Chorus2: design of genome-scale oligonucleotide-based probes for fluorescence in situ hybridization

Tao Zhang¹,²,*†, Guanqing Liu¹,², Hainan Zhao³, Guilherme T. Braz³ and Jiming Jiang³,⁴,⁵,*

¹Jiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Agricultural College of Yangzhou University, Yangzhou, China
²Key Laboratory of Plant Functional Genomics of the Ministry of Education/Joint International Research Laboratory of Agriculture and Agri-Product Safety of Ministry of Education of China, Yangzhou University, Yangzhou, China
³Department of Plant Biology, Michigan State University, East Lansing, MI, USA
⁴Department of Horticulture, Michigan State University, East Lansing, MI, USA
⁵Michigan State University AgBioResearch, East Lansing, MI, USA

Received 31 July 2020; revised 11 April 2021; accepted 19 April 2021.
*Correspondence (Tel +86 514 87977229; Fax +86 514 87996817; email zhangtao@yzu.edu.cn (TZ) and Tel +1 517 355 4043; email jiangjm@msu.edu (JJ))
†These authors contributed equally to this work.

Keywords: FISH, oligonucleotide, cytogenetics, genome research.

Summary

Oligonucleotide (oligo)-fluorescence in situ hybridization (FISH) has rapidly become the new generation of FISH technique in plant molecular cytogenetics research. Genome-scale identification of single-copy oligos is the foundation of successful oligo-FISH experiments. Here, we introduce Chorus2, a software that is developed specifically for oligo selection. We demonstrate that Chorus2 is highly effective to remove all repetitive elements in selection of single-copy oligos, which is critical for the development of successful FISH probes. Chorus2 allows to select oligos that are conserved among related species, which extends the usage of oligo-FISH probes among phylogenetically related plant species. We also implemented a new function in Chorus2 that allows development of FISH probes from plant species without an assembled genome. We anticipate that Chorus2 can be used in plants as well as in mammalian and other non-plant species. Chorus2 will broadly facilitate the design of FISH probes for various types of application in molecular cytogenetics research.

Introduction

Fluorescence in situ hybridization (FISH) was initially developed to map DNA sequences on chromosomes (Langer-Safer et al., 1982). FISH was introduced in plants in later 1980s (Schwarzacher et al., 1989) and gradually became the most important technique for plant cytogenetic research (Jiang, 2019; Jiang and Gill, 2006). Successful FISH experiments rely on robust DNA probes. For many years the 5S and 45S ribosomal RNA genes have been the most commonly used FISH probes because the rDNA generate strong FISH signals and can be used universally in all plant species (Fukui et al., 1994; Jiang and Gill, 1994; Leitch and Heslop-Harrison, 1992; Maluszynska and Heslop-Harrison, 1991; Schmidt et al., 1994). However, the number of rDNA loci as well as their chromosomal locations often vary among closely related species in some plant lineages (Badaeva et al., 1996; Datson and Murray, 2006; Fukui et al., 1994; He et al., 2020; Schubert and Wobus, 1985). Thus, the rDNA probes are not reliable markers for chromosome identification or comparative cytogenetic studies. Many tandemly repeated DNA sequences, which have similar structure as the rDNA sequences, are also robust FISH probes and have been used for chromosome identification in plant species (Kato et al., 2004; Pedersen and Langridge, 1997; Tang et al., 2014).

Single-copy DNA sequences or large-insert genomic clones containing single-copy sequences can also be used as FISH probes (Jiang and Gill, 2006). Single-copy sequence-based probes, however, have several limitations. Most single-copy sequence probes, typically containing <5 kb sequences, are time-consuming to develop and do not produce robust FISH signals (Franz et al., 1996; Jiang et al., 1996). By contrast, large-insert genomic clones, especially bacterial artificial chromosome (BAC) clones (Woo et al., 1994), generate strong FISH signals. BAC-based FISH has become a popular tool for chromosome identification in plants (Dong et al., 2000; Howell et al., 2002; Kim et al., 2002; Kulikova et al., 2001; Pedrosa et al., 2002). Nevertheless, most BACs from plant species with large complex genomes contain a high proportion of repetitive DNA sequences and cannot be used as FISH probes (Janda et al., 2006; Suzuki et al., 2012; Zhang et al., 2004). In addition, due to recent technology advances in genome mapping and sequencing, few labs will continue to invest on developing or maintaining of BAC libraries.

A new class of probe based on oligos has rapidly become the next-generation FISH probes in plants (Jiang, 2019). Oligos associated single-copy sequences specific to a chromosomal region or to an entire chromosome, or to a specific genotype (haplotype-specific) can be computationally identified from a species with a sequenced genome(s) (Braz et al., 2018; Han et al., 2015; Martins et al., 2019). These oligos can then be massively synthesized as a pool and labelled as a FISH probe (Beliveau et al., 2012; Han et al., 2015). Such oligo-based FISH probes overcome
most of the major limitations associated with traditional FISH probes: (1) Oligo-FISH probes can be designed from any plant species with a sequenced genome. If a target species is not sequenced, oligo probes designed from a genetically related species can possibly be used in FISH (Braz et al., 2018; Liu et al., 2020; Xin et al., 2020); (2) Each oligo-FISH probe is linked with a known chromosome or a known linkage group. The signal strength of each probe can be adjusted by including different number of oligos; (3) each synthesized oligo library provides enough template DNA for tens of thousands of FISH experiments. Thus, each library can be maintained as an infinite probe resource and shared by a research community (Han et al., 2015).

We previously developed an oligo selection software, Chorus, for designing oligo-FISH probes in plants (Han et al., 2015). We have now significantly upgraded this software as Chorus2, with the main goal to improve its repeat removal efficiency. We also implemented a new function that allow developing probes from plant species without an assembled genome. We conducted comparative analysis between Chorus2 and OligoMiner, a similar pipeline for oligo selection developed in mammalian species (Beliveau et al., 2018). We demonstrate that Chorus2 is a superior pipeline especially for plant species containing highly repetitive genomes.

Results and discussion

The Chorus2 pipeline for oligo selection

The Chorus2 package is implemented with an easy-to-use GUI (graphical user interface) and flexible command-line, and it can be run with Linux, macOS and Windows. Chorus2 uses python script Chorus.py to identify and pre-filter oligos (Figure S1a). A reference genome and a target sequence are required as input files. The target sequence can be a portion of a chromosome, or an entire chromosome, or an entire genome. The oligo filtering process is dependent on a k-mer method (Figure 1). Thus, identification of repetitive sequences is not dependent on a repeat-masked reference genome. This k-mer-based approach is more effective than RepeatMasker (http://www.repeatmasker.org) -based approach to identify and remove repeats, especially the repeats derived from decayed transposable elements (TEs). RepeatMasker may fail to identify repeats that are not well characterized or not assembled in a reference genome. In contrast, all potential repeats can be detected by the k-mer-based pipeline without a reference genome (Price et al., 2005). The implemented Jellyfish software identifies all given k-mers in the input sequences with an ultrafast speed. BWA (Li, 2013), a fast next-generation sequencing (NGS) aligner, is used to align oligos to the reference genome for oligo selection. Primer3 (Untergasser et al., 2012), a widely used programme for designing PCR primers, is implemented to perform thermodynamic analysis. After oligo filtering, Chorus2 outputs two files, one containing all filtered oligos identified from the genome and the second one containing a non-overlapped oligo list.

Repetitive DNA sequences are dominant components in many plant genomes. Genomic regions containing highly repetitive DNA sequences are difficult to assemble, which result in missing or collapsing of the DNA sequences in these regions in a reference genome. Oligos derived from collapsed repetitive sequences or sequences homologous to non-assembled repeats may hybridize to multiple genomic locations, which would increase the FISH background. Thus, a reference genome with misassemblies containing repetitive DNA sequences may impair the specificity of the selected non-overlapping oligos. We implemented an NGS filtering method to further improve the specificity of the selected oligos. First, all retained oligos from Chorus.py output are transmitted to ChorusNGSfilter script as input. A set of shotgun sequence reads from the target species is also required as an input file. Jellyfish is used to calculate a k-mer score (Figure 1), which represents the relative copy number of each oligo in input library. Each oligo will be assigned with a specific score. The ChorusNGSselect script is used to filter oligos based on their k-mer scores. If the k-mer score of an oligo deviates from the total k-mer score distribution, this oligo will be filtered out (Figure S1a). Chorus2 uses a user-friendly graphical user interface, named ChorusGUI, to facilitate oligo design (Figure S1b). Users can readily select the final set of oligos from a specific region with given density or strand with GUI using the ChorusPBGUI script (Figure S1c).

Specificity of FISH probes designed by Chorus2

We developed chromosome painting probes using both Chorus (the first version of the software) and Chorus2 for maize chromosome 1, which is the largest maize chromosome containing 301 megabase (Mb) sequences (Schnable et al., 2009). A total of 58 797 oligos were selected for the long arm of chromosome 1 using Chorus, which was implemented with BLAT for sequence alignment and RepeatMasker for repetitive DNA filtering. We selected 91 265 oligos for the entire chromosome 1 using Chorus2, including 51 283 oligos from the long arm. FISH using the Chorus-designed probe generated significant cross-hybridization signals on all maize chromosomes (Figure 2a), indicating insufficient elimination of repetitive DNA sequences. In contrast, FISH using Chorus2-designed probe generated strong signals on chromosome 1 with much weaker cross-hybridization on other maize chromosomes (Albert et al., 2019; Figure 2b).

We investigated why the Chorus-designed FISH probe produce extensive cross-hybridization to all chromosomes. Both probes were designed based on maize (B73) reference genome AGPv3. The latest version of reference genome (AGPv4) was developed using single-molecule sequencing technologies and high-resolution optical mapping (Jiao et al., 2017). Thus, the quality of AGPv4 is significantly improved compared to AGPv3. We aligned the two sets of oligos to the AGPv4 genome. We found that 2254 oligos generated by Chorus cannot be mapped to AGPv4 (unmapped oligos), and 213 oligos were mapped to multiple positions (multimappable oligos) (Table 1). In contrast, Chorus2 generated only 3 unmappable oligos and no multimappable oligo. The unmappable and multimappable oligos may be caused by the missing or incorrect assembly in the AGPv3 reference genome. Most noticeably, 869 oligos generated by Chorus were also mapped to other chromosomes or non-anchored contigs. Thus, the chromosome specificity of the oligos generated from Chorus2 is less dependent on the quality of the reference genome.

Repeat-associated oligos generated by Chorus

The cross-hybridization signals from the Chorus-designed FISH probe (Figure 2a) are most likely derived from oligos that are associated with repetitive DNA elements, suggesting inadequate removal of repetitive sequences by the Chorus pipeline. This is likely caused by ‘collapsing’ of repetitive sequences during sequence assembly (Salzberg and Yorke, 2005) and/or by some TEs undetected during repeat annotation. Chorus2 is more likely to overcome these shortcomings since it identifies repetitive
sequences based on raw NGS data, which represent an unbiased sequence composition of a genome. We performed k-mer analysis to detect potential repeat-related oligos. The k-mer analysis was conducted on all Chorus and Chorus2 oligos by using random shotgun sequences to examine the repetitiveness of each oligo.
Table 1 Summary of oligos designed by Chorus/Chorus2 mapped to the maize (Zea mays) AGPv4 reference genome

|                     | Chorus | Chorus2 |
|---------------------|--------|---------|
| Total oligos        | 58,797 | 51,283  |
| Uniquely mapped     | 56,330 | 51,280  |
| Multi-mapped        | 213    | 0       |
| Unmapped            | 2254   | 3       |
| Map to other chromos/contigs | 869 | 0 |
| Map to transposable elements (TEs) | 4931 | 3540 |

K-mer analysis

| Putative multi-copy unique mapped oligos | 10,620 | 32 |
| Putative multi-copy unmapped oligos      | 397    | 0  |
| Putative multi-copy multi-mapped oligos  | 139    | 0  |
| Putative single-copy TEs                 | 3780   | 3470 |
| Putative multi-copy TEs                  | 1209   | 2  |

A ‘K-mer score’ was calculated for each oligo to reveal its relative copy number in genome (see Methods). K-mer scoring initially suggested a total of 11,156 putative multi-copy oligos in the Chorus-designed oligo set. To further characterize these putative repeat-associated oligos, we mapped these oligos to the AGPv4 reference genome. Notably, 10,620 oligos were uniquely mapped to the genome. However, 139 oligos were mapped to multiple locations and 397 oligos were unmapped (Table S1), including 104 oligos mapped to two locations, 17 oligos to three locations and 18 oligos to four locations in the AGPv4 genome. In addition, the K-mer score was correlated with the copy number of these multi-copy oligos (Figure 3a).

We predicted that some of the multi-copy oligos are possibly associated with decayed TEs. We further annotated these oligos using the latest TE database (Anderson et al., 2019). We found that 91 of the 139 multi-copy oligos showed sequence homology with annotated TEs (Table S1). Most of these 91 oligos were mapped to different TEs in multiple positions in the reference genome. Most of the TEs are long terminal repeat (LTR) retrotransposons, which have high copy numbers in maize genome. These LTR-TE-related oligos appeared to be effectively detected and removed from the Chorus2 pipeline. For example, one LTR-TE-related Chorus-designed oligo is unique in the AGPv4 reference genome, but this oligo is mapped to three locations in the AGPv4 genome, one on chromosome 1 and two on chromosome 5 (Figure 3b). The K-mer score of the oligo is 15,940 and it was eliminated by the Chorus2 pipeline.

Conservation of oligos among genetically related species

Oligo-FISH is an excellent tool to investigate chromosome evolution among genetically related species (Bačovský et al., 2020; Bi et al., 2020; Braz et al., 2018; Xin et al., 2020). The applicability of an oligo-FISH probe in a different species will depend on the level of sequence conservation of the selected oligos among the two species (Braz et al., 2020). To extend the utility of designed oligo-FISH probes in multiple species, we added a module called ChorusHomo in Chorus2 that allows to select oligos conserved among related species. Briefly, the genome sequences of the reference species and a related target species are provided to Chorus2. Single-copy oligos are selected based on the sequences of the reference species. The oligos are then compared to the genome of the target species using BWA-MEM (Li, 2013) to identify conserved oligos.

The genetic distance between the target species and the reference species will be a key factor to determine the level of conservation of the selected oligos. Selection of highly conserved oligos will increase applicability of the FISH probes in distantly related species. However, this increased stringency will also reduce the number of oligos that can be selected. Chorus2 can generate a synteny map to show the positions of conserved oligos on the chromosomes of reference and target species. For example, we selected conserved oligos between potato and tomato genomes using ChorusHomo. A synteny map was generated showing the positions of the oligos conserved in the two genomes (Figure 4). Oligos selected from the syntenic regions can be used for both potato and tomato FISH experiments.

Oligo-FISH probe design without a reference genome

Although oligo-FISH probes designed from one plant species can be used in genetically related species, the quality of the FISH signals will depend on the level of sequence divergence between the two species. We previously developed two oligo-FISH probes for potato chromosomes 3 and 8, respectively. Each probe contained 27,392 oligos that were selected to uniformly cover the entire chromosomes. These two painting probes nearly uniformly hybridized to the corresponding potato chromosomes except for the centromeric regions, which are composed of highly repetitive DNA sequences (Gong et al., 2012; Figure 2c, Figure 5a). However, these two painting probes generated weak signals on the homoeologous chromosomes in a wild potato species Solanum etuberosum (Figure 2d). FISH signals were nearly not detectable in the pericentromeric regions on the S. etuberosum chromosomes, indicating that the pericentromeric sequences are highly diverged between the two species. Thus, the two potato probes are not ideal for chromosome painting in S. etuberosum due to the significant sequence divergence of these two species. Similarly, chromosome-specific painting probes developed in duckweed species did not generate robust signals in other species across duckweeds genera, most likely due to sequence divergence (Hoang et al., 2021).

We designed ChorusNoRef to develop oligo-FISH probes for plant species, such as S. etuberosum, which does not have a reference genome, however, a reference genome is available from a related species, potato. ChorusNoRef uses short sequence reads, such as those from Illumina sequencing, which can be readily generated with minimum cost. First, oligos from the species with a reference genome (referred as ‘Reference model’ hereafter) are designed by Chorus2. Second, short reads generated from the target species are mapped to the genome of the Reference model. The genomic sequences of the target species corresponding to the oligos from the Reference model are recovered by local assembly using sequence reads that overlap with the oligos from the Reference model. Finally, a new set of oligos are generated by replacing sequences of oligos from the Reference model with sequences from the target species (Figure S2).

We used ChorusNoRef to develop oligos in two wild potato species, S. etuberosum and S. jamesii. We first developed a set of 1,526,873 oligos from the Reference model, potato (DM genome v404 (Hardigan et al., 2016)), using Chorus2. We downloaded publicly available shotgun sequences (100 bp) from S.
etuberosum (46.9 million reads) and S. jamesii (57.6 million reads). These sequences cover approximately 5–6x of the potato genome (884 Mb). ChorusNoRef was then used to develop oligos in these two species. The pipeline generated a total of 483 675 oligos in S. etuberosum. We compared the sequences of these oligos with the corresponding potato oligos. Only 84 116 (17.4%) of the S. etuberosum oligos are identical to the corresponding potato oligos, including 9352 associated with chromosome 3 and 6115 with chromosome 8. However, when we compared to the sequences of the 27 392 oligos that were designed for potato chromosomes 3 and 8, we only found 2205 and 1681 identical S. etuberosum oligos, respectively. In addition, these oligos were relatively depleted in the pericentromeric regions (Figure 5b). These results explained why the two potato painting probes produced weak FISH signals on S. etuberosum chromosomes, especially in the pericentromeric regions (Figure 2d).

In addition to the 84 116 identical oligos, we found that 25 554 S. etuberosum oligos contain insertions compared to corresponding potato oligos, 14 409 with deletions, 203 697 with SNPs (Table 2). A total of 34 933 and 24 419 of these oligos were found on chromosome 3 and 8, respectively (Figure 5c). We recently demonstrated that 15 000 oligos are sufficient to generate FISH signals to cover maize chromosome 10 (150 Mb) (Martins et al., 2019). We found that 15 000 oligos are sufficient to generate FISH signals to cover maize chromosome 10 (150 Mb) (Martins et al., 2019). Thus, we expect that the numbers of oligos generated by ChorusNoRef will generate robust FISH signals on maize chromosome 10 (150 Mb) (Martins et al., 2019).

Comparison between Chorus2 and OligoMiner

OligoMiner (Beliveau et al., 2018) is a recently published tool that is customized for the genome-scale design of oligo-based FISH probes. Thus, we exploited the potential of using OligoMiner for oligo probe design in plants and compared the key functions between Chorus2 and OligoMiner.

Complete elimination of repetitive sequences is the most important factor for successful oligo-FISH probe designs. To compare the effectiveness of repeat elimination of the OligoMiner and Chorus2 pipelines, we designed 2 386 998 and 1 663 941 oligos from maize reference genome AGPv4 using OligoMiner and Chorus2 (both with default parameters), respectively (Table S2). Both software effectively designed and selected a mass of single-copy oligos (Figure 6a). However, OligoMiner generated 352 422 multi-copy oligos with a relative copy number greater than 1 (Figure 6a). We found that 63 714 of these multi-copy oligos are related with TEs. We aligned all oligos to maize AGPv4 reference genome by RMBlast program with e-value 1e-10. We found that 73 858 OligoMiner-designed oligos are mapped to multiple locations, with the most repetitive oligo mapped to 549 locations in the maize genome. In contrast, only 9116 Chorus2 designed oligos were mapped to multiple locations, with the most repetitive oligo mapped to 32 locations (Figure 6b). We investigated the chromosomal distribution of the top 5 repetitive oligos designed by OligoMiner and Chorus2, respectively. We found that the repetitive oligos from OligoMiner are nearly randomly distributed on all 10 maize chromosomes (Figure 6c), which would cause a strong FISH background.
RepeatExplorer2 is a NGS-based software for identification and characterization of repetitive elements (Novak et al., 2013). RepeatExplorer2 uses a sequence clustering algorithm to identify repeats de novo without depending on a reference genome and/or databases with known repeat elements. We used RepeatExplorer2 to identify the top 200 most repetitive DNA sequences in the maize genome. We then mapped the two sets of oligos to these 200 repetitive sequences with RMBlast. We found that 743 OligoMiner-designed oligos are mapped to 133 repetitive sequences. In contrast, only 38 Chorus2-designed oligos are associated with 31 repetitive sequences (Table S3). Therefore, OligoMiner has been successfully used in oligo-FISH probe design in mammalian species. However, the repeat removal system implemented in OligoMiner was not effective to process highly repetitive genomes such as the maize genome.

We also evaluated the performance of the two software, including running time and memory consumption. We used both software to design oligos in three different plant species, including Arabidopsis thaliana (TAIR10), rice (TIGR7) and maize (AGPv4). Chorus2 generated less number of oligos than OligoMiner (Table S2) in all three species. Chorus2 used less time to complete whole-genome probe design for Arabidopsis and maize (Table S4). OligoMiner used less memory. It is likely that OligoMiner designs probes for one chromosome at a time, which may save the computational resource. Chorus2 consumes relatively more memory but runs faster than OligoMiner on the same computer (Table S4). In addition, OligoMiner provides command-line version for users. Chorus2, however, uses a user-friendly GUI (graphical user interface) version that allows users design and select probes conveniently.

Development of oligo datasets

We have developed oligo datasets for both plant and animal species. These datasets can be used by laboratories that do not have bioinformatics capacity. These datasets can be used directly for oligo selection or oligo pool synthesis. The current datasets include nine commonly used species, including Arabidopsis thaliana, rice, maize, potato, barley, soybean, human, mouse and zebrafish (Table S5). The parameters used to design these oligos are provided in Methods. These oligo datasets are available online (http://zhangtaolab.org/download/oligo_datasets or http://jianglab.plantbiology.msu.edu/oligo_datasets.html).

Methods

Chorus2 workflow

The oligo selection procedure of Chorus2 includes four steps: (1) Genome sequence pre-filter; (2) Selection of unique oligos; (3) Thermodynamic analysis; (4) Further filter with NGS data. The detail of each step is described below.

Genome sequence pre-filter

In order to identify all single-copy oligos in a target genome, it is imperative to filter out all repetitive DNA sequences and homopolymer stretches. First, Jellyfish (Marcais and Kingsford, 2011) is used to count and build a k-mer index for the entire target reference genome. Second, each oligo \( O_i \), \( i = (1, 2, 3, \ldots, n) \) is divided into a set of k-mers \( \{ S_{ij} \}, j = (1, 2, \ldots, m, m = l - k + 1) \), \( l \) is oligo length, \( k \) is k-mer length and a oligo score \( C_i \) for oligo \( O_i \) is calculated:
The count and frequency of each k-mer in oligo \( O_i \) are denoted as \( S_{ij} \) and \( f_{ij} \), respectively. The oligo score is calculated as:

\[
T = lh - k
\]

where \( h \) is the homology (default is 75%), here homology means that an oligo can be aligned to another genomic location with at least 75% concordance. Oligos with \( C_i \) greater than \( T \) may target to more than one site under 75% homology in genome, thus these oligos are discarded. The default threshold is \( 16.75 \) which works well in many oligo-FISH works. In addition, an oligo should not contain more than 6 homopolymer stretches, such as AAAAAAA.

Selection of unique oligos

To identify single-copy oligos, the pre-filtered oligos are first mapped to reference genome using BWA software (Li, 2013). In this step, the BWA-MEM algorithm is used for alignment between the oligos and the reference genome. We modify the BWA mem default parameter to “-O, 0, -B, 0, -E, 0, -k, 5, -t”. Mismatch penalty (-B), gap penalty (-O) and gap extension penalty (-E) are set to 0, only matching score will be counted. Besides, minimum
was used for identification of maize genome size and library depth of shotgun reads. First, Jellyfish was utilized to calculate the frequency and depth of 17-mer. A histogram was drawn to illustrate the k-mer spectrum (Figure S3). The exponentially decreasing curve before 35 is noise which is derived from sequencing error. The main peak at 85 represents the sequencing depth of the shotgun library is approximately 85x. The hump beyond the main peak is caused by repetitive sequences. Thus those 17-mers with a depth around 85 are unique in the maize genome. Second, we calculated the k-mer score for all oligos using the NGS filter method mentioned above (see ‘Further filter with NGS data’ section). The k-mer score represents the relative copy number of oligos in the genome. The expected k-mer score of the single-copy oligo is 2,465 \([45 - 17 + 1] \times 85\). Last, we set a standard to divide all oligos into 3 groups based on the k-mer spectrum: Oligos with score less than 1015 \([45 - 17 + 1] \times 35\) are probably noise; Oligos with score between 1015 and 4060 \([45 - 17 + 1] \times 140\) are defined as single-copy oligos which appear once in the genome; Oligos with score greater than 4060 are regarded as multi-copy oligos. A relative copy number of each oligo was calculated using the following formula:

$$R_{Ci} = \frac{R_i}{E}$$

Where \(R_{Ci}\) is relative copy number, \(R_i\) is k-mer score of the oligo(i), \(E\) is expected k-mer score of the single-copy oligo.

Conservation of oligos between species

We used the default parameters of Chorus2 to design oligos of potato genome (DMv404) (Xu et al., 2011). We analysed the conservation between potato and tomato genome (SL3.0) (Sato et al., 2012) with ChorusHomo. Oligos were mapped to the target genome sequences using BWA with parameters ‘-O, 0, -B, 0, -E, 0, -c, 5, -t’. The alignment is output as SAM format and number of matches and mismatches were calculated from the MD tag. The similarity of oligos was calculated as the ratio of matches to the sum of matches and mismatches. Similarity of oligos greater than 90% was kept as conserved oligos. Synteny map was drawn using conserved oligos between potato and tomato in chromosome 4.

Oligo probe design in species without an assembled genome

Oligo probes for target species without reference genomes are able to be designed by ChorusNoRef. First, oligos from the species with a reference genome (referred as ‘Reference model’ hereafter) are developed using Chorus2 and putative repetitive sequences are removed with the assistance of shotgun reads from Reference model. Then, BWA is used to align random genomic reads from the target species to the Reference model, respectively. After that, bcftools is invoked to recover genomic sequences of the target species corresponding to the oligos from the Reference model by local assembly using sequence reads that overlap with the oligos from the Reference model. Next, a quality filter is performed by counting the reads support for each oligo in the Illumina library, oligos covered by less than 3 reads are discarded (referred as ‘low quality’) because they are insufficient to support the oligos. Finally, a new set of oligos are generated by replacing sequences of oligos from the Reference model with sequences from the target species. The homology of oligos from

| Type          | Solanum etuberosum | Solanum jameii |
|---------------|-------------------|----------------|
| Not found     | 1 043 198         | 1 026 092      |
| Low quality   | 155 899           | 79 312         |
| Insertion     | 25 554            | 22 965         |
| Deletion      | 14 409            | 18 162         |
| Identical     | 84 116            | 129 186        |
| 1 SNP         | 88 281            | 120 833        |
| 2 SNPs        | 57 368            | 70 007         |
| 3 SNPs        | 30 571            | 33 535         |
| 4 SNPs        | 14 647            | 15 103         |
| 5 SNPs        | 7202              | 6707           |
| 6 SNPs        | 3322              | 2985           |
| 7 SNPs        | 1458              | 1285           |
| 8 SNPs        | 592               | 487            |
| 9 SNPs        | 195               | 159            |
| 10 SNPs       | 53                | 44             |
| 11 SNPs       | 8                 | 10             |
| 15 SNPs       | --                | 1              |
target species and Reference model is calculated. Oligo probes that perfectly match to the Reference model or with SNPs are retained for FISH experiments.

Comparison between Chorus2 and OligoMiner

Three genomes were downloaded for Chorus2 and OligoMiner pipeline comparison (A. thaliana (Initiative, 2000); Rice (Kawahara et al., 2013); Maize (Jiao et al., 2017)). Shotgun sequencing data of three species were used for ChorusNGSfilter. OligoMiner was run with default parameters and LDA mode, except probe length is set to 45 nt. Chorus2 was run with default parameters. All Chorus2 designed oligos were further filtered with ChorusNGSfilter to keep oligos with k-mer score between 1015 and 4060, k-mer score and relative copy number of each oligo were calculated as mentioned above. Repetitive clusters of maize genome were identified by RepeatExplorer2 pipeline with 10 million paired reads for maize shotgun library SRR2960981 as input. Each RepeatExplorer2 cluster contains several repetitive contigs. Oligos from two software were mapped to RepeatExplorer2 identified contigs and maize AGPv4 reference genome using RMBlast program with parameters: ‘-task rmblastn -evalue 1e-10 -word_size 17 -outfmt 6’. Genome proportion of each repetitive contig was estimated from the results of RepeatExplorer2. Oligo mapped RepeatExplorer2 clusters were removed duplicates and genome proportion of the clusters was calculated by adding all proportions up. All plots were drawn with R or python.

**FISH**

FISH using the chromosome painting probes was performed in maize (Zea mays, inbred B73), in the doubled monoploid S. tuberosum Group Phureja clone DM1-3 516 R44 (DM), and in the wild species S. etuberosum (PI 558289). Root tips were harvested from greenhouse grown plants, treated with nitrous oxide (N₂O) gas at a pressure of 160 psi (~10.9 atm), and fixed in 3 ethanol:1 acetic acid solution. This treatment was 140 min for maize and 20 min for potato and S. etuberosum. The root tips were digested using an enzymatic solution composed of 4% cellulase (Yakult Pharmaceutical, Japan), 2% pectinase (Sigma-Aldrich Co.) and 2% pectolyase (Plant Media) for 1 h at 37°C. Slides were prepared using the stirring method (Ross et al., 1996).

FISH was performed following published protocols (Cheng et al., 2002). Probes labelled with biotin were detected using anti-biotin fluorescein conjugated (Vector Laboratories, Burlingame, California), whereas digoxigenin-labelled probes were detected by anti-digoxigenin rhodamine conjugated (Roche Diagnostics, Indianapolis, Indiana). Chromosomes were counterstained with 4,6-diamidino-2-phenylindole (DAPI) in VectaShield antifade solution (Vector Laboratories). The images were captured using a QImaging Retiga EXi Fast 1394 CCD camera attached to an Olympus BX51 epifluorescence microscope and processed with Meta Imaging Series 7.5 software. The final contrast of the images was processed using Adobe Photoshop software.
Development of oligos in plant and animal species

Oligos for six plant and three animal species were designed with Chorus2 using their respective reference genome. Default parameters were used in the processes of oligo design (Chorus2) and repeat identification (ChorusNGSfilter). Finally, oligos were filtered by k-mer score and min spacing distance (-d 20) using ChorusNGSselect command. Designed oligos were compressed and indexed by bgzip and tabix.

Acknowledgements

We thank for Xueming Yang for contribution in FISH experiments and the image of Figure 2a. This work was supported by the National Transgenic Major Project (2018ZX08020-003) and Funds from Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and Jiangsu Province Government Project (BK2018003) to TZ, grant IOS-1444514 from the National Science Foundation to JI.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

JI and TZ conceived the research; GTB conducted the FISH experiments. TZ, GL and HZ developed the Chorus2 software; GL and TZ analysed the data. JI, TZ and GL wrote the manuscript. All authors read and approved the final manuscript.

Data availability statement

Chorus2 software is deposited at: https://github.com/zhangtaolab/Chorus2. The software tutorial videos have been uploaded to both YouTube and bilibili (https://chorus2.readthedocs.io/en/latest/videos.html). A best practice manual is provided as supplementary file (File S1) to guide users to use the software step by step. Arabidopsis reference genome TAIR10 was downloaded from www.arabidopsis.org (Initiative, 2000). Rice reference genome TIGR7 was downloaded from http://rice.plantbiology.msu.edu/ (Kawahara et al., 2013). Maize reference genome B73 AGPv3 and AGPv4 were download from MaizeGDB (www.maizegdb.org) (Jiao et al., 2017). Potato reference genome DM v404 was downloaded from the PGSC Database (http://solanaceae.plantbiology.msu.edu/) (Xu et al., 2011). Tomato reference genome SL3.0 was downloaded from https://solgenomics.net/ (Sato et al., 2012). barley reference genome IBSC_v2 was downloaded from http://plantsensembl.org/Hordeum_vulgarum/ (Mascher et al., 2017). Soybean reference genome Gmax_ZH13_v2.0 was downloaded from https://bigd.big.ac.cn/gwv/Assembly/652/show (Shen et al., 2019). Human genome hg38, mouse genome mm10 and zebrafish genome danRer11 were downloaded from the UCSC Genome Browser Gateway website (https://hgdownload.soe.ucsc.edu/downloads.html) (Gonzalez et al., 2021). Genomic shotgun sequencing of A. thaliana was retrieved from NCBI Sequence Read Archive (SRA) under accession SRR5658649. Genomic shotgun sequencing of O. sativa was retrieved from NCBI SRA under accession SRR1630928. Genomic shotgun sequencing of Z. mays was retrieved from NCBI SRA under accession SRR2960981. Genomic shotgun sequencing of Solanum tuberosum, Solanum etuberosum and Solanum jamaicensi were retrieved from NCBI SRA under accessions SRR5349606, SRR5349573 and SRR5349574, respectively (Hardigan et al., 2017). Genomic shotgun sequencing of Hordeum vulgare was retrieved from NCBI SRA under accession ERR318755 (Monat et al., 2019). Genomic shotgun sequencing of Glycine max was retrieved from Genome Sequence Archive in Beijing Institute of Genomics (BIG) under accession CRR031689 (Shen et al., 2019). Genomic shotgun sequencing of Homo sapiens was retrieved from NCBI SRA under accession SRR1298980 (Altschuler et al., 2015; Sudmant et al., 2015). Genomic shotgun sequencing of Mus musculus was retrieved from NCBI SRA under accession SRR067844 (Broad Institute). Genomic shotgun sequencing of Danio rerio was retrieved from NCBI SRA under accession SRR10751463 (Freire et al., 2020). All designed oligo-FISH probe datasets are available at website http://zhangtaolab.org/download/oligo_datasets or http://jianglab.plantbiology.msu.edu/oligo_datasets.html.

References

Albert, P.S., Zhang, T., Semrau, K., Rouillard, J.M., Kao, Y.H., Wang, C.J.R., Danilova, T.V. et al. (2019) Whole-chromosome paints in maize reveal rearrangements, nuclear domains, and chromosomal relationships. Proc. Natl. Acad. Sci. USA, 116, 1679–1685.

Altschuler, D.M., Durbin, R.M., Abeasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P. et al. (2015) A global reference for human genetic variation. Nature, 526, 68–74.

Anderson, S.N., Stitzer, M.C., Brohammer, A.B., Zhou, P., Noshay, J.M., O’Connor, C.H., Hirsch, C.D. et al. (2019) Transposable elements contribute to dynamic genome content in maize. Plant J. 100, 1052–1065.

Bačovský, V., Čegan, R., Šimoníková, D., Hřebíková, E. and Holba, R. (2020) The formation of sex chromosomes in Silene latifolia and S. dioica was accompanied by multiple chromosome rearrangements. Front. Plant Sci. 11, 205.

Badaeva, E.D., Friebe, B. and Gill, B.S. (1996) Genome differentiation in Aeglops. 2. Physical mapping of SS and 185–265 ribosomal RNA gene families in diploid species. Genome, 39, 1150–1158.

Beliveau, B.J., Joyce, E.F., Apostolopoulou, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y.M. et al. (2012) Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Natl. Acad. Sci. USA, 109, 21301–21306.

Beliveau, B.J., Kish, J.Y., Nir, G., Sasaki, H.M., Saka, S.K., Nguyen, S.C., Wu, C.T. et al. (2018) OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. Proc. Natl. Acad. Sci. USA, 115, E2183–E2192.

Bi, Y.F., Zhao, Q.Z., Yan, W.K., Li, M.X., Liu, Y.X., Cheng, C.Y., Zhang, L. et al. (2020) Flexible chromosome painting based on multiplex PCR of oligonucleotides and its application for comparative chromosome analyses in Cucumis. Plant J. 102, 178–186.

Braz, G.T., He, L., Zhao, H., Zhang, T., Semrau, K., Rouillard, J.M., Torres, G.A. et al. (2018) Comparative oligo-FISH mapping. An efficient and powerful methodology to reveal karyotypic and chromosomal evolution. Genetics, 208, 513–523.

Braz, G.T., Martins, L.D., Zhang, T., Albert, P.S., Birchler, J.A. and Jiang, J.M. (2020) A universal chromosome identification system for maize and wild Zea species. Chromosome Res. 28, 183–194.

Cheng, Z.K., Buell, C.R., Wang, R.A. and Jiang, J.M. (2002) Resolution of fluorescence in situ hybridization mapping on rice mitotic prometaphase chromosomes, meiotic pachytene chromosomes and extended DNA fibers. Chromosome Res. 10, 379–387.

Datson, P.M. and Murray, B.G. (2006) Ribosomal DNA locus evolution in Nemesia: transposition rather than structural rearrangement as the key mechanism? Chromosome Res. 14, 845–857.
Dong, F.G., Song, J.Q., Naes, S.K., Helgeson, J.P., Gebhardt, C. and Jiang, J.M. (2000) Development and applications of a set of chromosome-specific cytogenetic DNA markers in potato. Theor. Appl. Genet. 101, 1001–1007.
Fransz, P.F., Stam, M., Montijn, B., Ten Hoopen, R., Wiegant, J., Kooter, J.M., Oud, O. et al. (1996) Detection of single-copy genes and chromosome rearrangements in Petunia hybrida by fluorescence in situ hybridization. Plant J. 9, 767–774.
Freire, C., Fish, R.J., Vilar, R., Di Sanza, C., Grzegorzi, S.I., Richter, C.E., Shavit, J.A. et al. (2020) A genetic modifier of venous thrombosis in zebrafish reveals a functional role for fibrinogen AalphaE in early hemostasis. Blood Adv. 4, 5480–5491.
Fukui, K., Ohmido, N. and Khush, G.S. (1994) Variability in rDNA loci in the genus Oryza detected through fluorescence in situ hybridization. Theor. Appl. Genet. 87, 893–899.
Gong, Z.Y., Wu, Y.F., Kobizkova, A., Torres, G.A., Wang, K., Iovene, M., Gonzalez, J.N., Zweig, A.S., Speir, M.L., Schmelter, D., Rosenbloom, K.R., Hardigan, M.A., Crisovan, E., Hamilton, J.P., Kim, J., Laimbeer, P., Leisner, C.P., Jiao, Y.P., Peluso, P., Shi, J.H., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., et al. (1996) Novel 5.8S, 18S, 26S ribosomal RNA gene loci: detection of single-copy genes and chromosome rearrangements in Solanum tuberosum by in situ hybridization. Genome, 45, 401–412.
Kulkova, O., Guattieri, G., Geurts, R., Kim, D.J., Cook, D., Huguet, T., de Jong, J.H. et al. (2001) Integration of the FISH pachytene and genetic maps of Medicago truncatula. Plant J. 27, 49–58.
Langer-Safer, P.R., Levine, M. and Ward, D.C. (1982) Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. USA, 79, 4381–4385.
Leitch, I.J. and Heslop-Harrison, J.S. (1992) Physical mapping of the 18S–5.8S–26S rDNA genes in barley by in situ hybridization. Genome, 35, 1013–1018.
Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. http://arxiv.org/abs/1303.3997. arXiv:1303.3997v2.
Liu, X.Y., Sun, S., Wu, Y., Zhou, Y., Gu, S.W., Yu, H.X., Yi, C.D. et al. (2020) Dual-color oligo-FISH can reveal chromosomal variations and evolution in Oryza species. Plant J. 101, 112–121.
Maluszynska, J. and Heslop-Harrison, J.S. (1991) Localization of tandemly repeated DNA-sequences in Arabidopsis thaliana. Plant J. 1, 159–166.
Marcia, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27, 764–770.
Martins, L.D., Yu, F., Zhao, H.N., Dennison, T., Lauter, N., Wang, H.Y., Deng, Z.H. et al. (2019) Meiotic crossovers characterized by haplotype-specific chromosome painting in maize. Nat. Commun. 10, 4064.
Mascher, M., Gundlach, H., Himmelbach, A., Beer, S., Twardziok, S.O., Wicker, T., Radvshuk, V. et al. (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature, 544, 427–433.
Monat, C., Padmarasu, S., Lux, T., Wicker, T., Gundlach, H., Himmelbach, A., Ens, I. et al. (2019) TRITEX: chromosome-scale sequence assembly of Triticaceae genomes with open-source tools. Genome Biol. 20, 284.
Novak, P., Neumann, P., Pech, J., Steinhauser, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics, 29, 792–793.
Pedersen, C. and Langridge, P. (1997) Identification of the entire chromosome complement of bread wheat by two-colour FISH. Genome, 40, 589–593.
Pedrosa, A., Sandal, N., Stougaard, J., Schweizer, D. and Bachmair, A. (2002) Chromosomal map of the model legume Lotus japonicus. Genetics, 161, 1661–1672.
Price, A.L., Jones, N.C. and Pezvner, P.A. (2005) De novo identification of repeat families in large genomes. Bioinformatics, 21, i351–i358.
Ross, K.J., Fransz, P. and Jones, G.H. (1996) A light microscopic atlas of meiosis in Arabidopsis thaliana. Chromosome Res. 4, 507–516.
Salzberg, S.L. and Yorke, J.A. (2005) Beware of mis-assembled genomes. Bioinformatics, 21, 4320–4321.
Sato, S., Tabata, S., Hirakawa, H. et al. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 485, 635–641.
Schmidt, T., Schwarzwarzer, T. and Heslop-Harrison, J.S. (1994) Physical mapping of rRNA genes by fluorescent in-situ hybridization and structural analysis of SS rRNA genes and intergenic spacer sequences in sugar beet (Beta vulgaris). Theor. Appl. Genet. 88, 629–636.
Schnable, P.S., Wray, D., Fulton, R.S., Stein, J.C., Wei, F.S., Pustelnak, S., Liang, C.Z. et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science, 326, 1112–1115.
Schubert, I. and Wobus, U. (1985) In situ hybridization confirms jumping nucleolus organizing regions in Allium. Chromosoma, 92, 143–148.
Schwarzwarzer, T., Leitch, A.R., Bennett, M.D. and Heslop-Harrison, J.S. (1989) In situ localization of parental genomes in a wide hybrid. Ann. Bot. 64, 315–324.
Shen, Y.T., Du, H.L., Liu, Y.C., Li, N.B., Wang, Z., Liang, C.Z. and Tian, Z.X. (2019) Update soybean Zhonghuang 13 genome to a golden reference. Sci. China Life Sci. 62, 1257–1260.
Spooner, D.M., Rodriguez, F., Polgar, Z., Ballard, L.E. and Jansky, S.H. (2008) Genomic origins of potato polyploids: GBSSI gene sequencing data. Crop Sci. 48, S27–S36.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y. et al. (2015) An integrated map of structural variation in 2,504 human genomes. Nature, 526, 75–81.

Sun, H.Q., Ding, J., Piednoel, M. and Schneeberger, K. (2018) findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics, 34, 550–557.

Suzuki, G., Ogaki, Y., Hokimoto, N., Xiao, L., Kikuchi-Taura, A., Harada, C., Okayama, R. et al. (2012) Random BAC FISH of monocot plants reveals differential distribution of repetitive DNA elements in small and large chromosome species. Plant Cell Rep. 31, 621–628.

Tang, Z.X., Yang, Z.I. and Fu, S.L. (2014) Oligonucleotides replacing the roles of repetitive sequences pAs1, pSc119.2, pTa-535, pTa71, CCS1, and pAWRC.1 for FISH analysis. J. Appl. Genet. 55, 313–318.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. (2012) Primer3 - new capabilities and interfaces. Nucleic Acids Res. 40, e115.

Woo, S.S., Jiang, J.M., Gill, B.S., Paterson, A.H. and Wing, R.A. (1994) Construction and characterization of a bacterial artificial chromosome library of Sorghum bicolor. Nucleic Acids Res. 22, 4922–4931.

Xin, H.Y., Zhang, T., Wu, Y.F., Zhang, W.L., Zhang, P.D., Xi, M.L. and Jiang, J.M. (2020) An extraordinarily stable karyotype of the woody Populus species revealed by chromosome painting. Plant J. 101, 253–264.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P. and Zhang, G. et al. (2011) Genome sequence and analysis of the tuber crop potato. Nature, 475, 189–195.

Zhang, P., Li, W.L., Fellers, J., Friebel, B. and Gill, B.S. (2004) BAC-FISH in wheat identifies chromosome landmarks consisting of different types of transposable elements. Chromosoma, 112, 288–299.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Workflow and graphic interface of Chorus2.
Figure S2 Flow chart of the ChorusNoRef pipeline.
Figure S3 The k-mer spectrum of Illumina shotgun sequence library SRR2960981.

Table S1 Summary of repeats-related oligos designed by Chorus.
Table S2 Oligos designed by Chorus2 and OligoMiner.
Table S3 Chorus2- and OligoMiner-designed oligos mapped to the top 200 repetitive clusters identified by RepeatExplorer2.
Table S4 Time and memory consumption by Chorus2 and OligoMiner.
Table S5 Information of designed oligo-FISH probes for nine species.

Supplementary Material A guide pipeline for designing Oligo-FISH probes in potato.