BIOINFORMATICS

Vol. 20 no. 2 2004, pages 140–147
DOI: 10.1093/bioinformatics/bth017

A strategy for assembling the maize (Zea mays L.) genome

Scott J. Emrich[1,2], Srinivas Aluru[1,2,7,8,*], Yan Fu[3,6], Tsui-Jung Wen[4], Mahesh Narayanan[7], Ling Guo[1,6], Daniel A. Ashlock[1,5,8] and Patrick S. Schnable[1,4,6,8,9]

[1]Bioinformatics and Computational Biology Graduate Program, [2]Department of Electrical and Computer Engineering, [3]Interdepartmental Genetics Graduate Program, [4]Department of Agronomy, [5]Department of Mathematics, [6]Department of Genetics, Development and Cell Biology, [7]Department of Computer Science, [8]L.H. Baker Center for Bioinformatics and Biological Statistics and [9]Center for Plant Genomics, Iowa State University, Ames, IA 50011, USA

Received on November 9, 2003; accepted on November 11, 2003

ABSTRACT

Summary: Because the bulk of the maize (*Zea mays* L.) genome consists of repetitive sequences, sequencing efforts are being targeted to its 'gene-rich' fraction. Traditional assembly programs are inadequate for this approach because they are optimized for a uniform sampling of the genome and inherently lack the ability to differentiate highly similar paralogs. **Results:** We report the development of bioinformatics tools for the accurate assembly of the maize genome. This software, which is based on innovative parallel algorithms to ensure scalability, assembled 730 974 genomic survey sequences fragments in 4 h using 64 Pentium III 1.26 GHz processors of a commodity cluster. Algorithmic innovations are used to reduce the number of pairwise alignments significantly without sacrificing quality. Clone pair information was used to estimate the error rate for improved differentiation of polymorphisms versus sequencing errors. The assembly was also used to evaluate the effectiveness of various filtering strategies and thereby provide information that can be used to focus subsequent sequencing efforts.

Contact: aluru@iastate.edu

INTRODUCTION

As the best-studied biological model for cereals and one of the world's most important crops, there is a strong rationale for sequencing the maize genome. Approximating the maize genome at 2500 million bases (MB) (Arumuganathan and Earle, 1991) makes it comparable in size with that of humans. Because 65–80% of the maize genome consists of tens of thousands of copies of large, highly homogenous retrotransposons (Bennetzen, 1996), many attendees at an international meeting convened in St Louis during 2001 (Bennetzen *et al*., 2001) were concerned that it would not be possible to assemble the maize genome using a shotgun-based sequencing approach. Instead, most attendees concluded that it would be more desirable to utilize various 'filters' prior to sequencing, so as to enrich for the 'gene-rich' fraction of the genome.

The National Science Foundation is funding two projects to compare three sequencing strategies. The first strategy of methyl filtration (MF) is based on the finding that retrotransposon sequences are greatly reduced in hypomethylated DNA (Rabinowicz *et al*., 1999; Meyers *et al*., 2001). The second strategy enriches for low-copy sequences by sequencing the high $C_o t$ (HC) fraction of the genome (Peterson *et al*., 2002; Yuan *et al*., 2003). The third obtains a 'random' sample of the genome by sequencing bacterial artificial chromosomes (BACs) and BAC ends.

Traditional assembly programs are inadequate since they are optimized for uniform sampling of the genome. Also, they cannot inherently differentiate among near-identical paralogs (NIPs) that arise via segmental duplications. We have recently established that the maize genome contains many NIPs (Wen *et al*., manuscript in preparation) that can be identified because they contain one or more cismorphisms (Hurles, 2002), i.e. polymorphisms between paralogs. Segmental duplications have complicated the assembly, annotation and analysis of the human genome. The segmental duplications in the human genome are being identified by virtue of their over-representation among randomly generated sequences (Bailey *et al*., 2002). This approach is not suitable for use in the maize genome because MF and HC sequences do not represent a random sample of the genome. We instead propose to exploit the nonuniformity of polymorphisms to identify NIPs that exhibit cismorphisms at rates less than the sequencing error rate.

Here, we report algorithmic, statistical and biological foundations developed for an accurate assembly of the maize genome. This software is based on innovative parallel algorithms and runs on multiprocessor platforms to ensure

*To whom correspondence should be addressed.

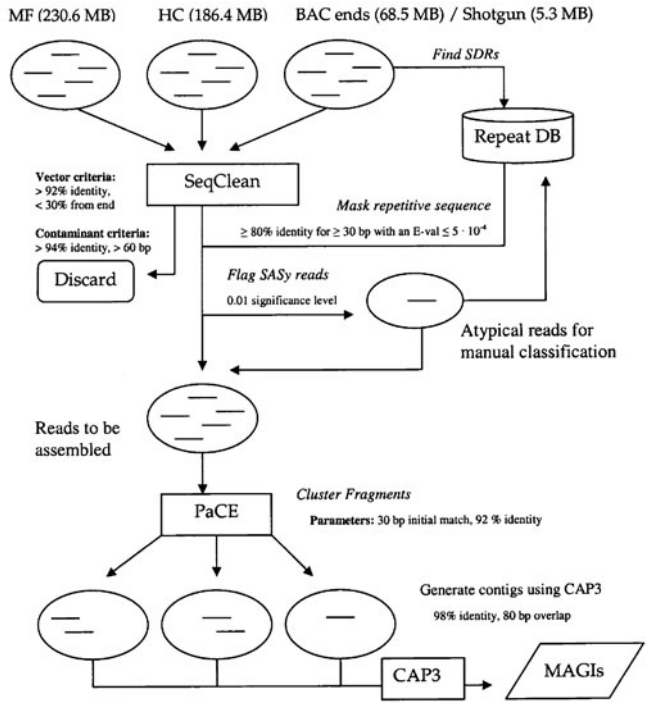

Fig. 1. Maize assembly pipeline.

scalability as the number of sequences increases. Our overarching goal is to produce an accurate assembly as quickly as new data becomes available. Our results therefore explore the peculiarities of current data and techniques, analytical methods for assessing errors and an assembly pipeline (Fig. 1) with computational and in progress biological verification.

METHODS

Input data, masking of low quality and contaminant sequences

Zea mays genomic survey sequences (GSSs) from the inbred line B73 obtained from MF, HC or 'random' sequencing approaches were downloaded from GenBank on July 27, 2003 and consisted of 730 974 fragments totaling 490.8 MB.

These sequences were first checked for sequence contamination and extensive simple repeats with the SeqClean script (http://www.tigr.org/tdb/tgi/software/). Vector contamination was trimmed by using the univec_core db (ftp://ftp.ncbi.nih.gov/pub/UniVec/). Contamination was identified via strong sequence similarity to any one of the following in GenBank: *Escherichia coli* K12 (U00096), bacteriophage phi_X174 (J02482), *Z.mays* chloroplast genome (NC_001666) and the draft *Z.mays* mitochondria NB genome (C.Fauron, University of Utah, personal communication). 16 478 sequences were completely masked, the majority of which were due to similarity to the mitochondrial genome. Because assembly including mitochondrial contamination produces contigs with equally high similarity to the mitochondria genome (data not shown), autosomal regions should not be discarded.

Determining and masking repetitive elements

A principal computational difficulty in assembling the maize genome is the abundance of repetitive elements. Implementing 'mathematically defined' repeats, as was performed during the assembly of the rice genome (Yu *et al*., 2002), may not be effective in maize due to the intentional sampling biases introduced by the use of 'gene rich' filters.

Since BAC end sequences provide a nearly uniform sample of maize genomic DNA (Meyers *et al*., 2001) statistical analysis of these sequences might provide additional uncharacterized high-copy elements for a repeat database. Then, such a database could be used to mask repeats prior to assembly. We term repetitive elements defined in this fashion 'statistically defined repeats' (SDRs). 74 442 maize BAC end sequences downloaded from GenBank in late June 2003 were grouped via single linkage clustering based on exact matches of at least 20 bases identified with our implementation of a generalized suffix tree (GST; Gusfield, 1997). The 1667 BAC end sequences (2.2%) that remain as singleton clusters are most likely low-copy within the maize genome, but interestingly only nine of these have maize expressed sequence tag (EST) hits by homology search (identity $\geq$98%, overlap $\geq$50 bp). The largest of these clusters may contain uncharacterized repeats. Therefore, clustering simply acts as a statistical sieve for BAC end sequences.

Maximal exact matches within the top 10% of maize BAC end clusters of size 30 or greater were then located using our implementation of the suffix tree data structure. The CAP3 sequence assembler (Huang and Madan, 1999) was then used to generate consensus sequences from these 'seed' matches. Maximal exact matches have been previously used to process BAC ends for repetitive sequences in an iterative approach using BLAST (Volfovsky *et al*., 2001). This new clustering approach is effective since it uses multiple suffix-tree algorithms to avoid alignment-based clustering on all possibilities.

A BLASTX verification of this method showed that over 99.5% of nr protein database matches to maize SDRs (minimum E-value of 1e $-$ 10) consist of retrotransposon-related sequence including putative HELITRON elements described previously (Lal *et al*., 2003). Within this statistical abstraction, however, the distinction between repetitive sequences and highly conserved proteins is blurred. To improve the overall quality of an assembly, any region of a SDR with at least 80% identity over 20 bases to genes that encode known plant proteins—that do not match any characterized repetitive sequence—were removed. In all, 265 relatively short SDRs (mostly $<$50 bp) met this criterion.

TIGR's Cereal Repeat Database 2.0 (http://www.tigr.org), the Wessler Laboratory's database of plant transposable

elements (Jiang *et al*., manuscript in preparation), maize SDRs and atypical repetitive sequences from previous assemblies were combined into an non-redundant repeat database. Sequences were pruned if there was another sequence that was at least 95% identical over 90% of the original sequence's length, resulting in a comprehensive repeat database containing 8595 sequences totaling 5.95 MB.

Masking prior to assembly uses a Perl script that relies upon standalone BLAST (Altschul *et al*., 1990). This approach is very similar to MaskerAid (Bedell *et al*., 2000), but uses a different search engine and was optimized for maize. BLAST hits with at least 80% identity over 30 bases with an associated minimum E-value of $5e-4$ are masked, along with any hit with 80% identity more than 60 bases. The latter criterion was added to mask AT-rich LTRs that do not pass the minimum E-value criteria due to their biased composition. Optimization was performed on shuffled fragments using multiple large random samples that were locally aligned against the repeat database. Based on these tests, the false negative rate of masking is very close to the minimum BLAST E-value used.

To determine the rate of false positive masking associated with this approach a set of gene-associated sequences that contain few repetitive elements was required; coding regions turned out to be the cleanest dataset available. Exons were located within 1036 *Z.mays* cDNA sequences downloaded from plantGDB (www.plantgdb.org) using a BLASTX database search against *nr* (minimum E-value $1e{-}10$). This search returned on average 1100 bases per cDNA. Twenty cDNAs were partially masked; 10 of these match a ∼50 bp region of *pl* transcription factor (AF015269); the remainder match non-repetitive coding regions of several genes (e.g., *waxy1* (K01965), *alcohol dehydrogenase1* (M27366), and *booster1* (AF326577)) that are present in the TIGR repeat database. These coding regions were presumably inadvertently included in the repeat database because they are adjacent to repetitive elements. All other matches were short and occurred at a frequency close to the rate of false negatives discussed above.

Final masking results for the three major types of maize genomic fragments can be found in Table 1, along with a comparison to TIGR's latest assembly when possible. In total 19.6% of this dataset is masked prior to determining overlaps. Since the numbers of reads are not equal between our method and TIGR's, it is not clear if our method is actually more restrictive than theirs. It is interesting to note, however, that our masking of random-like BAC ends is close to the estimated portion of the maize genome that is repetitive (Meyers *et al*., 2001) and is much higher than filtered MF and HC sequences.

IMM-based sequence classification

An empirically optimized likelihood ratio test was effective in locating Statistically Atypical Sequence (SASy) fragments within the maize dataset. Two separate IMMs were constructed using 'build-icm' within the GLIMMER package (Delcher *et al*., 1999), one from the sequence types considered atypical and another from randomly selected maize fragments with no similarity to these atypical sequences. Using IMMs, which are equivalent to multiple probabilistic suffix trees, is comparable with the work of Bejerano and Yona (2001) in successfully classifying protein families. The use of a likelihood ratio, however, allows the assignment of a p-value for each genomic fragment.

Table 1. Repeat masking of the three major sources of GSS. Percentage of bases masked in each sequence type

Sequence type	ISU MAGIs 2.3 (%)	TIGR AZMs 2.0 (%)
High $C_o t$	5.9	19
Hypomethylated	18.8	31
BAC ends	57.6	—
Total	19.6	25

MAGI, Maize Assembled Genomic Islands.

Although some prokaryotic homologs within the maize genome are also classified as SASy, all phage contamination is identified using SASies, as are potentially uncharacterized repetitive sequences with significant protein database matches to known transposable elements in *nr*. Examples include cinful polyprotein (AF114171, $1e{-}38$), gypsy-type retrotransposon (AF466203, $1e{-}34$), and putative Tam3-like transposon protein (AC079179, $5e{-}41$). Verified repetitive sequences are used to augment the repeat database while the remaining sequences are kept in the assembly pipeline.

Empirical determination of sampling biases and sequencing error rates

Potential sampling biases of the MF and HC filters were empirically determined using 73 maize genes with known structure within GenBank and seven maize genes sequenced by the Schnable lab (Yao *et al*., manuscript in preparation). The null hypothesis—that GSS sampling is uniform across the length of genes—was tested using a binomial test of the significance of the GSS starting location. There was little evidence for a potential 5′ versus 3′ bias. There was substantial bias, however, within the annotated gene structure. HC sequences appear to oversample non-exonic sequences, i.e. 5′ untranslated regions (UTRs), introns and 3′ UTRs, with stronger bias towards UTRs. For one-third of the 80 maize genes analyzed, MF sequences seem to oversample the entire gene structure (relative to the entire sequence record), and the exons of all seven Schnable laboratory genes are oversampled using these sequences. These results provide empirical evidence that targeted sequencing is non-uniformly sampling within maize genes.

Because clone pair sequencing is expected to generate the same sequence twice in the overlap, observed disagreements can be used to estimate empirically the sequencing error rate in maize GSS fragments. Average fractional disagreement rates of 0.0025 for HC and 0.0035 for MF sequences were

determined from an overall average of 434 bases per clone pair of 51 305 overlapping clone pairs. Both estimates can be modeled by exponential distributions.

Based on an exponential model fitted with an average error rate of 0.0025 (HC rate), a CAP3 assembly based on 98% identity should miss at most three overlaps out of 10 000 due to excessive sequencing errors. Because many of the errors are located in relatively low-quality regions (Phred score <30) more stringently end-trimming substantially reduces the error rate within GSS fragments (Fu *et al*., manuscript in preparation).

Assembly of non-uniform genomic fragments

Computing all pairwise alignments to determine overlapping fragments is computationally expensive. Sequence assembly programs, therefore, first determine pairs of sequences with a sufficiently long exact match. Alignment is then restricted to such pairs, which we term 'promising pairs'. Most assembly programs locate these exact matches using a lookup table based on substrings of a small, fixed length, and requires space proportional to the size of the input fragments. Under the assumption of uniform sampling of the genome to be assembled, this approach works well and generates $O(n)$ promising pairs that need to be collated and processed in some fashion, where n is the number of input fragments.

The problem with non-uniform sampling is that there are potentially a quadratic number of promising pairs. Even if time is not a constraint, the major obstacle is the memory required for storing $O(n^2)$ promising pairs. Using CAP3 as an example on a single processor of our IBM xSeries cluster, only 50 000 masked sequences can be assembled using 1 GB of RAM. Instead, we used the parallel EST clustering tool (Kalyanaraman, 2003), Parallel Clustering of ESTs (PaCE), to reduce significantly the problems inherent in non-uniform samples and established a pipeline that quickly assembles maize contigs. Clustering genomic fragments provides a method for reducing a large assembly problem into many, but smaller, assembly problems; an approach used by the recent Phusion (Mullikin and Ning, 2003) and ARACHNE (Batzoglou *et al*., 2002) assemblers for other eukaryotic organisms. PCAP (Huang *et al*., 2003) is an assembly program that also runs in parallel but uses lookup tables to generate promising pairs under the assumption of uniform sampling. A tool designed for EST clustering, however, is ideal for processing MF and HC filtering because of its inherent advantages in processing non-uniform samples resulting from differential sampling of transcripts. Hence, the algorithmic innovations developed for solving the EST clustering problem in PaCE can also be used effectively for complex genome assemblies.

The primary innovation in PaCE is that it identifies promising pairs in batches based on maximal exact matches using a distributed GST constructed from all of the sequences and their Watson–Crick complements. Therefore, PaCE never generates all the promising pairs at once. Storing the GST itself requires memory proportional to the size of the input sequences.

An additional benefit from using a GST is that exact matches of an arbitrary size k can be found for the detection of promising pairs without generating $(k-w+1)$ w-length matches as happens in using a lookup table based on w-long substrings. PaCE also generates promising pairs in decreasing order in the length of the maximal exact match. Intuitively, longer exact matches imply a higher chance of 'good' alignments; so by using this measure it is likely that the most significant promising pairs are generated first. This is a valid approach since the order in which clusters are merged does not impact the single linkage clustering technique used by PaCE. Generating pairs in decreasing order of match length as done in PaCE requires sorting the internal nodes of the GST by string depth and no extra space.

Another key innovation within PaCE is single linkage clustering which reduces the number of pairwise fragment alignments without sacrificing quality. Similar to locating SDRs within BAC end sequences, two fragment clusters are merged when a sufficient similarity score is detected between two members of these clusters. Since alignments are performed in decreasing order of maximal exact match length, a given pair might have already been put into the same cluster based on a previous merge and, thus, it is not necessary to perform the alignment. This often equates to less work, especially in the case of EST clustering or the related genome assembly problem where multiple sequences cover a particular 'gene island' due to biased sampling. An important observation is that in single linkage clustering a maximum of $n-1$ merges are possible since the biggest cluster possible is of size n. Even though there are as many promising pairs to investigate as maximal exact matches, only a linear number of alignments are sufficient to provide the optimal clustering of fragments where each PaCE cluster corresponds to a single maize contig. In practice, there are on average 1.08 contigs obtained per PaCE cluster due to the lower alignment threshold used in clustering.

Figure 2 illustrates the work performed by PaCE as the number of genomic fragments increases. The number of promising pairs grows quadratically as the number of fragments increases. A somewhat unexpected result, however, is that the difference in pair generation between masked and unmasked fragments grows linearly—indicating an underlying uniform distribution—by a factor of about five. A fragment-independent approach should remove most of these repetitive sequences where 'mathematical' repeat approaches might fail. Since PaCE clustering offsets unsuccessful alignments by processing them in parallel, all promising pairs can be processed quickly without assuming a uniform distribution of fragments.

Assembly and verification of PaCE clusters

A series of empirical tests was performed to obtain the optimal parameters that balance overall run-time with clustering

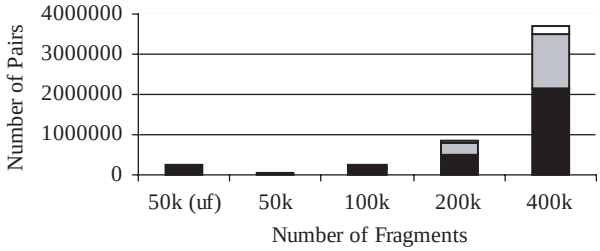

Fig. 2. The number of pairs generated by PaCE as a function of the size of randomly shuffled input. The black and white bars designate failed and successful promising pairs, respectively. Gray pairs are unaligned and represent a significant reduction in work.

quality. PaCE was run with a minimum initial exact match of 20, and a global alignment threshold of over 80% identity to determine ideal parameters for alignment and exact match criteria based on an assembly overlap of 95% identity.

From this experiment, it was determined that an exact match criterion of 30 bases had a false negative overlap rate of 0.001 while decreasing the number of pairs generated by a factor of four and this value is used in PaCE. Generating a CAP3-based assembly takes <24 h to complete using one processor of the IBM xSeries cluster and served as the basis for preliminary analyses, with a median cluster size of five and a maximum cluster composed of only 96 sequences. Unmasked fragments are used for assembly with the following CAP3 parameters: 98% identity, 80 bp overlap and 60 bp overhang. Using more stringent assembly options, as supported by empirical estimations of sequencing errors, allows our pipeline to potentially differentiate more paralogs within the maize genome as compared with a lower threshold. If, in a future implementation, all the 64 processors are used, assembly should take only 1/64th as much time, i.e. <30 min.

We note that once the sequences are partitioned by PaCE clustering, any assembly program could be used in parallel to generate contigs.

Near-identical paralogs

Over 8% of public human single nucleotide polymorphisms (SNPs) are potentially paralogous sequence variants (i.e. cismorphisms; Hurles, 2002) rather than actual SNPs (Cheung *et al*., 2003). Based on the analysis of EST assemblies followed by wet lab validation, we have established that the maize genome (namely, the inbred line B73) also contains a high frequency of NIPs (Wen *et al*., manuscript in preparation). It would be possible to prevent the misassembly of many NIPs by using more stringent CAP3 parameters as discussed, but this approach could separate some legitimate contigs as well.

A possible solution to this problem exploits the fact that differences within GSS fragments due to sequencing errors are random but those due to genetic divergence are not. A contig containing multiple sequences from each of two or more NIPs will have positions that have an apparently above average rate of sequencing error, and we term such positions coincidental polymorphisms (CPs). The presence of several CPs within a single CAP3 alignment should provide strong evidence that the alignment contains multiple members of a gene family (i.e. NIPs) under the assumption that errors are independent and identically distributed (i.i.d). The use of the multinomial distribution permits the construction of a model that can be used to provide a p-value for rejection of the null hypothesis 'this contig contains GSS fragments from a single gene'.

We note, however, that the assumption of uniform errors may not be valid for the following reasons. First, clone pairs, which were used for determining error rates, should not be treated as independent fragments for determining the statistical significance of CPs. In addition, the inclusion of low-quality ends of sequence reads is likely to result in an non-uniform distribution of errors. We therefore decided to examine single statistically significant columns within alignments of GSSs for validation of observed CPs (which should not be affected by these concerns) until we work out the proper test statistic that solves these problems.

Following the assembly of the maize genome, contigs that contained putative CPs were identified and flagged if the test statistic for any CP in that contig was less than 0.01. In total, 1108 contigs contained at least one statistically significant CP. The trace files of a sample of these were aligned using Sequencher 4.1 (e.g. Fig. 3). Most of the examined contigs that contained a single CP were validated following the manual examination of the sequencing trace file of each GSS included in the contigs. The putative CP-containing contigs that could not be validated by manual checking of trace files are likely false-positives because the putative CPs are located in regions of lower quality, typically at the ends of the GSS reads. As discussed above, more stringent trimming and incorporation of quality values should allow greater specificity of using combined evidence for prediction of CPs.

An automated version of this process is envisioned for future versions of the assembler that will both tentatively divide alignments that appear to contain NIPs and generate a log to alert an human expert to review suspicious clusters.

Sequence-based scaffolding of 'gene archipelagoes'

Maize 'gene islands' will eventually be linked into 'gene archipelagoes' for finishing and further analysis. We propose a computational approach, which we feel is an effective and useful tool for clustering gene islands into archipelagoes and bridging gaps induced by masking repetitive sequences (Yu *et al*., 2002).

Although this problem is being approached from two different perspectives, namely cDNA and protein evidence, the central idea remains the same: spliced exons can cover a larger portion of the genomic sequence. The main distinction between using protein or nucleotide sequences is the extent

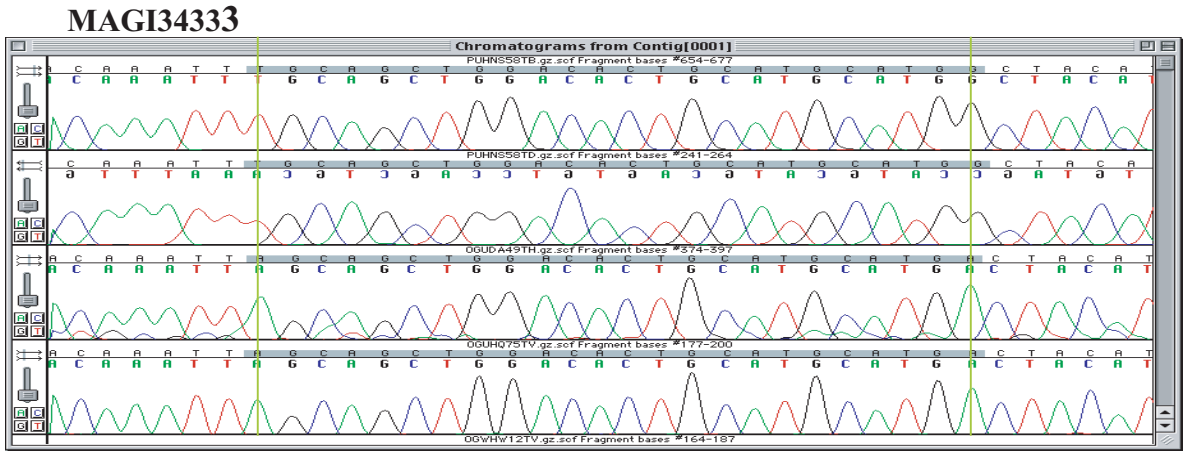

Fig. 3. Alignment of the sequence trace files for the CP-containing GSS Contig MAGI_34333. The grey vertical lines indicate two validated CP sites in this contig that has BLASTX matches to CCR4-associated factor related protein (E-value = 3e − 44).

of similarity one wishes to detect to generate 'scaffolds' based on spliced alignment. We term these approaches 'protein lookup' and 'EST lookup', and an important side benefit of this analysis is the annotation of contigs within the assembly.

Our current 'protein-lookup' approach uses BLASTX with yeast protein sequences, obtainable from EMBL, as the database. Yeast was chosen because almost all yeast proteins have been experimentally validated and if this method works on yeast it will also work on plant protein databases that are evolutionarily closer to maize.

BLASTX hits were considered 'valid' if the e-value of the alignment was lower than e−10. Of the 91 690 maize GSS contigs, 4008 met this criterion with matches to 1145 different yeast proteins. Protein lookup groups these GSS contigs that have the same most significant yeast protein match, and this reduced the set to 88 825 contig clusters. Of these, only 693 were non-singleton clusters.

To improve run-time for clustering in this manner a PaCE-like approach was taken. A GST tree is built in parallel including sequences of the protein database as well as the protein sequences produced by converting the GSS contigs using the six possible reading frames. Pairs of sequences of the form (database, contig) were generated in decreasing order of similarity and spliced alignments are performed on such pairs. The advantage of such a clustering-based approach is that it should reduce alignments computed when compared with an ALL versus ALL BLAST.

To test the analogous approach but based on ESTs (i.e. 'EST lookup'), spliced alignments of all GSS contigs (here, the term contig is used to refer to both contigs and singletons) and 3′ ESTs were performed using GeneSeqer (Usuka *et al.*, 2000). Several cDNA libraries were prepared from the same inbred line as were the GSS fragments, i.e. B73 (Wen *et al.*, manuscript in preparation) and this EST dataset contains 6270 singletons and 3751 contigs (i.e. 10 021 unique genes).

Table 2. The numbers of 2030 GSS contigs involved in each of four types of spliced alignment to EST contigs described within the text

GSS contig	EST	
	=1	>1
=1	1616	134
>1	220	60
Total	1836	194

A GeneSeqer alignment of GSS contig to an EST cluster was parsed out if it contained at least two qualifying spliced alignments (i.e. exons) with at least 98% sequence identity. A total of 2030 of the GSS contigs aligned to one or more EST clusters and these alignments were classified into four types: Type I, a single GSS contig had only a single EST match; Type II, multiple GSS contigs aligned with a single EST; Type III, a single GSS contig matched with many ESTs; Type IV, multiple GSS contigs aligned with multiple ESTs (Table 2).

Most alignments were of Type I (1616) and these provide evidence that a GSS contig contains a gene. A large number (220) of Type II alignments can be further subdivided into two types. Type IIa alignments involve the alignment of multiple GSS contigs to similar positions within the EST. These result likely from the presence of gene families, members of which have been correctly assembled into separate GSS contigs. Type IIb alignments involve non-overlapping (or that do not exhibit sufficient overlap to be joined into a single contig via our assembly parameters) GSS contigs that align with different portions of the EST cluster.

The 134 Type III alignments can also be subdivided further after detailed examination. Around 126 GSSs were

classified as Type IIIa where multiple ESTs are derived from paralogs or are alternatively spliced ESTs from the same genomic sequence. Type IIIb alignments contain multiple ESTs alignments to different regions of the contig. Because these are 3′ ESTs, this result suggests that a contig may contain more than one gene. The eight interesting contigs of this type were BLASTed against the rice genome (http://www.gramene.org, $E = 1e - 30$, total length of alignments $\geq$1500 bp) and six of these are strongly supported by micro-synteny in the rice genome. Since the average length of the eight GSS contigs or singletons involved in these alignments is only 2928 bp (range: 2033–4981 bp), this analysis provides evidence for the existence of 'gene archipelagoes' within the maize genome assembly.

Table 3. Statistics for the ISU MAGIs assembly

Starting data (no. of GSS)	195 233
% masked	25
% GC	44
Contigs	50 002
Average GSS per contig	2.56
No. of clustered clones	49 551
Average number of clones per contig	1.63
Average contig length (bp)	1003
Maximum contig length (bp)	6924
Total % masked post-assembly	28
Singletons	67 302

This assembly can be found at: www.plantgenomics.iastate.edu/maize.

DISCUSSION AND CONCLUSION

We have described a strategy for assembling the maize genome based on innovative parallel algorithms and statistical modeling of important criteria in the development of an assembly pipeline. This strategy requires neither a uniform sampling of the genome nor a definition of repeats by raw occurrences within fragment data. Therefore, it lays the groundwork for efficient assembly of purposefully non-uniform fragments sequenced to enrich for the 'gene-islands' within complex genomes.

A great deal of effort has gone into locating, testing and verifying a repeat database for masking. Inclusion of SDRs has added substantial breadth to this approach. We have also shown that IMMs can be effectively used to flag atypical fragments based on a statistical model of known repetitive sequences.

We have also begun to develop and use novel methods that will help scaffold and annotate these assemblies. Based on the apparent success of the PaCE algorithm in clustering and assembling maize genomic contigs, we are extending this model to multiple types of sequences, including proteins, contigs and ESTs. We also hope protein and EST lookup will generate bridges that span large intron-induced gaps, flag questionable contigs for analysis and provide an efficient means for annotating the assembly. Novel approaches were also developed to model background sequencing error rates to improve assembly parameterization. These are useful in detecting highly similar paralogs (NIPs) within the maize genome by differentiating between cismorphisms and sequencing errors.

Using this pipeline, 730 974 fragments were clustered in <4 h using 64 Pentium III 1.26 GHz processors of a commodity cluster. On this same cluster it would be possible to assemble the resuting clusters with CAP3 in approximately one-half hour. Recent enhancements to PaCE and the repeat database have reduced the time required to cluster over 830 000 sequences to under 2 h.

Table 3 contains information about our current assembly of 730 974 fragments. Although our assembly involved 3.6 times more fragments than the unpublished TIGR AZM 2.0 assembly (www.tigr.org/tdb/tgi/maize/), it consists of only twice as many contigs (91 690 versus 50 002). On the other hand, our assembly generated five times more clustered clones than the TIGR assembly (259 920 versus 49 551) did. This is not simply a case of redundant cloning of the same genomic regions because the average length of our contigs is 35% larger than the TIGR contigs (1,355 versus 1,003 bp). Therefore, as the number of filtered GSS sequences increases so does the overall coverage of the 'gene-rich' fraction of the maize genome.

In summary, we have generated a collection of over 90 000 ISU MAGIs (i.e. contigs) that total over 120 MB; 31 004 of these contigs exhibit a match to rice proteins, ~11 500 to maize proteins and 4008 to yeast proteins. Up-to-date details of our assembly—including a downloadable file containing the contigs and a facility to perform BLAST searches on our contig database—can be found at www.plantgenomics.iastate.edu/maize. We hope this early draft assembly will be of use to the scientific community, and expect that the efficient assembly methods reported here will allow for the quick generation of new drafts as the number of maize genome sequences increases.

ACKNOWLEDGEMENTS

We thank Anantharaman Kalyanaraman for helpful discussions and support regarding PaCE. This research was supported in part by competitive grants from the National Science Foundation to S.A. (award no. ACI-0203782) and to P.S.S, D.A.A and others (award nos DBI-9975868 and DBI-0321711). Support was also provided by Hatch Act and State of Iowa funds. S.J.E. was supported by an NSF-IGERT grant (award no. DGE-9972653).

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arumuganathan,K. and Earle,E.D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Bio. Rep.*, **9**, 208–218.

Batzoglou,S., Jaffe,D.B., Stanley,K., Butler,J., Gnerre,S., Mauceli,E., Berger,B., Mesirov,J.P. and Lander,E.S. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res.*, **12**, 177–189.

Bailey,J., Gu,Z., Clark,R., Reinert,K., Samonte,R., Schwartz,S., Adams,M., Myers,E., Li,P. and Eichler,E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.

Bedell,J.A., Korf,I. and Gish,W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.

Bejerano,G. and Yona,G. (2001) Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, **1**, 23–43.

Bennetzen,J.L. (1996) The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.*, **4**, 347–353.

Bennetzen,J.L., Chandler,V.L. and Schnable,P. (2001) National science foundation-sponsored workshop report. Maize genome sequencing project. *Plant Physiol.*, **127**, 1572–1578.

Cheung,J., Estivill,X., Khaja,R., MacDonald,J.R., Lau,K., Tsui,L.C. and Scherer,S.W. (2003) Genomic-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.*, **4**, R25.

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Huang,X., Wang,J., Aluru,S., Yang,S. and Hillier,L. (2003) PCAP: a whole-genome assembly program. *Genome Res.*, **13**, 2164–2170.

Hurles,M. (2002) Are 100 000 'SNP's' Useless? *Science*, **298**, 1509.

Kalyanaraman,A., Aluru,S., Kothari,S. and Brendel,V. (2003) Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res.*, **31**, 2963–2974.

Lal,S.K., Giroux,M.J., Brendel,V., Vallejos,C.E. and Hannah,L.C. (2003) The maize genome contains a helitron insertion. *Plant Cell*, **15**, 381–391.

Meyers,B.C., Tinge,S.V. and Morgante,M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.*, **11**, 1660–1676.

Mullikin,J.C. and Ning,Z. (2003) The Phusion assembler. *Genome Res.*, **13**, 81–90.

Peterson,D.G., Wessler,S.R. and Paterson,A.H. (2002) Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.*, **18**, 547–50.

Rabinowicz,P.D., Schutz,K., Dedhia,N., Yordan,C., Parnell,L.D., Stein,L., McCombie,W.R. and Martienssen,R.A. (1999). Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.*, **23**, 305–308.

Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.

Volfovsky,N., Haas,B.J. and Salzberg,S.L. (2001) A clustering method for repeat analysis in DNA sequences. *Genome Biol.*, **2**, RESEARCH0027.

Yuan,Y., SanMiguel,P.J. and Bennetzen,J.L. (2003) High-Cot sequence analysis of the maize genome. *Plant J.*, **49**, 249–255.

Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Deng,Y., Dai,L., Zhou,Y., Zhang,X., Cao,M. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.