JGI) in Walnut Creek, California and the HudsonAlpha Institute in Huntsville, Alabama. One 800 bp insert 2x250 Illumina fragment library (240x) was sequenced, giving 425,635,116 (Table S9). Illumina reads were screened for mitochondria, chloroplast, and PhiX contamination. Reads composed of >95% simple sequence were removed. Illumina reads <75bp after trimming for adapter and quality (q<20) were removed. For the PACBIO sequencing, a total of 36 P5C2 chips (4 hour movie time) and 41 P6C4 chips (10 hour movie time) were sequenced with a p-read yield of 59.09 Gb, with a total coverage of 118.18x (Table S10).

Genome assembly and construction of pseudomolecule chromosomes

An improved version 2.1 assembly was generated by assembling 4,768,857 PACBIO reads (118.18x sequence coverage) with the MECAT assembler [41] and subsequently polished using QUIVER [42]. The 425,635,116 Illumina sequence reads (240x sequence coverage) was used for correcting homozygous snp/indel errors in the consensus. This produced 110 scaffolds (110 contigs), with a contig N50 of 16.8 Mb, and a total genome size of 397.9 Mb (Table S11). A set of 36,061 syntenic markers derived from the version 2.2 *Setaria italica* release was aligned to the MECAT assembly. Misjoins were characterized as a discontinuity in the *italica* linkage group. A total of 15 breaks were identified and made. The *viridis* scaffolds were then oriented, ordered, and joined together into 9 chromosomes using syntenic markers. A total of 61 joins were made during this process. Each chromosome join is padded with 10,000 Ns. Significant telomeric sequence was identified using the TTTAGGG repeat, and care was taken to make sure that it was properly oriented in the production assembly.

Scaffolds that were not anchored in a chromosome were classified into bins depending on sequence content. Contamination was identified using blastn against the NCBI nucleotide collection (NR/NT) and blastx using a set of known microbial proteins. Additional scaffolds were classified as repetitive (>95% masked with 24mers that occur more than 4 times in the genome) (26 scaffolds, 1.2 Mb), alternative haplotypes (unanchored sequence with >95% identity and >95% coverage within a chromosome) (15 scaffolds, 1.0 Mb), chloroplast (3 scaffolds, 164.5 Kb), mitochondria (5 scaffolds,

344.9 Kb), and low quality (>50% unpolished bases post polishing, 1 scaffolds, 19.3 Kb). Resulting final statistics are shown in Table S8.

Finally, homozygous SNPs and INDELs were corrected in the release consensus sequence using ~60x of Illumina reads (2x250, 800 bp insert) by aligning the reads using bwa mem [43] and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool [44]. A total of 96 homozygous SNPs and 4,606 homozygous INDELs were corrected in the release. The final version 2.1 release contains 395.1 Mb of sequence, consisting of 75 contigs with a contig N50 of 11.2 Mb and a total of 99.95% of assembled bases in chromosomes (Table S12).

To assess completeness of the euchromatic portion of the version 2.1 assembly, a set of 40,603 annotated genes from the *S. italica* release was used for comparison. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The transcripts were aligned to the assembly using BLAT [15] and alignments ≥90% base pair identity and ≥85% coverage were retained. The screened alignments indicate that 39,441 (97.14%) of the *italica* genes aligned to the version 2.1 release. Of the unaligned 1,162 transcripts, 928 (2.28%) indicated a partial alignment, and 234 (0.58%) were not found in the version 2.1 release.

To assess the accuracy of the assembly, a set of 335 contiguous Illumina clones >20 Kb was selected. A range of variants was detected in the comparison of the clones and

the assembly. In 239 of the clones, the alignments were of high quality (< 0.01% bp error) with an example being given in Fig. S6 (all dot plots were generated using Gepard [45]). The remaining 96 clones indicate a higher error rate due mainly to their placement in more repetitive regions (Fig. S11). The major component of the error in the 96 repetitive clones was copy number variation, which affected 50 of the clones. These 50 clones accounted for >97% of all of the errors in the 335-clone set. Excluding the clones with copy number variation, the overall bp error rate in the 285 clone set is 0.0098% (1,043 discrepant bp out of 10,601,785).

SNP calling

598 diversity samples (Table S1) were used for diversity analysis. The samples were sequenced using Illumina paired end sequencing (2*151bp) at the Department of Energy Joint Genome Institute (JGI, Walnut Creek, CA, USA) and the HudsonAlpha Institute for Biotechnology (Huntsville AL, USA) using Hiseq 2500 and NovoSeq6000. The reads were mapped to *S. viridis* v2.1 using bwa-mem [46]. The bam was filtered for duplicates using picard (http://broadinstitute.github.io/picard), and realigned around indels using GATK [44]. Multi sample SNP calling was done using SAMtools mpileup [47] and Varscan V2.4.0 [48] with a minimum coverage of 8 and a minimum alternate allele frequency of four. An allele is confirmed to be homozygous or heterozygous using a binomial test significance at a p-value of 0.05. Repeat content of the genome was masked using 24bp kmers. Kmers that occur at a high frequency, up to 5%, were masked. SNP around 25bp of the mask were removed for further analysis. If an SNP has a coverage in 90% of the samples, and a MAF > 0.01, then only it was included for

further analysis. Imputation and phasing was done in Beagle V4.0 (Browning and Browning 2007). SNP Annotation was performed using snpEff [49].

Population structure

Population structure for both SNP and PAV data was estimated using fastStructure [20] and Admixture [21]. SNP markers were randomly subsetted to 50k by LD pruning (parameters: --indep-pairwise 50 50 0.5) in plink 1.9 [50], while shell genes (as determined by DAPC clustering) were extracted from the pan-genome. In both analyses, a single sample with a maximum membership coefficient (qi) of <0.7 was considered admixed. Only non-admixed samples were used for further analysis. For SNP markers, multidimensional scaling (MDS), identity by state (IBS), LD estimates (parameters: --r2 --ld-window-kb 500 --ld-window-r2 0) was performed in plink 1.9.

Search for *copia* elements

The *copia38* sequence (6.7 kb) was extracted from the genomic sequence of Seita.5G087200 using repbase (https://www.girinst.org/repbase/). Both the *copia38* sequence and *SvLes1* (Sevir.5G085400) sequence (both genomic sequence and CDS) were aligned to each of the pan-genome assemblies (n=605) using blat (-noHead -extendThroughN). From the blat results, each *copia38* alignment was checked whether it fell within the bounds of the *SvLes1* locus.

NJ tree construction

382 non-admixed samples along with *S. adhaerens* as outgroup were used for phylogenetics. Plink 1.9 was used for LD thinning of markers and a random set of 25000 SNP were converted to FASTA format. ClustalW [52] was used for building a NJ tree with 100 bootstraps.

Genomic signatures of local adaptation

Abrupt changes in allele frequency occur whenever a population establishes into a geographical niche. To detect genes that are under positive selection during adaptation within a subpopulation, we used four different population genetic estimates:

a) Ka/Ks [26], a gene-based method estimated using snpEff. A ratio greater than one implies positive selection.

b) Tajima's D [25], a frequency based method using custom python scripts. Genes with < 5% of TajimaD distribution are considered positively selected for adaptation.

c) Integrated haplotype score (iHS [27]), a LD based method estimated using Hapbin [53]. If a standardized absolute iHS value is > 95% of the distribution and is within the gene, it is considered selected.

d) Composite Likelihood Ratio (CLR [28]), based on site frequency spectra (SFS) was estimated using SweeD [54] for every 1Kbp. If the likelihood value is above > 95% of the distribution and is within a gene, it is considered selected. A gene is said to be positively selected for adaptation if it is significant in two or more of these estimates.

GO and KEGG pathway enrichment analysis

GO enrichment analysis of positively selected genes was performed using topGO [55,56], an R Bioconductor package, to determine overrepresented GO categories across biological process (BP), cellular component (CC) and molecular function (MF) domains. Enrichment of GO terms was tested using Fisher's exact test with $p<0.05$ considered as significant. KEGG [57] pathway enrichment analysis was also performed on those gene sets based on hypergeometric distribution test and pathways with $p<0.05$ were considered as enriched.

GWAS and validation of *SvLes1*

The GWAS population to assess seed shattering was planted in the greenhouse facility at Donald Danforth Plant Science Center in April 2014. About 200 accessions were chosen from the panel to perform the experiment (Table S13), with four replicates per accession. Shattering phenotype was measured by observing the amount of seed shattering after hand shaking of senesced dry plants. Individual plants were scored using a quantitative scale from 1 to 7. Genotypes were filtered at minor allele frequency > 5% for this population. GWAS was performed using a univariate mixed linear model from GEMMA [58], with centered kinship matrix. We used the Wald test p-value [58] for assessing significant peaks, but other p value estimates give similar results. SNP effects were identified using snpEff [49]. Deleterious effect of missense SNPs were predicted using PROVEAN [29] on both the reference and alternative allele against the NCBI nr protein database.

To knockout *SvLes1* we used the backbone pTRANS_250d as described by Cermák et al. [59]. The protospacer of the guide RNAs targeted the first and second

exons of the *SvLes1*, upstream to the predicted causal mutation to ensure knock out by frameshift (Figure S8). The binary vector was introduced into callus tissue using AGL1 agrobacterium. Tissue culture and transformation followed an established protocol for *S. viridis* [60]. T0 and T1 individuals were genotyped to identify newly acquired mutations near the targeted sites. A T2 homozygote *SvLes1-IN*S was obtained and confirmed by Sanger sequencing, together with homozygotes of the unedited reference line for comparison.

Tensile strength measurement

A10.1, *SvLes1-WT* (unedited control in the ME34v background) and *SvLes1-INS* seeds were treated with 5% liquid smoke overnight under room temperature and kept in wet moss at 4°C in the dark for 2-3 weeks. Seeds were sown in Metro mix 360 and grown in a greenhouse with a 14h light/10 h dark cycle, day/night temperatures of 28 and 22°C and relative humidity of 40-50%. Panicles from main stems were collected at 8, 10, 12, 14, 16, 18, 20, 24 and 29 days after heading (the apex of the panicles emerged from the leaf sheath). Tensile strength of the spikelet and pedicel junction was measured as described previously [8]. Briefly, panicles were hung upside down from a Mark-10 model M3-2 force gauge. Spikelets were pulled off individually from a panicle using forceps and the peak tension was recorded. Only the most developed spikelets from the central third of the panicle were used to minimize the effects of developmental variation of the spikelets. Six plants with 20 spikelets from each plant were used per genotype per day of measurement. For *SvLes1-WT* and *SvLes1-INS*, the plants in each genotype were offspring of two individual parent plants with the same allele.

Histology

Histological procedures followed [61]. Specifically, primary branches were collected from the central third of panicles 12 and 16 days after heading and fixed in FAA (37% formaldehyde: ethanol: H_2O: acetic acid = 10:50:35:5), followed by a dehydration series in 50%, 70%, 85%, 95%, 100%, 100% and 100% ethanol and 25%, 50%, 75%, 100%, 100% and 100% Histo-Clear (National Diagnostics) series with ethanol as solvent. Paraplast (Leica Biosystems) was then added to each vial of samples and kept overnight, heated at 42 °C, and placed in a 60 °C oven. The solution was replaced with molten Paraplast twice a day for 3 days. Samples were then embedded in paraffin using a Leica EG1150 tissue embedder, sectioned in 10 µm serial slices with a Leica RM2255 automated microtome, and mounted on microscope slides at 42 °C on a Premiere XH-2001 Slide Warmer. Sections were then deparaffinized, rehydrated, stained with 0.05% (w/v) toluidine blue O for 1.5 min, and then rinsed with water, dehydrated in ethanol, cleared with xylene and mounted with Permount Mounting Medium (Electron Microscopy Sciences) [16]. Images were taken using a Leica DM750 LED Biological microscope with ICC50 camera module and Leica Acquire v2.0 software.

Domestication selective sweep

Raw sequencing reads of foxtail millet lines were obtained from a previous study [34]. Because the average sequencing coverage in the earlier study (~0.5×) is much lower compared to our study, we chose 79 lines (Table S14) that have an estimated coverage of greater than 1× to maximize overlapping SNPs and perform analysis. Briefly, *S.*

italica sequences were quality trimmed using sickle [51], and aligned with bwa-mem to our *S. viridis* A10.1 genome. Multi-sample SNP calling was performed using samtools and Varscan with a minimum depth of 3. For *S.viridis*, the imputed, phased vcf was used for calculation of π, which uses high coverage. π calculation excluded missing samples. Shared SNPs between foxtail millet and *S. viridis* were combined, and missing data were imputed using Beagle 5.0 [62]. Nucleotide diversity values $\pi_{viridis}$ and $\pi_{italica}$ were then calculated using vcftools [63] at 100 kb window size. Using genome-wide nucleotide diversity as a reference, we used the program ms [64] to conduct 100,000 coalescent simulations to estimate the variation range of $\pi_{italica}/\pi_{viridis}$ under a domestication bottleneck model for a window of 20 kb. Strength of the bottleneck was determined by genome wide $\pi_{italica}/\pi_{viridis}$. The estimated ranges were then compared to observed values $\pi_{italica}/\pi_{viridis}$ to determine significance of domestication selective sweep regions.

Retrotransposon insertion in SvLes1

Copia38 sequence was obtained from the foxtail millet genomic sequence [39] near the ortholog of *SvLes1*, Seita.5G087200 (Si003873m.g). We confirmed the identity of *Copia38* and identified its LTR region by searching its sequence against repbase (https://www.girinst.org/repbase/). We used blastn [65] to identify close homologs of *Copia38* in Yugu1 [39] and A10.1 genomes. RaxML 8.2.9 (Stamatakis 2014) was used to construct phylogeny of *Copia38* homologs, and pairwise distances of close homologs to *Copia38* were calculated using Kimura 2 parameter model. Read mapping to Yugu1 genome follows similar procedure described previously. Paired end reads spanning

beyond the left and right junction point of *Copia38* were used to determine if the insertion occur in one line/accession or not (Figure S9).

BSA mapping for small leaf angle

The cross between TB159 and A10.1 used pollen of TB159 and follows the protocol described in [66]. F1 individuals were naturally self-pollinated to generate an F2 population. 30 F2 individuals were planted and phenotypically scored, and DNA from 30 small leaf angle individuals were pooled and sequenced. Sequences are available in the SRA at NCBI, BioProject number PRJNA527194 (to be released after publication). The analysis follows the methods described in a previous BSA study in *S. viridis* [37]. Identification of disruptive mutations and missense mutations with deleterious effects follows the same approach described in our GWAS study. Syntenic orthology between *SvLg2* and *liguleless2* in maize was examined and confirmed based on [67].

Acknowledgements

We thank Zhonghui Wang, Xiaoping Li, and Hui Jiang for their help in maintaining the diversity panel and data collection. This work was supported by NSF grants DEB-0115397, MCB-0110809, DEB-0108501, PGRP-0952185, and IOS-1557633 to EAK; and DE-SC0008769 to TPB. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions: Credit taxonomy: Conceptualization: PH, IB, MF, JS, TPB EAK,; Methodology: PH, JJ, DR, JS; Investigation: PH, SM, AH, JG, JJ, KB, AS, SS, MF, JW, YY, CC, JJ, HS, TK, TS; Resources: PH, MF, EAK; Writing - Original Draft: PH, KB, JJ, SM, EAK; Writing - Review & Editing: PH, SM, MF, IB, JS, TPB, EAK; Visualization: PH, JJ, EAK; Supervision: IB, TPB, JS, EAK; Project Administration: JS, DR, TPB, EAK; Funding Acquisition: JS, DR, TPB, EAK. Detailed contributions: JS: WGS assembly & sequencing project lead; JJ: map integration, chromosome assembly, analysis; JG: sequencing of BES, QC projects; HS, TK, TS: PacBio sequencing; SS: annotation; DR: JGI Eukaryotic Plant Program Leader; PH, SM, AH, MF: population genetics; PH and SM: GWAS, BSA; phenotyping; YY: histology and plant development.

Competing interests: The authors confirm that they have no competing interests.

Literature Cited

1 Grass Phylogeny Working Group. Phylogeny and subfamilial classification of the Poaceae. *Ann Missouri Bot Gard* **88**, 373-457 (2001).

2 Kellogg, E. A. *Poaceae*. In *Families and Genera of Vascular Plants*, K. Kubitzki, ed. Pp. 1-416 (Springer, 2015).

3 Soreng, R. J. *et al.* A worldwide phylogenetic classification of the Poaceae (Gramineae) II: An update and a coparison of two 2015 classifications. *J Syst Evol* **55**, 259-290, doi:10.1111/jse.12262 (2017).

4 Brutnell, T. P. *et al. Setaria viridis*: a model for C_4 photosynthesis. *Plant Cell* **22**, 2537-2544 (2010).

5 Doust, A. N. & Diao, X. The genetics and genomics of Setaria. Volume 19 in *Plant Genetics and Genomics: Crops and Models* (Springer International Publishing, Switzerland, 2017).

6 Doust, A. N., Kellogg, E. A., Devos, K. M. & Bennetzen, J. L. Foxtail millet, a sequence driven grass model system. *Plant Physiol* **149**, 137-141 (2009).

7 Yu, Y. & Kellogg, E. A. Inflorescence abscission zones in grasses: diversity and genetic regulation. *Ann Plant Rev* **1**, 1-35, doi:10.1002/9781119312994.apr0619 (2018).

8 Hodge, J. G. & Kellogg, E. A. Abscission zone development in *Setaria viridis* and its domesticated relative, *Setaria italica*. *Am J Bot* **103**, 998-1005, doi:10.3732/ajb.1500499 (2016).

9 Li, L. F. & Olsen, K. M. To have and to hold: selection for seed and fruit retention during crop domestication. *Curr Topics Devell Biol* **119**, 63-109 (2016).

10 Doust, A. N. *et al.* Beyond the single gene: How epistasis and gene-by-environment effects influence crop domestication. *Proc Natl Acad Sci USA* **111**, 6178-6183 (2014).

11 Doust, A. N., Mauro-Herrera, M., Francis, A. & Shand, L. Morphological diversity and genetic regulation of seed dispersal in grasses. *Amer J Bot* **101**, 1759-1769 (2014).

12 Yu, Y., Hu, H., Doust, A. N. & Kellogg, E. A. Divergent gene expression networks underlie morphological diversity of abscission zones in grasses. *New Phytol* (2019).

13 Chapman, J. A. Meraculous2: rast accurate short-read assembly of large polymorphic genomes. *arXiv*, 1608.01031 (2016).

14 Sonnhammer, E. L. & Ostlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* **43**, D234-239, doi:10.1093/nar/gku1203 (2015).

15 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202 (2002).

16 Kaas, R. S., Friis, C., Ussery, D. W. & Aarestrup, F. M. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* **13**, 577, doi:10.1186/1471-2164-13-577 (2012).

17 Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol* **79**, 7696-7701, doi:10.1128/AEM.02411-13 (2013).

18 Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* **36**, 6688-6719, doi:10.1093/nar/gkn668 (2008).

19 Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071, doi:10.1093/bioinformatics/btr521 (2011).

20 Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573-589, doi:10.1534/genetics.114.164350 (2014).

21 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).

22 Huang, P. *et al.* Population genetics of *Setaria viridis*, a new model system. *Mol Ecol* **23**, 4192-4295 (2014).

23 Kellogg, E. A., Aliscioni, S. S., Morrone, O., Pensiero, J. & Zuloaga, F. O. A phylogeny of *Setaria* (Poaceae, Panicoideae, Paniceae) and related genera, based on the chloroplast gene *ndhF*. *Intl J Plant Sci* **170**, 117-131 (2009).

24 Layton, D. J. & Kellogg, E. A. Morphological, phylogenetic, and ecological diversity of the new model species *Setaria viridis* (Poaceae: Paniceae) and its close relatives. *Amer J Bot* **101**, 539-557 (2014).

25 Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

26 Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496-503 (2000).

27 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72, doi:10.1371/journal.pbio.0040072 (2006).

28 Nielsen, R. & Wiuf, C. Composite likelihood estimation applied to Single Nucleotide Polymorphism (SNP) data. *ISI Conf Proc* (2005).

29 Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745-2747, doi:10.1093/bioinformatics/btv195 (2015).

30 Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936-1939 (2006).

31 Konishi, S. *et al.* An SNP caused loss of seed shattering during rice domestication. *Science* **312**, 1392-1396 (2006).

32 Zhou, Y. *et al.* Genetic control of seed shattering in rice by the APETALA2 transcription factor *SHATTERING ABORTION1*. *Plant Cell* **24**, 1034-1048 (2012).

33 Lin, Z. *et al.* Parallel domestication of the *Shattering1* genes in cereals. *Nat Genet* **44**, 720-724 (2012).

34 Jia, G. *et al.* A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat Genet* **45**, 957-961, doi:10.1038/ng.2673 (2013).

35 Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci U S A* **93**, 10274-10279 (1996).

36 Odonkor, S. *et al.* QTL mapping combined with comparative analyses identified candidate genes for reduced shattering in *Setaria italica*. *Front Plant Sci* **9**, 918, doi:10.3389/fpls.2018.00918 (2018).

37 Huang, P. *et al. Sparse panicle1* is required for inflorescence development in *Setaria viridis* and maize. *Nat Plants* **3**, 17054, doi:10.1038/nplants.2017.54 (2017).

38 Yang, J. *et al.* Brassinosteroids modulate meristem fate and differentiation of unique inflorescence morphology in *Setaria viridis*. *Plant Cell* **30**, 48-66, doi:10.1105/tpc.17.00816 (2018).

39 Bennetzen, J. L. *et al.* Reference genome sequence of the model plant *Setaria*. *Nat Biotech* **30**, 555-561 (2012).

40 Wang, Z. M., Devos, K. M., Liu, C. J., Wang, R. Q. & Gale, M. D. Construction of RFLP-based maps of foxtail millet, *Setaria italica* (L.) P. Beauv. *Theor Appl Genet* **96**, 31-36 (1998).

41 Xiao, C. L. *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**, 1072-1074, doi:10.1038/nmeth.4432 (2017).

42 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

43 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997v1 [q-bio.GN] (2013).

44 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

45 Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026-1028, doi:10.1093/bioinformatics/btm039 (2007).

46 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

47 Li, H. *et al.* The Sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

48 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

49 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).

50 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

51 A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) Available at https://github.com/najoshi/sickle. (2011).

52 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

53 Maclean, C. A., Chue Hong, N. P. & Prendergast, J. G. hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Mol Biol Evol* **32**, 3027-3029, doi:10.1093/molbev/msv172 (2015).

54 Pavlidis, P., Zivkovic, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* **30**, 2224-2234, doi:10.1093/molbev/mst112 (2013).

55 Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600-1607, doi:10.1093/bioinformatics/btl140 (2006).

56 topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0 v. 2.24.0 (http://bioconductor.org/packages/release/bioc/html/topGO.html, 2016).

57 Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27-30, doi:10.1093/nar/28.1.27 (2000).

58 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-824, doi:10.1038/ng.2310 (2012).

59 Cermák, T. *et al.* A multipurpose toolkit to enable advanced genome engineering in plants. *Plant Cell* **29**, 1196-1217, doi:10.1105/tpc.16.00922 (2017).

60 Van Eck, J., Swartwood, K., Pidgeon, K. & Maxon-Stein, K. *Agrobacterium tumefaciens*-mediated transformation of *Setaria viridis* in *Genetics and genomics of Setaria Plant Genetics and Genomics: Crops and Models* (eds A. Doust & X. Diao) 343-356 (Springer, 2016).

61 Ruzin, S. E. *Plant microtechnique and microscopy*. (Oxford University Press, 1999).

62 Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am J Hum Genet* **98**, 116-126, doi:10.1016/j.ajhg.2015.11.020 (2016).

63 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

64 Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).

65 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

66 Jiang, H., Barbier, H. & Brutnell, T. Methods for performing crosses in *Setaria viridis*, a new model system for the grasses. *J Vis Exp*, doi:10.3791/50527 (2013).

67 Schnable, J. C., Freeling, M. & Lyons, E. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol* **4**, 265-277 (2012).

Figure and Table legends

Fig. 1 a *Setaria viridis*, in its common highly disturbed habitat next to a road. **b-e** Assembly and gene presence/absence variation statistics for the 605 accessions of the *S. viridis* diversity panel. **b** Average library coverage. **c** Contig N50 (Kb). **d** Meraculous assembled bases. **e** Number of genes per library; red vertical line represents the number of genes necessary for a library to be included for PAV analysis (n=36,500).

Fig. 2 Population differentiation of *Setaria viridis* in North America. **a, b, d** data from SNPs. **c** data from PAV. Subpopulations color coded as: Central-north, purple, Central, red, Central-east, blue, West coast, green. **a** STRUCTURE analysis with K=4; Central (red) population is extensively admixed. **b** multidimensional scaling of the same data as in **a**, showing distinctiveness of the Central-north (purple) population. **c** distribution of subpopulations with admixed (Q<0.7) individuals removed; population assignment and distribution are highly similar to that shown by SNP data (**Fig. S2**). **d** Neighbor-joining tree of SNP data, rooted with *S. adhaerens* (outgroup).

Fig. 3 Phenotyping and GWAS of reduced shattering. **a** Manhattan plot of GWAS result, red line showing p=0.01 after correction. **b** Zoom in to peak on Chromosome 5; larger dots represent missense SNPs identified by snpEff. Different colors of missense SNPs indicate provean score range (blue for >-2.5, green for <-2.5 and >-4.1, red for <-4.1, -2.5 and -4.1 represents 80% and 90% specificity). Lower scores indicate higher likelihood of deleterious effects of the mutation. **c** Allele obtained, background,

shattering phenotype and allele effect on protein product. **d** Tensile strength measurements across panicle developmental stages with three alleles. **e,f** End point shattering phenotype of *SvLes1-WT* and *SvLes1-INS* in isogenic (ME034v) background.

Fig. 4 Phenotype and mapping of small leaf angle. **a, b** Small leaf angle phenotype in TB159. **c** BSA mapping result, red arrow indicating QTL. **d** Synteny analysis around *SvLg2* and maize *lg2* locus.

Fig S1 Plots of marker placements for each of the 9 chromosomes (scaffolds) of *S. viridis*.

Fig S2 Geographic distribution of subpopulations based on SNP variation. **a** Map Including all samples. **b** Map with admixed samples removed (qi > 0.7). Subpopulations indicated by colors as in **Fig 2**.

Fig S3 Over- and under-represented GO categories of genes for each subpopulation.

Fig S4 *SvLes1-mutant* (A10.1), *SvLes1-WT*, and *SvLes1-INS* have similar anatomical structures in the abscission zone. Spikelets from main panicles 12 days after heading, stained with 0.05% Toluidine blue O. **a,b** *SvLes1-mutant* (A10.1). **c,d** *SvLes1-WT.* **e,f** *SvLes1-INS.* **a,c,e** Scale bars = 100 µm. b,d,f Scale bars **=** 20 µm. En, endosperm; em, embryo. White dotted line, approximate position of AZ.

Fig S5 Phylogenetic tree of *Copia38* copies in A10.1 and Yugu1 geneomes. Red clade showing the recent explosion that contains the copy inserted into *SiLES1*. Gene models in green, *S. italica*; in blue, *S. viridis*.

Fig S6 Representative dot plots**. a** Dot plot of clone 114809 on a region of scaffold_1. This alignment is representative of the high quality clone alignments in 239 of the 365 available clones. **b** Dot plot of clone 113526 on a region of scaffold_3, which is representative of the 36 clones that landed in repetitive regions of the genome.

Fig S7 a gRNA protospacers and their position relative to allelic mutations and gene model. **b** Method to detect *Copia38* insertion.

Table S1 Metadata for 598 diversity lines. LIB, library; Q4 max, is the membership coefficient obtained from the output of STRUCTURE; Q4_subpop, subpopulation assignment using SNP data; Name, collector's name and number or other unique identifier; new name, name in germplasm collection, lat, latitude of original collection; long, longitude of original collection. Subpopulation assignments correspond to those in Figure and S1, viz. red - admixed - 95 samples; green - West coast - 222 samples; magenta - Central north - 148 samples; blue - South Central - 133 samples.

Table S2 PAV matrix.

Table S3 Gene observations per *S.viridis* subpopulation. 3,355 genes significantly ($p<0.05$) over- or under-represented for each of the three largest populations. Gene counts were tested for significant differences between expected and observed counts using a Pearson's chi squared test (df=2). Negative chi square values represent genes being observed less often than expected while positive chi square values represent genes observed more often than expected. Chi square values were converted to p values, then adjusted using a Benjamini-Hochberg adjustment to correct for multiple testing.

Table S4 SNP diversity estimates of subpopulations, showing high diversity in the Central-north population. Pi/bp, SNPs per base pair; private SNP, SNPs restricted to one population; IBS (average), identity by state for the population; LD >=0.7, percent of pairwise comparisons with LD>=0.7.

Table S5 Numbers of genes in each population found to be under selection using Tajima's D, Ka/Ks ratio, iHS statistic, and CLR.

Table S6 Genes with significant evidence for selection using four selection tests, by population. Populations correspond to those shown in Figure S2.

Table S7 Comparison of pairwise diversity in the region surrounding *SiLes1* vs. *SvLes1*. Diversity is significantly lower in the *Les1* region in *S. italica* than in *S. viridis*, p=0.0066 based on 100,000 coalescent simulations. Description, interval considered for

diversity estimates; total, total number of bp, centered on the gene; either side, number of bp upstream and downstream of the gene.

Table S8 Comparison of linkage disequilibrium (LD) in the region surrounding *SiLes1* vs. *SvLes1*. Intervals as in Table S7.

Table S9 Genomic libraries included in the *Setaria viridis* genome assembly and their respective assembled sequence coverage levels in the final release.

*Average read length of PACBIO reads.

Table S10 PACBIO library statistics for the libraries included in the *Setaria viridis* genome assembly and their respective assembled sequence coverage levels.

Table S11 Summary statistics of the initial output of the QUIVER polished MECAT assembly. The table shows total contigs and total assembled base pairs for each set of scaffolds greater than the size listed in the left hand column.

Table S12 Final summary assembly statistics for chromosome scale assembly.

Table S13 *Setaria italica* lines used for π *italica* - π *viridis* comparison and for test of selective sweeps. Data from Jia et al. (2013), downloaded from the Sequence Read Archive at NCBI.

Table S14 *Setaria viridis* samples scored for shattering phenotype. LIB, lbirary number; Name, original accession identified; New_name, internal Brutnell lab identified; Score_lowHatter, shattering indiex.