AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication

Baoxing Song[a,1], Santiago Marco-Sola[b,c], Miquel Moreto[b,d], Lynn Johnson[a], Edward S. Buckler[a,e,f,1], and Michelle C. Stitzer[a,g,1]

[a]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853; [b]Department of Computer Sciences, Barcelona Supercomputing Center, Barcelona 08034, Spain; [c]Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona 08193, Spain; [d]Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona 08034, Spain; [e]Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853; [f]Agricultural Research Service, US Department of Agriculture, Ithaca, NY 14853; and [g]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

Edited by Douglas Soltis, University of Florida, Gainesville, FL; received July 23, 2021; accepted November 15, 2021

Millions of species are currently being sequenced, and their genomes are being compared. Many of them have more complex genomes than model systems and raise novel challenges for genome alignment. Widely used local alignment strategies often produce limited or incongruous results when applied to genomes with dispersed repeats, long indels, and highly diverse sequences. Moreover, alignment using many-to-many or reciprocal best hit approaches conflicts with well-studied patterns between species with different rounds of whole-genome duplication. Here, we introduce Anchored Wavefront alignment (AnchorWave), which performs whole-genome duplication–informed collinear anchor identification between genomes and performs base pair–resolved global alignment for collinear blocks using a two-piece affine gap cost strategy. This strategy enables AnchorWave to precisely identify multikilobase indels generated by transposable element (TE) presence/absence variants (PAVs). When aligning two maize genomes, AnchorWave successfully recalled 87% of previously reported TE PAVs. By contrast, other genome alignment tools showed low power for TE PAV recall. AnchorWave precisely aligns up to three times more of the genome as position matches or indels than the closest competitive approach when comparing diverse genomes. Moreover, AnchorWave recalls transcription factor–binding sites at a rate of 1.05- to 74.85-fold higher than other tools with significantly lower false-positive alignments. AnchorWave complements available genome alignment tools by showing obvious improvement when applied to genomes with dispersed repeats, active TEs, high sequence diversity, and whole-genome duplication variation.

sensitive genome alignment | whole-genome duplication | genome comparison | transposable element variation | regulatory element alignment

Genome alignment tools are fundamental for comparative evolutionary analysis. Unlike initial genome sequencing efforts, which concentrated on cost-effective sequencing of model species, fulfilling the goal of sequencing a million eukaryotic reference genomes (1) adds many species with larger or repeat-rich genomes (2). Aligning those genomes provides a revolutionary opportunity to understand the evolution of eukaryotic genomes. Pipelines that have successfully aligned genomes of model species often do not work well among genomes with many complex variants, especially plant species. Although the alignment of genic regions is bolstered by their modest length and conservation of amino acid residues, genes comprise only a minority of the nucleotides in a genome. Distal regulatory regions can be recalcitrant to alignment, often pushed far away from the genes they regulate by transposable element (TE) insertions (3, 4). In addition, recursive whole-genome duplications (WGD) result in fractionation of duplicated copies by deletion or pseudogenization (5) and chromosomal rearrangement, further complicating genome alignments.

Seed-and-extend local alignment strategies (6, 7) have been widely successful for comparing model genomes. Such strategies generally trigger alignments with pairs of highly similar *k*-mers (seeds) from two genomes. This strategy can trigger false alignments when aligning genomes with many dispersed repeats or even fail to generate an alignment when repetitive genome sequences are masked. The seed-and-extend strategy often fails when aligning regulatory elements with essential functions, even though those sequences are expected to be conserved between species. For example, the core motifs of transcription factor–binding sites (TFBSs) (8–10) are 6.8 base pairs (bp) on average— much shorter than the seed size used in genome alignment. Furthermore, the presence of highly diverse fragments can limit alignment extension (6) and confound local

Significance

One fundamental analysis needed to interpret genome assemblies is genome alignment. Yet, accurately aligning regulatory and transposon regions outside of genes remains challenging. We introduce Anchored Wavefront alignment (AnchorWave), which implements a genome duplication informed longest path algorithm to identify collinear regions and performs base pair–resolved, end-to-end alignment for collinear blocks using an efficient two-piece affine gap cost strategy. AnchorWave improves the alignment under a number of scenarios: genomes with high similarity, large genomes with high transposable element activity, genomes with many inversions, and alignments between species with deeper evolutionary divergence and different whole-genome duplication histories. Potential use cases include genome comparison for evolutionary analysis of nongenic sequences and population genetics of taxa with large, repeat-rich genomes.

Author contributions: B.S., E.S.B., and M.C.S. designed research; B.S. and M.C.S. performed research; B.S., S.M.-S., M.M., L.J., and E.S.B. contributed new reagents/analytic tools; B.S. and M.C.S. analyzed data; and B.S., E.S.B., and M.C.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: esb33@cornell.edu, mcs368@cornell.edu, or bs674@cornell.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2113075119/-/DCSupplemental.

Published December 21, 2021.

alignment. In addition, alignment using the affine gap cost strategy does not model mechanisms that generate indels of different length distributions (11), and alignment extension terminates in front of long indels (i.e., TE presence/absence variants [PAVs]). Finally, most genome alignment approaches generate many-to-many alignments or are limited to one-to-one alignments (12), which may not reflect the true evolutionary history when comparing taxa with unshared WGD.

Here, we developed AnchorWave (Anchored Wavefront alignment), a whole-genome alignment method that utilizes genome collinearity and can be guided by differing levels of WGD to perform sensitive sequence alignment with high accuracy and recall long indels. These features provide a significant improvement compared to current methods when aligning genomes with enriched dispersed repeats, high sequence diversity, high transposon activity, or WGD variation. Some of these complex genomic variations that are challenging for alignment can be found in vertebrates and insects (13–15) but are widespread among plant genomes (16).

Results

AnchorWave leverages collinear regions to improve genome alignment (Fig. 1). Syntenic or collinear arrangements among taxa have been investigated by aligning protein-coding gene sequences (17). These collinear blocks are parsimoniously interpreted as being derived from a shared ancestor (17, 18). AnchorWave takes the reference genome sequence and gene annotation as input and extracts the reference full-length coding sequence (CDS) to use as anchors. We use a splice-aware sequence alignment program [minimap2 (19) or GMAP (20)] to lift over the start and end positions of the reference full-length CDS to the query genome (Fig. 1, step 1). AnchorWave then identifies collinear anchors using one of three user-specified algorithm options (Fig. 1, step 2) and uses a newly implemented two-piece affine gap cost strategy in a Wavefront Algorithm (WFA) library (21) to perform alignment for each anchor and interanchor interval (Fig. 1, step 4). Some interanchor regions cannot be aligned via WFA due to high computational costs. For these situations, AnchorWave either identifies novel anchors within long interanchor regions (Fig. 1, step 3) or, for those that cannot be split by novel anchors, aligns them using either the dynamic programming global sequence alignment function ksw_extd2 implemented in the minimap2 library (19) or a reimplemented sliding window approach (22) (Fig. 1, step 4). AnchorWave concatenates base pair sequence alignment for each anchor and interanchor region and outputs the alignment in Multiple Alignment Format (MAF) (Fig. 1, step 5).

To benchmark AnchorWave, we used partially synthetic and real genomes under a number of scenarios: small genomes with high similarity, large genomes with high TE activity, genomes with many inversions, and alignments between species with varying evolutionary divergence and WGD histories. We focused on the alignment performance in terms of TE alignment, sensitivity in putative regulatory sequence, and computational resources. To test AnchorWave for the alignment of highly similar genomes, we synthesized benchmark alignments by introducing variant calls of 18 *Arabidopsis* (*Arabidopsis thaliana*) accessions (23) to the TAIR10 reference genome. In these benchmark alignments, variants account for 2.13 to 3.60% of genome sites for each accession, including 1.22 to 2.38% of genome sites caused by variants longer than 50 bp (*SI Appendix*, Fig. S1). We synthesized genomes with these variant calls and aligned them against the TAIR10 reference genome using AnchorWave, minimap2 (19), LAST (12), MUMmer4 (24), and GSAlign (25) and compared the newly generated alignments with benchmark alignments. AnchorWave was the only algorithm that aligned chromosomes end to end and aligned highly diverse fragments that were not aligned in the benchmark (*SI Appendix*, Fig. S2), leading to a slight decrease in precision with only minimap2 ranking higher. AnchorWave had the highest F-score and recall (Fig. 2*A* and *SI Appendix*, Fig. S3 and Supporting Note 1).

To evaluate the performance for detecting long indels in repeat-rich genomes, we developed a benchmark by removing ~60% of long terminal repeat (LTR) retrotransposons from the maize (*Zea mays* L.) B73 v4 assembly (26). This synthetic TE-removed genome had 84,271 deletions with lengths ranging from 1,144 to 33,730 bp (*SI Appendix*, Fig. S4). We counted the number of TE deletions that could be correctly recalled by aligning the TE-removed genome against the reference genome (*SI Appendix*, Supporting Note 2). The genome alignment results from GSAlign, LAST, and MUMmer4 did not generate any variant longer than 1 kilobase pair (Kbp). Minimap2 (19) recalled ~21% of these long deletions correctly, likely benefiting from the usage of global alignment between adjacent anchors in a chain. AnchorWave recalled ~95% of these deletions correctly (Fig. 2*B*).

To evaluate the performance of the alignment approaches using a realistic polymorphism landscape where long indels are

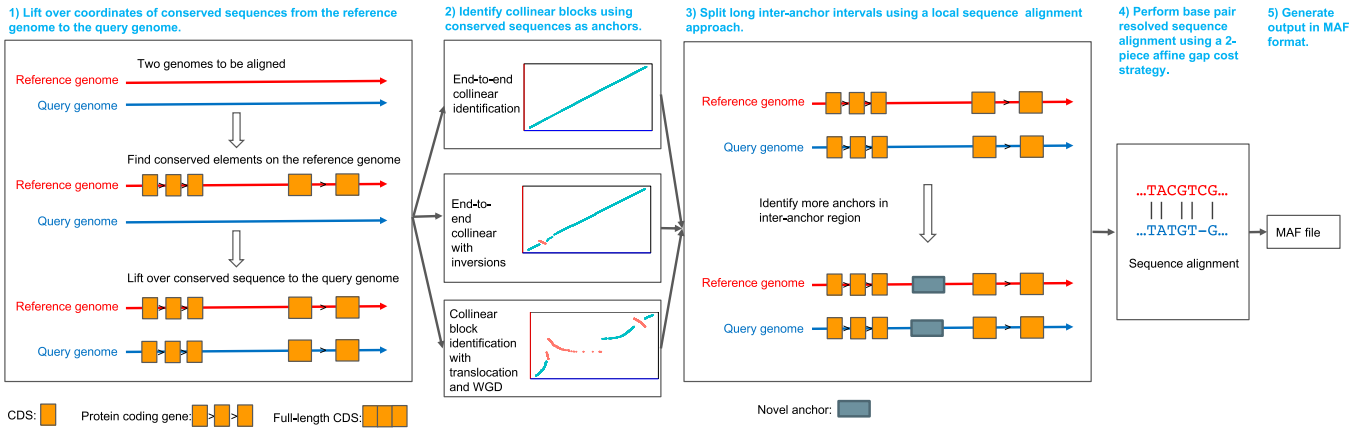

Fig. 1. Principle of the AnchorWave process. AnchorWave identifies collinear regions via conserved anchors (here, full-length CDS) and breaks collinear regions into shorter fragments (i.e., anchor and interanchor intervals). By merging shorter intervals together after performing sensitive sequence alignment via a two-piece affine gap cost global sequence alignment strategy, AnchorWave generates a whole-genome alignment. AnchorWave implements commands to guide collinear block identification with or without chromosomal rearrangements and provides options to use known WGD to inform the alignment.

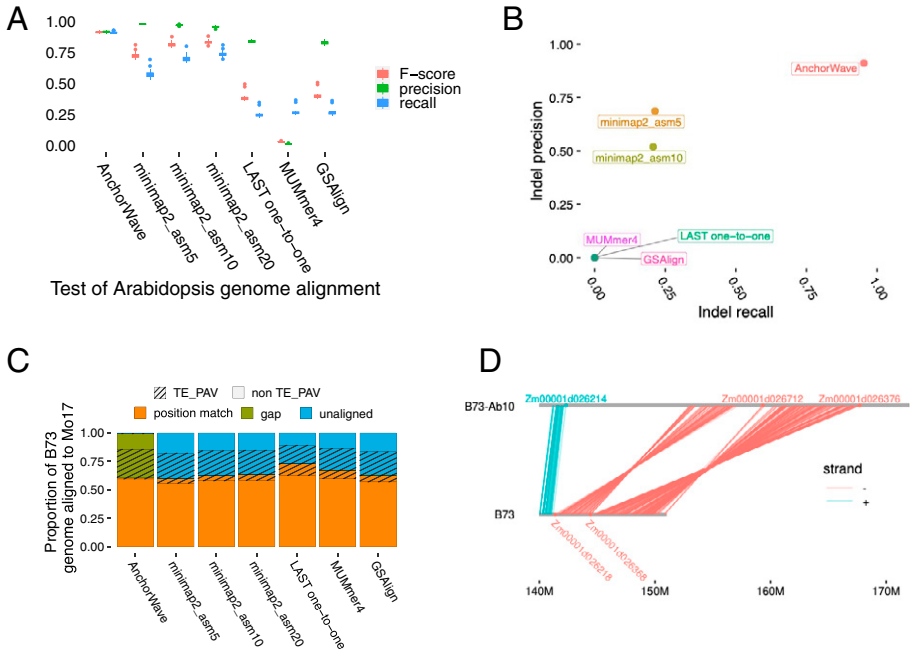

Fig. 2. Comparison of genome alignment tools using genomes of different individuals in the same species. (*A*) Comparison of the performance of genome alignment tools at variant sites of 18 *Arabidopsis* accession alignment benchmarks. Genome alignments were performed using minimap2 with preset options asm5, asm10, and asm20, which terminate extension in regions with 5, 10, and 20% sequence divergence, respectively. GSAlign and MUMmer4 alignments were performed with default parameters. LAST genome alignments with default parameters were termed as LAST many-to-many. LAST many-to-many alignments were processed with a chain and net procedure to generate LAST many-to-one alignments (each query genome nucleotide may be aligned multiple times, while each reference nucleotide can be aligned up to one time). LAST many-to-one alignments were filtered to generate LAST one-to-one alignments (*SI Appendix*, Supporting Note 1). (*B*) Recall and precision of TE deletions by aligning the TE-removed maize B73 genome against the reference genome. MUMmer4, GSAlign, and LAST one-to-one had zero recall ratio. (*C*) Overview of the maize B73 genome sites aligned to the maize Mo17 genome. In those TE regions which were previously reported as present in B73 and absent in Mo17 (TE PAV on the legend), no position match alignments are expected. A higher number of position matches in these regions (striped orange) indicates a higher false-positive ratio. (*D*) Two inversions were located using AnchorWave between the maize B73-Ab10 assembly and the B73 v4 reference genome.

mixed together with single nucleotide polymorphisms (SNPs) and short indels, we aligned the genomes of two maize individuals [B73 v4 and Mo17 (27), *SI Appendix*, Supporting Note 3]. AnchorWave aligned 61% of the B73 genome as position match (defined as an ungapped alignment, either matched or mismatched nucleotides, *SI Appendix*, Fig. S5), which is comparable to other tools (Fig. 2*C*). However, AnchorWave produced the lowest number of position matches in TE insertions previously reported to be specific to B73 (28). This suggests that AnchorWave generated the fewest false positives among compared tools, as these sites should not be matched (orange-striped region of Fig. 2*C*). Moreover, AnchorWave aligned much (37%) of the genome as deletions (gaps). Such gaps are expected for alignments between maize individuals in whom indel variation arises from TE PAVs. Anderson et al. (28) resolved a subset (15,182) of TE PAVs between these two genomes to base pair resolution using an alignment approach anchored on syntenic genes (28). AnchorWave increased this number of TE PAVs to 28,321, recovering 87% (13,181) (28). Other tools had almost zero recall ratios of TE PAVs (*SI Appendix*, Table S1), as they generated few gapped alignments (Fig. 2*C*).

The frequent presence of inversions in eukaryotic genomes poses an obstacle to the end-to-end alignment of chromosomes. By incorporating anchor strand information into the collinear identification approach, AnchorWave efficiently identifies inversions. As an example, we show the gene-level resolution of two neighboring inversions of abnormal chromosome 10 of maize, a cytologically known inversion that carries a meiotic drive locus (29) (Fig. 2*D*, *SI Appendix*, Fig. S6, and Dataset S1). Other recall cases of previously reported inversions using AnchorWave are included in *SI Appendix*, Supporting Note 4 and Dataset S2.

Genome collinearity breaks down as species diverge and as genomes fractionate after polyploidy events (30). We evaluated AnchorWave by aligning a number of representative genomes—autotetraploids, paleopolyploids, and genomes separated by multiple rounds of WGD.

Historically, genome assemblies have been produced as haploid chromosomes after collapsing heterozygous regions. Recent sequencing advances allow the haplotype-resolved assembly of polyploid genomes as shown for the tetraploid potato (*Solanum tuberosum*) (31). We aligned the tomato (*Solanum lycopersicum*) genome to the potato genome, allowing each tomato anchor to define four potato anchors. AnchorWave aligned 63.5% of the tomato genome as position matches or indels in the potato genome, ranking third highest, lower than that generated via LAST many-to-many (76.2%) and LAST many-to-one (66.5%) (*SI Appendix*, Fig. S7 and Supporting Note 5). Polyploidy is rare in vertebrates, but goldfish (*Carassius auratus*) (32) have undergone a WGD ∼14 Mya. Goldfish have twice as many chromosomes as zebrafish (*Danio rerio*) (33), but the size of their genome assemblies are similar (∼1.3 gigabase pairs [Gbp]), suggesting that extensive DNA loss has occurred along the goldfish lineage. Using parameters that allowed each zebrafish anchor to define up to two collinear blocks in goldfish, AnchorWave aligned ∼82.7% of the zebrafish genome sequence as position matches or indels (*SI Appendix*, Supporting Note 6), over twice as much as the second-highest, generated via LAST many-to-many. Often, the reduction to diploidy results in fractionation of subgenomes, a

signal observed in older WGDs. The maize lineage has undergone a WGD since its divergence with sorghum (34), but subsequent chromosomal fusions resulted in these species having the same chromosome number ($n = 10$). Allowing up to two collinear paths for each sorghum anchor (Fig. 3*A* and *SI Appendix*, Fig. S8 and Supporting Note 7), AnchorWave aligned a significantly larger proportion of the maize genome as position match or indels, 3.4 times that of the second-highest, generated via LAST many-to-many (Fig. 3*C*). Multiple rounds of WGDs can further complicate alignment. Soybeans (*Glycine max*) and common beans (*Phaseolus vulgaris*) share an ancient WGD ~56 Mya (35) and then diverged ~19 Mya (35). Subsequently, the soybean lineage had an allotetraploidy event ~13 Mya (36). We aligned the common bean genome to the soybean genome, allowing each common bean anchor to define two soybean anchors to account for the unshared WGD. AnchorWave aligned 78.06% of the soybean genome in collinear blocks with the common bean, which is ranked as the highest (*SI Appendix*, Fig. S9 and Supporting Note 8). For the above alignments, AnchorWave aligned more base pairs as indels than all the other alignment approaches and fewer base pairs as position matches than LAST many-to-many and LAST many-to-one.

An alignment that covers a large proportion of the genome may not be optimal, as false-positive alignment can always increase this proportion. Here, we assessed alignment quality based on biologically informed expectations about lack of sequence conservation in TEs and sequence conservation in putative regulatory sequences. TEs evolve more rapidly than their host genomes, with independent TE movement and sequence divergence among species (37). Thus, TE regions often reflect indels between genomes, as most have inserted more recently than species divergence. While 70 to 85% (38, 39) of the maize genome is composed of structurally recognizable TEs, almost all have estimated insertion times more recent than the divergence from sorghum (Fig. 3*B*). AnchorWave aligns less of the maize genome as position matches in sorghum than any LAST alignment and minimizes position matches located in maize TE regions (orange hatched region; Fig. 3*C*). Other investigated species lack information about the ages of individual TEs, but the turnover of LTR retrotransposons suggests that many TEs in tomato and soybean are younger than the divergence from the species to which we compare them (40). In the tomato–potato comparison, although LAST approaches aligned more of the genome than AnchorWave, much of that was position matches in TE sequence (*SI Appendix*, Fig. S7). In the soybean comparison, AnchorWave not only aligns more of the genome but aligned a smaller ratio in TE regions as position matches compared to non-TE regions (*SI Appendix*, Fig. S9). These TE comparisons show AnchorWave reduces false-positive alignments in repetitive regions with low expected conservation.

Conservation in genic regions is expected across species, but since AnchorWave uses genes to guide genome alignments, we did not use coding sequences to assess alignment quality.

Instead, we turn to putative regulatory sequences, which are expected to be more conserved and less affected by absence variants than the whole–genome background (41). The regulatory sequence of tomato and soybean has been investigated by identifying accessible chromatin regions through ATAC-seq (4, 42). AnchorWave aligned a higher proportion of accessible

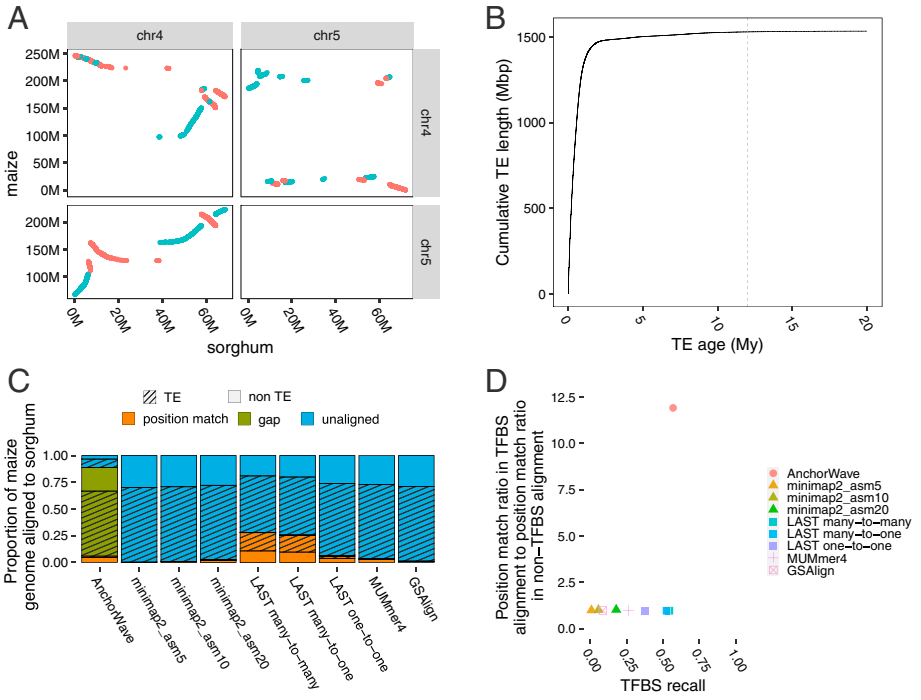

Fig. 3. A comparison of different genome alignment tools for aligning the maize B73 v4 and sorghum genomes. (*A*) The identified collinear anchors between the maize B73 v4 assembly and sorghum genome on chr4 and chr5. Each dot is plotted based on the start coordinate of the reference genome and query genome of each anchor. Collinear anchors on the same strand between the reference genome and query genome are shown in blue, otherwise red. (*B*) Cumulative distribution of TE length versus TE age in the B73 v4 maize genome with age measured in millions of years (My). The dashed line indicates 12 My, the estimated divergence time of maize and sorghum (34). TE age data are from Stitzer et al. (39); 371 TEs older than 20 My were not plotted, and the total length of these 371 TEs is 531 Kbp. (*C*) Sequence alignment between the maize B73 v4 genome and the sorghum genome. Minimap2, MUMmer4, and GSAlign generated many-to-many alignments. Since most maize TEs are not shared with sorghum, a higher number of position matches in maize TE regions (striped orange) indicates a higher false-positive ratio. AnchorWave aligns 88.7% of the maize genome to the sorghum genome, while the second highest is 28.0% generated by LAST many-to-many. (*D*) Comparison of the proportion of sites in maize TFBS that were aligned as a position match (recall) and the position match ratios (number of position match sites to number of aligned sites) in TFBS versus non-TFBS regions.

chromatin regions as position match than any other aligner. Moreover, AnchorWave alignments showed higher sequence identity across these regions, where tomato accessible chromatin regions show a 2.21-fold higher match ratio compared to the whole-genome background (*SI Appendix*, Fig. S7), and soybean accessible chromatin regions show 1.82-fold higher (*SI Appendix*, Fig. S9). All other aligners show nearly equal match ratios in ATAC-seq regions with the whole genome background. TFBSs have been identified directly for maize via chromatin immunoprecipitation sequencing (ChIP-seq). (8). AnchorWave aligned the highest proportion of maize TFBS regions (8) as position matches with sorghum and generated a higher match ratio in TFBS regions compared to the rest of the genome (Fig. 3*D*), suggesting that AnchorWave aligns these conserved functional sequences more accurately and comprehensively.

We further explored the performance of AnchorWave on various species at differing levels of divergence and polyploidy (43) and investigated the proportion of genomes that could be identified in collinear blocks (*SI Appendix*, Supporting Note 9 and Table S2). Maize and rice are two grasses with a greater (50 My) phylogenetic distance than that between maize and sorghum (34). AnchorWave identified 85.5% of the maize genome as collinear with the rice genome (44) (*Oryza sativa*, IRGSP-1.0), 3.5% less than that between maize and sorghum. *Arabidopsis* and chocolate diverged more recently than did *Arabidopsis* and grape (45) (*Vitis vinifera*, 12X), and while 61.8% of the *Arabidopsis* genome was identified as collinear with chocolate, only 50.0% was collinear with grape. Fractionation after WGD breaks collinear blocks, so we examined extreme paleopolyploid comparisons. Two rounds of WGD and one round of whole-genome triplication occurred between *Arabidopsis* and tomato, and AnchorWave detected only 36.3% of the *Arabidopsis* genome as collinear with tomato, despite a similar divergence time as between *Arabidopsis* and grape. Six rounds of WGDs separate maize and banana (46) (*Musa schizocarp*), and although the divergence time between maize and banana is smaller than that between *Arabidopsis* and tomato, only 11.6% of the maize genome was classified as collinear with banana. The joint action of sequence divergence and WGD resolution affects the existence of collinear blocks between species.

AnchorWave exploits the latest advances in sequence alignment using the WFA (21), which reduces the overall memory and computational requirements for global pairwise sequence alignment. To limit memory usage for long sequence comparisons, AnchorWave implements two strategies: identifying novel anchors within long interanchor regions and, for interanchor regions that lack sufficient homology, approximating alignment using either a banded approach or sliding windows. Alignments between *Arabidopsis* genomes typically used less than 10 Gb memory and only minutes of computing time, while the most resource intensive maize–sorghum alignment required 85 Gb memory and 130 h (*SI Appendix*, Tables S3–S5), and alignment can be parallelized when more memory is available. Although the execution time of AnchorWave is high for some experiments, it remains comparable when considering that most other methods fail to align many bases in the genome.

Discussion

Genome evolution across the tree of life has resulted in species with vastly differentiated ploidy, chromosome number, and genome organization. Beyond genome structure, species differ in the complexity and magnitude of nongenic repetitive sequences. Despite this widespread variation, genomes are always punctuated by evolutionarily conserved regulatory sequences and genes. AnchorWave makes use of this conservation, utilizing gene collinearity to guide genome alignment. This approach does not need highly similar seeds to trigger alignment in diverse regions and increases the sequence alignment sensitivity. AnchorWave uses collinear anchors to guide the alignment of repeat elements; thus, local duplications or translocations without enough anchors to be identified as a collinear block are expected to be aligned as indels. Chromosomal fusions after WGD can complicate the separation of subgenomes, but the WGD informed collinear blocks identification function in AnchorWave improves the alignment for genomes with WGD variation. The collinear anchor guided alignment, and the two-piece affine gap cost global sequence alignment strategy enabled the recall of long indels using AnchorWave. Based on the observation that long indels are generally derived from TE movement, while different mechanisms introduce shorter indel mutations (11), the two-piece affine gap cost strategy enables the alignment of short and long indels. The performance of other genome alignment implementations may improve by fine-tuning gap parameters (47), but, with the exception of minimap2, all other tested tools align long indels and short indels using the same gap penalty profile. This precludes parameter tuning as a solution to simultaneously optimize the alignment of long indels and short indels.

We highlight AnchorWave's ability to align several polyploid species to diploid relatives (Fig. 3 and *SI Appendix*, Figs. S8 and S9). Many of these species are ancient allopolyploids, and extensive fractionationation of subgenomes and chromosome fusions have occurred to generate the extant genomes (48). For example, thousands of maize genes have remained duplicated relative to sorghum (48), but, without an explicit model of WGD, most genome alignment strategies fail to align these regions. By generating alignment paths in each subgenome, AnchorWave allows interpretation of the conservation and evolution of these genes and their local regulatory regions. As much of this putative regulatory sequence is embedded between repetitive TEs, AnchorWave reduces false-positive alignments in this repetitive space.

While AnchorWave provides improved alignments for many complicated but real issues in genomics, relying on the existence of collinear regions can be a limitation. Collinearity is more limited in vertebrate systems than plants (49), yet, in comparisons between the human and mouse genomes, AnchorWave provided alignment over putative regulatory sequences (*SI Appendix*, Supporting Note 10). There are limits to aligning noncollinear sequences—although the autosomes of humans and chimpanzees share largely collinear genes, the gene-depauperate Y chromosome could not be aligned (*SI Appendix*, Supporting Note 11). Additionally, technical limitations such as fragmented genome assemblies can prevent the identification of collinearity (50), although this will likely pose less of a problem with advances in long-read sequencing.

When comparing genomes with different rounds of WGD, AnchorWave significantly increased the proportion of the genome that was aligned compared to one-to-one alignments and reduced false-positive alignments compared to many-to-many alignments (Fig. 3*C*). Compared to alignment approaches using a seed-and-extension strategy, AnchorWave increased sensitivity for putative regulatory sequences and could recall long indel variants. AnchorWave's collinear approach further reduces false-positive alignments from dispersed repeats. We showed that AnchorWave can generate whole-genome alignments, facilitating studies of the evolution of regulatory elements, TE polymorphisms, and chromosomal rearrangements such as inversions. AnchorWave even allows duplicated collinear blocks to be aligned, making it particularly relevant to plants, in which an estimated 35% of species are polyploids (16). AnchorWave complements available genome alignment tools by improving the genome alignment of many plant species.

Materials and Methods

Collinear Anchor Identification. AnchorWave takes the reference genome in FASTA format, the reference genome gene annotation in GFF3 format, and the query genome in FASTA format as input. AnchorWave extracts the full-length CDS from the reference genome using the reference genome and annotation. The start and end positions of the reference full-length CDS to the query genome are lifted over using a splice-aware sequence alignment program [minimap2 (19) was used in this manuscript]. AnchorWave then implements a longest-path dynamic programming algorithm to identify collinear anchors. Base pair resolution sequence alignments within each anchor and interanchor region are conducted using the two-piece affine gap cost global sequence alignment strategy, and these alignments are concatenated together to generate the alignment for each collinear block.

A longest-path dynamic programming approach is applied to a pair of chromosomes. On the reference chromosome is a list of n anchors:

$$Q = q_0, q_1, \ \ldots, q_{n-1}.$$

On the query chromosome is a list of m lift over hits:

$$T = t_0, t_1, \ \ldots, t_{m-1}.$$

For each anchor and its hit, q and t, the start and end positions, respectively, are identified from the SAM file output from the splice-aware setting of minimap2. We set up a list of o anchor matches:

$$M = m_0, m_1, \ldots, m_{o-1}.$$

Individual matches are defined as the following:

$$\begin{aligned} m_0 &= \left(q_{i0}, t_{j0}, matchScore_0, strand_0, chainScore_0, preIndex_0\right) \\ m_1 &= \left(q_{i1}, t_{j1}, matchScore_1, strand_1, chainScore_1, preIndex_1\right) \\ &\ldots \\ m_{o-1} &= \left(q_{io-1}, t_{jo-1}, matchScore_{o-1}, strand_{o-1}, chainScore_{o-1}, \ preIndex_{o-1}\right). \end{aligned}$$

t_{jk} is the lift over hit of q_{ik} identified via minimap2, $0 \le k < o$, $0 \le i < n$, $0 \le j < m$. $matchScore_k$ is the sequence similarity (ratio of the number of identical nucleotides to the length of the reference full-length CDS). If q_{ik} and t_{jk} are on the same strand, $strand_k$ is set as positive; otherwise, it is set as negative. $chainScore_k$ is initialized with $matchScore_k$. $preIndex_k$ is the index of the previous anchor on the chain and is initialized as -1, meaning that the current anchor is the first one on a chain.

Three longest-path approaches have been developed for genomes with different types of chromosomal rearrangements. The first two approaches expect chromosome-level assemblies and the homologous chromosomes in the input reference and query files to be named identically, as is common for intraspecific comparisons. Those two approaches try to align each pair of homologous chromosomes from the first base pair to the last base pair. The third approach is the most flexible and identifies collinear blocks using a local longest-path algorithm and then performs base pair resolution global alignment for each collinear block. The third method gains this flexibility by not requiring prior information about homologous chromosomes and may generate alignments for fewer base pairs compared to the first two approaches.

The first approach should be used when no chromosomal rearrangements or inversions occur between homologous chromosomes and the second when inversions disrupt collinearity. When a user has limited background knowledge on chromosomal contiguity of the genomes being aligned, the third approach can be used as the default choice.

Longest-path approach for genome sequences without inversions or rearrangements. The target is to select a subset of positive-strand, nonoverlapping anchor matches from M that give a maximum value of *chainScore* in which the positions of anchors on the reference and query chromosomes increase. The matches in M are first sorted in ascending order by reference anchor start positions. Then, with $0 \le e < f < o$, for a previous element, $m_e = (q_{ie}, t_{je}, matchScore_e, strand_e, chainScore_e, preIndex_e)$ and for a current element, $m_f = (q_{if}, t_{jf}, matchScore_f, strand_f, chainScore_f, preIndex_f)$, the list of matches is iterated, incrementing e and f, while the following conditions are true:

1. The strands of m_f and m_e are positive.
2. The end position of reference anchor q_{ie} is smaller than the start position of reference anchor q_{if}.
3. The end position of the query anchor t_{ie} is smaller than the start position of the query anchor t_{if}.
4. For each m_e and m_f pair, update m_f's *chainScore* based on the following:

$$\begin{aligned} &\text{If } matchScore_f + chainScore_e > chainScore_f: \\ &chainScore_f = matchScore_f + chainScore_e \\ &preIndex_f = \mathrm{e}. \end{aligned}$$

Starting from the match that has the maximum *chainScore*, AnchorWave will use the *preIndex* values to track back and produce a list of matches that give the highest score.

Longest path considering inversions. After sorting the matches in ascending order based on the reference anchor start positions, to consider inversions, we create *currentScore*, which is a cumulative score, and *maxScore*, which is the maximum value that *currentScore* has ever reached for each round of iteration.

When a match is encountered on the negative strand, we assign its *matchScore* to *currentScore*. If the next match is on the negative strand and the query start position is smaller than the query start position of the current one, we add this *matchScore* to the *currentScore*. Otherwise, we subtract this *matchScore* from the *currentScore*. When the *currentScore* drops below 0, the iteration is terminated, and the next iteration starts. If the maximum cumulative score (*maxScore*) is larger than a preset threshold for all matches in a kept group, we reverse their order in the list of matches. Those reversed matches have increasing query start positions and decreasing reference positions and thus remain on the diagonal.

A similar longest-path dynamic programming algorithm is applied to find an end-to-end chain as described for sequences without inversions or rearrangements, except that the end position of anchor q_{if} is smaller than the start position of anchor q_{ie} on the reference chromosome when $strand_e$ and $strand_f$ are negative. With $0 \le e < f < o$, for a previous element, $m_e = (q_{ie}, t_{je}, matchScore_e, strand_e, chainScore_e, preIndex_e)$ and a current element, $m_f = (q_{if}, t_{jf}, matchScore_f, strand_f, chainScore_f, preIndex_f)$, we iterate the list of matches, incrementing e and f, while the following conditions are true:

1. If the current match and previous match are on the negative strand, the end position of reference anchor q_{if} is smaller than the start position of reference anchor q_{ie}. Otherwise, the end position of reference anchor q_{ie} is smaller than the start position of reference anchor q_{if}.
2. The end position of the query anchor t_{ie} is smaller than the start position of the query anchor t_{if}.
3. For each m_e and m_f pair, update the m_f's chainScore based on the following:

$$\begin{aligned} &\text{If } \left(matchScore_f + chainScore_e > chainScore_f\right): \\ &chainScore_f = matchScore_f + chainScore_e \\ &preIndex_f = \ \mathrm{e}. \end{aligned}$$

Starting from the match that has the maximum *chainScore*, AnchorWave will use the *preIndex* values to track back and output the list of anchor matches that give the highest score.

For each pair of homologous chromosomes, at the base pair sequence alignment step, sequences between two anchors on opposite strands were skipped. If the first collinear anchor is on the negative strand, the sequence upstream the first anchor would not be aligned. If the last collinear anchor is on the negative strand, the sequence downstream the last anchor would not be aligned.

Longest path considering inversions, rearrangements, and WGDs. AnchorWave implements a function to constrain the alignment depths for both the reference and query genome, which is useful when there may be multiple collinear paths (i.e., genomes with chromosomal translocations, chromosome fusions, and varying numbers of rounds of WGDs). This function does not assume there are homologous chromosome pairs. Instead, it identifies homologous collinear blocks by applying a local longest-path algorithm on anchors.

The matches are sorted in ascending order based on the reference anchor start positions. Then, with $0 \le e < f < o$, for a previous element, $m_e = (q_{ie}, t_{je}, matchScore_e, strand_e, chainScore_e)$ and a current element, $m_f = (q_{if}, t_{jf}, matchScore_f, strand_f, chainScore_f)$, iterate the list of matches, incrementing e and f, while the following conditions are true:

1. The current match and previous match are on the same strand.
2. The end position of anchor q_{ie} is smaller than the start position of anchor q_{if} on the reference chromosome.
3. If $strand_e$ is positive, the end position of anchor t_{ie} is smaller than the start position of anchor t_{if} on the query chromosome. If $strand_e$ is negative, the end position of anchor t_{ie} is smaller than the start position of anchor t_{if} on the query chromosome.
4. For each m_e and m_f pair, update the m_f's *chainScore* based on the following:

$$\begin{aligned} &if \ \left(matchScore_f + max(0, \ chainScore_e + \ O \ + \ E \ * \ NumberOfGaps(e, \ f)\right) \\ &> chainScore_f): \\ &chainScore_f = \ \mathrm{matchScoref} + \ \max(0, \ \mathrm{chainScore_e} + \ \mathrm{O} \ + \ \mathrm{E} * \\ &\mathrm{NumberOfGaps(e, f)}) \end{aligned}$$

$preIndex_f = e$,

in which O is a gap opening penalty, and E is a gap extension penalty. Let a be the number of anchors between anchor q_{ie} and q_{if}, and let b be the number of anchors between t_{je} and t_{jf}.

$$NumberOfGaps(e, f) = (a + b + |a - b|)/d.$$

NumberOfGaps reflects the number of anchor mismatches and PAVs of anchors. $|a - b|$ is used to penalize differences in the number of anchors between two sequences (in contrast to a similar number of mismatched anchors). *d* is a settable parameter (by default: 3) to calculate a normalized value for the three parts (*a*, *b*, $|a - b|$); higher values of *d* allow more continuous chaining across gaps, which may introduce false-positive chains. To avoid extremely long indels or mismatches, if *a* or *b* is larger than a settable threshold *D* (by default: 30), the current chaining stops. The values of *d*, *D*, *O*, and *E* were set empirically by manually comparing the dot plots of minimap2 splice-aware mapping results and collinear anchors for all genomes used in this manuscript, and default values were used throughout the manuscript.

AnchorWave selects the chain with the maximum *chainScore*. If the maximum *chainScore* is larger than a settable threshold, then output the chain. To align genomes with WGD variants, all the reference and query anchors that fall into the chain range will be counted. If there are multiple hits for a reference anchor in the chain range, the reference anchor is counted once. Anchor matches that fall into chain ranges that are counted as larger than a settable alignment depth threshold are marked and will not be used for the next round of iteration. Then, the next iteration starts using all the unmarked matches until the maximum *chainScore* is smaller than a settable threshold.

The user-settable alignment depth thresholds are the parameters "-R" and "-Q." They are used to control the alignment depth for the reference genome and query genome, especially when the reference and query genomes have WGD variants. "-R" is the maximum alignment depth for the reference genome, and "-Q" is the maximum alignment depth for the query genome. For example, the maize genome has an additional round of WGD compared to the sorghum genome, so we set the maximum alignment depth of the maize genome as 1 and the maximum alignment depth of the sorghum genome as 2. We use parameters "-R 1 -Q 2" to align the sorghum genome against the maize genome.

Filtering Anchors to Improve Alignment Quality. A correct lift over of full-length CDSs from the reference genome to the query genome is central to the AnchorWave pipeline. If the full-length CDSs of two or more genes are identical, AnchorWave ignores all of them in the subsequent analysis because of their ambiguous mappings. We used the splice-aware function of minimap2 to lift over full-length CDSs across this manuscript. Minimap2 misaligns small exons (https://github.com/lh3/minimap2#limitations). To reduce this side effect, when extracting full-length CDSs, AnchorWave ignores CDS exon records <20 bp (although this limit is a user-settable parameter). All exons can be used when more accurate splice-aware aligners [e.g., GMAP (20)] are used for lift over.

We first identify full-length CDSs that might produce incorrect lift overs. In this study, we used minimap2 to map extracted full-length CDSs to the reference genome and then rank the hits of each full-length CDS based on similarity (ratio of the number of identical DNA base pairs to the length of the full-length reference CDS). To minimize the effect of tandemly duplicated anchors, AnchorWave uses two thresholds: a number of mapping hits threshold (e, 1 by default) and a similarity ratio threshold (*y*, 0.6 by default). Any full-length CDS with >*e* mapped hits on a single reference chromosome sequence is further investigated using the similarity ratio threshold. From a similarity-sorted list of all hits of this full-length CDS, we calculate the similarity ratio by dividing the *e* + *1* hit similarity by the highest hit similarity. If this ratio is above the similarity ratio threshold (*y*), we drop this full-length CDS and its hits on any chromosome. We then use the longest-path approach for genome sequences without inversions or rearrangements to further filter the hits. We compare the coordinates of remaining hits with the original GFF3 file; any full-length CDS with different coordinates between the original GFF3 file and lift over reference hits is placed on an unwanted list.

We then use minimap2 to lift over extracted full-length CDSs to the query genome. Full-length CDSs in the unwanted list are not used. Anchors are further filtered to reduce the impact of tandem duplications using the same approach and parameters as described for the reference genome. After filtering, the remaining anchors are fed into a user-specified longest-path algorithm to identify collinear blocks.

Identifying Additional Anchors to Reduce the Size of Interanchor Intervals. To improve sequence alignment quality and computational efficiency, additional anchors are needed in long interanchor intervals. AnchorWave used the "mm_map" function of the minimap2 library with a single-piece affine gap cost setting to perform local alignment in collinear interanchor regions. In each interanchor region, AnchorWave selects the mm_map primary alignment and specifies this as a novel anchor. This step is iterated until either all interanchor intervals are shorter than a settable threshold, the sequence similarity of the new primary alignment is lower than a threshold (by default, we do not filter novel anchors using similarity, and this is set as 0), or no new mm_map matches can be found.

Base Pair Resolution Sequence Alignment Using the Two-Piece Affine Gap Cost Strategy. Based on the assumption that pairs of sequences in each anchor or interanchor region are passed down from a common ancestor, AnchorWave performs base pair resolution global sequence alignment for each anchor and interanchor interval using the two-piece affine gap cost strategy (19). We implemented the two-piece affine gap cost strategy in the WFA (21) library, and sequence alignments are conducted using WFA by default. If the WFA library requires more memory than a preset threshold, the "ksw_extd2" functions implemented in minimap2 are called using a calculated bandwidth. The memory cost threshold of the WFA library is calculated using the "-w" parameter of AnchorWave. The bandwidth of ksw_extd2 is calculated using the "-w" parameter of AnchorWave and anchor/interanchor sequence length. For longer sequences with calculated ksw_extd2 bandwidth smaller than "-w," we implement the two-piece affine gap cost strategy with a sliding window approach (22), which generates approximate sequence alignments. The sliding window size (-w) was set as 38,000 in this study, which was also used as the minimum bandwidth for ksw_extd2.

Long indels are generally derived from TE movements, while different mechanisms introduce shorter indel mutations (11). Since most available genome alignment tools align long indels and short indels using the same gap penalty profile, long indels generally fail to be aligned. Here, we implemented the two-piece gap cost strategy to align long indels following equation 4 described by Li (19). Let the reference $r = r_0, r_1 \ldots r_{n-1}$ and the query $q = q_0, q_1 \ldots q_{m-1}$ be a pair of sequence fragments in an anchor or interanchor interval from the reference genome and the query genome, respectively, with length $|r| = n$ and $|q| = m$. O_1 is the first piece affine gap open penalty, E_1 is the first piece affine extend open penalty, O_2 is the second piece affine gap open penalty, and E_2 is the second piece affine extend open penalty. Let *l* be the gap length; $\min\{O_1 + |l| \times E_1, O_2 + |l| \times E_2\}$ was used as the indel penalty for the dynamic programming sequence alignment approach. We always assume $O_1 + E_1 < O_2 + E_2$. On the condition that $E_1 > E_2$, it applies cost $O_1 + |l| \times E_1$ to gaps shorter than $\lceil (O_2 - O_1)/(E_1 - E_2) \rceil$ and applies $O_2 + |l| \times E_2$ to longer gaps.

The values of the first piece affine gap cost and second piece affine gap cost were selected based on the finding that TE copies are longer than 50 bp (39) (*SI Appendix*, Fig. S10).

Data Sources and Methods for Validation. All the AnchorWave results shown in this manuscript are repeatable using AnchorWave version 1.0.0. The parameter "-IV" of the AnchorWave "genoAli" command was used to identify inversions between maize de novo assemblies against the maize B73 v4 reference genome. To align the sorghum genome against the maize genome, we used the "proali" command of AnchorWave with parameters "-R 1 -Q 2" to utilize the knowledge that the maize lineage has been through a WGD since its divergence with sorghum (34). The value of "-Q," "-R," and "-e" were also adjusted for other polyploid and paleopolyploid genome alignments using AnchorWave (see *SI Appendix*, Supporting Notes 1–11 for more details).

The values of asm5, asm10, and asm20 of the -x setting of minimap2 v2.16-r922 (19) were used separately. When aligning the TE-removed maize B73 genome sequence using a single thread, the setting asm20 gave an "insufficient memory" error on a computer with 2 terabytes of available memory, and we did not obtain the corresponding result.

The "lastal" function from the LAST toolkit v932 (12) was used to perform genome alignment with default parameters; the results were termed LAST many-to-many. Following previously described methods (51), the LAST many-to-many results were transformed into PSL format using the "maf-convert" command of LAST, and the psl files were fed into the chain-net pipeline, "axtChain -linearGap=loose," "chainMergeSort," "chainPreNet," "chainNet," "netToAxt," "axtSort," and "axtToMaf" in sequential order to generate the LAST many_to_one results. The LAST many_to_one results were further processed via the "last-split | maf-swap | last-split | maf-swap" command to

generate LAST one_to_one results. "last-split" and "maf-swap" are components of the LAST genome alignment toolkits.

The parameters –sam-short of MUMmer4 (24) were used to produce genome alignments in SAM format. We used the parameter "-fmt 1" of GSAlign (25) to perform genome alignments with default settings and output in MAF.

To calculate the proportion of the reference genome that was aligned and matched, all the alignments in MAF were reformatted into bam files using the "maf-convert sam" command of LAST and SAMtools v1.10 (52). We used the "depth" command of SAMtools to calculate how many base pairs of the reference genome were aligned. We used the "samtools depth | awk '$3>0{print $0}' | wc -l" command to calculate how many base pairs of a reference genome has a matched position in the query genome.

Because of the limit of available computational resources (maximum of 2 terabytes of memory), we split the maize B73 v4 genome and maize Mo17 genome into individual chromosomes and performed alignments using LAST and minimap2 for each pair of homologous chromosomes independently. When aligning chromosome 1 using minimap2 asm20, we set "-w 19" to reduce memory usage to less than 2 terabytes. The outputs of minimap2, MUMmer4, and GSAlign were filtered as one-to-one alignment for subsequent analysis using the "last-split | maf-swap | last-split | maf-swap" commands. More detailed description about the data and methods could be found in *SI Appendix*, Supporting Notes 1–11.

Data Availability. The source code of AnchorWave is hosted by GitHub at https://github.com/baoxingsong/anchorwave (53). All the commands, plotting scripts, and plot source data for this manuscript can be found at: https://github.com/baoxingsong/genomeAlignment (54).

ACKNOWLEDGMENTS. This project is supported by the United States Department of Agriculture Agricultural Research Service, NSF No. 1822330, NSF No. 1854828, the European Union's Horizon 2020 Framework Programme under the DeepHealth project [825111], the European Union Regional Development Fund within the framework of The European Regional Development Fund Operational Program of Catalonia 2014 to 2020 with a grant of 50% of total cost eligible under the DRAC project [001-P-001723], and National Natural Science Foundation of China No. 31900486. M.C.S. was supported by NSF Postdoctoral Research Fellowship in Biology No. 1907343. M.M. was partially supported by the Spanish Ministry of Economy, Industry, and Competitiveness under Ramón y Cajal (RYC) fellowship number RYC-2016-21104. We thank the members of the E.S.B. laboratory (Cornell University) for helpful discussions. We thank Travis Wrightsman (Cornell University) for suggesting the name AnchorWave and submitting it to bioconda and Hequan Sun (Max Planck Institute for Plant Breeding Research) for sharing the tetraploid potato assembly prerelease. We thank Merritt Khaipho-Burch, Qi Sun, Cheng Zou, Minghui Wang, Zack Miller, Sara Miller (Cornell University), Jeffrey Ross-Ibarra (University of California, Davis), and Shi Huang (University of California, San Diego) for helpful comments.

1. H. A. Lewin *et al.*, Earth BioGenome project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4325–4333 (2018).
2. M. Exposito-Alonso, H.-G. Drost, H. A. Burbano, D. Weigel, The Earth BioGenome project: Opportunities and challenges for plant genomics and conservation. *Plant J.* **102**, 222–229 (2020).
3. B. Wei *et al.*, Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. *BMC Genomics* **17**, 536 (2016).
4. Z. Lu *et al.*, The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* **5**, 1250–1259 (2019).
5. M. Freeling, M. J. Scanlon, J. E. Fowler, Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.* **35**, 110–118 (2015).
6. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
7. H. Li, N. Homer, A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11**, 473–483 (2010).
8. X. Tu *et al.*, Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat. Commun.* **11**, 5089 (2020).
9. R. C. O'Malley *et al.*, Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**, 1280–1292 (2016).
10. B. Song *et al.*, Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Res.* **31**, 1245–1257 (2021).
11. J. L. Bennetzen, J. Ma, K. M. Devos, Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95**, 127–132 (2005).
12. S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
13. Z. Li *et al.*, Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4713–4718 (2018).
14. G. Tsagkogeorga, J. Parker, E. Stupka, J. A. Cotton, S. J. Rossiter, Phylogenomic analyses elucidate the evolutionary relationships of bats. *Curr. Biol.* **23**, 2262–2267 (2013).
15. C. Sun *et al.*, LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol. Evol.* **4**, 168–183 (2012).
16. T. E. Wood *et al.*, The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13875–13879 (2009).
17. H. Tang *et al.*, Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**, 102 (2011).
18. D. Liu, M. Hunt, I. J. Tsai, Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinformatics* **19**, 26 (2018).
19. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
20. T. D. Wu, C. K. Watanabe, GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
21. S. Marco-Sola, J. C. Moure, M. Moreto, A. Espinosa, Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–463 (2020).
22. B. Song *et al.*, Complement genome annotation lift over using a weighted sequence alignment strategy. *Front. Genet.* **10**, 1046 (2019).
23. X. Gan *et al.*, Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**, 419–423 (2011).
24. G. Marçais *et al.*, MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
25. H.-N. Lin, W.-L. Hsu, GSAlign: An efficient sequence alignment tool for intra-species genomes. *BMC Genomics* **21**, 182 (2020).
26. Y. Jiao *et al.*, Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
27. S. Sun *et al.*, Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* **50**, 1289–1295 (2018).
28. S. N. Anderson *et al.*, Transposable elements contribute to dynamic genome content in maize. *Plant J.* **100**, 1052–1065 (2019).
29. R. J. Mroczek, J. R. Melo, A. C. Luce, E. N. Hiatt, R. K. Dawe, The maize Ab10 meiotic drive system maps to supernumerary sequences in a large complex haplotype. *Genetics* **174**, 145–154 (2006).
30. H. Tang *et al.*, Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
31. H. Sun *et al.*, Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *bioRxiv* [Preprint] https://doi.org/10.1101/2021.05.15.444292. Accessed 17 May 2021.
32. Z. Chen *et al.*, NISC Comparative Sequencing Program, De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Sci. Adv.* **5**, eaav0547 (2019).
33. K. Howe *et al.*, The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
34. Z. Swigonová *et al.*, Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923 (2004).
35. J. Schmutz *et al.*, A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).
36. J. Schmutz *et al.*, Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
37. C. Vitte, O. Panaud, LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenet. Genome Res.* **110**, 91–107 (2005).
38. M. B. Hufford *et al.*, De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *bioRxiv* [Preprint] https://doi.org/10.1101/2021.01.14.426684. Accessed 16 January 2021.
39. M. C. Stitzer, S. N. Anderson, N. M. Springer, J. Ross-Ibarra, The genomic ecosystem of transposable elements in maize. *PLoS Genet.* **17**, e1009768 (2021).
40. P. Jedlicka, M. Lexa, E. Kejnovsky, What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Front Plant Sci.* **11**, 644 (2020).
41. E. T. Dermitzakis, A. G. Clark, Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Mol. Biol. Evol.* **19**, 1114–1121 (2002).
42. K. A. Maher *et al.*, Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell* **30**, 15–36 (2018).
43. J. Schnable, E. Lyons, Plant paleopolyploidy (2015) https://doi.org/10.6084/m9.figshare.1538627.v1.
44. Y. Kawahara *et al.*, Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y.)* **6**, 4 (2013).
45. O. Jaillon *et al.*; French-Italian Public Consortium for Grapevine Genome Characterization, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
46. C. Belser *et al.*, Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **4**, 879–887 (2018).
47. Y. Wu *et al.*, A multiple genome alignment workflow shows the impact of repeat masking and parameter tuning on alignment of functional regions in plants. bioRxiv [Preprint] https://doi.org/10.1101/2021.06.01.446647. Accessed 2 June 2021.

48. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4069–4074 (2011).
49. A. Coghlan, E. E. Eichler, S. G. Oliver, A. H. Paterson, L. Stein, Chromosome evolution in eukaryotes: A multi-kingdom perspective. *Trends Genet.* **21**, 673–682 (2005).
50. T. Zhao, M. E. Schranz, Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2165–2174 (2019).
51. L. Kistler *et al.*, Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* **362**, 1309–1313 (2018).
52. H. Li *et al.*; 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. B. Song *et al.*, Data for "AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication." GitHub. https://github.com/baoxingsong/anchorwave. Deposited 15 October 2021.
54. B. Song *et al.*, Data for AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. GitHub. https://github.com/baoxingsong/genomeAlignment. Deposited 1 November 2021.