Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
GENOMICS
Genomics 89 (2007) 160–165
www.elsevier.com/locate/ygeno

Short Communication

Efficacy of clone fingerprinting methodologies

William M. Nelson [a], Jan Dvorak [b], Ming-Cheng Luo [b], Joachim Messing [c], Rod A. Wing [d], Carol Soderlund [a,*]

[a] *Arizona Genomics Computational Laboratory, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA*
[b] *Department of Plant Sciences, University of California at Davis, Davis, CA 95616, USA*
[c] *The Plant Genome Initiative at Rutgers, Waksman Institute, Rutgers, State University of New Jersey, Piscataway, NJ 08854, USA*
[d] *Arizona Genomics Institute, Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA*

Received 12 June 2006; accepted 18 August 2006
Available online 2 October 2006

Abstract

With the development of new high-information content fingerprinting techniques for constructing BAC-based physical maps, physical map construction is accelerating and it is important to determine which methodologies work best. In a recent publication (Z. Xu et al., 2004, Genomics 84:941-951), Xu et al. evaluated five different techniques (one agarose-based and four using multiple enzymes) and concluded that a two-enzyme technique was superior. In addition, they found that no benefit was gained from fingerprinting more than 10× coverage. In this paper we report our own extensive simulation results, which lead to contrasting conclusions. Our data indicate that the five-enzyme method known as SNaPshot is the most effective and that the assembly can in fact be significantly improved with greater than 10× coverage.

Keywords: Fingerprinting; Physical map; Simulation

Fingerprinted maps have been generated for over 20 years using a range of methods, which vary in the substrate, detection of fragments, and number of enzymes used. The most widespread approach is referred to as the agarose method, which was used for the human map [2]. Recently, high-information content fingerprinting (HICF) [3–7] has been used for large genomes. In all cases, the fingerprints are assembled using the Fingerprinted Contigs program (FPC) [8,9]. Though fingerprinted maps are created by only a small number of laboratories since they require special expertise and equipment, they benefit many scientists as they greatly aid the sequencing of large genomes and can be used to provide the locations of genes and other loci. Hence, it is important to determine the best method for building fingerprinted maps.

Since it would be very difficult to optimize all the different methods to compare them experimentally, it is advantageous to evaluate the methods with simulations, as the results provide a guide to the method(s) worth experimental optimization. Xu et al. [1] provided a simulation of perfect fingerprints comparing five methods that used between one and five enzymes. They found that the two-enzyme method of Zhang and Wing [10] worked better than the five-enzyme HICF method of Luo et al. [4]. We found this result counterintuitive since the five-enzyme method provides much more information than the two-enzyme method (described under Results), so it should discriminate better between false positive (F+) and false negative (F−) clone overlaps. Xu et al. did not provide access to their data or simulation code, so we developed our own simulations to verify their results. In contrast to their results, we found that the five-enzyme method worked best. As our results differ greatly from the published results, we feel it is important to the community to make ours available.

Xu et al. ran their simulations on three chromosomes and a concatenation of the chromosomes, using four genome coverage levels (5×, 8×, 10×, 15×), with the five methods, and two cutoff conditions, which control the approximate amount of error in an assembly (as explained below). We ran our simulations on the same three chromosomes, four genome coverages, and five methods. We also ran simulations on nine additional chromosomes, generated two sets of random clones for each chromosome, and tested three cutoff conditions. We automated the simulation and evaluation so that we could easily run tests on many sequenced genomes, parameters, and cutoff

[*] Corresponding author.
E-mail address: cari@agcol.arizona.edu (C. Soderlund).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.08.008

conditions. Our code and results from all datasets (comprising 1920 different FPC builds) are available at http://www.agcol.arizona.edu/software/fpc/sim, making our results fully verifiable and allowing other researchers to apply our methodology to other sequences if desired.

This article briefly describes the parameters that need to be considered, the five methods, and the results of our simulation compared to those of Xu et al. [1]. We discuss the possible reason for the disparity of results and also discuss a few of the more salient problems we found with their study. We also briefly discuss our experience with two whole genome maps created for maize, one with the agarose method and one with an HICF method.

Results

Methods and parameters

FPC has three parameters that need to be set for assembling the fingerprints into contigs: (1) The "tolerance" is the difference allowed between the sizes of two bands from different clones in order to call them shared (i.e., the same piece of DNA). (2) The "gel length" is the total number of possible values that the bands may have. (3) The fingerprints of each pair of clones are compared, and the probability that the shared bands are a coincidence is computed; if this coincidence score is below a given "cutoff," the clones are considered overlapping. It is important to set the cutoff to minimize the number of F+ and F− overlaps. The equation that computes the score uses the tolerance and gel length, along with the number of shared bands between the two clones.

We use the terminology of Xu et al. for referring to the fingerprinting methods by the number of enzymes used, that is, 1e to 5e. The enzymes and labeling for the methods are as follows (when bands are "labeled," only the labeled bands are detected):

1. The 1e method [11,12] uses the 6-cutter *Hin*dIII and all bands are detected.
2. The 2e method [10] uses the 4-cutter *Hae*III and the 6-cutter *Hin*dIII, and the overhang of the *Hin*dIII is labeled.
3. The 3e method [1] is similar to the 2e but has an additional 6-cutter (unlabeled) *Bam*HI digestion, and it uses a lower tolerance (as discussed below).
4. The 4e method [3] uses separate digestions of the 6-cutter/4-cutter pairs *Hin*dIII/*Hae*III, *Hin*dIII/*Rsa*I, and *Hin*dIII/*Dpn*I, and the *Hin*dIII ends are labeled by a different dye in each pair.
5. The 5e method [1,4] employs the 4-cutter *Hae*III and the 6-cutters *Bam*HI, *Hin*dIII, *Xba*I, and *Xho*I, and the 6-cutters create overhanging ends that are then labeled with four different dyes depending on the end base.

Methods 1e, 2e, and 3e each produce a single set of bands, while methods 4e and 5e generate three and four sets of bands, respectively. Multiple band sets are entered into FPC with a unique offset so that bands from different sets cannot be considered the same [3,7]. For example, as Table 1 shows, 5e has an average of 132 bands, but these are from four sets, resulting in an average 33 bands per set, where the four sets correspond to the four dyes and the corresponding offsets were (0, 10000, 20000, 30000).

In practice, the 1e method is run on an agarose gel and the 2e method (as specified in [1]) is run on a polyacrylamide gel. In both cases, the bands are scored manually with the aid of the Image program [13]. The 3e, 4e, and 5e methods are run on sequencing machines that produce trace files, and a threshold is set to distinguish automatically the true bands from the noise. Under experimental conditions this separation is imperfect, giving rise to false-positive and false-negative bands [7]; however, in both our results and those of Xu et al. this error has not been simulated, so they reflect the performance of the methods under ideal conditions.

The bands produced from sequencing machines have integer and fractional values and it has been shown that the sizing often discriminates between two bands of the same underlying size but with different base composition [4,7]. Since 3e, 4e, and 5e have fractional values, the band sizes and tolerance are multiplied by 10 (as FPC accepts only integers). We used a constant tolerance of 4 (0.4 bp) for the three methods that run on a sequencing machine, as determined by Luo et al. [4], whereas Xu et al. used a tolerance of 2, 5, and 2, respectively. Note that the tolerance of 2 for the 2e method is much less stringent than the tolerance of 4 used for the 3e, 4e, and 5e methods, since the 2e sizes are not multiplied by 10. This lower stringency reflects the fact that the 2e method is assumed to be manually scored.

Xu et al. did not provide gel lengths, though they did provide the range of band sizes for each method, hence, we were able to compute the correct gel lengths, as shown in Table 1 (capillary electrophoresis does not employ gels, but the term "gel length" continues to be used for the number of possible bands).

Table 1
Information about the five methods

Method	No. of sets[a]	Average No. of bands	Band size range (bp)	Gel length	Xu et al. tolerance[b]	Our tolerance
1e	1	30	600–16,000[c]	3,300	0.7%	7
2e	1	47	58–773	715	2	2
3e	1	47[d]	35–500	4,650[e]	2	4
4e	3	109[d]	75–500	12,750[e]	5	4
5e	4	132	35–500	18,600[e]	2	4

[a] Number of distinct sets of bands, where bands cannot be called shared between sets.

[b] Xu et al. and our tolerances are different for 1e since we used migration rates and Xu et al. used sizes. For the three methods using a sequencing machine, we used a constant 0.4 for each, as that is the value found by Luo et al. [4] and Nelson et al. [7] on the ABI 3100 and ABI 3730. Xu et al. simulated the 2e and 5e methods for the ABI 3100 and the 4e method for the ABI 377, stating that the tolerances were 0.2 and 0.5, respectively.

[c] Xu et al. used sizes, whereas we used migration rates. The sizes used by Xu et al. correspond to the migration rates 2293 to 565 that we used. The gel length of 3300 was used as that is the default for agarose.

[d] Xu et al. report 71.7 bands for the 3e method and 73.8 for the 4e method.

[e] Computation of gel lengths: 3e is (500–35) × 10 × 1; 4e is (500–75) × 10 × 3; 5e is (500–35) × 10 × 4.

Cutoff conditions

The same cutoff should not be used to compare the different fingerprinting methods because the coincidence scores for clone overlaps vary widely depending on the tolerance, gel length, and number of bands found in the different methods; instead, as was done by Xu et al., a method-independent “cutoff condition” should be used to determine the cutoff in each case. The cutoff condition is designed to control the amount of error in an assembly, so that the assemblies made by different methods have similar contig quality.

Xu et al. used a “1% Q” cutoff condition, in which a “Q” (questionable) clone is an FPC term for one that does not align well to the underlying consensus band (CB) map (i.e., approximate restriction fragment map). The 1% Q is a reasonable cutoff condition, but it should be realized that the number of Q clones is a very imprecise indicator of the actual number of assembly errors. Q clones may be caused by F+ overlaps [9]; however, the number of Q clones produced by a given F+ overlap varies greatly depending on random circumstances. A F+ overlap causing two contigs to be incorrectly merged on their ends may not produce any Q clones, whereas a F+ overlap causing contig A to be incorrectly incorporated into a central point of contig B will generally result in many Q clones because the clones from contig A will not align to the CB map constructed from the B clones; see Soderlund et al. [9] for further explanation. There may also be a few Q clones in a contig as a result of error in the fingerprints, bands incorrectly called shared due to the tolerance, repetitive bands, and the approximate nature of the FPC algorithm; in these cases, the basic ordering of the clones is still correct but the positions of the endpoints are less precise.

Because of the difficulty in determining one ideal cutoff condition, we simulated three different ones. The first cutoff condition, referred to as “5% chimeric,” was found by raising the cutoff by a factor of 5 as long as the number of chimeric contigs was still less than 5% of the total. The second cutoff condition, referred to as the “first F+,” is found by raising the cutoff by a factor of 5 until a F+ is found. The third condition is the “1% Q” of Xu et al., computed by raising the cutoff by a factor of 5 as long as the number of Q was still less than 1% of the total. The first two conditions use the underlying sequence data to detect the F+ overlaps and chimeric contigs. The third condition does not use information from the underlying sequence, but instead uses the FPC heuristic of Q clones. As described under Materials and methods, our simulation automatically tried different cutoffs and evaluated the results, hence avoiding the possibility of human error. Xu et al. also used a cutoff condition of “no Q’s,” which we did not use since it is too dependent on random factors, as mentioned above.

Automatic simulation

We have performed simulations of the same fingerprinting methods studied in [1], using 13 different sequenced chromosomes from four different species (*Arabidopsis,* rice, fly, human). The simulations were carried out using a fully automated system, which required a sequence as its only input. For each chromosome, four levels of clone coverage were simulated (5×, 8×, 10×, 15×), and two different random clone libraries were generated for each coverage to control for random factors of library selection. The simulated BACs were then digested in silico for the five different fingerprinting methods (see Materials and methods for details). For all 13 chromosomes, the gel lengths and our tolerances were used, as described in Table 1; in addition, the tolerances of Xu et al. were also tested for the 3 chromosomes studied in that paper.

Since the cutoff condition controls for erroneous contigs, the number of contigs determines the success of the assemblies, in that fewer contigs are better. The principal results of our simulations are contained in Table 2, which shows that for 32 cases at 10× coverage, the 5e method performed best in 13 cases, 4e performed best in 7 cases, 5e and 4e tied in 9 cases, and the remaining methods scored best (or tied) in only 5 cases

Table 2
Number of contigs obtained in simulations for 13 chromosomes at 10× simulated coverage, for two different randomly generated clone libraries

Chromosome	Lib	1e	2e	3e	4e	5e
Arab 1	1	25	22	**21**	**21**	**21**
	2	**27**	28	42	41	41
Arab 2	1	21	20	15	11	**2**
	2	22	11	5	6	**3**
Arab 2 (Xu Tol) [a]	1	21	20	16	11	**2**
	2	22	11	7	6	**3**
Arab 4	1	19	16	5	**4**	**4**
	2	19	22	4	**3**	**3**
Arab 4 (Xu Tol) [a]	1	19	16	18	**4**	8
	2	19	22	21	**3**	8
Fly 2L	1	21	11	5	**3**	**3**
	2	30	24	9	5	**3**
Fly 3L	1	26	24	12	5	**3**
	2	22	20	6	3	**3**
Human 18	1	106	70	22	14	**13**
	2	102	77	36	**14**	**14**
Human 19	1	45	123	**22**	41	41
	2	**37**	129	61	41	41
Human 20	1	33	72	29	22	**21**
	2	63	78	29	**21**	**21**
Human 21	1	39	40	21	**11**	16
	2	49	31	18	**8**	17
Human 22	1	45	66	41	38	**21**
	2	46	86	**21**	**21**	**21**
Human 22 (Xu Tol) [a]	1	45	66	44	22	**21**
	2	46	86	21	**21**	**21**
Rice 1	1	56	47	21	21	**20**
	2	67	56	35	24	**22**
Rice 2	1	47	29	13	**7**	9
	2	39	33	8	**7**	7
Rice 3	1	66	37	12	4	**3**
	2	46	41	14	**10**	10
No. of best		2	0	3	16	24

Assemblies were carried out using the “5% chimeric” cutoff condition, as described in the text. The best result for each row is in boldface (both results are in bold in case of a tie), and the bottom row indicates the total number of runs for which each method was best (or tied). Full data, including numbers of F+/F− overlaps, chimeric contigs, Q clones, and singletons, for these and many additional runs are available at (http://www.agcol.arizona.edu/software/fpc/sim).

[a] Used the tolerances for 2e–5e suggested by Xu et al.

altogether. The 4e and 5e methods therefore performed significantly better than the other three, and this finding was duplicated for all other coverages and for both of the other two cutoff conditions (additional data available at http://www.agcol.arizona.edu/software/fpc/sim).

Xu et al. tested three chromosomes and a concatenation of the three chromosomes and presented the accumulation of the results from four coverages, which is shown in Table 3. Also shown in Table 3 is the accumulation of our 16 datasets using the same four coverages. The "map score" was defined by Xu et al. to be a combination of the number of F+ overlaps, F− overlaps, Q clones, chimeric contigs, and total contigs; we do not feel that the map score provides accurate rankings because both the number of Q's and the number of F+ overlaps may vary widely between maps that have equal numbers of chimeric contigs. Since a single map score seems unworkable, we feel that the best way to evaluate mapping technologies is to seek minimum contig count subject to a fixed limit on the error. In Table 3, the results of the map score and number of contigs are shown. Xu et al. always ranked 2e first, and 5e ranked third, fourth, and fifth. Our simulations ranked 5e first based on number of contigs and second based on the map score; our 2e always ranked third, fourth, or fifth.

Xu et al. provided detailed results from human chromosome 22 (see Table 1 of [1]). For a 10× coverage and using the 1% Q cutoff condition, they determined that a 1×10^{-9} cutoff should be used for both 2e and 5e. We used their tolerances and their cutoff of 1×10^{-9}, and the results were very different. For example, counting the number of (contigs, F−, F+, chimeric, Q's) for the 2e method, Xu et al. counted (31, 0, 3, 1, 1) and we counted (79, 102, 112, 1, 1); for the 5e method, Xu et al. counted (30, 0, 8, 1, 14) and we counted (1, 2, 77,072, 1, 1462). Note that they computed the same cutoff for both 2e and 5e using the 1% Q cutoff condition, which is surprising as the 5e method generally uses a much lower cutoff; for example, when we ran this simulation with the 1% Q cutoff condition, it found 5×10^{-9} for the 2e method and 1×10^{-67} for the 5e method. For the complete results of this study, see http://www.agcol.arizona.edu/software/fpc/sim/noerror/chr22.html.

Table 3
Comparison of rankings

	Cutoff condition	Scoring scheme	Comparison
Xu et al.	No Q's	Map score [a]	2e>3e>1e>4e>5e
	No Q's	By contigs [b]	2e>3e>4e>1e>5e
	1% Q's	Map score	2e>3e>5e>1e>4e
	1% Q's	By contigs	2e>4e>3e>5e>1e
Nelson et al.	1% Q's	Map score	5e>4e>3e>1e>2e
	1% Q's	By contigs	5e>4e>3e>2e>1e
	5% chimeric	Map score	4e>5e>1e>2e>3e
	5% chimeric	By contigs	5e>4e>3e>1e>2e
	1st F+	Map score	4e>5e>1e>2e>3e
	1st F+	By contigs	5e>4e>1e>3e>2e

The first set is from Xu et al. [1] and the second set contains the results of this paper.

[a] Map score is a combined score of the number of contigs, F+ overlaps, F− overlaps, chimeric contigs, and Q clones, as defined by Xu et al.

[b] The results are scored by the number of contigs.

Table 4
Reduction in contig number with increasing clone coverage

	1e	2e	3e	4e	5e
5×	100%	100%	100%	100%	100%
8×	64%	64%	66%	50%	52%
10×	49%	50%	51%	38%	38%
15×	28%	30%	34%	27%	27%

For each chromosome and method, the contig numbers at each coverage level (using the "5% chimeric" cutoff condition) were converted to a percentage of the 5× value, and the results for all chromosomes were averaged to compute the entries.

It is also reported in [1] that increasing coverage beyond 10× does not improve the assemblies, but our data do not support this conclusion. As shown in Table 4, we found a steady reduction in contig number with increasing coverage for all methods, with 5e in particular forming on average 28% fewer contigs at 15× than at 10×.

Discussion

In a recent paper, Xu et al. [1] studied five different fingerprinting methods using both simulations and laboratory experiments. These authors arrived at two very unexpected conclusions for the simulated results: (a) the best fingerprinting method was the 2e method, despite the fact that it produces less information about the clones than the 3e, 4e, or 5e, and (b) coverage above 10× makes the assembly worse rather than better. Since physical mapping projects could base the choice of fingerprinting strategy on these assumptions, it is essential that these results are validated. As the code and data from their simulations were not available, we created our simulations as close as possible to their published description. Table 3 shows their results compared to ours, in which they consistently find the 2e is better than the 5e, and we consistently find the opposite. These extreme differences cannot be accounted for by implementation variations, which are discussed below.

We used the sequence from *Arabidopsis* chromosome 2, *Arabidopsis* chromosome 4, and human chromosome 22, as was used for the simulation by Xu et al. There are undoubtedly minor differences in the sequences, but considering that the sequences have been near finished since 1999 [14–16], it is doubtful they have changed enough to cause such radically different results. We created our simulated BACs in the same size range as theirs, and we both randomly selected the clones. A different random set of clones can make some difference in the occasional situation, for example, the difference between each two sets of clones for a given chromosome in Table 2; as is obvious, it does not radically change the results.

Simulation results will vary based on variations in the tolerance, gel length, and cutoff. As shown in Table 2, we used the same tolerances as Xu et al. on the three chromosomes that they analyzed; the one difference is with agarose, discussed below. Since they did not specify their gel lengths, we cannot be sure that we are using the same ones as they did. But since we computed the gel lengths based on their minimum and maximum size bands, they should be the same. By using the

1% Q cutoff condition (data available online), the cutoffs used should have been similar since they produced contigs with the same proportion of error. Hence, it is difficult to explain the large differences in results as being due to variations in parameters.

For 1e, we used migration rates with fixed tolerance, whereas they used sizes with a variable tolerance. By using migration rates, we were able to use a fixed tolerance for all methods. Also, the variable tolerance is not linear in the sizes, and it is not known exactly how to set it. For this reason, and since migration rates are the relevant variable for analysis using Image software, agarose fingerprinting projects generally use migration rates.

We observed significant discrepancies between our in silico digestions and theirs for both the 3e method and the 4e method. For the 3e method, we obtained 47 bands/clone on average (for human chromosome 22) compared to 71.7 reported in their Table 1. Our value seems more reasonable since it is close to that of the nearly identical 2e method; indeed, the purpose of the extra, unlabeled *Bam*HI digestion in 3e is unclear to us. Also, for the 4e method, we obtained 109 bands/clone compared to 73.8 reported in [1]. These two discrepancies do not change the conflicting results of whether 2e is better than 5e or vice versa.

Another possible difference concerns their map score, which uses the number of F− overlaps. An F− overlap occurs when two clones overlap, but are not detected as overlapping based on the FPC cutoff. If this is measured by any two clones that overlap by at least one band, it results in many F− overlaps. If it is measured by considering only the immediate adjacent clone, it provides a more meaningful number as the immediate adjacent F− overlap generally results in a contig splitting in two; we use this second definition. In Table 1 of [1], the parameter F− is reported as 0 for every entry, with no explanation. With any definition of F−, there must be at least one for every contig break that does not correspond to a physical gap in the clone library. They assembled the same set of clones with different cutoffs for the different methods, which resulted in different numbers of contigs, but the same number (0) of F− overlaps, which is not possible. Since our computations show F− overlaps greater than 0, which we can easily compute from the underlying sequence, our map score would be different from theirs even if all else was the same.

Xu et al. consistently find 2e better than 4e and 5e, and 3e better than 5e; we consistently find the opposite. These are radically different results. Logically, it would seem that 4e and 5e should perform better in simulations when there is no error except that introduced by the tolerance, i.e., ignoring experimental complexities. The reason for this is that 4e and 5e are approximately equivalent to several independent runs of 2e and 3e, using different enzymes, and with the information combined. For example, each separate band set of a 5e fingerprint contains a number of bands similar to that of a 2e fingerprint (33 vs. 30), has a similar total range of bands (465 vs. 715), and has considerably higher accuracy of band measurement (tolerance 0.4 bp vs 2 bp). The information content of a fingerprint is determined by these three factors, so the information contained in one set of a 5e fingerprint is at least as great as that in a whole 2e fingerprint, and the combination of all four sets in 5e provides four times as much information. A similar analysis holds for 5e compared to 3e, and 4e compared to 2e and 3e. The extra information in 5e and 4e should not lead to a worse assembly unless either (i) the fingerprints contain a higher proportion of error or (ii) the FPC assembly process is not suited to these fingerprints. The first possibility is certainly not the case in simulations; and the second possibility also does not hold as demonstrated by the simulations reported here.

In addition to a difference in the best method, our results also differed on the optimal coverage. Xu et al. found that as the number of clones increased from 10× to 15× and a 1% Q cutoff condition was used, there was not a significant reduction in contigs, and sometimes the number even increased. We found the opposite, i.e., contig number decreases significantly while the number of Q's remains beneath the 1% threshold. Nevertheless, it is the case that higher coverage levels are likely to generate more Q clones overall, since as the coverage increases, F+ overlaps are more likely. Since a F+ generally results in many Q clones, this situation is usually detected by FPC and can be fixed with a function called the "DQer." Hence, it is always beneficial to increase coverage when possible.

Though the focus of this study is on the differences between our simulation results, we note that laboratory experiments were also reported in Table 2 of [1], in which they found that 2e is superior to 3e and 5e. There are several puzzling aspects to these data. First, the clones were drawn from one contig of a physical map previously constructed using the 2e method [17,18], so it is surprising that the 2e method is reported to assemble into at least nine contigs from the same 157 clones. Furthermore, it is reported that the 5e method with cutoff 1×10^{-2} also assembles into nine contigs, but taking into account the Bonferroni correction for 157 clones one would expect more than 100 false positives with such a low cutoff, so that all clones should assemble into one contig (and in fact the Bonferroni correction generally underestimates the number of false positives because bands are not distributed uniformly [5]). Last, F− is again reported as 0 for all entries in Table 2 [1], without explanation.

We briefly report our experience with laboratory results comparing agarose with a variation of HICF. Nelson et al. [7] assembled two maize FPC maps with the same set of clones, one using the agarose method and one using a variation of HICF very similar to the 4e and 5e methods. The bands for the agarose method were manually selected, which can often lead to a very poor map; but in the case of the maize agarose map, two expert band-callers performed all the manual band selection. As a result, the agarose fingerprints for the maize map are of very high quality, with an average 6% error per clone. The HICF method was run on a sequencing machine by which the bands are automatically selected, and the fingerprints had an average 12.5% error per clone. As described in [7], the HICF method produced 2393 high-quality contigs and the agarose produced 6488 high-quality contigs.

Finally, as noted previously, no error modeling was used either in our simulations or in those of [1]. HICF projects studied to date have had greater error per fingerprint (i.e., more

F+ and F− bands) than seen in agarose projects, but additional simulation data available on our Web site, along with the results from the maize whole-genome project [7], show that HICF remains superior to agarose despite the error. However, the simulation with error also shows that the precision of clone coordinates is greatly improved by reduced error, and fewer contigs are formed at the same cutoff values, so that if substantially lower error could be demonstrated for one variant of HICF, then that variant would be preferable even at higher cost. Hence, a systematic study of the error rates of different HICF techniques and laboratory protocols would be very beneficial.

Materials and methods

Genome sequences

The *Arabidopsis* chromosomes (build 1/21/04) were downloaded from http://www.arabidopsis.org, fly (release 4.3) from http://www.flybase.org, rice (IRGSP version 4.0) from http://rgp.dna.affrc.go.jp/IRGSP/Build4/build4.html, and human from GenBank (Accession No. NC_000001-NC_000024, latest version as of 5/23/06).

Simulation procedure

Our automated simulation pipeline takes as input a chromosome sequence and then performs the following steps:

1. Run a Perl script to create simulated 5×, 8×, 10×, and 15× clone libraries of average insert size 150 kb (±30 kb), by listing all possible clones (i.e., pairs of *Hin*dIII cut sites separated by at least 120 kb and not more than 180 kb) and picking the requisite number at random. For each chromosome sequence, two distinct library sets were created using different seeds for the random number generator.
2. Run a Perl script to digest the clones in each simulated library in silico for each fingerprinting method, retaining band size ranges as specified in Table 1.
3. For all datasets, run a Perl script that iteratively runs FPC on the dataset using a binary search on the cutoff to locate the largest value at which the assembly meets the cutoff condition. A simulation mode was added to FPC to enable automated generation of the cutoff condition results for each trial assembly. A remark is added to each clone with its real coordinates; hence, FPC can easily compute the false positives, false negatives, and chimeric contigs.
4. Run a Perl script to generate a Web page of the results.

The software implementing this pipeline, along with additional data, is available at http://www.agcol.arizona.edu/software/fpc/sim.

Acknowledgments

This study was funded in part by NSF Grants 0211851 and 0213764 and by the Initiative for Future Agriculture and Food Systems Grant 2001-52100-11292 from the USDA Cooperative State Research, Education, and Extension Service.

References

[1] Z. Xu, S. Sun, L. Covaleda, K. Ding, A. Zhang, C. Wu, C. Scheuring, H.B. Zhang, Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage, and contig map quality, Genomics 84 (2004) 941–951.

[2] The International Human Genome Mapping Consortium, A physical map of the human genome, Nature 409 (2001) 934–941.

[3] Y. Ding, M.D. Johnson, R. Colayco, Y.J. Chen, J. Melnyk, H. Schmitt, H. Shizuya, Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting, Genomics 56 (1999) 237–246.

[4] M. Luo, C. Thomas, F. You, J. Hsiao, S. Ouyang, C. Buell, M. Malandro, P. McGuire, O. Anderson, J. Dvorak, High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis, Genomics 82 (2003) 378–389.

[5] B. Meyers, S. Scalabrin, M. Morgante, Mapping and sequencing complex genomes: let's get physical! Nat, Rev. Genet. 5 (2004) 578–588.

[6] W. Nelson, C. Soderlund. 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, 2005, pp. 285–306.

[7] W. Nelson, A. Bharti, E. Butler, F. Wei, G. Fuks, H. Kim, R. Wing, J. Messing, C. Soderlund, Whole-genome validation of high-information-contig fingerprinting, Plant Physiol. 139 (2005) 27–38.

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

[9] C. Soderlund, S. Humphray, A. Dunham, L. French, Contigs built with fingerprints, markers, and FPC v4.7, Genome Res. 10 (2000) 1772–1787.

[10] H.-B. Zhang, R.A. Wing, Physical mapping of the rice genome with BACs, Plant Mol. Biol. 35 (1997) 115–127.

[11] M.V. Olson, et al., Random-clone strategy for genomic restriction mapping in yeast, Proc. Natl. Acad. Sci. USA 83 (1986) 7826–7830.

[12] M. Marra, et al., High throughput fingerprint analysis of large-insert clones, Genome Res. 7 (1997) 1072–1084.

[13] J. Sulston, F. Mallett, R. Staden, R. Durbin, T. Horsnell, A. Coulson, Image analysis of restriction enzyme fingerprint autoradiograms, Comput. Appl. Biosci. 5 (1989) 101–106.

[14] X. Lin, et al., Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*, Nature 402 (1999) 761–768.

[15] K. Mayer, et al., Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*, Nature 402 (1999) 769–777.

[16] I. Dunham, et al., The DNA sequence of human chromosome 22, Nature 402 (1999) 489–495.

[17] Q. Tao, Y.-L. Chang, J. Wang, H. Chen, M.N. Islam-Faridi, C. Scheuring, B. Wang, D.M. Stelly, H.-B. Zhang, Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis, Genetics 158 (2001) 1711–1724.

[18] Y.-L. Chang, Q. Tao, C. Scheuring, K. Ding, K. Meksem, H.-B. Zhang, An integrated map of *Arabidopsis thaliana* for functional analysis of its genome sequence, Genetics 159 (2001) 1231–1242.