Available online at www.sciencedirect.com
ScienceDirect
Analytical Biochemistry 373 (2008) 78–87
ANALYTICAL BIOCHEMISTRY
www.elsevier.com/locate/yabio

# An automated, high-throughput sequence read classification pipeline for preliminary genome characterization

Philippe Chouvarine [a,b], Surya Saha [a,c], Daniel G. Peterson [a,b,d,*]

[a] *Mississippi Genome Exploration Laboratory, Mississippi State University, Mississippi State, MS 39762, USA*
[b] *Department of Plant & Soil Sciences, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA*
[c] *Department of Computer Science & Engineering, Mississippi State University, Mississippi State, MS 39762, USA*
[d] *Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA*

Received 9 July 2007
Available online 10 August 2007

**Abstract**

In the absence of a complete genome sequence, considerable insight into genome structure can be gained from survey sequencing of genomic DNA. To facilitate high-throughput characterization of genome structure based on shotgun sequence reads, we have developed an automated sequence read classification pipeline (SRCP). The SRCP uses a battery of novel and standard sequence analysis algorithms along with a sophisticated *decision tree* to place reads into ''best fit'' functional/descriptive categories. Once ''primed'' with genomic sequence data, the SRCP also permits estimation of gene/repeat enrichment afforded by reduced-representation sequencing techniques. To our knowledge, the SRCP is the only tool that has been designed to provide a description of a genome or a genome component based on sample sequence reads. In an initial test of the SRCP using sequence data from *Sorghum bicolor*, it was shown to provide results similar in quality to results generated by manual classification. Although the SRCP is not a replacement for manual sequence characterization, it can provide a rapid, high-quality overview of genome sequence content and facilitate subsequent annotation. The SRCP presumably can be adapted for analysis of any eukaryotic genome.


*Keywords:* DNA; Sequence analysis; Transposon; Genome; Bioinformatics; Computational analysis; Genomics; Comparative

Although complete genome sequencing represents an ideal means by which the genomes of organisms can be compared, it is not currently economically feasible for most eukaryotes. This is especially true for the numerous organisms that have large, highly repetitive genomes including many important plants and animals. With this said, sample sequencing of random genomic DNA can be used to gain considerable information about genome structure in lieu of a complete sequence [1,2]. However, it is often difficult for researchers to characterize the sequences they have obtained, especially if they have generated large sequence data sets for organisms for which previous sequencing research has been limited.

At present, numerous automated and semiautomated gene characterization programs are available [3,4]. Likewise, there are a growing number of programs designed to characterize repetitive elements [5–7]. However, to our knowledge, there is no program or pipeline designed to provide an overview of the sequence composition of an entire genome based on shotgun sequence reads. To permit such characterization, we have constructed a sequence read classification pipeline (SRCP)[1] in which a battery of exist-

* Corresponding author. Address: Department of Plant & Soil Sciences, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA. Fax: +1 662 325 8742.
*E-mail address:* dpeterson@pss.msstate.edu (D.G. Peterson).

[1] *Abbreviations used:* SRCP, sequence read classification pipeline; NCBI, National Center for Biotechnology Information; EST, expressed sequence tag; EMC, EST/mRNA/cDNA; BLAST, Basic Local Alignment Search Tool; TIGR, The Institute for Genomic Research; BLAT, BLAST-like alignment tool; XML, Extensible Markup Language; IIS, Internet Information Services; ASP, Active Server Pages; DTS, Data Transformation Services; FTP, File Transfer Protocol; SQL, Structured Query Language; XSLT, Extensible Stylesheet Language Transformations.

0003-2697/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2007.08.008

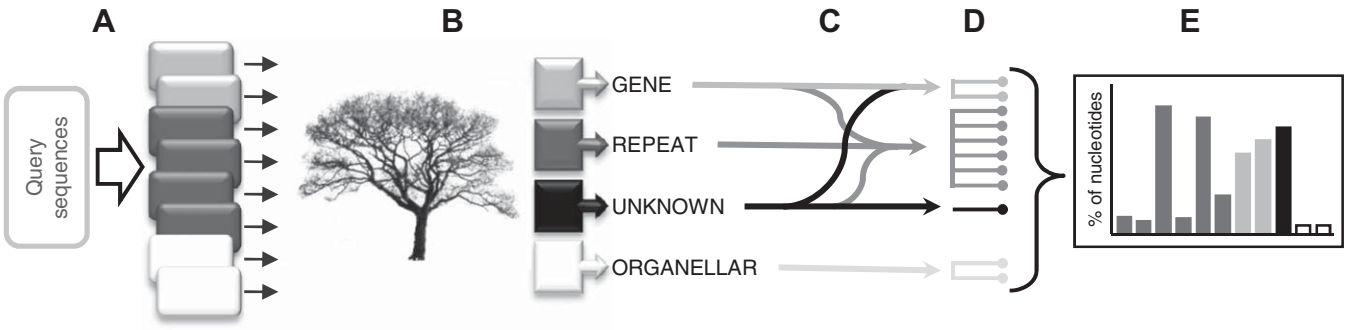


Fig. 1. General overview of the sequence read classification pipeline (SRCP). (A) Query sequences are compared using BLAST (Basic Local Alignment Search Tool) to the contents of two gene (light gray rectangles), four repeat (dark gray rectangles), and two organellar (white rectangles) highly curated, local sequence databases. (B) For each query sequence, data from the BLAST analyses are evaluated with a decision tree algorithm that places that sequence into a ''best fit'' descriptive gene, repeat, or organellar DNA category; those sequences that do not possess significant homology to sequences in any of the local sequence databases are classified as unknown. (C) Two independent algorithms interrogate those sequences classified as gene or unknown to see if they are possibly repetitive based on their frequency within the data set. Additionally, the unknown sequences are analyzed with *tblastn* to determine if they share significant homology with nontransposon proteins. Based on these secondary analyses, some query sequences are reclassified. (D) Each query sequence is placed into one of 11 final sequence categories. (E) The output of the SRCP is a graph (along with data and statistics) illustrating the composition of the query sequence set.

ing and novel algorithms are used to place random genomic query sequences into descriptive/functional sequence categories. The SRCP calculates the fraction of base pairs in each category, thus providing an overview of genome structure while facilitating initial annotation of query sequences (Fig. 1). In addition, the efficacy of reduced-representation sequencing techniques [8,9] can be assessed by comparing SRCP results for random genomic sequence with SRCP results for gene- or repeat-enriched DNA. With respect to basic configuration, the SRCP uses the program BLAST (Basic Local Alignment Search Tool) to query sequences against highly curated, custom local databases. The BLAST data are filtered, stored in a relational database, and analyzed to derive the final classification of each query sequence. The results of the analysis are available via a Web interface. The system is implemented as a series of Perl scripts, database scripts/queries, and dynamic Web pages.

## Materials and methods

### *General considerations*

1. Because our research is focused primarily on study of seed plants (Phylum Spermatophyta), we developed the SRCP for analysis of sequences from spermatophytes. However, the basic SRCP structure can be adapted for study of any organism or group of organisms.
2. The different sequence categories in the SRCP are based on those used by Peterson and co-workers [10].
3. The addresses of public Web pages and databases not generated as part of our research are given in Table 1.
4. Interested parties can obtain source codes and/or downloads of novel tools and access the contents of our local sequence databases at http://www.mgel.msstate.edu/tools.htm.

Table 1
Database and Web page addresses

| Database or Web page | Web address |
|---|---|
| NCBI | www.ncbi.nlm.nih.gov |
| Core Nucleotide DB | www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nuccore |
| EST DB | www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucest |
| Entrez Help Document | www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html |
| Display Formats | www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Display_Formats |
| Plastid Organelles | www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plastids.html |
| Viridiplantae Mitochondria | www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plants.html |
| The Inst. for Genomic Res. | www.tigr.org/ |
| TIGR Gene Index FTP site | ftp://ftp.tigr.org/pub/data/tgi/ |
| Canad. Bioinf. Help Desk | gchelpdesk.ualberta.ca |

5. The version of BLAST (Linux-ia32, Version 2.2.14) used in this pipeline was obtained from the National Center for Biotechnology Information (NCBI).

### *Technologies*

Traditionally, bioinformatics projects have used Linux/Unix platforms. However, there are a number of powerful and often neglected Windows-based software development technologies that afford rich functionality without extensive de novo programming. For this research, we developed a hybrid Linux and Windows system to use the

strengths of both operating systems. The power of the Linux operating system lies in its robustness, scalability, and high availability of compatible bioinformatics software. Therefore, we chose to run Linux on the computational server that runs bioinformatics tools. With respect to Windows tools, our database server runs SQL Server 2000 (SQL = Structured Query Language), and we use its built-in Data Transformation Services (DTS) for bulk upload of large XML (Extensible Markup Language) files containing BLAST results. We also use DTS to implement the classification logic of the pipeline (see below). Our Web server runs IIS (Internet Information Services) 6.0, which provides powerful native lock-down mechanisms. The freely available URLScan program (http://www.microsoft.com/technet/security/tools/urlscan.mspx) can be used to secure all versions of IIS. Running IIS allows us to use ASP.NET (ASP = Active Server Pages) for our Web interface. ASP.NET provides a collection of powerful and easily customizable Web controls, most notably the "data grid" control, which is ideal for displaying large data sets in a table structure with editable cells.

*Populating the repeat and organellar local databases*

For all repeat and organellar sequences, we currently download sequence information in the GenBank file format, which includes not only the sequence, its accession number, and its title, but detailed annotation and Internet links.

Spermatophyte transposon, rDNA, and centromere sequences were extracted from the NCBI Core Nucleotide Database by conducting searches using boolean text strings (Supplementary Table 1). Search results were used to create Transposon, rDNA, and Centromere local databases.

Chloroplast genome sequences were downloaded from NCBI's Plastid Organelles page and placed in the Chloroplast local database. Spermatophyte mitochondria sequences were downloaded from NCBI's Viridiplantae Mitochondria page and placed in the Mitochondria local database.

Each local database was assigned a version number containing the date it was populated and a two- or three-letter abbreviation indicating its contents (e.g., the first version of the Mitochondria local database was designated MC_2005-10-01). We update these local databases every 6 months.

Because many repeat sequences are found as annotated sections within larger genomic sequence entries (i.e., are not archived as individual GenBank entries), we developed a Perl script that extracts repeat regions and their annotations from select GenBank files. Extracted repeats were placed in an Annotated Repeat local database. Because of the large number of annotated repeats in plant whole-genome sequences, for this initial test we limited our extraction to manually annotated sequences available for *Sorghum bicolor*.

*Populating the "gene sequence" local databases*

Spermatophyte EST, cDNA, and mRNA (EMC) sequences were originally extracted from the NCBI EST Database and Core Nucleotide Database by conducting searches using a boolean search string (Supplementary Table 1). Because of the relatively large number of retrieved sequences, sequence data were downloaded in FASTA format [11] rather than in GenBank format. Downloaded sequences then were BLASTed (*blastn*) against the Chloroplast, Mitochondria, rDNA, Centromere, and Transposon local databases (see above). Any sequence exhibiting a significant hit (bit score = $S' \geqslant 60$) to one of these local databases was eliminated from the data set by Perl scripts. The remaining sequences were deposited in the EMC local database.

Spermatophyte "gene" sequences in FASTA format were downloaded from The Institute for Genome Research (TIGR) Gene Index FTP (File Transfer Protocol) site. Downloaded files were then scanned using a Perl script that eliminates those entries containing the following "repeat-affiliated" words in their titles (where asterisks indicate wild-card characters): retrovirus, retroelement, transpos*, gag, pol, polyprotein, env, reverse transcriptase, integrase, stowaway, MITE, miniature, copia, gypsy, RT, helitron, maverick, polinton, mul*, insertional, mitocondri*, chloroplast, capsid, and nucleocapsid. Remaining sequences were then BLASTed against the Annotated Repeats, Chloroplast, Mitochondria, rDNA, Centromere, and Transposon local databases. Sequences exhibiting a significant hit ($S' \geqslant 60$) to one or more of these databases were eliminated using the Perl scripts mentioned above. The remaining sequences were deposited in the Gene Index local database.

*Preparation of query sequences*

Random *S. bicolor* genomic shotgun sequences (GenBank Accession Nos. CW512190–CW514008) [12] were used as a sample "unfiltered" query sequence set. These 1819 sequences, collectively representing 1,088,783 bp, have a mean length of 599 bp (SE ± 38). To study the effect of sequence length on SRCP results, two representations of the sequence data were initially tested. The first representation contained the original GenBank sequences without any size adjustments (i.e., full-length query sequences); the second representation contained the same sequences digitally fragmented into 80- to 179-bp (average 105 bp ± SE 0.14) pieces, that is, short-length query sequences. The level of genome coverage of the short-length query sequence set was the same as that of the full-length query sequence set.

To further explore relationships between query sequence length and classification, a series of sequence subsets were prepared. Each subset contained DNA taken from the random *S. bicolor* genomic sequences used above. Names and details of the subsets are given in Supplementary Table 2.

To examine the ability of the SRCP to estimate gene and/or repeat enrichment afforded by Cot filtration (a

reduced representation sequencing technique), Cot-filtered sequences manually classified by Peterson et al. [10] (GenBank Accession Nos. AZ921847–AZ923007) were categorized by the SRCP following analysis of the unfiltered query sequences (see below).

*Analysis of random genomic DNA query sequences*

The basic steps in analysis of random genomic query sequences are outlined in Fig. 1. Specifics are illustrated in Fig. 2 and further detailed below.

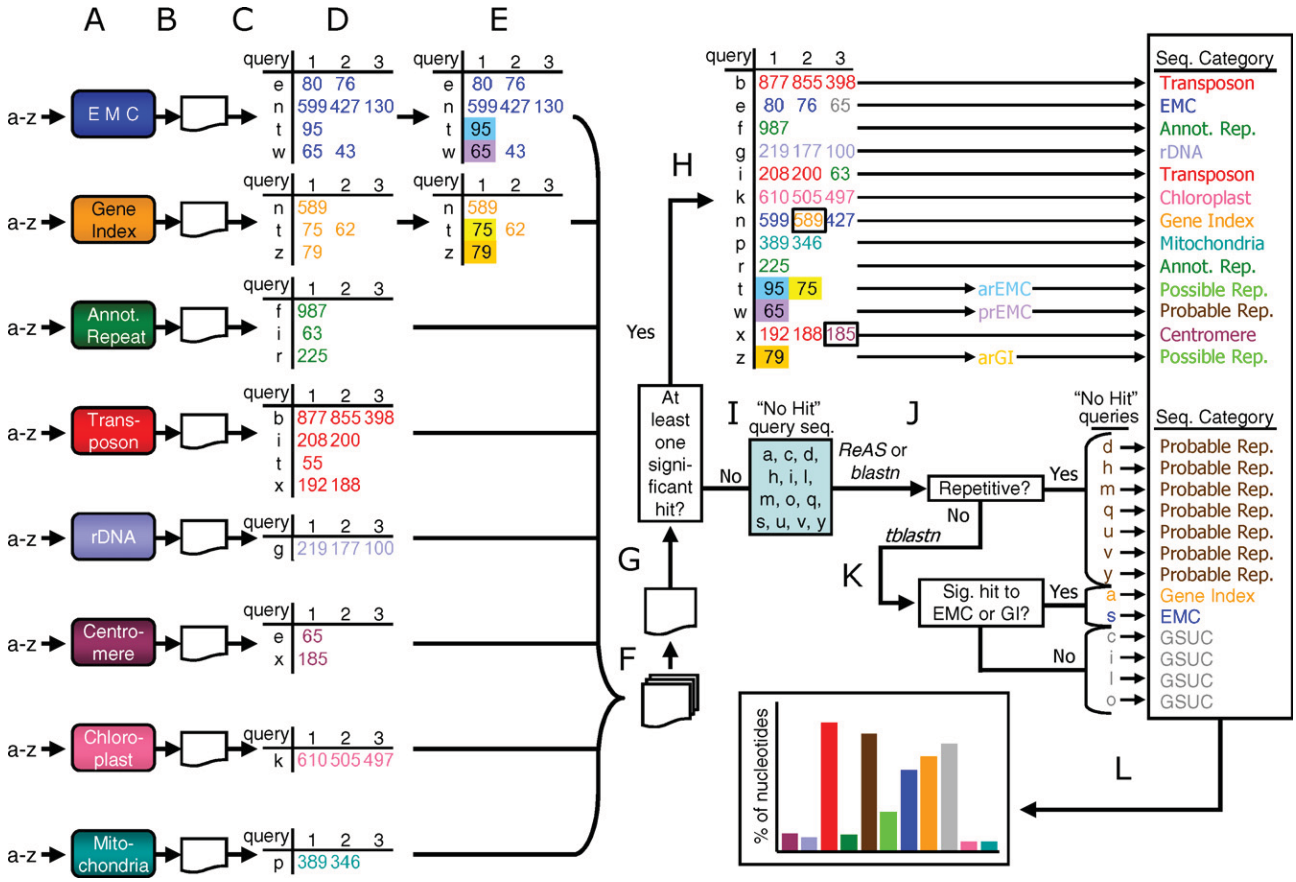


Fig. 2. Steps in categorization of random genomic query sequences. (A) A query sequence set is compared with sequences in the eight local sequence databases. In the diagram, the query sequence set is composed of 26 reads (represented by the lowercase letters a–z). BLAST (Basic Local Alignment Search Tool) parameters are set so that only the three most significant hits (if applicable) for each query sequence are recorded. (B) A Perl script removes unnecessary text and eliminates all hits with bit scores ($S'$) $< 45$ from the BLAST output files. (C) A script uploads the resulting "summary" files to an SQL Server database. (D) In the SQL database, the BLAST results from each local sequence database are stored in their own data table. In the diagram, each BLAST results table lists only the names of query sequences that produce a hit to a sequence in that local sequence database (left most column) and the bit scores of each query sequence's (up to three) most significant hits. In reality, the data tables contain highly detailed information including each hit's accession number(s), annotation, and alignment information with the query sequence. (E) As a means of detecting repetitive sequences in the EMC (EST/mRNA/cDNA) and Gene Index (GI) local sequence databases, an UPDATE query analyzes the EMC and Gene Index query BLAST data tables to see if multiple query sequences are recognizing the same local database entry, an indication that the entry and the query sequences may represent repetitive elements. On the basis of this analysis, some query sequences are marked as "Ambiguous Repetitive" (ar) or "Probable Repetitive" (pr). In the diagram, arEMCs, prEMCs, arGIs, and prGIs are represented by light blue, violet, gold, and yellow cells, respectively. (F) A UNION query integrates the information from all eight BLAST data tables. (G) A nested SELECT query eliminates hits with bit scores <60 and selects the best three hits from all of the data tables for each query sequence. Each query sequence with at least one $S' \geqslant 60$ hit is included in the query result set. (H) A decision tree assigns each query sequence in the query result set to a descriptive sequence category based on the (up to three) best hits for that sequence. The decision-making process is relatively complex. Rectangles mark instances in which the decision tree assigns a query sequence to a sequence category that differs from the name of the local sequence database to which that sequence shows its most significant hit. For simplicity, all query sequences that are "called" arEMCs or arGIs are assigned to the "Possible Repeat" category, whereas all those "called" as prEMCs or prGIs are assigned to the "Probable Repeat" category. (I) Query sequences that produce no significant hits to any of the local sequence databases are assigned to the temporary "No Hit" group. (J) Depending on the level of genome coverage, either *ReAS* [7] or *blastn* is used to compare "No Hit" sequences to each other. Those query sequences marked as repetitive by *ReAs* or exhibiting significant homology ($S' \geqslant 60$) to a number of other "No Hit" query sequences in excess of a mathematically defined threshold are placed in the "Probable Repeat" category. (K) Remaining "No Hit" query sequences are electronically "translated" by a Perl script into proteins representing each of the six potential reading frames. The program *tblastn* is then used to compare the translated "No Hit" query sequences into translated versions of the EMC and GI local sequence databases. If a translated "No Hit" sequence produces a significant ($S' \geqslant 60$) *tblastn* hit to the EMC and/or GI local sequence databases, it is reclassified based on the highest of its bit scores. If the highest EMC and Gene Index bit scores are equal, the "Gene Index" classification is selected. "No Hit" sequences that are not classified in step J or K are placed in the "Genome Sequences of Unknown Character" (GSUC) category. (L) The query sequence set is displayed in a histogram showing the percentage of base pairs found in each sequence category.

*Blast*

An entire query sequence set is BLASTed against each of the local databases (Fig. 2A). We set *–b* and *–v blastall* flags to 3 to collect only the top three hits for each sequence, minimizing the sizes of the resulting XML files, which, depending on the number of query sequences, may otherwise become unmanageably large.

The output XML files are processed with a Perl script that creates summary XML files. At this point, hits that do not satisfy a certain minimal bit score threshold may be filtered out using this Perl script. Summary files are then used by DTS scripts to bulk upload the data to an SQL Server database based on the corresponding XML Schema Definition files. Results from each local database BLAST comparison are stored in their own table (Figs. 2B and C).

*First-round detection of repeat sequences*

A common means used to assess the gene content of a batch of query sequences is comparison of the query sequences with ESTs. However, such an approach requires considerable caution as EST databases often contain numerous repetitive DNA sequences. Some of these repeats are simply organellar, rDNA, or genomic repeat sequences that were not eliminated during the mRNA isolation process. Others are the expressed regions of transposons such as retroelement genes. Because transposons are typically found in numerous copies per genome and contain only genes that promote their own propagation or movement, they are typically classified as repeats. Repeats are eventually "weeded out" of most EST databases, although it may be many years before the culling process is complete.

Recognition of the same EST database entry by multiple genomic query sequences is one means by which query sequence repetitiveness has been estimated and repeat sequence contaminants have been identified in "low-copy-sequence" databases [10,13,14]. In this regard, several SQL queries were used to identify EMC local database entries that were the top significant EMC hit for multiple query sequences. Assuming that query sequences in the EMC BLAST table represent single-copy genes, the average number of times a query sequence would represent a given gene can be predicted by dividing the number of query sequences in the EMC BLAST table by the predicted number of genes for the test organism. For example, in our analysis of the full-length sorghum query sequences, 972 query sequences exhibited their most significant hit ($S' \geqslant 60$) to the EMC local database. If sorghum has roughly 25,000 nonrepetitive gene sequences like *Arabidopsis* [15], the average expected number of hits by an EMC-recognized query sequence to any one of the hypothetical sorghum genes is ($972 \div 25{,}000 =$) 0.0389. The probability of multiple EMC-recognized query sequences recognizing a particular "single-copy EST" (≈gene) sequence by chance can be roughly estimated using the Poisson probability distribution function,

$$P(X) = \mu^x \div (e^{\mu} X!),$$

where $P =$ probability, $X =$ number of occurrences, and $\mu =$ is the population mean number of occurrences in a unit of space or time [16]. If $\mu = 0.0389$ (see above), the probabilities of two, three, four, and five EMC-recognized query sequences tagging the same single-copy EST by chance are $7.3 \times 10^{-4}$, $9.4 \times 10^{-6}$, $9.2 \times 10^{-8}$, and $7.1 \times 10^{-10}$, respectively.

In our implementation, the first value of $X$ to produce a $P(X)$ less than 0.01 can be represented by the variable $Y$. SQL queries mark a query sequence as an "Ambiguous Repeat EMC" if its most significant hit is to an EMC that is the most significant hit of $Y$ query sequences in the dataset. Any query sequence that has its most significant hit to an EMC that is the most significant hit for $> Y$ query sequences is classified as a "Probable Repeat EMC".

The repeat detection procedure is applied to the Gene Index local database as well with some query sequences being reclassified as "Ambiguous Repeat Gene Index" or "Probable Repeat Gene Index".

*Classification of query sequences with significant local database BLAST hits*

Local database BLAST results tables are combined in a UNION query. Query sequences with no significant local database hits are not included in the UNION query result set, but, rather, are given the temporary classification of "No Hit" and used to generate a corresponding FASTA file for further analysis (see below). For those query sequences with at least one significant local database hit, an SQL query (see Supplementary Materials, SQL Query) is used to determine the (up to) three best hits with bit scores $\geqslant 60$ for each query sequence from the UNION query result set (Figs. 2F–H).

A DTS script within SQL Server 2000 uses the output of the query above and runs it through a decision tree that places the results in a new table in which each query sequence with at least one hit has three sets of columns for its (up to) three best hits arranged from most significant to least significant (except in instances where two or more bit scores are equal). Generation of this combined results table allows each query sequence to be represented by a single record. Also, the classification calculations are performed only once and stored permanently in the results table precluding the need to run complex SQL SELECT queries over large data tables every time the results are fetched.

Each query sequence with at least one significant hit is classified into one of 11 different categories (see Fig. 2) using the decision tree algorithm mentioned above. The heuristics of this algorithm are presented below:

1. The TIGR Gene Index contains sequences that have been shown to code for protein (and, thus, are likely to actually represent genes), whereas there is no such

prerequisite for a sequence to be included in the EMC Local Database. Consequently, Gene Index is favored over EMC.
2. Because Gene Index and EMC local databases are likely to contain some repeat sequences, significant hits to organellar or repeat local databases are given priority over Gene Index and EMC hits.
3. If the first hit's bit score is at least 20% greater than the next two hits (if any) and the preceding heuristics are not violated, then the query sequence is classified based on the first hit's local database.
4. If the first and second hits or first and third hits are to the same local database, then the query sequence's classification is set to this local database.
5. If a query sequence is not classified in step 1, 2, 3, or 4, it is given the temporary classification of *Flag*. In the case of a *Flag* classification where the two best hits are to different repeat local databases (Ambiguous Repeat EMC, Ambiguous Repeat Gene Index, Probable Repeat EMC, Probable Repeat Gene Index, Annotated Repeats, Transposon, or Probable Repeats), the query sequence is classified by the local database to which it produces the highest bit score. The Probable Repeat local database is used only when analyzing reduced-representation sequences (see below).
6. If the classification is still *Flag* and the two best hits are to EMC and/or Gene Index, EMC is chosen if it has a higher bit score. Otherwise, Gene Index is chosen.
7. If the classification is still *Flag*, at least one of the hits is to Chloroplast, and none are to rDNA, then the classification is set to Chloroplast.
8. If the classification is still *Flag*, at least one of the hits is to rDNA, and none are to Chloroplast, then the classification is set to rDNA.
9. If the classification is still *Flag* and at least one of the hits is to Centromere with a bit score within 20% of the first hit's bit score, the query sequence is given the classification of Centromere.
10. If the classification is still *Flag* and all hits are to EMC, Gene Index, Ambiguous Repeat EMC, Ambiguous Repeat Gene Index, Probable Repeat EMC, or Probable Repeat Gene Index, the classification is set to the repetitive database with the highest bit score.
11. For simplicity, those query sequences classified as Ambiguous Repeat EMC or Ambiguous Repeat Gene Index are placed in the "Possible Repeat" category, whereas those query sequences classified as Probable Repeat EMC and Probable Repeat Gene Index are placed in the "Probable Repeat" category (see Fig. 2).

If *Flag* query sequences remain, they can be manually classified via the SRCP's Web interface or the decision tree algorithm can be modified. Although the decision tree algorithm described above resulted in automated classification of all *Flag* query sequences, other data and/or local database sets may produce unresolved flags indicating that fine tuning of the algorithm may be appropriate.

### *Identifying repeats in the "No Hit" query sequences*

The "No Hit" query sequence group can be further analyzed to identify novel repetitive elements based on their relative iteration in the query sequence set. If the genome coverage is at least $1.58X$, the "No Hit" query sequence group is analyzed using *ReAS* [7], an ab initio repeat-finding program that has proven especially robust in side-by-side comparisons with other database-independent repeat identification tools (our personal observations). However, the genome coverage in sample sequence-based genome characterization projects is often below the genome coverage levels necessary for most repeat analysis programs. Consequently, we developed a method to calculate which "No Hit" query sequences are probable repeats when genome coverage is below $1.58X$. First, we determine the $k$-mer length (sequence of length $k$) that will afford one chance in a thousand that two random query sequences will share an identical sequence of length $k$ for a genome of size $G$. This determination, based on Batzoglou [17], is made using the following logic:

1. There are four nucleotides in DNA; thus, the total number of potential $k$-mers is $4^k$.
2. Because of the double-stranded nature of DNA, a $k$-mer and its exact complement will be considered identical by *blastn*. This means that the number of "unique $k$-mers" is $4^k/2$.
3. Hence, the probability of a given "unique $k$-mer" occurring once in a genome of size $G$ is $2G/4^k$.
4. The probability of a specific "unique $k$-mer" occurring twice is $4G^2/4^{2k}$. The probability of any "unique $k$-mer" occurring twice is $2G^2/4^k$ [i.e. $(4G^2/4^{2k}) * 4^k/2$].
5. A 0.001 probability that two reads will share an identical sequence of length $k$ by chance is equivalent to $1000 * 2G^2/4^k$. Hence, the length of this unique $k$-mer is $k = \text{ceiling}(\log_4 G^2 + \log_4 2000)$.

The "No Hit" query sequences are BLASTed (*blastn*) against each other with the word size parameter set equal to the $k$ calculated as described above. Those query sequences that share a $k$-mer with one or more other "No Hit" query sequences are detected. We then use the Poisson distribution to determine a threshold contig depth $d$ [7] that is expected at error rate 0.1% for the level of genome coverage $\lambda$ as per the equation

$$p = (e^{-\lambda}\lambda^d) \div d!$$

Those query sequences that share a unique $k$-mer to $\geqslant d$ other "No Hit" query sequences (see Supplementary Table 3) are assigned to the "Probable Repeat" sequence category (Fig. 1). When genome coverage is $\leqslant 0.04X$ (and $d + 1 = 2$), the BLAST output file is parsed by a Perl script

that classifies query sequences as "Probable Repeats" if they have at least one hit to another query sequence, that is, share a unique *k*-mer. For data sets with coverage values between $0.05X$ and $1.57X$, we use another Perl script that classifies a query sequence as "Probable Repeat" only if it has at least the minimal number of hits sharing the same *k*-mer. "Probable Repeat" query sequences are then placed into a consolidated BLASTable local database of the same name. The Probable Repeats local database is used when analyzing sequences that have been generated through reduced-representation sequencing (see below).

### *Classification of remaining "No Hit" query sequences*

As shown in Fig. 2K, all remaining "No Hit" sequences are translated by the Perl script *three_frames.pl*, available from the Canadian Bioinformatics Help Desk, and compared with sequences in the EMC and Gene Index Local Databases using *tblastn* [18]. Such comparison can allow detection of potential gene orthologs that have undergone substantial divergence at the DNA level but have relatively conserved amino acid sequences. Those query sequences producing a significant *tblastn* hit ($S' \geqslant 60$) to an EMC or Gene Index entry are reclassified as described in Fig. 2. "No Hit" query sequences that do not produce a significant *tblastn* hit to EMC and/or Gene Index local databases are placed in the sequence category "Genome Sequence of Unknown Character." This part of the analysis is the most computationally expensive and may be performed using BLAT [19] and/or a computer cluster.

### *Output*

Once classification has been completed, summary statistics are calculated. They can be viewed or saved in an Excel file via a Web interface.

### *Contig assembly*

After classification, all query sequences are collectively analyzed using Phrap (www.phrap.org). An ACE file generated by Phrap is then parsed by Perl scripts that generate two summary XML files::one of the summary XML files contains data grouped by sequences and the other has data grouped by contigs. Both of the XML files include padded sequence data. These data are then bulk uploaded to the SQL Server database. A graphical interface has been designed to permit rapid visualization of contigs and the classification assigned to each query sequence within a contig. Desired outcomes of contig analysis include assembly of genes, characterization of repeat families, correction of potential erroneous classifications, and/or detection of improperly labeled/annotated GenBank/TIGR entries. With respect to error correction, visual inspection of assembly reads aided by color-coded classifications (Supplementary Fig. 1) allows rapid detection of query sequences that appear conspicuously out of place. If deemed appropriate, classifications can be changed and the source of the original classification traced back to the top three hits. It is anticipated that contigs visualized in this manner can potentially limit the *snowballing effect* of incorrect annotations and improve the quality of the local databases.

### *Analysis of reduced representation sequences*

Analysis of reduced-representation query sequences closely follows the scheme used for genomic query sequences (Fig. 2). However, the Probable Repeats local database (see above) generated after analysis of random genomic sequences is used as a ninth local database during the initial classification. Additionally, when analyzing "No Hit" query sequences, the genome size *G* is replaced by the fraction of the genome in a particular reduced-representation component. For example, according to Peterson et al. [10], the sorghum genome consists of highly repetitive, moderately repetitive, and single-/low-copy components that account for roughly 0.15, 0.41, and 0.24 of the genome, respectively. As the sorghum genome is about 760 Mb [20], the highly repetitive component of sorghum would contain 114 Mb of DNA (i.e., 0.15 * 760 Mb) while moderately repetitive and single/low-copy components would account for 311.6 and 182.4 Mb, respectively. To allow for consistent analysis of all reduced representation-enriched fractions, repetitive query sequences identified in reduced-representation data sets during the "No Hit" repeat analysis are not added to the Probable Repeats local database.

## Results and discussion

### *SRCP analysis of random genomic sorghum query sequences*

Initially, two representations of the same *S. bicolor* sequence set were analyzed by the SRCP. The first representation consisted of "full-length" genomic shotgun sequence reads of a size typical of trimmed reads produced via automated Sanger sequencing (mean length = 599 bp). The second representation consisted of the original full-length reads digitally fragmented into pieces between 80 and 179 bp in length (mean length = 105 bp) to simulate short read lengths such as those produced by 454 DNA sequencing [21]. The results of these analyses are summarized in Fig. 3A. As shown, shorter query sequence lengths resulted in an increase in the broadly defined Probable Repeats and Genome Sequence of Unknown Character categories with concomitant decreases in all other classes. This suggests that shortening query sequence length to about 100 bp often disrupts features that permit placement of query sequences into more narrowly defined categories, most notably EMC, Gene Index, and Transposon.

### *Comparison of Cot analysis and SRCP data*

Cot analysis is the study of the kinetics of DNA reassociation in solution. It can be used to learn much about

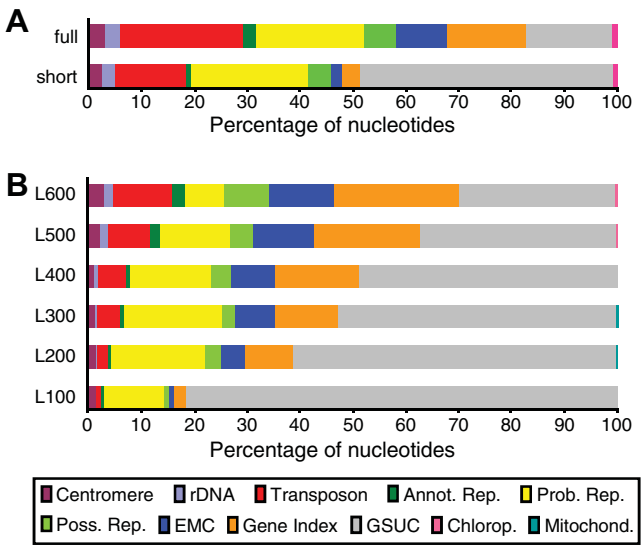


Fig. 3. SRCP-based classification of random sorghum genomic shotgun query sequences. (A) Classification of full-length query sequences (mean = 599 bp) versus short-length query sequences (mean = 105 bp). (B) Effect of query sequence length on classification. Six different query sequence lengths ranging from 100 to 600 bp were tested (see Supplementary Table 2).

the general structure of a genome, including genome size, number and size of kinetic components, amount of repetitive DNA, amount of single-/low-copy DNA, and kinetic complexities of unique and repeat components [22]. To permit comparison with Cot analysis data, percentages of Transposon, Annotated Repeat, Probable Repeat, Centromere, rDNA, and Possible Repeat categories were grouped together and deemed percentage repetitive genomic DNA. Conversely, the EMC and Gene Index categories probably represent single-/low-copy DNA and were grouped as such. The contents of the Genome Sequence of Unknown Character category may represent either low-copy and/or a combination of repetitive and low-copy sequences depending on the depth to which repeat components have been sequenced. With a sufficiently large sequencing depth or with fairly comprehensive repeat local databases, the rigorous repeat search conducted by the SRCP may afford a relatively high probability that sequences that end up in the Genome Sequence of Unknown Character category are also low-copy DNA. For this initial analysis, we conservatively assumed that 50% of the Genome Sequence of Unknown Character bases were low-copy DNA. Half the percentage of the Genome Sequence of Unknown Character category was added to the EMC and Gene Index percentages to yield a rough estimate of genomic single-/low-copy sequences. Based on SRCP analysis of full-length query sequences, the percentage of repetitive DNA in the *S. bicolor* genome is 58.2%, whereas short-length query sequence analysis provides a repeat value of 45.9%. A previous Cot analysis of sorghum [10] suggested that the genome is composed of at least 56% repetitive DNA, a value that falls within the range predicted by full- and short-length SRCP analyses. The percentage of single-/low-copy DNA as detected by SRCP analysis of full-length sorghum query sequences is 32.7%, whereas that of short-length query sequences is 29.4%. The Cot analysis suggested that single-/low-copy DNA makes up at least 24% of the sorghum genome. Considering the various biases inherent in Cot analysis and SRCP classification techniques, the similarity in repeat and low-copy sequence percentages between the two types of results is encouraging.

### *The effect of query sequence length on classification*

The SRCP uses an "all or nothing" approach, assigning every base in a query sequence to a "best-fit" sequence category. Although this is not a perfect classification solution, dissection and annotation of the parts of each query sequence would be a tremendous undertaking. As suggested in Fig. 3A, short query sequence lengths decrease the specificity of classification. Generation of single-read query sequence lengths beyond 600–700 bp is not currently feasible due to limitations of high-throughput capillary electrophoresis, but it is likely that increasing query sequence length much beyond this size would augment the chances that a repeat and a unique sequence occur on the same query sequence.

To further explore the effect of query sequence length on classification, we prepared sequence subsets with different query sequence lengths (Supplementary Table 2) and analyzed the subsets using the SRCP. The results of this analysis are summarized in Fig. 3B. In support of the observations made in analysis of the full-length and short-length query sequences, shorter query sequence lengths limit placement of sequences into gene and repeat classes. The L600 (600-bp sequence length) data set produces the highest levels of bases in the Gene (EMC and Gene Index) and Repeat (Transposon, Annotated Repeat, Probable Repeat, Possible Repeat, Centromere, and rDNA) categories. Compared with the results of the L600 analysis, the L500 set shows similar percentages of bases classified as EMC and Gene Index, but noticeable differences in how sequences are divided among repeat classes. Interestingly, the L600 set (Fig. 3B) shows fewer bases in repeat and low-copy classes compared with the full-length query sequences, which have a mean length of 599 bp (Fig. 3A). The full-length query sequence analysis involved roughly six times as much sequence data as the L600 analysis, and indeed, this may account for the observed differences. Although it is not clear what size query sequence will produce the most accurate description of a genome (and it is likely that optimal query sequence size may differ from genome to genome), our results suggest that 500- to 600-bp fragments provide an adequate compromise between length and classification specificity, while shorter sequences result in disruption of features that permit classification.

*Analysis of Cot-filtered DNA*

Reduced-representation sequencing techniques are methods that can be used to preferentially isolate and sequence a desired subset of DNA sequences from a larger population of sequences [8,9]. For example, some reduced-representation sequencing techniques are used to isolate and sequence gene-rich regions found within genomic DNA. Others may enrich for repeats or molecular markers. Examples of reduced-representation sequencing techniques include EST sequencing, methylation filtration [14], and Cot filtration [10].

If one is interested in evaluating reduced-representation sequencing-based enrichment using the SRCP, it is best if the SRCP is first used to analyze random genomic DNA from the same organism. This allows establishment of a "background" genome composition and results in generation of a Probable Repeat local database, which can be used to help identify repeats in the reduced-representation sequencing data.

To test the quality of SRCP classification versus manual classification, we first ran sorghum genomic query sequences through the pipeline to generate a Probable Repeat local database for sorghum. Then we used the SRCP to evaluate a set of Cot-filtered highly repetitive, moderately repetitive, and single-/low-copy sequences manually classified and described by Peterson and colleagues [10]. Peterson and colleagues made no attempt was to identify repeats and/or genes in the categories comparable to our "No Hit" group, preventing direct comparisons of repeat and low-copy contents. Consequently, we analyzed the "No Hit" sequences of Peterson et al. [10] with the algorithms depicted in steps J–L in Fig. 2 and made the assumption that 50% of bases given a final classification of Genome Sequence of Unknown Character were low-copy DNA. As with the random genomic DNA, the Cot-filtered sequences were analyzed as "full-length" query sequences (mean ± SE length = 177.5 ± 2.8 bp) and "short-length" query sequences (80–179 bp). The results of the full-length SRCP, short-length SRCP, and manual classification are summarized in Fig. 4. Of note, there is very little difference in the percentages of single-/low-copy and repetitive sequences detected using the three schemes.

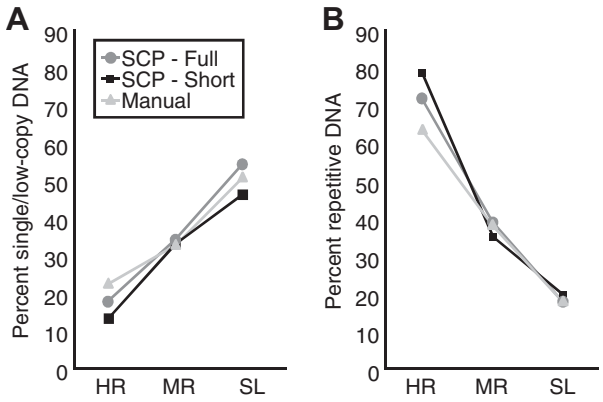


Fig. 4. Low-copy and repeat sequence contents of highly repetitive (HR), moderately repetitive (MR), and single/low-copy sorghum DNA libraries as determined by the SRCP and by manual classification. (A) The increase in low-copy DNA from HR to SL libraries as seen with the manually classified sequences is paralleled by the SRCP classification. (B) The decrease in repetitive DNA from HR to SL libraries as seen with the manually classified sequences is paralleled by the SRCP classification.

## Conclusions

The SRCP is an automated means through which genomes can be characterized based on sample shotgun sequencing. To our knowledge, it is the first pipeline designed for this purpose. Moreover, as demonstrated above, it can be used to determine the efficiency of reduced-representation sequencing in a manner that is as accurate as, and certainly much faster than, manual classification. Of note, careful adaptation of the SRCP may advance comparative genomics by affording a rapid means of evaluating divergence that has occurred in ostensibly related species. Although we developed our implementation for the study of higher plant genomes, the SRCP can be easily adapted for study of any group of organisms; the principal adjustment required for use of the SRCP for other subjects is modification of the boolean text strings used in building the local databases (Supplementary Table 1). Alternatively, one can use existing sequence databases, including those developed for model organisms. The implementation of the SRCP described in this article is based on the scale and demands of our current workloads. However, the design is such that it can readily be adapted for larger-scale projects. In such cases, sequence alignment might best be performed on a cluster running a parallelized version of BLAST (at least for alignments performed against the Gene Index and EMC local databases). Techniques such as Extensible Stylesheet Language Transformations (XSLT) may further speed up processing of large XML output files. Once the pipeline is established and performs all steps correctly, it can be further automated via script scheduling and bottleneck elimination in program flow. Additionally, the SRCP is designed to be easily coupled with other scripts that allow further utilization of the sequence data. Indeed, we have begun building a pipeline that will generate consensus sequences for transposons and classify these elements into families based on their sequence structures.

## Acknowledgments

This research was supported by Grant DBI-0421717 from the National Science Foundation, Grant 2006-34506-17290 from the U.S. Department of Agriculture Cooperative State Research, Education, and Extension Service, and a grant from the Mississippi Corn Promotion Board.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2007.08.008.

## References


[1] W.B. Strong, R.G. Nelson, Preliminary profile of the *Cryptosporidium parvum* genome: an expressed sequence tag and genome survey sequence analysis, Mol. Biochem. Parasitol. 107 (2000) 1–32.

[2] E.F. Kirkness, V. Bafna, A.L. Halpern, S. Levy, K. Remington, D.B. Rusch, A.L. Delcher, M. Pop, W. Wang, C.M. Fraser, J.C. Venter, The dog genome: survey sequencing and comparative analysis, Science 301 (2003) 1898–1903.

[3] A. Lomsadze, V. Ter-Hovhannisyan, Y.O. Chernoff, M. Borodovsky, Gene identification in novel eukaryotic genomes by self-training algorithm, Nucleic Acids Res. 33 (2005) 6494–6506.

[4] V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters, Genome Biol. 7 (Suppl. 1) (2006) S10–S12.

[5] Z. Bao, S.R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes, Genome Res. 12 (2002) 1269–1276.

[6] A.L. Price, N.C. Jones, P.A. Pevzner, De novo identification of repeat families in large genomes, Bioinformatics 21 (Suppl. 1) (2005) i351–i358.

[7] R. Li, J. Ye, S. Li, J. Wang, Y. Han, C. Ye, J. Wang, H. Yang, J. Yu, G.K. Wong, J. Wang, ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun, PLoS Comput. Biol. 1 (2005) e43.

[8] D.G. Peterson, Reduced representation strategies and their application to plant genomes, in: K. Meksem, G. Kahl (Eds.), The Handbook of Genome Mapping: Genetic and Physical Mapping, Wiley-VCH Verlag, Weinheim, 2005, pp. 307–335.

[9] A.H. Paterson, Leafing through the genomes of our major crop plants: strategies for capturing unique information, Nat. Rev. Genet. 7 (2006) 174–184.

[10] D.G. Peterson, S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, A.H. Paterson, Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery, Genome Res. 12 (2002) 795–807.

[11] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444–2448.

[12] J.A. Bedell, M.A. Budiman, A. Nunberg, R.W. Citek, D. Robbins, J. Jones, E. Flick, T. Rholfing, J. Fries, K. Bradford, J. McMenamy, M. Smith, H. Holeman, B.A. Roe, G. Wiley, I.F. Korf, P.D. Rabinowicz, N. Lakey, W.R. McCombie, J.A. Jeddeloh, R.A. Martienssen, Sorghum genome sequencing by methylation filtration, PLoS Biol. 3 (2005) e13.

[13] T.E. Bureau, P.C. Ronald, S.R. Wessler, A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes, Proc. Natl. Acad. Sci. USA 93 (1996) 8524–8529.

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

[15] The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant, *Arabidopsis thaliana*, Nature 408 (2000) 796–815.

[16] J.H. Zar, Biostatistical Analysis, Prentice-Hall, Upper Saddle River, NJ, 1996.

[17] S. Batzoglou, Computational Genomics: Mapping, Comparison, and Annotation of Genomes, Dissertation, Massachusetts Institute of Technology, 2000, p. 21.

[18] S.F. Altschul, M.S. Boguski, W. Gish, J.C. Wootton, Issues in searching molecular sequence databases, Nat. Genet. 6 (1994) 119–129.

[19] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.

[20] K. Arumuganathan, E.D. Earle, Nuclear DNA content of some important plant species, Plant Mol. Biol. Rep. 9 (1991) 208–218.

[21] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.

[22] R.J. Britten, D.E. Graham, B.R. Neufeld, Analysis of repeating DNA sequences by reassociation, Methods Enzymol. 29 (1974) 363–405.