Title

Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads.

Journal

Proceedings of the National Academy of Sciences of USA, 113(29)

Authors

Dong, Jiaqiang
Feng, Yaping
Kumar, Dibyendu
et al.

Publication Date

2016-07-19

DOI

10.1073/pnas.1608775113

Peer reviewed

Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads

Jiaqiang Dong[a], Yaping Feng[a], Dibyendu Kumar[a], Wei Zhang[a], Tingting Zhu[b], Ming-Cheng Luo[b], and Joachim Messing[a,1]

[a]Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; and [b]Department of Plant Sciences, University of California, Davis, CA 95616

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2015.

Contributed by Joachim Messing, June 1, 2016 (sent for review May 14, 2016; reviewed by Bin Han and Chad Nusbaum)

Haplotype variation not only involves SNPs but also insertions and deletions, in particular gene copy number variations. However, comparisons of individual genomes have been difficult because traditional sequencing methods give too short reads to unambiguously reconstruct chromosomal regions containing repetitive DNA sequences. An example of such a case is the protein gene family in maize that acts as a sink for reduced nitrogen in the seed. Previously, 41–48 gene copies of the alpha zein gene family that spread over six loci spanning between 30- and 500-kb chromosomal regions have been described in two Iowa Stiff Stalk (SS) inbreds. Analyses of those regions were possible because of overlapping BAC clones, generated by an expensive and labor-intensive approach. Here we used single-molecule real-time (Pacific Biosciences) shotgun sequencing to assemble the six chromosomal regions from the Non-Stiff Stalk maize inbred W22 from a single DNA sequence dataset. To validate the reconstructed regions, we developed an optical map (BioNano genome map; BioNano Genomics) of W22 and found agreement between the two datasets. Using the sequences of full-length cDNAs from W22, we found that the error rate of PacBio sequencing seemed to be less than 0.1% after autocorrection and assembly. Expressed genes, some with premature stop codons, are interspersed with non-expressed genes, giving rise to genotype-specific expression differences. Alignment of these regions with those from the previous analyzed regions of SS lines exhibits in part dramatic differences between these two heterotic groups.

shotgun DNA sequencing | haplotype variation | gene copy number | transposable elements | maize genome

Ultimately all traits are manifested in the gene content of the genome. Despite the complex nature of the regulation of mRNA synthesis and turnover, cDNA sequencing had been used as an economic approach to determine the gene content of the human genome (1). However, to achieve a better overview of all genes and their chromosomal organization, whole-genome sequencing of the human genome became essential (2). Still, the identification of all genes and their order in chromosomes of eukaryotic species has been hampered by the presence of repetitive DNA in large-size genomes. The portion of repetitive DNA in genomes is not only composed of transposable elements but also of gene families, which can vary in copy number even within the same species, generating haplotypes with changes in gene expression (3, 4).

It also has become clear that the distribution of genes and transposable elements is intermixed and one cannot sequence one or the other separately; they are contiguous in nature. This has been overcome by the construction of genomic libraries in the form of yeast and bacterial artificial chromosomes used to make physical maps before sequencing them individually (5, 6). An advantage was that such clone collections could be sequenced by large consortiums and organized as community efforts (7). A disadvantage was the lack of completion and the enormous cost. Because of the bias of restriction sites a significant portion of the genome could not be covered, and countless numbers of DNA preparations were needed, albeit use of robots comprised the bulk of the expenditures.

To reduce the cost and time to sequence a genome, the concept of whole-genome shotgun DNA sequencing had already been implemented 35 y ago and used for the sequencing of the first bacterial genome 20 y ago and the human genome 15 y ago (2, 7–9). The idea behind whole-genome shotgun DNA sequencing was that DNA fragments could be sequenced at random and then contiguous sequence information obtained by aligning overlapping sequence information (10). A critical aspect of the implementation was originally a cloning system for purification, which is now replaced by single-molecule sequencing (11). Although sequencing an entire viral genome based on restriction fragments was possible because of the lack of repetitive DNA, the original cloning system was already designed for sequencing sheared DNA instead of restriction fragments (10). Moreover, paired-end sequencing with oligonucleotide primers to overcome the presence of repetitive DNA also became available at the same time (12).

Despite the development of fluorescent capillary and, later, array solid-phase sequencing (13, 14), the length of sequencing reads became the major roadblock for resolving contiguous sequence information by overlaps. However, the recent development of the Pacific Biosciences (PacBio) single-molecule real-time (SMRT) technology permitting the generation of very long sequencing reads suggests a possible solution to this problem (15). Indeed, the human genome has been sequenced using this technology, facilitating the closure of still-existing gaps (15).

Significance

Gene copy number variation plays an important role in genome evolution and the penetrance of phenotype variations within a species. We have applied new sequencing and physical mapping strategies to obtain long chromosomal regions from a single DNA preparation in each method that comprise tandem repeated gene copies interspersed with transposable elements that comprise about 85% of the genome. This approach should reduce the time and cost to study haplotype variation of complex genomes like those from mammalian and plant species.

Author contributions: T.Z., M.-C.L., and J.M. designed research; J.D., D.K., T.Z., and M.-C.L. performed research; J.D., Y.F., D.K., W.Z., T.Z., M.-C.L., and J.M. analyzed data; and J.D., Y.F., W.Z., and J.M. wrote the paper.

Reviewers: B.H., Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences; and C.N., Broad Institute of MIT and Harvard.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database [KX247647 (z1A1), KX247648 (z1A2), KX247649 (z1B), KX247650 (z1C1), KX247651 (z1C2), and KX247652 (z1D)].

See Profile on page 7935.

[1]To whom correspondence should be addressed. Email: messing@waksman.rutgers.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1608775113/-/DCSupplemental.

Furthermore, full-length cDNAs of gene families, smaller genomes such as bacterial genomes, and a small plant genome have been sequenced this way (16–18). Real test cases, however, are those genomes that contain a very high percentage of repetitive DNA—even higher than the human genome (50%) (19)—such as that of maize (85%) (20). Although maize is smaller in total length [2.3 gigabases (Gb)] than the human genome (2.9 Gb), it has a higher proportion of retrotransposons, which are longer than the retroelements in the human genome. Indeed, the current reference genome for maize, which was based on three BAC libraries that were fingerprinted to develop a physical map (21), has a total length of pseudomolecules of 2.3 Gb and many incompletely sequenced BAC clones (20). In particular, it was difficult to sequence chromosomal regions with tandem gene clusters, which required the isolation of overlapping BAC clones different from the minimum tiling path and sequence them to a greater depth than was done for the entire genome. The single most critical parameter is the length of each sequence read to establish overlaps without the need of genomic clone libraries. Therefore, we tested the new SMRT technology to determine whether we could assemble chromosomal regions from one shotgun DNA sequencing dataset that would comprise large tandem gene copies.

An example of such chromosomal regions are the alpha zein genes that comprise the majority of seed storage proteins in maize endosperm (22). We have previously sequenced these regions in two maize inbreds, B73 and BSSS53, using overlapping BAC clones (23–25). Although we already know that maize inbreds are quite diverse, we thought that it might be quite useful to compare chromosomal regions between two different heterotic groups. Heterosis has been a major factor in increasing the yield of maize worldwide and contributed significantly to feeding the ever-increasing world population. When two maize inbreds from different heterotic groups are crossed, the hybrid produces a vigor superior to that of the two inbred parents. B73 and BSSS53 are inbreds belonging to the Iowa Stiff Stalk (SS) collection, whereas the Non-Stiff Stalk (NSS) represents a different heterotic subgroup (26). Therefore, we selected maize inbred W22, which belongs to the NSS subgroup and has been used for many genetic studies, for our analysis. In particular, many gene knockout collections have been generated in this genetic background, which will be critical for functional genomics studies of maize (27–29).

Genomic DNA from W22 was subjected to PacBio sequencing to a depth of about 40×. Because PacBio sequencing has a high but random error rate, they were corrected by alignment to yield consensus reads. After elimination of low-quality reads the corrected dataset still had genome coverage of 25×. Based on this high-quality single shotgun DNA sequencing dataset, we were able to use zein gene sequences as digital probes to assemble the entire collection of orthologous regions from W22. Furthermore, alignment with traditionally full-length sequenced cDNA resulted in 0.1% sequence deviation between the two datasets, indicating that autocorrection has worked sufficiently well to allow comparative genomic studies. To validate and improve the PacBio sequence assembly, we constructed a BioNano genome (BNG) map of maize inbred W22 using the BioNano Irys platform. Comparison of these chromosomal regions with the previously determined sequences indicates a greater degree of divergence due to retrotransposition and allelic variations between different inbred lines than expected, which provides us with new insights into the dynamic structure of the maize genome.

Results

Construction of a Genome Map and Whole-Genome Shotgun DNA Sequencing of Maize Inbred W22. Care was taken to select inbred W22 genomic DNA because even maize inbred designations could be based on diverse genotypes. However, a common source for maize genetics has been the inbred W22 from Wisconsin that we obtained from the laboratory of Hugo Dooner, Rutgers, The State University of New Jersey, so that any whole-genome restriction map would provide a unique utility for genetic research. Using the high-molecular-weight (HMW) genomic DNA of maize inbred W22, a total of 366 Gb of data (>100 kb) were obtained from 34 runs (398 unique scans), in which 726,362 raw molecules (>180 kb) corresponding to 207 Gb, representing 90× genome equivalent with an N50 length of 280 kb were used to de novo assemble a BNG map. These raw molecules were assembled into 705 BNG contigs, with N50 length of 4.43 megabases (Mb) and maximum BNG contig length of 16.2 Mb. The total BNG map assembly length is 2.1 Gb. Considering the nature of the BNG map assembly pipeline, similar repetitive regions of molecules always collapse together and produce an underestimated genome total length. We, therefore, normalized the coverage of BNX molecules to estimate the actual size of repetitive regions at a whole-genome scale (*Materials and Methods*). After normalization, the estimation of the BNG map total length reaches 2.2 Gb, which is about 96% coverage of a 2.3-Gb genome. This size is based on the estimate for the reference genome B73 (20). Although the BNG map relies, like BAC libraries, on the occurrence of specific oligonucleotide sequences in the genomic DNA, the coverage is surprisingly high and the contig length impressive, probably because it contains long stretches of retrotransposon blocks lacking such a sequence. Validation of a correct assembly should then become possible when contigs of shotgun DNA sequencing of chromosomal regions corroborate contigs of the BNG map because they constitute independent methods (discussed below) (16, 30).

The W22 genomic DNA was used for whole-genome sequencing using 128 SMRT cells on the PacBio RS II platform. We obtained 17.52 million reads with N50 length of 7.7 kb or a total of 93.47 billion bases, amounting to a depth of 40.7×, with a maximum read length of 47,824 bases. The raw reads were error-corrected using the hierarchical genome assembly process (31), resulting in a total of 58.93 billion consensus bases. Because of a large percentage of low-quality reads the N50 length dropped to 6.6 kb with a maximum read length of 46,632 bases. Still, this average read length is 26 times longer than Illumina and 8 times longer than ABI3730, providing us with significantly more contiguous information for shotgun DNA sequence assemblies.

Assembly of Chromosomal Regions Comprising Tandem Gene Copies. To test whether the self-corrected dataset with an N50 length of 6.6 kb is sufficient to achieve positional information of individual highly conserved gene copies interspersed with retrotransposons, we assembled overlapping sequences from the whole-genome shotgun DNA sequence dataset using specific gene sequences as baits. In maize, there are six loci of alpha zeins, each representing a subfamily, distributed over three chromosomes (23) (Fig. 1). We aligned the consensus PacBio reads to the known alpha zein loci sequences from both B73 and BSSS53 inbreds (23–25) using BLASTN. The mapped PacBio reads were separated using an empirical combination of percent identity and alignment length (Fig. S1). For the length of 3 kb, we chose 97% identity, for 2 kb 98%, and for 1 kb 99% because these three combinations led to coverage of around 20×–100×, which was close to the whole-genome coverage of ∼25×. Based on these collections, we formed sequence libraries that were locus-specific, reducing the computational footprint and time for the assembly of sequences. Celera Assembler 8.2 was applied to assemble individual zein gene clusters from the selected self-corrected PacBio reads as described in *Materials and Methods*.

Then, we assembled six long scaffolds comprising the larger zein gene clusters *z1A1*, *z1A2*, *z1B*, *z1C1*, *z1C2*, and *z1D* ranging from 143 kb to 601 kb in size (Fig. 2 and Fig. S2). The assembly strategies are shown in the flowchart of Fig. S3. More than 20,000 contigs can be mapped onto the zein-targeted BNG contigs, when using loose RefAlinger parameters. However, we

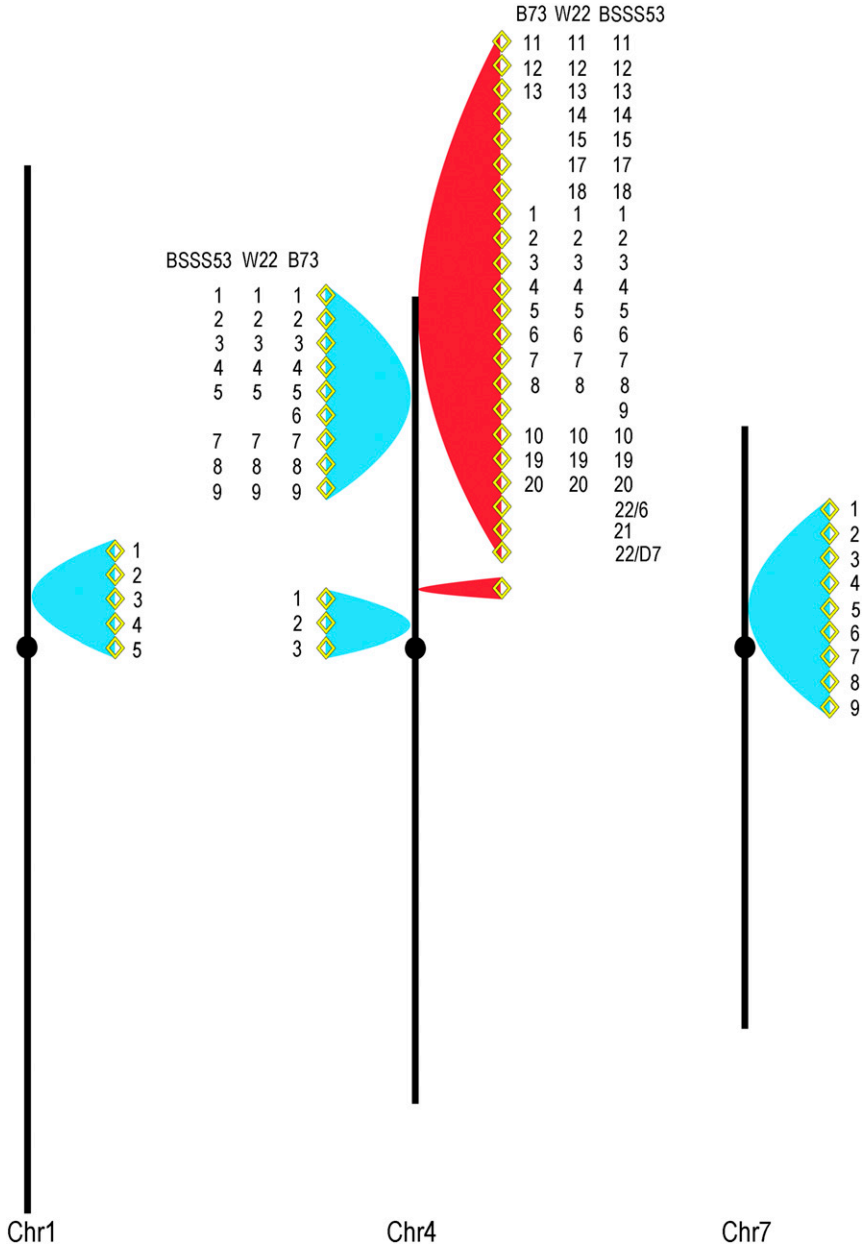

Fig. 1. Genomic distribution of alpha zein loci in three maize inbred lines. Zein gene copies at each locus in the genome are presented as yellow diamonds on a blue (19-kDa clusters) or red (22-kDa clusters) background. When copy number differs between three inbreds the zeins are numbered accordingly. Vertical bars represent maize chromosomes, from left to right, chromosome 1, chromosome 4, and chromosome 7.

only selected three to six supporting contigs. The criteria we used to select supporting contigs were as follows: (*i*) contigs containing zein or flanking genes or conserved retrotransposition events were given priority, (*ii*) contigs had to meet a confidence score of 4, and (*iii*) contigs were also selected to fill gaps (Fig. 3). The confidence scores of all mapped contigs are shown in boxplots, and supporting contigs are highlighted in red (Fig. S4). The corroboration of the genome map with our larger PacBio scaffolds is only possible if overlapping sequence information of shotgun reads is accurate enough for the assembly to work.

Gene Copy Number Variation Between W22 (NSS) and B73/BSSS53 (SS). Based on the assembly and on conserved flanking nonzein genes or conserved transposable elements in all three inbreds, we can account for all alpha zein gene copies in the NSS inbred W22. On maize chromosome 4S there are the *z1A1* and *z1A2* 19-kDa zein loci, on chromosome 7S the *z1B* 19-kDa zein locus, on chromosome 4S the *z1C1* and *z1C2* 22-kDa loci, and on chromosome 1S the *z1D* 19-kDa locus (Fig. 1). The *z1A1* scaffold is 337 kb long and collinear with the BNG map (Figs. 2*A* and 3). It contains eight copies of the 19-kDa zein genes. Inbred BSSS53 has the same copy number as W22 in this region, but B73 has one more copy. It was noted before that in B73 the *z1A1-5* gene copy duplicated 0.5 Mya to produce a new copy, *z1A1-6* (23). This duplication is missing in both BSSS53 and W22. The *z1A2* scaffold is 187 kb long and collinear with the BNG map (Fig. S2 and Fig. 3). All three inbred lines have three gene copies at the *z1A2* location (Fig. S2).

The W22 *z1B* scaffold is 444 kb long and is also collinear with the BNG map (Figs. 2*B* and 3). It contains nine copies of 19-kDa zein genes (Fig. 2*B*). There is no copy number variation (CNV) of *z1B* genes between these three inbreds. Allelic *z1B* gene copies

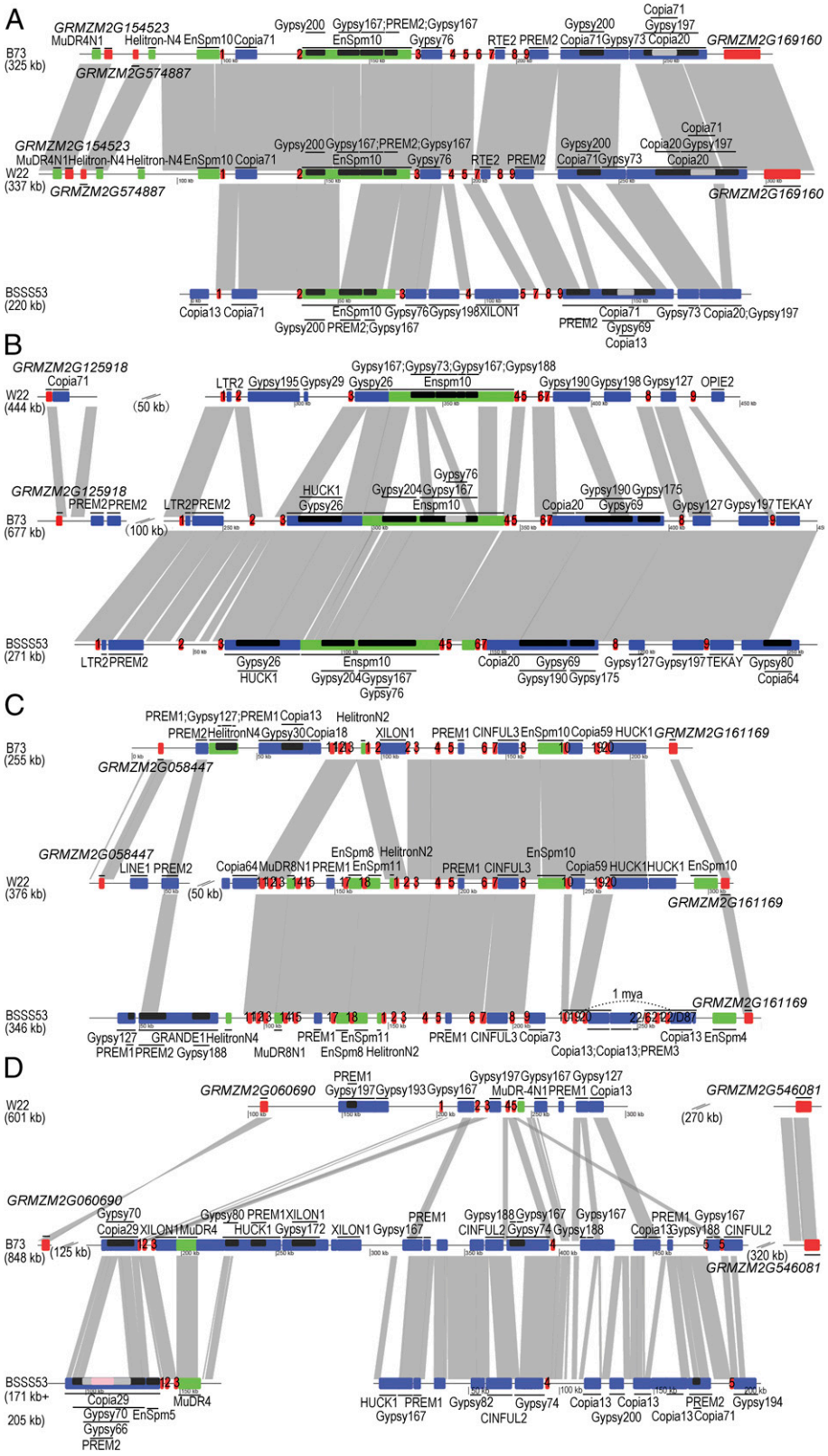

Fig. 2. Haplotype variability at four larger alpha zein loci. Zein genes are numbered from left to right, as red bars. Sequence conservation between these inbreds is represented as vertical gray lines. DNA TEs are depicted in green blocks, whereas REs are represented by blue blocks. Insertion either in DNA TEs or REs are marked as black side blocks. Nesting of REs is illustrated as gray and pink side blocks. (*A*) Haplotype variability of *z1A1* locus includes gene CNV. (*B*) The *z1B* locus of hte NSS line W22 differs from SS lines B73 and BSSS53, whereas the latter two are similar. (*C*) The W22 *z1C1* locus is a recombination of the other two haplotypes. (*D*) The *z1D* locus is the oldest one that suffers from many transposable elements insertions.

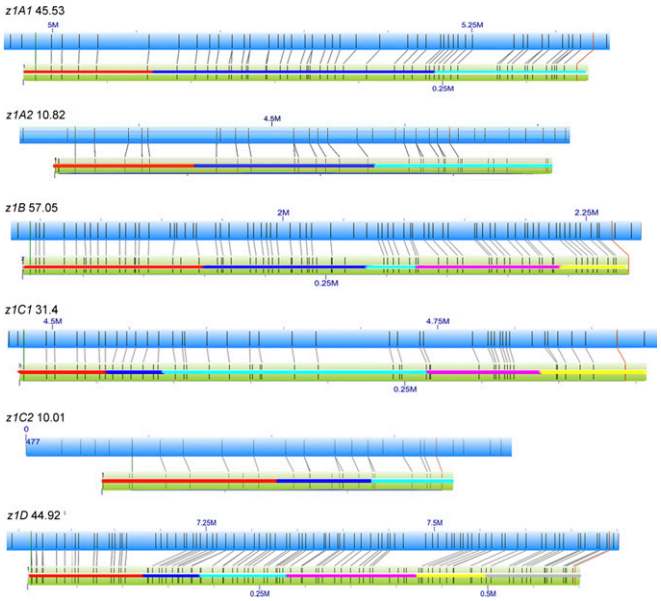

Fig. 3. Alignment of BioNano contigs with assembled PacBio scaffolds. BNG contigs are used as reference (blue bar), with which the scaffolds (green bars) are aligned. The black lines inside green and blues bars are the GCTCTTC sequences recognized by nickase Nt.BspQI. The colored lines on green bars represent supporting contigs for the assemblies. Junctions between colored bars can introduce shifts in the alignments because of gaps in the scaffolds. Contigs are chosen by an empirical confidence score cutoff. For instance, the cyan and yellow contigs contain *z1B* zein gene copies (third row). Because these contigs are rather short, each of them has a rather low score, and the threshold has been set as 4. However, they are contiguous because both contigs contain *z1B* zein gene copies in the right order. Therefore, the score of the scaffolds is much higher.

in B73 and BSSS53 are 100% identical with only two intact copies, *z1B-4* and *z1B-6*. However, in W22 there are three more intact copies besides the two copies found in the SS inbreds (Table S1). The difference is based on premature stop codons in B73, not found in W22. It should be noted that zein genes have a preponderance of glutamine codons, which can become stop codons with single C-to-T transitions.

The W22 *z1C1* scaffold is 376 kb long and is also collinear with the BNG map (Figs. 2*C* and 3). It contains 18 copies of zein genes that encode slightly larger proteins, 22 kDa versus 19 kDa, with an internal repeat of a block of 32 aa residues (32). Interestingly, the W22 *z1C1* locus seems to have undergone recombination between the other two haplotypes. The 5′ terminal region of the W22 *z1C1* locus is similar to that of BSSS53, having four more copies of *z1C1* genes than B73. The 3′ terminal of W22 *z1C1* locus is similar to that of B73, missing the *z1C1-9* gene copy that is present in BSSS53. There was also a recent segmental duplication of 20 kb at the 3′ terminal region of the *z1C1* locus in BSSS53 about 1 Mya, leading to an increase of three copies of *z1C1* genes in BSSS53, but not in B73 and W22. Besides the CNV, insertions also make a difference. There is a 697-bp insertion in the BSSS53 *z1C1-7* allele, resulting in a premature stop codon. Furthermore, the *z1C1-2* allele in B73 is interrupted with an 11-kb-long fragmented retrotransposon, XILON1, absent in W22 and BSSS53. The W22 *z1C2* scaffold is 143 kb long and is also collinear with the BNG map (Fig. S2 and Fig. 3). All three inbreds have a single gene copy at the *z1C2* location (Fig. S2).

It has been proposed that the younger ones arose through dispersed amplification (33). The *z1D* locus is the oldest alpha zein locus in maize. To illustrate the chronology of amplification at each locus, phylogenetic trees were constructed (Fig. S5). The W22 *z1D* represents the longest scaffold with 601 kb in our collection and is also supported by the BNG map (Figs. 2*D* and 3). All three inbreds have five copies of *z1D* genes (Fig. 2*D*). The W22 *z1D-3* is truncated and is only 269 bp long, whereas this allele in B73 has a normal length of 723 bp with a premature stop codon. Furthermore, the B73 *z1D-5* allele is interrupted with Gypsy188 and Gypsy167 elements in B73, but not in W22 and BSSS53. The five copies of *z1D* in W22 are clustered in a region about 50 kb long, but to link them to flanking nonzein genes we assembled a 12-fold larger scaffold. These flanking genes are not present in the previously sequenced BAC clones and separated from the zein cluster by large intergenic regions. In B73 and BSSS53, there are only three zein copies clustered together, whereas the other two members are dispersed over a larger region of about 200 kb. Because the W22 chromosomal region differed so strongly from B73 and BSSS53, we used suitable primer pairs to selectively amplify W22 fragments to validate their sizes by gel electrophoresis (Fig. S6 and Table S2), confirming contiguity of the W22 assembled *z1D* locus. Clearly this region has undergone the most dramatic changes and represents the most divergent haplotypes between SS and NSS inbreds.

Genotype-Specific Gene Expression. Previously, normalized full-length cDNAs of W22 have been sequenced with capillary electrophoresis (34). Because alpha zein gene copies do not have introns and each copy is sufficiently diverged from the others, we can directly align full-length cDNAs with the gene copies they are derived from. These alignments can provide us with an opportunity to calculate the error rates of PacBio consensus sequences at the single-nucleotide resolution. In total, there are 17 1-bp gaps and 12 mismatches found in all coding regions of the expressed alpha zein genes (Table S1). Thus, there is only about 0.1% single nucleotide difference between sequenced cDNAs and the genes from assemblies with autocorrected PacBio read sequences. Moreover, genomic sequences that differ from their corresponding full-length cDNA sequence can be easily corrected manually.

Another important use of the cDNA information is the determination of which zein gene copy is transcribed. Our results show that the different inbreds not only exhibit CNV but also genotype-specific expression patterns. In B73, there are seven *z1B* genes expressed, *z1B-1*, *-2*, *-3*, *-4*, *-5*, *-6*, and *-9* (23, 35). In W22, the number of expressed *z1B* genes is the same. However, *z1B-5* is not expressed in W22, whereas *z1B-8* is. The BSSS53 *z1C1-22/6* and *z1C1-22/D87* gene copies are expressed even in the absence of Opaque2, whereas in B73 and in W22 all 22-kDa alpha zeins are blocked in the *o2* background (24).

It was reported that older copies became damaged and new copies account for the majority of the expression of zein genes (23). The older copies at the W22 *z1A1* locus (i.e., *z1A1-1*, *-8*, and *-9*) are not expressed because of a premature stop codon (*z1A1-1*), a missing A at the start codon (*z1A1-8*), or severely truncated 3′ end (*z1A1-9* is only 340 bp long) in three inbred lines. Still, in some cases, we can find transcripts with a premature stop codon such as B73 *z1B-1*, *-3*, and *-9* (23), although their expression is lower than the intact genes. Interestingly, the *z1C2* locus is the only locus with a single copy gene, which is expressed in all three inbreds and also known for its allele of the semidominant *floury2* mutation (24).

Transposable Elements. Tandem amplification of zein gene copies at each locus was followed by the insertion of transposable elements, resulting in the divergence of each locus (23). The percentage of retrotransposon at the BSSS53 *z1A1* locus is 70%, whereas the percentage of transposon or retrotransposon space in B73 is similar to that of W22 (55%) (Table 1). However, transposon types and insertion patterns at the *z1A1* locus are different between W22 and BSSS53 (Fig. 2*A*). The RTE2 retrotransposon is inserted between the *z1A1-7* and *z1A1-8* gene

Table 1. Haplotype variability at the α-zein gene loci

Gene cluster	Inbred	REs, %	TEs, %	Genic, %	Conserved,* %
z1A1	B73	55	13	6	B73vsW22 94 86
	W22	55	13	6	W22vsBSSS53 69 67
	BSSS53	70	8	5	
z1A2	B73	73	2	2	W22vsB73 93 67
	W22	63	2	7	B73vsBSSS53 94 96
	BSSS53	69	1	3	
z1B	W22	81	3	4	W22vsB73 61 40
	B73	79	4	3	B73vsBSSS53 99 94
	BSSS53	78	5	3	
z1C1	B73	67	2	10	B73vsW22 55 42
	W22	56	16	12	W22vsBSSS53 51 55
	BSSS53	63	10	10	
z1C2	W22	68	12	2	W22vsB73 13 13
	B73	83	2	3	B73vsBSSS53 96 99
	BSSS53	70	2	8	
z1D	W22	73	4	7	W22vsB73 36 39
	B73	79	3	1	B73vsBSSS53 51 58
	BSSS53	73	9	1	

*"Conserved" means the proportion of conserved sequence in its corresponding region.

copies in W22 but not in BSSS53. There are also unique insertions of Gypsy198 retrotransposon between *z1A1-3* and *z1A1-4* genes and of XILON1 retrotransposon between *z1A1-4* and *z1A1-5* genes in BSSS53. The insertions of transposable elements at *z1A2* locus are quite similar between all three inbreds, except that there are several unique insertions in B73 compared with W22 (Fig. S2). Gypsy168 is the oldest insertion present in all three inbreds, dating back about 70 Mya (Table S3).

The *z1B* locus in the W22 inbred differs from that in the other two SS inbreds, whereas the latter two inbreds are quite similar. Although the percentages of retrotransposon and transposon are similar among all three inbreds (Table 1), the contents are quite different between W22 and B73 or BSSS53 (Fig. 2*B*). A major difference of B73 and BSSS53 *z1B* locus is a recent insertion of Copia64 retrotransposon in BSSS53 0.39 Mya (Table S3). However, there are a lot of unique insertions of transposons and retrotransposons between W22 and B73 (Table S3). Most of the differences are produced less than 3.5 Mya, later than the allotetraploidization of maize 4.8 Mya (36).

The W22 *z1C1* locus represents the recombination of the other two haplotypes (Fig. 2*C*). The percentage of retrotransposon in the *z1C1* locus is higher in B73 (67%) and BSSS53 (63%) compared with that in W22 (56%), but that of transposons is much higher in W22 (16%) than in the other two inbreds (Table 1). The W22 *z1C2* locus is quite different from the SS group with a number of unique insertions (Fig. S2).

Compared with the compactly clustered W22 *z1D* locus, there are many insertions in the intergenic regions in B73 and BSSS53 *z1D* loci (Fig. 2*D*). The percentage of retrotransposon in *z1D* locus is the highest in B73 (79%) and that of transposons is the highest in BSSS53 (9%) (Table 1). There are also some nested insertions that determine each haplotype (Table S3). For example, in BSSS53, retrotransposon PREM2 was inserted in a Gypsy66 retrotransposon 0.85 Mya, which was inserted in a Gypsy70 retrotransposon 2.54 Mya. This Gypsy70 retrotransposon was inserted in another retrotransposon, Copia29, 7.54 Mya; Copia29 was inserted into the maize genome 7.92 Mya. However, B73 does not have the two recent insertions that arose after allotetraploidization (36).

Besides the tandem insertions, duplication is another source of differences in orthologous regions. For example, in BSSS53, the recent duplication of the 20-kb segment at the 3′ terminal of *z1C1* locus also introduced another copy of Copia13 and LTR1 retrotransposons about 1 Mya (Fig. 2*C*).

Discussion

Genome Analysis Strategies. Maize is one of the most important crops worldwide, but it has been difficult to study the diversity of the species because of the dynamic structure of the genome. There have been two approaches to study its diversity. Survey short-read sequencing has been performed on a large number of inbred lines to achieve a global overview of genome structural variation (37). BAC libraries have been constructed and DNA fingerprinting has been used to reconstruct chromosomal regions with overlaps so that the sequences of these regions could be obtained from sequencing these BAC clones (21, 38). The latter approach leads to a more complete assembled product than the first one, but it carries a higher cost and more time.

Although the association of phenotypic variation among maize inbred lines with sequenced tagged markers has immense value in genome-wide association studies, the approach cannot replace the in-depth analysis of the chromosomal organization of traits like protein quality. Furthermore, analyses of haplotype variations of interwoven gene and repeat elements require the placement of repeated sequences in a location-specific manner. Here, we have taken the approach to analyze chromosomal regions of the maize genome with a focus on seed protein composition, which affects the nutritional quality of about 100 million tons of protein per year in worldwide production.

Reconstruction of Chromosomal Regions. We have used the new PacBio platform to sequence the maize inbred W22 by a single shotgun sequencing experiment because of its long read capabilities. To prove the value of this approach, we needed an independent method to test the assembly of overlapping sequence information. Here, we were able to take advantage of a new whole-genome restriction mapping method (39). The construction of a BNG map requires the isolation of HMW genomic DNA so that highly uniform linearization of DNA in nanochannels can be achieved (39). These molecules are then treated with the enzyme Nt.BspQI, which introduces nicks into the double-stranded DNA at the sequence GCTCTTC. These nicks are then repaired with a fluorescent tag. This nanochannel-based genome mapping technology, commercialized as the Irys platform, automatically images fluorescently labeled DNA molecules in this massively parallel nanochannel array. Distances between tags represent the equivalent of restriction fragments of defined sizes. The resulting restriction fragment patterns then allow us to reconstruct overlapping information and extend DNA molecules to their larger sizes. Moreover, in contrast to the DNA fingerprinting of BACs, only one DNA preparation is required, saving enormous costs in reagents and time. One limitation of this method is that it does not provide high-density information and therefore does not apply to rather small contigs. Furthermore, alignment shifts can occur when gap sizes of sequence contigs are unknown during scaffold building. Besides alignment shifts, a few restriction sites could be missing either due to incomplete digestion or the low sequencing error rate left after autocorrection. However, the larger scaffolds of PacBio assembled sequences and the genome map align very well as shown in Fig. 3. Such alignments would not work if either sequencing errors or lower stringency in the assembly algorithm would join DNA sequence reads incorrectly. Such a concern is based on the high error rate inherent in the PacBio sequencing system.

To correct the random errors, each base had to be sequenced multiple times to distinguish related sequences differing by SNP from random errors. Here, we used a redundancy of about 40-fold of the predicted genome size, which is a compromise between sequence depth and computing time that required up to 30 times 32 cores, equivalent to running over 900 processes for

almost 1 mo, totaling 192,522,542.21 central processing unit hours. Greater redundancy of sequence reads could be critical for total genome assembly but in our case was sufficient for the assembly of all chromosomal regions that were investigated. The resulting consensus sequences proved to be of high quality, although the redundancy dropped considerably to 25-fold genome coverage. Still, we were able to achieve assemblies of a gene-poor region such as the *z1D* locus on chromosome 1S with over 600 kb that could be validated by the BNG map, indicating that the depth could overcome the abundance of retrotransposon elements in the genome to concatenate the PacBio reads in a robust fashion. Moreover, we could use the alignments of orthologous regions from two other inbred lines, B73 and BSSS53, and the full-length cDNA sequences from W22 to validate the autocorrection of PacBio sequences. Therefore, the current PacBio platform should prove to be an advanced approach for the study of contiguous chromosomal regions of complex genomes.

Haplotype Variation. The in-depth comparison of the chromosomal regions of three maize inbreds exhibits diversities at each locus and might provide us with what we can expect from a larger whole-genome comparison. We should note that we could align larger regions of W22 with B73 because only a small region from BSSS53 was sequenced for each locus due to the cost of chromosome walking. The examples investigated here represent regions formed prior and after allotetraploidization. A case of distinction is the *z1B* locus, which is derived from one of the progenitors, whereas the *z1D* locus is derived from both and can be considered the one that the *z1B* locus is derived from. Previously, it has been reported that homologous regions have been under pressure to rearrange after polyploidization (38). Therefore, one might hypothesize that regions not duplicated due to allotetraploidization could be more stable. This seemed to be the case, when *z1B* from the SS lines B73 and BSSS53 was compared (23). However, comparison with W22 exhibits strong differences between the SS and the NSS lines mainly due to retrotransposition events. Although the orthologous regions do not exhibit any CNV, they differ in the number of genes that are transcribed between SS and NSS.

The region with most CNVs is the cluster with the 22-kDa alpha zein genes, the *z1C1* locus, on chromosome 4S. The number in all three lines is different. There are 14 copies in B73, 22 in BSSS53, and 18 in W22. The most recent duplicated copies are only present in BSSS53 at the 3′ end of the region. Interestingly, these copies switched their transcriptional regulation. The *z1C1-22/6* and *z1C1-22/D87* gene copies are expressed even in the absence of Opaque2 (24). Whereas in B73 and in W22 all 22 kDa alpha zeins are blocked in the *o2* background, there is still 22-kDa alpha zein gene expression in BSSS53, illustrating how CNV can affect the penetration of a phenotype. The comparison between the three inbreds also illustrates how recombination can generate new haplotypes with exchanges of blocks of alleles at the same time. Another case of recombination illustrates how a single allele can be duplicated in one haplotype but not the other two, as shown for the *z1A1* locus.

The oldest alpha zein gene cluster is the *z1D* locus on chromosome 1S, which is rather a gene-poor region as illustrated by the lack of flanking genes in close vicinity. The next upstream collinear gene was found at a distance of about 95 kb upstream of *z1D-1* in W22 and about 158 kb upstream of *z1D-1* in B73. The next downstream collinear gene in W22 and B73 is located about 320 kb away. However, the five gene copies in W22 are rather closely linked, whereas they are spread far apart in both B73 and BSSS53.

Conclusions

In this pilot study we demonstrated an effective method that has general utility for resolution of large complex repeats or tandem/dispersed gene family clusters. Here, by reconstructing zein clusters in maize inbred W22, we showed that the degree of increase in size in intergenic regions is highly variable and that distances can significantly enlarge in chromosomal regions, even in regions containing tandem gene clusters, although we did not observe any clear trend among SS and NSS subgroups because all three inbreds have unique zein gene cluster haplotypes. It seems likely that this range variability could be observed in zein clusters in other inbred lines, and in fact in other regions of the genome as well. Based on the results presented here, it should be straightforward to generate a whole-genome assembly of W22 from Pacific Biosciences and BioNano Genomics data given sufficient computational time and suitable assembly parameters. Finally, given the effectiveness of this approach in maize, we anticipate that it will be of general use with any complex genome including human and, in particular, cancer genomics, where structural changes can be dramatic.

Materials and Methods

DNA Sample Preparation for BioNano Genome Map Construction. The maize inbred line W22 was obtained from Hugo Dooner, Rutgers, The State University of New Jersey, and grown and self-crossed during the summer of 2014 in the Waksman Institute field space. HMW DNA was isolated from young leaves (grown in the dark) of maize inbred line W22 by Amplicon Express. The nicking endonuclease Nt.BspQ1 (New England BioLabs) was chosen to label high-quality HMW DNA molecules at specific sequence motifs (GCTCTTC) based on sequences of the publicly available maize inbred B73 genome (20). The nicked DNA molecules were then stained according to the instructions of IrysPrep Reagent Kit (BioNano Genomics).

De Novo BioNano Genome Map Assembly and Analysis. The stained DNA sample was loaded onto the nanochannel array of IrysChip (BioNano Genomics) and was automatically imaged by the Irys system (BioNano Genomics). Raw DNA molecules >100 kb were collected and converted into BNX files by AutoDetect software to obtain basic labeling and DNA length information. The filtered raw DNA molecules in BNX format were aligned, clustered, and assembled into the BNG map by using the BioNano Genomics assembly pipeline as described in previous publications (40, 41). The *P* value thresholds used for pairwise assembly, extension/refinement, and final refinement stages were 1×10^{-9}, 1×10^{-10}, and 1×10^{-10}, respectively. The initial BNG map was then checked for potential chimeric BNG contigs and was further refined. To compare the draft sequence assembly with the BNG map, sequences were digested in silico according to the restriction site of Nt.BspQ1 by using Knickers (BioNano Genomics). The alignment of sequence assemblies with the BNG map was computed with RefAligner, and the visualization of the alignment was performed with snapshot in IrysView. Software and packages used can be obtained from BioNano Genomics (bionanogenomics.com/support/software-updates/).

PacBio Sequencing. Maize inbred line W22 seeds were germinated in the greenhouse and young leaflets were used to prepare HMW genomic DNA for PacBio sequencing. Around 20 μg of high-quality genomic DNA was used to prepare the SMRTbell library. To select 20-kb double-stranded DNA fragments Bluepippin was used. The library was sequenced on PacBio RS II and processed with instrument software 2.3.

Autocorrection of PacBio Sequences. PacBio raw reads of the sequencing run were subjected to autocorrection to eliminate random sequencing errors. Two steps were used to overcome high base call errors in individual PacBio reads. In the first step, PacBio reads were preprocessed with SMRT portal to filter reads based on read length (500) and quality (80). In the second step, the overlapping algorithm reference implementation MHAP from PBcR v8.2 (wgs-assembler.sourceforge.net/wiki/index.php/PBcR) was used with "-sensitive" setting recommended for <50× coverage to self-corrected PacBio reads.

Assembly of Chromosomal Regions. Reference genomes of maize inbred B73 and BSSS53 were used to sort W22 specific zein cluster sequences and one to three collinear gene sequences flanking the zein cluster based on BLAST

alignment results. Reads belonging to each cluster are assembled separately using Celera Assembler 8.2. Assembled contigs containing zein genes and flanking genes were mapped onto the BNG map using RefAligner. To accurately locate zein gene clusters, stringent parameters of RefAlinger were chosen as follows: -T 1e-10 -endoutlier 1e-3 -outlier 1e-4 -biaswt 0 -sd 0.1 -res 2.9 -sf 0.2 -stdout -stderr -hashgen 5 3 2.4 1.5 0.05 5.0 1 1 1 -hash -hashdelta 50.

Then, BNG contig targets, <1Mb in length, were selected as reference for the next step of mapping. Because contigs of PacBio reads assembly are relatively short with N50 length of about 20 kb, the BNG contig target regions were usually sparsely covered when using stringent mapping parameters of RefAligner. Therefore, parameters of RefAligner were relaxed as follows: -T 1e-2 -endoutlier 0.5 -outlier 0.5 -biaswt 1 -sd 0.4 -res 1.9 -sf 0.5 -stdout -stderr -hashgen 5 1 3 2.0 0.1 5.0 2 2 4 -hash -hashdelta 50 -FP 2 -FN 0.3.

After this step of mapping, the contigs were selected, orientated, and linked. These raw assemblies were then mapped back onto BNG contigs using stringent RefAligner parameters. Only if then the full sequences could map to BNG contigs using stringent RefAligner parameters were they considered robust. Then, these assemblies were further improved, bridging gaps and merging overlapping regions. Sequences of all chromosomal regions have been deposited with GenBank under the following accession numbers: z1A1 as KX247647, z1A2 as KX247648, z1B as KX247649, z1C1 as KX247650, z1C2 as KX247651, and z1D as KX247652.

Annotation and Alignments of Allelic Regions. The assembled sequences were analyzed using various open source softwares available online: RepeatMasker for transposon (TE) searches (42), ExPASy to translate into protein (43), and the BLAST suite from the National Center for Biotechnology Information for homology searches and sequence comparison between pairs of inbred lines. Sequences were then manually annotated, and then haplotypes were aligned in pairwise comparisons. A threshold of 95% homology was set when comparing each two sequences. The annotation and alignment were subjected to an R package, genoPlotR, to graphically display a comparison alignment (44).

Molecular Clock of Retrotranspositions. To estimate the insertion time for the retrotransposons (REs), the left and right LTR sequences were aligned first with ClustalW software (45). To avoid any bias, any indels or sequencing gaps from the alignment were removed using Unipro UGENE software (46). Then the alignment was processed with MEGA4 software (47) to calculate the nucleotide synonymous substitution rates (Ks values). Default settings were changed to "Distance and Std. Err.," "Pairwise deletions," and "Kimura 2-parameter." The Ks values reported for LTRs were used to calculate the insertion time (48).

ACKNOWLEDGMENTS. This work was supported by the Selman Waksman Chair in Molecular Genetics (J.M.).

1. Adams MD, et al. (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252(5013):1651–1656.
2. Venter JC, et al. (2001) The sequence of the human genome. *Science* 291(5507): 1304–1351.
3. Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc Natl Acad Sci USA* 100(15):9055–9060.
4. Goettel W, Messing J (2009) Change of gene structure and function by non-homologous end-joining, homologous recombination, and transposition of DNA. *PLoS Genet* 5(6):e1000516.
5. Burke DT, Carle GF, Olson MV (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 236(4803):806–812.
6. Kim UJ, et al. (1996) A bacterial artificial chromosome-based framework contig map of human chromosome 22q. *Proc Natl Acad Sci USA* 93(13):6297–6301.
7. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
8. Gardner RC, et al. (1981) The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res* 9(12): 2871–2888.
9. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496–512.
10. Messing J, Crea R, Seeburg PH (1981) A system for shotgun DNA sequencing. *Nucleic Acids Res* 9(2):309–321.
11. Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci USA* 100(7):3960–3964.
12. Vieira J, Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* 19(3):259–268.
13. Smith LM, et al. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321(6071):674–679.
14. Sindelar LE, Jaklevic JM (1995) High-throughput DNA synthesis in a multichannel format. *Nucleic Acids Res* 23(6):982–987.
15. Clark TA, et al. (2012) Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res* 40(4):e29.
16. VanBuren R, et al. (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature* 527(7579):508–511.
17. Loman NJ, Pallen MJ (2015) Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* 13(12):787–794.
18. Zhang W, Ciclitira P, Messing J (2014) PacBio sequencing of gene families - A case study with wheat gluten genes. *Gene* 533(2):541–546.
19. Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet* 13(1):36–46.
20. Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.
21. Messing J, et al. (2004) Sequence composition and genome organization of maize. *Proc Natl Acad Sci USA* 101(40):14349–14354.
22. Wu Y, Messing J (2014) Proteome balancing of the maize seed for higher nutritional value. *Front Plant Sci* 5:240.
23. Miclaus M, Xu J-H, Messing J (2011) Differential gene expression and epiregulation of alpha zein gene copies in maize haplotypes. *PLoS Genet* 7(6):e1002131.
24. Song R, Llaca V, Linton E, Messing J (2001) Sequence, regulation, and evolution of the maize 22-kD α zein gene family. *Genome Res* 11(11):1817–1825.
25. Song R, Messing J (2002) Contiguous genomic DNA sequence comprising the 19-kD zein gene family from maize. *Plant Physiol* 130(4):1626–1635.
26. Liu K, et al. (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165(4):2117–2128.
27. Li Y, Segal G, Wang Q, Dooner HK (2013) Gene tagging with engineered Ds elements in maize. *Methods Mol Biol* 1057:83–99.
28. McCarty DR, et al. (2005) Steady-state transposon mutagenesis in inbred maize. *Plant J* 44(1):52–61.
29. Kolkman JM, et al. (2005) Distribution of Activator (Ac) throughout the maize genome for use in regional mutagenesis. *Genetics* 169(2):981–995.
30. Hastie AR, et al. (2013) Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. *PLoS One* 8(2):e55864.
31. Chin CS, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10(6):563–569.
32. Heidecker G, Messing J (1983) Sequence analysis of zein cDNAs obtained by an efficient mRNA cloning method. *Nucleic Acids Res* 11(14):4891–4906.
33. Xu J-H, Messing J (2008) Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. *Proc Natl Acad Sci USA* 105(38):14330–14335.
34. Lai J, et al. (2004) Characterization of the maize endosperm transcriptome and its comparison to the rice genome. *Genome Res* 14(10A):1932–1937.
35. Chen J, et al. (2014) Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol* 166(1):252–264.
36. Swigonová Z, et al. (2004) Close split of sorghum and maize genome progenitors. *Genome Res* 14(10A):1916–1923.
37. Lu F, et al. (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun* 6:6914.
38. Bruggmann R, et al. (2006) Uneven chromosome contraction and expansion in the maize genome. *Genome Res* 16(10):1241–1251.
39. Das SK, et al. (2010) Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res* 38(18):e177.
40. Cao H, et al. (2014) Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* 3(1):34.
41. Lam ET, et al. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* 30(8):771–776.
42. Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6:11.
43. Gasteiger E, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 31(13):3784–3788.
44. Guy L, Kultima JR, Andersson SG (2010) genoPlotR: Comparative gene and genome visualization in R. *Bioinformatics* 26(18):2334–2335.
45. Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chap 2:Unit 2.3.
46. Okonechnikov K, Golosova O, Fursov M; UGENE team (2012) Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* 28(8):1166–1167.
47. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24(8):1596–1599.
48. Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci USA* 101(34):12404–12410.