Research

A strategy for oligonucleotide microarray probe reduction

Alena A Antipova*†, Pablo Tamayo* and Todd R Golub*‡

Addresses: *Center for Genome Research, Whitehead Institute/Massachusetts Institute of Technology, and †Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ‡Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.

Correspondence: Todd R Golub. E-mail: golub@genome.wi.mit.edu

Published: 25 November 2002

Genome ***Biology*** 2002, **3(12)**:research0073.1–0073.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0073

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 16 August 2002
Revised: 23 September 2002
Accepted: 11 October 2002

Abstract

Background: One of the factors limiting the number of genes that can be analyzed on high-density oligonucleotide arrays is that each transcript is probed by multiple oligonucleotide probes of distinct sequence in order to magnify the sensitivity and specificity of detection. Over the years, the number of probes per gene has decreased, but still no single array for the entire human genome has been reported. To reduce the number of probes required for each gene, a robust systematic approach to choosing the most representative probes is needed. Here, we introduce a generalizable empiric method for reducing the number of probes per gene while maximizing the fidelity to the original array design.

Results: The methodology has been tested on a dataset comprising 317 Affymetrix HuGeneFL GeneChips. The performance of the original and reduced probe sets was compared in four cancer-classification problems. The results of these comparisons show that reduction of the probe set by 95% does not dramatically affect performance, and thus illustrate the feasibility of substantially reducing probe numbers without significantly compromising sensitivity and specificity of detection.

Conclusions: The strategy described here is potentially useful for designing small, limited-probe genome-wide arrays for screening applications.

Background

DNA microarrays have become commonplace for the genome-wide measurement of mRNA expression levels. The first described microarray for this purpose, the cDNA microarray, involves the mechanical deposition of cDNA clones on glass slides [1]. Although this strategy has proved highly effective, it has two limitations: cross-hybridization can occur between mRNAs and non-unique or repetitive portions of the cDNA clone; and the maintenance and quality control of large, arrayed cDNA libraries can be challenging. For these reasons, oligonucleotide microarrays have at least theoretical advantages. Short probes (25 nucleotides or longer) can be selected on the basis of their sequence specificity, and either synthesized *in situ* (by photolithography or inkjet technology) on a solid surface or conventionally synthesized and then robotically deposited.

The first oligonucleotide microarrays contained hundreds of distinct probes per gene in order to maximize sensitivity and specificity of detection [2]. Over the past few years, the

number of probes per gene has decreased as increasing amounts of sequence information have become available, probe-selection algorithms have improved, feature sizes have decreased and researchers have wanted to maximize the number of genes assayable on a single microarray. Nevertheless, no single array representing the entire human genome has been described. Furthermore, to date, no systematic high-throughput method has been published that can be used for reducing the number of probes per gene while maximizing the sensitivity and specificity of these reduced probe sets.

Several strategies for probe reduction could be considered. Probes could be selected at random, but given that different probes can have dramatically different hybridization properties, this random method would be likely to result in failure, at least for some genes. Alternatively, one could assess the fidelity of candidate probes by comparison to a gold standard of gene-expression measurement such as real-time quantitative PCR or Northern blotting. Such approaches, however, are not feasible at a genome-wide scale. We report here a generalizable, empiric strategy for probe reduction that eliminates 95% of probes, yet maximizes fidelity to the original microarray design.

Results and discussion

The experiments described here are based on HuGeneFL GeneChips commercially available from Affymetrix. These arrays contain approximately 282,000 25-mer oligonucleotide probes corresponding to 6,817 human genes and expressed sequence tags (ESTs) (a total of 7,129 probe sets). On average, each gene is represented by 40 probes: 20 'perfect match' probes that are complementary to the mRNA sequence of interest, and 20 'mismatch' probes that differ only by a single nucleotide at the central (13th) base. We refer to the perfect match/mismatch pair as a 'probe pair'. Each gene is thus represented by 20 probe pairs. Normally, these 20 probe pairs are consolidated into a single expression level (known as 'Average Difference') for each gene using GeneChip software (Affymetrix) which calculates a trimmed mean of the perfect match minus mismatch differences in order to incorporate some measure of non-specific cross-hybridization to mismatch probes [2]. Alternative methods for estimating message abundance have also been reported [3,4].

To reduce the number of probes per gene, we sought to identify the single probe pair for each gene that best approximated the Average Difference, a value that is based on all 20 probe pairs. To accomplish this, we first defined a training set of expression data derived from 141 human tumor samples of diverse cellular origins [5]. For each gene on the array, we generated a vector corresponding to the normalized Average Difference value across the 141 samples. Next, we calculated the perfect match minus mismatch value for each of the 20 individual probe pairs for each gene on the array (referred to hereafter as delta (Δ). In the final step, the 20 normalized Δs for each gene were ranked according to their degree of correlation with the Average Difference vector across the 141 training samples using Euclidean distance as the metric. The highest-ranking Δ (Δ_h) was chosen for further evaluation in an independent test set. A schematic for this procedure is shown in Figure 1.

The independent test set consisted of expression data derived from 176 tumor samples that were entirely non-overlapping with the training set. We determined the ability of the training-set-derived Δ_h values to approximate the Average Difference values in the independent test set, compared to randomly selected Δs. As shown in Figure 2, 79.3% (±3.0%) of Δ_h values were within twofold of their respective Average Difference value, as compared to 57.8% (± 5.1%) for randomly selected Δs. The relative error of the estimates was 0.8 (± 0.1) for Δ_h values and 2.7 (± 0.7) for randomly selected Δs. Overall, the distribution of Δ_h accuracies was distinct from randomly selected Δs ($p < 10^{-4}$, chi-squared test). This result indicates that the empirical selection of Δ_h is a better strategy for reducing probe numbers compared to random probe selection.

We next determined whether training-set-derived Δ_h values would be sufficient for pattern recognition and classification

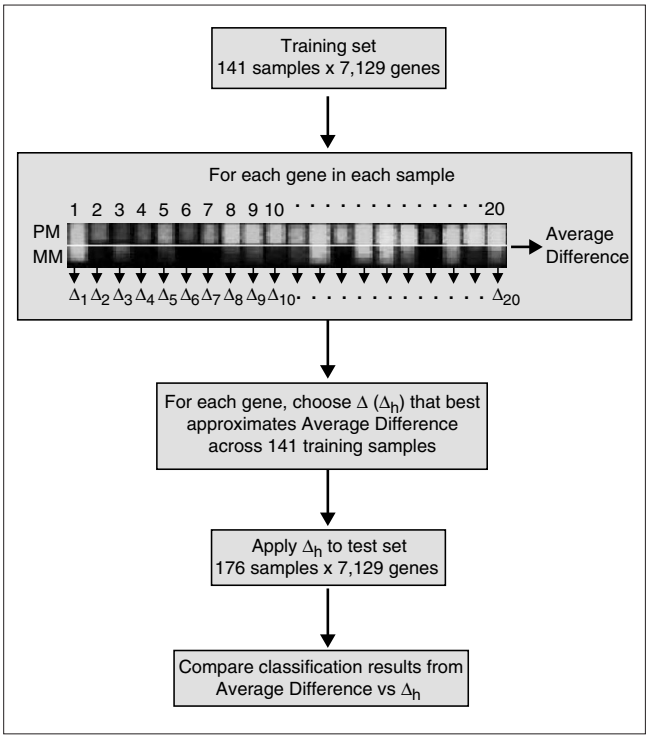

Figure 1
Schema for selection and evaluation of the single probe pair (Δ_h) that best approximates the Average Difference value derived from all 20 probe pairs. PM, perfect match; MM, mismatch. Δ = PM - MM.

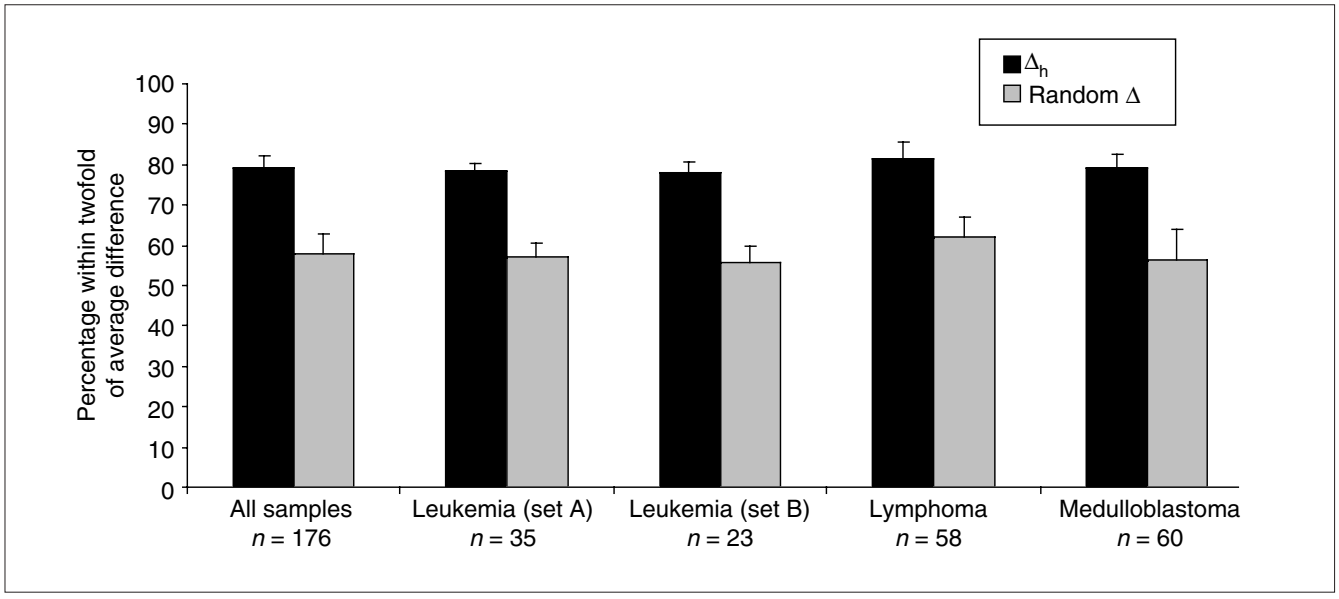

Figure 2
Comparison of Average Difference values with Δ_h and randomly selected Δs. For each of the datasets shown, the proportion of genes whose Δ_h value is within twofold of the Average Difference is shown by the black bars. The same comparison is shown for random Δs (gray bars). Error bars indicate standard deviation. Standard deviation shown reflects variations in the percentage of genes within twofold of the Average Difference between the 176 chips of the training set. Note that the Δ_hs better approximate the Average Difference compared to randomly selected Δs.

of the independent test set of samples. The 176 test samples fall into four binary classification problems: acute myeloid leukemia, AML, versus acute lymphoblastic leukemia, ALL (leukemia set A; $n = 35$); T-cell ALL versus B-cell ALL (leukemia set B; $n = 23$); diffuse large B-cell lymphoma survival prediction ($n = 58$); and medulloblastoma brain tumor survival prediction ($n = 60$), as described previously [6-8]. We used a *k*-nearest neighbors (*k*-NN) prediction algorithm [9] and applied it to these four classification problems using either the Average Difference values or the Δ_h values as the starting point. As shown in Table 1, classification accuracy based on Δ_h was nearly identical to that obtained using Average Difference values, despite the fact that 95% fewer probes were utilized. It should be noted that while Δ_h values more accurately approximated the Average Difference compared to random Δs (Figure 2), the random Δs also performed relatively well in these classification problems. It is possible, however, that classification performance would deteriorate when applied to more subtle classification problems, or when applied to samples of different tissue types. These results, taken together, demonstrate the feasibility of substantially reducing probe numbers without dramatically affecting performance.

Table 1

Classification accuracy using Average Difference, randomly selected Δs, and Δ_h values

Dataset	*n*	Classification problem	Error rate		
			Δ_h (%)	Random Δs (%)	Average Difference (%)
Leukemia (set A)	35	ALL vs AML	3	2 ± 1	3
Leukemia (set B)	23	T-ALL vs B-ALL	0	0 ± 1	0
Lymphoma	58	Cured vs fatal	26	29 ± 5	24
Medulloblastoma	60	Cured vs fatal	18	26 ± 4	24

Conclusions

In conclusion, the empirical approach to probe reduction presented here allows a systematic optimization of individual probe sets. Our studies specifically reinforce the notion that careful selection of probe pairs based on their hybridization behavior is a promising strategy for future chip design. Nevertheless, it remains likely that the use of multiple probes per gene will generate the most accurate and robust detectors. For diagnostic applications in particular, probe redundancy may significantly improve performance. For screening applications, however, the availability of small, limited-probe, genome-wide arrays could be useful.

Materials and methods

Datasets

The raw data analyzed here has been previously reported [6-8] and is available at [5].

Approximation of Average Difference

To estimate the percentage of genes with Δ_h values within 2-fold of the Average Difference, for each gene we compared the value of Δ_h with the Average Difference for this probe set. The percentage of genes within 2-fold of the Average Difference was then averaged over the 176 chips of the training set. To evaluate random probe selection, for each gene a Δ was chosen randomly and the percentage of genes within twofold of the Average Difference was similarly calculated. This process was repeated 20 times and then averaged. Values of both Average Difference and selected Δs were normalized and a threshold set at 100 units. Relative error for the estimates for Δ_h and randomly selected Δ values was calculated as $|\Delta$ - Average Difference$|$/Average Difference.

Rescaling

To account for minor variation in overall chip intensities, Average Difference values were scaled as previously described [8]. For Δ_h values, scaling was adjusted by a slope and intercept obtained from a least-squares linear fit of the Δ_h values for each chip compared to a randomly selected reference chip.

Classification

Average Difference and Δ values were clipped to minimum 20 and maximum 16,000 units. A variation filter was applied that excluded genes that did not vary at least threefold and 100 units across the entire dataset. To compare the classification accuracy for Δs and Average Difference, we applied a *k*-nearest neighbors (*k*-NN) [9] binary classifier, implemented in the software package GeneCluster 2.0 and available at [10], to each of the four classification problems as previously described [8]. Average Difference or Δ feature selection was performed with the signal-to-noise metric [6] $(\mu_{class\ 0} - \mu_{class\ 1})/(\sigma_{class\ 0} + \sigma_{class\ 1})$, where μ and σ represent the mean and standard deviation within each class, respectively, and the top-ranking features were fed into the *k*-NN algorithm. Performance was evaluated by leave-one-out cross-validation, whereby for each sample a prediction was made with a model trained on the remaining samples in the problem set, and the number of classification errors was tallied. Classifiers with variable numbers of features (1-100) and nearest neighbors (k = 3 or k = 5) were tested. The best-performing classification results are reported.

Acknowledgements

We thank Michael Angelo and Michael Reich for programming help, Sridhar Ramaswamy for providing datasets, and Eric Lander for helpful discussions.

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
2. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14:**1675-1680.
3. Schadt EE, Li C, Ellis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *J Cell Biochem Suppl* 2001, **Suppl(37):**120-125.
4. **Affymetrix, Statistical Algorithms Reference Guide, 2001** [http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf].
5. **Whitehead Institute, Center for Genome Research - Cancer Genomics Publications/Projects** [http://www-genome.wi.mit.edu/cancer/pubs/feature_reduction]
6. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
7. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, *et al.*: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8:**68-74.
8. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, *et al.*: **Prediction of central nervous system embryonal tumor outcome based on gene expression.** *Nature* 2002, **415:**436-442.
9. Dasarathy BV: *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* Washington, DC: IEEE Computer Society Press, 1991.
10. **Whitehead Institute, Center for Genome Research - Cancer Genomics Software** [http://www-genome.wi.mit.edu/cancer/software/software.html]