Supporting Information

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

Baoxing Song\textsuperscript{a*}, Santiago Marco-Sola\textsuperscript{b,c}, Miquel Moreto\textsuperscript{b,d}, Lynn Johnson\textsuperscript{a}, Edward S. Buckler\textsuperscript{a,e,f}\textsuperscript{*}, Michelle C. Stitzer\textsuperscript{a,g}\textsuperscript{*}

\textsuperscript{a}Institute for Genomic Diversity, Cornell University, Ithaca, NY, 14853, USA
\textsuperscript{b}Department of Computer Sciences, Barcelona Supercomputing Center, Barcelona, 08034, Spain
\textsuperscript{c}Departament d’Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, 08193, Spain
\textsuperscript{d}Departament d’Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain
\textsuperscript{e}Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, 14853, USA
\textsuperscript{f}Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, 14853, USA
\textsuperscript{g}Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

*Corresponding authors: Edward S. Buckler, Michelle C. Stitzer, Baoxing Song

Email: esb33@cornell.edu, mcs368@cornell.edu, bs674@cornell.edu

This PDF file includes:

- Fig. S1 to S10
- Titles for Datasets S1 to S2
- Table S1 to S5
- Supporting Note 1 to Note 11

Other Supporting materials for this manuscript include the following:

- Datasets S1 to 2
Fig. S1.

The length distribution of previously published variants for 18 Arabidopsis accessions. (a) Bur-0, (b) Can-0, (c) Ct-1, (d) Edi-0, (e) Hi-0, (f) Kn-0, (g) Ler-0, (h) Mt-0, (i) No-0, (j) Oy-0, (k) Po-0, (l) Rsch-4, (m) Sf-2, (n) Tsu-0, (o) Wil-0, (p) Ws-0, (q) Wu-0, (r) Zu-0.
Fig. S2.

A variant record from Arabidopsis bur_0.v7c.sdi and the inferred benchmark alignment in this region. (a) [left] The reference allele in the TAIR10 genome spans 150 bp, and the variant record is neither an SNP nor an indel. [right] The red region is not aligned in the benchmark, and the alignment across this region from AnchorWave is not expected to be consistent with the benchmark alignment, thus decreasing the precision of summarized AnchorWave alignments. Black base-pairs are 30bp upstream and downstream flanking regions, which are invariant between TAIR10 and the benchmark. (b) An IGV(1) screenshot showing the sequence alignments in this region generated by different tools. Tools except AnchorWave failed to perform sequence alignment in this region. Additional comparisons of different alignment tools at specific regions could be found at https://github.com/baoxingsong/genomeAlignment/blob/main/whereIsWrong.pdf.
Fig. S3.

Genome alignment metrics using the synthetic benchmark alignments of 18 Arabidopsis accessions.

The results of AnchorWave are ranked as highest or comparable to the highest values.
Fig. S4. Length distribution of LTR retrotransposon deletions removed from the maize B73 V4 reference genome.

Fig. S5. An example illustrating the concepts of (position) match sites, aligned sites, gap sites, and unaligned sites of the reference sequence used in this study. The T site of the reference sequence (ref) in orange and italic font of the reference sequence is aligned as a nucleotide mismatch, but a position match.
Collinear anchors between the maize B73 v4 and B73-AB10 genome. Each point was plotted at the start position of an anchor in the reference genome (B73) and the query genome (B73-AB10). Blue points represent anchors on the same strand in the reference genome and the query genome. Red points represent anchors on opposite strands in the reference genome and the query genome.
Fig. S7.

A comparison of different genome alignment tools for aligning the tomato SL2.5 (2) assembly and a potato tetraploid assembly (3). (A) Minimap2, MUMmer4, GSAlign generated many-to-many alignments. Proportions matching the tomato SL2.5 TE annotation (4) are shown with hatched lines. (B) Comparison of the proportion of tomato that were aligned as a position match (recall) and the position match ratios (number of position match sites to number of aligned sites) in accessible chromatin region (ACR) (5) v.s. non-ACR region.
Fig. S8.
Collinear anchors between the maize B73 v4 and sorghum genome assemblies. Each point was plotted at the start position of an anchor in the reference genome (maize) and the query genome (sorghum). Blue points represent anchors on the same strand in the reference genome and the query genome. Red points represent anchors on opposite strands in the reference genome and the query genome.
Fig. S9.

A comparison of different genome alignment tools for aligning the soybean v2.0 assembly (6) and a common bean genome assembly (7). (A) Minimap2, MUMmer4, GSAlign generated many-to-many alignments. (B) Comparison of the proportion of tomato that were aligned as a position match (recall) and the match ratios (number of match sites to number of aligned sites) in accessible chromatin region (ACR) v.s. non-ACR region. The soybean TE and ACR annotations were published by Lu et. al (8).
Fig. S10. The 2-piece affine gap cost strategy as a dynamic programming approach for sequence alignment.

The value of \( \min\{O_1 + |l| \cdot E_1, O_2 + |l| \cdot E_2\} \) is used as the gap penalty. The aim is to model different types of mutational mechanisms that introduce indels of different length distributions.
Dataset S1:
Collinear anchors identified using AnchorWave between the maize B73 v4 assembly and the maize B73-Ab10 assembly.

Dataset S2:
Collinear anchors identified using AnchorWave between the maize B73 v4 assembly and the maize SK assembly.

Table S1:
TE PAVs identified using different genome alignment tools. There are 341,426 TEs with complete target site duplication (TSD) sequences in the B73 genome and 305,022 TEs in the Mo17 genome. 15,182 TEs were identified as wholly absent in either B73 or Mo17 by Anderson et al. 2019 (9) (there, termed “site-defined” TEs).

|                  | Recall of TE PAVs identified by Anderson et al | Total TE PAVs identified |
|------------------|-----------------------------------------------|--------------------------|
| AnchorWave       | 13,181                                        | 28,321                   |
| minimap2_asm5    | 0                                             | 34                       |
| minimap2_asm10   | 0                                             | 55                       |
| minimap2_asm20   | 0                                             | 58                       |
| LAST             | 0                                             | 41                       |
| MUMmer4          | 0                                             | 36                       |
| GSAlign          | 0                                             | 42                       |
Table S2: Summary of the proportion of the genome being aligned by AnchorWave for all species investigated in this manuscript.

| Reference genome | Query genome | Proportion of reference genome in collinear blocks | Position match ratio | Nucleotide match percentage | Median divergence time (My) (10) | Number of rounds whole-genome duplication separation |
|------------------|-------------|-----------------------------------------------|----------------------|-----------------------------|---------------------------------|-----------------------------------------------------|
| maize B73 (2106.34Mbp) | maize Mo17 | 99.24% (2090.33Mbp) | 61.61% (1287.78Mbp) | 96.05% (1236.89Mbp) | NA | 0 |
| maize (2135.08Mbp) | sorghum | 88.70% (1893.77Mbp) | 6.66% (126.03Mbp) | 78.55% (99.00Mbp) | 12 | 1 |
| maize | rice | 85.20% (1819.00Mbp) | 3.72% (67.75Mbp) | 67.39% (45.66Mbp) | 48 | 1 |
| maize | banana | 11.61% (247.79Mbp) | 0.57% (1.41Mbp) | 52.86% (0.81Mbp) | 108 | 6 |
| arabidopsis (119.15Mbp) | cacao | 61.82% (73.66Mbp) | 44.07% (32.46Mbp) | 57.00% (18.50Mbp) | 90 | 2 |
| arabidopsis | grape | 50.96% (60.72Mbp) | 39.86% (24.20Mbp) | 54.52% (13.20Mbp) | 117 | 2 |
| arabidopsis | tomato | 36.17% (43.09Mbp) | 34.83% (15.01Mbp) | 50.69% (6.21Mbp) | 121 | 3 |
| zebrafish (1373.47Mbp) | goldfish | 83.40% (1145.40Mbp) | 27.38% (313.68Mbp) | 77.00% (323.75Mbp) | 83 | 1 |
| tomato (823.63Mbp) | potato | 63.52% (523.16Mbp) | 42.31% (221.33Mbp) | 82.56% (627.78Mbp) | 7.3 | 1 |
| soybean (978.42Mbp) | common bean | 78.06% (763.78Mbp) | 36.35% (277.67Mbp) | 68.04% (188.91Mbp) | 22 | 1 |
| human (3031.04Mbp) | mouse | 86.23% (2613.45Mbp) | 34.12% (891.90Mbp) | 71.07% (633.89Mbp) | 89 | 0 |
| human | chimpanzee | 95.73% (2901.54Mbp) | 91.85% (2664.98Mbp) | 98.52% (2625.46Mbp) | 6.4 | 0 |

1 The numbers of aligned based pairs were obtained by reformatting the alignments into bam format and using the samtools depth alignment.bam | wc -l command. The numbers of position match based pairs were obtained using the samtools depth alignment.bam | awk '$3>0{print $0}' | wc -l command.
2 (identical nucleotides)/(identical nucleotides + mismatch nucleotides). When there are multiple collinear blocks covering a specific region, coverage of each region is counted independently and then summed together.
3 Chromosomes 1-10 were analysed. Mitochondria and plastid sequences were not labeled in the Mo17 assembly. Other assembly contigs were also ignored.
4 Two rounds of doubling whole-genome duplication and one round of tripling whole-genome duplication.
5 The tomato genome is a monoploid assembly, and the potato genome is a heterozygous tetraploid assembly.
6 Chromosome Y was excluded for this analysis.
**Table S3:**
The CPU time and memory cost of testing software for aligning Arabidopsis synthetic genomes against the reference genome.

The computational costs were tested on a computer with 512Gb RAM and Intel(R) Xeon(R) Gold 6230 CPU.

Here, we used the single thread model for comparison purposes.

| software   | memory (Gb) | time (HH:MM:SS) | memory (Gb) | time (HH:MM:SS) | memory (Gb) | time (HH:MM:SS) | memory (Gb) | time (HH:MM:SS) | memory (Gb) | time (HH:MM:SS) |
|------------|-------------|-----------------|-------------|-----------------|-------------|-----------------|-------------|-----------------|-------------|-----------------|
| AnchorWave | 9.9Gb       | 00:06:13        | 1.0Gb       | 00:02:15        | 1.7Gb       | 00:02:30        | 1.5Gb       | 00:06:13        | 0.8Gb       | 00:01:41        |
| can_0      | 9.4Gb       | 00:08:15        | 1.3Gb       | 00:02:17        | 1.8Gb       | 00:02:28        | 1.5Gb       | 00:06:16        | 0.8Gb       | 00:01:43        |
| cs_1       | 9.5Gb       | 00:04:51        | 0.9Gb       | 00:01:55        | 0.9Gb       | 00:02:22        | 1.5Gb       | 00:06:07        | 0.8Gb       | 00:01:32        |
| edl_0      | 9.0Gb       | 00:06:05        | 1.2Gb       | 00:02:12        | 1.7Gb       | 00:02:29        | 1.5Gb       | 00:06:07        | 0.8Gb       | 00:01:51        |
| hi_0       | 9.9Gb       | 00:03:59        | 0.9Gb       | 00:01:58        | 0.9Gb       | 00:02:28        | 1.5Gb       | 00:07:38        | 0.7Gb       | 00:01:46        |
| kn_0       | 9.6Gb       | 00:05:57        | 1.1Gb       | 00:02:07        | 1.1Gb       | 00:02:30        | 1.5Gb       | 00:07:47        | 0.8Gb       | 00:01:39        |
| ls_0       | 9.6Gb       | 00:06:17        | 1.2Gb       | 00:02:10        | 1.7Gb       | 00:02:31        | 1.5Gb       | 00:07:46        | 0.8Gb       | 00:01:52        |
| mi_0       | 8.8Gb       | 00:05:27        | 1.0Gb       | 00:02:07        | 1.7Gb       | 00:02:27        | 1.5Gb       | 00:07:44        | 0.8Gb       | 00:01:48        |
| ns_0       | 9.2Gb       | 00:06:10        | 1.1Gb       | 00:02:08        | 1.7Gb       | 00:02:29        | 1.5Gb       | 00:07:44        | 0.8Gb       | 00:01:40        |
| oy_0       | 9.6Gb       | 00:05:31        | 1.2Gb       | 00:02:00        | 1.7Gb       | 00:02:23        | 1.5Gb       | 00:07:53        | 0.8Gb       | 00:01:37        |
| po_0       | 27Gb        | 00:03:21        | 1.1Gb       | 00:02:05        | 1.7Gb       | 00:02:27        | 6G          | 00:33:49        | 1.5Gb       | 00:07:55        |
| rna_4      | 9.9Gb       | 00:04:45        | 1.3Gb       | 00:02:18        | 1.7Gb       | 00:02:35        | 1.5Gb       | 00:08:15        | 0.8Gb       | 00:01:41        |
| sf_2       | 9.2Gb       | 00:06:59        | 1.4Gb       | 00:02:17        | 1.7Gb       | 00:02:31        | 9Gb         | 00:34:18        | 1.5Gb       | 00:08:22        |
| su_0       | 8.8Gb       | 00:06:38        | 1.4Gb       | 00:02:12        | 1.7Gb       | 00:02:28        | 13Gb        | 00:33:52        | 1.5Gb       | 00:09:09        |
| wil_2      | 9.4Gb       | 00:07:22        | 1.1Gb       | 00:02:09        | 1.7Gb       | 00:02:27        | 11Gb        | 00:35:32        | 1.5Gb       | 00:08:28        |
| ws_0       | 9.1Gb       | 00:06:17        | 1.3Gb       | 00:02:12        | 1.7Gb       | 00:02:27        | 11Gb        | 00:34:17        | 1.5Gb       | 00:08:40        |
| ws_0       | 9.3Gb       | 00:06:00        | 1.0Gb       | 00:02:15        | 1.7Gb       | 00:02:27        | 10Gb        | 00:33:24        | 1.5Gb       | 00:08:31        |
| zo_0       | 9.3Gb       | 00:06:21        | 1.2Gb       | 00:02:11        | 1.7Gb       | 00:02:28        | 12Gb        | 00:35:42        | 1.5Gb       | 00:08:45        |
The CPU time and memory cost of testing software for aligning the TE removed maize B73 genome against the maize B73 reference genome. The computational costs of AnchorWave, minimap2 asm5, minimap2 asm10, LAST, Mummer4 and GSAlign, were tested on a computer with 512 gigabyte RAM and Intel(R) Xeon(R) Gold 6230 CPU. Minimap2 asm20 was tested on a computer with 2 terabytes of memory and AMD EPYC 7702 Processor CPU, but gave an insufficient memory error. We could not access a machine with more memory installed.

| Aligner       | memory | time(HH:MM:SS) |
|---------------|--------|----------------|
| AnchorWave    | 56Gb   | 30:36:34       |
| minimap2_asm5 | 89Gb   | 27:06:59       |
| minimap2_asm10| 89Gb   | 25:59:57       |
| minimap2_asm20| >2000Gb| >24:00:00      |
| LAST          | 470Gb  | 223:16:54      |
| MUMmer4       | 25Gb   | 42:28:22       |
| GSAlign       | 9Gb    | 00:35:22       |
The CPU time and memory cost of testing software for aligning the sorghum genome against the maize B73 v4 genome assembly. The computational costs were tested on a computer with 512Gb RAM and Intel(R) Xeon(R) Gold 6230 CPU.

| Aligner        | memory | time(HH:MM:SS) |
|----------------|--------|---------------|
| AnchorWave     | 80Gb   | 131:12:25     |
| minimap2_asm5  | 7Gb    | 00:08:57      |
| minimap2_asm10 | 7Gb    | 00:08:00      |
| minimap2_asm20 | 10Gb   | 00:12:46      |
| LAST           | 21Gb   | 07:48:59      |
| MUMmer4        | 25Gb   | 00:39:38      |
| GSAlign        | 12Gb   | 00:34:51      |
Supporting Note 1:
Assessing the quality of an alignment is difficult, as the true evolutionary history of sequence orthology is unknown, and simulations do not recover many intricacies of sequence evolution\(^{(11)}\). We tested genome alignments using a published variant call dataset of 18 Arabidopsis accessions, originally conducted using a hybrid approach of read mapping and \textit{de novo} assembly\(^{(12)}\). The released variant calling results for 18 Arabidopsis individuals were downloaded in sdi format from http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/variants.SDI/.

Taking the TAIR10 genome as a reference, we generated synthetic genome sequences using the “pseudogeno” command of GEAN\(^{(13)}\) and benchmark genome alignments by replacing the reference alleles with alternative alleles for each accession separately.

To compare the newly generated genome alignments with benchmark genome alignments, we first counted the number of positions in the alignment that differed between the reference and the query genome (referred to as “variant sites”, Fig. S5). True positive variant sites in the alignment are reference sites that are shared at the same query genome position in both the benchmark and newly generated alignment. False positive variant sites are those present in the newly generated alignment but not in the benchmark alignment. False negative variant sites are those present in the benchmark alignment but not in the newly generated alignment.

We calculate precision as:
\[
\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}
\]

We calculate recall as:
\[
\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}
\]

The harmonic mean of precision and recall is used to calculate the balanced F-score:
\[
F - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

Most sites in these alignments are invariant, so we also calculate genome-wide precision and genome-wide recall in addition to those described above for variant sites. We calculate precision for genome-wide sites as the ratio of the number of aligned sites shared between the benchmark alignment and the newly generated alignment to the number of sites in the newly
generated alignment. Conversely, the ratio of the number of shared aligned sites to the number of sites in the benchmark is defined as the genome-wide recall. We calculate the F-score for genome-wide sites using genome-wide precision and recall with the formula shown above.

The TAIR10 reference genome sequence and genome annotation(14) were downloaded by commands:

```
wget https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes.gff
```

```
wget ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/TAIR10_chr1.fas
```

```
wget ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/TAIR10_chr2.fas
```

```
wget ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/TAIR10_chr3.fas
```

```
wget ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/TAIR10_chr4.fas
```

```
wget ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/TAIR10_chr5.fas
```

```
cat TAIR10_chr1.fas TAIR10_chr2.fas TAIR10_chr3.fas TAIR10_chr4.fas TAIR10_chr5.fas > tair10.fa
```

The variant calling results were downloaded from “http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/variants.SDI”.

For each accession, the reference alleles from the TAIR10(14) genome were replaced with alternative alleles to generate a synthetic genome using the “pseudogeno” command of GEAN(13) with default parameters. GEAN reported error messages for ler_0.v7c.sdi, no_0.v7c.sdi and oy_0.v7c.sdi, due to multiple variants located at the same position. We removed the following two records from ler_0.v7c.sdi manually:

```
Chr4 6020491 12 AAGACATCAATATCATCAGGAAAAATACTCATTCTATTATTAG
TAATGACTTAGAGAATAAACTACGAATACAAAA 4
```

We removed the following two records from no_0.v7c.sdi manually:

```
Chr5 8499298 -1
CAACAAGTTAAGATTTTAAAAATCCATTTTAAAAAAATTTTTTATTGGAAAAACTATATGGAAATTCAGGAAATTTTGAAA
```

```
AATTCATTTTAAAAGTTTTTTAATAACTGATAACGTACTACTTCAAAAACTGTAAACCAAATGCTAAGAATGCATATATGTTTCTGG
CAACAAAGAGATAACAGATA
GAGTGATAAAAGATAAATGCTACTATCTTTTGTTGGATATTCGCCAAACAGTTGAAAGTTTTAAGATCCAGCTATAGC
```

17
We removed the following two records from oy_0.v7c.sdi manually:

Chr2 7677637 7 ACAAGATAAATTTAAATATTTATTTTAGAGCTAAAGTAATGAAAAATGTTTCTAAACACTAGTTAAT 5

Chr2 7677637 8 - GCTAAATG 4

The synthetic genomes were generated using the "pseudogeno" function of GEAN (13):

gean pseudogeno -r tair10.fa -v bur_0.v7c.sdi -o bur_0.fa
gean pseudogeno -r tair10.fa -v can_0.v7c.sdi -o can_0.fa
gean pseudogeno -r tair10.fa -v ct_1.v7c.sdi -o ct_1.fa
gean pseudogeno -r tair10.fa -v edil_0.v7c.sdi -o edil_0.fa
gean pseudogeno -r tair10.fa -v hi_0.v7c.sdi -o hi_0.fa
gean pseudogeno -r tair10.fa -v kn_0.v7c.sdi -o kn_0.fa
gean pseudogeno -r tair10.fa -v ler_0.v7c.sdi -o ler_0.fa
gean pseudogeno -r tair10.fa -v mt_0.v7c.sdi -o mt_0.fa
gean pseudogeno -r tair10.fa -v no_0.v7c.sdi -o no_0.fa
gean pseudogeno -r tair10.fa -v oy_0.v7c.sdi -o oy_0.fa
gean pseudogeno -r tair10.fa -v po_0.v7c.sdi -o po_0.fa
gean pseudogeno -r tair10.fa -v rsch_4.v7c.sdi -o rsch_4.fa
gean pseudogeno -r tair10.fa -v sf_2.v7c.sdi -o sf_2.fa
gean pseudogeno -r tair10.fa -v tsu_0.v7c.sdi -o tsu_0.fa
gean pseudogeno -r tair10.fa -v wil_2.v7c.sdi -o wil_2.fa
gean pseudogeno -r tair10.fa -v ws_0.v7c.sdi -o ws_0.fa
gean pseudogeno -r tair10.fa -v wu_0.v7c.sdi -o wu_0.fa
The benchmark genome alignments were generated using commands:

```
anchorwave sdiToMaf -o col_bur.maf -r tair10.fa -s bur_0.sdi -q bur_0.fa
anchorwave sdiToMaf -o col_can.maf -r tair10.fa -s can_0.sdi -q can_0.fa
anchorwave sdiToMaf -o col_ct.maf -r tair10.fa -s ct_1.sdi -q ct_1.fa
anchorwave sdiToMaf -o col_edi.maf -r tair10.fa -s edi_0.sdi -q edi_0.fa
anchorwave sdiToMaf -o col_hi.maf -r tair10.fa -s hi_0.sdi -q hi_0.fa
anchorwave sdiToMaf -o col_kn.maf -r tair10.fa -s kn_0.sdi -q kn_0.fa
anchorwave sdiToMaf -o col_ler.maf -r tair10.fa -s ler_0.sdi -q ler_0.fa
anchorwave sdiToMaf -o col_mt.maf -r tair10.fa -s mt_0.sdi -q mt_0.fa
anchorwave sdiToMaf -o col_no.maf -r tair10.fa -s no_0.sdi -q no_0.fa
anchorwave sdiToMaf -o col_oy.maf -r tair10.fa -s oy_0.sdi -q oy_0.fa
anchorwave sdiToMaf -o col_po.maf -r tair10.fa -s po_0.sdi -q po_0.fa
anchorwave sdiToMaf -o col_rsch.maf -r tair10.fa -s rsch_4.sdi -q rsch_4.fa
anchorwave sdiToMaf -o col_sf.maf -r tair10.fa -s sf_2.sdi -q sf_2.fa
anchorwave sdiToMaf -o col_tsu.maf -r tair10.fa -s tsu_0.sdi -q tsu_0.fa
anchorwave sdiToMaf -o col_wil.maf -r tair10.fa -s wil_2.sdi -q wil_2.fa
anchorwave sdiToMaf -o col_ws.maf -r tair10.fa -s ws_0.sdi -q ws_0.fa
anchorwave sdiToMaf -o col_wu.maf -r tair10.fa -s wu_0.sdi -q wu_0.fa
anchorwave sdiToMaf -o col_zu.maf -r tair10.fa -s zu_0.sdi -q zu_0.fa
```

We extracted full-length CDS of the TAIR10 reference genome using the command:

```
anchorwave gff2seq -r tair10.fa -i TAIR10_GFF3_genes.gff -o cds.fa
```

We mapped full-length CDS to the reference genome using the splice-aware function of minimap2(15) with the command:
We mapped full-length CDS to each synthetic genome using the command:

```
minimap2 -x splice -t 4 -k 12 -a -p 0.4 -N 20 tair10.fa cds.fa > ref.sam
```

AnchorWave genome alignments were conducted via the command:

```
anchorwave genoAli -i TAIR10_GFF3_genes.gff -as cds.fa -r tair10.fa -a synthesis_genome.sam -ar ref.sam -s synthesis_genome.fa -v synthesis_genome.vcf -n synthesis_genome.anchors -o synthesis_genome.maf -f synthesis_genome.f.maf
```

We used minimap2 to align each synthetic genome against the reference genome with three genome alignment parameter sets using the commands:

```
minimap2 -x asm5 -t 1 -a tair10.fa synthesis_genome.fa > minimap2_asm5_synthesis_genome.sam
minimap2 -x asm10 -t 1 -a tair10.fa synthesis_genome.fa > minimap2_asm10_synthesis_genome.sam
minimap2 -x asm20 -t 1 -a tair10.fa synthesis_genome.fa > minimap2_asm20_synthesis_genome.sam
```

These output files in SAM format were reformatted into MAF using the “sam2maf” function implemented in AnchorWave:

```
anchorwave sam2maf -r tair10.fa -q synthesis_genome.fa -s minimap2_asm5_synthesis_genome.sam -o minimap2_asm5_synthesis_genome.maf
```

The LAST(16) genome alignments were conducted using commands:

```
lastdb col tair10.fa
faToTwoBit tair10.fa col.2bit
faSize -detailed tair10.fa > col.size
lastal col synthesis_genome.fa > synthesis_genome_lastal.maf
faSize -detailed synthesis_genome.fa > synthesis_genome.size
faToTwoBit synthesis_genome.fa synthesis_genome.2bit
maf-convert psl synthesis_genome_lastal.maf > synthesis_genome_lastal.psl
axtChain -linearGap=loose -psl synthesis_genome_lastal.psl -faQ -faT tair10.fa synthesis_genome.fa
synthesis_genome_lastal.chain
chainMergeSort synthesis_genome_lastal.chain > synthesis_genome_lastal.all.chain
```
The header of MAF files is not compatible between the LAST pipeline and the chain-net pipeline. We implemented "lastFinalToSplit.pl" to reformat the header of MAF files to make them compatible. This script has been released under "./src/tests/scripts/" of the AnchorWave source code repository. The output file "synthesis_genome_lastal.maf" is termed LAST many-to-many alignment, the output file "synthesis_genome_lastal_final.maf" is termed LAST many-to-one alignment, and the output of "synthesis_genome_lastal_final_split.maf" is termed LAST one-to-one alignment.

We performed alignments using MUMmer4(17) via the command:

```
nucmer -t 1 --sam-short=mumer.synthesis_genome.short.sam tair10.fa synthesis_genome.fa
```

We aligned synthesized genomes against the TAIR10 reference genome using GSAlign(18) via the command:

```
GSAlign -t 1 -r tair10.fa -q synthesis_genome.fa -t 178 -o synthesis_genome_gsalign -fmt 1
```

The scripts to summarize alignment recall, precision and F-score, "CompareMafAndCheckMafCoverage.py" and "comparemafandcheckmafcoverageAll.py", are available in the "./src/tests/scripts/" folder of the source code repository.

AnchorWave outperforms other approaches with the highest recall for variant sites (0.910 on average, the second-best is 0.737 achieved via minimap2 asm20) and for genome-wide
sites (0.992 on average, the second-best is 0.989, achieved via minimap2 asm20). Moreover, AnchorWave is the only implementation that performs end-to-end alignment, and the second-highest proportion of aligned sites (0.992) was achieved by minimap2 asm20. The precision of AnchorWave is lower than all parameter sets of minimap2 we tested (6% lower for variant sites and 0.7% lower genome-wide compared to minimap2 asm5, which is ranked the highest). For the genome-wide sites F-score, AnchorWave (0.9917 on average) is ranked as the third-highest and lower than that of minimap2 asm20 (0.9931 on average) and minimap2 asm10 (0.9922 on average). For variant sites, the average F-score of AnchorWave was 0.912, which is the highest one and the second highest one was 0.832 generated by minimap2 asm20 (Fig. S3).
Supporting Note 2:
The movement of TEs to new positions in the genome generates repeated interspersed sequences, which can cause false positive alignments. TE presence-absence variants terminate the extension of sequence alignment and reduce the proportion of the genome that is aligned as position matches. We tested the performance of AnchorWave for the detection of long indels in repeat-rich genomes by removing retrotransposons from the maize B73 v4 genome assembly and aligning the synthetic genome against the reference genome. We removed ~60% of long terminal repeat (LTR) retrotransposons from the maize B73 v4 genome, as these TEs make up the majority of genomic DNA. We used results from LTRharvest (19) with the parameters “-motif tgca -minlenltr 100 -maxlenltr 7000 -mindistltr 1000 -maxdistltr 20000 -similar 85 -motifmis 1 -mintsd 5 -overlaps best”, to generate a GFF3 file with entries for the LTR retrotransposon, as well as the two flanking target site duplications generated when the TE is inserted. We deleted the LTR retrotransposon as well as one of the two target site duplications in order to recapitulate the empty allele that would have existed before the TE was inserted. We merged adjacent TE deletion records together and formed a list of non-overlapping benchmark deletion records. We then aligned this TE-deleted genome to the B73 v4 genome to assess alignment quality in these highly repetitive regions. To evaluate the deletion recall ratios of the genome alignment approaches, we defined two deletion records as identical if by removing the sequence from the reference genome they could generate the identical synthetic genome sequences. This was done in order to avoid the impact of genomic variants that are represented in many different ways (13), e.g., when microhomology at the boundaries of a deletion means that both alignments are equally optimized.

The B73 v4 reference genome sequence and genome annotation (20) were downloaded using commands:

```
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-34/gff3/zea_mays/Zea_mays.AGPv4.34.gff3.gz
gunzip Zea_mays.AGPv4.34.gff3.gz
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-34/fasta/zea_mays/dna/Zea_mays.AGPv4.dna.toplevel.fa.gz
gunzip Zea_mays.AGPv4.dna.toplevel.fa.gz
```

To generate a GFF3 format LTR retrotransposon annotation, we use ltrharvest (19). First, we generate an index of the genome:
genometools-1.5.7/bin/gt suffixerator -db Zea_mays.AGPv4.dna.toplevel.fa.gz -indexname B73V4 -tis -suf -lcp -des -ssp -sds -dna -memlimit 48GB

Then, we search for LTR retrotransposons using ltrharvest:

Then, we search for LTR retrotransposons using ltrharvest:

```
 genometools-1.5.7/bin/gt ltrharvest -index B73V4 -gff3 B73V4.ltrharvest.gff3 -motif tgca -minlenltr 100 -maxlenltr 7000 -mindistltr 1000 -maxdistltr 20000 -similar 85 -motifmis 5 -xdrop 5 -overlaps best -longoutput -out B73V4.fa > B73v4.ltrharvest.out
```

Finally, since genometools names sequences with an internal identifier, we convert the sequence names in GFF3 file back to chromosome names:

```
wget https://raw.githubusercontent.com/mcstitzer/maize_v4_TE_annotation/master/ltr/mask_subtract/convert_ltrharvest_seq_gff_to_contignames.py
```

```
python2 convert_ltrharvest_seq_gff_to_contignames.py B73V4.ltrharvest.gff3 > B73V4.ltrharvest.contignames.gff3
```

We use this GFF3 to generate a FASTA file of the genome with these TEs removed (B73V4.pseudomolecule.subtract1.fa), by removing both the LTR retrotransposon and one of the target site duplications generated when it inserted in the genome. This is implemented in the “gffToMaf, c1” function located within “/src/tests/impl/TEGffToAlignment.cpp” of the AnchorWave source code repository. This function replaces each nucleotide within TE regions of maize B73 v4 genome with “-”. After all the nucleotides within TE regions are replaced, this function removes the “-” characters and outputs the remaining nucleotides in FASTA format. We created a function “gff2vcf” under “/src/tests/impl/TEGffToAlignment.cpp” to create the variant records file “B73V4.pseudomolecule.ltrharvest.contignames.gff3.contigpositions.vcf”. This function replaces each nucleotide within TE regions of maize B73 v4 genome with “-” and generates a pairwise sequence alignment and then performs variant calling. The length distribution of deletion records in “B73V4.pseudomolecule.ltrharvest.contignames.gff3.contigpositions.vcf” is shown in (Fig. S4).

We aligned “B73V4.pseudomolecule.subtract1.fa” against the B73 v4 reference genome using AnchorWave, minimap2, LAST, MUMmer4, and GSAlign using the same settings as described in Supporting Note 1. The genome alignment files in SAM(21) format were transformed into MAF format using the “sam2maf” function implemented in AnchorWave, and the function “maf2vcf” of AnchorWave was used to generate variants in VCF format(22). We implemented the “evaluateTEAlignment” function to compare generated VCF files with the benchmark file “B73V4.pseudomolecule.ltrharvest.contignames.gff3.contigpositions.vcf”.

```
```
As discussed in the main text, AnchorWave shows the highest recall of these TE deletions. Besides AnchorWave, minimap2 is the only method tested that could recall long indels. This may be due to minimap2’s usage of global alignment between adjacent anchors in a chain, and use of a 2-piece affine gap penalty. There is no obvious length distribution difference between correctly recalled deletions and incorrectly recalled deletions for both minimap2 and AnchorWave (Figure SN1).

Figure SN1
The length distribution of TE deletion records being correctly and incorrectly recalled using AnchorWave and minimap2.
Supporting Note 3:

We downloaded the maize Mo17 genome in FASTA format and renamed chromosomes to be consistent with the B73 genome FASTA file using the commands:

```
wget https://download.maizegdb.org/Zm-Mo17-REFERENCE-CAU-1.0/Zm-Mo17-REFERENCE-CAU-1.0.fa.gz
```

```
gunzip Zm-Mo17-REFERENCE-CAU-1.0.fa.gz
```

```
sed -i 's/>chr/>/g' Zm-Mo17-REFERENCE-CAU-1.0.fa
```

The B73 V4 reference genome sequence and genome annotation were downloaded as described in Supporting Note 2. The reference full-length CDSs were extracted and mapped to the reference genome and the query genome:

```
anchorwave gff2seq -r Zea_mays.AGPv4.dna.toplevel.fa -i Zea_mays.AGPv4.34.gff3 -o cds.fa
```

```
minimap2 -x splice -t 6 -k 12 -a -p 0.4 -N 20 Zm-Mo17-REFERENCE-CAU-1.0.fa cds.fa > cds.sam
```

```
minimap2 -x splice -t 6 -k 12 -a -p 0.4 -N 20 Zea_mays.AGPv4.dna.toplevel.fa cds.fa > ref.sam
```

We performed genome alignment using the following command:

```
anchorwave genoAli -i Zea_mays.AGPv4.34.gff3 -as cds.fa -r Zea_mays.AGPv4.dna.toplevel.fa -a cds.sam -ar ref.sam -s Zm-Mo17-REFERENCE-CAU-1.0.fa -n anchors -o anchorwave.maf -f anchorwave.f.maf -1 1 -IV
```

The minimap2 and MUMmer4 genome alignments were reformatted into MAF format using the "anchorwave sam2maf" function. The minimap2, MUMmer4, and GSalign results were purified into one-to-one alignment using the "last-split | maf-swap | last-split | maf-swap" as described in Supporting Note 1.

The recall of TE PAVs was evaluated using a script available under the AnchorWave source code repository at "./src/tests/scripts/evaluateTePavALignment.pl". This script evaluates whether deletions from a sam file overlap precisely the positions of features in an input bed file. We used site-defined TE PAVs that differ between B73 and Mo17 from Anderson et al. 2019 (9) (https://github.com/SNAnderson/maizeTE_variation/raw/master/non-redundant_TEs_4genomes_1Feb19.txt.gz), adding the length of a target site duplication (TSD) appropriate for the TE superfamily to generate a bed file.
Supporting Note 4:

We downloaded the maize B73-Ab10 genome (23) in FASTA format, and renamed the entry names to be consistent with the B73 genome file using the commands:

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/902714155.1_Zm-B73_AB10-REFERENCE-NAM-1.0b/GCA_902714155.1_Zm-B73_AB10-REFERENCE-NAM-1.0b_genomic.fna.gz
```

```
gunzip GCA_902714155.1_Zm-B73_AB10-REFERENCE-NAM-1.0b_genomic.fna.gz
```

```
sed -i 's/, whole genome shotgun sequence//' GCA_902714155.1_Zm-B73_AB10-REFERENCE-NAM-1.0b_genomic.fna
```

```
GCA_902714155.1_Zm-B73_AB10-REFERENCE-NAM-1.0b_genomic.fna
```

The B73 V4 reference genome sequence and genome annotation were downloaded as described in Supporting Note 2. The reference full-length CDSs were extracted and mapped to the reference genome and query genome as introduced in Supporting Note 3.

We detected collinear anchors using the following command:

```
anchorwave genoAli -i Zea_mays.AGPv4.34.gff3 -r Zea_mays.AGPv4.dna.toplevel.fa -as cds.fa -a cds.sam -ar ref.sam -s
```

```
GCA_902714155.1_Zm-B73_AB10-REFERENCE-NAM-1.0b_genomic.fna -n anchorsiv -lV -ns
```

We further test the performance of AnchorWave for inversion detection by comparing with inversions reported between the maize B73 and SK lines (24). We downloaded the maize SK genome in FASTA format and renamed entry names to be consistent with the B73 genome file using commands:

```
wget ftp://download.big.ac.cn/gwh/Plants/Zea_mays_the_genome_of_SK_GWAACS00000000/GWAACS00000000.genome.fasta.gz
```

```
gunzip GWAACS00000000.genome.fasta.gz
```

```
sed -i 's/>.*Chromosome/>/g' GWAACS00000000.genomefasta
```

```
sed -i 's/>Zea mays genome assembly, chromosome:/>g' GWAACS00000000.genomefasta
```

The reference full-length CDSs were extracted and mapped to the reference genome and the query genome as described in Supporting Note 3.

The CDS mapping on the query genome is visualized as Figure SN2.
The anchor matches between the maize B73 V4 genome assembly and the SK assembly. Blue dots represent anchors on the same strand between the reference genome and the query genome. Red dots represent anchors on opposite strands between the reference genome and the query genome.

We detected collinear anchors using the following command:

```
anchorwave genoAli -i Zea_mays.AGPv4.34.gff3 -r Zea_mays.AGPv4.dna.toplevel.fa -as cds.fa -a cds.sam -ar ref.sam -s GWAACS00000000.genome.fasta -n anchorsiv -IV -ns
```

The collinear anchors in the output file (anchorsiv) were plotted for each chromosome separately.
Collinear anchors identified between the B73 v4 assembly and the SK assembly. Collinear anchors on the positive strand are colored with semi-transparent blue, and the negative strand anchors are colored with red. Previously reported inversions are labeled with arrows.

AnchorWave identified 11 out of 13 reported inversions (24). For the two previously reported inversions it failed to identify, AnchorWave detected colinear gene anchors in the region, suggesting these regions may not be true inversions. The first uncalled inversion was previously reported in the range of Chr3:86.8 Mbp-88.3 Mbp of the B73 v4 assembly and Chr3:85.2 Mbp-87.0 Mbp of the SK assembly. We found 7 full-length CDS collinear anchors located in the range of Chr3:87.0 Mbp-88.1 Mbp of the B73 v4 assembly and Chr3:88.5 Mbp-89.6 Mbp of the SK assembly. These anchors are located on the same strand of the B73 assembly and the SK assembly. The second inversion was previously reported in the range of Chr4:168.3 Mbp-169.1 Mbp of the B73 v4 assembly and Chr4:165.1 Mbp-167.0 Mbp of the SK assembly. We found 7 full-length CDS collinear anchors located in the range of Chr4:168.6 Mbp-169.0 Mbp of the B73 v4 assembly and Chr4:170.2 Mbp-170.7 Mbp of the SK assembly. Again, these anchors are located on the same strand in the B73 assembly and the SK assembly. We found an additional inversion in an adjacent upstream region of the B73 assembly (Chr4:165.9 Mbp-166.6 Mbp) and from Chr4:168.4 Mbp-169.0 Mbp in the SK assembly. These results indicate that using full-length CDSs as anchors provides sensitivity to capture long inversions.
Supporting Note 5:
The tomato SL2.5 genome sequence, genome annotation and a potato tetraploid assembly were downloaded via the commands:

```bash
wget ftp://ftp.ensemblgenomes.org/pub/release-41/plants/fasta/solanum_lycopersicum/dna/Solanum_lycopersicum.SL2.50.dna.toplevel.fa.gz
wget ftp://ftp.ensemblgenomes.org/pub/release-41/plants/gff3/solanum_lycopersicum/Solanum_lycopersicum.SL2.50.41.gff3.gz
gunzip *gz
wget https://websafe.mpipz.mpg.de/d/B4J4W17Lp4/new_He1.fa
wget https://websafe.mpipz.mpg.de/d/Upxtb7j1P/new_He2.fa
wget https://websafe.mpipz.mpg.de/d/h4P80fhlxRL/new_St1.fa
wget https://websafe.mpipz.mpg.de/d/14ww42hcd8/new_St2.fa
cat new_He1.fa new_He2.fa new_St1.fa new_St2.fa > potato.fa
```

The reference full-length CDSs from tomato were extracted and mapped to the reference genome and query genome:

```bash
anchorwave gff2seq -r Solanum_lycopersicum.SL2.50.dna.toplevel.fa -i Solanum_lycopersicum.SL2.50.41.gff3 -o tomato.cds.fa
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Solanum_lycopersicum.SL2.50.dna.toplevel.fa tomato.cds.fa > tomato.sam
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 potato.fa tomato.cds.fa > potato.sam
```

The full-length CDS mapping hits in the file potato.sam are plotted as Figure SN4:

![Figure SN4](image)

Figure SN4

The anchor matches between the tomato SL2.5 genome assembly and the potato tetraploid assembly. Blue dots represent anchors on the same strand between the reference genome and the query genome. Red dots represent anchors on the opposite strands between the reference genome and the query genome.

The genome alignment using AnchorWave was conducted via:

```bash
anchorwave proali -i Solanum_lycopersicum.SL2.50.41.gff3 -as tomato.cds.fa -r Solanum_lycopersicum.SL2.50.dna.toplevel.fa -a potato.sam -ar tomato.sam -s potato.fa -n potato.anchors -o potato.maf -f potato.f.maf -Q 1 -R 4
```
Collinear anchors between the tomato genome and the potato genome are plotted in Figure SN5:

Collinear anchors between the tomato SL2.5 genome assembly and the potato tetraploid assembly.

- Blue points represent anchors on the same strand between the reference genome and the query genome.
- Red points represent anchors on opposite strands between the reference genome and the query genome.

The alignments was reformatted into bam format:

```
maf-convert sam potato.maf | samtools view -O BAM --reference Solanum_lycopersicum.SL2.50.dna.toplevel.fa - | samtools sort - > potato.anchorwave.bam
```

The sequence alignments using LAST, MUMmer4, minimap2, and GSAlign were conducted with the same settings as introduced in Supporting Note 1.

We counted the number of position match sites using the command:

```
samtools depth potato.anchorwave.bam | awk '{if ($3>0) print $0} ' | wc -l
```

We counted the number of aligned sites using the command:

```
samtools depth potato.anchorwave.bam | wc -l
```

The number of gap sites is calculated as the number of aligned sites minus the number of position match sites, and the number of unaligned sites is the tomato genome size minus the number of aligned sites.
The tomato SL2.5 TE annotation file has been released at:
https://solgenomics.net/jbrowse_solgenomics/?data=data%2Fjson%2FSL2.50&loc=SL2.50ch01%3A9859544..88694298&tracks=DNA%2CRepeatScout&highlight= . The REPET, RepeatScout, and Tallymer tracks were downloaded separately, and merged together as sl2.5.TE.bed. Supplementary dataset 1 from https://doi.org/10.1105/tpc.17.00581 was organized as ATAC_tomato.2.5.bed manually. They were further re-organized using

```bash
bedtools (25):
cat sl2.5.TE.bed | awk '{print($1"\t"$2"\t"$3)}' > sl2.5.TE.simple.bed
bedtools subtract -A -a ATAC_tomato.2.5.bed -b sl2.5.TE.simple.bed > ATAC_tomato.2.5.nonTE.A.bed
```

The commands used to summarize alignments in TE and ATAC peak regions will be described in Supporting Note 7.
Supporting Note 6:
The common goldfish (Carassius auratus) experienced a whole-genome duplication since diverging from the common ancestor with zebrafish, and collinearity between the zebrafish genome assembly and the goldfish genome assembly exists (26). As the goldfish genome assembly size is comparable to that of zebrafish genome assembly (27), fractionation and loss of some homologous zebrafish genome fragments have occurred in goldfish. Each of the two subgenomes of goldfish is found in separate linkage groups; this absence of subgenome chromosome fusion makes the genome comparison more straightforward than that of other whole-genome duplications, such as the maize-sorghum genome comparison.

We downloaded the zebrafish genome assembly, genome annotation, and the goldfish genome assembly using the following commands:

wget ftp://ftp.ensembl.org/pub/release-102/fasta/danio_rerio/dna/Danio_rerio.GRCz11.dna.primary_assembly.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/gff3/danio_rerio/Danio_rerio.GRCz11.102.chr.gff3.gz
wget https://research.nhgri.nih.gov/goldfish/download/carAur01.sm.fa

Following previous work (26), to avoid heterozygous assembled regions, we used only linkage group assemblies when comparing genomes. The linkage group assemblies were extracted from the original FASTA file using the following command:

head -15923983 carAur01.sm.fa > goldfish.fa

We extracted the zebrafish full-length CDS and mapped full-length CDS to the zebrafish reference genome and goldfish query genome as introduced in Supporting Note 1. The mapping matches are plotted in Figure SN6.
Figure SN6

Anchor matches between the zebrafish genome and the goldfish genome. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

We performed genome alignment using AnchorWave with command:

```
anchorwave proali -i Danio_rerio.GRCz11.102.chr.gff3 -r Danio_rerio.GRCz11.dna.primary_assembly.fa -a carAur.sam -as cds.fa -ar ref.sam -s goldfish.fa -n align1.anchors -R 2 -Q 1 -ns
```

Collinear anchors between the zebrafish genome and the goldfish genome are plotted in Figure SN7:
Collinear anchors between the zebrafish genome assembly and the goldfish genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

We split the goldfish genome into two subgenomes via SeqKit(28) using commands:

```bash
seqkit grep -n -p LG1 goldfish.fa > goldfish1.fa
seqkit grep -n -p LG2 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG3 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG4 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG5 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG6 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG7 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG8 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG9 goldfish.fa >> goldfish1.fa
```
seqkit grep -n -p LG10 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG11 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG12 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG13 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG14 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG15 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG16 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG17 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG18 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG19 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG20 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG21 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG22 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG23 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG24 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG25 goldfish.fa >> goldfish1.fa
seqkit grep -n -p LG26 goldfish.fa > goldfish2.fa
seqkit grep -n -p LG27 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG28 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG28B goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG29 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG30 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG30F goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG31 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG32 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG33 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG34 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG35 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG36 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG36F goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG37 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG37M goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG38 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG39 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG40 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG41 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG42 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG42F goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG43 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG44 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG44F goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG45 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG45M goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG46 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG47 goldfish.fa >> goldfish2.fa
seqkit grep -n -p LG48 goldfish.fa >> goldfish2.fa
We aligned the two subgenomes of goldfish separately against the zebrafish genome using LAST, minimap2, MUMmer4, and GSAAlign with the same parameters as described in Supporting Note 1. The minimap2 and MUMmer4 genome alignments were reformatted into MAF format using the “anchorwave sam2maf” function. The minimap2, MUMmer4, and GSAAlign results were filtered to one-to-one alignments using the “last-split | maf-swap | last-split | maf-swap” as described in Supporting Note 1. We then merged the alignment of the two subgenomes together. The proportion of the zebrafish genome aligned as position match and gap are plotted in Figure SN8.
Figure SN8

Comparison of the proportion of sites in the zebrafish genome aligned to the goldfish genome using AnchorWave and other genome alignment tools.
Supporting Note 7:
The maize B73 v4 genome sequence and genome annotation were obtained as introduced in Supporting Note 2, and the sorghum genome (29) was downloaded via the commands:

```
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/sorghum_bicolor/dna/Sorghum_bicolor.Sorghum_bicolor_NCBIv3.dna.toplevel.fa.gz

```

```
gunzip Sorghum_bicolor.Sorghum_bicolor_NCBIv3.dna.toplevel.fa.gz

```

The reference full-length CDSs from maize were extracted and mapped to the reference genome and query genome:

```
anchorwave gff2seq -r Zea_mays.AGPv4.dna.toplevel.fa -i Zea_mays.AGPv4.34.gff3 -o cds.fa

```

```
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Sorghum_bicolor.Sorghum_bicolor_NCBIv3.dna.toplevel.fa cds.fa > cds.sam

```

```
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Zea_mays.AGPv4.dna.toplevel.fa cds.fa > ref.sam

```

The genome alignment using AnchorWave was conducted via:

```
anchorwave proali -i Zea_mays.AGPv4.34.gff3 -as cds.fa -r Zea_mays.AGPv4.dna.toplevel.fa -a cds.sam -ar ref.sam -s Sorghum_bicolor.Sorghum_bicolor_NCBIv3.dna.toplevel.fa -n anchors -R 1 -Q 2 -o alignment.maf -f alignment.f.maf

```

The alignment was reformatted into bam format:

```
maf-convert sam anchorwave.maf | sed 's/Sorghum_bicolor.Sorghum_bicolor_NCBIv3.dna.toplevel.fa.//' | sed 's/Zea_mays.AGPv4.dna.toplevel.fa.//' | samtools view -O BAM --reference Zea_mays.AGPv4.dna.toplevel.fa - | samtools sort > anchorwave.bam

```

The sequence alignments using LAST, MUMmer4, minimap2, and GSAAlign were conducted with the same settings as introduced in Supporting Note 1.

Fig. S5 illustrates how we define position match sites, aligned sites, gap sites, and unaligned sites on the reference genome in this study.

We counted the number of position match sites using the command:

```
samtools depth alignment.bam | awk '$3>0 (print $0)' | wc -l

```

We counted the number of aligned sites using the command:

```
samtools depth alignment.bam | wc -l

```
The number of gap sites is calculated as the number of aligned sites minus the number of position match sites, and the number of unaligned sites is the maize genome size minus the number of aligned sites.

The maize TFBS annotation (all_reproducible_peaks_summits_merged.bed) is available as “Supporting Data 5” from a previous publication (30). We used the following commands to count the number of reference sites being aligned, and those aligned as position match in maize TFBS regions:

```
SAMtools depth alignment.bam -b all_reproducible_peaks_summits_merged.bed | wc -l
SAMtools depth alignment.bam -b all_reproducible_peaks_summits_merged.bed | awk '$3>0{print $0}' | wc -l
```

The maize TE annotation file has been released at: https://github.com/mcstitzer/maize_TEs/blob/master/B73.structuralTEv2.fulllength.2018-09-19.gff3.gz. We use the following command to transform it into bed format:

```
grep -v "#" B73.structuralTEv2.disjoined.2018-09-19.gff3 | awk '{print $1"	"$4-1"	"$5}'} > B73.structuralTEv2.disjoined.2018-09-19.bed
```

We used the following commands to count, in maize TE regions, the number of reference sites being aligned and aligned as position matches:

```
SAMtools depth alignment.bam -b B73.structuralTEv2.disjoined.2018-09-19.bed | wc -l
SAMtools depth alignment.bam -b B73.structuralTEv2.disjoined.2018-09-19.bed | awk '$3>0{print $0}'} | wc -l
```
Supporting Note 8:

The soybean genome sequence, genome annotation\(^6\) and the common bean genome assembly\(^7\) were downloaded via the commands:

```
wget ftp://ftp.ensemblgenomes.org/pub/release-41/plants/gff3/glycine_max/Glycine_max.Glycine_max_v2.0.41.gff3.gz
wget ftp://ftp.ensemblgenomes.org/pub/release-41/plants/fasta/glycine_max/dna/Glycine_max.Glycine_max_v2.0.dna.toplevel.fa.gz
wget ftp://ftp.ensemblgenomes.org/pub/release-51/plants/fasta/phaseolus_vulgaris/dna/Phaseolus_vulgaris.PhaVulg1_0.dna.toplevel.fa.gz
```

The reference full-length CDSs from soybean were extracted and mapped to the reference genome and query genome:

```
anchorwave gff2seq -r Glycine_max.Glycine_max_v2.0.dna.toplevel.fa -i Glycine_max.Glycine_max_v2.0.41.gff3 -o gm.cds.fa
```

```
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Glycine_max.Glycine_max_v2.0.dna.toplevel.fa gm.cds.fa > gm.sam
```

```
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Phaseolus_vulgaris.PhaVulg1_0.dna.toplevel.fa gm.cds.fa > pv.sam
```

The full-length CDS mapping hits in pv.sam file was plotted as Figure SN9:

![Figure SN9](image-url)
The anchor matches between the soybean genome assembly and the common bean assembly. Blue dots represent anchors on the same strand between the reference genome and the query genome. Red dots represent anchors on the opposite strands between the reference genome and the query genome.

By looking into the above plot, we observe some anchors have two copies in a single chromosome from soybean, especially chromosome 5 and chromosome 11. These likely arise from chromosome fusions following the shared whole genome duplication between common bean and soybean. We changed the value of the -e parameter of the anchorwave proali command from its default value of 1 to 2 for genome alignment, allowing two anchor matches on a single chromosome:

```
anchorwave proali -i Glycine_max.Glycine_max_v2.0.41.gff3 -r Glycine_max.Glycine_max_v2.0.dna.toplevel.fa -as gm.cds.fa -a pv.sam -ar gm.sam -s Phaseolus_vulgaris.PhaVulg1_0.dna.toplevel.fa -n pv.anchors -o pv.maf -f pv.f.maf -R 1 -Q 2 -e 2
```

Collinear anchors between the soybean genome and the common bean genome are plotted in Figure SN10:

**Figure SN10**
Collinear anchors between the soybean genome assembly and the common bean assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

The alignments were reformatted into bam format:

```
maf-convert sam pv.maf | samtools view -O BAM --reference Glycine_max.Glycine_max_v2.0.dna.toplevel.fa | samtools sort |>
pv.bam
```

The sequence alignments using LAST, MUMmer4, minimap2, and GSAlign were conducted with the same settings as introduced in Supporting Note 1.

The soybean TE and accessible chromatin regions (ACRs) annotations are from a previous publication(8) and further organized using bedtools.

```
bedtools merge -i Soybean_10days_leaf_ACRs.bed > Soybean_10days_leaf_ACRs_uniq.bed
bedtools sort -i Gm.RM_TE_annotations.bed | bedtools merge > Gm.RM_TE_annotations_uniq.bed
```
Supporting Note 9:
Alignment with AnchorWave relies on the presence of collinear blocks between species. We investigated the power of AnchorWave for aligning species with different levels of divergence. We selected several species to span different phylogenetic distances and rounds of whole-genome duplication as illustrated at:

https://genomevolution.org/wiki/index.php/Whole_genome_duplication.

Maize and rice are among the most divergent grass species with available genome assemblies.

We downloaded the rice genome using commands:

```
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/oryza_sativa/dna/Oryza_sativa.IRGSP-1.0.dna.toplevel.fa.gz
gunzip Oryza_sativa.IRGSP-1.0.dna.toplevel.fa.gz
```

We mapped maize full-length CDS sequence using minimap2 and identified collinear blocks using AnchorWave:

```
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Oryza_sativa.IRGSP-1.0.dna.toplevel.fa maize_cds.fa > maize.rice.sam
```

```
anchorwave proali -i Zea_mays.AGPv4.34.gff3 -r Zea_mays.AGPv4.dna.toplevel.fa -as maize_cds.fa -a maize.rice.sam -ar
maize_cds_maize.sam -s Oryza_sativa.IRGSP-1.0.dna.toplevel.fa -n B73_rice.anchorspro -R 1 -Q 2 -t 18 -o maize.rice.maf
```

1.82 Gbp of the maize genome was identified in collinear blocks with the rice genome.

Figure SN11
Collinear anchors between maize B73 v4 genome assembly and the rice genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

There have been six rounds of whole-genome duplication since maize and banana diverged. We downloaded the banana genome and aligned the banana genome against the maize genome.

We downloaded the banana genome and aligned the banana genome against the maize genome.

```
wget https://www.genoscope.cns.fr/externe/plants/data/Mschizocarpa_chromosomes.fasta
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Mschizocarpa_chromosomes.fasta maize_cds.fa > maize.banana.sam
```

```
anchorwave proali -i Zea_mays.AGPv4.34.gff3 -r Zea_mays.AGPv4.dna.toplevel.fa -as maize_cds.fa -a maize.banana.sam -ar maize_cds_maize.sam -s Mschizocarpa_chromosomes.fasta maize.cds.fa -n B73_banana.anchorspro -R 8 -Q 8 -t 18 -o maize.banana.maf
```

**Figure SN12**

Collinear anchors between the maize B73 v4 genome assembly and the banana genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

The collinear blocks between maize and banana were short, thus AnchorWave genome alignment between maize and banana did not recall homologous blocks. This represents an upper bound for divergence via whole-genome duplication between species for applying AnchorWave.
The divergence time between Arabidopsis (*Arabidopsis thaliana*) and cacao (chocolate) is longer than that between maize and banana, but the species are only separated by two rounds of whole-genome duplication. We downloaded the chocolate genome and aligned the chocolate genome against the Arabidopsis genome.

```
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/theobroma_cacao_matina/dna/Theobroma_cacao_20110822.dna.toplevel.fa.gz

gunzip Theobroma_cacao_matina.Theobroma_cacao_20110822.dna.toplevel.fa.gz

minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Theobroma_cacao_matina.Theobroma_cacao_20110822.dna.toplevel.fa
tair10_cds.fa > ara.chocolate.sam
```

**Figure SN13**

Collinear anchors between the Arabidopsis TAIR10 genome assembly and the chocolate genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

There are higher proportions of collinear blocks found by AnchorWave between the Arabidopsis genome assembly and the chocolate genome assembly than between the maize genome assembly and the banana assembly. This indicates that whole-genome duplications and subsequent decay and fractionation may be one of the key mechanisms that break collinearity, and should be considered when selecting a genome alignment approach.

There are the same number of whole-genome duplications between Arabidopsis and grape as there are between Arabidopsis and chocolate. However, the divergence time between
Arabidopsis and grape is longer than that between Arabidopsis and chocolate. We downloaded the grape genome and aligned the grape genome against the Arabidopsis genome.

```bash
wget ftp://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/vitis_vinifera/dna/Vitis_vinifera.12X.dna.toplevel.fa.gz
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Vitis_vinifera.12X.dna.toplevel.fa tair10_cds.fa > ara.grape.sam
anchorwave proali -i TAIR10_GFF3_genes.gff -r tair10.fa -as tair10_cds.fa -a ara.grape.sam -ar tair10.sam -s Vitis_vinifera.12X.dna.toplevel.fa -n ara_grape.anchorspro -R 1 -Q 4 -t 18 -o ara_grape.maf
```

**Figure SN14**
Collinear anchors between the Arabidopsis TAIR10 genome assembly and the grape genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Those red dots represent anchors on the opposite strands between the reference genome and the query genome.

Fewer Arabidopsis base-pairs are found in collinear blocks with the grape genome than with the chocolate genome. Collinearity may decay along with genomic divergence and reduce the applicability of AnchorWave for genome alignment.

The divergence time between Arabidopsis and tomato is greater than Arabidopsis and grape, and the tomato lineage has an additional paleohexaploidy. We downloaded the tomato genome and aligned the Arabidopsis genome against the tomato genome.

```bash
wget ftp://ftp.ensemblgenomes.org/pub/release-41/plants/fasta/solanum_lycopersicum/dna/Solanum_lycopersicum.SL2.50.dna.toplevel.fa.gz # tomato
gunzip Solanum_lycopersicum.SL2.50.dna.toplevel.fa.gz
minimap2 -x splice -t 11 -k 12 -a -p 0.4 -N 20 Solanum_lycopersicum.SL2.50.dna.toplevel.fa tair10_cds.fa > ara.tomato.sam
anchorwave proali -i TAIR10_GFF3_genes.gff -r tair10.fa -as tair10_cds.fa -a ara.tomato.sam -ar tair10.sam -s Solanum_lycopersicum.SL2.50.dna.toplevel.fa -n ara_tomato.anchorspro -ns -R 3 -Q 4
```
Figure SN15

Collinear anchors between the Arabidopsis TAIR10 genome assembly and the tomato genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Those red dots represent anchors on the opposite strands between the reference genome and the query genome.

Only 36.17% of the Arabidopsis genome is found in collinear blocks with tomato, again highlighting the consequences of divergence, and the importance of collinearity for AnchorWave alignment.

A guideline document to check collinearity and select appropriate parameters of “-Q” and “-R” can be found under the AnchorWave source code repository.
Supporting Note 10:

We tested AnchorWave to perform genome alignment between the human genome (hg38) and the house mouse (*Mus musculus*) genome (mm39).

We downloaded the mouse genome using the following commands:

- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr1.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr2.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr3.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr4.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr5.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr6.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr7.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr8.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr9.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr10.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr11.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr12.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr13.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr14.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr15.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr16.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr17.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr18.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chr19.fa.gz`
- `wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chrM.fa.gz`
wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/chromosomes/chrX.fa.gz

gunzip *gz

cat chr*fa > mm39.fa

We downloaded the human genome annotation using the following commands:

wget ftp://ftp.ensembl.org/pub/release-102/gff3/homo_sapiens/Homo_sapiens.GRCh38.102.gff3.gz

We downloaded the human genome using the following commands:

wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.1.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.2.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.3.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.4.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.5.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.6.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.7.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.8.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.9.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.10.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.11.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.12.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.13.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.14.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.15.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.16.fa.gz
wget ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.17.fa.gz
We extracted the human full-length CDS and mapped to the human genome and mouse genome separately using the following commands:

```
anchorwave gff2seq -r hg38.fa -i Homo_sapiens.GRCh38.102.gff3 -o cds.fa
```

```
minimap2 -x splice -t 10 -k 12 -a -p 0.4 -N 20 hg38.fa cds.fa > ref.sam
```

```
minimap2 -x splice -t 10 -k 12 -a -p 0.4 -N 20 mm39.fa cds.fa > cdss.sam
```

The genome sequence alignment was conducted using the following command:

```
anchorwave proali -i Homo_sapiens.GRCh38.102.gff3 -r hg38.fa -a cds.sam -as cds.fa -ar ref.sam -s mm39.fa -n mm39.anchors -R
```

The collinear anchors are plotted in Figure SN16.
Collinear anchors between the human (hg38) genome assembly and mouse (mm39) genome assembly. Each point is plotted at its start position of each anchor on the reference (hg38) genome and the query genome (mm39). Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

On the autosomes and the X chromosome, 2.613Gbp (86.23%) of the human genome is aligned using AnchorWave. This is larger than that of commonly used chain and net (33.06%, 1.002Gbp) (Figure S12). The number of sites being aligned as position matches is comparable between these two methods (AnchorWave:891.9 Mbp, chain and net: 892.3 Mbp).
Comparison of the proportion of human (hg38) genome sites being aligned to mouse (mm39) using AnchorWave versus genome chain and net genome alignment. The chain and net genome alignment is downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg38/vsMm39/hg38.mm39.synNet.maf.gz.

To identify overlap with putative regulatory sequence, we downloaded ENCODE human candidate cis-regulatory elements via commands:

```
wget http://hgdownload.soe.ucsc.edu/gbdb/hg38/encode3/ccre/encodeCcreCombined.bb
bigBedToBed encodeCcreCombined.bb encodeCcreCombined.bed
sed -i -E 's/^chr//g' encodeCcreCombined.bed
bedtools sort -i encodeCcreCombined.bed | bedtools merge | grep -v "^Y" > encodeCcreCombined_merged.bed
```

AnchorWave aligned 121.15 Mbp (48%) of cis-regulatory elements as position match, which is slightly lower than the downloaded chain and net alignment (131.68 Mbp 52%).
Supporting Note 11:
We tested AnchorWave for collinear block identification between the human genome (hg38) and the chimpanzee (Pan troglodytes) genome (panTro3).

We downloaded the chimpanzee genome using the following commands:
```
wget ftp://ftp.ensembl.org/pub/release-102/fasta/pan_troglodytes/dna/Pan_troglodytes.Pan_tro_3.0.dna.toplevel.fa.gz
```

We downloaded the human genome and genome annotation as described in Supporting Note 10.

We extracted the human full-length CDS and mapped to the human genome and chimpanzee genome separately using the following commands:
```
anchorwave gff2seq -r hg38.fa -i Homo_sapiens.GRCh38.102.gff3 -o cds.fa
```

```
minimap2 -x splice -t 10 -k 12 -a -p 0.4 -N 20 hg38.fa cds.fa > ref.sam
```

```
minimap2 -x splice -t 10 -k 12 -a -p 0.4 -N 20 Pan_troglodytes.Pan_tro_3.0.dna.toplevel.fa cds.fa > cds.sam
```

The full-length CDS mapping is visualized as Figure SN17, and the Y chromosome is shown in Figure SN18. This pattern is consistent with previous reports of collinearity between human and chimpanzee(31).

![Figure SN18](image)
Anchor matches between the human (hg38) genome assembly and the chimpanzee (panTro3) genome assembly. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on the opposite strands between the reference genome and the query genome.

Figure SN19
Anchor matches between the human (hg38) Y chromosome and the chimpanzee (panTro3) Y chromosome. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on opposite strands between the reference genome and the query genome.

The genome sequence alignment was conducted using the following command:
```
anchorwave proali -i Homo_sapiens.GRCh38.102.gff3 -r hg38.fa -a cds.sam -as cds.fa -ar ref.sam -s Pan_troglodytes.Pan_tro_3.0.dna.toplevel.fa -n 2panTro3.anchors -R 1 -Q 1 -o 2panTro3.maf -f 2panTro3.f.maf
```

The collinear anchors are plotted in Figure SN19.
Figure SN20

The plot of identified collinear anchors between human (hg38) genome assembly and chimpanzee (proTro3) genome assembly. Each point was plotted at the start position on the reference (hg38) genome and the query genome (proTro3) of an anchor. Blue points represent anchors on the same strand between the reference genome and the query genome. Red points represent anchors on the opposite strands between the reference genome and the query genome.

We downloaded the widely used chain and net pipeline genome alignment of panTro3 against hg38 and transformed the alignment into bam format using the following commands:

```
wget http://hgdownload.cse.ucsc.edu/goldenpath/hg38/vsPanTro3/hg38.panTro3.synNet.maf.gz

gunzip hg38.panTro3.synNet.maf.gz

maf-convert sam hg38.panTro3.synNet.maf | sed 's/hg38.chr//g' | sed 's/panTro3.chr//g' > hg38.panTro3.synNet.sam

samtools depth 2panTro3_noY.bam | awk '$1 != "Y" && $3 != "Y" {print $0}' > hg38.panTro3.synNet_noY.sam

`
On the autosomes and the X chromosome, 2.902Gbp (95.73%) of the human genome is aligned using AnchorWave; this is larger than that of chain and net (88.97%, 2.697Gbp) (Figure SN20). The number of sites aligned as position matches is comparable (AnchorWave:2.665Gbp, chain and net: 2.688Gbp). The proportion of nucleotides identical as position match sites between the AnchorWave alignment and the downloaded chain and net alignment from UCSC website are very close to each other (98.5% from AnchorWave and 98.7% from chain and net alignment).

The position match ratio between the human and the chimpanzee genome is significantly higher than that between the maize line B73 and the maize line Mo17. Further, the DNA nucleotide identity ratio at position match sites between maize B73 and maize Mo17 (96.0% using AnchorWave) is lower than that between the human genome and the chimpanzee genome (98.5% using AnchorWave). This is consistent with the widely cited observation that maize lines are more different genetically than a human and a chimpanzee.

AnchorWave aligned 245.78 Mbp (97.13%) of candidate cis-regulatory elements as position match, which is slightly higher than the UCSC released chain and net alignment (244.11 Mbp, 96.59%).

```
samtools depth 2panTro3_noY.bam | awk '$3>0 {print $0}' | wc -l
```

```
samtools depth hg38.panTro3.synNet_noY.bam | wc -l
```

```
samtools depth hg38.panTro3.synNet_noY.bam | awk '$3>0 {print $0}' | wc -l
```
Figure SN21

Comparison of the proportion of human (hg38) genome sites aligned to chimpanzee (panTro3) using AnchorWave versus genome chain and net genome alignment. The chain and net genome alignment is downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg38/vsPanTro3/hg38.panTro3.synNet.maf.gz.
Figure SN22 Comparison of the proportion of human (hg38) genome candidate cis-regulatory elements being aligned to chimpanzee (panTro3) using AnchorWave versus genome chain and net genome alignment.
References:

1. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192 (2013).

2. Tomato Genome Consortium, The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635–641 (2012).

3. H. Sun, et al., Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. bioRxiv, 2021.05.15.444292 (2021).

4. O. Jouffroy, S. Saha, L. Mueller, H. Quesneville, F. Maumus, Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genomics 17, 624 (2016).

5. 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).

6. J. Schmutz, et al., Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183 (2010).

7. J. Schmutz, et al., A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Genet. 46, 707–713 (2014).

8. Z. Lu, et al., The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants 5, 1250–1259 (2019).

9. S. N. Anderson, et al., Transposable elements contribute to dynamic genome content in maize. Plant J. 100, 1052–1065 (2019).

10. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819 (2017).

11. D. Earl, et al., Alignathon: a competitive assessment of whole-genome alignment methods. Genome Res. 24, 2077–2089 (2014).

12. X. Gan, et al., Multiple reference genomes and transcriptomes for Arabidopsis thaliana.
1571  
*Nature* **477**, 419–423 (2011).

1572  13. B. Song, Q. Sang, H. Pei, X. Gan, F. Wang, Complement genome annotation lift over using a weighted sequence alignment strategy. *Front. Genet.* **10**, 1046 (2019).

1574  14. P. Lamesch, *et al.*, The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–10 (2012).

1576  15. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

1578  16. S. M. Kielbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).

1580  17. G. Marçais, *et al.*, MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).

1582  18. H.-N. Lin, W.-L. Hsu, GSAlign: an efficient sequence alignment tool for intra-species genomes. *BMC Genomics* **21**, 182 (2020).

1584  19. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

1586  20. Y. Jiao, *et al.*, Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).

1588  21. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

1590  22. P. Danecek, *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

1592  23. J. Liu, *et al.*, Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol.* **21**, 121 (2020).

1594  24. N. Yang, *et al.*, Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* **51**, 1052–1059 (2019).

1596  25. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic
features. *Bioinformatics* **26**, 841–842 (2010).

26. Z. Chen, *et al.*, De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication. *Sci Adv* **5**, eaav0547 (2019).

27. K. Howe, *et al.*, The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).

28. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).

29. A. H. Paterson, *et al.*, The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).

30. X. Tu, *et al.*, Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat. Commun.* **11**, 5089 (2020).

31. J. F. Hughes, *et al.*, Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).