Whole-Genome Validation of High-Information-Content Fingerprinting[1]

William M. Nelson[2], Arvind K. Bharti[2], Ed Butler, Fusheng Wei, Galina Fuks, HyeRan Kim, Rod A. Wing, Joachim Messing, and Carol Soderlund*

Arizona Genomics Computational Laboratory, BIO5 Institute (W.M.N., C.S.) and Arizona Genomics Institute, Department of Plant Sciences (E.B., F.W., H.K., R.A.W.), University of Arizona, Tucson, Arizona 85721 (W.M.N., C.S.); and The Plant Genome Initiative at Rutgers, Waksman Institute, Rutgers, State University of New Jersey, Piscataway, New Jersey 08854 (A.K.B., G.F., J.M.)

Fluorescent-based high-information-content fingerprinting (HICF) techniques have recently been developed for physical mapping. These techniques make use of automated capillary DNA sequencing instruments to enable both high-resolution and high-throughput fingerprinting. In this article, we report the construction of a whole-genome HICF FPC map for maize (*Zea mays* subsp. *mays* cv B73), using a variant of HICF in which a type IIS restriction enzyme is used to generate the fluorescently labeled fragments. The HICF maize map was constructed from the same three maize bacterial artificial chromosome libraries as previously used for the whole-genome agarose FPC map, providing a unique opportunity for direct comparison of the agarose and HICF methods; as a result, it was found that HICF has substantially greater sensitivity in forming contigs. An improved assembly procedure is also described that uses automatic end-merging of contigs to reduce the effects of contamination and repetitive bands. Several new features in FPC v7.2 are presented, including shared-memory multiprocessing, which allows dramatically faster assemblies, and automatic end-merging, which permits more accurate assemblies. It is further shown that sequenced clones may be digested in silico and located accurately on the HICF assembly, despite size deviations that prevent the precise prediction of experimental fingerprints. Finally, repetitive bands are isolated, and their effect on the assembly is studied.

Restriction fingerprint clone maps are a crucial part of large genome research projects. The maps enhance the value of clone arrays by assembling them into contigs, which can be anchored to chromosomal locations using genetic markers. Clones of particular interest can be isolated, and for clone-by-clone sequencing, a minimal tiling path of clones can be selected (Engler et al., 2003).

Whole-genome restriction fragment maps were first published in 1986. Coulson et al. (1986) mapped *Caenorhabditis elegans* cosmid clones using a two-enzyme method, which produced small fragments that could be run on an acrylamide gel. The result was many high-resolution (i.e. precision) fragments covering a subset of the clone. Also, Olson et al. (1986) mapped *Saccharomyces cerevisiae* using a complete digest method, which produced large fragments that were run on agarose gels. The result was low-resolution fragments covering most of the clone. This agarose method was improved by Marra et al. (1997), and the FingerPrinted Contigs (FPC) software package was developed by Soderlund et al. (1997); this approach was used extensively, including for construction of the human (International Human Genome Sequencing Consortium, 2001), *Arabidopsis thaliana* (Marra et al., 1999), and rice (*Oryza sativa*; Chen et al., 2002) physical maps.

The agarose method has been very successful, but the throughput is limited by the need for human bandcalling of the gel images. Even with assistance from Image software (Sulston et al., 1989), this is a time-consuming process, which also requires considerable skill; however, an automated alternative has recently been described by Fuhrmann et al. (2003). Another drawback of agarose fingerprinting is that the fingerprints contain relatively little information; this is because the fragments are large and few in number and their sizing is not precise.

To address these problems, a set of new methods has been developed, known collectively as high-information-content fingerprinting (HICF; see Ding et al., 1999, 2001; Luo et al., 2003). These methods emerged from the acrylamide-based methods of Coulson et al. (1986) and Brenner and Livak (1989), with further adaptations to take advantage of automated sequencing technology, leading to substantial increases in both the throughput and sensitivity of fingerprinting.

All HICF methods share certain features. First, the characteristics of DNA sequencing machines require that HICF fragments be quite small, currently 500 bp or less. Second, dye labeling is employed to make the fragments detectable by the sequencers. The labels, which are dideoxy terminators (ddNTP) with a specific

[1] This work was supported by the National Science Foundation Plant Genome (grant no. 0211851). Additional support for multithreading of FPC was provided by the National Science Foundation (grant no. 0213764).

[2] These authors contributed equally to the paper.

* Corresponding author; e-mail cari@agcol.arizona.edu; fax 520–626–4824.

www.plantphysiol.org/cgi/doi/10.1104/pp.105.061978.

fluorescent dye for each base, are end-labeled to recessed 3′-OH ends generated by either a type II or type IIS restriction enzyme. Clones are digested with one or more of these enzymes, along with a 4-cutter to reduce the fragment sizes to the required size range (6-cutters may also serve this purpose; Ding et al., 1999). The fragments are end-labeled, and only fragments with at least one labeled end are detected, so that the fragments in an HICF fingerprint do not cover the entire clone.

HICF methods can be divided into two classes, depending on whether they use type II or IIS enzymes for fragmentation. Type II enzymes cut within their recognition sequence, resulting in a predetermined overhang sequence and a fixed label. Therefore, each dye label requires digestion by a different type II enzyme, and several digestion reactions may be needed depending on enzyme compatibility. By contrast, type IIS enzymes cut outside their recognition sequence and leave an undetermined overhang, so that digestion with a single type IIS enzyme produces fragments that can be labeled with any florescent ddNTP. For further review, see Meyers et al. (2004) and Nelson and Soderlund (2005).

The type II method has recently been elaborated by Luo et al. (2003), who adapted the ABI SNaPshot SNP detection kit for use with HICF and tested it on rice clones using an ABI 3100 sequencer. The type IIS method has been elaborated by DuPont (M. Morgante, personal communication) and by Ding et al. (2001). DuPont used the enzymes *Ear*I and *Taq*I to construct a proprietary HICF map of maize inbred line Mo17, using the gel-based ABI 377, while Ding et al. used *Hga*I and *Rsa*I and tested the method with human clones, also using the ABI 377. The reported success of the Mo17 map provided the impetus for this work, as it was decided to construct a publicly available HICF phase I (unedited) map of maize inbred line B73 using the DuPont technique to establish a framework for whole-genome sequencing and to aid in the finishing of the agarose-based phase II (edited) map (Coe et al., 2002).

Accordingly, the three bacterial artificial chromosome (BAC) libraries that had been used to construct the agarose map were refingerprinted with HICF and assembled in FPC. To our knowledge, the resulting assembly is the first whole-genome HICF map to be described in the literature and provides a far larger and more realistic demonstration of the method than previous studies. In addition, since the same BAC libraries were used, an unambiguous comparison between the HICF and agarose methods can be made, with the conclusion that HICF is considerably superior at forming contigs. The HICF assembly did also prove very helpful for the manual finishing of the agarose-based map; this will be described further below and in detail elsewhere (F. Wei, personal communication).

The building of an FPC map occurs in three stages. (1) The complete build forms the initial contigs. (2) One or more iterations of the DQer and end-merger functions are run, where the DQer removes many of the false-positive joins, and the end-merger removes many of the false-negative joins. (3) The remaining false positives and false negatives are found and fixed by manual editing. The HICF map has been taken through the first two stages on a 22× coverage of the maize genome. Previous studies only performed the complete build on a small set of clones; Ding et al. (1999) assembled 98 BACs from chromosome 22, Ding et al. (2001) assembled 555 BAC clones from human chromosome 16, and Luo et al. (2003) assembled a 26 clone contig and a 58 clone contig from two different regions of rice chromosome 10. In all three cases, there was sequence to confirm the assembly. Our analysis of 464,544 BAC clones on a highly repetitive genome brings the robustness of this technique to a whole new level.

To build a map using a new fingerprint method with FPC, there are three salient issues: (1) the fingerprints must be of high quality, (2) the FPC parameters must be tuned for the data, and (3) the method for creating compatible in silico digest fingerprints must be established. In this article, these issues are addressed for the maize HICF assembly, and the general characteristics of the fingerprint data are compared with results of Ding et al. (2001) and Luo et al. (2003). Additionally, the agarose and HICF methods are compared using a set of clones fingerprinted by both methods, and an estimate of the false-negative and false-positive merges is presented. Lastly, the effect of repetitive sequence on the maize genome assembly is studied.

RESULTS

Fingerprinting Maize Inbred B73 by HICF

To take advantage of the increased throughput possible with capillary DNA sequencers, such as the ABI 3700, the type IIS HICF method based on *Ear*I and *Taq*I was adapted for capillary separation. This process utilizes the type IIS 6-cutter *Ear*I, which creates ends that can be labeled, and the type II 4-cutter *Taq*I, which serves to reduce the size of the *Ear*I fragments. *Taq*I leaves a two-base overhang where the first base is G, which if labeled would not be informative since too many fragments would be labeled. Therefore only C, T, and A overhangs are end-labeled in this method, using the fluorescent dyes ddGTP (blue), ddATP (green), and ddTTP (yellow), respectively (Fig. 1). Most labeled fragments are derived from *Ear*I digestion on one end and *Taq*I on the other, but *Ear*I-*Ear*I fragments also occur at a low frequency. A red dye (ROX) is also employed for the internal sizing standard fragments, which are run in each capillary to provide a size ladder.

Using this technique, we fingerprinted 464,544 clones from three different BAC libraries of maize inbred B73 (Table I). The data set was generated in <1

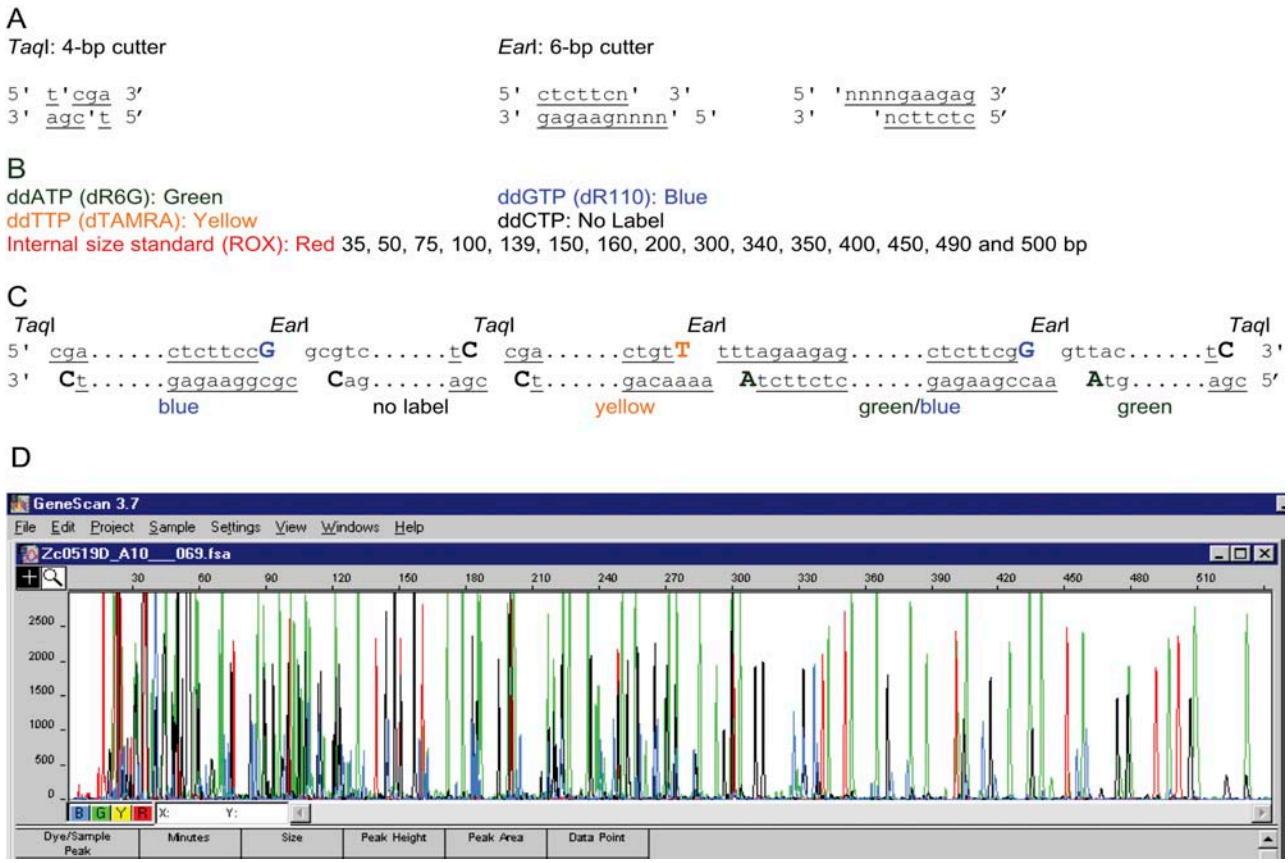

Figure 1. Three-color fluorescent fingerprinting based on the type IIS restriction enzyme *EarI* and 4-cutter *TaqI*. Shown are the (A) restriction enzyme cutting patterns, (B) the assignments of dye label to ddNTP and the ROX (−250 bp) standard that was used, (C) an example of enzyme cutting and labeling, and (D) a sample HICF trace (ZMMBBc0519B22) displayed in Genescan.

year, testifying to the high throughput made possible by HICF. In addition, the same DNA preparations were used to sequence 472,682 BAC ends for these clones, providing a thorough and random sampling of the entire maize genome and enhancing both the agarose and the HICF maize maps by providing anchor points for locating other sequences on them (Messing et al., 2004).

Fingerprint Processing and Determination of FPC Parameters

HICF fragment detection results in an electropherogram, or graph, of the intensity of the various color labels detected by the instrument (see Fig. 1). Peaks in the electropherogram correspond to fragments, but in general there are many spurious peaks, which must be separated from the correct peaks by a process known as scoring (the nature of these extra peaks will be discussed further below). Scoring proceeds by choosing a threshold for each color and rejecting all fragments whose peak height falls below the threshold. We used a scoring strategy derived by DuPont (M. Morgante, personal communication) in which, after performing a complete build on each; considerable experimentation with parameters, the threshold was set equal to 25% of the height of the sixth highest peak in the given color. The resulting fingerprint then consists of approximately 35 fragments in each color, and each fragment is counted only once, i.e. there is no attempt to recognize especially high (or wide) peaks in order to infer a doubled fragment.

Several quality checks were applied to the fingerprints, namely: (1) the well should not be empty; (2) no more than one of the expected vector bands should be missing after scoring (requiring all vector bands is too stringent due to imperfect reproducibilty, as described below); (3) no more than one spurious peak in the standards channel higher than a valid standards peak (spurious standards peaks arise from bleed-through of strong peaks in other colors and suggest that other channels may also contain excessive spurious peaks); (4) no more than 1000 peaks before scoring; and (5) at least 25 and no more than 250 bands after scoring. The percentages of clones removed by each check were (1) 3.1%, (2) 5.2%, (3) 2.3%, (4) 0.1%, and (5) 1.7%. A total of 403,638 fingerprints passed these tests, for an overall success rate of 87%. The average number of scored bands per fingerprint was 107, and since the average clone size is 150 kb (Table I), there is approximately

Table I. *Characteristics of the three BAC libraries of Zea mays subsp. mays cv B73 that were used in both the agarose-based map and the HICF map*

Library	Made by	No. of BACs	384-Well Plates	BACs with no Inserts	Average Insert Size	Genome Equivalent[a]	Genomic DNA Partially Digested with	Cloned into	Vector Size *bp*	References
ZMMBBb (NSF B73)	J.P. Tomkins, CUGI	247,680	1 to 645	991 (0.4%)	136 kb	14.2×	*Hin*dIII	*Hin*dIII site of pBeloBAC11	7,507	Tomkins et al. (2002)
ZMMBBc (CHORI 201)	P.J. de Jong, CHORI	110,592	1 to 288	0 (0%)	163 kb	7.6×	*Eco*RI + *Eco*RI Methylase	*Eco*RI site of pTARBAC2.1	13,397	Cone et al. (2002); Yim et al. (2002)
ZMMBBc (CHORI 201)	P.J. de Jong, CHORI	110,592	289 to 576	2,375 (2.1%)	167 kb	7.8×	*Mbo*I	*Bam*HI site of pTARBAC1.3	13,462	Cone et al. (2002); Yim et al. (2002)
Total		468,864	1,221	3,366 (0.7%)	150 kb	29.6×				

[a]Genome coverage based on 2,365 Mb (Bennett and Laurie, 1995).

one band for every 1.4 kb of sequence. Note that this is not the same as the average size of a detected fragment, which is approximately 200 bp.

FPC only takes one set of integer values per clone as input; hence, the size/color pairs generated by HICF must be converted for FPC, using the technique of Ding et al. (2001). Each band was multiplied by 20 and the fractional part discarded. An offset was then added to each band as follows: 20,000 to yellow bands, 10,000 to green, and 0 to blue. Since we used only fragments in the range 75 to 500 bp, the result of this conversion was bands occupying the ranges 1,500 to 10,000, 11,500 to 20,000, and 21,500 to 30,000. Note that no band from one range can match any other range, so bands originating from different labels are kept separate.

A description of the FPC parameters is provided in the FPC assembly section of ''Materials and Methods.'' The FPC gel length parameter must be set to the total number of possible band values. Adding the ranges above, this comes to 25,500 for this HICF methodology. The FPC tolerance parameter specifies how close two bands must be to be considered matching. From the vector band measurements described below, this value was estimated at 0.35 bp, and since all bands have been multiplied by 20, the tolerance must also be multiplied, resulting in tolerance = 7 for FPC.

Reproducibility

Reproducibility of fragment sizes is measured by identifying identical fragments in many different fingerprints and computing the standard deviation of their sizes in the different fingerprints. For this purpose, vector fragments are ideal because they are contained in every fingerprint from a given library. Table II summarizes the observed size data for the vector fragments in the maize project; all standard deviations lie between 0.07 and 0.12 bp, with an overall average of 0.1 bp. The sizing of fragments on ABI 3700 sequencers is therefore highly reproducible, as observed previously for the ABI 3100 sequencer (Luo et al., 2003).

Table II also illustrates unusual characteristics of fragment sizing by ABI sequencers. First, the sizes are not integers even though the actual fragments have integral sizes. Furthermore, the reported sizes are quite inaccurate, being larger than predicted by between 0.9 and 4.1 bp for the fragments in Table II (for details on how the in silico predictions are made, see ''Materials and Methods''). This inaccuracy has also been observed for ABI 3100 (Luo et al., 2003) and ABI 377 (Ding et al., 2001). These anomalies will be discussed further below, but it is important to stress that the inaccuracy of the fragment sizes does not affect their high reproducibility and therefore does not hinder contig assembly.

Equally important is reproducibility of the overall band pattern (i.e. false positives and false negatives). This was measured by repeatedly fingerprinting the same plate of 96 maize clones, and it was found that on average 75% of the bands were shared between replicate fingerprints of the same clone; Table III shows this concretely in an alignment between six randomly chosen replicates for one of the clones. This relatively low reproducibility may be an artifact of scoring, but numerous variations were tried without substantial improvement. One difficulty that is illustrated in Table III is that false peaks can be higher than true peaks, making it impossible to eliminate them based on peak height alone.

The imperfect reproducibility of band patterns has a detrimental impact on the assembly and leads to the presence of many Q (questionable) clones (see ''Materials and Methods'' for a more in-depth description of Q clones). In agarose mapping projects, Q clones have been used to indicate possible chimeric contigs, and the DQer function of FPC is designed to break up chimeric contigs based on this evidence; however, the occurrence of many extra Q clones in HICF renders this test much less informative.

Table II. *Vector fragments from 464,544 maize HICF fingerprints*

The top row shows the in silico prediction, and each entry shows the average sizes observed on each of the five machines, with the SD in parentheses. Only peaks with height at least 100 were used. Machines 1 and 2 processed only CUGI clones, while 3, 4, and 5 processed both CUGI and CHORI clones. The yellow 131-bp fragment was present only in the ZMMBBb library, while the blue 272-bp fragment was present only in the ZMMBBc libraries. Sizing discrepancies of up to 4.1 bp are seen between the predicted and observed fragment sizes.

Machine	Green 159 bp	Blue 261 bp	Green 197 bp	Green 352 bp	Blue 252 bp	Blue 272 bp	Yellow 131 bp
ABI 3700–1	162.2 (0.1)	263.9 (0.09)	200.9 (0.06)	353.0 (0.09)	256.1 (0.09)		132.6 (0.09)
ABI 3700–2	162.2 (0.1)	263.9 (0.11)	200.9 (0.07)	352.9 (0.1)	256.1 (0.11)		132.6 (0.1)
ABI 3700–3	162.2 (0.12)	264.0 (0.07)	200.9 (0.06)	352.9 (0.08)	256.2 (0.08)	273.9 (0.11)	132.6 (0.09)
ABI 3700–4	162.2 (0.11)	263.9 (0.09)	200.9 (0.06)	352.9 (0.09)	256.1 (0.09)	273.8 (0.11)	132.6 (0.09)
ABI 3700–5	162.2 (0.11)	263.9 (0.08)	200.9 (0.06)	352.9 (0.09)	256.2 (0.09)	273.8 (0.11)	132.6 (0.09)
Overall	162.2 (0.11)	263.9 (0.09)	200.9 (0.07)	352.9 (0.10)	256.1 (0.1)	273.8 (0.11)	132.7 (0.09)

HICF Fingerprint Assembly

The build cutoff was initially chosen to be 1e-45, but it was found that assemblies on this data set yielded one large contig containing almost all the clones. Since 1e-45 is a reasonably stringent cutoff, requiring approximately 58% overlap between fingerprints, some type of contamination was suspected. Well-to-well contamination within plates was judged to be the most likely and was clearly seen in a few cases, so two screens were applied to eliminate it. First, all clones overlapping another clone on the same plate at cutoff 1e-45 were removed. Second, all clones having >175 bands were removed because

Table III. *Alignment of six randomly chosen replicate fingerprints of clone ZMMBBb0032A15, illustrating the difficulties of scoring HICF data*

Only bands from 11,500 to 20,000 (i.e. deriving from green-labeled fragments) are shown. The height of the peak is shown in parentheses. G indicates a gap where there was no matching band. Traces were scored as described in the text. Large differences can be seen between the heights of peaks both in the same traces and in different traces, and some apparently spurious peaks are higher than correct peaks, e.g. the boldface pair in the last replicate.

Six Randomly Chosen Replicate Fingerprints of Clone ZMMBBb0032A15					
11628(5278)	11627(984)	11627(669)	11628(963)	11628(622)	11626(3924)
11771(3778)	11770(770)	11770(642)	11769(733)	11770(461)	11769(3195)
11939(3915)	11938(735)	11938(581)	11939(724)	11937(377)	11937(3179)
12025(7925)	12024(1943)	12023(1349)	12024(1832)	12023(979)	12022(6569)
12613(4428)	12611(1005)	12611(648)	12612(991)	12612(513)	12612(3581)
12688(1155)	12687(418)	12689(166)	12687(407)	G	G
12795(3693)	12793(672)	12793(511)	12794(779)	12794(381)	12794(2914)
12996(3682)	12994(584)	12997(599)	12993(502)	G	G
13246(2176)	13245(1082)	13244(578)	13245(1013)	13246(452)	13244(2864)
13903(1922)	13903(566)	13902(215)	13904(541)	13901(156)	**13902(1362)**
14019(4067)	14019(1280)	14018(585)	14021(1227)	14018(532)	14019(3266)
14269(3053)	14269(1770)	14269(483)	14271(1697)	14267(483)	14269(2755)
14310(3319)	14310(1222)	14309(520)	14311(1156)	14308(485)	14310(2956)
G	G	G	G	G	**15166(1963)**
15386(3228)	15386(1226)	15388(575)	15387(1207)	15385(534)	15387(2873)
15633(3094)	15634(1434)	15632(430)	15633(1322)	15630(458)	15635(2817)
15661(2121)	15661(1090)	15660(185)	15662(1014)	15658(211)	15662(1523)
G	G	G	G	15869(197)	G
15922(1992)	15921(1203)	15922(169)	15922(1154)	15917(186)	15923(1394)
16088(2817)	16090(912)	16090(330)	16091(875)	16088(388)	16091(2517)
G	16311(472)	G	16315(438)	G	G
16659(1629)	16656(949)	16655(190)	16660(906)	16655(245)	16656(1455)
16727(2820)	16725(1721)	16727(441)	16729(1657)	16727(546)	16726(2808)
16777(2221)	16774(1244)	16775(338)	16779(1186)	16775(401)	16775(2394)
16912(1654)	16911(695)	G	16912(702)	G	G
17060(2791)	17059(1449)	17060(334)	17059(1379)	17059(403)	17061(2478)
17598(2004)	17597(628)	G	17593(625)	G	G
17675(2082)	17674(1166)	17673(265)	17674(1085)	17672(388)	17675(2017)
18388(1513)	18387(763)	G	18386(732)	G	G
18773(2588)	18770(852)	18771(222)	18768(902)	18768(300)	18767(1439)
G	G	18796(155)	G	18799(234)	G
19395(2651)	19396(1494)	19399(392)	19396(1363)	19397(657)	19393(2653)
19513(1038)	19513(435)	G	19514(400)	G	G

contaminated clones generally will have a large number of bands. Altogether, 53,115 clones were removed by these screens, the vast majority of which were undoubtedly perfectly valid; however, it is worth discarding some data to avoid chimeric contigs, and the removed clones can later be placed back onto the map after contig merging is completed.

These screens reduced the false-positive merges greatly, but building the full data set at 1e-45 still resulted in an unacceptably large chimeric contig. The most likely explanation is that some contamination still remained, but repetitive bands may also play a role (see below). To minimize both of these problems, the initial build was performed at a very stringent cutoff and then contigs were end-merged at successively higher cutoffs (i.e. lower stringencies). In order to prevent a single contaminated clone from causing a merge, each end-merge was confirmed by two completely separate pairs of overlapping clones. To enable this process, new options were added to the Ends→ Ends function in FPC v7.2 (see ''Materials and Methods'').

The initial build was performed at 1e-70 and required only 22 h using two additional enhancements to FPC v7.2, shared-memory multiprocessing and precomputation (see ''Materials and Methods''). The initial build had 11,245 contigs, and these were then end-merged at 12 successively lower cutoffs, terminating at 1e-21. The DQer was used after each merge to break up all contigs containing >15% Q clones. All possible singleton clones were merged with the existing contigs before the 1e-43 end-merge. The final build contains 350,253 clones, which based on an average insert size of 150 kb and genome size of 2.365 Gb (Bennett and Laurie, 1995) corresponds to a genome coverage of 22×. The assembly has 1,500 contigs covering 1.9 Gb, or 83% of the genome (using the previous estimate of 1.4 kb per band). The HICF assembly has not been manually edited, but as described below, was used to assist in manual editing of the agarose map.

We note that singletons were merged into the assembly only once because this was found to increase the number of Q clones in the map substantially. It appears that many clones that are singletons at the original high-stringency cutoff are so because their fingerprints are of lower quality and not because they come from regions of low coverage.

Use of HICF for Finishing of Agarose Map

Manual editing of the agarose maize map was well under way before the HICF assembly was completed, and it was therefore decided to use the HICF assembly to further improve the agarose map, instead of manually editing the HICF assembly. Several other information sources were also available for this editing, and each edit (i.e. contig split or merge) was confirmed by at least two of the following types of evidence: (1) inspection of the agarose fingerprints, (2) markers, (3) synteny with rice, or (4) the HICF assembly. Hence, the agarose phase II map (F. Wei, personal communication) contains the most accurate contigs and contig order possible with current information, and in particular it contains a great deal of information beyond that provided by the HICF or agarose initial assemblies alone, making it reasonable to use it as a standard against which to test the quality of HICF or agarose assemblies.

Comparison of Agarose and HICF Methods

Having both agarose and HICF fingerprints of the same clone libraries provides a unique opportunity to compare the effectiveness of these two mapping techniques with all variables related to genome or library characteristics removed. Though the same BAC libraries were used for both agarose and HICF maps, the specific clones having successful fingerprints differed, so a subset of 199,446 clones was chosen that had successful fingerprints in both methods. For each method, a complete build was then performed with subsequent iterations of end-merging and DQing. The agarose analysis allowed five Qs per contig, whereas the HICF allowed 15% per contig. After the build, and then after each end-merge/DQer step, the contigs were compared to the manually edited agarose map. Any contig formed from two contigs in the manually edited map was counted as a false merge; note that this metric will only be incorrect if the two contigs should really be merged, but due to the extensive editing of the agarose map, this will be rare. The agarose build was started at 1e-12, while the HICF was started at 1e-70, and then both were end-merged several times and evaluated on the basis of contig number and false merges. As shown in Table IV, we stopped the end-merging of the agarose contigs after two iterations, as there were already 128 false joins, whereas the HICF continued successful end-joining until 1e-15 for which it had 93 false merges. At these stopping points, the agarose map had 128 false joins and 6,488 contigs compared to HICF with 93 false joins and 2,303 contigs. Note that we experimented with different parameters for both the agarose and HICF analysis and found these parameter sets to result in the least false joins and least contigs for this genome.

Although the cutoffs used for the initial builds in the two methods are very different, they in fact correspond to nearly the same stringency. The agarose cutoff of 1e-12 requires approximately 70% overlap of fingerprints, while the HICF cutoff 1e-70 requires 74%. The difference between the two methods is seen in the final cutoffs used, because the HICF cutoff 1e-15 requires only 31% overlap, while the agarose cutoff 1e-10 still requires 63% overlap.

Simulated HICF

An unsatisfactory aspect of the HICF assembly is its large number of Q clones. The HICF phase I map

Table IV. *Side-by-side comparison of agarose and HICF assemblies using an identical set of 199,446 maize clones, showing that HICF generates many fewer contigs as compared to agarose and does not increase the number of false contigs*

Both assemblies were done starting with a complete build, followed by iterations of the DQer and end-merger routines.

Build	Contigs	Qs >*N*[a]	False Joins[b]	Numbers of Contigs of Different Sizes					
				≥100	99:50	50:26	24:10	<10	Singletons[c]
HICF[d]									
Initial 1e-70	11,627	35	1	12	388	1,689	3,798	5,740	31,160
Merge 1e-50	8,094	118	3	80	668	1,657	2,655	3,034	
Merge 1e-40	5,722	269	6	208	847	1,364	1,661	1,641	
Merge 1e-30	3,931	476	13	408	816	911	939	858	
Merge 1e-20	2,659	609	47	599	611	518	462	469	
Merge 1e-15	2,393	623	93	615	552	472	372	382	
Agarose									
Initial 1e-12	7,420	0	84	159	924	1,682	2,166	2,593	10,588
Merge 1e-11	7,153	0	88	176	932	1,648	2,041	2,356	
Merge 1e-10	6,488	0	128	232	971	1,503	1,729	2,033	

[a]The agarose contigs were not allowed to have more than five Q clones; hence, there is never more than five Qs in the third column. As discussed in the text, agarose has less error in the data, so this stringency is acceptable. In constrast, the HICF analysis had to allow 15% Q clones due to the error in its fingerprints. [b]False contigs were determined by comparison with the manually edited agarose maize map, with contigs containing five or more clones from two different manually edited contigs being scored as false. [c]Singletons were not merged after the initial build, so their number remains constant. [d]Additional HICF end-merges at 1e-65, 1e-60, 1e-55, 1e-45, 1e-35, and 1e-25 are not shown.

contains 11% Q clones, as compared to 0.5% in the agarose phase I map. To verify that this resulted from fingerprint error rather than a problem with FPC processing, several simulations were performed using 16 Mb of maize genomic sequence created by concatenating 93 sequenced maize clones, with gaps removed. A simulated 22× BAC library was generated from this sequence, and the resulting 2,495 BACs were HICF-fingerprinted in silico and assembled in FPC. At 1e-50 and a tolerance of 7, it built into two contigs with no Q clones, demonstrating that realistic, error-free HICF data assemble properly in FPC. The effect of fingerprint error was also tested by replacing 12.5% of the correct bands with random bands (this value was chosen to match the 75% reproducibility observed for the actual maize data). With this simulated error, the clones assembled at 1e-50 and tolerance of 7 into 26 contigs having 7.8% Q clones (after DQing as in the real build). This large difference in both contig number and Q clones shows the advantages to be gained by improving the reproducibility of HICF fingerprints.

Placing Sequence onto an HICF Physical Map

If a genome has an agarose FPC map, then sequenced clones can be digested in silico and placed on the map. If a sequenced clone has been fingerprinted, then the placement of the in silico clone should be near the fingerprinted clone, which verifies that the correct clone was sequenced and that the sequence is accurate. Or, if the sequenced clones are not in the FPC map, the in silico fingerprint provides a way to locate those clones on the map, which otherwise would not be possible without fingerprinting them. Any sequences that are located on the map can then be used to anchor additional sequences, such as sequenced markers, cDNAs, or genomic survey sequence. This procedure has been used extensively in the maize and rice agarose maps (Soderlund et al., 2002; Engler et al., 2003; Pampanwar et al., 2005).

With HICF, placement of in silico fingerprints on the FPC map becomes problematic because of the errors in fragment sizing (see Table II). There is no obvious pattern to the errors, and, as will be discussed below, they appear to depend on the sequence of the fragments. It therefore is currently impossible to carry out HICF digestions in silico and reproduce the exact sizes seen in the experimental fingerprints; however, it turns out that knowledge of the distribution of the size errors permits an approximation that is sufficient for placing sequence onto the FPC map.

To obtain the distribution of sizing errors, a reasonably large number of fragments are needed for which both the actual and the experimentally measured size are known. These were obtained by in silico digestion of 22 sequenced maize clones that had already been fingerprinted experimentally. The clones chosen were in phase II sequencing, had four or fewer pieces, and had successful experimental fingerprints in both HICF and agarose methods; in addition, it was verified that the agarose in silico fingerprint could be placed correctly on the agarose map. The clones were digested in silico for HICF and aligned to the experimental HICF fingerprints using a dynamic-programming algorithm (see ''Materials and Methods''), and the sizing error was then determined for each pair of matched fragments. The resulting distribution is shown in Figure 2. The peak is near +3 bp, and the distribution is not extremely wide, having a SD of 1.3 bp. This suggests that if each in silico band is offset by ap-

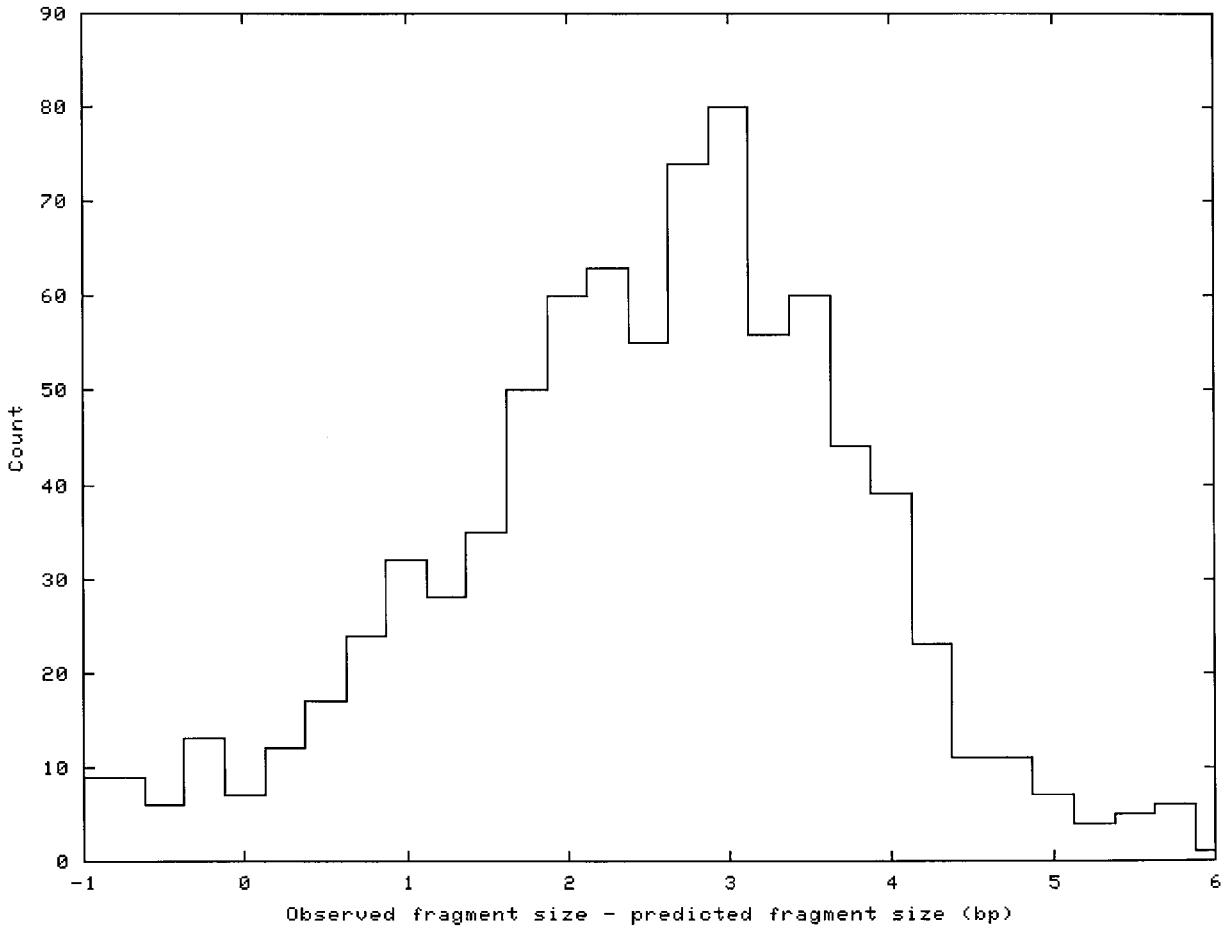

Figure 2. Distribution of fragment sizing error from ABI 3700 sequencers, as determined by comparing in silico size predictions with observed fragment sizes for 22 sequenced maize clones (see text). The data are plotted as a histogram with a bin size of 0.25 bp.

proximately 3 bp, the in silico fingerprints can then be placed on the map using an acceptably small tolerance.

To verify that in silico fingerprints can in fact be placed accurately using this technique, the optimal values of offset and tolerance were first determined through experimentation to be 2.85 and 1.35 bp, respectively. The 22 sequenced clones were then digested in silico, the offset was added to each of their bands, and the fingerprints were added to the FPC map using the Keyset→FPC function with tolerance 27 (= 1.35 × 20) and cutoff 1e-08. All 22 clones placed to the correct contigs, and for nine of them, their best match based on the Sulston score was the original fingerprinted clone. In comparison, 13 of them matched best to their original clone in the agarose map. (The best match is not always to the original clone due to either fingerprint or sequence error.) The experiment shows that it is possible to place sequenced clones with reasonable accuracy onto an HICF map, and this removes a significant drawback to the method. This result has been derived for ABI 3700 data, but it seems likely to hold for other capillary instruments.

The fragment alignments derived above can be put to further use in determining whether the sizing of fragments depends on the sequence of the fragments. Sequence dependence was previously inferred by Luo et al. (2003) on the ABI 3100 from unequal migration of identically sized but complementary fragment pairs (such pairs occur in SNaPshot when two identical 6-cutters cut with no intervening 4-cutter). Our data confirm this on the ABI 3700 with a larger number of samples. The data set contains 390 pairs of in silico fragments that have the same size but different sequence, and in 218 of these cases, the two fragments are matched to different experimental bands; that is, they were separated by >0.35 bp, which corresponds to the FPC tolerance of 7. Therefore, it appears that same-sized fragments of differing sequence migrate with different mobility through capillary electrophoresis in approximately 56% of cases. This varying mobility then gives significance to the decimal part of the measured fragment sizes.

Repetitive Bands

The maize genome is known to be highly repetitive (Flavell et al., 1974; Meyers et al., 2001; Messing et al., 2004), containing numerous high-copy families of long terminal repeat retrotransposons that cover approximately 60% of the maize genome. Furthermore, many gene families (35%) are organized in tandemly arranged copies, e.g. the storage protein genes (Song and Messing, 2003). Such a large amount of repetition could present a serious obstacle to the construction of a physical map, but two factors reduce its impact. The first factor is sequence divergence, which causes repetitive elements to differ if their amplification occurred sufficiently far in the past. Since the maize amplifications have been spread over the past 5 million years (Swigonová et al., 2005), significant divergence is expected. The second factor is the tendency of retroelements to insert inside one another (San Miguel et al., 1996; Swigonová et al., 2005). This can cause some restriction fragments to contain pieces of several retroelements, possibly randomizing them (Luo et al., 2003); however, since the HICF fragments are very small, it is unclear how many will be affected by this.

Given these factors, it is difficult to estimate how many repetitive bands to expect in the fingerprints, but a search of all the band data reveals 113 bands that are present in at least 10% of the clones. Of these, only 10 are found in >30% of the clones, with the most common band occurring in 51% of clones. Since there are several retroelement families common enough to be represented in nearly every clone (Meyers et al., 2001; Messing et al., 2004), it is clear that the mitigating factors have drastically reduced the numbers of repetitive bands.

For physical mapping, one would like to estimate the cutoff at which false overlaps due to repeat bands begin to affect the FPC assembly. For this purpose, all clones containing at least 40 of these bands were extracted and studied in isolation, yielding a set of 1,563 clones, with the most repetitive clone containing 65 repeat bands. These clones formed a false contig at cutoff 1e-43 consisting of 11 clones from 11 different contigs of the actual HICF map, and at 1e-35, this chimeric contig grew to include 97 clones from 63 different actual contigs. This indicates that repeats begin to have an effect in this range of cutoffs, although the build process based on end-merging plus DQing evidently reduces these effects since the HICF map was end-merged at 1e-21 without forming such a highly chimeric contig.

One important question is whether it is worthwhile to screen out repetitive bands, as is done with vector bands. The screening must be tested experimentally, and in the case of maize, it is found to be harmful. Following the same steps used for the actual build, the screened clones form 2,193 contigs at 1e-21, which is considerably more than the 1,500 obtained without screening. Furthermore, the screened contigs contain more errors than the unscreened, as estimated by comparison with the phase II agarose map.

DISCUSSION

This study demonstrates that the HICF method may be successfully applied to assemble a deep-coverage BAC library of a large genome containing highly repetitive (but diverged) sequences. The data collection time is greatly reduced as compared to the agarose method, and the number of gaps in the resulting assembly is also much lower. Both of these advantages are likely to increase in the future as automatic sequencers become more powerful and their application to physical mapping receives greater commercial support.

A drawback to this HICF methodology is the large number of Q clones that are generated. It is possible that faulty peak scoring causes this problem, but after extensive testing of different scoring parameters and algorithms, we do not believe this is likely. The most likely culprit is therefore inconsistency in enzyme digestion, which may be especially problematic for type IIS enzymes (Gardner et al., 1982; Ding et al., 1999). The use of a single reaction buffer for the simultaneous digestion by two or more restriction enzymes (depending on the method type) in a single reaction tube may also result in inconsistencies.

We have shown that, despite the errors in fragment sizing, sequenced clones may be digested in silico and placed correctly onto an HICF map; however, with accurate sizing it would be possible to do this with very small sequences (having as few as 10 bands) while still using sufficiently stringent cutoffs to prevent false-positive matches. This would allow, for example, accurate location of repeat elements by in silico digestion of catalogued repeat sequences. It is therefore important to better understand the physical basis of the sizing discrepancies and to apply that knowledge to improve in silico digestion. If the sizing error does depend on the nucleotide sequence of the restriction fragments, as indicated by our results and those of Luo et al. (2003), then it also provides a small benefit in that some fragments that could not be distinguished based on size and label alone are separated by the capillary electrophoresis because of their differing sequences.

One significant difference between the HICF and agarose methods is that doubled peaks are not scored in HICF. The reason for this is that the large variability of peak heights makes it difficult to distinguish the possible doubled peaks using height. Neglect of doubled peaks leads in theory to less accurate clone ordering, but since simulations indicate that approximately 5% of HICF peaks should be doubled, this is not a large source of error; nevertheless, it would be desirable to develop techniques for recognizing and scoring these peaks correctly.

A substantial obstacle to assembling the HICF maize map was the presence of contamination. This is a larger problem in HICF than in agarose because the large number of Q clones arising from noise makes it difficult for the DQer to effectively identify chimeric contigs, which are generally identified by a large number of Q clones. To overcome this problem, several screens were applied and a stepwise build process was developed, with the initial build starting at a cutoff stringent enough to reduce the likelihood of residual contamination causing false merges.

In the future, the power of HICF is likely to be increased considerably by the development of size standards extending beyond 500 bp (DeWoody et al., 2004; http://www.bioventures.com/products/mapmarker/index.php). This will allow fingerprints to contain larger fragments, perhaps as large as 1,000 bp or more. These large fragments will be much less common than smaller ones, and their matches will be highly significant signals of overlap. To take full advantage of this, it may be necessary to modify the Sulston overlap formula to correctly handle nonuniform distributions of bands. One proposal for accomplishing this has been tested (Hatfield, 2002), and additional ideas are under study.

The HICF FPC map is available for download and viewing using the WebAGCoL tools (Pampanwar et al., 2005) from www.genome.arizona.edu and www.agcol.arizona.edu. The FPC software is available at www.agcol.arizona.edu.

MATERIALS AND METHODS

Fingerprinting Reaction

The entire HICF data set comprises 403,638 fingerprints, which were made from three BAC libraries of *Zea mays* subsp. *mays* cv B73. The bacterial cultures were inoculated from glycerol stocks (384-well) using a Q-Bot (Genetix USA) in 1.7 mL of 2× YT medium (96-deep-well plates) and grown for 19 h at 900 rpm in a rotary shaker with a very small orbital radius (3 mm). BAC DNA was isolated using the standard alkaline lysis method (Birnboim and Doly, 1979) using Whatman filters. The DNA was pelleted using isopropanol and finally dissolved in 1 mM Tris·Cl (pH 8). The type IIS HICF method, originally designed for a gel-based sequencer (ABI 377) at DuPont, was adapted for use with a capillary sequencer (ABI 3700) with a one-eighth dilution protocol. About 100 ng of template was added to a master mix containing 4 units *Ear*I, 4 units *Taq*I, 1× NEB buffer 2 (New England Biolabs), 17.9 nM ddATP-dR6G, 17.5 nM ddGTP-dR110, 18.8 nM ddTTP-dTAMRA, 1 unit *Taq* FS (Applied Biosystems), and 0.2 μM ddCTP (Roche Diagnostics). The fingerprinting reaction (15 μL) was incubated at 37°C (1 h) followed by 1 h at 72°C in thermocyclers.

Capillary Electophoresis

The unincorporated dyes (ddATP-dR6G, ddGTP-dR110, and ddTTP-dTAMRA) were removed from the above-labeled fingerprinting reactions by ethanol precipitation. The cleaned up reactions were air-dried and finally dissolved in 100% formamide containing 0.02× Genescan-500 ROX (16 fragments of internal lane standard). The labeled samples were run on ABI 3700 automated capillary sequencers using POP6 polymer running across 50-cm capillary arrays. Prior to running the BAC samples, the ABI 3700 machines were spectrally calibrated by labeling 400 ng pBluescript II SK+ (Stratagene) with the same master mix as described above. This control vector gives five labeled fragments (399-bp blue, 216- and 380-bp green, and 145- and 152-bp yellow bands). The spectral calibration was done for dye set F using the parameter file MtxStd{AnyDyeSet}.par and the default calibration module with a slight modification, i.e. cuvette temperature was changed from 40°C to 46°C. The maximum peak height for G, A, T, and ROX corresponded to spectral bin numbers 3, 6, 9, and 10, respectively. The HICF trace output was extracted using ABI GeneScan v3.7.1 analysis software with the default parameters with some modifications (analysis range from 1,200–10,000, size call range 35–500 bp, and GS 500-250.szs for auto analysis of the standards). The 250-bp internal size standard was not used for the analysis because of its anomalous migration in capillary sequencers. Genescan-500 ROX was originally designed to achieve high precision in molecular sizing of DNA fragments in the 35- to 500-bp range for gel-based instruments. The manufacturer (Applied Biosystems) does not recommend using the 250-bp band for analysis on data generated by capillary sequencers.

FPC Assembly

The FPC assembly algorithm compares the fingerprints of each pair of clones, where the fingerprint for a clone is a list of integers and each integer represents a band (i.e. fragment). For each pair, it counts the number of bands that are the same within a user-supplied tolerance. It then computes the probability that the shared bands are just a coincidence. FPC has two equations the user can select from; the equation that is generally used is the Sulston score (Sulston et al., 1988). It uses a variable referred to as gel length, which is the total number of possible bands. If two clones have a score below the user-supplied cutoff, they are said to overlap (note that the lower the overlap score, the more significant the overlap).

The quality of the FPC map is very dependent on the quality of the data, as is best understood by briefly considering the algorithm (Soderlund et al., 1997). After FPC clusters clone into contigs, it computes a consensus band (CB) map and aligns the clones to the CB map. A clone may have extra bands that will not align or missing bands that cause gaps in its alignment. If the sum of extra and missing bands comes to >50% of the clone's total band count, it is marked as a Q clone.

One source of Q clones is false-positive overlaps. That is, if contig A is being built, and one of its clones falsely overlaps with a clone in contig B, all clones in contig B are incorrectly incorporated into contig A. The clones that do not belong will generally not align well to the CB map and will become Q clones. There is nothing wrong with these clones other than their incorrect location. This type of Q clone is an indicator of false overlaps, and the DQer function of FPC takes advantage of this to automatically break up bad contigs based on their Q content.

Q clones can also be caused by noise within the fingerprints, which leads to extra or missing bands that degrade the alignment to the CB map. These Q clones are highly detrimental for two reasons. First, they make it difficult to use Q clones to detect false joins. It becomes hard to set a DQer threshold that will break up bad contigs but not break up good contigs that happen to contain some lower-quality fingerprints. Second, since the alignment of clones to the CB map is degraded, the precision of clone coordinates is reduced (Soderlund et al., 2000). This becomes a problem when accurate location is needed, e.g. for picking a minimal tiling path.

FPC v7.2 Changes for Increased Speed

Two changes were made to FPC to increase the speed of assembly, a need that became particularly acute with HICF because the large number of bands in HICF fingerprints slows the computations. Rapid assembly allows experimentation with parameters such as cutoff and tolerance, which is important for any FPC mapping project because the optimal parameters are not known in advance.

The first change was to implement shared-memory multiprocessing in the assembly algorithm, allowing for an N-fold speedup on a machine with N processors. The second change was to add a Precompute option, which causes all possible clone overlap scores to be computed in advance and stored in a table. Scores of clone overlaps are thereafter found using a fast table lookup, providing a twofold speedup for 100-band fingerprints. (This option is not needed when there are <60 bands and the Sulston score is used, since an equally fast optimization for this case is already implemented.) The table occupies 7 Mb of RAM for the maize HICF map and grows with the cube of the maximum number of bands per clone.

With these enhancements, a clean build of the maize HICF project requires 22 h on a Dell PowerEdge 6650 having four Intel Xeon 2.8-GHz processors; this assembly would previously have required >1 week.

FPC v7.2 New Features for Assembly

Several features were added or enhanced to meet needs encountered in the maize HICF map. One crucial change was to lower the minimum allowed cutoff from 1e-37 to 1e-99, as required for most HICF assemblies.

There were also some enhancements to the DQer. The DQer executes the following loop three times: It decreases the cutoff exponent by M and reanalyzes all contigs with more than N (or N%) Q clones, and splits them into multiple contigs when necessary. The default values N = 5 and M = 1 are typically suitable for agarose projects, but since HICF tends to have more Q clones that are not from chimeric contigs, the N is now allowed to be a percentage, e.g. N = 10% would allow a contig to have 10% Q clones before being subject to reassembly. Also, when using a low cutoff such as 1e-70, setting M = 1 would only reanalyze at 1e-71, 1e-72, and 1e-73, which will not break up many Q contigs. Therefore, M may now be set by the user, so, e.g. a step size of M = 5 will reanalyze at 1e-75, 1e-80, and 1e-85.

Lastly, automatic end-joining was implemented to enable the stepwise build process described in the text. This option builds on the existing Ends→Ends function, which compares all clones at the ends of contigs and provides a report suggesting pairs of contigs to merge. Previously, all merges had to be done manually, but now if the Auto option is selected, then the merges will be performed automatically. Since this eliminates manual verification of merges, it is only safe to use with cutoffs for which few bad merges are expected. Besides the Auto mode, an additional important parameter called Match was added to Ends→Ends. This parameter controls how many unique clone overlaps are required for a merge; for example, if Match = 2, then two completely different overlapping clone pairs (i.e. involving four clones total) are required for a contig merge. This prevents a single contaminated clone from causing a false merge. In the Auto mode, no merges will be performed for a contig if more than four possible merges were detected for that contig; this also helps to prevent incorrect merges.

All other functions in FPC work exactly the same for both agarose and HICF fingerprints.

FPC v8.0 Improvements

FPC v8.0 was released as this article went to press and contains important improvements to the features described above. (1) The Ends→Ends and CB map algorithm have been parallelized for shared-memory multiprocessors. This completes the parallelization of all of the commonly used, processor-intensive FPC functions. (2) The Ends→Ends function no longer recomputes the CB maps in v8.0, but simply joins the contigs at their ends. Also, it only needs to be run only once, instead of in stages as described for the maize HICF assembly. The new Ends→Ends has been tested on the data set of Table IV and found to perform equivalently to that in v7.2.

In Silico HICF Digestion

Figure 1 shows the recognition sequences and cutting pattern for the *Ear*I and *Taq*I enzymes used for the maize HICF fingerprinting as well as the association of dye color to the labeled overhanging base. In silico digestion is mostly a straightforward application of these rules, but certain details must be handled correctly. First, *Ear*I is not palindromic, and its forward- and reverse-complement recognition sequences must be searched separately. As a result, it cuts twice as often as a palindromic 6-cutter. Second, *Taq*I is sensitive to bacterial DAM methylation and will not cut when its recognition sequence (TCGA) overlaps a DAM site (GATC; see also Luo et al., 2003). Finally, *Ear*I and *Taq*I sites can overlap, in which case reaction kinetics (incubation first at 37°C and then at 72°C) implies that *Ear*I should cut first, disabling the *Taq*I site. All of these effects are confirmed by vector bands; in particular, the green 159-bp fragment in Table II arises because a methylated *Taq*I site is skipped, and the blue 252-bp fragment would be missing if *Taq*I were allowed to cut before *Ear*I.

A final complication arises because of the lack of detection of doubled peaks in HICF. When in silico digestion results in a double band, it must be decided whether to drop one of the copies or keep both. If the two fragments have different sequences, then our results indicate that approximately half of the time there will be two different bands observed in the experimental fingerprint; therefore, it makes sense to retain the double band. However, if the fragments have identical sequences, then one copy should be dropped.

Alignment of in Silico and Experimental Fingerprints

Alignment of in silico and experimental fingerprints was performed using a dynamic-programming algorithm to maximize the number of individual band matches. Matches were permitted in the range -1 bp $\leq$ (experimental $-$ predicted) ≤ 6 bp, a range estimated from vector data. A total of 2,021 matches were found, and they were estimated to be 82% correct based on vector bands for which the correct experimental value is known. In order to minimize the number of false-positive matches in the data set, a subset was selected for which both experimental and predicted bands were at least 3 bp from their nearest neighboring band. This additional margin of safety reduces ambiguous matches and raises the correctness as estimated by vector matches to 97%. This reduced set of 841 matches was then used for the analyses described in "Results."

Sequence data of the 22 BACs can be found in the GenBank/EMBL data libraries under the following accession numbers: AC146795, AC146811, AC145228, AC148112, AC148083, AC146810, AC146975, AC145227, AC148243, AC146950, AC146812, AC148099, AC146763, AC148100, AC148110, AC145261, AC148234, AC148163, AC146813, AC145481, AC148350, and AC148082.

ACKNOWLEDGMENTS

This project is supported by the National Science Foundation Plant Genome Research Program (grant no. 0211851). We would like to thank Michele Morgante, Kevin Fengler, Frank You, Ming-Cheng Luo, Phillip San Miguel, and Jose Luis Goicoechea for valuable discussions. We would also like to thank Danielle Yost, Diana Stum, Steve Young, Steve Kavchok, Gladys Keizer, and Amy B. Nelson for much of the hard work that made the HICF maize map possible. Jamie Hatfield and Gaurav Gupta implemented the multithreaded implementation of FPC, without which the assemblies would have been prohibitively time consuming.

Received February 25, 2005; revised May 20, 2005; accepted May 20, 2005; published September 12, 2005.

LITERATURE CITED

Bennett MD, Laurie DA (1995) Chromosome size in maize and sorghum using EM serial section reconstructed nuclei. Maydica **40:** 199–204

Birnboim HC, Doly J (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res **7:** 1513–1523

Brenner S, Livak KJ (1989) DNA fingerprinting by sampled sequencing. Proc Natl Acad Sci USA **86:** 8902–8906

Chen M, Presting G, Barbazuk B, Goicoechea J, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, et al (2002) An integrated physical and genetic map of the rice genome. Plant Cell **14:** 537–545

Coe E, Cone K, McMullen M, Chen SS, Davis G, Gardiner J, Liscum E, Polacco M, Paterson A, Sanchez-Villeda H, Soderlund C, Wing R (2002) Access to the maize genome: an integrated physical and genetic map. Plant Physiol **128:** 9–12

Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG, et al (2002) Genetic, physical, and informatics resources for maize. On the road to an integrated map. Plant Physiol **130:** 1598–1605

Coulson A, Sulston J, Brenner S, Jonathan K (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans.* Proc Natl Acad Sci USA **83:** 7821–7825

DeWoody JA, Schupp J, Kenefic L, Busch J, Murfitt L, Keim P (2004) Universal method for producing ROX-labeled size standards suitable for automated genotyping. Biotechniques **37:** 348, 350, 352

Ding Y, Johnson MD, Chen WQ, Wong D, Chen YJ, Benson SC, Lam JY, Kim YM, Shizuya H (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. Genomics **74:** 142–154

Ding Y, Johnson MD, Colayco R, Chen YJ, Melnyk J, Schmitt H, Shizuya H (1999) Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. Genomics **56:** 237–246

Engler FW, Hatfield J, Nelson W, Soderlund CA (2003) Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res **13:** 2152–2163

Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet **12:** 257–269

Fuhrmann DR, Krzywinski MI, Chiu R, Saeedi P, Schein JE, Bosdet IE, Chinwalla A, Hillier LW, Waterston RH, McPherson JD, Jones SJ, Marra MA (2003) Software for automated analysis of DNA fingerprinting gels. Genome Res **13:** 940–953

Gardner RC, Howarth AJ, Messing J, Shepherd RJ (1982) Cloning and sequencing of restriction fragments generated by *Eco*RI*. DNA **1:** 109–115

Hatfield J (2002) Analyzing restriction fragments for contig assembly. Master's thesis. Clemson University, Clemson, SC

International Human Genome Sequencing Consortium (2001) A physical map of the human genome. Nature **409:** 934–941

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics **82:** 378–389

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res **7:** 1072–1084

Marra MA, Kucaba TA, Sekhon M, Hillier LW, Martienssen R, Chinwalla A, Crocket J, Fedele J, Grover H, Gund C, et al (1999) A map for sequence analysis of the *Arabidopsis thaliana* genome. Nat Genet **22:** 265–270

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA **101:** 14349–14354

Meyers BC, Scalabrin S, Morgante M (2004) Mapping and sequencing complex genomes: Let's get physical! Nat Rev Genet **5:** 578–588

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res **11:** 1660–1676

Nelson W, Soderlund C (2005) Software for restriction fragment physical maps. *In* K Meksem, G Kahl, eds, The Handbook of Genome Mapping: Genetic and Physical Mapping. Wiley-VCH, Weinheim, Germany, pp 285–306

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R, Frank T (1986) Random-clone strategy for genomic restriction mapping in yeast. Proc Natl Acad Sci USA **83:** 7826–7830

Pampanwar V, Engler F, Hatfield J, Blundy S, Gupta G, Soderlund C (2005) FPC tools for rice, maize, and distribution. Plant Physiol **138:** 116–126

San Miguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768

Soderlund C, Engler F, Hatfield J, Blundy S, Chen M, Yu Y, Wing R (2002) Mapping sequence to rice FPC. *In* P Wang, J Wang, C Wu, eds, Computational Biology and Genome Informatics. World Scientific Publishing, Singapore, pp 59–80

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res **10:** 1772–1787

Soderlund C, Longden I, Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci **13:** 523–535

Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. Proc Natl Acad Sci USA **100:** 9055–9060

Sulston J, Mallett F, Durbin R, Horsnell T (1989) Image analysis of restriction enzyme fingerprint autoradiograms. Comput Appl Biosci **5:** 101–106

Sulston J, Mallett F, Staden R, Durbin R, Horsnell T, Coulson A (1988) Software for genome mapping by fingerprinting techniques. Comput Appl Biosci **4:** 125–132

Swigonová Z, Bennetzen JL, Messing J (2005) Structure and evolution of the r/b chromosomal regions in rice, maize, and sorghum. Genetics **169:** 891–906

Tomkins JP, Davis G, Main D, Yim Y, Duru N, Musket T, Goicoechea JL, Frisch DA, Coe EH Jr, Wing RA (2002) Construction and characterization of a deep-coverage bacterial artificial chromosome library for maize. Crop Sci **42:** 928–933

Yim YS, Davis GL, Duru NA, Musket TA, Linton EW, Messing JW, McMullen MD, Soderlund CA, Polacco ML, Gardiner JM, Coe EH Jr (2002) Characterization of three maize bacterial artificial chromosome libraries toward anchoring of the physical map to the genetic map using high-density bacterial artificial chromosome filter hybridization. Plant Physiol **130:** 1686–1696