Assembly and analysis of complex plant genomes

Scott Joseph Emrich
Iowa State University

Retrospective Theses and Dissertations

Follow this and additional works at: http://lib.dr.iastate.edu/rtd
Part of the Bioinformatics Commons, and the Computer Sciences Commons

Recommended Citation
http://lib.dr.iastate.edu/rtd/15533

This Dissertation is brought to you for free and open access by Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Assembly and analysis of complex plant genomes

by

Scott Joseph Emrich

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Srinivas Aluru, Co-major Professor
Patrick S. Schnable, Co-major Professor
David Fernández-Baca
Xiaoqiu Huang
Daniel Nettleton

Iowa State University
Ames, Iowa
2007

Copyright © Scott Joseph Emrich, 2007. All rights reserved.
DEDICATION

To my family, friends, mentors and Viola Miner
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ix</td>
</tr>
<tr>
<td>GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Thesis Organization</td>
<td>2</td>
</tr>
<tr>
<td>Large-scale genome assembly</td>
<td>2</td>
</tr>
<tr>
<td>A unified approach to clustering and assembly</td>
<td>3</td>
</tr>
<tr>
<td>Applications</td>
<td>4</td>
</tr>
<tr>
<td>Maize genome sequencing and assembly</td>
<td>4</td>
</tr>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>Gene-enrichment</td>
<td>5</td>
</tr>
<tr>
<td>Maize Assembled Genomic Islands (MAGIs)</td>
<td>6</td>
</tr>
<tr>
<td>Assembly validation and analysis</td>
<td>8</td>
</tr>
<tr>
<td>Large-scale experimental validation of MAGIs</td>
<td>9</td>
</tr>
<tr>
<td>Maize gene number</td>
<td>9</td>
</tr>
<tr>
<td>Massively parallel sequencing</td>
<td>10</td>
</tr>
<tr>
<td>LCM-454 gene discovery</td>
<td>10</td>
</tr>
<tr>
<td>LCM-454 SNP discovery</td>
<td>11</td>
</tr>
<tr>
<td>References</td>
<td>12</td>
</tr>
<tr>
<td>A STRATEGY FOR ASSEMBLING THE MAIZE</td>
<td>14</td>
</tr>
<tr>
<td>(ZEA MAYS L.) GENOME</td>
<td>14</td>
</tr>
<tr>
<td>Abstract</td>
<td>14</td>
</tr>
<tr>
<td>Introduction</td>
<td>15</td>
</tr>
<tr>
<td>Methods</td>
<td>16</td>
</tr>
</tbody>
</table>
Input data, masking of low quality and contaminant sequences .................. 16
Determining and masking repetitive elements ........................................ 16
IMM-based sequence classification ......................................................... 20
Empirical determination of sampling biases and sequencing error rates ........ 21
Assembly of non-uniform genomic fragments ........................................... 22
Assembly and verification of PaCE clusters .............................................. 24
Near-Identical Paralogs ........................................................................... 25
Sequence-based scaffolding of “gene archipelagos” ................................... 27
Discussion and Conclusion ....................................................................... 29
Acknowledgments .................................................................................... 31
References ............................................................................................... 31

ASSEMBLING GENOMES ON LARGE-SCALE

PARALLEL COMPUTERS ................................................................. 34
Abstract ................................................................................................. 34
Introduction ............................................................................................ 35
Genome Assembly .................................................................................... 38
Our Clustering-based Parallel Framework for Genome Assembly ................ 40
Clustering Fragments .............................................................................. 42
Generating Promising Pairs ..................................................................... 44
Parallel Generalized Suffix Tree Construction ......................................... 49
Experimental Results ............................................................................. 51
Detecting Overlaps And Managing Clusters In Parallel ............................. 51
Run-time and space complexity ............................................................... 58
Experimental Results ............................................................................. 59
Maize Genome Assembly ........................................................................ 60
Whole Genome Shotgun and Environmental Assemblies .......................... 63
Assembly of the Drosophila Genome ....................................................... 63
Assembly of Environmental Genome Sequences ..................................... 64
# NEARLY IDENTICAL PARALOGS: IMPLICATIONS FOR MAIZE (*Zea mays* L.) GENOME EVOLUTION

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>93</td>
</tr>
<tr>
<td>Introduction</td>
<td>94</td>
</tr>
<tr>
<td>Methods</td>
<td>96</td>
</tr>
<tr>
<td>Locating and validating NIPs in collections of maize ESTs and GSSs</td>
<td>96</td>
</tr>
<tr>
<td>Annotation of NIPs:</td>
<td>97</td>
</tr>
<tr>
<td>NIP expression assays</td>
<td>97</td>
</tr>
<tr>
<td>Genetic mapping of NIPs</td>
<td>97</td>
</tr>
<tr>
<td>Locating NIPs within Arabidopsis</td>
<td>98</td>
</tr>
<tr>
<td>Results</td>
<td>99</td>
</tr>
<tr>
<td><em>In silico</em> detection of maize NIPs</td>
<td>99</td>
</tr>
<tr>
<td>Experimental validation of EST-based CP sites</td>
<td>99</td>
</tr>
<tr>
<td>NIPs discovered within a partial maize genome assembly</td>
<td>101</td>
</tr>
<tr>
<td>Expression of NIPs</td>
<td>101</td>
</tr>
<tr>
<td>Frequency of NIPs</td>
<td>104</td>
</tr>
<tr>
<td>Both members of many NIP families are expressed</td>
<td>104</td>
</tr>
<tr>
<td>Members of many NIP families exhibit differential expression</td>
<td>106</td>
</tr>
<tr>
<td>Genomic organization of maize NIPs</td>
<td>106</td>
</tr>
<tr>
<td>Arabidopsis NIPs</td>
<td>107</td>
</tr>
<tr>
<td>Discussion</td>
<td>107</td>
</tr>
<tr>
<td>The maize genome contains a high frequency of NIPs</td>
<td>107</td>
</tr>
<tr>
<td>Origins of NIPs</td>
<td>108</td>
</tr>
<tr>
<td>Why does maize have more NIPs than Arabidopsis?</td>
<td>109</td>
</tr>
<tr>
<td>NIPs and genetic markers</td>
<td>112</td>
</tr>
<tr>
<td>Evolutionary implications of NIPs</td>
<td>112</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>113</td>
</tr>
<tr>
<td>References</td>
<td>113</td>
</tr>
</tbody>
</table>
## GENE DISCOVERY AND ANNOTATION USING LCM-454 TRANSCRIPTOME SEQUENCING

Abstract ............................................................................................................................... 120
Introduction ........................................................................................................................... 121
Methods ................................................................................................................................ 122
  Isolation of SAM mRNA .......................................................................................................... 122
  454-EST sequencing and processing ...................................................................................... 122
  Comparisons of 454-ESTs to public sequence databases ...................................................... 123
  Annotation using 454-EST sequences ................................................................................... 124
  Validation of expression of orphan genes ............................................................................. 125
  Estimating the rate of sequencing errors in 454-ESTs ......................................................... 125
  Estimating the 3'-enrichment of 454-ESTs .......................................................................... 125
Results .................................................................................................................................... 126
  Gene discovery and annotation using 454 sequencing .......................................................... 126
  Validation of orphan expression ............................................................................................ 128
Discussion .............................................................................................................................. 130
Acknowledgments .................................................................................................................. 133
References .............................................................................................................................. 133

## SNP DISCOVERY VIA 454 TRANSCRIPTOME SEQUENCING

Abstract ................................................................................................................................. 136
Introduction ............................................................................................................................ 137
Methods .................................................................................................................................. 139
  Isolation of SAM mRNA and 454 sequencing ...................................................................... 139
  Identification of B73 reference sequences for 454 ESTs ....................................................... 139
  Multiple sequence alignments and SNP detection of 454 sequence data ............................ 140
  SNP parsing .......................................................................................................................... 141
Results ..................................................................................................................................... 141
  Assignment of Mo17 and B73 SAM ESTs to maize genomic anchor sequences .................... 142
ACKNOWLEDGMENTS

First and foremost, I would like to thank my scientific mentors in chronological order. I have been told that my first job was taking care of rats as a lab assistant, and I am sure by doing so my mother and her colleagues instilled biology in me at a young age. I especially thank Drs. Jacqueline Emrich, Joseph Rubertone, Hans Bender, Reiner Class, Christopher Hand and George Tusynski at Hahnemann/Drexel University for letting me play with pipets and tissue cultures more than they probably should have.

At Loyola College, I was very lucky to have excellent undergrad research advisers who provided substantial training in interdisciplinary science. Drs. Mary Lowe (Physics) and Arthur Delcher (Computer Science) carefully helped me acquire both experimental and computational expertise when I had nearly none. I hope to someday be as thorough as Dr. Lowe in analyzing data and as innovative as Dr. Delcher in developing algorithms.

I came to Iowa State with an interdisciplinary thesis in mind, and without the guidance and inspiration of my graduate advisers, Drs. Srinivas Aluru and Patrick Schnable, this work would not have been possible. Both have the ability to think clearly and critically in their respective fields, and were confident enough in a potential collaboration to let a young PhD student play with some maize genomic data. Individually, Dr. Aluru has spent substantial time helping me think about computational problems and has prepared me well for faculty life. Dr. Schnable has spent at least an hour a week helping me think about data more critically, and by doing so has pushed me to become a better computational biologist. In addition, I attribute my improvements in scientific writing to his tutelage including substantial advice provided on all six papers included in this thesis. Both have provided excellent opportunities during my time at ISU and have given me a great foundation to my research career.
Second, I would like to thank my coauthors. In the beginning, there was the original Maize Assembled Genomic Island (MAGI) team that met during the summer of 2003: Dr. Srinivas Aluru, Dr. Daniel Ashlock, Dr. Patrick Schnable, Dr. Tsui-Jung (TJ) Wen, Dr. Yan Fu, Dr. Ling Guo, and Mahesh Narayanan. Subsequent collaborations with TJ and Yan on NIPs and MAGI3.1, respectively, were both fun and scientifically rewarding. I’d like to thank Dr. Ananth Kalyanaraman for several interesting collaborations including porting the MAGI framework to the IBM BlueGene/L supercomputer and Dr. Brad Barbazuk (Danforth Plant Science Center) for excellent discussions and our collaboration on using 454 sequencing to address fundamental problems in plant genomics. Li Li has been an excellent collaborator and has done outstanding experimental work on our joint projects.

Life at Iowa State would not be the same without my colleagues, and I have been fortunate to have been a member of both the Aluru and Schnable labs. My primary lab was in Coover Hall, which I have shared with the following Alurians: Chad Brewbaker, Benjamin Jackson, Ananth Kalyanaraman, Pang Ko, Srikanth Komarina, Sarah Orley, Yogy Namara, Olga Nikolova, Stepan Rajko, Abinav Sarje, Sudip Seal, Xiao Yang and Dr. Jaroslaw Zola. I especially thank Sudip, Pang, Ananth, Ben and Sarah who have been here nearly as long as I and for countless good times over those years. The members of the Schnable lab have accepted me as their own, and I have had excellent discussions and collaborations with many. I especially thank Dr. Kazihiro Ohtsu, Debbie Chen, Ruth Swanson, Jun Cao, Yi Jia, Karthik Viswanathan, Eddy Yeh and Sarah Hargreaves.

This research was initially supported by a National Science Foundation Integrative Graduate Education Training Grant (NSF-IGERT award number DGE-9972653). More recently, I have been supported by the NSF/DOE/USDA Maize Genome Sequencing project (DBI-0527192). Much of the experimental work in this thesis has been funded by the National Science Foundation, ISUs Plant Science Institute, and the Donald Danforth Plant Science Center; additional support was provided by Hatch Act and State of Iowa funds.
GENERAL INTRODUCTION

Each organism has its own “cookbook” called a genome that contains recipes called genes whose expression into RNA and proteins is required for cellular processes and phenotypic characteristics. Deciphering complete genomes is an important goal of many biological communities. Because of current experimental limitations, biochemical procedures collectively known as sequencing are only capable of accurately determining only hundreds of bases per reaction (≈500–1000 nucleotides). To extend the reach of sequencing to entire genomes, numerous short fragments are sequenced from randomly distributed locations of a larger molecule. These fragments are then combined to form the original sequence through a computational process called assembly.

Concurrent advances in high-throughput sequencing and assembly techniques have led to the successful completion of increasingly larger and more complex genomes. Even so, assembling the tens of millions of fragments typical of a large genome project still places enormous demands on computational resources. For example, Celera Genomics reported that it took 20,000 CPU hours to assemble 27.27 million human genome fragments [Venter et al. (2001)]. To accommodate these requirements, most assembly projects have been carried out by specialized teams running software developed for serial assembly on high-end workstations with tens of gigabytes of main memory and rudimentary parallelization approaches. Further, only a few draft genome sequences can be produced per project, of which a final build is made publicly available months to years after initial sequencing.

The large runtime and memory requirements of genome assembly make it an ideal problem for parallel processing. In fact, many important problems in computational genomics can substantially benefit from high performance computing including but not limited to large-
scale gene discovery, novel sequencing technologies, high-throughput genome annotation and comparative genomics. Prior to work presented in this thesis, however, previous solutions achieved at most a modest degree of parallelism.

Thesis Organization

Here, we present a total of six papers published or in print that can be broken into three distinct parts: (1) a method for large-scale genome assembly [Emrich et al. (2004); Kalyanaraman et al. (2007)]; (2) maize genome analysis [Fu et al. (2005); Emrich et al. (2007a)]; and (3) preliminary results of applying massively parallel pyrosequencing to important problems in plant genomics [Emrich et al. (2007b); Barbazuk et al. (2007)]. All have or can benefit from high performance computational biology. In this chapter, we introduce each component and provide a general overview of the contributions contained in these manuscripts.

Large-scale genome assembly

Most assembly algorithms piece together a genome using information preserved between fragments derived from overlapping regions. While finding these overlaps is compute-intensive, overlaps are also often stored in memory or on disk during assembly. When using traditional sequencing approaches, which sample the genome uniformly, the number of overlaps is expected to be proportional to the number of sequenced fragments.

The initial work presented in this thesis was motivated by a changing scenario where alternative strategies [Rabinowicz et al. (1999); Yuan et al. (2003)] were used to non-uniformly sequence gene-rich regions of the maize genome, which imposed a much higher storage requirement because these overlaps scaled quadratically. In this section, we summarize our primary contribution: an approach for large-scale assembly that adequately addresses the computational challenges posed by uniformly and non-uniformly sampled sequence data. Specifically, we describe the theory and adaptation of an efficient parallel clustering algorithm to the assembly of maize genome sequence data, whole genome shotgun data, and metagenomic data obtained from environmental samples.
A unified approach to clustering and assembly

Many assemblers follow the overlap-layout-consensus paradigm, in which pairs of overlapping fragments are used to build long contiguous stretches of the genome called contigs. Under uniform random sampling it is expected that each nucleotide of the genome is present in a constant number of fragments. As a result, the number of valid overlapping fragments in any sequencing project is linear, assuming there is a way to identify them. Now suppose that we can no longer assume a uniform sample. In this case we may have to test all pairs of fragments. For example, consider a large jigsaw puzzle mostly of a prairie; this task will be very labor-intensive because of the substantial number of viable “green” piece combinations.

Just like assembly, clustering algorithms determine sequence relatives based upon overlaps. The difference between these two problems is illustrated in the following example. During clustering, if A overlaps with B, and B overlaps with C, we can deduce that A is related to C without having to compare A and C. Assemblers, on the other hand, must directly determine if A consistently overlaps with C, which is required for accurate genome reconstruction. Intuitively, it follows that a solution to clustering is easier to compute than its assembly counterpart — as it is typically easier to group pieces than to assemble them correctly.

In this thesis, we present a unified view of these two problems using clustering as an initial preprocessing step for assembly. The rationale of this approach is as follows: sequencing approaches often do not capture everything and, as a result, disjoint contigs are produced by assembly algorithms. For example, the selective sequencing of the maize genome was hypothesized to produce a large number of genomic “islands” that contained most of the genic sequences. Sequence clustering provided a method for decomposing this large maize assembly problem into many, but smaller, assembly problems [Emrich et al. (2004)], each of which corresponded to single island or contig. In general, this cluster-then-assemble approach shifts the computational burdens of large-scale sequence analysis to the clustering phase, while benefiting from and not duplicating the painstakingly built-in biological expertise present in conventional bioinformatic tools such as assemblers.

There are two additional advantages to this paradigm. First, clustering limits the sub-
sequent peak memory usage to the memory required to post-process the largest subproblem. Second, breaking the problem into clusters allows trivial parallelization; each cluster can be individually processed on a different processor. When applied to the problem of sequence assembly, both of these properties facilitate the generation of assemblies that are consistent with conventional assembly program results, except that the maximum solvable problem size is larger and overall speed is significantly enhanced. In order for this approach to be valid, however, any overlap considered significant by the assembler should also be considered acceptable by the clustering algorithm; therefore, in practice the overlap criteria are less stringent than those used during assembly to ensure consistency.

Applications

This strategy has been shown to be applicable in conventional genome assemblies [Kalyanaraman et al. (2007)], the explanation of which is as follows. The coverage of sequencing—typically between five and seven—can be plugged into the Lander-Waterman equation [Lander and Waterman (1988)] to determine the expected number of clusters that result from random sampling. A real-life example is the Celera human genome assembly that resulted in 221,036 contigs each spanning ∼11.7 Kbp on an average and the longest contig spanning ∼1.2 million bp (i.e., only 0.48% of the longest chromosome) [Venter et al. (2001)]. Even assembling a small genome like that of N. meningitis (∼2.18 Mbp) generated 149 contigs with an average length of 14 Kbp [Pevzner et al. (2001)]. If the sequences that compose each contig can be accurately partitioned via clustering, each can be processed independently in parallel.

Maize genome sequencing and assembly

Introduction

Over the past few thousand years, the domestication and spread of maize throughout the Americas has led to an immense source of phenotypic variation including kernel color, cob size and other important nutritional traits such as sugar content. It was once believed that collinearity, or the preservation of the positions of genes in related species, within the
cereal crops (e.g., rice, wheat, barley and maize) would facilitate the discovery of economically
important genes faster than traditional approaches. Therefore, draft sequences of the much
smaller rice genome (430 million bases) were completed and international sequencing projects
were begun. Based on incoming sequence data from smaller intervals of these species, however,
biologists now believe that the genomic and evolutionary differences between maize and rice
— and even between multiple maize subspecies [Brunner et al. (2005)] — are unique and
interesting enough to warrant genome sequencing. The inbred line B73, which was developed
at Iowa State University and is the genetic ancestor of commercially important lines, was the
initial choice for sequencing.

**Gene-enrichment**

Because 65-80% of the maize genome consists of large, highly homogenous retrotransposons,
many attendees at an international meeting convened in St. Louis during 2001 were concerned
that it would not be possible to assemble the maize genome using a shotgun-based sequencing
approach [Bennetzen et al. (2001)]. 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.

There are two primary genome reduction sequencing strategies that have been successfully
used for maize and are now being applied to other plants including wheat, pine and sorghum.
The first strategy, *Methyl Filtration* (MF) [Rabinowicz et al. (1999)], discards the portions
of the genome that are methylated. The second strategy, *High C₀t* sequencing (HC) [Yuan et al.
(2003)], utilizes hybridization kinetics to isolate low-copy fractions of a genome. Each of these
techniques is explained in detail below.

Methylation occurs at certain nucleotides and is important in multiple biological processes
including gene silencing, or turning off transcription, of certain genomic regions. In particular,
it has been shown that retrotransposons and other repetitive sequences in plants tend to be
predominantly methylated. By sequencing the unmethylated regions, the sampled sequences
should mostly originate from gene-rich stretches of the genome. The interesting aspect of this
sequencing approach is that it only requires a special strain of *E. coli* that can recognize only a single methyl group per fragment; the rest of the sequencing protocol is similar to traditional genome sequencing.

The HC sequencing approach is somewhat more complex than MF because it relies on biochemical instead of biological filtration. Repetitive sequences will hybridize more often in a heterogeneous mixture of single-stranded genomic DNA fragments; consequently, removing double-stranded sequences after some elapsed time enriches for lower-copy sequences. Consider, as an example, a bag of marbles. Suppose this bag has more red marbles than blue ones, say by a ratio of 9:1 and our game consists of reaching into the bag and pulling out two marbles at random. If the colors match, we discard them; otherwise, we place the marbles back in the bag. It should be clear that in the beginning of this exercise we will remove many more red pairs (81% chance) than blue pairs (1% chance). It follows that our original population can be normalized from a ratio of 9:1 to 1:1 over time. Even if we stop this exercise earlier, we still enrich for underrepresented objects because overrepresented objects are preferentially removed. Because HC selection may remove non-repetitive sequences as the solution approaches equilibrium, multiple time intervals are processed to maximize low-copy sequence recovery and minimize the loss of highly similar genic sequences (e.g., gene families).

**Maize Assembled Genomic Islands (MAGIs)**

To assemble the maize genome it was necessary to develop a scalable solution that employed mechanisms to minimize assembly artifacts due to the presence of repetitive elements and that also accounted for the non-uniform sampling of the genome due to gene-enrichment. Traditional assembly programs were deemed inadequate by us because they were optimized for uniform sampling of the genome and as a result did not work well for maize. Moreover, they could not inherently differentiate near-identical paralogs (NIPs) that arose via segmental duplications and other mechanisms [Emrich et al. (2004)].

The overarching goal of our early maize assembly work was to produce an accurate assembly as quickly as new data became available. In our pipeline sequences are cleaned, repeat
masked and clustered based on defined overlap criteria. The sequences within clusters are then unmasked and assembled using CAP3 [Huang and Madan (1999)] into one or more contigs. Relative to our initial maize genome assembly [Emrich et al. (2004)], current assemblies incorporate further improvements in the quality of the input sequences, the repeat masking process, and the use of clone pair information [Fu et al. (2005); Kalyanaraman et al. (2007)].

Even though gene enrichment selectively samples disparate regions of the genome, the highly repetitive nature of the maize genome may lead to excessive merges of unrelated regions based on common repeats. To solve this problem, a modified transitive closure clustering algorithm was used in order to locate Statistically-defined repeats, or SDRs, that could later be used to mask the repetitive sequences and thus eliminate the formation of large clusters [Emrich et al. (2004)]. SDRs were a significant contribution to maize genome assembly and analysis and without them the cluster-then-assemble approach works poorly as a result of numerous retrotransposons in the maize genome.

Because both previous datasets were relatively small, their memory requirements both fit into the 64 GB available on a Pentium III cluster in our laboratory. To meet the estimated memory requirement of ~100 GB on the entire maize collection and to accelerate the clustering process, a total of 1,024 dual-processor nodes of the IBM BlueGene/L supercomputer — each with 512 MB RAM and 700 MHz CPUs — were used. This allowed the clustering to complete in 2 hours. In comparison with the initial MAGI assembly, this clustering computed almost ten times larger number of alignments in just the same parallel run-time, using 16 times as many, but less powerful processors.

The most recent assembly based upon the BlueGene/L clustering completed in under 8.5 hours on 40 processors of the Pentium III cluster and resulted in a total of 163,390 maize genomic islands (or contigs) formed by two or more input sequences, and 536,377 singletons (i.e., sequences that did not assemble with any other sequence). On an average, each cluster assembled into 1.1 contigs; given that the CAP3 assembly is performed with a higher stringency, this result indicates the high specificity of our clustering method and its utility in breaking the large assembly problem into disjoint pieces of easily-manageable sizes.
The Maize Genome Project

The U.S. NSF, DOE and USDA are currently supporting large-scale sequencing of the maize genome. Unfortunately, the predominance of repeats within this genome make it unlikely that a cheaper shotgun sequencing approach alone will be able to accurately reconstruct the maize genome given current experimental and computational limitations, many of which were uncovered in our pioneering work on maize genome assembly and analysis. Because the highly similar repeats are expected to be scattered throughout the maize genome, however, breaking the genome into smaller chunks (e.g., Bacterial Artificial Chromosomes, or BACS) was hypothesized to reduce their overall effect on maize genome assembly. As such, an incremental sequencing approach could increase assembly accuracy while reducing the computational work required by performing thousands of smaller “mini” genome assemblies.

The stated goal of the NSF, DOE and USDA joint effort is explicit in its special emphasis on identifying and locating all genes and their associated regulatory regions along the genome. Note that unlike previous sequencing projects, such as the human genome project, the emphasis has shifted from knowing the complete genome to a large collection of genomic contigs whose order and orientation along maize chromosomes is known. Although not explicitly included in this thesis, many of the experiments used to justify the current sequencing approach to these funding agencies were performed at Iowa State University using knowledge, techniques and resources (including our cereal repeat database) developed in this thesis. As such, our work has had a substantial impact on final result of the B73 maize genome sequencing effort, expected to be finished in early 2008, and continually informs other plant sequencing projects as our high performance computing ideas are being applied to future projects including sorghum, soybean and Brachypodium, the latter of which is important in cellulosic ethanol production.

Assembly validation and analysis

The interdisciplinary work presented in this thesis was a product of a collaboration between computational and plant scientists centered around the ISU MAGI genome assembly. In addition to contributions in plant genome assembly, we have been using these results to enable
biological discovery in plants. In this section we describe computational methods developed to decipher unique features of the maize genome, many of which have significant evolutionary implications.

**Large-scale experimental validation of MAGIs**

Interestingly, a substantial number of predicted genes in the MAGI assemblies were not supported by early maize expression data. Further, many of these predicted genes were “orphans”, i.e., there was no available orthologous sequence present in GenBank.

To test the validity of these computational observations, a large-scale gene discovery project was undertaken. Concomitantly, the growing experimental data were used to validate the biological validity of the MAGI assembly framework. Two techniques were devised: alignments to known BAC sequences and a novel approach that subdivided primers into control and experimental groups to test whether the rate of experimental success was significantly lower in regions supported only by assembly-derived decisions. Based on this computational analysis, we determined that a large number of our MAGI assemblies accurately reflected the structure of the maize genome.

**Maize gene number**

The estimated number of maize genes prior to the availability of MAGIs was on the order of 50,000-60,000, based primarily on the nearly completed shotgun assembly of the rice genome. This subject was interesting from multiple evolutionary perspectives. If maize did have more genes than rice, was that a potential reason for the vast phenotypic and adaptive variability exhibited by modern maize? Did duplication have a role in these processes?

Computational experiments were devised and performed on the MAGI3.1 assembly to address these questions. The first of these experiments established accurate lower and upper bounds on the number of expressed genes in the maize genome and for the first time established that maize likely had more genes than rice. The latter of these experiments used the assembly to locate approximately 300 putative recent duplications, many of which were
tested experimentally. We concluded that approximately 1% of all maize genes were recently duplicated, and more importantly many of these exhibited differential expression in multiple tissues. In addition, this observation supported an interesting mechanism where maize could carry more alleles than the typical diploid genome and could therefore substantially influence the adaptability of maize. Large-scale gene duplication also increases the recovery of rare, yet beneficial, alleles. As such, this work substantially impacted understanding and analysis of the evolution of modern maize.

**Massively parallel sequencing**

One of the interesting open questions in computational genomics is how to utilize sequences that are inexpensive to generate, yet currently ill-suited for complex genome assembly. A potential solution, considered by us and others, is to filter repetitive regions from consideration using previously mentioned gene-enrichment techniques.

One of the oldest enrichment techniques is Expressed Sequence Tag (EST) sequencing where mRNA transcribed in cells is converted into cDNA, cloned, and sequenced. Recent advances in this technology used isolation technology such as Laser Capture Microdissection (LCM; reviewed by Schnable et al. (2004)) to enrich for transcripts present in specific tissues. Even so, these approaches were labor intensive and as a result expensive.

We, on the other hand, hypothesized that the technology commercialized by 454 Life Sciences [Margulies et al. (2005)] could be used to inexpensively directly sequence cDNA without cloning and its associated costs. Further, we postulated that transcripts present in the developmentally important shoot-apical meristem (SAM) of maize could be interesting given all above ground tissue in this plant derive from these omnipotent cells. A single sequencing run was first performed on the inbred line B73.

**LCM-454 gene discovery**

To determine its effectiveness, we also decided to ask the simple question: “How effective was LCM-454 in sequencing maize genes?” Maize LCM-454 ESTs were compared to multiple
sequencing collections including a collection of Apex ESTs, which were enriched for SAM transcripts, and all available maize EST sequences in the end of 2005. Surprisingly, over 70% of these ESTs were not captured by the previous Apex EST sequencing project and nearly 30% were not captured by extensive EST sequencing of multiple maize varieties. To address the concern that these were a result of sampling the middle regions of genes previously captured, we devised an experiment that showed that the LCM-454 sequencing performed was substantially 3’ enriched. Given that many maize ESTs were sequenced from the 3’ end, we concluded that overestimation of novelty was minimal.

A total of 27 “orphan” genes were experimentally validated, many of which were indetectible in other tissues including meristem rich-immature ears. We concluded that LCM-454 could recover rare, potentially tissue-specific genes.

**LCM-454 SNP discovery**

We have shown that LCM-454 is also an efficient strategy for the resequencing of maize genes expressed in additional inbred lines. A single sequencing run was performed on the inbred Mo17, which along with B73 form the primary parental lines of most commercial corn in the U.S. and the basis of the maize communities genetic map. Using a similar number of ESTs, we were able to predict over 36,000 putative SNPs among almost 10,000 MAGIs. These were stringently post-processed by Dr. Brad Barbazuk at the Danforth Plant Science center to obtain over 7,000 high-confidence maize SNPs. Based on experimental validation, we conservatively estimated that we have found at least 4,900 SNPs distributed amongst over 2,400 maize genes.

Of greater importance, we devised a computational strategy that showed not only the ability to obtain SNPs using LCM-454 EST data, but such SNPs can be accurately converted into genetic markers that will be useful in marker assisted selection (MAS) programs, quantitative genetic studies and to enhance our understanding of genome organization and function. In addition, SNPs can be used for genome-wide association studies that assign genes to specific functions or traits. These are of great importance not only for maize and cereal biology, but
also in the numerous orphan crops where current funding levels do not support large-scale sequencing efforts.

References


A STRATEGY FOR ASSEMBLING THE MAIZE
(ZEA MAYS L.) GENOME

A paper published in Bioinformatics\(^1\)

Scott J. Emrich, Srinivas Aluru, Yan Fu, Tsui-Jung Wen, Mahesh Narayanan,
Ling Guo, Daniel A. Ashlock and Patrick S. Schnable

Abstract

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. Here, 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 GSS fragments in 4 h using 64 Pentium III 1.26 GHz processors of a commodity cluster. Algorithmic innovations are used to significantly reduce the number of pairwise alignments 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.

\(^1\)Reprinted with permission of Bioinformatics, 2004, 20:140–147.
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 2,500 million bases (MB) [Arumuganathan and Earle (1991)] makes it comparable in size to 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_0t (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 established that the maize genome contains many NIPs [Emrich et al. (2007)] that can be identified because they contain one or more cismorphisms [Hurles (1998)], 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 scalability as the number of sequences increases. Our overarching goal is to produce an accurate assembly as quickly as new data become available. Our results therefore explore the peculiarities of current data and techniques, analytical methods for assessing errors and an assembly pipeline (Figure 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: *E. coli* K12 (U00096), bacteriophage phiX174 (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 done during
Figure 1  Maize assembly pipeline
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 1,667 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 EST hits by homology search (identity ≥98%, overlap ≥50bp). 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 twenty 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 Version 2.0 (http://www.tigr.org), the Wessler Laboratory’s database of plant transposable elements [N. Jiang, University of Georgia (personal communication)], maize SDRs and atypical repetitive sequences from previous assemblies were combined into a 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 8,595 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 thirty bases with an associated minimum E-value of $5e^{-4}$ are masked, along with any hit with 80% identity over 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 1,036 $Z.mays$ 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 1,100 bases per cDNA. Twenty cDNAs were partially masked; 10 of these match an 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
Table 1  Repeat masking of the three major sources of GSSs. Percentage of bases masked in each sequence type.

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>ISU MAGIs&lt;sup&gt;a&lt;/sup&gt; 2.3</th>
<th>TIGR AZMs&lt;sup&gt;b&lt;/sup&gt; 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>High C&lt;sub&gt;0&lt;/sub&gt;t</td>
<td>5.9</td>
<td>19</td>
</tr>
<tr>
<td>Hypomethylated</td>
<td>18.8</td>
<td>31</td>
</tr>
<tr>
<td>BAC Ends</td>
<td>57.6</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>19.6</td>
<td>25</td>
</tr>
</tbody>
</table>

<sup>a</sup>MAGI, Maize Assembled Genomic Island
<sup>b</sup>AZM, Assembled Zea mays

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 to the work of Bejerano and Yona (2001) in successfully classifying protein families. The use of a likelihood ratio, however, allows the assignment of a <i>p</i>-value for each genomic fragment.

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<sup>−38</sup>), gypsy-type retrotransposon (AF466203,
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. (2005)]. 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 over sample non-exonic sequences, *i.e.* 5' UTRs, introns and 3' UTRs, with stronger bias towards UTRs. For one-third of the eighty 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 by 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 empirically estimate 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 3 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. (2004)].
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 require 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 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 1GB of RAM. Instead, we used the parallel EST clustering tool [Kalyanaraman et al. (2003)], Parallel Clustering of ESTs (PaCE), to significantly reduce the problems inherent in non-uniform samples and established a pipeline that quickly assembles maize contigs. Clustering genomic fragments provides a method of 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 HIC 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 generalized suffix tree (GST) constructed from all of the sequences and their Watson-Crick complements. Therefore, PaCE never generates all of
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 of 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 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 been 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
Figure 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. Grey pairs are unaligned, and represent a significant reduction in work.

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 runtime with clustering 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 5 and a maximum cluster composed of only ninety-six sequences. Unmasked fragments are used for assembly with the following CAP3 pa-
rameters: 98% identity, 80bp overlap, 60bp 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 to a lower threshold. If, in a future implementation, all the 64 processors are used, assembly should take only 1/64th as much time, \( i.e., <30 \text{ min} \).

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 SNPs are potentially paralogous sequence variants (\( i.e., \) cis-morphisms, [Hurles (1998)]) 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 [Emrich et al. (2007)]. 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 i.i.d (independent and identically distributed). 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 a 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, 1,108 contigs contained at least one statistically significant CP. The trace files of a sample of these were aligned using Sequencher 4.1 (see Figure 3 for an example). 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 a human expert to review suspicious clusters.
Table 2 The numbers of 2,030 GSS contigs involved in each of four types of spliced alignment to EST contigs described within the text.

<table>
<thead>
<tr>
<th>GSS Contig EST = 1</th>
<th>EST &gt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>= 1</td>
<td>1616</td>
</tr>
<tr>
<td>&gt; 1</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td>1836</td>
</tr>
</tbody>
</table>

Sequence-based scaffolding of “gene archipelagos”

Maize “gene islands” will eventually be linked into “gene archipelagos” for finishing and further analysis. We propose a computational approach, which we feel is an effective and useful tool for clustering gene islands into archipelagos 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 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 $1e^{-10}$. Of the 91,690 maize GSS contigs, 4,008 met this criterion with matches to 1,145 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 runtime for clustering in this manner a PaCE-like approach was taken. A
The generalized suffix 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 6 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 to an ALL vs. 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 [Emrich et al. (2007)] and this EST dataset contains 6,270 singletons and 3,751 contigs (i.e., 10,021 unique genes).

A GeneSeqer alignment of a 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 2,030 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 many ESTs; Type IV, multiple GSS contigs aligned with multiple ESTs (Table 2).

Most alignments were of Type I (1,616) 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 likely result 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 to different portions of the EST cluster.

The 134 Type III alignments can also be further subdivided after detailed examination. 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 8 interesting contigs of this type were BLASTed against the rice genome (http://www.gramene.org, \( E=1e^{-30} \), total length of alignments \( \geq 1,500 \) 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 2,928 bp (range: 2,033–4,981bp), this analysis provides evidence for the existence of “gene archipelagos” within the maize genome assembly.

**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 based on 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 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 \(<\ 4\) hours using 64 Pentium III 1.26
Table 3  Statistics for the ISU MAGI 2.3 assembly

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting data, no. of GSSs</td>
<td>730,974</td>
</tr>
<tr>
<td>Input masked, %</td>
<td>19.6</td>
</tr>
<tr>
<td>No. of contigs</td>
<td>91,690</td>
</tr>
<tr>
<td>No. clustered clones</td>
<td>259,920</td>
</tr>
<tr>
<td>Average GSSs per contig</td>
<td>4.12</td>
</tr>
<tr>
<td>Average clones per contig</td>
<td>2.83</td>
</tr>
<tr>
<td>Contig % GC</td>
<td>44.5</td>
</tr>
<tr>
<td>Average contig length, bp</td>
<td>1,355</td>
</tr>
<tr>
<td>Maximum contig length, bp</td>
<td>8,489</td>
</tr>
<tr>
<td>No. of singletons</td>
<td>353,558</td>
</tr>
</tbody>
</table>

GHz processors of a commodity cluster. On this same cluster it would be possible to assemble the resulting 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 (http://www.tigr.org/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 sequences increases so does the coverage of the “gene-rich” portion 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 4,008 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.
Acknowledgments

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 number: ACI-0203782) and to P.S.S, D.A.A and others (award numbers: 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 number DGE-9972653).

References


ASSEMBLING GENOMES ON LARGE-SCALE PARALLEL COMPUTERS

A paper accepted by the Journal of Parallel and Distributed Computing

Ananth Kalyanaraman1, Scott J. Emrich2, Patrick S. Schnable, Srinivas Aluru

Abstract

Assembly of large genomes from tens of millions of short genomic fragments is computationally demanding requiring hundreds of gigabytes of memory and tens of thousands of CPU hours. The advent of high throughput sequencing technologies, new gene-enrichment sequencing strategies, and collective sequencing of environmental samples further exacerbate this situation. In this paper, we present the first massively parallel genome assembly framework. The unique features of our approach include space-efficient and on-demand algorithms that consume only linear space, and strategies to reduce the number of expensive pairwise sequence alignments while maintaining assembly quality. Developed as part of the ongoing efforts in maize genome sequencing, we applied our assembly framework to genomic data containing a mixture of gene enriched and random shotgun sequences. We report the partitioning of more than 1.6 million fragments of over 1.25 billion nucleotides total size into genomic islands in 2 hours on 1,024 processors of an IBM BlueGene/L supercomputer. We also demonstrate the effectiveness of the proposed approach for traditional whole genome shotgun sequencing and assembly of environmental sequences.

1Primary author; developed the software PaCE and ported it to BlueGene/L
2Primary author; developed the genome assembly application of PaCE
Introduction

Each cell in a living organism contains one or more long DNA sequences called chromosomes, collectively known as the genome. Contained within the genome are DNA sequences called genes that code for proteins and RNA molecules, which perform various cellular functions in an organism. Deciphering an entire genome sequence and identifying regions within it that are genes and regulatory elements is of fundamental importance in molecular and functional genomics. Genome sequencing also forms the basis for the rapidly expanding field of comparative genomics, which attempts to study genome evolution and unravel genome structure through cross-genome comparisons.

Genomes span multiple length scales—from a few tens of thousands of nucleotides in viruses to millions of nucleotides in microbes to billions of nucleotides in complex eukaryotic organisms such as plants and animals. Because DNA is double stranded, its length is measured in units called base pairs, denoted bp. The biochemical procedure of determining the nucleotide sequence of a DNA molecule is called sequencing. Accurate sequencing is experimentally viable only up to hundreds of base pairs (≈ 500–1,000 bp). To extend the reach of sequencing to genomic scales, long genomic stretches are sampled at uniform random locations by a procedure called shotgun sequencing. This results in numerous short DNA fragments that can be sequenced using conventional techniques. If this procedure is directly applied to an entire genome, it is called Whole Genome Shotgun (WGS) sequencing. After generating and sequencing such fragments, the target genome is computationally assembled from them. The primary information used during assembly is the pairwise overlaps that exist between fragments derived from the same region of the genome. Pairwise overlaps are detected by computing alignments between the corresponding pairs of fragments using standard dynamic programming approaches. Because such overlaps could also result from fragments derived from different but repetitive parts of the genome, fragments are typically sequenced in pairs from either end of longer DNA sequences (or sub-clones) of approximate known length (≈ 5,000 bp). Knowledge of the distances between paired fragments, known as clone mate information, is useful in detecting repeat-induced overlaps, but only for repeats shorter than sub-clone lengths.
Concomitant with advances in sequencing strategies and the undertaking of numerous genome sequencing projects, many genome assembly programs have been developed: Arachne [Batzoglou et al. (2002)], Atlas [Havlak et al. (2004)], CAP3 [Huang and Madan (1999)], Celera Assembler [Myers et al. (2000)], Euler [Pevzner et al. (2001)], GigAssembler [Kent and Haussler (2001)], PCAP [Huang et al. (2003)], Phrap [Green (1994)], Phusion [Mullikin and Ning (2003)] and TIGR Assembler [Sutton et al. (1995)]. Despite advances in hardware speeds and memory capacities over the same period, assembling genomes from the tens of millions of fragments typical of large sequencing projects places enormous demands on computational resources, with most of the run-time and memory spent in detecting overlaps using alignment algorithms and recording them. It is common for such work to be carried out by specialized teams on workstations with tens of gigabytes of main memory using manual efforts to partition the problem, a week or more of compute time, and disks for storing intermediate results. While this should make genome assembly an ideal application for parallel processing, most assemblers are serial and the few that take advantage of parallel processing do so in a rudimentary fashion—using multiple processors to accelerate the stage of the assembler that deals with computing large numbers of pairwise overlaps and/or manually partitioning the problem and launching multiple jobs on different processors. For example, one human genome assembly took 20,000 CPU hours for ∼27 million fragments using ten 4-processor SMP clusters each with 4 GB RAM, along with a 16-processor NUMA machine with 64 GB shared memory [Venter et al. (2001)].

Shotgun sequencing has been carried out for increasingly larger sized genomes over the past two decades, starting from the ∼50,000 long genome of the virus bacteriophage λ [Sanger et al. (1982)] to the recent sequencing of mouse, human and chimpanzee genomes that are 2.5 to over 3 billion nucleotides long. Current targets for large-scale genome sequencing include economically important plant crops such as maize, sorghum, soybean and wheat. In addition to their large sizes, sequencing and assembly of the genomes of these plants is considered particularly challenging because of the abundance of repeats in them. For instance, repeats are estimated to span 65-80% of the maize genome, which has an estimated size of 2.5–3 billion
nucleotides [Arumuganathan and Earle (1991)]. While the previously sequenced genomes contain repeats albeit at a smaller scale, repeats in maize are much harder to resolve due to very high sequence identity resulting from their short evolutionary history. On the other hand, the genes are estimated to occupy only 10-15% of the genome, mostly outside the repeat content [Bennetzen et al. (2001)]. To meet the goal of deciphering this relatively smaller “gene space” in highly repetitive genomes, biologists have designed experimental techniques such as Methyl Filtration in plants [Rabinowicz et al. (1999)] and High-Cot sequencing [Yuan et al. (2003)] that are expected to bias fragment sampling towards gene-rich regions [Palmer et al. (2003); Whitelaw et al. (2003)]. Similar gene-enrichment sequencing is also underway for sorghum [Bedell et al. (2005)] and loblolly pine [Peterson (2004)].

Traditionally, genome assemblers are designed with the expectation that fragments are obtained through uniform random sampling. For \( n \) fragments, it can be argued that their memory and run-time requirement is \( \Theta(n) \) for uniform sampling but is \( \Theta(n^2) \) in the worst-case for non-uniform sampling or when a significant fraction of fragments show mutual overlaps due to repeats, though the effect is not as bad in practice. As a concrete illustration, our experiments with the CAP3 assembler on a workstation with 2 GB RAM showed that just 80,000 maize fragments saturated the memory.

In this paper, we present the first massively parallel genome assembly framework. Our approach guarantees a worst-case \( O(n) \) total space complexity despite gene-enrichment and repeats. Like other assemblers, our method generates a selected set of pairs of fragments to which alignment computations are restricted. However, we generate such pairs in \( O(1) \) amortized time per pair and do so in a prioritized order, which is used to drive a heuristic strategy that significantly reduces the number of pairwise alignments computed without affecting quality. We demonstrate the effectiveness, scalability, and biological validity of our approach using data generated from diverse sequencing approaches: (a) random shotgun sequencing data from the fruit fly Drosophila psuedoobscura [Richards et al. (2005)], (b) a mixture of shotgun, gene enriched, and Bacterial Artificial Chromosome (BAC) derived fragments from maize, and (c) shotgun fragments from thousands of bacterial genomes from the Sargasso Sea environmental
sample [Venter et al. (2004)]. We present detailed experimental analysis for all three cases using a 1,024 node BlueGene/L supercomputer.

The rest of the paper is organized as follows: In Section 2, we present an overview of conventional genome assembly strategies. Section 3 contains a high level overview of our parallel assembly framework. The algorithmic and implementation details of our framework and scaling results using maize data are presented in Sections 4 through 7. Application of our framework to maize gene enriched sequence data, which resulted in the first publicly available assembly of this data, is presented in Section 8. Section 9 contains additional validation using random shotgun sequencing and environmental sample sequencing. Section 10 contains our concluding remarks.

**Genome Assembly**

Many assemblers follow a three phase “overlap-layout-consensus” paradigm. The first phase is the time-dominant phase, in which pairs of “significantly” overlapping fragments are detected. The overlap between a pair of fragments need not be an exact match due to errors in sequencing and natural genetic variations, if multiple individuals are selected in sequencing. The standard method for accounting these is to compute an optimal alignment between the fragments, allowing for insertions, deletions and substitutions. The alignment computation uses a dynamic programming algorithm [Gotoh (1982); Needleman and Wunsch (1970); Smith and Waterman (1981)] that takes time proportional to the product of the lengths of the fragments being aligned. When the number of fragments is in the millions, it is computationally impractical to apply this algorithm to all pairs of fragments.

Given that the fragments are small ($\approx 700–800 \text{ bp}$), and sampled from the much larger genome ($10^4–10^7$ times larger), only a small subset of the pairs will actually overlap. To exploit this, assembly algorithms use filters to generate a reduced subset of pairs and limit alignment computation to those pairs. A key property of a filter is that every pair of fragments that pass the alignment test should also be generated by the filter. Another important property is to directly generate pairs of fragments instead of testing every pair to see if it satisfies the
specified criterion. All of the filters are based on identification of exact matches.

Given the low rate (~1-2%) of errors, sequencing artifacts, and other variations, any good alignment is expected to contain long exact match regions, though the converse is not necessarily true. The most frequently used filter is to generate pairs that have an one or more exact match of a specified length, say $w$. Such pairs are easily identified using a lookup table constructed for all $w$-length substrings within each fragment [Pearson and Lipman (1988)]. A downside to this approach is that a long exact match of length $l$ reveals itself as $(l - w + 1)$ matches of length $w$; in practice, there could be many overlaps with matches spanning hundreds of nucleotides, while $w$ is kept as small as 10 or 11 because the size of the lookup table is exponential in $w$.

In the second phase, a layout consistent with the detected overlaps is constructed. The extent to which a genome is sampled in a sequencing project is indicated by coverage, which is the ratio of the total length of all the fragments to the estimated size of the genome. The coverage denotes the average number of fragments that contain a nucleotide. Typically a five to seven fold coverage (denoted 5X to 7X) is used for very large genomes. However, it cannot be guaranteed that each nucleotide in the genome is spanned by one or more fragments. Therefore, the final assembly typically consists of a large number of contiguous stretches called contigs interspersed by unsampled regions. As an example, in the human genome project [Venter et al. (2001)], using whole genome shotgun sequencing resulted in an initial assembly with over 221,000 contigs, and the largest contig spanned only under 2 million nucleotides of the genome. During the third phase, contigs are constructed from the layout on a consensus basis and/or by taking the available nucleotide-level sequencing quality values into account.

The order and orientation of the contigs along the chromosomes is later determined using a process called scaffolding. Programs developed for this purpose are called scaffolders. Targeted biological experiments are then used to “fill” the gaps in sequencing, and this process is known as finishing.

In the layout construction phase of the assembler, overlaps are typically sorted and processed in decreasing order of their quality. Sorting entails storing all overlaps, implying a linear
space complexity only if the fragments uniformly sample the target genome and spurious alignments due to repeats can be successfully avoided. When gene-enrichment strategies are used on highly repetitive genomes, these assumptions are no longer valid — the gene-enriched fragments correspond to a non-uniform sampling over the genic regions (as demonstrated in Emrich et al. (2004)), and even the small fraction of repetitive sequences that survive the initial screening is substantial because of their high initial frequency. Under these circumstances, the number of significantly overlapping pairs of fragments is expected to grow quadratically, although the effect is not as bad in practice because a majority of the fragments may contain characterized repeats that can be detected and “masked” in prior to assembly.

**Our Clustering-based Parallel Framework for Genome Assembly**

As mentioned before, sequencing gaps and regions difficult to sequence result in hundreds of thousands of contigs. Taking advantage of this, we propose a *cluster-then-assemble* approach that partitions the input fragments into “clusters” such that each cluster contains fragments constituting one or more contigs. The goal of clustering is to form as many clusters as possible while making sure that fragments that belong to the same contig are never split across clusters. Ideally, there should be as many clusters as contigs. Thus, clustering can be viewed as decomposing the assembly problem into a large number of much smaller, independent assembly problems. This framework is illustrated in Figure 1.
Our cluster-then-assemble approach has several advantages: If we develop parallel methods for clustering, the subsequent assembly tasks are trivially parallelized by distributing the clusters across multiple processors and running multiple instances of a serial assembler in parallel. This approach has allowed us to focus on developing parallel methods while benefiting from and not duplicating the painstakingly built-in biological expertise of current assemblers. The space and other limitations of these assemblers will now not be a limiting factor because of the relatively small size of each cluster. Furthermore, this allows one to generate assemblies consistent with what would have been generated by any conventional assembler, except that the problem size reach and speed are significantly enhanced.

Our main contributions in space optimality, run-time efficiency and parallel methodology lie in the clustering framework, and are described in detail in the following sections. The following attributes are essential for success in the clustering strategy:

**Correctness:** Fragments that belong to a contig should not be split across clusters because the subsequent assembly step is carried out independently, and there is no way of combining fragments from multiple clusters. For this reason, less stringent overlap criteria are used during clustering than during final assembly. This way, fragments belonging to the same contig are never separated into different clusters. However, this also implies that fragments from multiple contigs could be combined in a cluster. Even so, this is not a problem because the serial assembly on each individual cluster will detect such discrepancies, thereby guaranteeing the correctness of the overall assembly result.

**Speed:** To effectively utilize a large number of processors during the assembly phase, the fragments must be split into as many clusters as possible. Some measures of interest are the average number of contigs per cluster and the size of the largest cluster. Note that the former is purely a measure of clustering effectiveness while the latter highly depends on the input fragments. As the maximum cluster size determines parallelism in the assembly phase, it is important to experimentally demonstrate that it will not be a limiting factor.

While the large number of initial contigs in shotgun sequencing projects provides sufficient rationale for the proposal strategy, other sequencing techniques should result in even larger
number of clusters. Gene-rich sequencing should generate a large number of contigs that correspond to the many sparsely located “genomic islands” from which the fragments were originally derived. Similarly, collective sequencing of thousands of bacterial genomes from environmental samples should further magnify the number of clusters.

Clustering Fragments

Keeping in mind the goals of clustering, we formulate the clustering problem as follows: Two fragments are said to overlap if there is a “high quality” alignment between a suffix of one and a prefix of the other, also known as suffix prefix alignment. Two fragments are said to belong to the same cluster if and only if they overlap or there exists a chain of overlaps connecting them. Because of the transitive implication, this formulation may permit fragments with inconsistent overlaps to be clustered as illustrated in Figure 2(a). Resolving such inconsistencies is deferred until assembly. An advantage of allowing transitive clustering is the following observation: regardless of how a set of fragments sample an underlying genomic island, there exists a linear number of overlapping pairs that is sufficient to arrive at their clustering (see Figures 2(b) and 2(c)). While it is not possible to predict these in advance, the heuristic algorithm described below reduces run-time by increasing the likelihood that such pairs are identified early, without affecting the correctness of the final clustering.

We use the term promising pair to denote a pair of fragments that has a maximal match\(^3\) of length no smaller than a cutoff \(\psi\). The clustering algorithm is as follows: Let \(n\) denote the number of genomic fragments. Initially, each fragment is considered to be in a cluster by itself.

Pair Generation Heuristic: Promising pairs are generated in the non-increasing (henceforth, “decreasing” for convenience) order of their maximal match lengths.

Alignment Heuristic: Each generated pair is aligned only if the constituent fragments currently belong to two different clusters. If the alignment test succeeds, then the two clusters are merged into one. Otherwise, the clusters are left intact, and so the alignment effort is wasted.

\(^3\)A “maximal match” is an exact match between two fragments that cannot be extended on either side to result in a longer match.
Figure 2  Illustration of clustering: (a) Three fragments clustered because of transitivity despite not sharing consistent overlaps, i.e., \((f_1, f_2)\) and \((f_2, f_3)\) overlap, but \((f_1, f_3)\) do not overlap as depicted by the oval and rectangular regions. Parts (b) and (c) show genomic regions (shown in thick lines) with uniform and non-uniform sampling, respectively. In either case, a linear number of pairwise overlaps (shown in dotted lines) is sufficient to cluster the fragments. Note that such a combination of overlaps need not be unique.

The process of merging is continued until all promising pairs are considered. Figure 3 outlines our clustering strategy.

In the above clustering scheme, the number of merges is no more than \(n - 1\), though in the worst case a quadratic number of pairs \(O(n^2)\) could be aligned before arriving at the final clustering. Generating pairs based on maximal matches, as opposed to fixed length matches using lookup tables, helps in two ways: (i) it limits the number of times a promising pair is generated to the number of distinct maximal matches in it (as proved in Section ), instead of the considerably larger number of fixed length matches shared by the fragments; and (ii) it provides an effective way to distinguish among promising pairs, in terms of the expected overlap quality — longer the maximal match, higher the likelihood of surviving the alignment test. Therefore, processing pairs in this order is expected to result in early cluster merges, thereby significantly reducing the chance of a pair being selected for alignment work as the execution progresses.

Our clustering scheme primarily relies on the above two heuristics to drastically reduce the number of pairs aligned from \(\Theta(n^2)\) in the worst case to \(<< O(n^2)\) in practice. Note that the final clustering remains unchanged regardless of the order in which the pairs are processed because of the transitive closure property. Therefore, these are only run-time heuristics and do not affect the correctness of the algorithm.
Algorithm 1  Fragment Clustering

**Input:** Set $S = \{f_1, f_2, \ldots, f_n\}$ of $n$ sequences  
**Output:** A partition $C = \{C_1, C_2, \ldots, C_m\}$ of $S$, $1 \leq m \leq n$

1. Initialize Clusters: 
   
   \[ C \leftarrow \{ \{f_i\} \mid 1 \leq i \leq n\} \]

2. FOR each pair $(f_i, f_j)$ with a maximal match of length $\geq \psi$ 
   generated in non-increasing order of maximal match length 
   \[ C_p \leftarrow \text{Find}(f_i) \]
   \[ C_q \leftarrow \text{Find}(f_j) \]
   IF $C_p \neq C_q$ THEN
   \[ \text{score} \leftarrow \text{Align}(f_i, f_j) \]
   IF score is significant THEN
   \[ \text{Union}(C_p, C_q) \]

3. Output $C$

Figure 3  Our clustering strategy. Operations on the set of clusters are performed using the Union-Find data structure.

Generating Promising Pairs

In this section, we provide the serial version of our pair generation algorithm, and in Sections and we elaborate on the parallel version. We use the Generalized Suffix Tree (or GST; see Chapter 5 of [Aluru (2005)]) data structure for identifying maximal matches (of length $\geq \psi$) between fragments. The GST for a set of strings is a compacted trie of all suffixes of all the strings, and occupies space proportional to the input size. Our algorithm to generate promising pairs uses the GST built on all input fragments and their complementary strands\(^4\). Complementary strands are included because fragments could have been sequenced from either strand of the genomic DNA. For convenience, we use ‘fragment’ to refer to both types of sequences.

A naive approach to generate fragment pairs in decreasing order of maximal match length is to first identify all such fragment pairs and then sort them by their maximal match lengths. This scheme would, however, require $O(n^2)$ space, because sorting entails storing the pairs.

\(^4\)A complementary strand of a DNA fragment is obtained by reversing it and substituting $A \leftrightarrow T$ and $C \leftrightarrow G$. DNA is a double stranded molecule where the two strands are antiparallel.
Instead, we developed an algorithm with linear space complexity that directly generates the promising pairs in the sorted order without storing them. This is achieved through an “on-demand” scheme in which pairs are generated one at a time and are either immediately aligned or discarded as dictated by the current clustering. Eliminating the need to store promising pairs and pairwise alignment scores is key to achieving linear space.

Fragments are represented as strings over the alphabet \( \Sigma = \{A, C, G, T\} \). Let \( s[i] \) denote the character in position \( i \), and \( s(i) \) denote the suffix starting from position \( i \) of string \( s \). Positions are numbered starting at 1. Let \( s[i..j] \) denote the substring from positions \( i \) to \( j \) in \( s \). Let \( |s| \) denote the length of \( s \). Let \( N = \Sigma_{n=1}^\infty(|f_i|) \). The average length of a fragment is \( \approx 700 - 800 \), making \( N \approx 700n - 800n \).

For a node \( u \) in the GST, let \( \text{path-label}(u) \) denote the concatenation of edge labels along the path from the root to \( u \), and \( \text{string-depth}(u) \) denote its length. By definition of a GST, the path-label of every leaf is a suffix in at least one string. To ensure that each suffix of a string is represented by a leaf, a unique termination symbol ‘$’ is appended to each input string. Let \( \text{leaf}(s(i)) \) denote the leaf corresponding to suffix \( s(i) \), and \( \text{subtree}(u) \) denote the set of suffixes corresponding to all leaves in \( u \)’s subtree.

**Definition 1** A string \( \alpha \) is a maximal match between fragments \( f_i \) and \( f_j \) if and only if \( \exists f_i(k) \) and \( f_j(l) \) such that

- \( f_i[k..k + |\alpha| - 1] = f_j[l..l + |\alpha| - 1] = \alpha \),
- **Right maximality**: If \( f_i[k] \neq $ \) and \( f_j[l] \neq $ \), \( f_i[k + |\alpha|] \neq f_j[l + |\alpha|] \),
- **Left maximality**: If \( k \neq 1 \) and \( l \neq 1 \), \( f_i[k - 1] \neq f_j[l - 1] \).

**Definition 2** For a node \( u \) and \( c \in \Sigma \), let \( \ell_c(u) = \{f_i(j) \mid f_i(j) \in \text{subtree}(u); j > 1; f_i[j - 1] = c\} \), and \( \ell_\lambda(u) = \{f_i(1) \mid f_i(1) \in \text{subtree}(u)\} \). These are collectively called “lsets” at \( u \).

Basically, the lsets at \( u \) represent a partition of all suffixes with leaves in \( u \)’s subtree based on their preceding characters. The algorithm to generate fragment pairs with maximal matches is based on the following key observation.

\(^5\)In our implementation, we generate pairs in bounded size batches rather than one at a time to maximize the communication bandwidth, as explained in Section.
Lemma 1 Fragments \( f_i \) and \( f_j \) share a maximal match \( \alpha \) if and only if

C1. \( \exists \ u \) such that \( \text{path-label}(u) = \alpha \).

C2. \( \exists \ k \) and \( l \) such that \( f_i(k), f_j(l) \in \text{subtree}(u) \).

C3. If \( u \) is not a leaf, \( f_i(k) \in \text{subtree}(u') \) and \( f_j(l) \in \text{subtree}(u'') \), where \( u' \) and \( u'' \) are two different children of \( u \).

C4. If \( k \neq 1 \) and \( l \neq 1 \), \( f_i(k) \in \ell_{c'}(u) \) and \( f_j(l) \in \ell_{c''}(u) \), for \( c', c'' \in \Sigma \) and \( c' \neq c'' \).

Proof: \( (\Rightarrow \ C1,C2,C3,C4) \) The fact that \( \alpha \) is a match between \( f_i \) and \( f_j \) implies that \( \exists \ f_i(k) \) and \( f_j(l) \) that share the prefix \( \alpha \). This implies that the path-labels of \( \text{leaf}(f_i(k)) \) and \( \text{leaf}(f_j(l)) \) have \( \alpha \) as a common prefix. Moreover, the right-maximality of \( \alpha \) implies that there will exist an internal node \( u \) that is the lowest common ancestor between \( \text{leaf}(f_i(k)) \) and \( \text{leaf}(f_j(l)) \), unless \( \alpha = f_i(k) = f_j(l) \). The left-maximality of \( \alpha \) ensures that \( f_i(k) \) and \( f_j(l) \) are in \( lsets \) corresponding to two different characters at \( u \), unless \( k = 1 \) or \( l = 1 \).

\( (\Leftarrow \ C1,C2,C3,C4) \) C1 and C2 imply that \( f_i(k) \) and \( f_j(l) \) share \( \alpha \) as a prefix, and therefore \( \alpha \) is a match between \( f_i \) and \( f_j \). Moreover, the fact from C3 that \( \text{leaf}(f_i(k)) \) and \( \text{leaf}(f_j(l)) \) are in two different children of \( u \) implies that the leading characters in the edge-labels to \( u' \) and \( u'' \) are different, thereby implying the right-maximality of \( \alpha \). Similarly, C4 implies the left-maximality of \( \alpha \).

Thus, identifying all pairs that satisfy C1...C4 at a node \( u \) translates into identifying all pairs with the maximal match \( \text{path-label}(u) \). Pairs can be generated in decreasing order of their maximal match lengths by processing the nodes in the decreasing order of their string-depths.

The Pair Generation Algorithm

S1. Compute the GST of all \( n \) fragments. Section provides a parallel algorithm.

S2. Sort the nodes with string-depths \( \geq \psi \) in decreasing order of string-depths, and “process” the nodes in that order. Use step S3 for a leaf and S3 for an internal node.

S3. Let \( u \) be the leaf node. First, the \( lsets \) at \( u \) are computed directly from its labels. As right-maximality is automatically satisfied at a leaf, it suffices to check for C4. Thus,
pairs at $u$ are obtained by computing $\bigcup \ell_c(u) \times \ell_{c'}(u)$, where $c < c'$ or $c = c' = \lambda$. An arbitrary ordering of characters in $\Sigma \cup \{\lambda\}$ is assumed in order to limit generating a pair to one of its forms: either $(f_i(k), f_j(l))$ or $(f_j(l), f_i(k))$.

S4. Let $u$ be the internal node. Processing the nodes in the sorted order ensures that $u$ is processed only after all its children are processed. Therefore, by induction let us assume that the lsets at every child of $u$ are already computed. As $\text{subtree}(u')$ at any child $u'$ of $u$ is given by $\bigcup_{c \in \Sigma \cup \{\lambda\}} \ell_c(u')$, there is no need to compute it explicitly. The set of pairs at $u$ is obtained by computing $\bigcup \ell_c(u') \times \ell_{c'}(u'')$, for every pair of different children of $u$, $u' < u''$ (to satisfy C3), and and $c < c'$ or $c = c' = \lambda$ (to satisfy C4). Again, an arbitrary ordering of characters and child nodes is assumed to prevent generating redundant copies of pairs. After generating pairs from $u$, its lsets are computed from the lsets of its children as follows: $\ell_c(u) = \bigcup_{u'} \ell_c(u')$.

Figure 4 shows an example illustrating the pair generation algorithm at an internal node.

**Lemma 2** The run-time complexity of the pair generation algorithm is $O(N + \text{# promising pairs generated})$. The space-complexity is $O(N)$.

**Proof:** Step S1 is achieved in $O(N)$ time and space (Chapter 5; Aluru (2005)). Radix sort can be used for S2 because the number of nodes and string-depth value are bound by $O(N)$. The overall cost for computing the lsets at all leaves is $O(N)$ because it requires examining at most one character for each suffix. Generating pairs by computing a cross-product within the leaf’s lsets (as shown in S3) takes $O(1)$ time per pair. Similarly, for an internal node $u$, generating pairs by computing a cross-product across lsets of different children (as shown in S4) also takes $O(1)$ time per pair. The lsets at each node are maintained as linked lists to allow constant time union operations for creating each $\ell_c(u)$ from its counterparts in $u$’s children. Given that there are at most $|\Sigma| + 1$ lsets and $|\Sigma| + 1$ children for each node, the total cost of generating the lsets at an internal node is $O(|\Sigma|^2)$, which is bound by the number of pairs generated at $u$. Thus, the amortized cost for generating each pair at an internal node is also $O(1)$. The linked list implementation of lsets ensures that the overall space required to store the lsets is restricted to $O(N)$.

$\blacksquare$
Figure 4 Detection of maximal match pairs at an internal node $u$. Part (a) shows three fragments sharing a match $\alpha$. Part (b) shows the node $u$ in the GST with path-label $\alpha$ with all its children. Each column under a child shows its lsets corresponding to characters in $\Sigma \cup \{\lambda\}$. Solid arrows denote the maximal match pairs generated by our algorithm.

The above scheme generates all maximal matches (of length $\geq \psi$) between each pair of fragments. This is needed if pairwise alignment computations are anchored to the maximal matches. If arbitrary suffix prefix alignments are computed, then it is wasteful to generate the same pair multiple times. In such a case, the algorithm can be improved to reduce the number of duplicate generations of the same fragment pair. Instead of partitioning the suffixes in a node’s subtree, we now partition the corresponding fragments into its lsets — i.e., the definition of lsets changes. Formally, $f_i \in \ell_c(u)$ only if $\exists f_i(k) \in \text{subtree}(u)$ such that $f_i[k-1] = c$ or $k = 1$. Before generating pairs at an internal node $u$, the lsets at $u$’s children are scanned such that all but one arbitrary occurrence of a fragment are removed. We call this process “duplicate elimination”, following which the pair generation algorithm is run as before, except
that the generated pairs are of the form \((f_i, f_j)\) instead of \((f_i(k), f_j(l))\).

There is a key difference between this partitioning scheme and the previous. Because an arbitrary occurrence of a fragment is retained at a node \(u\), the \(lsets\) at \(u\) may no longer be unique if \(u\) is an internal node\(^6\). As it is only the retained occurrence that participates in pair generation, the set of pairs generated at \(u\) may also no longer be unique. Therefore, it is possible that a pair containing \(\text{path-label}(u)\) as a maximal match is not generated at \(u\). However, this is not a problem because it can be guaranteed that our algorithm generates the fragment pair at least once. The proof is as follows: Of the maximal matches between \((f_i, f_j)\), consider a largest maximal match \(\alpha\), and let \(u\) be the node with that path-label. For the case of leaf, the proof is trivial. If \(u\) is an internal node, any pair of occurrences of \(f_i\) and \(f_j\) retained by the duplicate elimination process will constitute a valid pair that our algorithm will generate at \(u\); otherwise, \(\alpha\) cannot be a longest maximal match, a contradiction.

The above improvement guarantees that a fragment pair is generated at most once under a node \(u\). But there could be multiple distinct maximal matches between the same fragment pair, each corresponding to a different node in the GST. Therefore, in the worst case, a pair is generated at most as many times as the number of distinct maximal matches between between the fragments. As for run-time complexity, the eliminating duplicates under \(u\) takes time proportional to the sum of the sizes of \(lsets\) of all its children, which is bounded by size of \(lsets\) at \(u\) times \(|\Sigma|+1\) — a term that is bound by the number of pairs generated at \(u\), assuming \(|\Sigma|\) is a small constant.

**Parallel Generalized Suffix Tree Construction**

There are no provably optimal and practically efficient parallel algorithms for suffix tree construction suited for distributed memory parallel computers. We developed the following algorithm that works well in practice. Let \(p\) denote the number of processors.

The first step is to sort all suffixes based on their \(w\)-length prefixes, where \(w \leq \psi\). Partition the fragments such that each processor has approximately \(\frac{\Sigma}{p}\) nucleotides. Through a linear

\(^6\)At a leaf, a fragment can occur at most once in its \(lsets\) — ie., \(\exists\) only one combination for \(lsets\).
scan, each processor partitions the suffixes of its fragments into $|\Sigma|^w$ buckets based on their first $w$ characters. The suffixes are then globally redistributed such that those belonging to the same bucket are in the same processor, and the number of suffixes per processor is approximately $\frac{N}{p}$. While adversarial input such that only one bucket contains all $N$ suffixes can be easily constructed, this poses no difficulty in practice because of the following expectation: if input sequence data is random, the expected number of distinct $w$-substrings in them is $|\Sigma|^w$.

While this may not entirely hold in practice, the fragments are still diverse enough because of sequence cleaning and repeat masking. Therefore, a value between 10–12 for $w$ is expected to generate millions of buckets sufficient to be distributed in a load balanced manner even for thousands of processors. Empirically, a value of $w = 11$ was found appropriate for the data and the range of processors we tested (up to 1,024 processors).

The next phase consists of constructing for each bucket, a compacted trie of all its suffixes. Each of these represents a subtree in GST rooted at a node with string depth $\geq w$. We construct each trie in a depth-first manner as follows: Partition all suffixes in the bucket into at most $|\Sigma|$ sub-buckets based on their respective $(w + 1)^{th}$ characters. This is recursively applied for each sub-bucket by examining characters in subsequent positions until all suffixes are separated or their lengths exhausted. In the worst case, this procedure visits all suffixes to their full lengths, resulting in a run-time of $O\left(\frac{N \times l}{p}\right)$, where $l$ is the average length of an input fragment. We now have a distributed representation of the GST as a collection of subtrees containing all nodes at depth $\geq w$. The top portion of the GST is not needed for pair generation.

The main challenge in this scheme is acquiring the fragments required to construct the local subtrees. Storing all fragments with suffixes in local buckets requires $O\left(\min\{\frac{N \times l}{p}, N\}\right)$ space in the worst case, which is not a scalable solution. Space can be reduced by constructing one subtree at a time, and loading all fragments required for a subtree from disk prior to its construction. Given that disk latencies are in the millisecond range for random accesses as required here, we developed an alternative to take advantage of the high bandwidth interconnection network typical of large scale parallel computers such as the BlueGene/L.
Each processor partitions its buckets into variable-sized batches, such that the fragments required to construct all buckets in each batch would occupy $\Theta\left(\frac{N}{p}\right)$ space. Before constructing a batch, all fragments needed for its construction are fetched through two collective communication steps — the first to request the processors that have the required fragments, and the second to service the request. The processor that has a given fragment is determined in constant time by recalling the initial distribution of the fragments. A processor may exhaust all its batches, in which case it continues to participate in the remaining communication rounds to serve requests from other processors.

In the above communication based solution, each processor receives $O\left(\frac{N}{p}\right)$ characters from all other processors per communication step. However, the size of the buffer used to send fragments to other processors may exceed $O\left(\frac{N}{p}\right)$. This is because requests from different processors may intersect, in the worst case over all of $O\left(\frac{N}{p}\right)$ local data; the likelihood of this scenario increases with the number of processors. We resolved this issue by implementing a customized Alltoallv, which ensures $O\left(\frac{N}{p}\right)$ size for the buffers by doing $p - 1$ sends and receives instead of one collective communication.

**Experimental Results**

We studied the performance of our GST construction algorithm by varying the number of processors from 256 to 1,024. Each dual-processor node of the BlueGene/L system was used in co-processor mode, i.e., one processor was used for computation and the other processor was used for communication. Experiments were conducted on two subsets of the maize data, with sizes 250 and 500 million nucleotides that comprised 322,009 and 649,957 fragments, respectively. Figure 5 shows the parallel run-times and their breakdown into communication and computation times, all of which show linear scaling with both processor and input sizes.

**Detecting Overlaps And Managing Clusters In Parallel**

Once a distributed representation of the GST for all input fragments is constructed in parallel, each processor can generate promising pairs from its portion of the GST using the
algorithm described in Section \( \text{Section } \). As pairs are generated, they need to be checked against the current clustering, allocated for alignment if necessary, and the alignment results interpreted to update the current clustering. To implement these tasks in parallel, we designed an iterative solution with one master and \( p - 1 \) worker processors.

In addition to the load balancing concerns typical in a single master multiple worker setup such as keeping all the worker processors busy and the master processor available most of the time, our master-worker model presents other unique challenges. The worker processors in our model, in addition to processing the tasks (by aligning pairs), also generate tasks (by generating pairs). Thus, care must be taken that the rate of work generation is neither too fast to result in a memory overflow (because a batch of pairs needs to be stored until the master processor decides if they should be aligned) nor too slow to result in unnecessary processor wait times. Moreover, as not all generated pairs are necessarily selected for alignment, it is important to regulate the rate of pair generation in order to maintain a steady rate in alignment computation. Another concern may arise when processors start to run out of pairs to generate from their portion of the GST as execution progresses. Henceforth, we call such processors \textit{passive} while those that still have pairs to generate are called \textit{active} processors. In the interest of maintaining

![Figure 5](image-url)

**Figure 5** Parallel run-times for constructing GST on inputs of sizes: (a) 250 million, and (b) 500 million nucleotides.
parallel efficiency, it is necessary to keep passive workers busy computing alignments. Also, allocating pending alignment computations to passive workers ahead of any active worker can help balance pair generation and pairwise alignment computation dynamically.

With the above goals in mind, we designed an iterative solution with responsibilities as shown in Figure 6. The master and worker processors interact iteratively until all promising pairs are generated and all alignments identified as necessary have been computed. The master processor is responsible for maintaining the clusters, selecting and allocating pairs for alignment computation, and load balancing. Each worker processor is responsible for generating promising pairs from its local GST portion in decreasing order of maximal match length, computing alignments for pairs allocated by the master processor, and reporting the alignment results to the master processor. To reduce communication setup costs, the worker processors send pairs in batches instead of one pair at a time. Similarly, the master processor also allocates pairs for alignment computation and collects their results in batches.
The master and workers store and maintain the following information.

**Information at the Master Processor:**

- **Clusters:** The current set of clusters maintained using the Union-Find data structure. This allows the operations of finding the cluster containing a given fragment, and the merging of two clusters to run in amortized time given by the inverse of Ackermann’s function, a constant for all practical purposes.

- **Pending Work Buf:** A fixed size buffer to temporarily store the pairs selected but not yet dispatched for alignment computation. This is implemented as a queue.

- **Idle Workers:** A list of all passive workers that do not have any alignment work allocated to them. This is implemented as a queue.

**Information at a Worker Processor** \( P_i \):

- **GST\(_i\):** The local portion of GST.

- **New Pairs Buf:** A fixed size buffer to temporarily store newly generated promising pairs that have not yet been sent to the master processor. This is implemented as a queue.

The following messages are exchanged between the master and an arbitrary worker processor \( P_i \) during one iteration:

- **AW:** A new batch of alignment work allocated by master to \( P_i \). The number of pairs sent in each batch, called *batch size*, is a fixed, user-specified parameter with value \( b \). AW is implemented as an array.

- **r:** The number of promising pairs to be sent by \( P_i \) during its next communication with the master.

- **NP:** A batch of new promising pairs sent by \( P_i \) to the master processor. This is implemented as an array.
• **AR:** A list of alignment results sent by $P_i$ to the master processor. The results are for the alignments computed over the most recent batch of pairs allocated by the master to $P_i$. AR is also implemented as an array.

Figures 7 and 8 detail the algorithms for the master and a worker processor, respectively. In each iteration, the master processor polls for messages from any of the workers. When a message arrives from a worker $P_i$, the master updates *Clusters* using the alignment results that are satisfactory, scans the batch of newly generated pairs from $P_i$, and adds only those pairs for which alignments are necessary to *Pending Work Buf*. It then repeatedly extracts batches of size $b$ from *Pending Work Buf*, dispatching each batch to an idle worker. If all workers become idle, then it signals the end of clustering. If no more idle workers remain and if there is more work left in the *Pending Work Buf*, then the next batch of $b$ pairs are allocated to $P_i$. In the same message, the master also piggybacks the number of new pairs, $r$, that it expects to receive from $p_i$ in its next communication; $r$ is given by: $\min\left\{ \frac{|NP|}{|NP|'} \times \frac{p}{p_{active}} \times b, \frac{|Pending Work Buf|}{p_{active}} \right\}$, where $p_{active}$ denotes the number of active processors. The main idea is to request as many pairs as necessary to expect that $b$ of them would be selected for alignment computation, while not overflowing *Pending Work Buf*. In other words, this load balancing strategy aims at regulating the inflow of work so as to keep the outflow roughly constant.

In each iteration, a worker processor generates as many new promising pairs as requested by the master processor and sends them in a message along with the results of the latest alignments it computed. While waiting for the master to reply, the worker computes alignments on the batch of pairs allocated by the master during the previous iteration. This is effective in masking the communication wait time with computation. If alignment computation is completed before the master replies, then the worker processor resumes from its earlier state of pair generation and generates fresh batches of promising pairs from its local GST portion until either a message from the master arrives or its temporary store *New Pairs Buf* is full. If a worker becomes passive, it keeps itself busy by computing alignments that the master allocated.
Algorithm 2 Algorithm for Master Processor

1. Clusters ← Initialize such that each fragment is in a cluster of its own
   $p_{active} ← p$
   Idle_Workers ← ∅
2. REPEAT
   Blocking Receive until message from an arbitrary processor $P_i$
   NP ← new promising pairs
   AR ← alignment results
   IF NP = ∅ AND $P_i$ is active THEN
     Mark $P_i$ as passive
     Decrement $p_{active}$
   Update Clusters based on AR
   NP′ ← Identify pairs in NP that need alignment computation
   Add NP′ into $Pending\_Work\_Buf$
   $r ← \min\left\{ \frac{|NP|}{|NP|^2} \times \frac{p}{p_{active}} \times b, \frac{|Pending\_Work\_Buf|}{p_{active}} \right\}$
   FOR EACH $P_j ∈ Idle\_Workers$ DO
     s4AR ← Dequeue $\min\{b, |Pending\_Work\_Buf|\}$ pairs
     IF AR $\neq ∅$ THEN
       Send AR to $P_j$
       Remove $P_j$ from Idle_Workers
     AR ← Dequeue $\min\{b, |Pending\_Work\_Buf|\}$ pairs
     IF AR $\neq ∅$ OR $r > 0$ THEN
       Send (AR, $r$) to $P_i$
     ELSE
       Insert $P_i$ into Idle_Workers
   UNTIL all workers become idle
3. Send termination signal to all workers
4. Output Clusters

Figure 7 Algorithm for the master processor. Bold font indicates a communication step.
Algorithm 3 Algorithm for a Worker Processor $P_i$

1. $AW \leftarrow$ Generate next $b$ promising pairs from $GST_i$
2. $AR \leftarrow$ Compute alignments on $AW$
3. $AW \leftarrow$ Generate next $b$ promising pairs from $GST_i$
4. $NP \leftarrow$ Generate next $b$ promising pairs from $GST_i$
5. $r \leftarrow b$
6. REPEAT
   Send $NP$ and $AR$ to master
   $AR \leftarrow$ Compute alignments on $AW$
   $(AW,r)\leftarrow$ Non-blocking Receive from master
   REPEAT
     Generate $r$ pairs from $GST_i$ and add to $New\_Pairs\_Buf$
   UNTIL message arrives from master OR $New\_Pairs\_Buf$ is full
   IF no message from master THEN
     $(AW,r)\leftarrow$ Blocking Receive until master sends a message
     $NP \leftarrow$ Extract $r$ pairs, first from $New\_Pairs\_Buf$ and then from $GST_i$ if necessary
   UNTIL no more promising pairs to generate from $GST_i$
7. REPEAT
   $AW \leftarrow$ Blocking Receive from master
   $AR \leftarrow$ Compute alignments on $AW$
   Send $AR$ to master
   UNTIL master sends termination signal

Figure 8 Algorithm for each worker processor. Bold font indicates a communication step.
Run-time and space complexity

The parallel work of the master-worker phase for performing pair generation, alignment and clustering is proportional to $O(K + K' \times l^2 + K \times A(n))$, where $K$ is the number of promising pairs generated, $K'$ is the number of pairs selected for alignment, $l$ is the average length of a fragment, and $A()$ is the inverse of the Ackermann function required to access the union-find data structure for clustering. Although $K$ and therefore, $K'$ are $O(n^2)$ in the worst-case, the maximal match based definition of promising pairs and the algorithmic heuristics in our approach ensure that they are $<< O(n^2)$ in practice. For example, for $n \approx 1.6$ million maize fragments, $K \approx 48$ million pairs and $K' \approx 22$ million pairs.

The overall space complexity for each worker processor is $O\left(\frac{N}{p}\right)$. For the master processor, the union find data structure is implemented as an array of $n$ integers and, therefore the space complexity is $O(n)$. Even for systems with a low memory footprint per processor such as the IBM BlueGene/L that has only 512 MB per node, this implementation allows upto a value of 128[s5] million for $n$, assuming 4 bytes per integer.
Experimental Results

We studied the performance of our master worker implementation on the BlueGene/L (see Figure 9). The results show a better scaling for the larger (500 million) input than the smaller (250 million) input. Upon increasing the number of processors from 256 to 1,024, we observe relative speedups of 2.6 for the 250 million input and 3.1 for the 500 million input. Further investigation revealed that the percentage average idle time for the processors increased from 16% on 256 processors to 26% on 1,024 processors on the 250 million input, and from 9% to 16% for the 500 million input — indicating that the processor size needs to be quadrupled with doubling the problem size to maintain parallel efficiency. This is expected because of the near quadratic growth in the alignent workload with input size. Note that a full sequencing project will generate over 22 billion nucleotides (30 million fragments each about 750 bp long), on which tens of thousands of processors can be utilized with our scheme. In fact, we tested an 8,192-processor run on an input containing as small as 1.15 billion nucleotides, and observed a total run-time of 75 minutes.

Figure 9b shows the number of promising pairs generated as a function of the input size. This figure also shows the effectiveness of our clustering heuristic in significantly reducing the number of alignments computed. For the entire maize data, which has 1,607,364 fragments of total size 1.252 billion nucleotides, only about 40% of the pairs generated are aligned. However, less than 1% of the pairs aligned contributed to merging of clusters, indicating the presence of numerous medium-sized (∼ 100 bp) repeat elements that survived initial screening procedures. Growth in the number of promising pairs is a direct reflection of the expected worst-case quadratic growth in the maize data. The number of promising pairs generated and the relative savings in the alignment work are highly data sensitive. For example, we observed that only 22% of generated pairs were aligned on a different data [Kalyanaraman et al. (2003)].

Currently, the master processor is designed to handle one request at a time. Messages arriving concurrently from multiple processors are therefore buffered at the MPI level on the master node. Message sizes can range from tens to hundreds of kilobytes depending on the requests made by the master processor, implying that the MPI buffer at the master node can
potentially overflow for larger number of processors. To avoid message losses, our implementation uses \texttt{MPI\_Send} that sends a message to the master processor only after a corresponding receive has been posted. Using \texttt{MPI\_Send}, however, indicated a performance degradation of about 30\% as opposed to using \texttt{MPI\_Isend} or \texttt{MPI\_Send} both on the BlueGene/L and a Myrinet Pentium cluster. An alternative is to change the underlying design to allow scaling the number of master processors with processor size. This will also significantly enhance the availability of the master processors. With the current single master implementation, we observed a gradual decrease in its availability (idle time) from 90\% to 70\% when the processor size was increased from 256 to 1,024.

There are a few key challenges in designing a multiple-master approach: If the set of clusters is replicated on all master processors, then a mechanism is required to dynamically monitor and update the local clustering to keep it consistent with the several other remote clusterings. Otherwise, the amount of unnecessary alignment work could increase because the latest clustering is not available to take advantage. Alternatively, if the clusters are partitioned across master processors, then a scheme to address the clustering of fragments maintained in different master processors is required.

\textbf{Maize Genome Assembly}

Our initial parallel assembly framework was designed to carry out assemblies of gene-enriched fragments derived from maize to make them available to the maize genetics community as quickly as possible [Emrich et al. (2004)]. Newer versions of the assembly were generated as additional sequences became available. The analysis results presented in this paper are based on maize genomic data composed of 3,124,130 fragments. This includes 852,838 Methyl-Filtrated (MF) [Rabinowicz et al. (1999)] and High-C\textsubscript{0}t (HC) [Yuan et al. (2003)] fragments. The MF strategy is based on the elimination of bacterial colonies containing methylated sub-clones, which are typically non-genic regions in plants. The HC strategy utilizes hybridization kinetics to enrich for lower copy sequences, which in case of maize are mostly genic regions. Also available are fragments from WGS sequencing and another strategy called \textit{Bacterial Artificial}
Table 1 Maize genomic fragment data types and size statistics: Methyl–filtrated (MF), High-C<sub>0</sub>t (HC), Bacterial Artificial Chromosome (BAC) derived, and Whole Genome Shotgun (WGS).

<table>
<thead>
<tr>
<th>Fragment Type</th>
<th>Number of Fragments</th>
<th>Total length (in millions)</th>
<th>Number of Fragments</th>
<th>Total length (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>411,654</td>
<td>335</td>
<td>349,950</td>
<td>288</td>
</tr>
<tr>
<td>HC</td>
<td>441,184</td>
<td>357</td>
<td>427,276</td>
<td>348</td>
</tr>
<tr>
<td>BAC</td>
<td>1,132,295</td>
<td>964</td>
<td>425,011</td>
<td>307</td>
</tr>
<tr>
<td>WGS</td>
<td>1,138,997</td>
<td>870</td>
<td>405,127</td>
<td>309</td>
</tr>
<tr>
<td>Total</td>
<td>3,124,130</td>
<td>2,526</td>
<td>1,607,364</td>
<td>1,252</td>
</tr>
</tbody>
</table>

Chromosome (BAC) sequencing, in which long genomic sequences (∼150,000–200,000 bp) are cloned in bacterial vectors, and their ends and internal regions are individually sampled through sequencing. A summary of the entire maize data is provided in the first three columns of Table 1.

As with any assembler, the first step in our framework is to “preprocess” the input fragments: raw fragments obtained from sequencing strategies can be contaminated with foreign DNA elements known as vectors, which are removed using the program Lucy [Chou and Holmes (2001)]. In addition, we designed a database of known and statistically-defined repeats [Emrich et al. (2004)] and screened all fragments against it. The matching portions are masked with special symbols such that our clustering method can treat them appropriately during overlap detection. The last two columns in Table 1 show the results of preprocessing the data using our repeat masking and vector screening procedures. As expected, preprocessing invalidates a significant number of shotgun fragments (∼60-65%) because of repeats, while most of the fragments resulting from gene-enrichment strategies are preserved. An efficient masking procedure is important because unmasked repeats cause spurious overlaps that cannot be resolved in the absence of paired fragments spanning multiple length scales.

The results of applying our parallel genome assembly framework are as follows: Preprocessing the 3,124,130 fragments downloaded from GenBank took 1 hour by trivially parallelizing
on 40 processors of a Myrinet Pentium cluster with 1.1 GHz Pentium III processors and 1 GB RAM per processor. Our clustering method partitioned the resulting 1,607,364 fragments (over 1.25 billion nucleotides) in 102 minutes on 1,024 nodes of the BlueGene/L. Construction of the GST, containing over 2.5 billion leaf nodes, took only the first 13 minutes. We used CAP3 [Huang and Madan (1999)] for assembling the fragments in each resulting cluster. This assembly step finished in 8.5 hours on 40 processors of the Myrinet Pentium cluster through trivial parallelization.

Our clustering resulted in a total of 149,548 clusters containing two or more input fragments, and 244,727 singletons. Singletons are fragments that do not cluster with any other fragment because of sharing no overlap and/or having a high repetitive content that was masked during preprocessing. The average number of input fragments per non-singleton cluster is 9.00, while the maximum is 86,369, or 5.37% of the input size. It should be emphasized that targeted and effective masking of repeats has significant influence on the largest cluster size [Emrich et al. (2004)]. On an average, each cluster assembled into 1.1 contigs. Given that the CAP3 assembly was performed with a higher stringency, this result indicates the high specificity of our clustering method and its usefulness in breaking the large assembly problem into disjoint pieces for conventional assembly. The results of this assembly are available at http://www.plantgenomics.iastate.edu/maize under “Download”.

While the above results demonstrate the effectiveness of the proposed massively parallel framework, it is important to ensure the correctness of the assembly. Validation was carried out using a number of computational and experimental methods. Alignments to ten highly accurate maize genes (74kb of highly finished sequence), as described previously [Fu et al. (2005)], indicated a less than 1 nucleotide in 10,000 was incorrect relative to the benchmark. These few inconsistencies are residual errors from the sequencing process. Even so, this assembly approaches the quality standard agreed upon for the human genome. For extensive validation of the assembly parameters used during this study and large-scale experimental verification of predicted novel maize genes on an earlier version of maize data with less than a million fragments, the reader is referred to Fu et al. (2005). Perhaps the most comprehensive validation
is the successful use of our assemblies by hundreds of researchers over the last three years in a variety of projects. Because the most accurate benchmarks are the DNA molecules in a maize cell, each success has confirmed our reconstruction of the maize genome to be correct.

**Whole Genome Shotgun and Environmental Assemblies**

To demonstrate the effectiveness of the proposed approach for other types of sequencing projects, we consider whole genome shotgun (WGS) sequencing, and the more recent approach of simultaneously sequencing fragments from thousands of bacterial genomes in environmental samples.

**Assembly of the Drosophila Genome**

To demonstrate the effectiveness of our cluster-then-assemble approach to conventional whole genome shotgun sequencing, we reassembled the recently sequenced genome of the fruit fly *Drosophila pseudoobscura*. This fruit fly species diverged from the model organism *Drosophila melanogaster* approximately 25–55 million years ago [Richards et al. (2005)], and its genome contains approximately 205 million nucleotides. It was sequenced using a WGS approach by the Baylor College of Medicine [Richards et al. (2005)].

The initial dataset consisted of 2,686,355 sequences downloaded from GenBank’s trace archive. After trimming using Lucy under default parameters, a total of 2,666,207 high quality fragments were obtained totaling 1.81 billion nucleotides. The corresponding coverage provided by these trimmed reads (8.8X) is consistent with the observed coverage in the *D. pseudoobscura* assembly (9.1X; Richards et al. (2005)), suggesting that trimming will not affect the resulting number of contigs. Estimation of repeats and masking of repeat sequences is key to producing a manageable largest cluster size. Repeats can be identified through their statistical over-representation in a random sample. Because WGS fragments themselves comprise a random sample, we used 32,462 randomly chosen fragments (0.1X coverage) to predict 5,407 high-copy sequences in this fruit fly genome. Repeat masking resulting in 2,074,483 fragments that comprised 1.37 billion unmasked bases. Clustering of this masked data took 3.1 hours on 1,024
nodes of BlueGene/L. A total of 32,893 non-singleton clusters and 174,277 singleton clusters were generated. The average cluster size is 57.77 and the largest cluster is composed of 140,307 fragments (about 6.76% of the total).

To demonstrate the importance of repeat masking to the proposed cluster-then-assemble approach, we performed clustering without repeat masking. Not only did clustering take 24 hours on 1,024 BlueGene/L nodes due to the large number of pairwise alignments forced by the repeats, almost 50% of the fragments were combined into one large cluster.

Biological validation of our assembly was carried out by using the published genome as benchmark. We aligned over 1.3 million random WGS fragments to the draft *D. pseudoobscura* assembly using BLASTN [Altschul et al. (1990)] (95% identity over 80% of the fragment’s length). The lower identity was used to compensate for sequencing errors (95% identity) and lower coverage to accommodate residual vector sequence (80% of the fragment’s length) that may be present in low quality sequences. Only the best match was stored for further analysis. Significantly, 27,830 out of 28,185 clusters post-masking (98.7%) map to a single benchmark sequence. This result suggests that even prior to assembly our clustering methodology in tandem with masking achieves high specificity.

**Assembly of Environmental Genome Sequences**

Most bacterial species cannot be currently studied under laboratory conditions. They can, however, be subjected to modern DNA sequencing approaches and such “metagenomics” approaches have recently been applied to “survey” entire microbial communities [Venter et al. (2004)]. Assembly of heterogeneous samples of DNA is an open problem convoluted by closely related bacterial sequences present in a sample. Even so, such complications do not affect clustering and, although an assembler would have the difficult task of differentiating highly similar sequences, deconvolution would be made easier by reducing individual problem sizes using clustering.

We tested the effectiveness of our clustering approach on the largest environmental WGS dataset currently available: 1.66 million fragments obtained from the Sargasso Sea [Venter
et al. (2004)]. After removing ubiquitous sequences from the sample by masking, a total of 825,696 clusters were obtained generated including 129,741 non-singleton clusters. Although comprehensive validation of resulting contigs would be difficult because of inherent complexities, our cluster-then-assemble approach could enhance any future environmental assembler. This is especially true given the extensive diversity in such samples; the Sargasso Sea dataset contains over 1,800 unique species, many of which could generate one or more clusters.

Table 2 shows further experimental results on our clustering of *Drosophila* fragments and environmental sequences. The total run-time is about the same in both cases because the numbers of the pairs aligned are roughly the same (∼26-27 million pairs). The table also shows the savings achieved by our clustering strategy in the number of pairs selected for alignment computation — for the environmental data ∼57% of the pairs generated are not selected for alignment, while for the *D. pseudoobscura* data this fraction is 65%.

**Conclusions and Future Directions**

We presented the design and development of an efficient clustering-based framework for genome assembly on massively parallel distributed memory computers. We demonstrated the effectiveness of our approach on random shotgun sequencing, gene-enriched sequencing, and sequencing of environmental samples. Run-time results on a 1,024 node BlueGene/L show significant reduction in assembly turnaround times, from weeks to less than a day. Faster turnaround times also encourage more experimentation, which is useful as a host of parameters

<table>
<thead>
<tr>
<th></th>
<th>Number of fragments</th>
<th>Total length (in billion bp)</th>
<th>Clustering time (in minutes)</th>
<th>Promising pairs (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GST</td>
<td>Total</td>
<td>Accepted</td>
<td>Rejected</td>
</tr>
<tr>
<td><em>D. pseudoobscura</em></td>
<td>2,074,483</td>
<td>1.36</td>
<td>15</td>
<td>187</td>
</tr>
<tr>
<td>Sargasso Sea</td>
<td>1,660,141</td>
<td>1.47</td>
<td>28</td>
<td>199</td>
</tr>
</tbody>
</table>
influence the final assembly. Perhaps equally important, the proposed scheme fully automates assembly for large-scale sequencing projects when compared to the previous approaches that often require manual intervention for data partitioning, storing of intermediate results and running multiple programs.

An important contribution of our framework is the assembly of gene-enriched maize fragments, which are frequently being used by many plant scientists. Experiments indicate that the run-time behavior of our clustering solution shows good scaling. Our key contributions in space-optimality and a heuristic-based clustering scheme to significantly reduce alignment computations will play a crucial role in the large-scale applicability of our framework in the context of the maize genome and many other complex genomes of economically important plant crops. To give a perspective — our current implementation requires 80 bytes for every input nucleotide, implying that we can scale up to $\approx 8$ million fragments for every 1,024 BlueGene/L nodes (each with 512 MB). This would enable us to cluster 30 million fragments on about 4,000 nodes. Moreover, we conducted a few preliminary experiments on 8,192 nodes and the scaling results are encouraging. We believe that a continued improvement of our algorithmic techniques on large-scale parallel computers will provide a robust and efficient platform for many impending large-scale genome projects such as for sorghum and pine, which also involve gene-enrichment sequencing.

The effectiveness of our clustering approach can be further enhanced by resolving inconsistent overlaps during cluster formation. By reducing the largest cluster size, this will increase available parallelism during the assembly phase. Even with the limitations of single linkage clustering, the partitioning of original data is sufficient to allow assembly software to run on commonly available desktop machines. Recent advances in high throughput sequencing, such as the 454 GS20 sequencer that can sequence nearly 200,000 fragments in about 4 hours, are causing a significant gap between data generation and data processing capabilities. We believe parallel approaches such as the one presented here should become increasingly more important as high throughput sequencing techniques become mainstream.
Acknowledgments

We thank Sam Ellis, Kurt Pinnow and Brian Smith of IBM Rochester, for facilitating our access to the BlueGene/L supercomputer and commenting on the manuscript. This research was supported in part by NSF grant DBI-0527192 and an IBM Ph.D. Fellowship.

References


QUALITY ASSESSMENT OF MAIZE ASSEMBLED GENOMIC ISLANDS AND LARGE-SCALE EXPERIMENTAL VERIFICATION OF PREDICTED NOVEL GENES

A paper published in the Proceedings of the National Academy of Science, USA

Yan Fu\textsuperscript{2}, Scott J. Emrich\textsuperscript{3}, Ling Guo, Tsui-Jung Wen,
Daniel A. Ashlock, Srinivas Aluru, and Patrick S. Schnable

Abstract

Recent sequencing efforts have targeted the gene-rich regions of the maize (Zea mays L.) genome. We report the release of an improved assembly of maize assembled genomic islands (MAGIs). The 114,173 resulting contigs have been subjected to computational and physical quality assessments. Comparisons to the sequences of maize bacterial artificial chromosomes suggest that at least 97\% (160 of 165) of MAGIs are correctly assembled. Because the rates at which junction-testing PCR primers for genomic survey sequences (90–92\%) amplify genomic DNA are not significantly different from those of control primers (≈91\%), we conclude that a very high percentage of genic MAGIs accurately reflect the structure of the maize genome. EST alignments, \textit{ab initio} gene prediction, and sequence similarity searches of the MAGIs are available at the Iowa State University MAGI web site. This assembly contains 46,688 \textit{ab initio} predicted genes. The expression of almost half (628 of 1,369) of a sample of the predicted genes that lack expression evidence was validated by RT-PCR. Our analyses suggest that the

\textsuperscript{1}Reprinted with permission of \textit{PNAS}, 2005, 102:12282–12287
\textsuperscript{2}Primary author, performed all experimental validation
\textsuperscript{3}Primary author, performed all computational experiments and statistical analysis
maize genome contains between $\approx 33,000$ and $\approx 54,000$ expressed genes. Approximately 5% (32 of 628) of the maize transcripts discovered do not have detectable paralogs among maize ESTs or detectable homologs from other species in the GenBank NR nucleotide/protein database. Analyses therefore suggest that this assembly of the maize genome contains approximately 350 previously uncharacterized expressed genes. We hypothesize that these “orphans” evolved quickly during maize evolution and/or domestication.

**Introduction**

Maize (*Zea mays* L.) is the best-studied model for cereal biology and one of the world’s most important crops. Most of the maize genome consists of highly repetitive sequences; consequently, the genes in this plant comprise only 10–15% of its genomic DNA [Bennetzen et al. (2001); Martienssen et al. (2004)]. Because of its large repetitive fraction, the National Science Foundation funded the Maize Genomics Consortium to test two distinct filtration strategies for sequencing the “gene-rich” portion of the maize genome: methylation filtration (MF) and high $C_9t$ (HC) selection. To date, these pilot projects have generated and deposited into GenBank 450,166 MF sequences, 445,541 HC sequences, and 50,877 random shotgun sequences as genomic survey sequences (GSSs). MF and HC strategies have proven effective in selectively recovering maize genes not captured by EST projects [Whitelaw et al. (2003); Palmer et al. (2003)].

The assembly of these GSSs into genomic contigs significantly increases their utility. Our group developed a genome assembly pipeline based on innovative parallel algorithms that can quickly assemble hundreds of thousands of nonuniformly generated genomic fragments, such as MF and HC sequence reads, in a few hours [Emrich et al. (2004)]. A key advantage of our parallel genome assembly pipeline is that the speed with which assemblies can be generated allows experimentation on the assembly process *per se*. Specifically, this speed makes it possible to determine the effects of different assembly parameter values on the quality of the resulting assemblies.

Three research groups currently provide publicly available partial maize genome assemblies
based on the GSS data [The Institute for Genomic Research (TIGR), Plant Genome Database, and our group]. To our knowledge, none of these assemblies has been subjected to systematic studies into the quality of the resulting genomic contigs, nor have attempts been made to validate the structures of potentially novel maize genes found in these assemblies that have to date eluded discovery via the extensive maize EST projects. Structure validation will provide data that can be used to design strategies to assemble the maize genome [National Plant Genome Initiative (2005)].

The current study reports improvements to the quality of the sequence data used for assembly and the assembly pipeline used to generate our maize assembled genomic islands (MAGIs). Computational and biological quality assessments indicate that a high percentage of the MAGIs accurately reflect the structure of the maize genome. In addition, we estimate that this assembly of the maize gene space has “tagged” >6,900 expressed genes that previously lacked evidence of transcription and that almost 350 of these genes are “orphans”; i.e., they do not exhibit similarity to genes in other species. This large-scale application of RT-PCR for the verification of the expression of predicted monocot genes is a step to developing a framework for the subsequent annotation of the entire maize transcriptome. Based on the results of these RT-PCR experiments, we estimate that the B73 genome contains between ≈33,000 and ≈54,000 expressed genes.

Methods

Maize GSS retrieval, trimming and repeat masking

Genomic Survey Sequence (GSS) sequence and quality score files generated by the Maize Genome Sequencing Consortium (Danforth Center, TIGR, Purdue University, and Orion Genomics) from the Zea mays inbred line B73 were downloaded from the National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov/pub/TraceDB) in late September 2003. This untrimmed, raw dataset consisted of 880,404 fragments totaling 857 MB and was subsequently trimmed with LUCY [Chou and Holmes (2001)]. The trimming parameters used for these GSSs were Bracket [20 0.003], Window [10 0.01], and Error [0.005 0.002]. Approx-
imately 240,000 bacterial artificial chromosome (BAC) end reads were similarly downloaded from GenBank and processed to locate additional maize statistically defined repeats [Emrich et al. (2004)] that were used for repeat masking.

Validation of MAGI Assemblies by Using Sequenced Maize BACs

Sixteen entire maize B73 BAC sequences (GenBank accession nos. AC144717, AF448416, AF464738, AF466202, AF466203, AF466646, AF466931, AF546189, AY325816, AY371488, AY146791, AY180107, AY180106, AY271636, AY530952, and AY530951) downloaded from GenBank on July 24, 2004, were used as benchmarks to test the structures of MAGIs. These BACs were aligned with MAGIs by using BLASTN with the low complexity filter turned off. Only MAGIs that had BLAST alignments of ≥99% identity and alignment lengths of ≥400 bp were analyzed. The overlapping region between two BACs (accession nos. AY325816 and AF464738) resulted in five pairs of identical MAGI/BAC alignments. Only one member of each pair was analyzed.

RNA Isolation and Reverse Transcription

RNA samples of maize inbred line B73 were isolated from various treatments and/or tissues (see Supporting Materials and Methods, which is published as supporting information on the PNAS web site). First-strand cDNA was synthesized with SuperScript II reverse transcriptase (RT) with Oligo-dT priming (Invitrogen). The resulting product was then treated with DNase I (Invitrogen) and purified for PCR by following a previously described protocol that prevents genomic DNA contamination [Flohr et al. (2003)].

Touchdown PCR Amplification and Direct Sequencing of RT-PCR Products

Primers for genomic and RT-PCRs were designed with PRIMER3 (see Supporting Materials and Methods for details) [Rozen and Skaletsky (2000)]. For cDNA and genomic DNA templates, PCRs were incubated for 2 min at 92°C, followed by 10 cycles of denaturation at 94°C for 30s, annealing for 30s, and elongation at 72°C for 1 min and another 24 cycles of 94°C for
30s, 61 °C for 30s, and 72 °C for 1 min and a final 10-min extension at 72 °C. The annealing temperature was decreased by 0.8 °C per cycle during the first 10 cycles from 69 °C to 61 °C to increase the specificity of the amplification. PCR cleanup plates (Millipore) were used to purify PCR products for single RT bands. QIAquick spin columns (Qiagen, Valencia, CA) were used to purify individual bands for double RT bands. Each purified sample was sequenced from both directions. The sequences of the RT-PCR products are available from the authors upon request.

Gene Content Analyses

The sequences of the assembled B73 3’ Iowa State University Maize ESTs build [Yao et al. (2005)] were used to assess gene coverage by querying MAGIs with low \( (e^{-30}) \) and high \( (e^{-100}) \) stringency \( E \)-value criteria and with the low-complexity filter turned off. FGENESH (Softberry, Mount Kisco, NY) was used for \textit{ab initio} gene prediction with monocot parameters and the GC option that uses all potential GC donor splice sites [Yao et al. (2005)]. Evidence for the transcription of predicted gene models was obtained by querying another larger build of assembled maize transcripts (see \textit{Supporting Materials and Methods}) using BLASTN \( (E \)-value cutoff, \( e^{-10} \)).

Annotation of Maize Genes Without Evidence of Expression

The MAGI 3.1 assembly was initially screened against the Plant Genome Database maize tentative unique genes downloaded in September, 2003, using GENESEQER [Usuka et al. (2000)] as described by Yao et al. (2005). A sample of MAGIs that exhibited FGENESH predictions but that did not have GENESEQER EST alignments were subjected to RT-PCR. The predicted genes tested by RT-PCR were later compared with the above-mentioned assembly of all maize ESTs by using TBLASTN with a criterion of an \( E \)-value of \( \leq e^{-10} \). RT-PCR primer pairs designed from predicted genes that exhibited significant matches to maize transcripts were used as controls for RT efficiency within the mRNA sources used in this study. The remaining candidates were then run against TIGR plant Gene Indices (see \textit{Supporting
Materials and Methods for details). Significant matches were determined by using BLASTN and TBLASTN with a criterion of an $E$-value of $\leq e^{-10}$. The cDNA sequences of predicted genes without matches to the plant transcripts were also compared with the GenBank NR protein and nucleotide database (June 2005) with BLASTX and TBLASTX, respectively, and a very conservative $E$-value cutoff ($e^{-4}$). Predicted genes that exhibit matches only to maize and that did not align to transposons or annotated genes were deemed novel.

Display of MAGI Annotation

GBROWSE 1.61 was downloaded from the Generic Model Organism Database web site [Stein et al. (2002)] and installed on an Apple Mac OS 10.3 system. The CAP3 assembly output files [Huang and Madan (1999)], GENESEQER alignments using Iowa State University B73 assembled 3’ EST data [Yao et al. (2005)], FGENESH predictions, BLASTX hits ($E$-value cutoff, $e^{-10}$), and PRIMER3 results were parsed into GFF files by using PERL and AWK scripts. All GFF files were loaded into MYSQL database for GBROWSE display.

Results and Discussion

Assembly of MAGI Version 3.1

To assemble the maize gene space, it was necessary to develop a scalable solution that used mechanisms to minimize assembly artifacts caused by the presence of repetitive elements and that also accounted for the nonuniform sampling of the genome due to gene enrichment [Emrich et al. (2004); Springer et al. (2004)]. In our pipeline, sequences were cleaned, repeat-masked, and clustered by using PACE [Kalyanaraman et al. (2003)] based on defined overlap criteria. The sequences within clusters were then unmasked and assembled with CAP3 into one or more contigs. Relative to our prior maize genome assembly [Emrich et al. (2004)], the assembly presented here (MAGI 3.1) incorporates further improvements in the quality of the input sequences and the repeat masking process, and it uses clone pair information during clustering.
When assembling a genome sequenced with a shotgun cloning approach, sequence errors in the input data tend to “average out” if a sufficient degree of redundancy exists. As compared with the shotgun approach, nonuniform genome sampling approaches (e.g., MF and HC enrichment) could lead to higher rates of sequence errors within poorly sampled regions. Therefore, before the assembly of version 3.1, we conducted an analysis of a sample of publicly available MF and HC sequences to determine the sources and locations of sequencing errors relative to a benchmark set of 10 genes totaling $\approx 79$ kb of highly finished sequence [Fu et al. (2004)]. This study demonstrated that the average rate of errors per base in a sample of unassembled MF and HC GSSs could be reduced 6-fold (to $3.6 \times 10^{-4}$) by applying more stringent trimming parameters with minimal loss of gene content. These parameters were applied to all input sequences used in assembling MAGI Version 3.1.

Another of the improvements of MAGI 3.1 versus MAGI 2.3 was the use of an updated version of our nonredundant repeat database for repeat masking. Because repeats are overrepresented in the genome, they should also be overrepresented within a random sample of genomic fragments. Available BAC end sequences are not a random sample of the maize genome but are substantially more representative than sequences obtained by gene enrichment (e.g., MF or HC selection). Consequently, we first masked an updated collection of BAC end sequences by using known repeats to enrich for lower-copy repetitive sequences. These masked data were then subjected to single-linkage clustering to generate statistically defined repeats. This analysis resulted in the recovery of additional repetitive sequences, which were incorporated into version 2.0 of the MAGI repeat database. A larger fraction of unfiltered shotgun and BAC end data are classified as repetitive by using these new statistically defined repeats (74% versus 57.6%) relative to the previously reported repeat database (version 1.0), a value that better correlates with the estimated frequency of repetitive sequences in the maize genome [Meyers et al. (2001)].

The third improvement of the MAGI 3.1 pipeline over that of MAGI 2.3 relates to the use of clone pair information. Sequencing both ends of a cloned fragment of DNA generates two sequences with known physical proximity features; this information is especially useful to
help the assembler resolve highly similar repeats found in complex genomes. In our pipeline, paired sequences that contain at least 100 bases of nonrepetitive DNA are grouped together and provided to PACE as initial clusters, thereby preserving all relevant clone pair information. Although a large percentage of these PACE clusters yield single contigs, proximity constraints sometimes provide evidence that clusters should be split into two or more contigs during assembly. A comparison of this build to the previously reported MAGI 2.3 is presented in Table 1. Both builds are available from the authors upon request.

Quality Assessment of MAGIs

A combination of computational and wet-laboratory approaches (illustrated in Figure 1) was developed to assess the quality of our current partial maize genome assembly. In the following sections, we demonstrate that the contigs in the latest MAGI assembly are of high quality.

MAGI Validation: Comparisons to BAC Sequences

The sequences of 16 published maize B73 BACs were used as a benchmark for validating the structures of MAGIs. A BLAST search returned 173 nonredundant alignments between these 16 BACs and MAGIs. To determine whether these alignments verify the structure of a MAGI,
Figure 1 Illustrations of computational and wet-laboratory strategies used for MAGI validation. (A) The consistency of MAGIs was assayed via alignment to maize B73 BACs. A set of potential MAGI/BAC alignments was identified by using BLAST (see Methods). The dashed lines mark portions of the MAGI that fail to match the BAC sequence. MAGIs were deemed to be inconsistent if they had a total overhang length (combined length of dashed lines) of >20 bp. The overhangs associated with four of the six consistent MAGI/BAC pairs that have sizes of between 6 and 20 bases can be recognized as incompletely trimmed vector sequences on a terminal GSS of a MAGI (Table 5). Four of the consistent MAGI/BAC pairs have overhangs of <6 bases, which may also be derived from incompletely trimmed vector. Terminal MAGI/BAC alignments of the type shown on the right do not provide evidence of inconsistency. Six such cases were identified. (B) Comparison of genomic PCR success rates: Within a MAGI, each primer pair annealed to the same GSS (set 1), two GSSs from the same clone (set 2), or two GSSs from different clones (set 3). Set 1 primer pairs served as a control to assess the success of primer design and PCR. Sets 2 and 3 primer pairs were used to validate the structure of MAGIs.
we define the concept of consistency (Figure 1A). Consistent MAGI/BAC pairs contain no more than 20 bases of a MAGI that do not align to the BAC (i.e., the sum of the two potential overhangs).

We excluded from subsequent analyses the eight inconsistent alignments that involved only a single GSS within a MAGI because these alignments do not test fragment assembly errors and are instead most likely due to misalignment of repetitive sequences. Indeed, all eight alignments of this type included repetitive sequences contained with the TIGR repeat database 4.0 (Table 5, which is published as supporting information on the PNAS web site).

Removing these eight inconsistent repetitive alignments left 165 MAGI/BAC alignments for validation (Table 5). Of these 165 MAGI/BAC alignments, 95.2% (157) are consistent. Because we observed evidence of the collapse of Near Identical Paralogs (NIPs) in the MAGI 2.3 build [Emrich et al. (2004)], we hypothesized that at least some of the eight inconsistent MAGIs detected in the current study could also have arisen via the collapse of NIPs into a single MAGI. Potential paramorphisms (polymorphisms between paralogs) in GSSs that comprise a MAGI have been reported previously [Emrich et al. (2004); Fu et al. (2004)]. Consequently, the trace files of the GSSs used to assemble each of the eight inconsistent MAGIs were examined manually. The GSSs associated with four inconsistent MAGIs are 100% identical and therefore exhibit no evidence of NIP collapse (Table 6, which is published as supporting information
on the PNAS web site). The structures of three of these inconsistent MAGIs (nos. 41789, 84169, and 107229) were validated by genomic PCR (data not shown). Hence, the apparent inconsistencies associated with these three MAGIs appear to be a consequence of aligning MAGIs to highly similar but inappropriate BACs. The origin of the fourth inconsistent MAGI (no. 53496) is not known. In contrast, and consistent with the hypothesis that at least some of the inconsistent MAGIs arise because of NIP collapse, the GSSs used to assemble the remaining four inconsistent MAGIs (nos. 8097, 22812, 39419, and 89783) exhibited at least one putative paramorphism (Table 6). In the case of MAGI 89783, which encodes cis-zeatin O-glucosyltransferase, this hypothesis regarding the origin of inconsistent MAGIs is further supported by the presence in the maize inbred line B73 of two highly similar (98.3% nucleotide identity) cis-zeatin O-glucosyltransferase genes, ciszog1 and ciszog2 (accession nos. AF318075 and AY082660) [Veach et al. (2003)]. Significantly, the putative paramorphisms observed in the GSSs that comprise MAGI 89783 match those that distinguish ciszog1 and ciszog2 (Figure 2). Further support for the hypothesis that at least some of the inconsistent MAGIs arise via NIP collapse is provided by the observation that the rate of inconsistent MAGI/BAC alignments that contain putative paramorphisms (2.4%, 4 of 165) is similar to the observed rate of NIPs in the maize genome, i.e., \( \approx 1\% \) (S.J.E., T.-J.W., M. D. Yandeau-Nelson, Y.F., L. Li, L.G., H.-H. Chou, S.A., D.A.A., and P.S.S., unpublished data). These results suggest that the misassembly can be caused not only by highly homologous transposons but also by nearly identical nontransposon genes. The prevention of the misassembly in future assemblies of the maize genome will require access to very high-quality sequence data and the application of stringent assembly parameters.

**MAGI Validation: Genomic PCR**

The computational analyses described above suggest that, at minimum, \( \approx 97\% \) of MAGIs are correctly assembled. This observation is based on the hypothesis that if two independent assemblies (BAC and MAGI) agree, both are most likely correct. Note, however, that this is a conservative estimate; inconsistent MAGI/BAC alignments could also arise because of
biological idiosyncrasies within the maize genome. To provide an estimate that incorporates such uncertainty, PCR amplification was used to independently estimate the proportion of MAGI assemblies that accurately reflect the structure of the maize genome (Figure 1B). To first estimate the rate of false-negative PCR amplification, pairs of control primers that span predicted introns were designed that anneal to a single GSS (Figure 1B, set 1; see also Methods). Each of these pairs of primers was used to conduct touchdown PCR on genomic DNA from the inbred line B73. As shown in Table 7, which is published as supporting information on the PNAS web site, ≈86% (1,165 of 1,358) of these control primers yielded a single PCR product of the size expected based on the positions at which the primers anneal to the GSS. Another 5% (68 of 1,358) of the control primers yielded a doublet PCR product, one of which was the expected size. Consistent with the structure of the maize genome [Blanc and Wolfe (2004)], these doublets probably arise via the amplification of pairs of paralogous sequences. PCR failures [i.e., primer pairs that yielded either no band (6%) or multiple bands/smears (3%)] probably reflect problems in primer design, e.g., attempts to amplify multigene families.

Junction-testing primers were used to experimentally determine the quality of MAGIs. Pairs of junction-testing primers are those in which each member of a primer pair can anneal to either different GSSs within the same clone (Figure 1B, set 2) or to different clones in a single MAGI (Figure 1B, set 3). As such, these primer pairs can be used to test the assembly junctions of the GSSs that comprise a given MAGI. Approximately 90.9% (512 of 563; Figure 1B, set 2) and 92.5% (99 of 107; Figure 1B, set 3) of the junction-testing primer pairs yielded a single or doublet PCR product of the expected size (Table 7). Hence, the success rates of the junction-testing primers (90–92%) are similar to that of the control primers (≈91%). Based on a Z test for difference of two proportions, there is no statistical support for the hypothesis that the success rates of these classes of primers differ. We therefore conclude that a very high percentage of the GSS junctions reported in genic MAGIs are correct (i.e., they accurately reflect the structure of the maize genome).
Sequence Fidelity of MAGIs

By aligning GSSs to a benchmark set of 10 genes totaling ≈79 kb of highly finished sequence identified trimming parameters that reduced the rate of sequencing errors in a sample of GSSs from $2.3 \times 10^{-3}$ to $3.6 \times 10^{-4}$ [Fu et al. (2004)]. As mentioned above, these trimming parameters were used in the MAGI 3.1 build. We report here that the MAGIs corresponding to these 10 control genes have a sequencing error rate of $1 \times 10^{-4}$. The reduction in the rate of sequencing errors observed in MAGIs relative to GSSs is probably a consequence of the resampling of some base positions within MAGIs as compared with single-pass GSSs. About half (82 of 165) of the consistent MAGI/BAC alignments described above exhibit 100% identity, and only 213 bp of 274,689 bp ($7.7 \times 10^{-4}$) within consistent alignments exhibit disagreements between the sequences of a MAGI and its respective BAC. The almost 8-fold difference between the estimated rates of sequencing errors in MAGIs obtained through alignments to BACs ($7.7 \times 10^{-4}$) and alignments to the set of 10 control genes (i.e., $1 \times 10^{-4}$) may reflect higher sequencing errors in the BACs or the inappropriate alignment of a MAGI to a BAC that contains a NIP of a gene present in that MAGI.

Genic Content of MAGIs

Determining how successfully the MF and HC filtration strategies have sampled the gene space of the maize genome is complicated by the fact that a complete inventory of maize genes is not available. Even so, several computational experiments suggest that the MF and HC GSSs have captured a large fraction of the maize gene space. For example, these GSSs have been shown to tag all members of small collections of known maize genes [Springer et al. (2004); Fu et al. (2004)]. In addition, ≈11% of the contigs in an assembly consisting of approximately one-fifth of the GSSs used in the MAGI 3.1 assembly exhibit similarity (BLAT settings: 95% identity and ≥20% of contig length) to the TIGR Plant Gene Index (3). Furthermore, ≈560,000 MF GSSs exhibit similarity to ≈65% of the nonrepeat, nonhypothetical maize genes detected on published BACs (BLAT settings: 98% identity and ≥90% of read length) [Whitelaw et al. (2003)].
Table 2  FGENESH-derived gene prediction in all 114,173 MAGIs

<table>
<thead>
<tr>
<th>Type of predictions</th>
<th>Total (%)^a</th>
<th>With transcription evidence in maize^b (%)^c</th>
<th>Containing repeats^d (%)^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete gene models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With intron</td>
<td>13,597 (29.1)</td>
<td>9,096 (66.9)</td>
<td>1,423 (10.5)</td>
</tr>
<tr>
<td>Without intron</td>
<td>6,638 (14.2)</td>
<td>3,918 (59.0)</td>
<td>770 (11.6)</td>
</tr>
<tr>
<td>Subtotal</td>
<td>20,235 (43.3)</td>
<td>13,014 (64.3)</td>
<td>2,193 (10.8)</td>
</tr>
<tr>
<td>Incomplete gene models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacking first exon</td>
<td>10,937 (23.4)</td>
<td>8,085 (73.4)</td>
<td>1,477 (13.5)</td>
</tr>
<tr>
<td>Lacking last exon</td>
<td>10,861 (23.3)</td>
<td>6,268 (57.7)</td>
<td>1,491 (13.7)</td>
</tr>
<tr>
<td>Lacking first and last exon</td>
<td>4,655 (10.0)</td>
<td>3,228 (69.3)</td>
<td>715 (15.4)</td>
</tr>
<tr>
<td>Subtotal</td>
<td>26,453 (56.7)</td>
<td>17,581 (66.5)</td>
<td>3,683 (13.9)</td>
</tr>
<tr>
<td>Total no. of predictions</td>
<td>46,688 (100)</td>
<td>30,595 (65.5)</td>
<td>5,876 (12.6)</td>
</tr>
</tbody>
</table>

^a^The percentage of indicated types of predicted gene models/total number of gene predictions.  
^b^Predicted transcript matches either a maize expressed gene or maize cDNA sequence (BLASTN; E-value cut-off, e^{-10}).  
^c^The percentage of predictions that contain the indicated type of database match/number of the indicated type of gene model predictions.  
^d^Each predicted coding sequence was screened against the nucleotide MAGI repeat database using BLASTN (E-value cutoff, e^{-10}). Predictions with at least one database match were deemed to be repetitive.

To estimate gene coverage within our MAGI 3.1 assembly, we used a set of assembled 3’ reads of maize ESTs from the inbred B73 that presumably corresponds to unique genes [Yao et al. (2005)]. Of the 19,454 unigenes in this set, 14,606 (76%) match at least one MAGI using BLAST with a stringent E-value cutoff of e^{-100}. Although it is not possible to directly compare these results to the previously reported estimates because of differences in algorithms and significance criteria, it is clear that the MAGIs contain a high percentage of known maize genes.

Genes can be detected not only by means of alignments to the sequences of known genes as was done above but also by ab initio gene prediction software. We previously used a set of >1,300 maize gene sequences to compare the performance of three ab initio gene prediction programs (FGENESH, GENEMARK.HMM, and GENSCAN), each of which had been trained on maize. In this analysis, FGENESH performed the best, although GENEMARK.HMM also performed well [Yao et al. (2005)]. These results are consistent with the observation that FGENESH was the most successful program for gene prediction in rice [Yu et al. (2002)]. With
the 114,173 MAGIs as input, FGENESH returned 46,688 gene predictions, of which only \( \approx 13\% \) contained repetitive sequences (Table 2). Approximately 34\% (16,093) of the predicted cDNAs had no hits against assembled maize ESTs or maize cDNAs (see Materials and Methods). As an additional measure of gene content, another 9,323 MAGIs did not contain a prediction but did exhibit similarity to known ESTs and/or proteins. Hence, >47\% of all MAGIs in build 3.1 contain a gene or predicted gene.

**Display of Annotated MAGIs**

Annotated MAGIs can be viewed at the Iowa State University MAGI web site. An example is shown in Figure 3, which is published as supporting information on the PNAS web site. Layouts of individual GSSs from parsed CAP3 output are color-coded for convenience. Sequence-based annotations against protein databases were performed with BLASTX against the Protein Information Resource International Protein Database (version 79.00). Gene structures predicted by FGENESH and GENESEQER are also displayed (Methods). Primers used in this study were also entered into the MAGI 3.1 GBROWSE database. The entire membership of this assembly can be downloaded along with the contigs per se.

**RT-PCR Validation of Predicted Transcripts**

As discussed above, FGENESH analysis of MAGIs resulted in the prediction of \( \approx 16,100 \) genes that do not match known maize transcripts. We designed pairs of intron-spanning primers to test whether 1,590 of these *ab initio* predicted novel genes are transcribed. Another batch of 438 pairs of primers from *ab initio* predictions that do have significant BLAST hits to maize transcripts were also designed as a control. Because paralogs and nonspecific amplification can complicate the verification of putative genes by RT-PCR, we tested each pair of primers by conducting PCR on B73 genomic DNA. Approximately 86\% (1,737 of 2,028) of these reactions yielded single genomic PCR bands of the size expected based on the positions at which the primers anneal to the corresponding MAGI (Table 7). The rates at which primers designed to amplify predicted genes with and without transcription evidence amplified single
Table 3  RT-PCR results for all primer pairs that yielded a single genomic PCR band

<table>
<thead>
<tr>
<th>RT-PCRs</th>
<th>Band pattern</th>
<th>BLAST results$^a$, n (%)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>Total, n (%)</td>
</tr>
<tr>
<td>RT-PCR-positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 band</td>
<td>&lt;</td>
<td>125 (34.0)</td>
<td>370 (27.0)</td>
<td>495 (28.5)</td>
</tr>
<tr>
<td></td>
<td>=</td>
<td>32 (8.7)</td>
<td>165 (12.1)</td>
<td>197 (11.3)</td>
</tr>
<tr>
<td>2 bands</td>
<td>&lt;&lt;</td>
<td>14 (3.8)</td>
<td>35 (2.5)</td>
<td>49 (2.8)</td>
</tr>
<tr>
<td></td>
<td>≤</td>
<td>18 (4.9)</td>
<td>58 (4.2)</td>
<td>76 (4.4)</td>
</tr>
<tr>
<td>Subtotal</td>
<td></td>
<td>189 (51.4)</td>
<td>628 (45.9)</td>
<td>817 (47.0)</td>
</tr>
<tr>
<td>RT-PCR-negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No band</td>
<td></td>
<td>134 (36.4)</td>
<td>582 (42.5)</td>
<td>716 (41.2)</td>
</tr>
<tr>
<td>1 band</td>
<td>&gt;</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2 bands</td>
<td>≥</td>
<td>5</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>Others$^b$</td>
<td></td>
<td>40 (10.9)</td>
<td>125 (9.1)</td>
<td>165 (9.5)</td>
</tr>
<tr>
<td>Subtotal</td>
<td></td>
<td>179 (48.6)</td>
<td>741 (54.1)</td>
<td>920 (53.0)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>368 (100)</td>
<td>1,369 (100)</td>
<td>1,737 (100)</td>
</tr>
</tbody>
</table>

$^a$BLAST results indicate primer pairs derived from predicted genes that do (+) or do not (-) have significant BLASTIN ($E$-value cutoff, $e^{-10}$) hits against all maize transcripts. <, The RT PCR band is smaller than genomic PCR band; =, the RT-PCR band is the same size as the genomic PCR band; <<, both RT-PCR bands are smaller than the genomic PCR band; ≤, one RT-PCR band is smaller than genomic PCR band and the other one is the same size as the genomic PCR band; >, the RT-PCR band is larger than the genomic PCR band. Sequence analyses established that five of five RT-PCR products of this type do not exhibit similarity to the predicted genes from which the PCR primers were designed; ≥, at least one of the two RT PCR bands is larger than the genomic PCR band.

$^b$The gel analyses of RT products yielded more than two visible bands or a smear.

PCR products were similar: 84% (368 of 438) and 85% (1,369 of 1,590), respectively. The 1,737 primer pairs were also subjected to RT-PCR using a diverse cDNA pool as template (Methods). Reactions that yielded single bands that were smaller than or equal in size to the PCR product from genomic DNA template or that yielded double bands, one of which was smaller than or equal in size to the PCR product from genomic DNA template, were considered RT-positive. Reactions that yielded any other outcomes were deemed RT-negative. Approximately 51% (189 of 368) of the BLASTN-positive set and 46% (628 of 1,369) of the BLASTN-negative set of Table 3 were RT-positive.

To determine the specificity of these RT reactions, we sequenced >160 PCR products from RT-positive reactions (Table 8, which is published as supporting information on the PNAS web site). These analyses demonstrated that ≈94% of these RT products were derived from
Table 4  Evidence of transcription of FGENESH predicted genes

<table>
<thead>
<tr>
<th></th>
<th>Maize transcripts</th>
<th>Plant transcripts</th>
<th>NR databases databases</th>
<th>Maize matches only</th>
<th>Transposons</th>
<th>Novel</th>
<th>Total (a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT+</td>
<td>256</td>
<td>300</td>
<td>37</td>
<td>3</td>
<td>32</td>
<td>628</td>
<td></td>
</tr>
<tr>
<td>RT</td>
<td>296</td>
<td>236</td>
<td>66</td>
<td>11</td>
<td>132</td>
<td>741</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>552</td>
<td>536</td>
<td>103</td>
<td>14</td>
<td>164</td>
<td>1,369</td>
<td></td>
</tr>
</tbody>
</table>

\(a\)Maize transcript values show TBLASTIN hits against maize transcripts (see Supporting Materials and Methods) with an \(E\)-value cutoff of \(e^{10}\). Values for the plant transcripts show BLASTIN and TBLASTIN hits against TIGR plant gene indices (see Supporting Materials and Methods), with an \(E\)-value cutoff of \(e^{10}\). Values for the NR databases show BLASTX and TBLASTX hits against the GenBank NR nucleotide database and protein database, respectively, with \(e^4\) as the \(E\)-value cutoff. The NR databases column does not include predicted genes that match only maize sequences. Most of the 103 predicted genes in this column match cereal retroelements. Entries in the maize and plant transcripts and NR databases columns did not exhibit matches to the databases shown to the left. For example, the 536 sequences with BLAST hits to plant transcripts did not exhibit matches to maize transcripts. Predicted genes that exhibited matches only to maize entries in the GenBank NR database and that did not align with transposons were deemed novel. RT+, RT-positive; RT-, RT-negative.

the predicted genes from which the primers had been designed (data not shown). Thus, it was possible to verify the expression of 43\% \([94\% \times (628/1,369)]\) of predicted genes that lack evidence of transcription in maize \((i.e., the BLASTN-negative set in Table 3)\). Consequently, the MAGIs have probably “tagged” >6,900 \([43\% \times (46,688-30,595)]\) (Table 2) expressed genes that previously lacked evidence of transcription. Because only half of the control genes for which evidence of transcription already exists in maize (the BLASTN-positive set in Table 3) were RT-positive in this experiment, we conclude that our RT-PCRs did not sample the entire maize transcriptome. Hence, our estimate of the number of predicted genes that are expressed is highly conservative.

Annotation of RT-Validated Genes

Of the 628 RT-positive predicted genes that previously lacked evidence for expression in the maize transcriptome, 256 (41\%) exhibit significant TBLASTN hits to maize transcripts (Table 4) and are therefore probably paralogs of maize genes for which evidence of transcription exists. Another 337 (300 + 37; 54\%) of the remaining genes exhibit significant similarity to plant transcripts and nonmaize genes or proteins in the GenBank NR DNA/protein databases \((Methods)\).
Significantly, after carefully removing sequences that exhibit similarity to transposons that are often responsible for overestimations of gene numbers in complex plant genomes [Bennetzen et al. (2004)], >5% (32 of 628) of the RT-positive predicted genes are novel based on very conservative criteria (Table 4; see also Table 9, which is published as supporting information on the PNAS web site). In all, ≈12% (164 of 1,369) of the predicted genes are novel and the expression of 20% (32 of 164) of the novel genes could be verified by RT-PCR experiments. Hence, the MAGIs are conservatively expected to contain ≈350 expressed novel genes or orphans [94% x (32/1,369) x (46,688-30,595)].

**Estimation of the Number of Maize Genes**

Based on available EST data ≈30,600 of the ≈46,700 predicted gene models in our assembly are expressed; moreover, we have shown that RT-PCR can conservatively validate the expression of ≈40% (94% x 46%) of the remaining ≈16,100 gene models (Tables 2 and 3). Taken together, these results imply that our partial maize genome assembly contains at least 37,100 genes [30,600 + (40% x 16,100)]. It is, however, possible that some of the 26,453 incomplete gene models in Table 2 do not represent unique genes. A more conservative estimate of the number of maize genes is therefore provided by considering only gene models that contain a last exon (and which could therefore be detected in our set of 3’ EST unigenes) and for which there is evidence of expression (21,099 = 13,014 + 8,085) plus the at least 40% of gene models that lack expression evidence but would be confirmed via RT-PCR experiments based on our experience (4,071 = 40% x 10,073). Dividing this sum (25,170 = 21,099 + 4,071) by the 76% of 3’ unigenes that can be identified among the MAGIs (E = e^{-100}) yields a lower bound of ≈33,000 genes. If we assume each nonrepetitive gene model from Table 2 is unique and expressed (40,812 = 46,688-5,876) and divide by 76%, the upper bound for the number of nonrepetitive genes in the maize genome is ≈54,000.
Conclusion

The gene enrichment strategies that have been validated by using the maize genome are likely to be applied to the genomes of other large-genome plants. Indeed, preliminary enrichment projects have already been reported for the wheat [Li et al. (2004)] and sorghum [Bedell et al. (2005)] genomes, and a gene enrichment project has been funded for pine. The assembly of the nonuniform genomic fragments that are generated by gene enrichment strategies poses special challenges, which we have addressed previously [Emrich et al. (2004)].

The current study provides two metrics (one strictly computational and the other based on large-scale PCR experiments) by which the quality of genome assemblies can be evaluated. Applying these metrics to our partial maize genome assembly demonstrates that gene-enriched sequences can be assembled into high quality contigs that facilitate biological discovery. For example, the application of large-scale RT-PCR using primers designed based on MAGIs made it possible to obtain expression data for hundreds of predicted genes.

Interestingly, these experiments also uncovered evidence for the existence of \( \approx 350 \) expressed maize genes that do not have homologs in other species. We hypothesize that these orphans are quickly evolving genes that played important roles during maize evolution and/or domestication. As such, these orphans present attractive targets for reverse genetics experiments.

Acknowledgements

We thank Jia Yi for help with RNA isolation and pilot experiments and two anonymous reviewers for useful suggestions. This research was funded in part by National Science Foundation Plant Genome Program Competitive Grants DBI-9975868 and DBI-0321711 and by the Hatch Act and funds from the state of Iowa. S.J.E. was supported in part by National Science Foundation Integrative Graduate Education and Research Traineeship fellowship DGE-9972653.
References


NEARLY IDENTICAL PARALOGS: IMPLICATIONS FOR MAIZE (ZEA MAYS L.) GENOME EVOLUTION

A paper published in Genetics¹

Scott J. Emrich, Li Li, Tsui-Jung Wen, Marna D. Yandeau-Nelson, Yan Fu, Ling Guo, Hui-Hsien Chou, Srinivas Aluru, Daniel A. Ashlock, and Patrick S. Schnable

Abstract

As an ancient segmental tetraploid, the maize (Zea mays L.) genome contains large numbers of paralogs that are expected to have diverged by a minimum of 10% over time. Nearly identical paralogs (NIPs) are defined as paralogous genes that exhibit ≥98% identity. Sequence analyses of the “gene space” of the maize inbred line B73 genome, coupled with wet lab validation, have revealed that, conservatively, at least ≈1% of maize genes have a NIP, a rate substantially higher than that in Arabidopsis. In most instances, both members of maize NIP pairs are expressed and are therefore at least potentially functional. Of evolutionary significance, members of many NIP families also exhibit differential expression. The finding that some families of maize NIPs are closely linked genetically while others are genetically unlinked is consistent with multiple modes of origin. NIPs provide a mechanism for the maize genome to circumvent the inherent limitation that diploid genomes can carry at most two “alleles” per “locus.” As such, NIPs may have played important roles during the evolution and domestication of maize and may contribute to the success of long-term selection experiments in this important crop species.

¹Reprinted with permission of Genetics, 2007, 175:429–439
Introduction

The grasses (Poaceae) are a highly adaptable family of monocotyledonous plants that have been independently domesticated by several human civilizations. Maize (Zea mays L.) is a hypothesized ancient segmental tetraploid, and it is estimated that nearly one-third of all modern maize genes have a paralogous sequence [Blanc and Wolfe (2004)]. More recently, the expected divergence of the segmental allotetraploid event has been revised from the original 15–30% [Gaut and Doebley (1997)] to 10–20% [Blanc and Wolfe (2004)] on the basis of maize ESTs.

Genomewide duplications are generally believed to provide raw material for evolutionary innovation [Ohno (1970)] and as such they have played important roles in the evolution of both plants and vertebrates (reviewed by Durand (2003); Moore and Purugganan (2005)). In contrast to the diverged paralogs produced via ancient duplications, detailed analyses of the human genome have identified nearly identical sequences that were inadvertently collapsed, or condensed into a single contiguous region, during genome assembly [Bailey et al. (2002); Cheung et al. (2003); She et al. (2004)].

Tandem duplications are common among plant species [Zhang and Gaut (2003)]. Indeed, Messing et al. (2004) have estimated that approximately one-third of maize genes are tandemly duplicated. Few of these tandem duplications are similar enough that they would collapse during genome assembly. Several tandem duplications of maize have been well characterized, including, \( R-r \) [Robbins et al. (1991)], \( Rp1 \) [Richter et al. (1995)], \( P1 \) [Zhang et al. (2005)], and \( A1-b \) [Yandeau-Nelson et al. (2006)]. Such duplications can be generated via unequal recombination [Richter et al. (1995); Yandeau-Nelson et al. (2006)]. In contrast, the transposition of Mu-like transposons in rice [Pack-MULEs; Jiang et al. (2004); Juretic et al. (2005)] and Helitrons in maize [Lal et al. (2003); Brunner et al. (2005); Lai et al. (2005); Lal and Hannah (2005); Morgante et al. (2005)], which have incorporated fragments of unrelated genes, can generate dispersed genic duplications. Although as many as 11% of all maize gene fragments are unique to a specific inbred line [Morgante et al. (2005)], the extent to which these gene duplications are functional is not known.
Because the maize inbred line B73 is homozygous at essentially all loci and its “gene space” has been extensively sequenced, it is an ideal candidate for beginning to study the extent, causes, and evolutionary significance of recent duplications in this complex genome. Toward this end, assemblies of B73 ESTs and gene-enriched Genome Survey Sequences (GSSs) were examined for the appearance of “polymorphic” nucleotide positions, which we term candidate paramorphisms [CPs; Emrich et al. (2004); Fu et al. (2005)]. If a specific CP site is not due to a sequencing error or residual heterozygosity, we term this site a paramorphism [PM; Fu et al. (2004)]. A paramorphism provides evidence of the existence of highly similar genomic loci and is strong evidence of a recent duplication without respect to the underlying duplication mechanism. We have termed a subset of such regions nearly identical paralogs (NIPs) if they exhibit ≥98% identity, are genic, and are not transposons or other repetitive sequences.

On the basis of highly conservative criteria, we estimate that ≈1% of genes in the B73 maize genome have at least one NIP, and nearly all of these exhibit >99% identity. In addition, we determined that many of these highly similar loci in the maize genome are genetically linked. Because Mu elements do not preferentially move to linked sites [Lisch et al. (1995)], this result implies either that Helitrons preferentially insert into neighboring locations or that other mechanisms were involved in the origins of these genetically linked NIPs. The observed frequency of NIPs is substantially higher in maize than in the model dicotyledon, Arabidopsis thaliana, suggesting that this phenomenon is not universal in plants. Most importantly, we also report that members of many NIP families are differentially expressed. We hypothesize that the high frequency of NIPs in combination with their diverse expression patterns may have provided a selective advantage during the domestication and the genetic improvement of maize by classical plant breeders and may play a fundamental role in the success of long-term selection experiments (e.g., Laurie et al. (2004)).
Methods

Locating and validating NIPs in collections of maize ESTs and GSSs

EST sequences were generated from three B73 cDNA libraries constructed by Fang Qiu (Iowa State University) with the advice of the Bento Soares laboratory (University of Iowa). A total of 32,229 EST sequences and their corresponding trace files were deposited in GenBank after removing short inserts and other irregularities. These B73 EST sequences were first assembled with CAP3 [Huang and Madan (1999)] using >98% similarity in detected overlaps, a minimum overlap size of 50 bp, and 60 bp as the clipping parameter. Potential NIPs were then identified by detecting contigs with CPs composed of at least two different nucleotides, each of which is supported by two independent EST reads, within CAP3 multiple sequence alignments.

We later endeavored to locate NIPs within “gene-enriched” maize genomic data [Palmer et al. (2003); Whitelaw et al. (2003)] using an updated version of our maize assembled genomic islands [MAGIs; Emrich et al. (2004); Fu et al. (2005)]. We use the same CP-detection heuristic described above for EST NIPs, but we restricted these analyses to only methyl-filtered (MF) clones because ≈40% of current high-CT clones contain cloning artifacts [Fu et al. (2004)]. In addition, we required that each CP variant be supported by at least two independent MF clones. On the basis of the criteria used to assemble the MAGIs [Fu et al. (2005)], only CP-competent intervals that exhibit ≥98% identity are recovered.

Even with the conservative criteria described above, it was possible that some CPs resulted from sequencing errors. Primer3 [Rozen and Skaletsky (2000)] was used to design primers ≈250 bp from each side of targeted CP sites. Genomic DNA was isolated from B73 seedling leaves using the protocol of Dietrich et al. (2002) and was PCR amplified using these CP-flanking primers. The resulting PCR products were analyzed via agarose gel electrophoresis. Single-band PCR products were then subjected to direct sequencing using the same CP-flanking PCR primers or were subcloned using a TOPO TA cloning kit (Invitrogen, Carlsbad, CA) followed by sequencing with the T7 and T3 primers.
Annotation of NIPs:

GBrowse (V1.61) was downloaded from the Generic Model Organism Database website and installed using a MySQL database at its core. The CAP3 assembly output files, CP-competent intervals, CP sites, primers used to validate CPs, GeneSeqer alignments (at least one exon of similarity of $\geq 95\%$ identity, $\geq 50$ bp length), FGENESH predictions, and BLASTX hits (PIR-PSD v.79.00; $E$-value $\leq 1e^{-10}$) were converted into GFF files using PERL and AWK scripts for display on the MAGI website (http://magi.plantgenomics.iastate.edu). CP-competent intervals were deemed genic if the MAGI contained a nonrepetitive gene model within 500 bp of the CP prediction. Repetitive models were excluded on the basis of protein matches to well-characterized transposons in GenBank.

NIP expression assays

Forty-six validated MAGI-NIPs with at least one predicted exon were analyzed; 42 yielded a single genomic PCR band with the expected size. These were then subjected to touchdown RT-PCR using the pooled inbred line B73 cDNA, very similar to that described previously [Fu et al. (2005)]. In addition, RNA samples were also isolated from various tissues, organs, and developmental stages of the B73 inbred line similar to those described by Qiu et al. (2003). Reactions that yielded single bands that were not larger than the genomic PCR product were sequenced. If the sequence of a RT-PCR product had a double peak at the paramorphic site, we concluded that both members of the NIP family are expressed. If in a given source of RNA only a single peak was observed at a paramorphic site, we concluded that only that member was expressed in that sample. Only if identical results were obtained from two independent biological replications did we conclude that the two members of a NIP family were differentially expressed. In almost all instances, the results from the two replications were consistent.

Genetic mapping of NIPs

NIPs were genetically mapped using 91 recombinant inbreds (RIs) of the intermated B73 x Mo17 (IBM) mapping population [Lee et al. (2002)]. CP validation primers that amplified B73
but not Mo17 DNA templates (*i.e.*, plus/minus markers) were identified via gel electrophoresis. If a pair of NIPs is tightly linked genetically, the RIs will segregate 1:1 for the presence and absence of the B73-derived PCR product; conversely, if a pair of NIPs is unlinked genetically, the RIs will segregate 3:1 for the presence and absence of the B73-derived PCR product. NIPs with segregation ratios that fall between 1:1 and 3:1 were deemed to be loosely linked genetically. To position the tightly linked NIPs on the genetic map, the RI genotype scores for each NIP-derived marker were directly compared to the RI scores of all of the ≈3500 genetic markers on a genetic map developed by us [IBM_IDP+MMPmap4; Fu et al. (2006)].

**Locating NIPs within Arabidopsis**

A total of 190,978 *A. thaliana* ESTs were downloaded from dbEST (GenBank) in June 2004, and 50 bp were trimmed from each end to reduce false positives associated with low-quality sequences. These ESTs were then clustered using PaCE [Kalyanaraman et al. (2003)] under default parameters, and contigs were generated using CAP3 from each resulting cluster as previously described. Polymorphic sites with representation in ≥25% of participating ESTs, which also violated random expectation for sequencing errors (*P* < 0.01), were selected; 28 primer pairs were designed to flank the 24 previously unreported duplications using Primer3. Successful reactions, which yielded a single band (*N* = 25), were sequenced and the corresponding trace files were analyzed.

In addition, all 68 low-copy Arabidopsis gene pairs that have rates of synonymous substitution (*K*<sub>a</sub>) <2% [Lynch and Conery (2000); Moore and Purugganan (2003)] were analyzed. Using the 02/28/2004 Arabidopsis gene annotation from The Arabidopsis Information Resource (http://www.arabidopsis.org), each potential NIP pair was checked to ensure that both members were genic and were annotated as distinct loci. Pairs that met these initial criteria were then compared using BLAST; candidates without a highly similar (>98% identity) continuous alignment were manually aligned and validated where possible. The genetic distances between members of a NIP family were determined by multiplying the physical distance that separates them by the centimorgan/megabase values reported by Zhang and Gaut (2003).
Results

**In silico detection of maize NIPs**

Nearly identical sequences are subject to being erroneously “collapsed” into single sequences during genome assembly. Collapsed segmental duplications within the human genome assembly were identified by virtue of their overrepresentation among randomly generated sequences [Bailey et al. (2002)], and it has been estimated that >8% of public human single nucleotide polymorphisms (SNPs) are potentially paramorphisms rather than actual SNPs [Cheung et al. (2003)].

Evidence for the existence of NIPs in the inbred maize B73 genome was first sought in EST data. A total of 32,229 3' EST sequences generated by us from the B73 inbred line were assembled into 3975 contigs and 6804 singleton ESTs. To be considered a CP, each of the two nucleotides must be supported by at least two independent sequence reads. Because this conservative heuristic qualifies only a subset of an assembly for locating putative NIPs, we term such regions “CP competent.” Of the 3975 EST contigs generated by CAP3 [Huang and Madan (1999)], 1659 were CP competent. To further analyze the correctness of these CP predictions, all 1659 candidates were manually inspected and the respective trace files were analyzed; following these analyses, 78 contigs were deemed promising.

**Experimental validation of EST-based CP sites**

*In silico* predicted CP sites could arise erroneously due to sequencing errors. We therefore endeavored to experimentally validate many of the putative NIPs. A total of 75 primer pairs flanking predicted CP sites were designed from the 78 EST contigs; 54 of these primer pairs amplified a single band from B73 genomic DNA. These PCR products were sequenced. Only those CP sites that exhibited overlapping sequence trace peaks were considered to be “validated.” Overlapping trace peaks were mostly of equal intensity, although in a few instances the relative intensities were consistent with differential NIP copy number in the maize genome. Of the 54 sequenced EST contigs that contained putative CPs, 9 could be validated in this manner.
Those CP sites that were validated via sequencing provide evidence in B73 of either residual heterozygosity or NIPs. The strategy outlined in Figure 1 was employed to distinguish between these possibilities. All nine validated EST contigs were analyzed in 20 individual selfed progeny from their B73 parent plant and in a pool of 20 individual progeny from 4 additional B73 parent plants (a total of 80 plants). If the validated CPs arose via the presence of residual heterozygosity, overlapping and nonoverlapping sequence trace peaks should segregate among the selfed progeny. No evidence of residual heterozygosity was detected. We therefore conclude that B73 exhibits a very low level of residual heterozygosity. We further conclude that 0.5% (9/1659) of the analyzed EST contigs is derived from NIPs.
NIPs discovered within a partial maize genome assembly

For purposes of NIP detection, ESTs are valuable because they are expressed and therefore inherently meet one of the criteria for classifying a duplicated sequence as a NIP (i.e., expression). On the other hand, because introns may be more diverged than ESTs, genomic regions from which these cDNAs are transcribed may not exhibit sufficient nucleotide identity (>98%) to be classified as NIPs. In addition, CPs can be identified only in genes for which at least four ESTs have been captured.

To address these limitations and to identify more NIPs in the maize genome, we endeavored to locate CPs within version 3.1 of our MAGIs [Fu et al. (2005)], which consists of 114,173 contigs. Because MAGIs include introns, the selection of MAGI-derived NIPs is even more stringent than for EST-based NIPs. A total of 15,375 MAGIs contain at least four overlapping clones and are therefore CP competent; 289 of these competent contigs exhibit at least one CP.

Primer pairs that flank CP sites for 280 of the 289 candidate MAGIs were designed, of which 231 amplified a single band from B73 genomic DNA. Sequence analyses of these amplicons validated a total of 258 paramorphisms (PMs) in 116 PM-containing MAGIs (Figure 2; see also Supplemental Figure 1 at http://www.genetics.org/supplemental) via a strategy identical to that used to validate NIPs identified from EST contigs. In several cases, primer pairs appeared to amplify multiple amplicons as evidenced by numerous multiple peaks in the sequence trace files. This suggests that a somewhat more distant paralog was also being amplified. Although at least one CP site was confirmed in these cases, to be conservative, these MAGIs were not included in subsequent analyses and calculations.

Expression of NIPs

Evidence for the expression of each of the 116 PM-containing MAGIs was sought via EST alignments, FGENESH predictions, and BLASTX results (Methods; Figure 2). The 84 PM-containing MAGIs for which evidence of gene expression was obtained were deemed to be NIPs (Supplemental Table 1 at http://www.genetics.org/supplemental). These 84 NIPs contain a
Figure 2  An example of a validated NIP (MAGL21152). The membership and layout of MF GSSs, a CP-competent interval (∼900 bp), and the trace file for a 150-bp subinterval of the CP-competent interval (the bottom chromatograph) are shown relative to the two paramorphisms highlighted.
Table 1  NIP pairs for which RT-PCR validated expression of both members

<table>
<thead>
<tr>
<th>MAGI ID</th>
<th>Annotation</th>
<th>Position</th>
<th>Paramorphisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. NIPs with EST support</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33361</td>
<td>Class III peroxidase 70 precursor</td>
<td>2571, 2586</td>
<td>G ... T</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A ... C</td>
</tr>
<tr>
<td>43016</td>
<td>Putative proteosome subunit</td>
<td>825</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>53926</td>
<td>Putative cytochrome P450</td>
<td>2594, 2635</td>
<td>T ... G</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C ... A</td>
</tr>
<tr>
<td>58637</td>
<td>Putative membrane related</td>
<td>651</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>65202</td>
<td>Hypothetical protein</td>
<td>3315</td>
<td>G</td>
</tr>
<tr>
<td>80184</td>
<td>Receptor-like kinase-like</td>
<td>1442</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>86866</td>
<td>Putative acyltransferase</td>
<td>1669</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>89568</td>
<td>NA</td>
<td>1085</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>97955</td>
<td>Putative nitrate reductase apoenzyme</td>
<td>1684</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>100946</td>
<td>Putative trehalose-6-phosphate synthase</td>
<td>715</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>B. NIPs with only FGENESH support</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21152</td>
<td>Putative strictosidine synthase</td>
<td>904, 909, 975, 999</td>
<td>G ... A ... T ... A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A ... G ... C ... G</td>
</tr>
<tr>
<td>36788</td>
<td>NA</td>
<td>1176</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>45574</td>
<td>Protein kinase</td>
<td>1448</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>67751</td>
<td>Putative S-receptor kinase</td>
<td>1048, 1087, 1122</td>
<td>T ... G ...G</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C ... A ... C</td>
</tr>
<tr>
<td>85672</td>
<td>NA</td>
<td>330</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>89009</td>
<td>Pentatricopeptide repeat-containing</td>
<td>862</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>98934</td>
<td>Putative cytochrome P450</td>
<td>652, 660, 785</td>
<td>C ... T ... A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G ... C ... T</td>
</tr>
<tr>
<td>95980</td>
<td>AKIN 1-like protein</td>
<td>2126</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>101406</td>
<td>Terpene syntase 5 related</td>
<td>1263</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

*aBLASTX search against UniRef protein database using $e^{-10}$ as E-value cutoff.

bThe presence of "..." between paramorphisms indicates that sites are not adjacent.
total of 170 validated paramorphic sites, which are located in both coding and noncoding regions.

Of the 44 NIPs that could be assigned functions via significant BLASTX matches, 10 are predicted kinases and 3 are predicted transcription factors and/or contain a zinc-finger domain. The remaining 31 NIPs are involved in a wide variety of biochemical pathways (e.g., metabolism, nitrogen utilization, and DNA methylation). We therefore conclude that NIPs are not restricted to a limited number of biological functions.

**Frequency of NIPs**

The experiments described above identified 84 genic MAGIs that contain one or more paramorphisms and are therefore classified as NIPs. Of the 15,375 CP-competent MAGIs, 12,012 appear to be genes on the basis of their lack of similarity to transposons and evidence of expression. The CP-competent intervals associated with the 84 validated NIPs exhibit $\geq 98\%$ nucleotide identity, include both coding and noncoding sequences, and can be as long as 2.6 kb (Supplemental Figure 2 at http://www.genetics.org/supplemental). Because $<80\%$ (231/289) of the CP-containing MAGIs were analyzed, we conservatively estimate that 0.9% $[84/(12,012 \times 0.8)]$ of the genes in this assembly have a NIP.

**Both members of many NIP families are expressed**

Forty-six NIPs that contained at least one exon or putative exon (Methods) were selected for analysis. Touchdown PCR was performed using both genomic DNA and pooled cDNA isolated from various tissues and organs of the inbred line B73. A total of 29 NIPs yielded a single band from both PCR reactions, of which 25 could be confirmed to be derived from the target NIP via sequencing. As shown in Table 1, these sequencing experiments provided evidence that both members of 20 NIP families ($80\%; 20/25$) are expressed (Methods). For the remaining 5 NIPs ($20\%; 5/25$), only one copy could be shown to be expressed. This is, however, a highly conservative assay for the expression because only a portion of the transcriptome was sampled. We conclude that both members of at least four-fifths of NIP families are expressed.
Table 2  NIP pairs for which RT-PCR validated expression of both members

<table>
<thead>
<tr>
<th>MAGI ID</th>
<th>Paramorphism sites</th>
<th>cDNA sample no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>21152</td>
<td>G .. A .. T .. A</td>
<td>+ ND</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A .. G .. C .. G</td>
<td>+ ND</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43016</td>
<td>G</td>
<td>+ ND</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>+ ND</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53926</td>
<td>T .. G</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C .. A</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>ND</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65202</td>
<td>G</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>ND</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>+</td>
<td>ND</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67751</td>
<td>T .. G .. G</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C .. A .. C</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80184</td>
<td>T</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86866</td>
<td>C</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89568</td>
<td>C</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>97955</td>
<td>C</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100946</td>
<td>G</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>ND</td>
<td>ND</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>+ ND</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>ND</td>
<td>ND</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>ND</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Members of many NIP families exhibit differential expression

Ten NIP families in which both members were expressed were further analyzed using RNA samples extracted from 16 different developmental stages of various tissues and organs. Members of 8 (80%) of these 10 NIP families were differentially expressed in at least one RNA sample (Table 2). We conclude that the members of many expressed NIP families are differentially expressed.

Genomic organization of maize NIPs

To begin to define the molecular events that give rise to NIPs, it would be useful to know the relative positions of members of NIP families within the maize genome. These experiments were conducted by using PCR primers that flank paramorphisms to amplify genomic DNA from the inbreds B73 and Mo17 and the IBM RIs derived from a cross between B73 and Mo17. Most of the 84 NIP primer pairs could amplify both B73 and Mo17 and the resulting amplicons from these two inbreds were the same size at the resolution afforded by gel electrophoresis. However, B73 genomic DNA but not Mo17 was amplified when 14 of the primer pairs were used in PCR. This indicates either that the corresponding Mo17 NIPs exhibit a high degree of sequence or structural polymorphism relative to the B73 NIPs from which the PCR primers were designed or that the Mo17 genome does not contain the corresponding NIP, a result that would extend the violations of genomic colinearity among maize inbreds initially observed by Fu and Dooner (2002) and extended by others [Brunner et al. (2005); Lai et al. (2005); Lal and Hannah (2005)]. Using the PCR primers that amplify B73 NIPs but not Mo17 to genotype the IBM RIs, it was possible to determine the positions of the members of all 14 NIP families relative to each other (Methods). The members of 7 and 2 NIP families were tightly and loosely linked, respectively (see Supplemental Table 1 at http://www.genetics.org/supplemental). The members of an additional 5 NIP families were unlinked genetically.
Arabidopsis NIPs

Although Arabidopsis has a much smaller genome than maize, it is also thought to have undergone an ancient polyploidization event [Vision et al. (2000)]. To compare the relative rates of NIPs in these two model plants, we sought EST-based NIPs in Arabidopsis using the Columbia ecotype. Of the 33 initial EST clusters analyzed that contained at least one statistically significant CP, 7 were found to have already been reported to be transcribed from two or more copies in the Arabidopsis genome; however, the inclusion of introns for all seven of these genes results in <98% identity. A total of 117 CPs were tested in 24 of the 26 novel Arabidopsis NIPs using primer pairs that successfully amplified a single band of DNA from Columbia genomic template (25 primer pairs total); 100 were definitively established as false positives. The remaining 17 putative CP sites could not be verified as negative due to low-quality sequence reads. Hence, there is no evidence that any of the Arabidopsis EST clusters surveyed here represent novel collapsed paralogs.

To confirm this observation, we located NIPs among all 68 low-copy Arabidopsis gene pairs that have rates of synonymous substitution ($K_s$) that are $<2\%$ [Lynch and Conery (2000); Moore and Purugganan (2003)]. Only 39 pairs meet the NIP criteria and are annotated as distinct loci (Methods), which is consistent with the EST result. Of these NIP families, 28 are located $<10$ cM apart (Methods). Of the remaining 11 NIP families, 9 of these are located on different chromosomes.

Discussion

The maize genome contains a high frequency of NIPs

Plant genomes contain large numbers of paralogs, many of which are tandemly arrayed [Sun et al. (2001); Yuan et al. (2002); Messing et al. (2004)]. In addition, maize contains a substantial degree of intraspecies diversity for gene content [Fu and Dooner (2002)]. At least some of the intraspecific violations of genetic colinearity are due to “hitchhiking” gene fragments that have been duplicated by active transposons [Brunner et al. (2005); Lai et al.
(2005); Lal and Hannah (2005); Morgante et al. (2005)). Potentially, these duplications of genic sequences have significant evolutionary implications. The extent to which these duplications are functional is, however, under debate [Juretic et al. (2005)].

It has previously been reported that several pairs of NIPs are expressed. These include the genetically unlinked ciszog1 and ciszog2 genes [Swigonova et al. (2005)], the tightly linked p1 and p2 genes [Zhang et al. (2000)], and the locally duplicated zein seed storage protein gene families that exhibit 98% identity [Song et al. (2001)]. This study demonstrates that most NIPs are expressed and that individual members of many NIP families exhibit differential expression patterns. Given their high degree of sequence identity, it likely that these different expression patterns are controlled by sequence variation outside the NIPs or differing epigenetic states, including local chromatin structure. Taken together, this study provides the first conclusive evidence that substantial numbers of hypomethylated duplications have successfully diversified their expression profiles and may therefore have unique functional roles.

**Origins of NIPs**

Following duplication, gene pairs would be expected to decay into NIPs. Although transposons can “capture” gene sequences and duplicate them via transposition, Mu elements do not preferentially insert at genetically linked sites [Lisch et al. (1995)]. It is therefore unlikely that Pack-MULEs [Jiang et al. (2004)] would be able to generate the large proportion of genetically linked NIPs observed in this study. Similarly, unless Helitrons [Lal et al. (2003); Brunner et al. (2005); Lai et al. (2005); Lal and Hannah (2005); Morgante et al. (2005)] preferentially insert in nearby locations, tandemly arrayed NIPs are unlikely to have arisen via the action of Helitrons. We therefore consider several alternative mechanisms that could generate NIPs.

Unequal recombination between repetitive sequences that flank genes can generate gene duplications [Babcock et al. (2003)]. In humans, such processes are thought to be responsible for ≈30% of the recent segmental duplications [Zhou and Mishra (2005)]. Unequal recombination occurs between the long terminal repeats of rice retrotransposons [Ma et al. (2004); Ma and Bennetzen (2006)]. Tandem gene duplications generated via this mechanism would
be flanked by repeats of high identity. A \(\approx 10\)-kb segment of BAC clone ZMMBBb0483G05 deposited in GenBank (accession no. AC157776) by the McCombie laboratory contains two pairs of tandemly duplicated NIPs; each pair of NIPs exhibits \(>99.5\)% identity. Significantly, conserved repeats (as defined by the Iowa State University MAGI Cereal Repeat Database 3.1; Fu et al. (2005)) are located between and flanking the duplications. The positioning of these repeats is consistent with duplication via unequal pairing between the repeats.

More exotic mechanisms of NIP generation are also possible. For example, break-induced replication at stalled replication forks could stimulate the production of segmental duplications (Figure 3A, iii) and rearrangements in regions of genomic instability [Koszul et al. (2004); Zhou and Mishra (2005)]. Gene conversion or similar mechanisms may have also homogenized diverged paralogs. Because many of the characterized maize gene conversion events have conversion tracts \(>1\) kb (reviewed by Yandeau-Nelson et al. (2005)), it is possible that this mechanism could generate NIPs. In support of this hypothesis, we have recently observed that the duplicate \(gl8\) genes (\(gl8a\) and \(gl8b\)), which reside on syntenic regions of different chromosomes and therefore presumably originated during the ancient allotetraploidization event, exhibit a degree of nucleotide identity [96%; Dietrich et al. (2005)] that is substantially higher than the 80\% identity expected for ancient paralogs [Blanc and Wolfe (2004)]. Because tandemly arrayed paralogs undergo frequent recombination [Yandeau-Nelson et al. (2006)], gene conversion can also maintain a high degree of nucleotide identity between them [Zhang et al. (2005)].

While it is not currently possible to identify the mechanism by which a given NIP pair was generated, it is likely that multiple mechanisms are involved. It may be possible to decipher these mechanisms once the maize genome sequence has been completed by locating the specific sequence signatures that are associated with each duplication mechanism (Figures 3 and 4).

**Why does maize have more NIPs than Arabidopsis?**

We conservatively estimate that the maize genome contains at least 500 NIPs. In contrast, we identified <10\% of this number of NIPs in the Arabidopsis genome (\(N = 39\)). This is true even though the Arabidopsis genome contains Helitrons [Kapitonov and Jurka (2001)], which
Figure 3  Mechanisms of gene duplication. Unequal pairing between flanking repeats (A, ii) can occur between homologs or sister chromatids, but probably at a lower rate. Transposon-mediated duplication can generate genetically tightly linked (A, i) and unlinked (B, i) NIPs. Unlinked NIPs could reside on separate chromosomes as depicted in (B, i) or could be at least 50 cM apart on the same chromosome. (B) Genetically unlinked NIPs are shown on two separate chromosomes (I and II). Unlinked NIPs can result from duplications of entire chromosomes (B, ii) or large segments of chromosomes that subsequently diverge (i.e., chromosomal rearrangements and gene loss or gain). Both linked and unlinked gene duplications might also occur by currently uncharacterized mechanisms. Boxes, thick lines, and solid circles represent genes, nongenic repeats, and centromeres, respectively.
Figure 4  A proposed mechanism for the evolution of gene duplications and the generation of NIPs and totally identical paralogs (TIPs). Genetically linked (A) and unlinked (B) duplication events generate TIPs that can diverge over time to produce NIPs. NIPs can be homogenized back into TIPs via nonallelic gene conversion or can further diverge. More diverged paralogs might also be homogenized into TIPs, but likely at a lower rate (dashed line). Shaded boxes represent genes and vertical lines within the boxes represent paramorphisms.
duplicate genes in maize [Brunner et al. (2005); Lai et al. (2005); Lal and Hannah (2005); Morgante et al. (2005)].

The frequency of NIPs within a species depends on the rates of four parameters: the rate and timing of initial duplication events, the rate at which NIPs decay (mutation rate), and the rates of gene loss and gene conversion. Hence, the lower frequency of NIPs in Arabidopsis as compared to maize could be a consequence of a lower rate of gene duplication. Alternatively, if gene conversion is a dominant mechanism for gene duplication, the fact that only ≈12.616.6% of Arabidopsis genes are members of tandemly arrayed gene families [Zhang and Gaut (2003)] as compared to ≈35% of maize genes [Messing et al. (2004)] may contribute to the observed differences in NIP content between these species.

NIPs and genetic markers

NIPs can complicate the development of SNP-based genetic markers. This is because an apparent “SNP” identified via comparisons of ESTs or shotgun sequences from two inbreds may represent a paramorphism rather than a true SNP. Unlike SNPs, paramorphisms will not necessarily exhibit Mendelian segregation; therefore, it may not be possible to convert them into informative genetic markers. Indeed, such an explanation has been invoked to explain the inability to convert a fraction of human “SNPs” into genetic markers [Fredman et al. (2004)].

Evolutionary implications of NIPs

An individual diploid genome can contain at most two alleles of a given locus. NIPs provide a mechanism for a maize plant to include more than two “alleles” of a given gene within its genome and the differential expression of members within a NIP family can increase the plasticity of the transcriptome. Hence, the genetic diversity provided by NIPs may contribute to the environmental stability of maize. NIPs may also serve as a reservoir of genetic variability upon which selection can act because recombination between highly similar paralogs can generate new “alleles” that condition novel phenotypes [Zhang et al. (2005)]. Finally, the existence of multiple copies of a given sequence (i.e., NIPs) increases the probability of recovering
rare favorable mutations. As such, NIPs may have facilitated the domestication of maize and may contribute to the continuing success of long-term selection experiments in closed maize populations [Laurie et al. (2004)] and maize breeding in general.

**Acknowledgments**

We thank Amy Nienaber and Elizabeth Hahn for technical assistance with the NIP detection experiments in maize and Arabidopsis, respectively; Karthik Viswanathan and Mu Zhang for computational support; Hsin D. Chen and Josh Shendelman for genetic mapping; Sang-Duck Seo for assistance in figure design; and An-Ping Hsia for helpful comments. Scott Emrich was supported in part by a National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship fellowship (award no. DGE-9972653). This research was funded in part by competitive grants from the NSF Plant Genome Program (DBI-9975868 and DBI-0321711); additional support was provided by the Hatch Act and State of Iowa funds.

**References**


GENE DISCOVERY AND ANNOTATION USING LCM-454 TRANSCRIPTOMOME SEQUENCING

A paper published in Genome Research\textsuperscript{1}

Scott J. Emrich, W. Brad Barbazuk, Li Li, and Patrick S. Schnable

Abstract

454 DNA sequencing technology achieves significant throughput relative to traditional approaches. More than 261,000 ESTs were generated by 454 Life Sciences from cDNA isolated using laser capture microdissection (LCM) from the developmentally important shoot apical meristem (SAM) of maize (\textit{Zea mays} L.). This single sequencing run annotated \(>25,000\) maize genomic sequences and also captured \(\approx 400\) expressed transcripts for which homologous sequences have not yet been identified in other species. Approximately 70\% of the ESTs generated in this study had not been captured during a previous EST project conducted using a cDNA library constructed from hand-dissected apex tissue that is highly enriched for SAMs. In addition, at least 30\% of the 454-ESTs do not align to any of the \(\approx 648,000\) extant maize ESTs using conservative alignment criteria. These results indicate that the combination of LCM and the deep sequencing possible with 454 technology enriches for SAM transcripts not present in current EST collections. RT-PCR was used to validate the expression of 27 genes whose expression had been detected in the SAM via LCM-454 technology, but that lacked orthologs in GenBank. Significantly, transcripts from \(\approx 74\%\) (20/27) of these validated SAM-expressed “orphans” were not detected in meristem-rich immature ears. We conclude that the

\textsuperscript{1}Reprinted with permission of Genome Res., 2007, 17:69–73.
coupling of LCM and 454 sequencing technologies facilitates the discovery of rare, possibly cell-type-specific transcripts.

Introduction

Although genome sequencing technology has become progressively more efficient over the past decade, the sequencing of complex genomes remains expensive. Expressed Sequence Tag (EST) sequencing provides an attractive alternative to whole-genome sequencing because this technique produces sequences of the transcribed portions of genes at a fraction of the cost of sequencing complete chromosomes. Even so, because genes are differentially expressed, multiple tissues must be sampled, and, when using traditional (Sanger) methods, these EST projects require substantial investments in library construction and sequencing, particularly if the goal is to capture rare transcripts.

Recently, 454 Life Sciences developed a scalable, highly parallel DNA sequencing system that is 100 times faster than standard sequencing methods and is capable of sequencing >200,000 fragments per 4-h run [Margulies et al. (2005)]. This increase in throughput comes at the expense of read length. On average, 454 sequence reads are only $\approx 100$ bp in length, and in addition, this technology does not capture read-pair information [Margulies et al. (2005)]. Hence, the assembly of 454 sequences from samples that contain large amounts of repetitive DNA such as eukaryotic genomes may prove problematic for conventional fragment assembly programs.

In contrast, the read-length limitation associated with 454 technology is less of a concern for transcriptome sequencing and analysis. This is because transcriptomes are smaller than the genomes from which they are derived and typically contain less repetitive DNA. Using laser-capture microdissection (LCM) (for review, see Schnable et al. (2004)) to isolate transcripts that accumulate in specific cell types has the potential to further reduce the size of a target transcriptome. Because 454 technology avoids expensive cloning-based library construction, it is feasible to sequence a wide variety of LCM-derived cDNA samples, thereby increasing the recovery of highly specialized transcripts. Moreover, 454 technology combined with LCM is
particularly well suited for EST-based gene discovery because it generates hundreds of thousands of tags per run, greatly increasing the chances of capturing rare transcripts.

Here, we report the sequencing of cDNA extracted from developmentally important Shoot Apical Meristem (SAM) cells [Baurle and Laux (2003); Guyomarc’h et al. (2005)] using the LCM-454 approach. A single 454 sequencing run was able to annotate >25,000 maize genomic sequences and capture transcripts from nearly 400 “orphan genes” [Fu et al. (2005)]. Interestingly, experimental validation suggests that not only are orphan transcripts discovered using the LCM-454 approach, but most of these genes are undetectable in cDNA samples from other tissues including meristem-rich immature ears. LCM-454 sequencing is, therefore, an efficient gene-discovery platform when applied to highly specialized organs such as the SAM.

Methods

Isolation of SAM mRNA

Maize (*Zea mays* inbred line B73) SAM tissue, which included Plastochron0 (P0) and P1, was extracted from ≈10 14-d-old seedlings. This was achieved with modifications to the paraffin-embedding technique described by Kerk et al. (2003) and the Laser Capture Microdissection (LCM) technique described by Nakazono et al. (2003). Full details are described elsewhere (Ohtsu et al., in prep.). A highly repeatable T7 RNA polymerase-based RNA amplification was performed as described by Nakazono et al. (2003) with some modifications to generate sufficient SAM cDNA for sequencing. Because a poly(T) primer was used for amplification, the resulting cDNA was enriched for the 3’-ends of transcripts.

454-EST sequencing and processing

Approximately 15 µg of LCM-derived cDNA was submitted to 454 Life Sciences, who ensured sample quality by checking the SAM cDNA on a 2% agarose gel and an Agilent bioanalyzer. The cDNA sample was then fractionated into smaller pieces (300–500 bp) that were subsequently polished (blunted). Short adapters were then ligated on to each resulting fragment, which provide priming sequences for both amplification and sequencing, forming the
basis of the single-stranded template library. Finally, one sequencing run was performed using the method of Margulies et al. (2005) and resulted in 288,992 EST sequences. 454 Life Sciences have helped submit these sequences to the NCBI trace archive (http://www.ncbi.nlm.nih.gov/, accession nos. DW724699–DW985434), where they are available for independent analysis. These sequences were subsequently trimmed using Lucy [Chou and Holmes (2001)] under default settings with the exception that sequences as short as 50 bp were not discarded; this returned 260,887 high-quality sequences, which we have submitted to dbEST and used for annotation. In addition, poly(A/T) tails were removed from raw 454 sequences with SeqClean (http://www.tigr.org/tgi/software) using default settings to ascertain the novelty of these sequences using longer, albeit lower quality, reads. In addition, contaminating sequences (150 sequences; 0.05% of total) were removed by SeqClean based on similarity to the Escherichia coli K12 (GenBank accession no. U00096) and Lactococcus lactis (GenBank accession no. AE005176) genomes and GenBank’s Univec database.

Comparisons of 454-ESTs to public sequence databases

Maize ESTs (N = 656,696) were downloaded from GenBank in December 2005 and processed using SeqClean as described above. After eliminating 9011 contaminating or low-quality sequences, 29,615 maize ESTs (MESTs) sequenced by us from diverse cDNA libraries were extracted based on the presence of a poly(T) prefix of at least 10 bp; these were used to assess 3’-enrichment and putative sampling biases. For annotation purposes, another subset of 31,036 ESTs sequenced by us from a cDNA library generated by M. Scanlon’s group (University of Georgia) from mRNA isolated from vegetative apices was extracted. The Apex ESTs were assembled using CAP3 [Huang and Madan (1999)] to generate unigenes using the following parameters: -p 98 -o 100 -y 20 -h 5.

The 454 SAM ESTs with poly(A/T) tails removed were compared to the 647,685 high-quality, unassembled maize ESTs, the maize Apex unigenes, ISU MAGIs version 3.1 (including singletons), maize chloroplast (GenBank accession no. X86563) and mitochondrial genome sequences (GenBank accession no. AY506529), and the ISU Cereal Repeat Database (http://
magi.plantgenomics.iastate.edu) using BLAST. Nucleotide alignments with either an $E$-value $\leq 1e^{-8}$ or $>70\%$ identity over 50\% of the EST length were deemed to have been previously discovered, providing a highly conservative estimate of novel gene discovery. The following TIGR Plant Gene Indices (http://www.tigr.org/tdb/tgi) downloaded in December 2005 were similarly searched for matches: HVGI release 9 (barley), OGI release 16 (rice), SBGI release 8 (sorghum), SOGI release 2 (sugar cane), and TAGI release 10 (wheat). Candidates were also compared to the Arabidopsis genome (ATH1.1con.01222004; http://www.arabidopsis.org), finished rice chromosome sequences (GenBank AP008207-AP008218), and the TIGR dicot gene indices used by Fu et al. (2005) to locate homologous sequences among plant ESTs.

Evidence of expression of SAM genes was determined by locating reciprocal best hits between predicted maize genes [Fu et al. (2005)] and Lucy-trimmed 454-ESTs requiring a minimum $E$-value of $1e^{-20}$. Potential homologs were located among the monocot gene indices described above, and repeats were located against the MAGI Cereal Repeat Database v 3.1; both analyses used the novelty criteria previously described [Fu et al. (2005)]. Similarly, all putative orphan genes were compared to the GenBank nr database (BLASTN and BLASTX) and to the est.others database (BLASTN) on January 8, 2006 using netBLAST (blastcl3).

**Annotation using 454-EST sequences**

All 114,173 contigs from the partial maize inbred line B73 genome assembly MAGI 3.1 [Fu et al. (2005)] were aligned to Lucy-trimmed 454 SAM ESTs using GeneSequer and its maize-specific splice models [Usuka et al. (2000)]. Only alignments consisting of at least one exon of at least 50 bp in length and with identity $\geq 95\%$ over at least 80\% of the length of the 454-EST were used as evidence of expression. ESTs with $>50$ bp of repetitive sequence, as determined by a previously described masking procedure [Emrich et al. (2004)], were ignored when the number of expressed MAGIs was calculated.
Validation of expression of orphan genes

RT-PCR and sequencing were conducted as described by Fu et al. (2005) on three pools of cDNA generated as described by Fu et al. (2005). The first pool was derived from amplified RNA isolated from SAMs via LCM as described above. The second pool was a complex mixture generated from multiple tissues harvested from B73 maize plants 79 d after planting in Ames, Iowa during the summer of 2005. The third pool was generated from immature, unpollinated top ears harvested from the inbred B73 59 d after planting (ears measured 1.25–2.5 cm in length). Based on RT-PCR results obtained using a pair of tub6 primers that flank a 100-bp intron, these cDNA samples are free of detectable genomic DNA contamination.

Estimating the rate of sequencing errors in 454-ESTs

To estimate the rate of sequencing error in the ESTs generated by 454 Life Sciences, we aligned all ESTs to a collection of FGENESH-predicted maize cDNAs [Fu et al. (2005)] using BLASTN and only used the best hit with an $E$-value $< 10^{-10}$. For all comparisons, at least 90% of the length of a 454 read had to match its corresponding benchmark to be considered a valid alignment. Although any disagreement is not conclusive proof of an error, we have shown that the MAGI-based maize cDNAs are of high enough quality [Fu et al. (2005)] that these disagreements are likely errors in the 454 sequences.

Estimating the 3'-enrichment of 454-ESTs

A set of 8852 MAGIs was selected based on their alignment to 29,615 3' maize ESTs with discernable poly(A/T) tails [Fu et al. (2005)]. Only the 5575 of these MAGIs that had an experimentally determined poly(A) site within 50 bp of the predicted termination of transcription were tested for alignment to the LCM-454 ESTs. A total of 32,075 LCM-454 ESTs aligned to these predicted genes. The LCM-454 ESTs were also directly aligned to the 29,615 3'-ESTs. A total of 36,258 LCM-454 ESTs aligned to the 3'-ESTs.
Results

Gene discovery and annotation using 454 sequencing

As of December 2005, >650,000 maize EST sequences obtained from diverse tissues and genotypes had been deposited in GenBank, including sequences derived from libraries prepared from specialized structures such as the vegetative shoot apex. The apex contains both newly formed leaves and SAM cells that initiate all above-ground tissue in plants. The developmentally important SAM cells, however, comprise only a very small portion of the apex. Consequently, it is difficult to capture rare SAM-specific transcripts by sequencing ESTs from an apex library.

One means to obtain rare transcripts from specific cell types (e.g., those that comprise the SAM) is to extract and clone mRNA from individual cell types using LCM [Asano et al. (2002)]. This approach, however, requires a significant investment in cDNA sequencing including library construction. As a potential alternative, we attempted to discover rare transcripts by directly sequencing cDNA using the high-throughput 454 sequencing platform. Maize cDNA was extracted from multiple SAMs using LCM as described by Nakazono et al. (2003), amplified, and sequenced by 454 Life Sciences. After removing poly(A/T) tails from these reads (Methods), the ≈261,000 resulting SAM ESTs had an average length of 101 bp.

The 454-ESTs were BLASTN-aligned to a variety of maize sequence databases (Table 1). In total, >93% of the 454 SAM EST sequences matched maize ESTs, GSSs, repeats, or organelle genomes. We and colleagues had previously generated ≈31,000 ESTs from a cDNA library prepared from hand-dissected maize apices (Methods). The 454-ESTs were aligned to the ≈18,560 unigene transcripts assembled from these Apex ESTs (Methods). More than 70% of the SAM 454-ESTs did not align to the Apex ESTs from this SAM-enriched library. GenBank contains >600,000 additional maize ESTs (Methods). More than 30% of the 454-ESTs did not align to this extensive collection of ESTs. These results indicate that this 454 sequencing run captured ESTs from many maize genes without previous evidence of expression.

We previously assembled ≈880,000 “gene-enriched” B73 genomic sequences into Maize As-
Table 1  Genic sequences captured by a single 454 run compared to other gene-enriched sequencing approaches

<table>
<thead>
<tr>
<th>Source database</th>
<th>No. of matching 454-ESTs</th>
<th>No. of novel 454-ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>UGA-ISA Apex unigenes (N=18,558)</td>
<td>73,145</td>
<td>187,591 (71.9%)</td>
</tr>
<tr>
<td>GenBank Maize ESTs (N = 647,685)</td>
<td>179,912</td>
<td>80,824 (31.0%)</td>
</tr>
<tr>
<td>ESTs + ISU MAGI 3.1 (GSS) (N = 862,158)</td>
<td>239,113</td>
<td>21,623 (8.3%)</td>
</tr>
<tr>
<td>ESTs + ISU MAGI 3.1 + organelle genomes + repeats (N=877,431)</td>
<td>244,328</td>
<td>16,408 (6.3%)</td>
</tr>
<tr>
<td>ESTs + ISU MAGI 3.1 + organelle + monocot ESTs (N=1,282,226)</td>
<td>245,339</td>
<td>15,397 (5.9%)</td>
</tr>
</tbody>
</table>

sembled Genomic Islands (MAGIs) (http://magi.plantgenomics.iastate.edu). Previous alignments between the 114,173 MAGIs and a unigene set composed of the ≈419,000 maize ESTs available in GenBank prior to February 2004 provided evidence that ≈20,900 MAGIs contain at least portions of expressed genes [Fu et al. (2005)]. Similar alignments of the 454 SAM ESTs provide evidence that ≈25,800 MAGIs contain at least portions of expressed genes. Significantly, 15,521 of these ≈25,800 MAGIs did not have prior expression evidence from the alignments to the ≈419,000 maize ESTs, which included the Apex ESTs. These results suggest that the representation of rare and/or SAM-specific transcripts has been enriched by the deep sequencing of cDNA isolated from SAM tissue. Hence, we conclude that LCM-454 sequencing is an efficient approach for the large-scale validation of gene expression.

We previously reported [Fu et al. (2005)] that ≈5% of expressed maize genes are “orphans” relative to known sequence databases including GenBank and dbEST. Consistent with this previous observation, we estimate that relative to current plant databases (Methods), ≈15,400 (6%) of the 454 SAM ESTs were transcribed from orphan genes. Because ESTs are differentially expressed and full-length cDNAs are not available, it is difficult to determine exactly how many unique SAM-expressed genes are orphans. It is possible, however, to estimate the
overall frequency of orphans by confirming the expression of a sample of genes. A total of 9944 of the predicted maize genes described by Fu et al. (2005) were deemed, based on 454-ESTs data, to be expressed in the SAM (Methods). Of these, 914 (9%) do not have homologous sequences in monocot EST databases (Methods). Of these, 390 genes do not have matches to non-EST databases, including repeat databases (Methods). Hence, a single 454 sequencing run provided EST-based support for the expression of >9000 SAM-expressed genes, of which 390 are nonrepetitive orphans.

Validation of orphan expression

RT-PCR was used to confirm the expression of a sample of the orphan genes detected among the 454-ESTs. A set of 42 MAGIs that contained orphan FGENESH-predicted genes was selected for analysis that (1) aligned to 454-ESTs, (2) contained at least one intron, and (3) yielded primers that met our design criteria. Criteria 2 and 3 were used to be consistent with a prior study of maize orphans [Fu et al. (2005)]. As in the previous study, PCR primers were designed based on FGENESH-predicted exonic sequences in each gene. Initially, PCR amplification was performed using B73 genomic DNA as a template. A total of 38 of the 42 primer pairs yielded genomic PCR products of the expected sizes. To obtain an independent test of whether these orphan genes are indeed expressed, the 38 primer pairs were then used to conduct PCR experiments on three pools of cDNA derived from (1) SAMs, (2) meristem-rich immature ears, and (3) multiple tissues (Methods). If a single RT-PCR band was obtained, it was sequenced. Of the 38 primer pairs, 27 produced RT-PCR products that were of the correct size and whose sequence matched the MAGIs from which the primers were designed. All 27 of these orphans were expressed in the SAM (Figure 1). Based on these results, we conclude that many of the orphans detected among the 454-ESTs are, indeed, expressed. Eleven of the 27 orphans were expressed in at least one of the other two cDNA pools. Interestingly, 20/27 (74%) of the RT-positive orphan transcripts that were detected in the SAM were not detected in the meristem-rich immature ears. This could be because of the substantial enrichment of meristems in the SAM sample and/or the existence of genes that are expressed in the SAM
Figure 1  Experimental validation of the expression of orphan genes. (A) Test for genomic DNA contamination of cDNA. Primers that flank a 100-bp intron in the maize beta-tubulin6 (tub6) gene were used to amplify genomic DNA (lane 3), SAM cDNA (lane 2), immature ear cDNA (lane 4), and the complex cDNA pool (lane 5). (B) Examples of orphans with validated expression patterns. Primers designed to amplify (lanes 2-4) MAGI\_80343, (lanes 5-7) MAGI\_60450, (lanes 8-10) MAGI\_75030, and (lanes 11-13) MAGI\_30050 were used to amplify (lanes 2,5,8,11) SAM cDNA, (lanes 3,6,8,12) immature ear cDNA, and (lanes 4,7,9,13) the pooled cDNA sample. (C) Summary of RT-PCR results for the 27 orphan genes. (+) Indicates that an RT-PCR product of the correct size was detected. Lane 1 of panels A and B contains the One KB Plus size standard (GIBCO BRL). Because primer dimers present in some lanes were cropped in both panels, the smallest size standard band shown is 200 bp.
but not in the reproductive meristems present on the immature ears. In either case, this result provides further evidence for the value of coupling LCM and 454 sequencing for gene discovery.

Discussion

Reductions in reagent volumes of Sanger sequencing reactions have substantially reduced costs without affecting read lengths or sequence accuracy [Smailus et al. (2005)]. Because of diminishing returns, replacement technologies are required to achieve additional cost savings and make possible grand challenges such as the “$1000 genome” and the complete characterization of all expressed genes of an organism and their respective splice forms. Recently, 454 Life Sciences released a proprietary sequencing technology that quickly provides vast amounts of sequence data without the need to clone DNA prior to sequencing, further reducing the total effort required for large-scale sequencing projects. The reads obtained with 454 technology are, however, much shorter than traditional Sanger reads and are subject to a higher rate of base-calling errors, particularly in association with homopolymer runs.

This study provides experimental data that demonstrate the value of using 454 technology to sequence expressed sequences present in specific cell types isolated using laser capture microdissection (LCM). Because of its reduced size relative to the entire genome, an LCM-derived transcriptome can be more efficiently sampled, and therefore covered, by 454 sequencing. In addition, reducing the complexity of the transcriptome prior to sequencing by restricting cDNA recovery to specific tissues of interest was expected to increase the recovery of rare, tissue-specific transcripts. Approximately 261,000 454-ESTs were generated from LCM-collected SAM tissue. Only 70% of the 454 SAM ESTs align to ≈648,000 maize ESTs. All potentially novel LCM-454 ESTs were aligned to the complete set of MAGIs. This corrected for LCM-454 ESTs derived from the same gene, but that did not overlap. These analyses validated the expression of >15,000 MAGIs that did not have prior evidence of expression.

As alluded to above, if a given gene is sampled by multiple nonoverlapping ESTs, the number of unique transcripts will be overestimated. Some traditional EST projects address this problem by sequencing the 3’-ends of cDNAs. It is not possible to specifically sequence
the 3'-ends of cDNAs using 454 sequencing technology. Even so, our LCM-454 EST project greatly enriched for 3'-sequences and thereby minimized the overestimation of the number of unique transcripts in the SAM.

The 3' enrichment achieved via LCM-454 sequencing is a consequence of the procedure used to amplify RNA from LCM-collected tissue (Methods), which results in relatively short cDNA fragments (≈200–600 bp), all of which included the 3'-terminus of the corresponding transcripts. Prior to 454 sequencing, cDNAs are sheared. But because the target shear size is close to the size of our amplified cDNAs, most of our cDNAs were probably not sheared, or if sheared were removed via size selection prior to sequencing. Hence, we expected that a large percentage of our cDNAs were sequenced from their 3'-termini.

To test the degree to which our 454-ESTs were 3'-enriched, we identified a set of 3'-ESTs and a set of predicted maize genes that align to these 3'-ESTs (Methods) and then examined the distributions of LCM-454 EST alignments along the lengths of these genes. Using the 3'-ESTs (average length 565 bp), the beginning of the 454-EST/3'-EST alignment is within the first 20 bp upstream of the poly(A) site in 41% of the alignments, within the first 100 bp in 76% of the alignments, and within the first 300 bp in >95% of the alignments. Results for the substantially longer FGENESH-predicted genes (average length of 1039 bp) that aligned to LCM-454 ESTs were similar; the beginning of the 454-EST/MAGI alignment was within the first 20 bp upstream of the poly(A) site in 40% of the alignments, within the first 100 bp in 66% of the alignments, and within the first 300 bp in 90% of the alignments. This substantial 3'-enrichment provides confidence that the number of novel transcripts detected in this study is not substantially overestimated.

Current estimates suggest that up to 5% of expressed maize genes are “orphans” [Fu et al. (2005)] that is, they match no genes isolated to date from any species. Previously, the expression of hypothetical orphan genes has been detected via large-scale efforts to specifically amplify associated transcripts from cDNA preparations [Fu et al. (2005); Xiao et al. (2005)]. In contrast, a single run of SAM 454-ESTs was able to detect the expression of ≈400 expressed orphans; the expression of many of the tested orphans was validated via RT-PCR.
Consequently, we conclude that the combination of LCM and 454 sequencing technologies is an efficient approach to discover and annotate genes.

Given the ease with which hundreds of thousands of ESTs can be generated, 454 technology makes it possible to obtain relative expression data on thousands of genes. Several high-throughput, sequencing-based quantitative expression analysis techniques are already available, most notably SAGE [Velculescu et al. (1995)] and MPSS [Brenner et al. (2000)]. Because both of these prior technologies produce short sequence signatures from discrete regions of transcripts, they provide a sensitive indicator of relative expression levels [Meyers et al. (2004)]; however, these techniques cannot provide sequence data over substantial portions of cDNAs and are therefore less well suited for applications such as SNP detection. In contrast, 454 sequencing could potentially recover virtually the entire template via “shotgun” sequencing of the transcriptome, and these tags are inherently better suited for discriminating the expression of members of highly conserved gene families because they are longer in length. Even so, under some circumstances it may be desirable to sequence SAGE libraries with 454 technology to leverage the advantages of both approaches to analyze expression digitally.

Following LCM, and prior to sequencing, we amplified RNA using a poly(T) primer. This procedure yielded fragments that are 3’-enriched relative to the entire transcriptome. The advantages of this 3’-enrichment are that it provides a better estimate of the numbers of unique transcripts within a particular transcriptome and greater depth of coverage is achieved in the 3’-ends of transcripts. The resulting data are well suited for gene discovery and in silico Northerns because transcripts are sampled at rates independent of their lengths. On the other hand, to obtain the sequence of a complete transcriptome, it would be desirable to avoid this 3’-enrichment by using random primers, rather than a poly(T) primer, to amplify the RNA following LCM. Our coverage modeling (data not shown) suggests that the ends of cDNAs will not be efficiently captured via 454 technology. Even so, 454 sequencing technology can efficiently capture the bulk of a transcriptome for use in applications such as gene discovery, annotation, and the discovery of polymorphisms. This is particularly true if transcriptome size is controlled by analyzing appropriate cell types, organs, or tissues via LCM.
Acknowledgments

We thank Kazuhiro Ohtsu and Marianne Smith for preparing the SAM-specific cDNA used for 454 sequencing, Xiaolan Zhang of the Mike Scanlon laboratory (University of Georgia) for preparing the Apex EST library, Ruilian Zhou and other members of the Schnable Laboratory for sequencing this library, and Kazuhiro Ohtsu and Yan Fu and three anonymous reviewers for helpful comments on this manuscript. This research was supported by grants from the National Science Foundation (DBI-0321595 and DBI-0527192), ISUs Plant Science Institute, the Donald Danforth Plant Science Center, and Pioneer Hi-Bred; additional support was provided by Hatch Act and State of Iowa funds.

References


Emrich, S. J., Aluru, S., Fu, Y., Wen, T. J., Narayanan, M., Guo, L., Ashlock, D. A.,


SNP DISCOVERY VIA 454 TRANSCRIPTOME SEQUENCING

A paper accepted by the Plant Journal

W. Brad Barbazuk\textsuperscript{1}, Scott J. Emrich\textsuperscript{2}, Hsin D. Chen, Li Li, and Patrick S. Schnable

Abstract

A massively parallel pyrosequencing technology commercialized by 454 Life Sciences Corporation was used to sequence the transcriptomes of shoot apical meristems isolated from two inbred lines of maize using laser capture microdissection (LCM). A computational pipeline that uses the POLYBAYES polymorphism detection system was adapted for 454 ESTs and used to detect SNPs (Single Nucleotide Polymorphisms) between the two inbreds. Putative SNPs were computationally identified using 260,000 and 280,000 454 ESTs from the B73 and Mo17 inbreds, respectively. Over 36,000 putative SNPs were detected within 9,980 unique B73 genomic anchor sequences (MAGIs). Stringent post processing reduced this number to >7,000 putative SNPs. Over 85% (94/110) of a sample of these putative SNPs were successfully validated via Sanger sequencing. Based on this validation rate, this pilot experiment conservatively identified >4,900 valid SNPs within >2,400 maize genes. These results demonstrate that 454-based transcriptome sequencing is an excellent method for the high-throughput acquisition of gene-associated SNPs.

\textsuperscript{1}Primary author: developed experiment, applied PolyBayes and analyzed data

\textsuperscript{2}Primary author: developed experiment, developed computational validation and analyzed data
Introduction

SNPs (Single Nucleotide Polymorphisms) are single base differences between haplotypes. Once discovered, SNPs can be converted into genetic markers that can be inexpensively assayed in a high-throughput manner [Gut (2001); Kwok (2001)]. Due to their abundance, it is possible to use SNP-based markers to generate very dense genetic maps [Rafalski (2002)]. Such maps can be used to conduct marker assisted selection (MAS) programs, construct the specific genotypes required for quantitative genetic studies, as well as to enhance our understanding of genome organization and function and address fundamental questions related to evolution and meiotic recombination. SNPs can also be used for genome wide linkage disequilibrium and associations studies that assign genes to specific functions or traits. Furthermore, transcript associated SNPs can be used to develop allele-specific assays for the examination of cis-regulatory variation within a species [Cowles et al. (2002); Bray et al. (2003); Guo et al. (2003); Pastinen et al. (2004); Stupar and Springer (2006)].

Although SNPs can be identified by sequencing candidate genes from a set of individuals that represent diversity in the species of interest, this is neither high throughput nor inexpensive. Alternative approaches used during the construction of the human SNP map included identifying sequence polymorphisms within overlapping BAC clones derived from different individuals and shotgun sequencing of genomic fragments [Sachidanandam et al. (2001)]. This approach is not, however, usually possible because most genome sequencing projects use DNA extracted from highly similar or inbred individuals. Instead, SNP-based markers are typically mined from whole genome sequences or Expressed Sequence Tags (ESTs) obtained from genetically diverse individuals. For example, SNPs have been identified by comparing genomic sequences from two or more genetically distinct inbred lines of mouse [Wiltshire et al. (2003)], the Indica and Japonica subspecies of rice [Feltus et al. (2004)], the Columbia and Landsberg ecotypes of Arabidopsis [Jander et al. (2002)], and different lines of maize [Yamasaki et al. (2005)]. EST collections from genetically dissimilar individuals have similarly been mined for SNPs in humans [Marth et al. (1999)], pine [Dantec et al. (2004)], barley [Kota et al. (2001, 2003)] cassava [Lopez et al. (2005)] and maize [Batley et al. (2003)].
The latest maize genetic map (IBM_IDP_bd map, ver4) contains over 3,000 gene-based PCR markers distributed across the 2.5 Gbp genome [Fu et al. (2006)]. Even so, this map is not dense enough to support high resolution mapping applications and association genetics, particularly given the decay of linkage disequilibrium outside of maize genes [Tenaillon et al. (2001); Ching et al. (2002)]. Additionally, because the maize inbred B73 is being hierarchically sequenced, a higher density genetic map would be invaluable for anchoring each sequenced Bacterial Artificial Chromosome or BAC contig to its proper place in the genome. Increasing the marker density of this crop therefore has applications for accurately assembling this highly complex genome and ultimately in improving agricultural traits.

Maize is genetically very diverse; SNP and indel polymorphism frequencies between inbred lines and landraces average one variation per 124 and 28 bases for coding regions [Ching et al. (2002)] or all associated regions [Tenaillon et al. (2001)], respectively. We were particularly interested in identifying SNPs between B73 and the inbred Mo17. These two inbreds represent two of the major heterotic groups and were historically the parental lines of much of the commercial corn grown in the U.S. These inbreds are also the parents of the IBM RILs (Recombinant Inbred Lines) that were used to develop the maize genetic community’s high-resolution genetic maps.

The size and complexity of the maize genome make it unlikely that a second inbred will be sequenced in the immediate future. Although there are currently over 650,000 maize EST sequences available in GenBank, nearly all of these were drawn from a small subset of inbred lines, principally B73, W23, and Oh43A. Hence, the identification of B73/Mo17 SNPs requires the development of Mo17 EST sequence resources. Although genome sequencing technology has become progressively more efficient, EST projects require substantial investments in library construction and sequencing efforts to achieve the overall coverage required to locate SNPs.

Recently, 454 Life Sciences reported a highly parallel DNA sequencing system that is 100 times faster than standard sequencing methods and is capable of providing over 20 Mbp of sequence in a single four-hour run [Margulies et al. (2005)]. Increased throughput comes at the expense of read length (100 bp average length) and the absence of clone pair information,
making it less attractive for whole genome sequencing of complex genomes. However, 454 sequencing of maize cDNAs obtained from Shoot Apical Meristem tissue isolated by laser capture microdissection (LCM; reviewed by Schnable et al. (2004)) has recently been shown to be an effective method for tagging tens of thousands of maize genes without cloning and its associated costs [Emrich et al. (2007a)]. Therefore, 454-based sequencing of the B73 and Mo17 SAM transcriptome was expected to provide a collection of diverse ESTs that could support high throughput computational identification of gene associated SNPs. Because 454 reads contain more sequence errors than do reads generated by traditional sequencing technology [Margulies et al. (2005)], it was not, however, clear whether 454-based ESTs could be used for SNP discovery.

Here, we describe the generation of over 280,000 Mo17 SAM ESTs using 454 sequencing technology, the development of an efficient computational SNP mining pipeline based on the POLYBAYES sequence polymorphism detection tool, and the subsequent identification of over 7,000 putative Mo17/B73 SNPs within expressed sequences, a subset of which has been experimentally validated.

**Methods**

Isolation of SAM mRNA and 454 sequencing

Maize SAM cDNA isolation, 454 sequencing and raw sequence processing was previously described [Emrich et al. (2007a)]. A single GS-20 run produced 260,887 (28.8 Mbp) and 287,917 (30.7 Mbp) B73 and Mo17 SAM ESTs, respectively.

Identification of B73 reference sequences for 454 ESTs

Mo17 454 ESTs were initially mapped to a specific contig or singleton (217,773 total) from the MAGI 3.1 partial genome assembly of the maize inbred line B73 [Fu et al. (2005)] using best BLASTN matches ($1e^{-8}$ minimum E-value). Although 'best hit' criteria were used, it is possible that some 454 ESTs align to paralogous genomic fragments, especially given the partial nature of the MAGI assembly. To compensate, we used POLYBAYES (see below),
which includes an internal paralog filter and should identify and discard these instances. These ESTs were also aligned to MAGIs using GeneSeqer and its maize-specific splice models [Usuka et al. (2000)] for display on the MAGI website (http://magi.plantgenomics.iastate.edu). Only alignments consisting of at least 50 bp in length and with identity greater than or equal to 95% over at least 80% of the length of the 454 EST were used to annotate genomic sequences.

**Multiple sequence alignments and SNP detection of 454 sequence data**

Custom PERL scripts were written to create a pipeline to process MAGI3.1 anchor sequences and their associated B73 and Mo17 454 EST sequences for detecting SNPs. Anchored MSAs were produced by CROSS\_MATCH with the following parameters: `discrep_lists` `-tags -masklevel 5 -gap_init -1 -gap_ext -1`. Low initiation (`-gap_init`) and gap extension (`-gap_ext`) were used to increase alignment tolerance between the short 454 ESTs and the unplaced MAGI3.1 genomic anchors. Sequence polymorphisms were detected by POLYBAYES using the following parameters: `-anchorBaseQualityDefault 34 -memberBaseQualityDefault 18 -maskAmbiguousMatches -nofilterParalogs -priorParalog 0.03 -thresholdNative 0.75 -screenSnps -considerAnchor -noconsiderTemplateConsensus -prescreenSnps -priorPoly 0.01 -thresholdSnp 0.5`. Default anchor quality values (34) were based on a previous assessment of sequence error rates within the MAGI3.1 assembly [Fu et al. (2005)]. Default quality values of 18 were assigned to the 454 reads. This corresponds to an error rate of $\approx 1/65$, which overcompensates for the error rate observed for current 454 sequencing [Margulies et al. (2005); Emrich et al. (2007a)]. Although each base within the 454 sequence reads is given a quality score, these scores are only reliable when confirmed within independent sequences covering the same region. Because CROSS\_MATCH aligns each sequence individually to the anchor during MSA construction, and POLYBAYES assesses base quality on an individual basis, use of a stringent default rather than the base quality information provided by 454 Life Sciences is expected to increase the accuracy of polymorphism detection.
SNP parsing

Mo17 and B73 are inbreds, and thus should be mono-allelic at every base position. Custom PERL scripts were written to parse the POLYBAYES output. POLYBAYES identifies indel polymorphisms. Because indels are a common form of 454 sequencing error, only base substitutions were considered during this analysis. MAGI3.1 assemblies contain a low frequency of base substitutions propagated during shotgun sequencing of the High C\textsubscript{0}t selected maize genomic DNA [Fu et al. (2004)]. High C\textsubscript{0}t selected maize DNA sequences account for only a portion of the MAGI3.1 assembly sequence, but unidentified base substitutions within these regions could increase the number of false polymorphisms detected. Strict parsing rules ensured that potential MAGI3.1 sequence errors were avoided when B73-454 EST sequences are present in the multiple alignment. In cases where B73-454-ESTs are not present in the multiple alignment, SNPs called within regions of the MAGI3.1 assemblies containing High C\textsubscript{0}t selected DNA were avoided.

Results

The shoot apical meristem ultimately gives rise to all above-ground tissues. Thus, it is expected that many rare and developmentally important transcripts are present in the SAM transcriptome. Indeed, we have demonstrated that 454 sequencing of maize SAM cDNA captures fragments of thousands of genes, including many that may be expressed only rarely or only in the SAM [Emrich et al. (2007a)].

Using 454 sequencing we previously generated from the B73 inbred a collection of 260,887 high-quality SAM ESTs with an average length of 101 bp [Emrich et al. (2007a)]. Using the same methodology a collection of 454 SAM ESTs was generated from the maize inbred Mo17. After trimming polyA/T tails the 287,917 resulting SAM ESTs from Mo17 had an average length of 100 and consisted of 30.7 Mbp.
Table 1  **Summary of Multiple Sequence Alignments (MSAs) between MAGI3.1 anchors and B73 and Mo17 454-ESTs.** For this analysis, 454 sequences were initially mapped using BLAST to individual MAGIs, which later served as the template on which these MSA were computed by CROSS_MATCH.

<table>
<thead>
<tr>
<th></th>
<th>All ESTs</th>
<th>All B73 ESTs</th>
<th>All Mo17 ESTs</th>
<th>Both B73 and Mo17 ESTs</th>
<th>Only B73 ESTs</th>
<th>Only Mo17 ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of MAGIs</strong></td>
<td>48,063</td>
<td>33,567</td>
<td>34,928</td>
<td>20,432</td>
<td>13,135</td>
<td>14,496</td>
</tr>
<tr>
<td><strong>Bases covered</strong></td>
<td>8,897,508</td>
<td>4,989,045</td>
<td>5,798,933</td>
<td>1,890,459</td>
<td>3,098,586</td>
<td>3,908,463</td>
</tr>
<tr>
<td><strong>Coverage Depth</strong></td>
<td>1.8X</td>
<td>2.3X</td>
<td>2.3X</td>
<td>8.4X</td>
<td>1.3X</td>
<td>1.3X</td>
</tr>
</tbody>
</table>

**Assignment of Mo17 and B73 SAM ESTs to maize genomic anchor sequences**

MAGIs are maize genomic sequence assemblies [Fu et al. (2005)] composed of gene-enriched B73 genomic survey sequences [GSSs; Whitelaw et al. (2003)]. Because these sequences are highly accurate [1 disagreement per 10,000 bp; Fu et al. (2005)] and comprehensive (>75% of all maize genes are present), they comprise an excellent collection of B73 reference sequences for SNP detection. Attempts were made to align each of the 260,887 B73 and 287,917 Mo17 454 ESTs to the MAGI version 3.1 partial maize B73 genome assembly using a two-step approach. The initial preprocessing step uses BLAST to save time and improve accuracy by grouping together individual 454 SAM ESTs that preferentially align to a single MAGI template. This analysis assigned 432,431 of the 454-ESTs (207,294 B73 and 225,137 Mo17) to 48,063 MAGIs (Table 1). Of these MAGIs 20,432 aligned to both B73 (N=120,662) and Mo17 (N=135,249) ESTs. An additional 14,496 and 13,135 MAGIs aligned to only Mo17 (N=89,888) or only B73 (N=86,632) ESTs, respectively. The MAGI assembly sequences identified above served as templates upon which associated 454-ESTs were multiple aligned by CROSS_MATCH (P. Green, unpublished).

Doing so produced 48,063 anchored multiple sequence alignments (MSAs) that covered a total of 8,897,508 MAGI template bases with 454-ESTs. Approximately 5M and 5.8M anchor template bases were sampled by B73 and Mo17, respectively, while slightly fewer than 1.9M bases were sampled by both inbreds (Table 1). The relative proportions and average sequence
Table 2  Average coverage of nucleotides represented within B73-454-ESTs, Mo17-454-ESTs and MAGI3.1 anchored multiple sequence alignments. Although only one Mo17 EST would be sufficient to predict a SNP, increased sampling depth as shown is expected to increase accuracy by filtering out sequencing errors in either the 454 or MAGI data. Depths not separated below were pooled for analysis.

<table>
<thead>
<tr>
<th>454-EST component depths</th>
<th>Number of nucleotides</th>
<th>Average Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mo17</td>
<td>B73</td>
<td></td>
</tr>
<tr>
<td>1X</td>
<td>≥1X</td>
<td>1,092,570</td>
</tr>
<tr>
<td>≥1X</td>
<td>1X</td>
<td></td>
</tr>
<tr>
<td>2X</td>
<td>≥2X</td>
<td>326,095</td>
</tr>
<tr>
<td>≥2X</td>
<td>2X</td>
<td></td>
</tr>
<tr>
<td>≥3X</td>
<td>2X</td>
<td>134,386</td>
</tr>
<tr>
<td>≥3X</td>
<td>≥3X</td>
<td>471,794</td>
</tr>
</tbody>
</table>

depths (coverage) of the 1.9 Mbp MAGI nucleotides sampled by B73 and Mo17 454-ESTs are presented in Table 2. Although it is theoretically possible to identify putative B73/Mo17 SNPs across the entire region of the MAGI3.1 sequence space that was simultaneously sampled by B73 and Mo17 454-ESTs (≈1.9 Mbp), analysis of those regions that contain deeper sequencing coverage for both inbreds is expected to yield putative SNPs that are more likely to be valid. We therefore defined a high confidence set of bases on the MAGI anchor that was sampled to ≥ 3X by Mo17 ESTs and to ≥ 2X by B73 ESTs. With the inclusion of the MAGI3.1 anchor sequence (B73), these bases are sampled a minimum of 3 times for both inbreds. This set comprises 42% (606,180) of the simultaneously sampled sequence space (Table 2).

Polymorphism detection with POLYBAYES

Putative SNPs were identified from the MSA using the POLYBAYES polymorphism software package [Marth et al. (1999)]. POLYBAYES uses a Bayesian statistical model that considers depth of coverage, sequence quality and an a priori expected polymorphism rate to determine the probability that polymorphic sites within an MSA are SNPs rather than disagreements resulting from either sequencing errors or the alignment of paralogous (rather than
allelic) sequences [Marth et al. (1999)]. 454 sequencing technology is susceptible to indel-type errors [Margulies et al. (2005)] and the resulting ESTs exhibit an overall rate of sequencing error of approximately 1.5% [Emrich et al. (2007a)]. To address the issue of indel-type errors, we used MAGI assemblies as templates on which 454 SAM ESTS were aligned. Template-based MSAs such as these are often correct even in the presence of abundantly expressed or alternatively spliced transcripts [Marth et al. (1999)], and are therefore more likely to overcome the technical issues associated with 454 ESTs.

POLYBAYES identifies single base substitutions as well as single base insertions and deletions. However, because of the high number of indel errors associated with 454 technology [Margulies et al. (2005)] only base substitutions (i.e., SNPs) were considered in the current analysis. Initially, a total of 36,006 putative SNPs (p=0.5) were detected within 9,980 unique MAGI anchor sequences. This number of putative SNPs is expected to overestimate the diversity present in SAM expressed genes in the two maize inbreds. Because Mo17 and B73 are inbreds they should be mono-allelic at every base position, with relatively rare exceptions caused by nearly identical paralogs [NIPs; Emrich et al. (2007b)]. Hence, the observation that many of the putative SNPs discovered initially are multi-allelic within Mo17, B73 or both, suggests that many are false positives due to sequencing errors. With this in mind, we purposefully set the SNP probability low (p = 0.5) and filtered the putative SNPs using the following rules designed to substantially decrease the rate of false positives within the context of this study:

1. Polymorphic sites require a minimum of 2X representation in the Mo17-454-ESTs.

2. All Mo17 base calls at sites that were polymorphic between Mo17 454 ESTs and the B73 MAGI anchors were expected to be identical. This ensures mono-allelism within the Mo17-454-ESTs.

3. When B73-454-EST sequences also align across polymorphic sites that pass Rules #1 and 2, all of the B73-454-ESTs and the MAGI3.1 anchor base calls must agree. This avoids polymorphisms resulting from incorrect MAGI base calls or NIPs within B73.

4. To reduce the possibility of a erroneous base in the MAGI anchor mimicking a true SNP,
regions of the MAGI assemblies composed of sequences from High $C_0^t$ selected clones that are not covered by B73 ESTs were avoided because 40% of High $C_0^t$ clones contain cloning artifacts that mimic SNPs [Fu et al. (2004)].

Applying these stringent rules to the raw SNP data returned 7,016 putative B73/Mo17 SNPs distributed among 3,403 MAGIs. The numbers of 454 ESTs that cover these polymorphic sites range from only two Mo17-454-ESTs to at least three B73 and three Mo17 ESTs (Table 3).

For completeness, Table 3 presents all polymorphism data. The total numbers of polymorphic bases sampled by only one or two Mo17-454-ESTs and/or B73-454-ESTs are displayed in rows 1 and 2, respectively; these were removed from further consideration. The numbers of putative SNPs that pass the above rules and their associated MAGIs are presented in Rows 3-12 of Table 3. Rows 3 and 4 illustrate the total number of polymorphic sites sampled simultaneously by a minimum of 3 Mo17-454-ESTs, 2 B73-454-ESTs and the B73 MAGI3.1 anchor. This represents the highest confidence data set, with a minimum sampling depth of 3X for both inbreds. Rows 5-12 display putative SNPs at sites with decreasing depths of coverage, which are expected to represent decreasingly confident data sets. This expectation is supported by their corresponding POLYBAYES assigned SNP probabilities (pSNP) (Supplemental Materials). The number of potential SNPs, the number of their associated MAGI anchors for each B73/Mo17 sampling depth, and the total number (additive) of potential SNPs and the number of unique MAGIs anticipated by systematically including data sets (starting with row 3) is also presented in Table 3. In summary, after single 454 GS20 sequencing runs of B73 and Mo17 SAM cDNA our computational polymorphism mining strategy identified over 7,000 putative SNPs (published as Supplementary Material in the *Plant Journal*).

**Validation of SNPs**

A set of 110 putative B73/Mo17 SNPs were subjected to validation by sequencing via Sanger technology the corresponding alleles that had been PCR amplified from B73 and Mo17 genomic DNA. Detailed results of these validation experiments were published as Supplemen-
Table 3  Number of putative SNPs, depth at each SNP site by inbred, and estimates of the total number of maize genes that contain at least one putative SNP between the B73 and Mo17 inbreds in this SNP dataset. Polymorphic bases sampled by only one Mo17-454-EST displayed in rows 1 and 2 were removed from consideration. In contrast, rows 4 and 5 illustrate polymorphic sites with a minimum sampling depth of 3X for both inbreds and, as a result, have the highest confidence. Subcategories not separated by lines were pooled for analysis.

<table>
<thead>
<tr>
<th>454-EST component depths of MSAs</th>
<th>No. putative SNPs</th>
<th>No. MAGI3.1 anchors&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Additive SNP No.</th>
<th>Additive minimum estimate of SNP-containing genes&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mo17</td>
<td>B73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1X</td>
<td>1X</td>
<td>1,762</td>
<td>1,154</td>
<td>–</td>
</tr>
<tr>
<td>1X</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2X</td>
<td>≥2X</td>
<td>1,648</td>
<td>1,039</td>
<td>–</td>
</tr>
<tr>
<td>≥2X</td>
<td>2X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥3X</td>
<td>≥3X</td>
<td>1,452</td>
<td>900</td>
<td>1,452</td>
</tr>
<tr>
<td>≥3X</td>
<td>2X</td>
<td>565</td>
<td>404</td>
<td>2,017</td>
</tr>
<tr>
<td>≥3X</td>
<td>1X</td>
<td>717</td>
<td>513</td>
<td>2,734</td>
</tr>
<tr>
<td>2X</td>
<td>≥3X</td>
<td>537</td>
<td>372</td>
<td>3,271</td>
</tr>
<tr>
<td>2X</td>
<td>2X</td>
<td>546</td>
<td>363</td>
<td>3,817</td>
</tr>
<tr>
<td>2X</td>
<td>1X</td>
<td>1,045</td>
<td>707</td>
<td>4,862</td>
</tr>
<tr>
<td>≥3X</td>
<td>0</td>
<td>481</td>
<td>283</td>
<td>5,353</td>
</tr>
<tr>
<td>2X</td>
<td>0</td>
<td>1,673</td>
<td>830</td>
<td>7,016</td>
</tr>
</tbody>
</table>

<sup>a</sup>MAGIs are gene-enriched assemblies that are likely to contain only a single gene or gene fragment.

<sup>b</sup>Numbers represent a non-redundant collection at each row.
tary Material in the *Plant Journal*. The overall rate of validation was over 85% (94/110). Most of the SNPs selected for tested represent sites with at least moderate levels of B73/Mo17 coverage. Over 88% (85/96) of SNPs sampled by $\geq$ Mo17-454-ESTs and $\geq$ 2 B73-454-ESTs (Table 3, rows 3 and 4) were validated. Fewer of the lesser confidence SNPs were assayed; these exhibit a collective validation rate of 64% (9/14). Using the above validation rates and data presented in Table 3, the number of SNPs that could be validated was estimated. This analysis suggests that 4,984 computationally identified B73/Mo17 SNPs represent 'true' polymorphisms, and that these are distributed within 2,472 MAGIs. The average sizes of the MAGI assemblies suggest they contain only one (or a portion of one) maize gene. Because these polymorphisms were mined from cDNA sequences derived from mRNA and conservatively filtered, we estimate that this analysis identified at least 4,900 valid SNPs within at least 2,400 maize genes.

**Discussion**

Once discovered, SNPs have a wide variety of applications in biological research. One means to discover SNPs is to align ESTs from more than one genotype. LCM-454 sequencing enables efficient deep sampling of ESTs obtained from specific cell-types [Emrich et al. (2007a)], but suffers from the disadvantage of higher error rates than Sanger sequencing. Even so, this study demonstrates that it is possible to use ESTs obtained via LCM-454 sequencing to achieve high-throughput SNP discovery. Over 260,000 Mo17 ESTs were obtained from a single GS-20 sequencer run on cDNA isolated from SAM tissue, and over 7,000 putative SNPs were identified relative to B73 genomic and 454 EST sequences. A subset of these SNPs were validated via direct sequencing of PCR products amplified from B73 and Mo17 genomic DNA.

Putative SNPs are identified as mismatches between aligned sequences, and several computational tools for SNP identification are available [Nickerson et al. (1997); Marth et al. (1999); Manaster et al. (2005); Wang and Huang (2005); Weckx et al. (2005); Zhang et al. (2005)]. Our SNP discovery pipeline implements POLYBAYES, which has been used to identify SNPs in several studies [Marth et al. (1999); Useche et al. (2001); Dantec et al. (2004); Pavy et al. (2006)]. We assigned default values to 454 sequences based on an empirical evaluation of the
Figure 1 A portion of the CROSS_MATCH produced, template-driven, padded alignment between B73 and Mo17 454 EST sequences and the high-quality MAGI105195 sequence assembly constructed from B73 maize GSS sequence that serves as an alignment template. The G/A polymorphism occurs at position 2846 of the template. Because these insertions are not supported by other sequences they are easily identified as errors by the POLYBAYES pipeline and are not called as polymorphisms.

base error rate rather than using the relatively new 454 quality scores. As a result, sequence depth and relative allele proportions have the greatest influence on polymorphism detection and, based on this observation, potential SNPs were filtered by examining these statistics at each polymorphic site. The highest confidence polymorphisms are those that are minimally covered by both Mo17 and B73 sequences to 3X. Experimentally >88% of these sites could be validated as being polymorphic, and are assigned prior probability scores (P SNP) of at least 0.997 by POLYBAYES.

POLYBAYES is designed to use template-driven MSAs, in which sequences are scaffolded across a high quality template sequence that serves as an anchor. In addition to being highly accurate [Marth et al. (1999)], this approach eliminates the need to perform de novo assemblies of 454 ESTs that are complicated by the short lengths of 454 reads. Furthermore, gaps and insertions in this template-driven multiple sequence alignment approach are propagated throughout all members so 454 semi-random indels can be easily identified and ignored (Figure 1). Finally, the ability of POLYBAYES to use quality scores during SNP detection provides
the option of utilizing 454 sequence calls once they are better accepted by the research community, or if Sanger sequences are also used, or if the base accuracy of the template is suspect. In all of these cases, the availability of accurate base quality data could improve the accuracy of SNP detection.

We estimate that our SNP collection contains at least 4,984 valid SNPs within 2,472 genes. This estimate is based on an observed validation rate of 0.88 for polymorphic sites minimally sampled to 3X by each inbred and the assumption that all other depth classes of polymorphism have a conservative validation rate of 0.64. A subset of 2,017 high confidence SNPs were detected within B73 genomic sequence that was sampled by a minimum of 2 B73-ESTS and a minimum of 3 Mo17 454-ESTs (Table 3). The size of this reduced sequence space is 621,956 bp (Table 2), providing an observed polymorphism rate of at least 1/300. This rate is only about one half of that previously reported in maize coding sequence [Ching et al. (2002)]; however, this published rate is based on only 18 genes and may not be representative of the genome. Furthermore, the conservative parameters used in this study are expected to underestimate polymorphism rates. Specifically, in the absence of 454 quality information we required that B73 and Mo17 inbreds both be monoallelic at each nucleotide before calling a putative SNP. In fact, 17,671 instances where either inbred (or both) exhibits bi-allelism were initially ignored to simplify polymorphism detection and subsequent validation. These were further parsed to identify putative SNPs where the B73 and/or Mo17 major allele frequencies are ≥ 0.75, and each major allele is represented at least 3 times within the MSA. There are 879 such cases (Supplemental Table 1) that if all were validated would increase our polymorphic rate to at most 1/214 bp.

All of the polymorphic sites discussed in this study were detected by comparing the sequences obtained from single 454 GS-20 sequencer runs on cDNA obtained from Mo17 and B73 SAM tissue. Additional sequencing runs would be expected to increase the proportion of the transcriptome sequence space covered, and perhaps most importantly, increase the overall depth of coverage. Consequently, additional sequencing runs would be expected to increase the confidence of at least a fraction of the putative SNPs that are currently poorly supported due
to insufficient sampling depth (Table 3). Increased depth would also lend additional support to the identification of NIPs, a process that is particularly dependent on deep sampling.

Maize is a globally important crop and a model system for the study of genome structure, evolution, and genetics. Between 5,000 and 10,000 years ago the wild grass teosinte was domesticated to produce modern maize. Domestication resulted in a population bottleneck that reduced allelic diversity in maize relative to teosinte. Over the past decade, the analysis of DNA sequence polymorphism data to detect signatures of genes that were involved in domestication and subsequent selection has become a well-established approach \[e.g. \] Wang et al. (1999); Whitt et al. (2002); Tenaillon et al. (2004); Wright et al. (2005); Yamasaki et al. (2005).

The maize genome is composed of approximately 2.5 billion bases and contains an estimated 50,000 genes [Fu et al. (2005)]. The vast majority of this genome is composed of a small number of highly repetitive retrotransposons [Bennetzen (1996); SanMiguel et al. (1996); Meyers et al. (2001); Whitelaw et al. (2003)]. Hence, it has not been economically feasible to conduct whole-genome scans for SNPs by sequencing multiple maize haplotypes. But the 454 EST-based SNP mining procedure described here, which is focused on a specific transcriptome using LCM, provides the underpinnings for a high-throughput SNP discovery platform than could be used to cost effectively identify genes that exhibit signatures of having been involved in the domestication or improvement of maize and other large-genome crops, and that would therefore be potential targets for improving agriculturally relevant traits.

Acknowledgments

We thank Ruth Swanson-Wagner (Iowa State University) for preparing the SAM-specific Mo17 mRNA used for 454 sequencing. This research was supported by grants from the National Science Foundation (DBI-0321595, DBI-0321711, and DBI-0527192), ISU’s Plant Science Institute and Pioneer Hi-Bred International, Inc.; additional support was provided by the Hatch Act, State of Iowa and the Donald Danforth Plant Science Center.
References


An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Research*, 14:1812–1819.


tion and genotyping of single nucleotide substitutions using fluorescence-based resequencing. 


GENERAL CONCLUSIONS

Molecular biology continues to shift toward data-driven hypotheses. Small questions related to a specific pathway or process are still undertaken; however, these experiments are often aided and sometimes greatly improved based upon the availability of complete genome sequences. To accomplish this goal, an intimate alliance has been formed between traditional biologists and computer scientists to process and analyze large-scale sequence data resources.

Maize, which is discussed in depth in this thesis, has importance both as an important food crop and a feedstock for ethanol production. While future projects are unlikely to be more complex than the human and maize genomes, these approaches will still require a significant amount of computational resources. The parallel clustering framework presented in this thesis has begun to shift this burden to supercomputing platforms that can effectively address these computationally demanding problems without the complete overall of well-tested assembly techniques.

As the benefit of genomics-driven biology becomes more and more important, there is a push to provide rudimentary survey sequences of the genes within a species until resources can be allocated to completing the entire genome. For this reason genome reduction and EST sequencing will continue to represent valuable approaches, especially for non-model organisms whose genomes are unlikely to be sequenced in the near future. Unfortunately, both of these sequencing approaches are subject to non-uniformity that further aggravate the computational resources needed to solve these problems. We believe that a cluster-then-assemble is a viable solution and demonstrated its usefulness toward the creation of the highly accurate maize assembled genomic islands (MAGIs) along with performing large-scale 454 EST analysis.
There appear to be at least two intriguing challenges based upon the ideas we have explored within this thesis. First, as the availability of diverse large-scale sequencing data continually accelerates, can the scientific community devise accurate and appropriate algorithms and experiments to extract as much information as possible on diverse topics including molecular evolution, functional genomics, and comparative biology to name only a few. It is our belief that an increased push towards high performance computing to answer many of these questions is essential and is the goal of our future work. Second, would the sequencing capabilities and algorithms ever improve to the point that genome assembly may well become inexpensive? This question remains to be answered.