BMC Bioinformatics

Methodology article

Oliz, a suite of Perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3' untranslated region

Hao Chen* and Burt M Sharp*

Address: Department of Pharmacology, University of Tennessee Health Science Center, Memphis, TN 38163 USA

E-mail: Hao Chen* - hchen@utmem.edu; Burt M Sharp* - bsharp@utmem.edu

*Corresponding authors

Published: 6 October 2002

BMC Bioinformatics 2002, **3**:27

Received: 7 April 2002

Accepted: 6 October 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/27

Keywords: oligonucleotide microarray, Perl, UniGene

Abstract

Background: Identifying reliable oligonucleotide sequences for use in microarray experiments is a complex process. Two key issues are the accuracy of the input sequences and the specificity of the oligonucleotide sequences.

Results: We provide a suite of Perl scripts that facilitates the search for gene-specific oligonucleotides for microarray experiments. Genes of interest are first identified in the form of UniGene clusters. The sequences of these clusters were extracted and assembled into contigs to increase their accuracy. The 3' untranslated region (3'UTR) of the contig was parsed. Then, multiple 50mer oligonucleotide sequences with similar melting temperature were obtained from each 3'UTR. These sequences were analyzed for gene specificity. Five Cy3-labeled cDNAs were used to empirically verify the specificity of a set of 1814 50mers.

Conclusion: Oliz can be used to select oligonucleotide sequences for microarrays. Oliz is freely available for academic users at [http://www.utmem.edu/pharmacology/otherlinks/oliz.html]

Background

DNA microarrays usually involve the hybridization of labeled cDNA samples to a set of complementary DNA (either PCR products or synthetic oligonucleotides) fixed onto solid media. Spotting presynthesized oligonucleotide has many advantages, such as high sensitivity, convenience, and cost effectiveness. Most importantly, the use of oligonucleotide probes circumvents the high error rate that is associated with the PCR amplification of bacterial clones [4,6].

The starting point in the design of oligonucleotide microarrays is the identification of short DNA sequences that can be used as probes for the genes of interest. Obviously, all sequences should be gene specific and have similar melting temperature (Tm). We have been interested in using the 3' untranslated region (3'UTR) as the target region for the design of oligonucleotide probes primarily because of the relatively high specificity of this region [8] and the availability of sequence information (in the form of Expressed Sequence Tags, ESTs).

Frist, our approach involves the identification of genes of interest in the form of UniGene clusters. The sequences of these clusters were retrieved and assembled into contigs. Then, the 3'UTRs were parsed from the contigs. Finally, oligonucleotide sequences of 50 nucleotides with similar

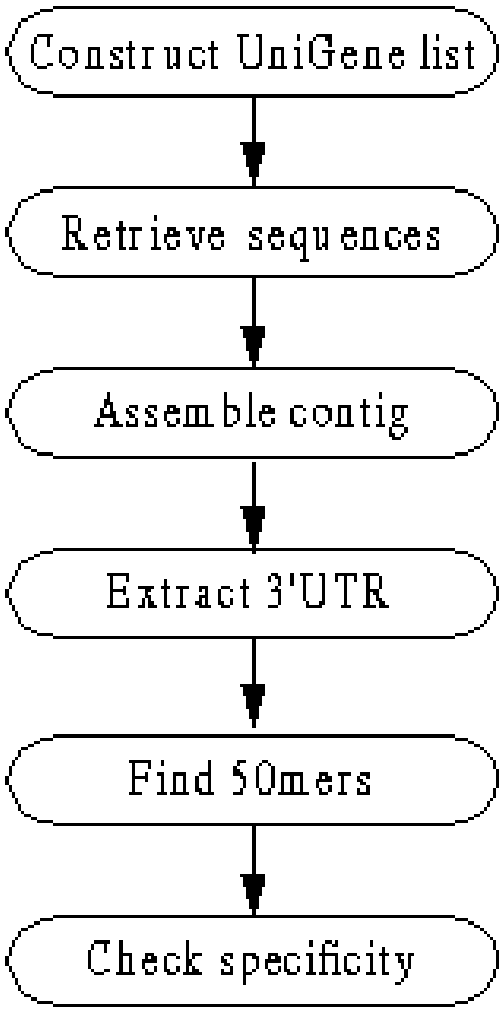

Figure 1
The work flow of the Oliz program

melting temperature (Tm) and GC content were selected and screened for specificity (Figure 1).

Results

The Oliz suite was written in Perl (v.5.6) and was tested on the RedHat Linux (v.7.1) operating system. Oliz has four modules. The UNI module extracts UniGene clusters, which are assembled into contig(s) by the CONTIG module. Then, the UTR module parses the 3'UTRs of the contigs, and selects multiple 50mer sequences that are within the selected range for GC content (45–50%) and Tm (76°C ± 5). Lastly, the UNIQ module performs blast searches on the 50mers to ensure their gene specificity.

Step 1. UniGene retrieval and contig assembly

A list of selected UniGenes is first compiled and used as the input file for the UNI module. The sequences contained in these UniGene clusters are extracted by the UNI module. To achieve this function, the UNI module requires a file that contains all the UniGene sequences of the species of interest. This file is available from NCBI's FTP site [ftp://ftp.ncbi.nih.gov/repository/UniGene]. The name of this file follows the convention of "species.seq.all". Then, the CONTIG module assembles each of the clusters into a contig using the CAP3 program[2]. Due to the high error rate in both the sequence and annotation of ESTs, clusters contains only one EST sequence are excluded from further analysis.

Step 2. Parsing 3'UTR

The UTR module performs several tasks. Initially, it determines the orientation of a contig by comparing it to a reference sequence, such as those provided by the NCBI RefSeq project [5] (1st priority), or GenBank sequences with coding region annotations (2nd priority), or sequences with polyA tails (3rd priority). It is generally assumed that these sequences are in 5'-3' orientation. When the above approaches fail to identify the orientation of the contig, its cluster identifier is sent to a separate file. The orientation of these contigs can be obtained manually by cross-referencing to their homologues in other species, and then be incorporated into the results.

The UTR module's main function is to parse the 3'UTR of the contigs, according to the coding region annotation in the reference sequence. The length of the 3'UTR varies from gene to gene. Based on the average length of the transcripts obtained from oligo dT primed cDNA synthesis, we decided to target the last 500 bases of the 3'UTR as the region for the selection of 50mer oligonucleotides. In addition, the UTR module generates several HTML files to facilitate visual inspection of the results. These files contain links to the UniGene cluster sequences, the contigs and the 3'UTRs.

Step 3. Generating 50mer oligonucleotides with close Tms

The EMBOSS prima program was used to select 50mer oligonucleotides with similar Tms for each 3'UTR. The Tm was set at 76 ± 5°C based on the average Tm for 50mers. The resulting 50mer sequences were saved as a comma-separated text file, ready for processing by the UNIQ module.

Step 4. Similarity search

One of the advantages of using the 3'UTR as the target region for hybridization is that this region has been under less evolutionary pressure to remain constant. However, this does not guarantee that all 50mers selected from this region are gene specific. Therefore it is necessary to identi-

fy potentially similar sequences in other genes. The UNIQ module automates the blastn search, analyzes the blastn results, and decides whether to retain or discard a particular 50mer based on the set criteria.

The UNIQ module runs blastn searches using a local database constructed using sequences obtained from NCBI. While analyzing the sequences identified by blastn, it disregards accession numbers that are found in the same UniGene cluster as the 50mer. Matches that are oriented complementary to the 50mer also are ignored, and only sense/sense pairs are analyzed further. The orientation of the blastn matches is apparent when they are known genes. EST hits are judged based on their "clone_end" annotation.

Kane et. al. [3] reported that specificity of a 50mer oligonucleotide requires that it is less than 75% similar to all non-target transcripts. In addition, when it is 50–75% similar to a non-target transcript, the similar region must not include a stretch of sequence of greater than 15 contiguous bases. Since blast only returns part of the sequence where a match is found (usually less than 50 nucleotides), it is necessary for the UNIQ module to retrieve the entire matching sequence before calculating the overall sequence similarity. The guidelines reported by Kane et al. are then followed to determine whether candidate 50mers are acceptable.

Occasionally all the candidate sequences generated by EMBOSS prima were disqualified when compared to one EST entry. This is a difficult issue, insofaras these ESTs may represent unknown genes, implying that the candidate 50mer is not gene specific. However, the apparent similarity may simply be caused by errors in the EST. When this occurs, the UNIQ module performs another blastn search that excludes all the ESTs from the database. The accession number of the EST in question is provided in the output file and a detailed log file for each oligonucleotide sequence is also provided.

Experimental verification of the specificity of the 50mer oligonucleotides

A set of 1816 rat specific 50mer oligonucleotide sequences was obtained using the methods described above. Most of these genes are known to be expressed in the central nervous system. These oligonucleotides were spotted in duplicate onto TeleChem SuperAmine slides.

A total of 10 oligonucleotide sequences were chosen at random to test the specificity of the 50mers designed by Oliz. PCR primers were designed for each of these 10 sequences in order to amplify a fragment of approximately 100–450 base pairs that include the 50mer. Reverse transcription PCR (RT-PCR) was performed using mixed rat brain mRNA. Five of these ten primer pairs amplified a single product with the expected length. A second PCR reaction was performed on these 5 RT-PCR products to selectively amplify the antisense strand while incorporating amino allyl dUTP. These antisense DNAs then were labeled with Cy3 fluorescent dye, and were used for microarray hybridization. Each microarray slide was only probed with one Cy3-labeled DNA.

All of the five Cy3-labeled cDNAs hybridized to their expected spots. The subgrids (13×8 spots) that contain the specific hybridized spots are shown in Figure 2. Two spots on the array, known to have green autofluorescence (not shown), were excluded from the analysis. Depending on the specific cDNA sequence, there were 0–4 additional spots that had detectable fluorescence. This represents $0–4/1814 \times 100\% = 0–0.22\%$ of all spots. No non-specific spots was associated with more than one cDNA. The intensity of these spots are shown in Table 1. Since the duplicates within each microarray slide had similar intensity values, the average intensity is given. The level of non-specific binding ranged from approximately 5–15% of specific binding; only 25% of the non-specific spots had fluorescence binding values that were > 10% of specific binding. The magnitude of non-specific binding was not related to the level of specific binding.

Discussion

Our approach to the selection of 50mers for microarray experiments is unique in two aspects: 3'UTR sequences are derived from contigs assembled from UniGene clusters, and the specificity of the final 50mer oligonucleotide is verified computationally. We also experimentally verified the specificity of five Cy3-labeled cDNAs against a microarray containing 1816 50mers designed by Oliz.

We focused on designing 50mers from the 3'UTR, in part because it has a relatively high degree of gene specificity [8] due to low evolutionary pressure. More importantly, in many species including the rat, the 3'UTR is the only region where sequence is available for many genes. Many of these 3'UTRs exist in the form of ESTs, which are known to have high error rates in both their sequences and orientation annotations. Thus, instead of identifying genes of interest in the form of a single sequence entry, we choose target genes from UniGene clusters. Assembling the individual sequences in each cluster into a contig appears to provide the most accurate sequence information available. We also implemented algorithms to verify the orientation of the contig.

The assurance of gene specificity is a critical issue in the design of oligonucleotide microarrays. Kane et al. [3] reported that 50mer oligonucleotides can detect non-target transcripts that are marginally similar to target transcripts.

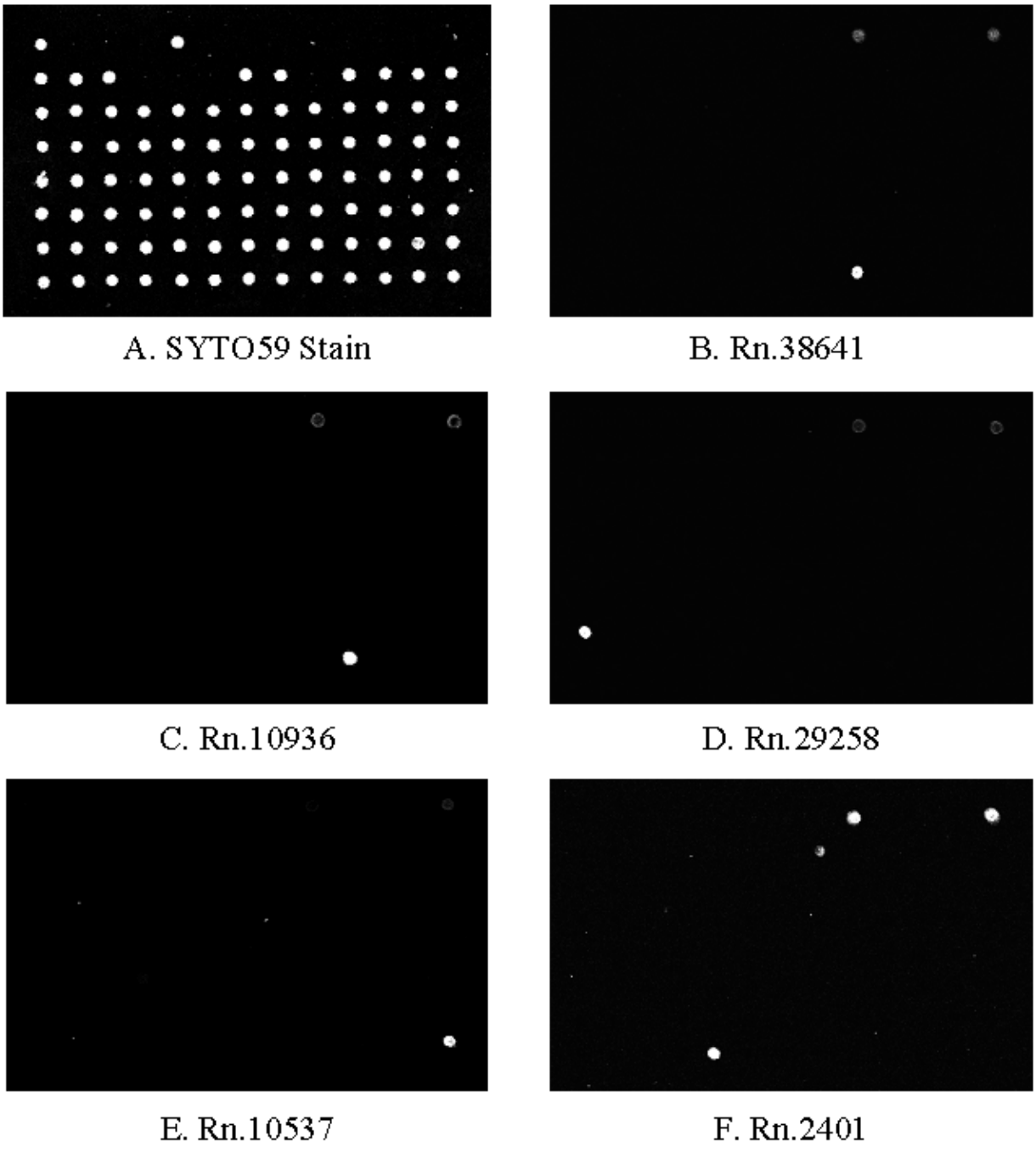

Figure 2
Verification of specificity. Microarrays were printed with 1816 oligonucleotides in duplicate. Each sub-grid has 13 columns and 8 rows. (A) SYTO59 DNA staining demonstrating the layout of one sub-grid. The spot in the upper right corner and the spot four columns to the left were printed with Cy3-labeled oligonucleotides for orientation purposes. Some positions near the top of each sub-grid were printed with buffer only, and are thus not visible on SYTO59 staining. (B)-(F): cDNAs contain the complementary strand of one of the 50mers on the array were amplified by PCR and labeled with Cy3. Hybridization of these cDNAs to the microarray shows that only the complementary spot had fluorescent signal in most cases. cDNA for Rn.2401 also hybridized to another spot in the same sub-grid, but with lower intensity.

Table 1: Intensity values of specific and non-specific binding

Spot Identification		Intensity	Percentage*
Specific	**Non-specific**		
Rn.10573		1945618	
	Rn.7181	127311	6.7
	Rn.2759	110929	5.84
	Rn.22143	95109	5.01
Rn.38641		999528	
	Rn.45596	112537	11.26
	Rn.17591	92120	9.22
	Rn.4089	83486	8.35
	Rn.44292	66381	6.64
Rn.2401		647796	
	Rn.22775	65428	10.1
	Rn.25764	48190	7.44
	Rn.63672	35700	5.51
	Rn.9056	33209	5.13
Rn.10936		1827034	
	Rn.3252	275348	15.07
Rn.29258		2145519	
	None		

Percentage = non-specific intensity / specific intensity × 100

The UNIQ module avoids these 50mers by performing similarity searches using the blastn program. Blastn lists genes with significant sequence similarity to the query sequence. However, only the portion of the sequence that matches closely to the query is provided in the blast results. To allow the calculation of overall similarity, the complete sequence for each match was retrieved and then truncated to a 50mer corresponding to the query sequence. Those 50mers that reached the similarity threshold (as described previously) were discarded.

One caveat to our approach is that the accuracy of the specificity calculation is influenced by the completeness of the database used for the homologue search. The ideal database should contain all the expressed sequences (transcriptome) of the species under investigation. Thus, for species like rat, where the genome is to large extent unknown, verifying microarray results with an independent methodology (e.g. real-time PCR) is essential.

The total number of cDNA species that one oligonucleotide binds to is an index of it's specificity. However, this is difficult to measure directly, if at all. Thus, to experimentally assess the specificity of oligonucleotides selected by Oliz, we amplified 5 cDNAs using PCR, labeled each with Cy3, and hybridized each to separte microarrays containing a duplicate set of 1816 oligonucleotides. The oligonucleotides on the microarray were synthesized according to the sequences selected by Oliz, and each of the PCR products contain a segment that is complementary to one of the 50mers on the microarray. The results showed that all of the 5 cDNAs hybridized to their corresponding 50mer oligonucleotides (Figure 2). Depending on the cDNA, 0–4 additional spots out of 1814 spots (excluding 2 known autofluorescent spots) also had a detectable hybridization signal. These are considered to be non-specific binding. The average intensity of these non-specific signal was only 8% of the specific binding. These data suggested that the oligonucleotide sequences selected by Oliz have the specificity that is needed for microarray experiments. However, the low level signals may mimic those emitted by some low abundance genes. Thus further emphasis the importance of verifying microarray results by independent means such as PCR.

Conclusions

We implemented a computational solution to select 50mer oligonucleotide sequences for microarray experiments. This solution is based on several well-recognized public domain bioinformatics tools. We also provided experimental evidence that the 50mers selected by our program have the necessary specificity required for microarray experiments. Our implementation is available

Table 2: Primer sequences:

Rn.10936 Growth hormone – releasing receptor	
5' primer	TGCACTCAAACTGCCATGCC
3' primer	TCCCCTTCAGAGCCACATTGAC
Rn.29258 ATP synthase, H+ transporting	
5' primer	CAACTCTTCTCCTATGCGATTC
3' primer	TTTGTCTTCTCTCTGTCAAACC
Rn.2401 cysteine rich protein	
5' primer	AAGGGAGAGCCTAATGAATACC
3' primer	TGATAAATCAGCTTCAACAGCC
Rn.38641 Recoverin	
5' primer	GTGAAGGGAAGGATAAAGGAAAAG
3' primer	TGGGATTATCACAGGGCTAC
Rn.10573 nitrix oxide synthase 1, neuronal	
5' primer	TTGACTAAATTCGGACACACACG
3' primer	AGGAAGAAGGACCAGGACAC

at [http://www.utmem.edu/pharmacology/otherlinks/oliz.html] for non-commercial use.

Materials and Methods

Oliz program

The Oliz suite was written in Perl (v.5.6) and was tested on a computer (Pentium IV 1.4 GHz processor, 785 Mb RAM) operating on RedHat Linux (v.7.1).

Several publicly available bioinformatics tools were used, including CAP3 [2] for contig assembly, blastn [1] program for similarity searching, Clustalw [7] for pair-wise alignment, EMBOSS prima for sequence selecting, and Bioperl [http://www.bioperl.org] for sequence manipulation.

Oliz source code is available free of charge for academic users from [http://www.utmem.edu/pharmacology/otherlinks/oliz.html] Detailed instruction on installation and execution is distributed with the source code.

Microarray manufacturing

A total of 1816 50mer unmodified oligonucleotides were synthesized by Integrated DNA Technologies (Skokie, IL) based on sequences selected by the Oliz program. These oligonucleotides were then diluted to 30 uM in 1× spotting buffer (TeleChem, Sunnyvale, CA) and were spotted in duplicate onto SuperAmine slides (TeleChem) using MicroGrid II (BioRobotics, Woburn, MA) arrayer.

Preparation of Cy3-labeled cDNA

Messenger RNA was extracted from mixed rat brain tissue using Trizol reagent (Invitrogen, Carlsbad, CA) following manufacturer's instructions. cDNA was synthesized using MuLV reverse transcriptase (ABI, Foster City, CA) from 2 ug of brain total RNA. PCRs were performed using AmpliTaq DNA Polymerase (ABI) and 1 ul cDNA. The sequences of the 5 PCR primers pairs that produced a unique band at the expected molecular weights are listed in Table 2. These PCR products were re-amplified using 1 ul of the first round PCR product with a dNTP mixture (3 mM dTTP, 3 mM 5-3 aminoallyl-2'-dUTP, 10 mM each dATP, dCTP, dGTP). The concentration of the forward primers were reduced to 1/100th (0.008 uM) of the reverse primer to yield a product that is mainly antisense strand. These amino allyl labeled cDNAs were purified using a QiaQuick column (QiaGen, Valencia, CA), and then vacuum dried. The samples then were reconstituted in 4.5 ul labeling buffer (Ambion, Austin, TX), and mixed with 2.5 ul water and 3 ul Cy3 dye (Amersham, Piscataway, NJ). This mixture was incubated in the dark at room temperature for 30 min, and was again purified by using QiaQuick column. The final product was eluted in 80 ul water (pH7.5), and DNA concentration was measured using a spectrophotometer. These Cy3-labeled cDNAs were diluted to a final concentration of 10 pM in a hybridization buffer containing 5 × SSC, 5 × Denhardt's solution, and 0.01% SDS.

Microarray hybridization and analysis

Microarray slides were washed in 0.2% SDS and then twice in water. Each wash was 2 min. Slides were then dipped in 100% ethanol for 30 sec and were spun dry (1000 rpm for 2 min). A lifterslip (Erie Scientific, Portsmouth, New Hampshire) was first laid on top of the microarray and then 80 ul of the Cy3-labeled cDNA was added to the space between the lifterslip and the microarray. This assembly was incubated at 45 C overnight. After hybridization, the slides were washed in 1 × SSC with 0.2% SDS for 4 min, 0.1 × SSC with 0.2% SDS for 4 min, and 0.1 × SSC twice for 2 min. The arrays were spun dry and were scanned using a confocal microarray scanner

(BioRad, Hercules, CA). Imagepro software (Mediacybernetics, Silver Spring, MD) was used for spot intensity analysis.

Authors'contributions

HC conceived of the project (together with BMS), coded the Oliz program, carried out the PCR and microarray hybridization experiments, analyzed the data, and drafted the manuscript. BMS conceived of the project (together with HC), participated in its design and coordination, and edited and revised the manuscript.

Acknowledgments

The authors would like to thank Dr. William Taylor for his valuable advice in the development of these scripts. The authors also would like to thank Dr. Taylor and Mr. Bill Orr for the manufacture of microarrays and for suggestions regarding hybridization conditions. This work was supported by NIH grant DA03977 (B.M.S.)

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410
2. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9:**868-877
3. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.** *Nucleic Acids Res* 2000, **28:**4552-4557
4. Knight J: **When the chips are down.** *Nature* 2001, **410:**860-861
5. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140
6. Taylor E, Cogdell D, Coombes K, Hu L, Ramdas L, Tabor A: **Sequence verification as quality-control step for production of cDNA microarrays.** *Biotechniques* 2001, **31:**62-65
7. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680
8. Yazaki J, Kishimoto N, Nakamura K, Fujii F, Shimbo K, Otsuka Y, Wu J, Yamamoto K, Sakata K, Sasaki T, *et al*: **Embarking on rice functional genomics via cDNA microarray: use of 3' UTR probes for specific gene expression analysis.** *DNA Res* 2000, **7:**367-370