A Compartmentalized Approach to the Assembly of Physical Maps

Serdar Bozdag*, Timothy J Close† and Stefano Lonardi*

*Department of Computer Science and Engineering
University of California, Riverside, CA 92521
{sbozdag,stelo}@cs.ucr.edu

†Department of Botany and Plant Sciences
University of California, Riverside, CA 92521
timothy.close@ucr.edu

Abstract

We propose a novel compartmentalized method for the assembly of physical maps from fingerprinted clones. Our assembler exploits the presence of genetic markers at the global level to improve the accuracy of the assembly. Experimental results on the genome of rice and barley demonstrate that the compartmentalized assembler produces significantly more accurate maps, and that it can detect and isolate clones that induce chimeric contigs.

I. Introduction

A physical map is a linear ordering of a set of clones encompassing a chromosome. Physical maps can be generated by first digesting clones with a restriction enzyme such as *Eco*RI, and then detecting their overlaps by matching the lengths of the fragments, called *bands*, produced by the digestion. There are two mainly used methods to read the bands, namely agarose gel-based [24] and high information content fingerprinting (HICF) [10], [11], [19]. In the former, digested fragments are run on an agarose gel to determine their sizes. In contrast, the latter uses multiple restriction enzymes and the fragments are run on a capillary gel electrophoresis.

Physical maps have been historically one of the cornerstones of genome sequencing projects. For instance, in clone-by-clone sequencing, first a physical map is constructed from fingerprinted clones; then, a set of minimally overlapping clones that span the entire genome, called *minimal tiling path* (MTP) is selected; finally, the clones in the MTP are sequenced one by one [14]. The clone-by-clone method has been used to sequence several genomes including *A. thaliana* [21], *H. sapiens* [16], and *O. sativa* [9], [29]. In several recent whole-genome shotgun sequencing projects, physical maps have been also employed to validate and improve the sequence assemblies [30]. This latter strategy has been used, e.g., in the assembly of *M. musculus* [15], *R. norvegicus* [17], and *G. gallus* [25].

For very large and highly repetitive genomes, physical maps that are augmented with "landmarks" such as genetic markers or expressed sequenced tags (ESTs) can be used for *targeted sequencing*. In targeted sequencing, only a region of interest of the genome is sequenced. For instance, one could focus on the gene-rich regions, like in the ongoing sequencing projects of *Z. mays* [7] and *S. bicolor* [8]. Physical maps are essential not only in genome sequencing, but they can also provide a robust infrastructure required by many applications in genomics such as marker assisted breeding, map based cloning of interesting genes, and high throughput EST mapping just to name a few.

Despite an extensive corpus of algorithmic studies in the eighties and nineties on the problem of assembling physical maps from fingerprinting data (see, e.g., [6], [12], [13]), nowadays almost all physical mapping projects rely on a software called FingerPrint Contigs (FPC) [27]. FPC implements an algorithm called *consensus band* (CB) that constructs a physical map using a combination of greedy and heuristic approaches. At the core of the CB algorithm, clones are assigned to contigs based on a coincidence score, called *Sulston score*, which measures the probability that two clones share a given number of restriction fragments (*bands*) according to a simple probabilistic model [28]. For each contig, the algorithm then builds a consensus band map, which is a coordinate system to which clones are aligned. Each distinct band represents one CB unit, and

the length of a clone on a CB map is the number of its unique bands aligned to the CB map.

FPC does not attempt to resolve all the conflicts arising in the assembly of the physical map, but instead provides interactive features that expert users can employ for manual editing. As it turns out in practice, manual editing is an inevitable step in any physical mapping project. The manual editing is tedious, very time-consuming and requires a significant expertise. Clearly, the required amount of manual intervention depends on the initial quality of the physical map produced by the algorithm.

In an attempt to decrease the amount of manual work, i.e., in order to produce more accurate maps, we propose an alternative approach to the assembly that exploits the presence of markers at the global level[1]. Typically, FPC is run on the entire set of fingerprinted clones (approach hereafter called *standard method*). Since fingerprinting data obtained by band sizing from agarose gel or capillary electrophoresis may be inaccurate, the standard method often produces misassembled contigs. If markers are available, as it is usually the case in large genomic projects, a *compartmentalized* assembly is possible. The main idea is to try to assemble first and independently from each other, subset of clones that are *more likely to be truly overlapping*. The markers allow us to determine which clones are more likely to be overlapping. In the physical maps discussed in this paper, the markers are obtained by hybridizing pools of short oligonucleotide probes to a BAC clone library.

Given the popularity and the trust of the scientific community in FPC, our algorithm relies on it and uses it as a subroutine. First, FPC is run independently on each set of BAC clones identified by the probes in each pool and then intermediate assemblies are merged into a single assembly. Second, clone-based and contig-based redundancies are removed from the merged assembly. Third, we use both FPC and a novel algorithm of ours to merge contigs iteratively. FPC's merge process is based on shared bands between contigs, whereas our algorithm is based on shared clones between contigs. The general strategy behind the design of our assembler is "be conservative first". For example, in the beginning of the assembly we merge contigs only if we are quite sure while later we allow riskier moves.

In the experimental section, we report on the assembly of the physical map of two plants, namely rice and barley. Real fingerprinting data is available for both plants. Regarding markers, real hybridization data is available for barley, while rice hybridization data was simulated *in silico*. For both plants, we constructed the physical maps using the standard and the compartmentalized method.

We compared the accuracy of the maps produced by the two methods using a variety of evaluations. We also compared these maps to the manually edited physical maps of rice and barley. Our evaluations show that the compartmentalized method produces significantly more accurate maps than the standard method. In addition, our method is capable of detecting and isolating clones which induce chimeric contigs in the physical maps constructed by the standard method.

[1]FPC can exploit the presence of markers only at a *local* level. When two clones share a marker, they are merged using a higher cutoff than the default.

II. The compartmentalized method

The first step in the compartmentalized method is to run FPC independently on each subset of clones. Clones in each subset (hereafter called *clone sets*) are clones that contain some genetic marker, e.g., they are positive for a pool of probes in some hybridization experiment. Clone sets are not necessarily disjoint.

As stated, our compartmentalized method uses FPC as a subroutine. Since FPC does not offer all of its functionalities in batch mode, we instrumented it so to enable batch mode processing of functions such as END-MERGER, DQER, and REBUILD-CONTIGS. Except for this, we did not make any other modification to the internal code of FPC. FPC's key parameters such as cutoff, tolerance, and fromEnd can be set by the user as usual.

The compartmentalized method consists of five phases, as follow.

A. Initial contig assembly

(A1) Assemble clone sets. FPC's BUILD-CONTIGS procedure is run on the clones of each clone set, one by one. This step generates a "mini" physical map (i.e., contigs and singletons) for each clone set.

(A2) Concatenate physical maps. Next, we concatenate the files containing the maps corresponding to each clone set into a single project. Recall that in general, clone sets are not necessarily disjoint; however FPC cannot handle multiple instances of a clone with the same name. In this phase, we rename multiple copies of the same clone occurring in distinct clone set maps, by adding a distinct suffix. By the end of the assembly, this redundancy is removed completely (i.e., all clones are unique). This renaming process is transparent to the user, since clones are eventually relabeled back to their original names.

B. Redundancy removal

Once all projects are concatenated into a single project, there may be redundant clones as well as redundant

contigs. In this step, this redundancy is removed. Both actions will be repeated in phase C and D.

(B1) Eliminate redundant contigs. A contig is called *redundant* if all of its clones (excluding Q-clones) are completely contained in another contig. Q-clones are clones for which more than 50% of their bands do not align to the CB map [23]. By computing the number of common clones between all contig pairs, redundant contigs are eliminated. In particular, if there are multiple identical contigs, only one of them is kept alive. In this step, all Q-clones that belong to a redundant contig are moved to the singleton set.

(B2) Eliminate redundant clones. A clone is defined to be *redundant* if either (1) it is a singleton and it also occurs in a contig or (2) it occurs multiple times in the singleton set or (3) it occurs multiple times in the same contig. All redundant clones are reduced to one clone in this step.

C. FPC processing

In this phase, the main FPC procedures are run iteratively on the merged project. Steps (C2)–(C6) are repeated a few times until convergence. For more details on FPC functionalities please refer to [23], [27].

(C1) Resolve Q-clones. We run the procedure DQER that reduces the number of Q-clones in an attempt to split the incorrectly merged contigs. DQER runs the CB algorithm on contigs that contain more than $q\%$ of Q-clones, where q is a user-supplied input parameter.

(C2) Merge contigs. We execute the procedure END-MERGER that merges two contigs A and B if M distinct pairs of *end clones*, one of which is in A and the other in B, match each other with a Sulston score lower than the cutoff value. A clone in a contig is an *end clone* if it is within *fromEnd* CB units from one of the ends of the contig, where *fromEnd* is a user-supplied input parameter [22]. To avoid making wrong merges early in the process, we run END-MERGER with increasingly lower values of M (6 for the first iteration, 4 for the second, and 3 for the following iterations).

(C3) Eliminate redundant contigs. See (B1)

(C4) Eliminate redundant clones. See (B2)

(C5) Rebuild contigs. We execute the procedure REBUILD-CONTIGS at this point because END-MERGER does not update the CB map (in FPC v8.0 or above [22]). REBUILD-CONTIGS executes the CB algorithm on the current version of the contigs in order to improve the clone ordering.

(C6) Resolve Q-clones. See (C1)

D. Post-processing

In this fourth phase, a novel algorithm to merge contigs is used and the redundancy present in the physical map is removed completely. Step (D2)-(D4) are repeated a few times until convergence.

(D1) Eliminate redundant Q-clones. A *redundant Q-clone* is a Q-clone that occurs as a non-Q-clone in another contig. The removal of redundant Q-clones is performed only in this phase, since DQER resolves most of the Q-clones in the main processing phase.

(D2) Merge contigs. Recall that END-MERGER merges two contigs if a given number of their end clones overlap with a Sulston score lower than the cutoff. However, in the compartmentalized method, contigs may still share several common clones. Clearly, contigs that share many common clones should be merged. Our MERGE-SIMILAR-CONTIGS algorithm works as follow. For all contig pairs (c_1, c_2) for which $S = c_1 \cap c_2 \neq \emptyset$, the probability that they share clones in S (according to an i.i.d. model) can be obtained as follows

$$p(c_1, c_2) = \frac{\prod_{i=1}^{|S|} \binom{f_{S_i}}{2} \cdot \binom{|M|-2|S|}{|c_1|+|c_2|-2|S|}}{\binom{|M|}{|c_1|+|c_2|}}$$

where M is the multiset of all clones, and f_{S_i} is the number of copies of the i-th element in S in the physical map. Given these probabilities and a specified threshold T_p, we build a directed acyclic graph $G = (V, E)$, where V is the set of contigs that share at least one clone with some other contig, and $E = \{(u, v) | p(u, v) \leq T_p$ and $|u| \leq |v|\}$. When $p(u, v) \leq T_p$ and $|u| = |v|$, source and destination of the edge are selected randomly. We merge contig u to contig $m(u)$, where

$$m(u) = \begin{cases} u & \text{if } outdeg(v) = 0 \\ m(argmin_{(u,v)\in E} p(u, v)) & \text{otherwise} \end{cases}$$

MERGE-SIMILAR-CONTIGS is run until no further merging is possible. As in step (C2), the threshold T_p is increased at each iteration until it reaches a user-supplied maximum (0 for the first iteration, 1e-30 for the second, and 1e-15 for the following iterations).

(D3) Eliminate redundant contigs. See (B1)

(D4) Eliminate redundant clones. See (B2)

(D5) Move redundant clones to the singleton set. After merging contigs, there may be still some clones that occur in multiple contigs. Since the location of these clones in the physical map is ambiguous, they are moved to the singleton set.

E. Finalizing

In this phase, final adjustments are done on the physical map. We reorder the clones and try to resolve any Q-clone introduced in the last phase.

(E1) Rebuild contigs. See (C5)

(E2) Resolve Q-clones. See (C1)

F. Dataset

We used the genomic data of two plants, namely barley and rice, to compare our compartmentalized approach to the standard method.

For barley, HICF fingerprinting data was obtained as part of our NSF funded project [20]. The total number of BAC clones that were successfully fingerprinted is 47,499. About a dozen research groups around the world contributed hybridization data, including our group. We used OLIGOSPAWN [32] to design 12,467 36-mer oligonucleotide (overgo) probes from a dataset of 53,799 barley unigenes [1]. A unigene is obtained as a product of assembling several ESTs. Probes were grouped in 70 pools of usually 192 overgos each, with a maximum of 310 overgos in a single pool. In total there were 1,434 pools; the vast majority were pools containing only one to a few probes processed by colleagues at many locations using a variety of methods, whereas the vast majority of probes were contained in these 70 large pools using a uniform method in our work [20]. The barley BAC library screened against the pools of overgos is a Morex library covering 6.3 genome equivalents [31]. The average insert size is 106 kb. The average number of restriction fragments (bands) is 92.

Since the barley genome has not been sequenced yet, we had to resort to an organism with a known genome for our comparative evaluations. We used the agarose gel-based fingerprinting data and the manually edited physical map of rice obtained from [2] for this purpose. The fingerprinting data was real, but the hybridization data was simulated *in silico*, as explained next. We used again OLIGOSPAWN to design 36-mer unique overgo probes from rice unigene dataset (build 62) obtained from NCBI [4] containing 46,381 unigenes. For about 70% of unigenes, at least one unique overgo probe was designed. We generated 146 pools of rice overgo probes by randomly selecting 200 probes in each pool (except for the last pool, which had 55 probes). To model the hybridization, we decided that if a probe had a perfect match to a BAC clone with 30 or more consecutive bases (out of 36), we considered it a positive hybridization.

In order to carry out the hybridization of rice BAC clones to overgo probes *in silico*, we obtained the sequences of rice clones indirectly by uniquely locating their BAC end sequences (BES) on the rice genome. There were 59,430 rice BAC clones for which BESs were available [3], but only 65% of them had both BESs sequenced. We BLASTed the BESs against the rice genome (fourth release [5]) and filtered out the low-scoring BLAST hits. If a BAC clone had at least one pair of good BLAST hits, it was selected for further analysis. For each selected BAC clone, we checked all possible pairs of left and right BES hits. The coordinates were assigned only when there was only one pair for which (1) the hits were on the same chromosome, (2) the distance between them was consistent with the typical length of a BAC clone, and (3) the orientations of the alignment for the two ends are opposite to each other. If more than one pair met the criteria (1-3), we declared that the location of that clone in the genome could not be determined. Following this procedure, we obtained 26,469 rice BAC clones for which the sequence could be uniquely determined.

We verified the correctness of this procedure by matching the sequences obtained by our method against the small subset of 3,413 BAC clones sequenced by the International Rice Genome Sequencing Project (IRGSP). When we aligned the sequences obtained by our method against the actual sequenced BAC clones using MUMmer [18], only 0.8% of the sequences turned out to be misaligned.

The final dataset of clones for which a sequence was uniquely determined and the fingerprinting data was available contains 22,508 clones (about 10x genome equivalence). The average insert size of these clones is 145 kb and average number of bands is 29.

III. Experimental results and discussion

We applied both the standard and the compartmentalized methods to rice and barley data. The tolerance parameter used in the compartmentalized assembler is the same used in the standard method. This is because the tolerance should be set according to the quality of fingerprinting data [23] and both methods use the same data. The cutoff value is also the same in both methods because the cutoff only depends on genome size [26] and genome composition [23]. We set the parameters in our experiments based on the physical mapping project of rice [9] and barley [20].

TABLE I. FPC statistics of standard, compartmentalized, and manually edited barley and rice physical maps. [a] A Q-contig is a contig that contains at least one Q-clone. [b] Number of contigs that contain at least 15% of Q-clones.

	Clones	Contigs	Singletons	Q-contigs[a]	Q-contig[b]
Barley (Standard)	47,449	7,127	9,634	669	60
Barley (Compartmentalized)	47,449	7,246	13,984	**433**	20
Barley (Manual)	47,449	6,579	4,355	494	**6**
Rice (Standard)	22,486	1,918	860	8	0
Rice (Compartmentalized)	22,486	1,942	1,148	5	0
Rice (Manual)	68,531	179	2,661	**0**	**0**

A. FPC statistics

Table I shows some statistics about the standard, compartmentalized, and manually edited physical maps of rice and barley. The manually edited physical map of rice obtained from [2] contains more clones than the standard/compartmentalized maps because in the latter we used only the subset of clones for which a unique location in the rice genome could be determined. The physical maps of barley contain more Q-contigs (i.e. contigs that contain at least one Q-clone) than the physical maps of rice mostly because of the fingerprinting method as discussed in [22].

According to the statistics produced by FPC, the compartmentalized assembler produces physical maps which contain less Q-contigs than the standard method. Since the manual maps have been extensively edited by experts, it is not surprising that it contains less contigs and Q-contigs than standard/compartmentalized maps. We also observe that for both plants, there are more singletons in the compartmentalized physical map. More interestingly, the singleton set in the standard map of rice is a subset of the singleton set in the compartmentalized map. For barley, about 92% of the singletons in the standard map are also singletons in the compartmentalized map.

When we concentrated our attention on the extra singletons in the compartmentalized map of rice, we were able to determine that 78.1% of these extra singletons were *misplaced* in the standard physical map of rice (see Section III-B for definition of a misplaced clone). This statistics demonstrates that our method is capable of detecting and isolating problematic clones.

B. Comparative evaluations of the physical maps for rice

Since the coordinates of the clones on the rice genome for the 22,508 selected clones in our library are known, more precise comparative evaluations of the two methods are possible for rice than barley. Next, we report on four evaluation metrics to compare the maps produced by the compartmentalized and the standard method, as well as the manually edited map.

Evaluation I (Clone coordinates). We first cluster the clones in each contig according to their locations in the genome. For each contig, two clones are assigned to the same cluster if they are on the same chromosome and the distance between them is smaller than a given threshold. We clustered the clones with several values of the threshold (1 kb to 100 kb) and the results turned out very similar (data not shown). This suggests that two clones are assigned to different clusters usually because they are on different chromosomes. In the following, we show the evaluation results based on clone clustering with 1 kb threshold.

After clustering the clones in each contig, we compute the *cluster score* which is defined as the percentage of clones in the largest cluster. For example, a cluster score of 90% means that 90% of the clones in a contig are on the same chromosome and relatively close to each other. Then, a cluster score for the whole map is computed as the weighted mean of the cluster scores of all contigs in the physical map, using the contig size (i.e., number of clones in a contig) as the weighting factor. The cluster score of each map is shown in Table II. According to the weighted cluster score, the compartmentalized method produces better maps than the standard method.

Once the clone clustering was completed, we also computed the number of misplaced clones and misassembled contigs in each physical map. If the large majority (70% in this evaluation) of the clones in a contig belong to a single cluster then we call *misplaced* the rest of the clones. A contig is called *misassembled* if it contains at least one misplaced clone. As shown in Table II, the

compartmentalized method produces a smaller number of misplaced clones and misassembled contigs than the standard method.

A further analysis on misplaced clones showed that the set of misplaced clones in the compartmentalized rice map is completely contained in the set of misplaced clones in the standard rice map. The compartmentalized assembler isolates 97.4% of the additional misplaced clones in the standard map to the singleton set. This shows that our method can detect and isolate clones that are otherwise misplaced by the standard method. These latter misplaced clones are usually the main responsible for connecting contigs that should not be connected and creating *chimeric* contigs.

We were unable to evaluate the manually edited physical map, since most of the clones in this map cannot be uniquely located in the rice genome.

Evaluation II (Clone order). It is well known that FPC does not order clones within a contig very reliably [23]. Nonetheless, since we have the coordinates of rice clones, we can compute an ordering score for each contig. We define the *ordering score* of a contig as the absolute value of Pearson's product-moment correlation coefficient between the ranking of its clones in the genome and the order of its clones in the contig.

The rankings of clones in the genome are obtained from their coordinates if they belong to the same chromosome. If two clones belong to two different chromosomes then the clone with lower chromosome number has lower ranking than the ranking of the other. For this evaluation, we computed a global ordering score as the weighted mean of the ordering score of all contigs in the physical map, using the contig size as the weighting factor.

The results in Table II show that the compartmentalized method produces contigs where the clone ordering is better than the standard method, probably due to the smaller number of misplaced clones and misassembled contigs.

Evaluation III (Minimal tiling path). As mentioned in the introduction, the minimal tiling path (MTP) of a physical map is a critical component in many genome sequencing projects. Thus, the overall quality of the MTP is a good metric to evaluate physical maps. In this evaluation, first we computed an MTP for both the standard and the compartmentalized physical maps by using the most recent version of FPC (v8.5.3 as the time of writing) with default parameters. Then, we compared the number of the MTP clones, the coverage of the MTP clones on the genome, and the percentage of the consecutive MTP clones that truly overlap on the genome.

The results shown in Table III illustrate that both maps use essentially the same number of clones, but the MTP of the compartmentalized physical map covers almost 1% more of the genome than the MTP of the standard physical map. We also observe that in the physical map obtained by the standard method a higher number of the consecutive MTP clones do not overlap on the genome.

Evaluation IV (Overlapping clones). In our final evaluation, we focus on the set of overlapping clones on the genome. For each pair of clones that are actually overlapping, we check whether they are in the same contig (counted as true positive) or not (counted as false negative). More precisely, only clones that overlap by at least 100 kb are considered in the evaluation. Because, given the parameter set we used, FPC can possibly join two clones if they overlap by at least 70% of their length (100 kb is approximately 70% of their average clone length) [23]. If one or both clones are in the singleton set, this pair is added to the singletons count.

The results in Table IV show that for rice although the true positive rate in the standard map is a little higher than the compartmentalized map, the former suffers from a much higher false negatives rate. Note that the 2% additional false negatives in the standard map are "moved" to the singleton set by the compartmentalized assembler, as we argued previously. We also observe that the manually edited physical map is much better than the compartmentalized and the standard physical maps. This is not surprising given that the manually edited physical map of rice has been curated for more than five years. Furthermore, one should keep in mind that this measure favors physical maps with smaller number of contigs. In the extreme case, a physical map in which all clones were assigned to one single contig would beat all the maps shown here according to this evaluation.

C. Evaluation results for the physical map of barley

Since the barley genome has not been sequenced yet, none of the evaluations explained above can be carried out. We were able, however, to obtain a small dataset from Institute of Plant Genetics and Crop Plant Research (IPK) that gives about 140 lists of BAC clones that hybridized to a single oligonucleotide probe. Some of the pools that we used in the physical mapping consisted of only one probe. By using the BACs identified by the probes in these pools, we were able to extent the dataset to 239 lists. The assumption is that all the clones in each list should overlap.

For each clone set that is identified by a probe, we first computed the contig ID that contains majority of the clones in the set. Then for all clones in the set, we computed the number of clones that were either in that

TABLE II. Global ordering and cluster score of the standard and the compartmentalized physical maps of rice.

	Cluster score	Misplaced clones	Misassembled contigs	Ordering score
Standard	96.43%	675	493	0.8252
Compartmentalized	**97.56%**	**444**	**356**	**0.8426**

TABLE III. A comparison among standard, compartmentalized, and manual physical maps of rice based on their MTPs. "True overlaps" represents the percentage of consecutive MTP clones that overlap on the genome. [c] Since we do not have coordinates for about 50% of the MTP clones, we could not compute the actual coverage.

	MTP clones	Coverage (%)	True overlaps (%)
Standard	2,791	84.90	84.31
Compartmentalized	2,792	**85.24**	**86.94**
Manual	3,365	N/A[c]	86.00

TABLE IV. Evaluation results for standard, compartmentalized, and manually edited physical maps of rice (based on overlapping clones) and barley (based on genetic markers).

	True positive (%)	False negative (%)	Singletons (%)
Rice (Standard)	88.91	8.53	2.56
Rice (Compartmentalized)	88.61	**6.46**	4.94
Rice (Manual)	**92.09**	7.26	0.65
Barley (Standard)	73.69	12.26	14.05
Barley (Compartmentalized)	71.90	**8.56**	19.54
Barley (Manual)	**83.91**	11.11	4.98

contig (counted as true positive), or in another contig (counted as false negative), or in the singleton set. This evaluation is very similar to the one based on overlapping BAC clones performed for rice. However, in this case the dataset is not as reliable. BAC clones identified by the same probe may not necessarily be overlapping (for instance, if BAC clones overlap a repeat region or a gene family). Although this evaluation is not very reliable, it is still rather informative, since it is not biased toward any map.

The results shown in Table IV for barley illustrate that the compartmentalized map has fewer errors than the standard map. There are about 1.8% extra true positives in the standard map, but about 3.5% more false negatives than the compartmentalized map. In other words, the compartmentalized method is able to isolate some clones to the singleton set that are otherwise misplaced by the standard method. Although the true positive rate of the manually edited map is higher than the true positive rate of the compartmentalized map, the compartmentalized map has less errors than the manually edited map. This suggests that starting the manual editing process from the compartmentalized map would reduce considerably the manual intervention.

IV. Conclusions

We proposed a novel compartmentalized approach to the construction of physical maps from fingerprinted clones. The compartmentalized method exploits globally the presence of genetic markers and constructs more accurate physical maps. Consequently, we argue that the compartmentalized method reduces the amount of manual editing that is an inevitable step in any physical mapping project. Additionally, we showed that the MTP produced from the compartmentalized physical map is more reliable, and that should help clone-by-clone sequencing projects.

V. Acknowledgments

The authors would like to thank the members of Prof. Carol Soderlund's group, in particularly Dr. William Nelson for helpful discussions regarding the FPC software. They are also grateful to the people in Prof. Rod Wing's lab, in particularly Dr. Andrea Zuccolo and José Luis Goicoechea for their help with rice BESs and physical map. Finally, the authors would like to thank Dr. Ming-Cheng Luo for providing the fingerprinting data of barley. This project was supported in part by NSF CAREER IIS-0447773 and NSF DBI-0321756.

References

[1] HarvEST home page. http://www.harvest-web.org/.

[2] Rice physical map data. ftp://ftp.genome.arizona.edu/pub/fpc/rice/.

[3] Rice BAC library home page. http://www.genome.arizona.edu/stc/rice.

[4] NCBI rice unigenes home page. ftp://ftp.ncbi.nih.gov/repository/UniGene/Oryza_sativa/.

[5] TIGR rice genome sequence version 4. ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_4.0.

[6] ALIZADEH, F., KARP, R. M., NEWBERG, L. A., AND WEISSER, D. K. Physical mapping of chromosomes: A combinatorial problem in molecular biology. *Algorithmica 13*, 1 (1995), 52–76.

[7] BARBAZUK, W., BEDELL, J., AND RABINOWICZ, P. Reduced representation sequencing: a success in maize and a promise for other plant genomes. *Bioessays 27*, 8 (2005), 839–848.

[8] BEDELL, J. A., BUDIMAN, M. A., AND *et al*. Sorghum genome sequencing by methylation filtration. *PLoS Biol 3*, 1 (2005), e13.

[9] CHEN, M., PRESTING, G., AND *et al.* An integrated physical and genetic map of the rice genome. *Plant Cell 14*, 3 (Mar. 2002), 537–545.

[10] DING, Y., JOHNSON, M. D., CHEN, W. Q., WONG, D., CHEN, Y. J., BENSON, S. C., LAM, J. Y., KIM, Y. M., AND SHIZUYA, H. Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics 74*, 2 (June 2001), 142–154.

[11] DING, Y., JOHNSON, M. D., COLAYCO, R., CHEN, Y. J., MELNYK, J., SCHMITT, H., AND SHIZUYA, H. Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. *Genomics 56*, 3 (Mar. 1999), 237–246.

[12] EVANS, G. A., AND LEWIS, K. A. Physical mapping of complex genomes by cosmid multiplex analysis. *Proceedings of the National Academy of Sciences 86*, 13 (1989), 5030–5034.

[13] FASULO, D. P., JIANG, T., KARP, R. M., SETTERGREN, R., AND THAYER, E. C. An algorithmic approach to multiple complete digest mapping. *Journal of Computational Biology 6*, 2 (1999), 187–208.

[14] GREEN, E. Strategies for the Systematic Sequencing of Complex Genomes. *Nature Reviews Genetics 2* (2001), 573–583.

[15] GREGORY, S., SEKHON, M., AND *et al.* A physical map of the mouse genome. *Nature 418* (2002), 743–750. 10.1038/nature00957.

[16] INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. A physical map of the human genome. *Nature 409* (2001), 934–941. 10.1038/35057157.

[17] KRZYWINSKI, M., WALLIS, J., AND *et al.* Integrated and Sequence-Ordered BAC- and YAC-Based Physical Maps for the Rat Genome. *Genome Res. 14*, 4 (2004), 766–779.

[18] KURTZ, S., PHILLIPPY, A., DELCHER, A., SMOOT, M., SHUMWAY, M., ANTONESCU, C., AND SALZBERG, S. Versatile and open software for comparing large genomes. *Genome Biology 5*, 2 (2004), R12.

[19] LUO, M., THOMAS, C., YOU, F., HSIAO, J., OUYANG, S., BUELL, C., MALANDRO, M., MCGUIRE, P., ANDERSON, O., AND DVORAK, J. 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.

[20] MADISHETTY, K., CONDAMINE, P., MOSCOU, M., SVENSSON, J., ZHENG, J., WANAMAKER, S., BHAT, P., RODRIGUEZ, E., WALIA, H., BOZDAG, S., REZNIK, J., LE, H., LUO, M.-C., JIANG, T., LONARDI, S., WITT, H., YOU, F., KLEINHOFS, N. S. A., COOPER, L., GILL, K., MUEHLBAUER, G., WISE, R., AND CLOSE, T. J. Towards a physical map of the barley "gene space". in preparation, 2007.

[21] MARRA, M., KUCABA, T., AND *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nat Genet 22*, 3 (July 1999), 265–270.

[22] NELSON, W., BHARTI, A., BUTLER, E., WEI, F., FUKS, G., KIM, H., WING, R., MESSING, J., AND SODERLUND, C. Whole-Genome Validation of High-Information-Content Fingerprinting. *Plant Physiol. 139*, 1 (2005), 27–38.

[23] NELSON, W., AND SODERLUND, C. Software for restriction fragment physical maps. In *The Handbook of Genome Mapping: Genetic and Physical Mapping*, K. Meksem and G. Kahl, Eds. Wiley-VCH, 2005, pp. 285–306.

[24] OLSON, M., DUTCHIK, J., GRAHAM, M., BRODEUR, G., HELMS, C., FRANK, M., MACCOLLIN, M., SCHEINMAN, R., AND FRANK, T. Random-Clone Strategy for Genomic Restriction Mapping in Yeast. *PNAS 83*, 20 (1986), 7826–7830.

[25] REN, C., LEE, M., YAN, B., DING, K., COX, B., ROMANOV, M., PRICE, J., DODGSON, J., AND ZHANG, H. A BAC-Based Physical Map of the Chicken Genome. *Genome Res. 13*, 12 (2003), 2754–2758.

[26] SCALABRIN, S., MORGANTE, M., AND MEYERS, B. Mapping and Sequencing Complex Genomes: Let's get Physical! *Nature Reviews Genetics 5* (2004), 578–588. 10.1038/nrg1404.

[27] SODERLUND, C., HUMPHRAY, S., DUNHAM, A., AND FRENCH, L. Contigs Built with Fingerprints, Markers, and FPC V4.7. *Genome Res. 10*, 11 (2000), 1772–1787.

[28] SULSTON, J., MALLETT, F., STADEN, R., DURBIN, R., HORSNELL, T., AND COULSON, A. Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci. 4*, 1 (1988), 125–132.

[29] TAO, Q., CHANG, Y., WANG, J., CHEN, H., ISLAM-FARIDI, M., SCHEURING, C., WANG, B., STELLY, D., AND ZHANG, H. Bacterial Artificial Chromosome-Based Physical Map of the Rice Genome Constructed by Restriction Fingerprint Analysis. *Genetics 158*, 4 (2001), 1711–1724.

[30] WARREN, R. L., VARABEI, D., AND *et al.* Physical map-assisted whole-genome shotgun sequence assemblies. *Genome Res 16*, 6 (June 2006), 768–775.

[31] YU, Y., TOMKINS, J., WAUGH, R., FRISCH, D., KUDRNA, D., KLEINHOFS, A., BRUEGGEMAN, R., MUEHLBAUER, G., WISE, R., AND WING, R. A bacterial artificial chromosome library for barley (*Hordeum vulgare L.*) and the identification of clones containing putative resistance genes. *Theor. Appl. Genet. 101* (2000), 1093–1099.

[32] ZHENG, J., SVENSSON, J. T., MADISHETTY, K., CLOSE, T. J., JIANG, T., AND LONARDI, S. OligoSpawn: a software tool for the design of overgo probes from large unigene datasets. *BMC Bioinformatics 7*, 7 (2006).