UC Berkeley

UC Berkeley Previously Published Works

Title

De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes.

Permalink

https://escholarship.org/uc/item/3sf8f9gr

Journal

Science (New York, N.Y.), 373(6555)

ISSN

0036-8075

Authors

Hufford, Matthew B
Seetharam, Arun S
Woodhouse, Margaret R
et al.

Publication Date

2021-08-01

DOI

10.1126/science.abg5289

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

PLANT GENOMICS

De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes

Matthew B. Hufford[1], Arun S. Seetharam[1,2], Margaret R. Woodhouse[3], Kapeel M. Chougule[4], Shujun Ou[1], Jianing Liu[5], William A. Ricci[6], Tingting Guo[7], Andrew Olson[4], Yinjie Qiu[8], Rafael Della Coletta[8], Silas Tittes[9,10], Asher I. Hudson[9,10], Alexandre P. Marand[5], Sharon Wei[4], Zhenyuan Lu[4], Bo Wang[4], Marcela K. Tello-Ruiz[4], Rebecca D. Piri[11], Na Wang[6], Dong won Kim[6], Yibing Zeng[5], Christine H. O'Connor[8,12], Xianran Li[7], Amanda M. Gilbert[8], Erin Baggs[13], Ksenia V. Krasileva[13], John L. Portwood II[3], Ethalinda K. S. Cannon[3], Carson M. Andorf[3], Nancy Manchanda[1], Samantha J. Snodgrass[1], David E. Hufnagel[1,14], Qiuhan Jiang[1], Sarah Pedersen[1], Michael L. Syring[1], David A. Kudrna[15], Victor Llaca[16], Kevin Fengler[16], Robert J. Schmitz[5], Jeffrey Ross-Ibarra[9,10,17], Jianming Yu[7], Jonathan I. Gent[6], Candice N. Hirsch[8], Doreen Ware[18,4], R. Kelly Dawe[5,6,11]*

We report de novo genome assemblies, transcriptomes, annotations, and methylomes for the 26 inbreds that serve as the founders for the maize nested association mapping population. The number of pan-genes in these diverse genomes exceeds 103,000, with approximately a third found across all genotypes. The results demonstrate that the ancient tetraploid character of maize continues to degrade by fractionation to the present day. Excellent contiguity over repeat arrays and complete annotation of centromeres revealed additional variation in major cytological landmarks. We show that combining structural variation with single-nucleotide polymorphisms can improve the power of quantitative mapping studies. We also document variation at the level of DNA methylation and demonstrate that unmethylated regions are enriched for cis-regulatory elements that contribute to phenotypic variation.

Maize is the most widely planted crop in the world and an important model system for the study of gene function. The species is known for its extreme genetic diversity, which has allowed for broad adaptation throughout the tropics and intensive use in temperate regions. Nevertheless, most current genomic resources are referenced to a single inbred, B73, which contains only 63 to 74% of the genes and/or low-copy sequences in the full maize pan-genome (*1*–*4*). Moreover, there is extensive structural polymorphism in noncoding and regulatory genomic regions that has been shown to contribute to variation in numerous traits (*5*). In recent years, additional maize genomes have been assembled, which has allowed limited characterization of the species' pan-genome (*2*, *6*–*10*). However, comparisons across genome projects are often confounded by differences in assembly and annotation methods.

The maize nested association mapping (NAM) population was developed to study the genetic architecture of quantitative traits (*11*). Twenty-five founder inbred lines were strategically selected from a larger association panel (*12*) to represent the breadth of maize diversity, including lines from the non–stiff-stalk temperate heterotic group; lines from tropical and subtropical regions of Africa, Asia, and the Americas; and both sweet corn and popcorn germplasm (*13*). Each NAM parental inbred was crossed to B73 and selfed to generate 25 populations of 200 recombinant inbred lines that combine the advantages of linkage and association mapping for important agronomic traits (*14*). Biological infrastructure continues to be developed around these lines [e.g., (*15*, *16*)], but comprehensive genomic resources are needed to fully realize the power of the NAM population.

Consistency and quality of genome assemblies

Here, we describe assembled and annotated genomes for the 25 NAM founder inbreds and an improved reference assembly of B73 (table S1). The 26 genomes were sequenced to high depth (63–85×) by PacBio long-read technology, assembled into contigs by a hybrid approach (*17*), scaffolded by Bionano optical maps, and ordered into pseudomolecules by using linkage data from the NAM recombinant inbred lines and maize pan-genome anchor markers (*4*). Assembly and annotation statistics improve upon nearly all available maize assemblies, including the previous B73 reference genome (*18*), with the total length of placed scaffolds (2.102 to 2.162 billion base pairs) at the estimated genome size of maize, a mean scaffold N50 of 119.2 Mb [contig N50 of 25.7 million base pairs (Mbp)], complete gene space [mean of 96% complete benchmarking universal single-copy orthologs (BUSCO)] (*19*), and, on the basis of the LTR Assembly Index (mean of 28) (*20*), full assembly of the transposable element (TE)–laden portions of the genome (Table 1 and table S2). Improvements in contiguity and completeness can be attributed to recent advances in sequence and optical map data, as well as more-effective assembly algorithms (*21*).

Gene identification and diversity in gene content

We sequenced mRNA from 10 tissues for each inbred. These data were used for evidence-based gene annotation of each line, which was then improved by using B73 full-length cDNA and expressed sequence tags. The evidence set was augmented with ab initio gene models and the gene structures refined for all accessions through phylogeny-based methods. This pipeline revealed an average of 40,621 (SE = 117) protein-coding and 4998 (SE = 100) noncoding gene models per genome. Most genes share orthologs with the grass (Poaceae) family and species in the Andropogoneae tribe of grasses, which includes maize and sorghum (Fig. 1A). The accuracy of the annotations, which was measured by the congruence between annotations and supporting evidence (annotation edit distance) (*22*), is higher than that of previous reference maize annotations (fig. S1) (*2*, *6*, *10*, *18*, *23*).

We next assessed the gene catalog of the pan-genome. Genes with high sequence similarity, located within blocks of homologous sequence in pairwise comparisons, were grouped together as one pan-gene. In many instances, a gene was not annotated by our computational pipeline, yet at least 90% of the gene was present in the correct homologous location; when this occurred, the pan-gene was considered present (fig. S2A) (*17*), even though in some cases, the absence of

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA. [2]Genome Informatics Facility, Iowa State University, Ames, IA 50011, USA. [3]USDA-ARS Corn Insects and Crop Genetics Research Unit, Iowa State University, Ames, IA 50011, USA. [4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [5]Department of Genetics, University of Georgia, Athens, GA 30602, USA. [6]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA. [7]Department of Agronomy, Iowa State University, Ames, IA 50011, USA. [8]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA. [9]Center for Population Biology, University of California, Davis, CA 95616, USA. [10]Department of Evolution and Ecology, University of California, Davis, CA 95616, USA. [11]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA. [12]Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN 55108, USA. [13]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA. [14]Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA, 50010, USA. [15]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA. [16]Corteva Agriscience, Johnston, IA 50131, USA. [17]Genome Center, University of California, Davis, CA 95616, USA. [18]USDA-ARS NAA Robert W. Holley Center for Agriculture and Health, Agricultural Research Service, Ithaca, NY 14853, USA.
*Corresponding author. Email: kdawe@uga.edu

Downloaded from http://science.sciencemag.org/ on August 5, 2021

Table 1. Quality metrics for genome assemblies and gene model annotations. Darker shading indicates higher quality. The NAM lines are shaded on the basis of their primary grouping (gold, stiff-stalk heterotic group; blue, non–stiff-stalk heterotic group; gray, mixed tropical-temperate ancestry; purple, popcorn; orange, sweet corn; green, tropical). Hp301 and P39 have the lowest amounts of TR-1 and subtelomere repeats, respectively. Our methods can overestimate assembly when repeats are in low abundance (*17*).

	B73_v4	B73_v5	B97	Ky21	M162W	Ms71	Oh43	Oh7B	M37W	Mo18W	Tx303	HP301	P39	Il14H	CML52	CML69	CML103	CML228	CML247	CML277	CML322	CML333	Ki3	Ki11	NC350	NC358	Tzi8
Assembly Size (Mb)	2134	2182	2193	2172	2184	2214	2177	2165	2192	2223	2216	2141	2139	2125	2308	2225	2162	2301	2215	2191	2219	2231	2216	2274	2291	2227	2271
Contig N50 (Mb)	1.2	52.36	49.77	19.07	27.81	34.1	28.63	13.62	39.62	24.98	27.97	35.6	35.78	19.64	11.2	21.34	11.34	9.553	11.43	6.255	30.49	28.82	16.18	31.4	49	25.94	11.61
Scaffold N50 (Mb)	10.69	160.85	137.68	115.38	111.38	98.45	105.60	140.13	105.37	111.10	99.16	135.87	147.88	135.80	92.05	107.57	129.92	108.07	101.10	98.85	102.20	99.84	107.93	110.07	100.66	98.95	100.58
Pseudomolecule % N	1.43	0.175	0.156	0.306	0.175	0.23	0.121	0.407	0.087	0.338	0.314	0.198	0.117	0.158	0.936	0.296	0.241	1.207	0.459	0.426	0.144	0.146	0.392	0.121	0.072	0.175	0.567
BUSCO (% complete)	95.70	95.76	95.69	96.04	96.04	95.97	95.76	95.76	95.97	96.60	95.83	95.63	95.76	95.63	95.76	95.90	95.56	96.25	96.32	96.18	95.42	96.32	96.67	95.83	96.18	96.18	96.11
LTR Assembly Index (LAI)	26.68	27.84	28.06	28.08	28.09	27.91	27.89	28.04	28.09	27.81	27.71	28.05	27.61	27.83	27.92	28.34	28.3	27.93	28.44	27.95	28.44	28.33	28.27	27.64	27.96	28.22	27.9
CentC (% assembled)	17.52	87.94	38.89	56.78	44.54	75.47	66.73	47.55	79.84	75.75	76.42	65.64	69.96	55.31	45.25	56.92	54.95	55.32	47.43	29.33	69.96	69.24	43.82	74.21	64.18	52.69	62.81
Knob180 (% assembled)	5.651	18.63	8.24	8.96	7.89	7.79	4.34	8.35	6.91	10.73	7.49	22	22.24	55.54	4.89	2.21	10.57	3.97	3.65	3.2	4.73	3.83	3.44	12.71	13.9	5.61	2.7
TR-1 (% assembled)	23.01	89.43	66.41	15.67	36.98	25.13	8.26	36.76	79.14	13.93	34.96	110.8	42.22	86.81	6.93	3.17	8.45	3.3	5.54	9.55	4.08	11.42	10.76	20.3	12.6	5.29	2.48
rDNA arrays (% assembled)	0.352	9.41	7.16	10.71	8.76	6.05	6.5	9.74	7.5	16.34	6.38	13.54	6.53	6.89	13.5	9.77	8.33	5.64	8.48	8.02	11.56	11.64	7.33	3.97	10.9	9.44	7.88
Subtelomere (% assembled)	1.963	90.31	69	52.2	73.43	60.75	60.06	48.66	70.63	70.47	68.71	65.19	116.5	97.04	19.65	85.34	34.06	85.15	52.76	32.95	94.6	70.27	86.82	91.63	88.51	64.07	83.64
Gene Length (average)	4163	4477	4403	4371	4403	4349	4408	4332	4377	4327	4278	4405	4232	4337	4477	4446	4436	4318	4564	4348	4280	4432	4439	4442	4445	4406	4382
Genic Space Annotated (%)	8.03	8.17	8.12	8.23	8.38	8.12	8.25	7.97	8.17	8.05	7.97	8.2	8.22	8.24	7.93	8.07	8.23	7.9	8.36	8.04	7.94	8.04	8.26	7.8	7.86	7.88	8.07

annotation may reflect fractionation and/or pseudogenization.

Across the 26 genomes, a total of 103,033 pan-genes were identified. Previous analysis reported ~63,000 pan-genes on the basis of transcriptome assemblies of seedling RNA sequencing (RNA-seq) reads from 500 individuals (*1*). The superior contiguity of our assemblies and the application of both ab initio and evidence-based annotation using RNA-seq from a diverse set of 10 tissues likely account for the increased sensitivity. More than 80% of pan-genes were identified within just 10 inbred lines on the basis of a bootstrap resampling of genomes (Fig. 1B). When considered separately, temperate and tropical lines have differentiated sets of pan-genes but show a comparable rate of pan-gene increase as lines are added, suggesting they have similar gene-content diversity (Fig. 1B).

Pan-genes, excluding tandem duplicates (*17*), were classified as core (present in all 26 lines), near-core (present in 24 to 25 lines), dispensable (present in 2 to 23 lines), and private (present in only 1 line) (Fig. 1C). The portion of genes classified into each of these groups was consistent across genotypes, with an average of 58.41% (SE = 0.07%) belonging to the core genome, 8.23% (SE = 0.05%) to the near-core genome, 31.75% (SE = 0.09%) to the dispensable genome, and 1.60% (SE = 0.08%) private genes (Fig. 1C; fig. S2, B and C; and table S3). In total, 32,052 genes are in the core or near-core portion of the pan-genome, and 70,981 are genes in the dispensable or private portion. The core genes (and gene families enriched for core genes) (table S4) are generally from higher phylostrata levels (i.e., Viridiplantae and Poaceae), whereas those in the near-core and dispensable sets either share orthologs only with closely related species or are maize specific (fig. S2E). Some private genes may be spurious annotations that result from imperfect masking of repeat sequences, as most core and near-core genes are syntenic to sorghum (57.78%), whereas this is rarely the case for dispensable and private genes (1.83% syntenic). Core genes were expressed in more tissues (Fig. 1D) and had higher transcript abundance (fig. S2F) when compared with genes present in fewer individuals. However, across the relatively small number of tissues (eight or more per line) profiled for this analysis, 18% of dispensable and 32% of private genes were expressed in at least one tissue. A total of 16,751 pan-genes were tandemly duplicated in at least one genome, of which 7040 were duplicated in a single genome. On a per-gene basis in genomes with at least one tandem duplicate, the average copy number is 2.20 (SE = 0.01) (fig. S2D).

Partial tetraploidy and tempo of fractionation

The maize ancestor underwent a whole-genome duplication (WGD) allopolyploidy event 5 to 20 million years ago (Fig. 2A) (*24*, *25*). Evidence for WGD is found in the existence of two separate genomes that are broken and rearranged yet still show clear synteny to sorghum (*24*, *26*). Many duplicated genes have since undergone loss, or fractionation, reducing maize to its current diploid state (*26*, *27*). Furthermore, fractionation is biased toward one homoeologous genome (M2, more fractionated) over the other (M1, less fractionated) (*26*). The M1 and M2 subgenomes are composed almost exclusively of core (87.25%) and near-core (6.19%) pan-genes (Figs. 1C and 2A). The broad architecture of syntenic regions relative to sorghum is consistent across the NAM genomes (fig. S3).

Given the ancient time frame of the WGD in maize and the rapid tempo of fractionation observed in other species (*28*, *29*), little variation in the retention of specific homoeologs is expected at the species level. In fact, prior work in temperate maize suggested that most fractionation occurred before domestication (*6*, *30*). However, our diverse set of genomes allows for a more complete characterization of fractionation within the species. Because fractionation can occur at the level of small deletions (*27*, *31*), we evaluated both partial and complete homoeolog loss beginning with a conservative set of 16,195 maize pan-orthologs. We determined that 7043 were single-copy orthologs, in which the homoeologous gene was likely deleted before maize speciation (Fig. 2A). In addition, we identified 4576 homoeologous pairs (Fig. 2A), of which 2155 had the same exon structure of the sorghum ortholog in both homoeologs. In 1281 pairs, at least one copy of the gene differed from its sorghum ortholog but did not vary among NAM lines, likely representing fractionation that pre-dated *Zea mays*. Another 1140 pairs varied across the genomes in their pattern of exon retention, segregating for deletions or structural differences in at least one copy of the gene. This segregating set was manually curated (data S1) to remove loci where exons or flanking sequence could not be confidently identified (Fig. 2A), resulting in a curated set of 494 homoeolog pairs segregating for fractionation, which represents >10% of pairs present in the pan-genome. Of these, 281 M2 homoeologs had exon loss compared with 236 M1 homoeologs, a 19% difference ($P < 0.05$, χ^2 test), which suggests ongoing biased fractionation. Analysis of gene ontology terms revealed putative functional differences between fully fractionated and segregating fractionated loci (fig. S4 and data S1).

Population genetic theory predicts that mutations segregating within a species, such as the segregating fractionation deletions we have identified, arose within the past $4N_e$ generations, where N_e represents the effective population size of the species. Using the N_e of the maize progenitor teosinte as an upward bound for maize [$N_e = 150,000$; (*32*)], we can infer that most segregating fractionation arose within the past 600,000 generations. Therefore, most segregating fractionation substantially postdates the WGD. Theory also predicts that

Downloaded from http://science.sciencemag.org/ on August 5, 2021

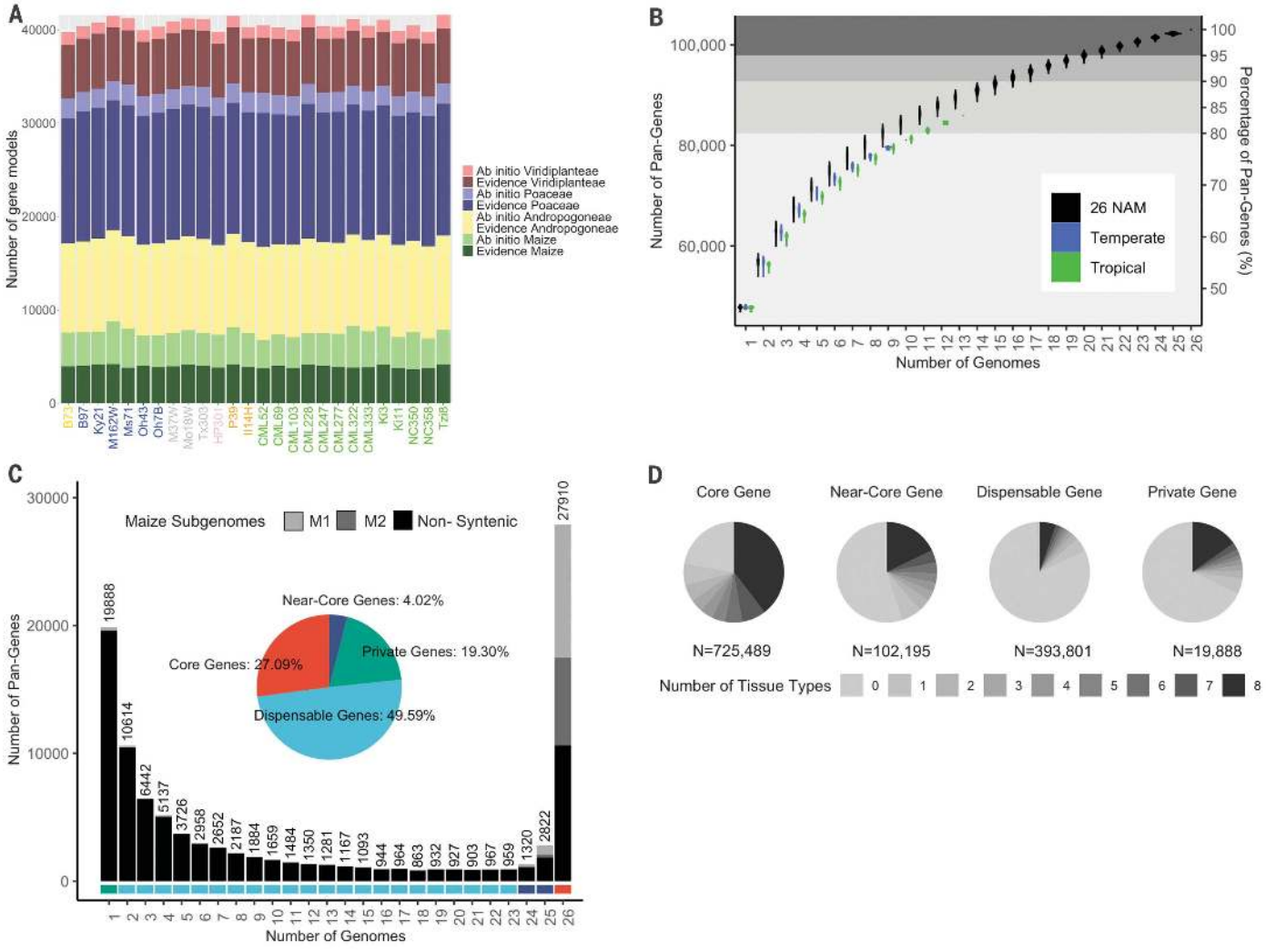

Fig. 1. Pan-genome analysis of the gene space. (**A**) Pan-genes categorized by annotation method and phylostrata. Genes annotated with evidence have mRNA support, whereas ab initio genes are predicted on the basis of DNA sequence alone. Genes within progressing phylostrata [species *Z. mays* (maize), tribe Andropogoneae, family Poaceae, kingdom Viridiplantae] are more conserved. (**B**) Number of pan-genes added with each additional genome assembly. Order of genomes being added into the pan-genome was bootstrapped 1000 times. Tropical lines include CML52, CML69, CML103, CML228, CML247, CML277, CML322, CML333, Ki3, Ki11, NC350, NC358, and Tzi8; temperate lines include B73, B97, Ky21, M162W, Ms71, Oh43, Oh7B, HP301, P39, and Il14H. (**C**) Proportion of pan-genes in the core, near-core, dispensable, and private fractions of the pan-genome. For (B) and (C), tandem duplicates were considered as a single pan-gene and coordinates were filled in when a gene was not annotated, but an alignment with >90% coverage and 90% identity was present within the correct homologous block. (**D**) Number of tissues with expression (reads per kilobase per million reads > 1) for each gene in each genome on the basis of their pan-genome classification. Tissues in this analysis include root, shoot, V11 base, V11 middle, V11 tip, anther, tassel, and ear.

rare deletions should be younger than those segregating at intermediate frequency. We constructed the unfolded site frequency spectrum (SFS) of segregating fractionation deletions and compared this with the unfolded SFS of noncoding single-nucleotide polymorphisms (SNPs) using sorghum to define the ancestral state (Fig. 2B). The data reveal a similar frequency distribution in deletions and SNPs, with a preponderance of rare variants in both, suggesting that a subset of fractionation may be quite young, with diploidization potentially continuing in modern maize. We also evaluated patterns of co-exon retention in non–stiff-stalk temperate, tropical, and flint-derived maize, observing population-specific fractionation (Fig. 2C). This variation in homoeolog retention at the population level confirms previous suppositions about the tempo of fractionation (*33*) and may reflect relaxed constraint on retained homoeologs after the domestication and migration of maize to temperate climates.

The repetitive fraction of the pan-genome

TEs were annotated in each assembly by using structural features and sequence homology (*34*). Individual TE libraries from each inbred were then combined to form a pan-genome library, which was used to identify TE sequences missed by individual libraries. The annotations reveal that DNA transposons and LTR retrotransposons constitute 8.5 and 74.4% of the genome, respectively (table S5 and fig. S5). A total of 27,228 TE families were included in the pan-genome TE library, of which 59.7% were present in all 26 NAM founders, and 2.5% were specific to one genome (fig. S6). The average percentages of intact and fragmented TEs were 30.5 and 69.5% (SE = 0.06%), respectively. As reported previously, *Gypsy* LTR retrotransposon families are more abundant in pericentromeric regions, whereas *Copia*

Downloaded from http://science.sciencemag.org/ on August 5, 2021

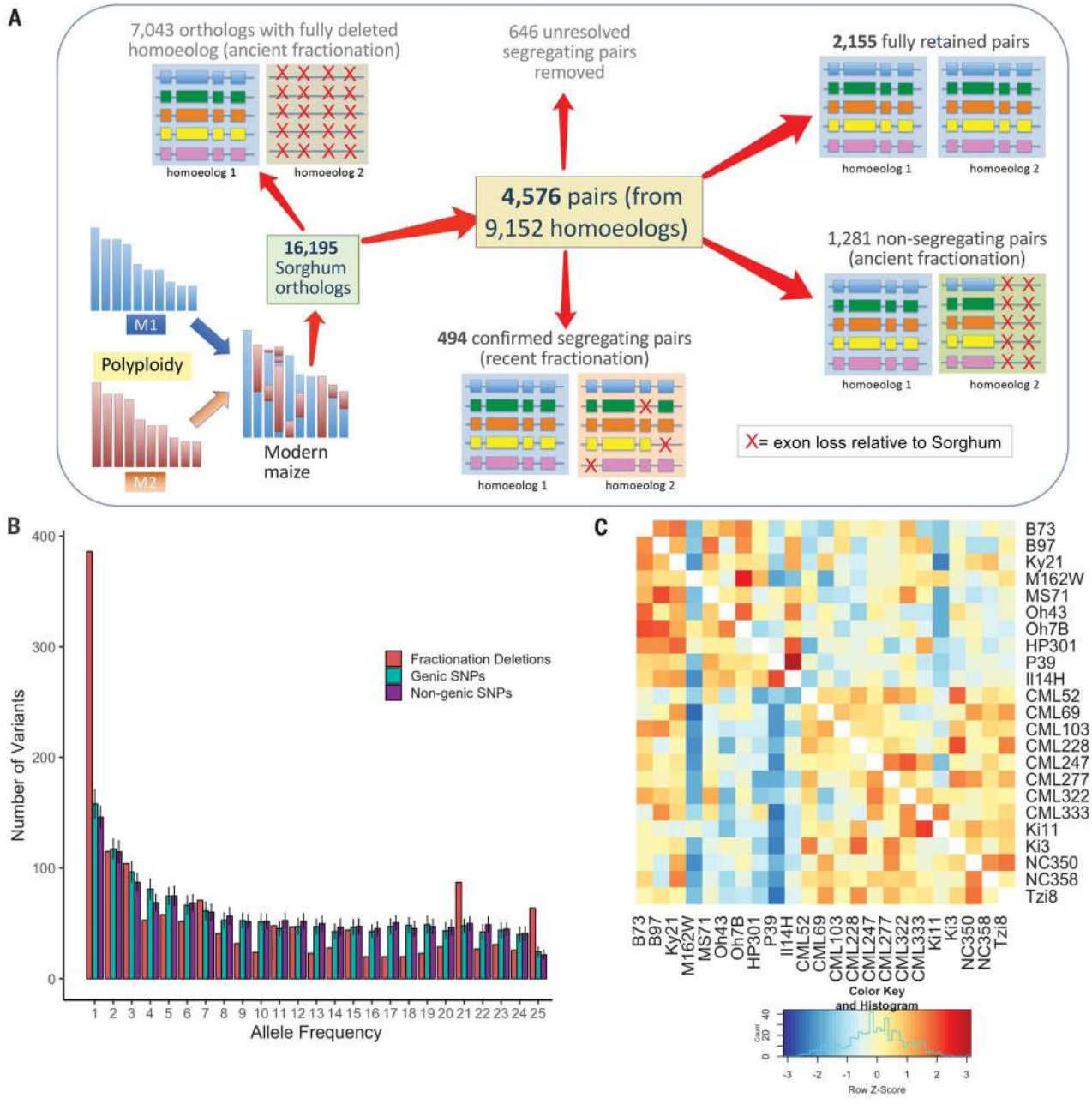

Fig. 2. The tempo of fractionation in maize. (**A**) Schematic showing how genes were categorized. A total of 16,195 conservatively chosen orthologs were subdivided into classes, representing retained pairs, ancient fractionation, and recent fractionation. (**B**) Unfolded SFS of segregating exon loss and noncoding SNPs (genic and nongenic) by using sorghum to define the ancestral state. (**C**) Heatmap of the number of co-retained exons between any two NAM lines. Lines with mixed ancestry (M37W, Mo18W, Tx303) are excluded. Colors indicate the *z*-score (the difference measured in standard deviations between a single pairwise comparison and all others in the row).

LTR retrotransposons are enriched in the gene-dense chromosome arms (fig. S7) (*35*). Tropical lines have significantly more *Gypsy* elements than temperate lines (P = 0.002, t test), with mean *Gypsy* content of 1018 and 988 Mbp, respectively (table S5 and fig. S5). This may reflect increasing constraint on *Gypsy* proliferation in temperate lines that have, on average, smaller genomes (Table 1).

In some maize lines, >15% of the genome is composed of tandem repeat arrays, including the centromere repeat CentC, the two knob repeats knob180 and TR-1, subtelomere, and telomere repeats (*36*, *37*). Repeats of this type remain a major impediment to assembly. A mean of 60% of CentC, 70% of the 4-12-1 subtelomeric sequence (*38*), 28.9% of TR-1, 1% of knob180, and 0.09% of ribosomal DNA repeat units were incorporated in the final assemblies (Table 1).

A total of 110 (of 260) functional centromeres identified by CENH3 chromatin immunoprecipitation sequencing (*39*, *40*) were fully assembled, and of these, 88 are gapless (fig. S8A) (*40*). Chromosomes with very long CentC arrays (such as chromosomes 1, 6, and 7) often have assembly gaps, and the precise location of the centromere could not be determined. In other cases, the centromeres include fully assembled small CentC arrays or the functional centromeres are located to one side of the CentC tracts in regions dominated by retrotransposons (Fig. 3A). By projecting all

Downloaded from http://science.sciencemag.org/ on August 5, 2021

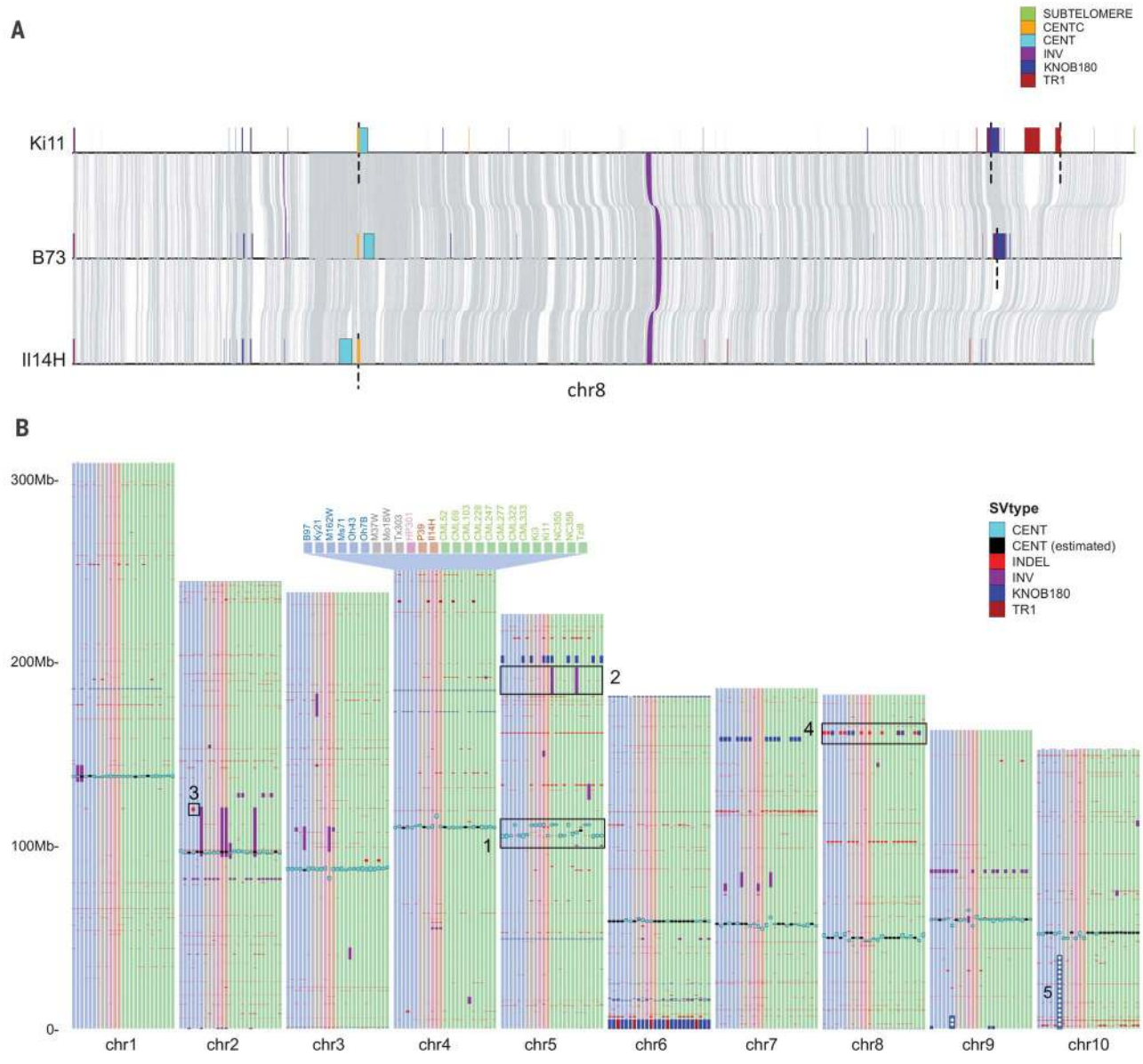

Fig. 3. Structural variation in the NAM founders. (**A**) Pairwise alignments between Ki11, B73, and Il14H on chromosome 8. Gray links represent syntenic aligned regions; gaps of unknown size (scaffold gaps) are marked by dashed lines. INV, inversion. (**B**) Large (>100 kbp) SVs, centromeres, and knobs across the NAM lines versus the B73 reference. The subset of SVs larger than 1 Mbp were manually curated, and only those containing genes are represented. Features 1 to 5 highlight major SVs: (1) multiple centromere movement events; (2) a major inversion previously hypothesized on the basis of suppressed recombination; (3) a large deletion in the Ms71 inbred; (4) knob polymorphism; (5) reciprocal translocation between chromosome 9 and 10 in the Oh7B inbred (both segments placed in their standard positions for display).

centromere locations onto B73, we were able to identify 12 centromere movement events (3 on chr5 and chr9 and 2 on chr3, chr8, and chr10), which clarifies and extends prior evidence for centromere shifting (Fig. 3B and fig. S8B) (*39*). The variation in CentC abundance and positional polymorphism made it possible to gaplessly assemble at least two variants of all 10 centromeres (fig. S8A).

Both knob180 and TR-1 arrays are subject to meiotic drive and accumulate when a chromosome variant known as Abnormal chromosome 10 (Ab10) is present (*37*, *41*). Although Ab10 is absent from modern inbreds, its legacy remains in the form of many large knobs. Most knob180 and TR-1 repeat arrays were identified in midarm positions (81.9%), where meiotic drive is most effective. Long knob180 and TR-1 repeat arrays can occur separately but are more frequently intermingled in fragmented arrays along with transposons (Fig. 3A and fig. S9) (*42*). Analysis of classical (cytologically visible) knobs on chromosomes 1S, 2S, 2L, 3L, 4L, 5L, 6L, 7L, 8L, and 9S revealed that their locations are syntenic and that several are composed of a series of disjointed smaller knobs (Fig. 3A and fig. S10). In some lines, knobs are not visible cytologically but can still be detected as smaller arrays at the sequence level; however, many show strict presence-absence variation among the NAM founder inbreds.

Tandem repeat arrays are also commonly found at the ends of chromosome arms (table S6). Among the 520 chromosome ends, 57.9% contained knob180 repeats, and 30.5% contained

Downloaded from http://science.sciencemag.org/ on August 5, 2021

subtelomere repeats. At least 65.6% of chromosome ends were fully assembled as indicated by the presence of telomere sequences.

Structural variation and impact on phenotype

Comparative analyses among the NAM genotypes to B73 revealed a cumulative total of 791,101 structural variants (SVs) >100 bp in size. Tropical lines, which are the most divergent from B73, include a substantially higher number of SVs than temperate lines (mean = 32,976 versus 29,742; $P = 0.00013$) (tables S7 and S8). SVs are more common on chromosome arms where recombination is highest (fig. S11), similar to SNPs and other forms of genetic variation (*43*). Almost half (49.6%) of SVs were <5 kbp in size, with 25.7% being <500 bp. Across all size classes, SVs are skewed toward rare variants (fig. S12). Several large SVs were found segregating within the 26 NAM genomes (Fig. 3B), including 35 distinct inversion polymorphisms and 5 insertion-deletion polymorphisms >1 Mbp. For example, a 14.6-Mbp inversion on chromosome 5 in the CML52 and CML322 lines, which was previously hypothesized on the basis of suppressed recombination in the NAM RILs (*11*), is confirmed in this study through assembly. Additionally, there is a 1.9-Mbp deletion with seven genes on chromosome 2 in the MS71 inbred and a 1.8-Mbp deletion with two genes on chromosome 8 found in eight lines. Our data also capture a very large reciprocal translocation (involving >47 Mbp of DNA) between the short arms of chromosomes 9 and 10 in Oh7B that had been previously detected in cytological studies (Fig. 3B) (*38*).

The high proportion of rare SVs in maize suggests that these may be a particularly deleterious class of variants, as observed in other species (*44*, *45*). Indels and inversions occur in regions that have 49.8% fewer genic base pairs than the genomic background. Furthermore, SVs are 17% less likely to be found in conserved regions than SNPs (odds ratios of 0.27 and 0.58 for SVs and SNPs, respectively; Fisher's exact test; $P < 0.001$). Approximate Bayesian computation modeling revealed that selection against SVs is at least as strong as that against nonsynonymous substitutions (fig. S13) (*17*). These results suggest that, when they occur, SVs are particularly consequential and relevant to fitness.

To estimate the phenotypic impact of SVs, we assessed the genetic basis of 36 complex traits (*14*) using 71,196 filtered SVs in 4027 recombinant inbred lines derived from the NAM founder inbreds (fig. S14A) (*11*). The analysis revealed that SVs explain a high percentage of phenotypic variance for disease traits (60.10 ~ 61.75%) and less for agronomic or morphological (20.04 ~ 61.04%) and metabolic traits (4.79 ~ 26.78%). Much of the phenotypic variation was also explained by SNPs, which were much more numerous (288-fold more) relative to our conservative set of SVs (fig. S14A). When the SNP and SV data were integrated into one linear mixed model, the combined markers only slightly surpassed values from SNPs, consistent with the fact that most SVs are in high linkage disequilibrium with SNPs (fig. S14A).

We also carried out genome-wide association analyses (GWASs) to identify specific SVs contributing to phenotypic variation for the same suite of traits (fig. S14, B to G). Among the detected GWAS signals, 93.05% overlapped with those identified with SNPs, and 6.95% were specific to SVs (no significant SNP detected within 5 Mbp of significant SVs). There

Downloaded from http://science.sciencemag.org/ on August 5, 2021

Fig. 4. UMR variation across the NAM founders. (**A**) Annotation of the *Miniature seed1* gene in the Mo18W inbred. An image from the MaizeGDB browser shows gene, TE, and UMR tracks. TE tracks are color-coded by superfamily: green-gray, long terminal repeats; red, terminal inverted repeats; and blue, long interspersed nuclear elements. The gray vertical lines show 2.5-kbp intervals. (**B**) Annotation and underlying methylation data for *Miniature seed1* in the B73 inbred. The insertion of a *Gypsy* element moved part of the proximal UMR to a position 14 kbp upstream from the TSS. Methylation tracks indicate base pair–level methylation values from 0 to 100%. Asterisks indicate gaps in coverage, which are visible in separate tracks (fig. S28). (**C**) Relationship between methylation and gene expression. UMRs were mapped to B73 to identify UMRs that overlap with TSS. The *y* axis indicates the ratio of transcripts per million (TPM; compared with B73) when the region is methylated (red) or unmethylated (teal).

was a significant enrichment of SVs associated with phenotypes in genic regions (z = 8.022, P < 1.04e-15) (fig. S15). The most significant association between an SV and a trait not identified with SNP markers was a quantitative trait locus for northern leaf blight on chromosome 10 (fig. S14F). This SV is within a gene encoding a thylakoid lumenal protein; such proteins could be linked to plant immunity through the regulation of cell death during viral infection (*46*). We anticipate that the effects of SVs may be even more pronounced in larger association panels, where extensive historical recombination may help disentangle their effects from nearby SNPs.

Disease resistance in plants is frequently associated with SV in the form of tandem arrays of resistance genes. Complex arrays of resistance genes are retained, potentially through birth-death dynamics in an evolutionary arms race with pathogens or through balancing selection for the maintenance of diverse plant defenses (*47*). Nucleotide-binding, leucine-rich-repeat (NLR) proteins provide a common type of resistance. Our data reveal that there are fewer NLR genes in maize than in other Poaceae (fig. S16) and that most NAM lines have lost the same clades of NLRs as sorghum (fig. S17). Only one line (CML277) retains the MIC1 NLR clade, which is particularly fast-evolving in Poaceae (*48*). Nevertheless, there is clear NLR variation among the NAM lines (fig. S18), and tropical genomes contain a significantly higher number of NLR genes than temperate genomes (t test, P = 0.006), suggesting ongoing coevolution with pathogens, particularly where disease pressure is high.

The annotated NLR genes were significantly enriched for overlap with SVs (bootstrap permutation test, P < 0.001). An extreme example is found at the *rp1* (resistance to *Puccinia sorghi1*) locus on the short arm of chromosome 10, which is known to be highly variable (*49*). We observed exceptional diversity in the NAM lines with as few as 4 *rp1* copies in P39 and as many as 30 in M37W (table S9). However, because of its repetitive nature, only 18 NAM lines have gapless assemblies of the *rp1* locus.

SVs linked to transposons have been shown, through the modulation of gene expression, to underlie flowering-time adaptation in maize during tropical-to-temperate migration (*50*, *51*). Our SV and TE-annotation pipelines identified the adaptive *CACTA*-like insertion that was previously reported upstream of the flowering-time locus *ZmCCT10* (*51*). We also surveyed 173 genes linked to flowering-time (*52*, *53*) and discovered three genes (*GI15*, *ZCN10*, and *Dof21*) with TE-derived SVs <5 kbp upstream of their transcription start sites (TSSs). These SVs distinguish temperate from tropical lines (t < −2.346, P < 0.0358) (fig. S19) and show significant correlation (F > 8.658, P < 0.001) with expression levels.

Discovery of candidate cis-regulatory elements through DNA methylation

On the basis of sequence alone, it can be difficult to identify functional sequences in the intergenic spaces. One approach is to score for unmethylated DNA, which provides both a tissue-independent indicator of gene regulatory elements and evidence that annotated genes are active (*5*, *54*, *55*). We sequenced enzymatic methyl sequencing libraries from each NAM line and identified methylated bases in three sequence contexts, CG, CHG, and CHH (where H is A, T, or C). Results are consistent across genes and transposons, demonstrating the quality of the libraries (figs. S20 and S21). There is minor variation in total methylation across inbreds, with CML247 being noteworthy for uniformly lower CG methylation in several tissues (fig. S22). Such natural variation in methylation is also observed in *Arabidopsis* ecotypes (*56*).

Each of the three methylation (m) contexts reveals information on the locations of repeats, genes, and regulatory elements. mCHH levels are generally low except at heterochromatin borders, whereas mCHG and mCG are abundant in repetitive regions. Both mCHG and mCG are depleted from regulatory elements, and mCHG is depleted from exons (*57*). However, mCG is often present in exons (Fig. 4) (*58*). Thus, to identify unmethylated regions (UMRs) that correspond to regulatory elements and gene bodies, we defined UMRs using a method that takes into account mCHG and mCG but does not exclude high mCG–only regions (the term UMR is used for simplicity; some regions contain CG methylation). Comparison of the 26 methylomes revealed uniformity in number and length of UMRs, averaging ~180 Mbp in total length in each genome (figs. S23 and S24). To confirm the accuracy of the UMR data, we also identified accessible chromatin regions using ATAC sequencing for each inbred. We expect chromatin to be accessible mainly in the subset of genes expressed in the tissue sampled (primarily leaves) and to show concordance with UMRs. The data reveal that a mean of 99% of genic and 96% of nongenic (distal) accessible chromatin regions overlap with UMRs in each genome (figs. S25 and S26).

To assess methylation diversity, we mapped UMRs from all inbreds to the B73 genome. Approximately 95% of genic UMRs overlap across genomes in pairwise comparisons (fig. S27). UMR polymorphism is higher in the intergenic space, particularly among UMRs >5 kbp from genes, where typically ~75% of UMRs overlap (fig. S27). Even when the UMR sequence is conserved, its position relative to the closest gene may vary substantially among inbreds. This is exemplified by the *Miniature Seed1* gene, in which a UMR proximal to the promoter in Mo18W is displaced nearly 14 kbp upstream in B73 by a single *Huck* element (*Gypsy* LTR superfamily) (Fig. 4). The *Huck* insertion is present in 23 of 26 genomes, and in 2 of these (Oh43 and CML322), additional nested TE insertions increased the distance between the gene and the UMR to 27 kbp. Although UMR polymorphism correlates with genetic distance across NAM lines (fig. S29), UMRs from Tzi8 were not substantially shared with other tropical genomes.

Adaptive variation in DNA methylation has been observed in maize (*59*), most likely through effects on gene expression. To estimate how well UMRs predict transcription, we identified a conservative subset of UMR overlapping genes that were unmethylated in B73 but methylated in at least one other methylome. These differentially methylated regions were strongly correlated with differences in gene expression (Fig. 4 and fig. S30). We further evaluated the enrichment of significant GWAS SNPs across 36 traits in UMRs. From genome-wide estimates, UMRs show 2.50- to 3.26-fold enrichment across traits for significant associations. Roughly 18% of SNPs identified by GWAS lie outside of genic regions but within UMRs (table S10), which is consistent with the view that UMRs can be used to identify functional, noncoding regions (*5*, *54*, *55*).

Summary

Our analysis of 26 genomes uncovered variation in both the genic and repetitive fractions of the pan-genome. Tropical, temperate, and flint-derived popcorn and sweet corn germplasm are differentiated in distinctive ways, including their pan-gene complement, homoeolog retention after polyploidy, abundance of TEs, NLR disease-resistance gene copy number, and methylation profiles. The available data will have broad utility for genetic and genomic studies and facilitate rapid associations to phenotyping information. For example, the genic presence-absence variation that was identified in this study may be imputed across additional mapping populations to clarify its contribution to heterosis through complementation (*60*). More generally, these resources should motivate a shift away from the single-reference mindset to a multireference view in which any one of 26 inbreds, each with different experimental and agronomic advantages, can be deployed for the purposes of basic discovery and crop improvement.

REFERENCES AND NOTES

1. C. N. Hirsch *et al.*, *Plant Cell* **26**, 121–135 (2014).
2. C. N. Hirsch *et al.*, *Plant Cell* **28**, 2700–2714 (2016).
3. M. Jin *et al.*, *Sci. Rep.* **6**, 18936 (2016).
4. F. Lu *et al.*, *Nat. Commun.* **6**, 6914 (2015).
5. W. A. Ricci *et al.*, *Nat. Plants* **5**, 1237–1249 (2019).
6. S. Sun *et al.*, *Nat. Genet.* **50**, 1289–1295 (2018).

Downloaded from http://science.sciencemag.org/ on August 5, 2021

7. G. Haberer *et al.*, *Nat. Genet.* **52**, 950–957 (2020).
8. N. Yang *et al.*, *Nat. Genet.* **51**, 1052–1059 (2019).
9. G. Lin *et al.*, *Genome Biol.* **22**, 175 (2021).
10. N. M. Springer *et al.*, *Nat. Genet.* **50**, 1282–1288 (2018).
11. M. D. McMullen *et al.*, *Science* **325**, 737–740 (2009).
12. S. A. Flint-Garcia *et al.*, *Plant J.* **44**, 1054–1064 (2005).
13. J. Yu, J. B. Holland, M. D. McMullen, E. S. Buckler, *Genetics* **178**, 539–551 (2008).
14. J. G. Wallace *et al.*, *PLOS Genet.* **10**, e1004845 (2014).
15. S. R. Eichten *et al.*, *Plant Cell* **25**, 2783–2797 (2013).
16. R. J. Schaefer *et al.*, *Plant Cell* **30**, 2922–2942 (2018).
17. See supplementary materials.
18. Y. Jiao *et al.*, *Nature* **546**, 524–527 (2017).
19. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, *Bioinformatics* **31**, 3210–3212 (2015).
20. S. Ou, J. Chen, N. Jiang, *Nucleic Acids Res.* **46**, e126 (2018).
21. S. Ou *et al.*, *Nat. Commun.* **11**, 2288 (2020).
22. K. Eilbeck, B. Moore, C. Holt, M. Yandell, *BMC Bioinformatics* **10**, 67 (2009).
23. M. Law *et al.*, *Plant Physiol.* **167**, 25–39 (2015).
24. Z. Swigonová *et al.*, *Genome Res.* **14** (10a), 1916–1923 (2004).
25. X. Wang *et al.*, *Mol. Plant* **8**, 885–898 (2015).
26. J. C. Schnable, N. M. Springer, M. Freeling, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4069–4074 (2011).
27. M. R. Woodhouse *et al.*, *PLOS Biol.* **8**, e1000409 (2010).
28. J. C. Schnable, M. Freeling, E. Lyons, *Genome Biol. Evol.* **4**, 265–277 (2012).
29. T. Mandáková, S. Joly, M. Krzywinski, K. Mummenhoff, M. A. Lysak, *Plant Cell* **22**, 2277–2290 (2010).
30. A. B. Brohammer, T. J. Y. Kono, N. M. Springer, S. E. McGaugh, C. N. Hirsch, *Plant J.* **93**, 131–141 (2018).
31. H. Tang *et al.*, *Genetics* **190**, 1563–1574 (2012).
32. T. M. Beissinger *et al.*, *Nat. Plants* **2**, 16084 (2016).
33. F. Cheng *et al.*, *Nat. Plants* **4**, 258–268 (2018).
34. S. Ou *et al.*, *Genome Biol.* **20**, 275 (2019).
35. R. S. Baucom *et al.*, *PLOS Genet.* **5**, e1000732 (2009).
36. P. Bilinski *et al.*, *PLOS Genet.* **14**, e1007162 (2018).
37. R. K. Dawe *et al.*, *Cell* **173**, 839–850.e18 (2018).
38. P. S. Albert, Z. Gao, T. V. Danilova, J. A. Birchler, *Cytogenet. Genome Res.* **129**, 6–16 (2010).
39. K. L. Schneider, Z. Xie, T. K. Wolfgruber, G. G. Presting, *Proc. Natl. Acad. Sci. U.S.A.* **113**, E987–E996 (2016).
40. N. Wang, J. Liu, W. A. Ricci, J. I. Gent, R. K. Dawe, *Genetics* **217**, iyab020 (2021).
41. K. W. Swentowsky *et al.*, *Genes Dev.* **34**, 1239–1251 (2020).
42. J. Liu *et al.*, *Genome Biol.* **21**, 121 (2020).
43. J.-M. Chia *et al.*, *Nat. Genet.* **44**, 803–807 (2012).
44. H. J. Abel *et al.*, *Nature* **583**, 83–89 (2020).
45. E. V. Leushkin, G. A. Bazykin, A. S. Kondrashov, *Genome Biol. Evol.* **5**, 514–524 (2013).
46. S. Seo *et al.*, *Plant Cell* **12**, 917–932 (2000).
47. H. Mizuno *et al.*, *Sci. Rep.* **10**, 872 (2020).
48. P. C. Bailey *et al.*, *Genome Biol.* **19**, 23 (2018).
49. S. H. Hulbert, J. L. Bennetzen, *Mol. Gen. Genet.* **226**, 377–382 (1991).
50. C. Huang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E334–E341 (2018).
51. Q. Yang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16969–16974 (2013).
52. Z. Dong *et al.*, *PLOS ONE* **7**, e43450 (2012).
53. Y.-X. Li *et al.*, *Plant J.* **86**, 391–402 (2016).
54. R. Oka *et al.*, *Genome Biol.* **18**, 137 (2017).
55. P. A. Crisp *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 23991–24000 (2020).
56. T. Kawakatsu *et al.*, *Cell* **166**, 492–505 (2016).
57. J. I. Gent *et al.*, *Genome Res.* **23**, 628–637 (2013).
58. A. J. Bewick, R. J. Schmitz, *Curr. Opin. Plant Biol.* **36**, 103–110 (2017).
59. G. Xu *et al.*, *Nat. Commun.* **11**, 5539 (2020).
60. N. M. Springer *et al.*, *PLOS Genet.* **5**, e1000734 (2009).
61. M. B. Hufford *et al.*, HuffordLab/NAM-genomes: publication. prerelease, Zenodo (2021); https://zenodo.org/record/4781590).

ACKNOWLEDGMENTS

We appreciate the sequencing services provided by the University of Arizona, Oregon State University, Brigham Young University, and the University of Georgia, as well as coordination among sequencing centers provided by Pacific Biosciences. We appreciate the early contributions to maize genome assembly in the cloud from DNAnexus. The authors further acknowledge the High Performance Computing facility at Iowa State University [partially funded by the National Science Foundation (NSF) 1726447], Minnesota Supercomputing Institute, the Georgia Advanced Computing Resource Center, Cold Spring Harbor Laboratory high-performance computing center (NIH S10 OD0286321-01), and the participants of the Virtual Maize Annotation Jamboree who evaluated the initial gene predictions for benchmarking and improvements in the final gene annotations. **Funding:** Primary support for this work came from a generous grant from the NSF (IOS-1744001). Additional support came from NSF IOS-1546727 to C.N.H., USDA 2018-67013-27571 to C.N.H., USDA-ARS 8062-21000-041-00D, NSF IOS-1127112 and NIH-OD S10 OD028632 to D.W., NSF IOS-1546719 to M.B.H., NSF IOS-1822330 to J.R.-I. and M.B.H., USDA Hatch project (CA-D-PLS-2066-H to J.R.-I.), NSF IOS-1856627 to R.J.S., an NSF Postdoctoral Fellowship in Biology (DBI-1905869 to A.P.M.), NSF Graduate Research Fellowships (1650042 to A.I.H. and 1744592 to S.J.S.), NSF Research Traineeship (DGE-1545463) to Iowa State University (trainee S.J.S.), USDA-ARS 58-5030-8-064 to M.B.H. and C.M.A., USDA-ARS project 5030-21000-068-00D to C.M.A. and M.R.W., and NSF IOS-1546657 to J.Y. **Author contributions:** Conceptualization—R.K.D., D.W., M.B.H., C.N.H., and J.I.G.; Data curation—M.R.W., A.S.S., K.M.C., S.O., J.L., S.W., A.P.M., Z.L., B.W., M.K.T.-R., J.L.P., E.K.S.C., and C.M.A.; Formal analysis—A.S.S., M.R.W., K.M.C., S.O., J.L., W.A.R., T.G., A.O., Y.Q., R.D.C., S.T., A.P.M., A.I.H., S.W., Z.L., B.W., M.K.T.-R., R.D.P., Y.Z., C.H.O., X.L., A.M.G., E.B., J.L.P., N.M., S.J.S., Q.J., S.P., M.L.S., K.F., and J.I.G.; Funding acquisition—R.K.D., D.W., M.B.H., C.N.H., J.I.G., and R.J.S.; Investigation—D.w.K., D.A.K., N.W., D.E.H., V.L., K.F., and J.I.G.; Methodology—M.B.H., D.W., C.N.H., A.S.S., M.R.W., K.F., W.A.R., J.L., R.J.S., J.I.G., J.R.-I., J.Y., and R.K.D.; Project administration—R.K.D., M.B.H., D.W., C.N.H., and J.I.G.; Software—A.S.S., D.E.H., N.M., and S.O.; Supervision—M.B.H., D.W., R.K.D., C.N.H., J.I.G., K.V.K., R.J.S., J.R.-I., and J.Y.; Visualization—M.R.W., A.S.S., J.L., W.A.R., Y.Q., K.M.C., S.O., R.D.P., S.J.S., C.N.H., and J.I.G.; and Writing—M.B.H., R.K.D., C.N.H., and J.I.G. **Competing interests:** R.J.S. is a cofounder of REquest Genomics, LLC, a company that provides epigenomic services. All other authors declare no competing interests. **Data and materials availability:** Genome assemblies and annotations can be accessed at https://maizegdb.org/NAM_project and http://maize-pangenome.gramene.org. Raw data used for the assemblies including PacBio, Illumina, and Bionano data are available through ENA BioProject IDs PRJEB31061 and PRJEB32225. RNA-seq data are available at ENA ArrayExpress E-MTAB-8633 and E-MTAB-8628. Enzymatic methyl sequencing reads are available at ENA ArrayExpress E-MTAB-10088. ATAC-sequencing reads are available under NCBI GEO accession GSE165787. Other files, tables and supplemental data can be found in CyVerse https://datacommons.cyverse.org/browse/iplant/home/shared/NAM/NAM_genome_and_annotation_Jan2021_release. Links to the NLR trees can be found at https://itol.embl.de/shared/xCJbI9ndshEK. Scripts used to generate and analyze data are available as a Zenodo repository (*61*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/373/6555/655/suppl/DC1
Materials and Methods
Figs. S1 to S30
Tables S1 to S10
References (*62*–*176*)
Data S1

View/request a protocol for this paper from *Bio-protocol*.

14 January 2021; accepted 24 June 2021
10.1126/science.abg5289

Downloaded from http://science.sciencemag.org/ on August 5, 2021

Science

De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes

Matthew B. Hufford, Arun S. Seetharam, Margaret R. Woodhouse, Kapeel M. Chougule, Shujun Ou, Jianing Liu, William A. Ricci, Tingting Guo, Andrew Olson, Yinjie Qiu, Rafael Della Coletta, Silas Tittes, Asher I. Hudson, Alexandre P. Marand, Sharon Wei, Zhenyuan Lu, Bo Wang, Marcela K. Tello-Ruiz, Rebecca D. Piri, Na Wang, Dong won Kim, Yibing Zeng, Christine H. O'Connor, Xianran Li, Amanda M. Gilbert, Erin Baggs, Ksenia V. Krasileva, John L. Portwood II, Ethalinda K. S. Cannon, Carson M. Andorf, Nancy Manchanda, Samantha J. Snodgrass, David E. Hufnagel, Qiuhan Jiang, Sarah Pedersen, Michael L. Syring, David A. Kudrna, Victor Llaca, Kevin Fengler, Robert J. Schmitz, Jeffrey Ross-Ibarra, Jianming Yu, Jonathan I. Gent, Candice N. Hirsch, Doreen Ware and R. Kelly Dawe

Science **373** (6555), 655-662.
DOI: 10.1126/science.abg5289

An a-maize-ing set of genomes

Maize is an important crop that is cultivated worldwide. As maize spread across the world, selection for local environments resulted in variation, but the impact on differences between the genome has not been quantified. By producing high-quality genomic sequences of the 26 lines used in the maize nested association mapping panel, Hufford *et al.* map important traits and demonstrate the diversity of maize. Examining RNA and methylation of genes across accessions, the authors identified a core set of maize genes. Beyond this core set, comparative analysis across lines identified high levels of variation in the total set of genes, the maize pan-genome. The value of this resource was further exemplified by mapping quantitative traits of interest, including those related to pathogen resistance.

Science, abg5289, this issue p. 655

ARTICLE TOOLS http://science.sciencemag.org/content/373/6555/655

SUPPLEMENTARY MATERIALS http://science.sciencemag.org/content/suppl/2021/08/04/373.6555.655.DC1

REFERENCES This article cites 173 articles, 39 of which you can access for free
http://science.sciencemag.org/content/373/6555/655#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2021, American Association for the Advancement of Science

Downloaded from http://science.sciencemag.org/ on August 5, 2021