Optimized detection of sequence variation in heterozygous genomes using DNA microarrays with isothermal-melting probes

David Gresham[a,b,c], Bo Curry[d], Alexandra Ward[a,b], D. Benjamin Gordon[d], Leonardo Brizuela[d], Leonid Kruglyak[b,e,f], and David Botstein[a,b,1]

[a]Department of Molecular Biology and [b]Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; [c]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003; [d]Agilent Technologies, Santa Clara, CA 95051; and [e]Department of Ecology and Evolutionary Biology and [f]Howard Hughes Medical Institute, Princeton University, Princeton, NJ 08544

Contributed by David Botstein, December 14, 2009 (sent for review October 6, 2009)

The use of DNA microarrays to identify nucleotide variation is almost 20 years old. A variety of improvements in probe design and experimental conditions have brought this technology to the point that single-nucleotide differences can be efficiently detected in unmixed samples, although developing reliable methods for detection of mixed sequences (e.g., heterozygotes) remains challenging. Surprisingly, a comprehensive study of the probe design parameters and experimental conditions that optimize discrimination of single-nucleotide polymorphisms (SNPs) has yet to be reported, so the limits of this technology remain uncertain. By targeting 24,549 SNPs that differ between two *Saccharomyces cerevisiae* strains, we studied the effect of SNPs on hybridization efficiency to DNA microarray probes of different lengths under different hybridization conditions. We found that the critical parameter for optimization of sequence discrimination is the relationship between probe melting temperature (T_m) and the temperature at which the hybridization reaction is performed. This relationship can be exploited through the design of microarrays containing probes of equal T_m by varying the length of probes. We demonstrate using such a microarray that we detect >90% homozygous SNPs and >80% heterozygous SNPs using the SNPScanner algorithm. The optimized design and experimental parameters determined in this study should guide DNA microarray designs for applications that require sequence discrimination such as mutation detection, genotyping of unmixed and mixed samples, and allele-specific gene expression. Moreover, designing microarray probes with optimized sensitivity to mismatches should increase the accuracy of standard microarray applications such as copy-number variation detection and gene expression analysis.

DNA/DNA hybridization | sequence discrimination | single-nucleotide polymorphisms | melting temperature | probe design

The original motivation for the development of DNA microarrays by the group of Edwin Southern was the identification of DNA sequence variation (1). Early studies by Southern and others showed that when short single-stranded DNA probes are affixed to a solid surface, the efficiency with which they form duplexes with single-stranded DNA in free solution is sensitive to the presence of single-base-pair mismatches. This made it feasible to detect the presence of single-nucleotide polymorphisms (SNPs) in a DNA sample on the basis of hybridization efficiency to DNA probes of known sequence. The ability to discriminate DNA sequence using microarrays of many hundreds of thousands of oligonucleotide probes underpins a number of DNA microarray applications, including multiplex genotyping of SNPs (2), mutation detection (3–5), and resequencing by hybridization (6). Mutation detection using microarrays remains a cheap and simple means of characterizing nucleotide variation in small genomes (1, 4); however, the extent to which this approach is extensible to more complex genomes or the detection of heterozygous mutations remains unclear. Additional emerging applications make use of SNP-specific DNA probes including global studies of allele-specific gene expression (7) and quantitative genotyping of pooled samples for bulk-segregant genetic mapping (8–10).

Despite the myriad applications of DNA sequence discrimination using DNA microarrays, a comprehensive empirical study of the parameters important for optimizing sequence discrimination on microarrays has not been performed. Furthermore, probe design rules that are relevant for DNA microarrays intended for other uses, such as gene expression analysis (11), DNA barcode measurements (12), or detection of copy-number variation, are not necessarily relevant for sequence discrimination applications.

A large body of literature regarding the thermodynamics of duplex formation (13–18) is relevant to DNA microarray design. The free energy of duplex formation (ΔG°) is best estimated by a nearest-neighbor (NN) model, which assumes that the stability of a given base pair depends on the identity and orientation of the neighboring base pair (17). Empirically determined enthalpic (ΔH°) and entropic (ΔS°) values have been determined for all 10 NNs, and therefore ΔG° is readily determined using the relationship $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$. The ability to discriminate DNA sequence on the basis of hybridization requires that the difference between the free energy of hybridization of perfectly matched duplex (ΔG°_{PM}) be significantly less than the free energy of hybridization of mismatched DNA (ΔG°_{MM}). For duplex formation, a useful metric is the melting temperature of the duplex (T_m), which is the temperature at which half the DNA strands are in a double-helix state. The T_m of a given sequence is calculated according to the relationship $T_m = \Delta H^\circ \times 1000/(\Delta S^\circ + R \times \ln(C_T/x)) - 273.15$, where R is the gas constant (1.9872 cal/K mol), C_T is the total molar strand concentration, and x = 4 for non-self-complementary duplexes (19). Thermodynamic parameters have been determined for all possible mismatches, the majority of which are destabilizing of duplex formation (19) and thus decrease the T_m of the mismatched duplex. Thus, in principle, it should be possible to estimate the ideal probe design such that the T_m of the matched duplex is much greater than the T_m of the mismatched duplex. In practice, however, for multiplex scenarios, other factors must be considered: in particular the specificity of the probe within a genomic context and the fact that there are additional reactions competing with the bimolecular reaction necessary for duplex formation. Furthermore, the vast majority of thermodynamic studies of duplex formation has been performed in

Author contributions: D.G., B.C., D.B.G., L.B., L.K., and D.B. designed research; D.G. and A.W. performed research; D.G. and B.C. analyzed data; and D.G. and D.B. wrote the paper.

Conflict of interest statement: B.C., D.B.G. and L.B. are, as indicated, Agilent employees. They performed research as part of this employment.

Freely available online through the PNAS open access option.

Database deposition: The sequence reported in this paper has been deposited in the Gene Expression Omnibus (GEO) at NCBI under series number GSE19319.

[1]To whom correspondence should be addressed. Email: botstein@genomics.princeton.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0913883107/DCSupplemental.

solution, whereas microarrays involve one strand affixed to a solid substrate and one strand in free solution. The effects of this asymmetry, and the importance of hybridization conditions, substrate concentration, and signal to noise, require empirical determination.

The purpose of this study was to identify the optimized microarray design and hybridization conditions for discriminating sequence variation on microarrays. We sought especially to determine whether an optimized microarray design makes it feasible to detect heterozygous mutations in a diploid genome on the basis of hybridization efficiency. We made use of two fully sequenced yeast genomes that contain 24,549 sequence-verified SNPs to test the effect of single-nucleotide mismatches on hybridization efficiency. We identified a relationship between the hybridization temperature and the T_m of the probe, regardless of probe length, that maximizes the sensitivity of a DNA probe to mismatches. We used this finding to guide construction of a DNA microarray in which probes are designed to have a homogeneous T_m of ≈57 °C by varying their length between 16 and 35 nucleotides. Using this isothermal microarray design, we demonstrate the sensitivity of the SNPScanner algorithm for the detection of homozygous and heterozygous mutations. The optimized design parameters identified in this study should prove useful for guiding future microarray design for a variety of sequence-specific applications.

Results

To study the parameters that are important for sequence discrimination using DNA microarrays, we made use of the complete genome sequences available for two strains of *Saccharomyces cerevisiae*: the S288c reference sequence (hereafter, the reference genome) and the RM11-1a sequence (hereafter, the nonreference genome). Our previous analysis had identified 24,549 sequence-verified SNPs between these two strains that are separated by at least 25 nucleotides (4). To study the effect of microarray probe length on sensitivity to mismatches, we designed three different test microarrays each containing DNA oligonucleotides of length 20, 25, or 30 bases that were tiled in an overlapping manner across SNP sites. The ~240,000 DNA probes were designed to be perfectly complementary to the reference genome. The position of each probe relative to the SNP was systematically altered so that all possible mismatched positions within the probe were equally represented across the array. In addition, two probes were designed to flank, but not cover, each SNP. These probes covered regions that have identical sequence in the reference and nonreference genomes (see *Methods*).

Probe T_m Determines Optimal Conditions for Sensitivity to Mismatches. To test systematically the effect of probe length and hybridization conditions on sequence discrimination, we cohybridized reference (Cy5-labeled) and nonreference (Cy3-labeled) genomic DNA to the three test microarrays containing probes of 20-, 25-, or 30-nucleotide length. Hybridization experiments were performed at 5 °C increments from 45 to 65 °C (Table S1). For all experiments, the set of probes that spanned nonpolymorphic regions of the genome (between 42,867 and 49,587 probes depending on the microarray) was used to normalize the microarrays.

We first performed experiments on test microarrays using DNA from haploid yeast strains. To assess the sensitivity of probes for each microarray to mismatches, we determined the median ratio (expressed as a $\log_2$ value) of all probes that contain a polymorphic site in the genomic DNA sample regardless of the exact position of the SNP in the probe (Fig. 1*A*). These experiments demonstrate that the sensitivity of duplex formation to mismatches increases with decreasing probe length under identical hybridization conditions. Furthermore, performing hybridization reactions at higher temperatures increases the sensitivity of hybridization to mismatches for all probe lengths. We performed the same series of experiments using DNA from a diploid produced by mating the reference and nonreference strains together, which ensures that the genomic DNA is heterozygous at all 24,549 SNP sites. At heterozygous sites, one allele is perfectly complementary to the probe and one allele contains a mismatch. The optimal ratio that can be expected in these cases is 0.5 ($\log_2 = -1$). As with haploid DNA, we observed increased sensitivity to mismatches with both decreased probe length and increased hybridization temperature for heterozygous DNA (Fig. 1*B*).

Previous studies have found that where the mismatch occurs in the probe is a major determinant of the perturbation on hybridization (4, 20). Namely, more central mismatches have much great effect on hybridization than those occurring at the terminal positions. We found that this effect holds for all probe lengths (Fig. S1). In free solution, terminal mismatches have been reported to have a stabilizing effect on duplex formation (19). In contrast, we found that all terminal mismatches on a microarray result in decreased hybridization.

Another well-known parameter affecting duplex formation is the proportion of bases in the probe that are either guanines or cytosines (%GC). This metric is often used as a proxy for probe T_m. We computed the melting temperatures for all probes on the three microarrays using NN parameters (18). T_m and %GC content are correlated for all probe lengths and, in general, probe-melting temperatures increase with probe length (Fig. S2). Within each microarray probe, T_ms are widely distributed with a standard deviation of ≈5 °C (Table 1).

We determined the relationship between sensitivity to mismatches and T_m at different hybridization temperatures for each microarray (Fig. 2). For this purpose we used only those central positions of DNA probes that are most sensitive to mismatches. In general, there is reduced sensitivity to sequence mismatches with increased probe T_m for all hybridization temperatures. We also observed that discrimination is reduced at probe T_ms well below the hybridization temperature. The optimal relationship between probe T_m and hybridization temperature occurs where

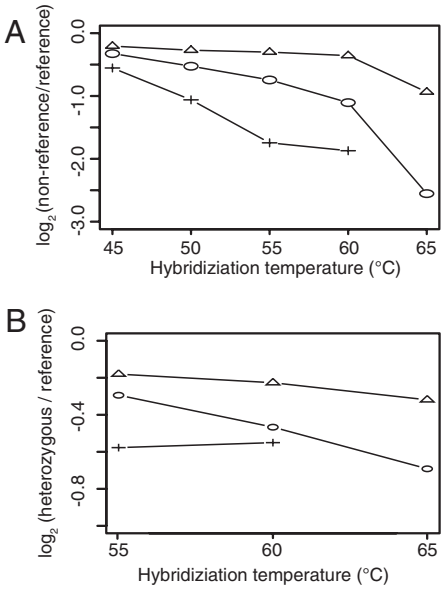

Fig. 1. Sequence discrimination depends on probe length and hybridization temperature. (*A*) When hybridization is performed at a given temperature, probes of length 20 nucleotides (plus signs) exhibit enhanced sensitivity to mismatches over probes of 25 nucleotides (circles) and 30 nucleotides (triangles). Discrimination improves with increased temperature for all probe lengths. (*B*) Hybridization of DNA from a diploid yeast that is heterozygous at all 24,549 SNP sites shows a similar trend with respect to both probe length [20-mers (plus signs), 25-mers (circles), and 30-mers (triangles)] and hybridization temperature.

GENETICS

Table 1. Melting temperature (T_m) for microarrays

Microarray Probe Length	Melting Temperature (°C) (Mean ± 1 SD)
20 nucleotides	50.6 ± 5.7
25 nucleotides	59.6 ± 5.3
30 nucleotides	65.9 ± 5.1
16–35 nucleotides	57.7 ± 0.84

the smoothed curve reaches a minimum. This optimum, which is clearest for hybridization temperatures greater than 50 °C, appears to occur when probes have a melting temperature ≈5 °C lower than the temperature at which the hybridization reaction is performed. This relationship is independent of probe length, as it is observed in microarray experiments using probes of 20- (Fig. 2*A*), 25- (Fig. 2*B*), and 30- (Fig. 2*C*) nucleotide length.

A comparison of hybridization efficiencies for reference and nonreference DNA makes clear the basis of this relationship: for any given hybridization temperature and probe length, as the probe melting temperature increases, total hybridization increases (Fig. S3). The thermodynamic cost of a mismatch is maximized when the hybridization temperature is 5 °C higher than the melting temperature of the probe. This penalty is reduced for probes of higher melting temperature. For probes with T_m much lower than the hybridization temperature, the hybridization efficiency of perfectly matched DNA is reduced, and thus the thermodynamic cost of a mismatch is less pronounced.

Performance of Microarrays with Isothermal-Melting Variable-Length Probes. To exploit this newly discovered relationship between probe T_m and sensitivity to mismatches, we designed a DNA microarray for which we aimed to establish uniform probe melting temperatures by varying the length of the probes. We designed probes with a target T_m of 57 °C, computed using NN parameters (17), by varying the probe length between 16 and 35 nucleotides (see *Methods*), and tiled them across the 24,549 SNPs that differ between the reference and nonreference genomes. The modal probe length for this microarray is 24 nucleotides (Fig. S4) and the T_ms for all probes on the array are tightly distributed with a standard deviation nearly one order of magnitude less than that of arrays with fixed probe length (Table 1). By cohybridizing reference and nonreference DNA at 55, 60, and 65 °C, we found that the best temperature for hybridization was 60 °C, consistent with our observations using fixed probe length microarrays (Table S2). We infer, from the fixed length data, that hybridization at 62 °C might be slightly better still.

We performed four hybridization experiments using haploid DNA and four hybridizations using heterozygous diploid DNA at 60 °C. Experimental results were highly reproducible (pairwise correlations > 0.87). We confirmed that the relationship between a mismatched position within a probe and sensitivity holds for an isothermal probe design for both haploid (Fig. S5*A*) and hetero-

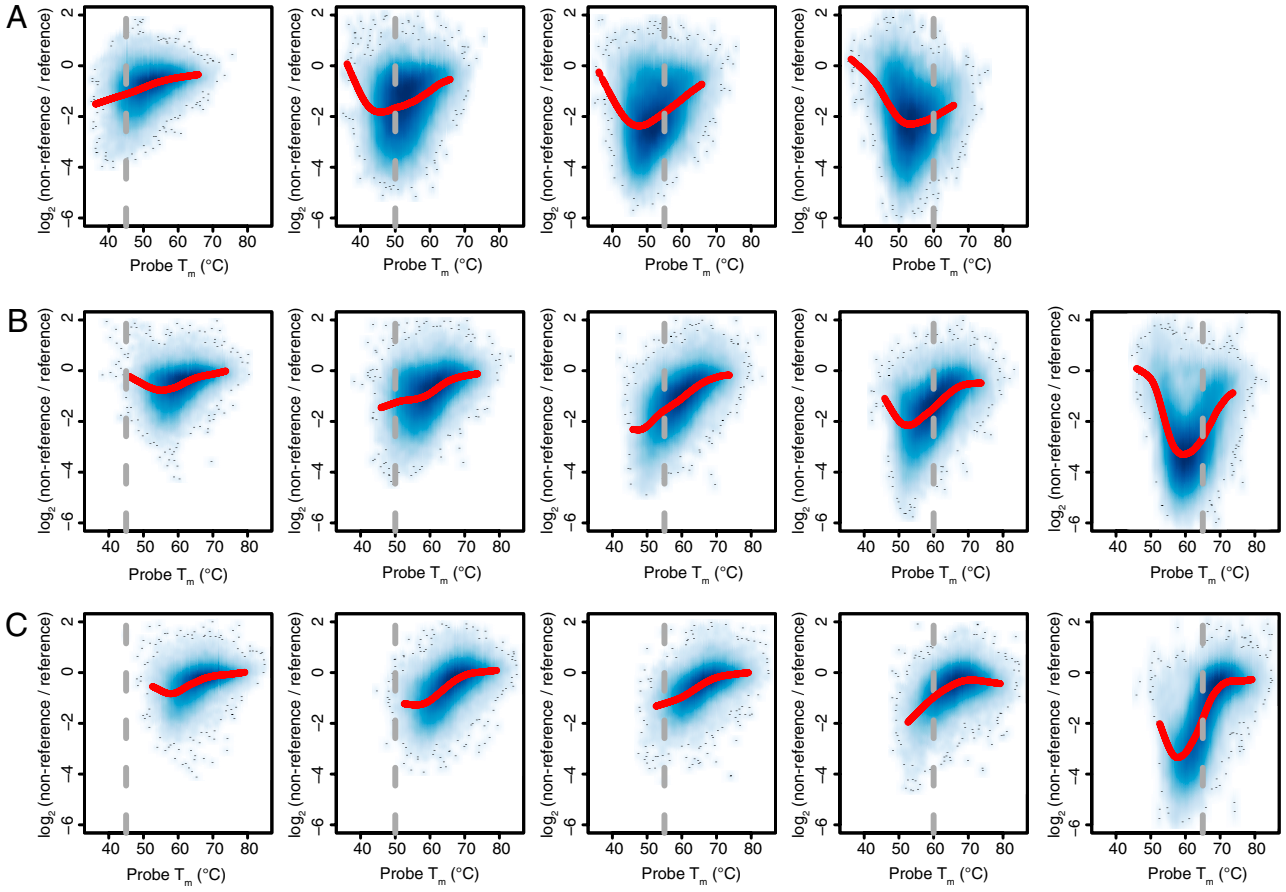

Fig. 2. The relationship between sensitivity to sequence mismatches and DNA probe melting temperature for probes of length (*A*) 20, (*B*) 25, and (*C*) 30 nucleotides. For each hybridization experiment performed at increasing temperatures from 45 to 65 °C (indicated in gray), the distribution of $\log_2$ ratios for probes in which the SNP results in a mismatch in the central region of the probe is shown. Darker blue regions represent a greater density of points from a total of 34,000–42,000 points in each experiment. To summarize the data, we superimposed a spline fit to the data (red line).

zygous (Fig. S5*B*) DNA. Consistent with an optimization of sequence discrimination by isothermal probes, ratios are very close to the theoretical optimum of $\log_2$ ratios of −1 for heterozygous genomes (Fig. S5*B*). For both haploid and heterozygous samples, probes behave predictably for lengths between 19 and 30 nucleotides. Shorter probes (16–18 nucleotides) have extremely high % GC, whereas longer probes (31–35 nucleotides) have extremely high %AT content (Fig. S6*A*), and in both cases behave less predictably. Using our T_m-matched design parameters with a target T_m of 57 °C, 225,100 (95.5%) probes are between 19 and 30 nucleotides in length.

We examined the effect of each possible mismatched base pairing on hybridization efficiency (Fig. 3). The maximal effect of a mismatch occurs when a mutation in the sample DNA, present in free solution, results in a mismatch with a cytosine in the probe. C-C mismatches have the greatest effect, followed by C-T and C-A mismatches, which have a greater effect than all other possible mismatches. This is consistent with known thermodynamic properties of mismatches in which mismatches with C are the weakest (19). This effect is not symmetrical, as T-C and A-C mismatches in which the A or T are in the DNA probe result in significantly less perturbation of hybridization. In fact, this is true of all pairs of symmetrical mismatches due to fact that the computed ratio is dependent on both the identity of the mismatched base pair and the perfectly matched base pair determined by the base in the probe. The smallest effect of mismatches is observed when a T or A is mismatched with G. This is consistent with the known promiscuity of G, as it forms the strongest mismatches (19).

Microarrays with Isothermal-Melting Probes Efficiently Detect Heterozygous Mutations. Previously, we developed the SNPScanner algorithm, which accurately detects the presence of >85% of SNPs in haploids using an array of fixed probe length (25 nucleotides) with an average of 4-base-pair spacing between probes (4). We tested the performance of the SNPScanner algorithm on isothermal arrays by performing holdout analyses from individual hybridization experiments. We performed multiple tests in which we trained a model using data from 23,549 randomly selected SNPs and tested the detection ability of the algorithm on sets of 1,000 random test SNPs held out from the training set. The SNPScanner algorithm computes the likelihood that a site in the genome is polymorphic. For ratiometric data that are $\log_2$-transformed, the likelihood calculation reduces to

$$L_k = \log_{10} e \sum \frac{2 \cdot x_i \cdot \mu_p - \mu_p}{2\sigma^2}$$

The variance of ratios differs for different probe lengths for cohybridized identical DNA sequences on isothermal arrays (Fig. S6*B*). Therefore, we employed a probe-length-specific variance (σ^2) measure for the likelihood calculation determined from a microarray to which reference DNA had been hybridized in both channels. The likelihood is computed for site k in the genome using the experimentally determined intensity (x) in probe i and summed for all probes containing site k. μ_p is the modeled value for a SNP in probe i complementary to site k. For each training/test set, we estimated our false negative rate to be the fraction of the 1,000 test SNPs that we failed to detect. To estimate our false positive rate, we used the same holdout procedure for training but tested detection of the 1,000 SNPs in a cohybridization experiment of differentially labeled reference DNA in which we expect to detect no SNPs.

We performed 10 independent tests per hybridization experiment on 4 replicate hybridizations. Our 40 tests used an average of 165,744 probes to train the algorithm and an average of 9,033 probes to predict the presence of 1,000 SNPs. Of these probes, 2,000 flanked but did not cover an SNP. Therefore, our test set comprised an average of 7 probes per SNP, which is the same probe density as the genome-wide tiling array used in our initial study (4). We observed an average true positive rate of 92.3%. Over 90% of the SNPs were detected with a $\log_{10}$ likelihood score greater than 2 and the magnitude of likelihoods ranged to values over 100 (red line in Fig. 4). These results held for data from four independent hybridization experiments (Fig. S7*A*).

We performed the same holdout procedure in which we applied the trained algorithm to random selections of 1,000 known SNP sites from a self-self hybridization. From 10 tests of 1,000 sites, we found one example of a $\log_{10}$ likelihood value

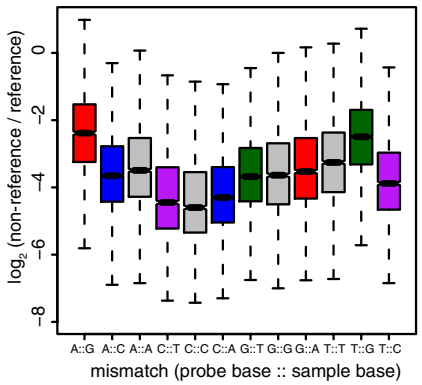

Fig. 3. Differential effect of mismatches on annealing efficiency at isothermal probes. The median ratio for the interquartile region of probes of all lengths is plotted for each possible mismatch. The first nucleotide is the base present in the probe. The second nucleotide is the base present in the genomic DNA sample. Homo-mismatches are in gray. Symmetrical hetero-mismatches are shown in the same color.

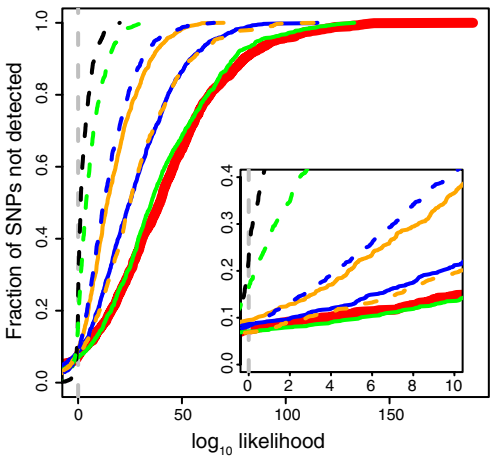

Fig. 4. SNPScanner performance using ratiometric data from an isothermal microarray. Using an average of 7 probes per SNP site, the SNPScanner algorithm accurately predicts 93% of haploid SNPs, with a $\log_{10}$ likelihood value greater than 0 (red line). Decreasing the average number of probes interrogating each SNP to an average of 4.8 (green line), 3.4 (blue line), or 2.0 (orange line) reduces the overall magnitude of the likelihood values for SNP predictions, but does not appreciably increase the false negative rate. The position of the SNP in the probes affects the fraction of SNPs not detected (i.e., false negative rate; see *Inset*): For SNP positions in the outer 30th percentile (green dashed line; 2.1 probes/SNP), 836/1000 SNPs are predicted. We only detect 792/1000 SNPs when the SNPs only occur in the outer 20th percentile of probes (black dashed line; 1.5 probes/SNP). Constraining SNP positions to the outer 50th (blue dashed line; 3.5 probes/SNP) or 70th percentile (orange dashed line; 5.0 probes/SNP) did not increase the false negative rate.

greater than 0, indicating that our false positive rate is of the order 10^{-4}. Using a threshold $\log_{10}$ likelihood value of 2 results in 0 false positives and >90% true positives.

We sought to determine the required number of probes for accurate detection of SNPs by excluding subsets of probe data from the test set of SNPs. For this purpose, the training set contained the full complement of probe information. As expected, reducing the number of independent measurements reduced the magnitude of the total likelihood values (Fig. 4). However, we found that we were still able to detect over 90% of test SNPs using as few as two probes per SNP. To assess the effect of the placement of SNPs within probes on the detection quality, we constrained mismatched sites for test SNPs to increasingly terminal regions of probes. We found that a significant decrease in the fraction of SNPs detected only occurs once mismatches are constrained to the outer 30th percentile of probes (83% true positives; 836/1,000 SNPs predicted; see Fig. 4 *Inset*).

Previously, our attempts to detect heterozygous SNPs with the SNPScanner algorithm using a tiling array with a fixed probe length of 25 nucleotides had proven unsuccessful. We investigated whether the increased specificity of T_m-matched probes makes it feasible to detect the presence of heterozygous SNPs. We performed the same test procedure by withholding data for 1,000 SNPs and training the algorithm with the rest of the data. We were able to predict an average of 829.5/1,000 SNPs from 10 independent tests each from four independent hybridizations. Likelihood scores were much smaller in magnitude than haploid SNP prediction (Fig. S7*B*), consistent with the decreased difference in $\log_2$ ratio between expected polymorphic and nonpolymorphic values. Whereas over 80% of heterozygous SNPs are predicted from a single hybridization, the false positive rate when the same method is applied to self-self hybridization data is around 8% (Fig. 5). The false positive rate can be reduced by imposing higher cutoffs: Using a cutoff score of 2 results in 75% true positive calls with a 2.3% false positive rate. Although the false positive rate is prohibitively high on a genome-wide scale, the use of additional heuristic criteria to filter SNP calls such as those used in our original report of SNPScanner (4) can potentially reduce the total number of SNP calls to a more manageable number when applied on a whole-genome scale.

Discussion

We have discovered that the sensitivity of a DNA probe to a single mismatch is maximized when hybridization is performed at a temperature ≈2–5 °C higher than the probe T_m. We used this discovery to guide the design and construction of an isothermal probe design in which probe length is varied between 16 and 35 nucleotides to ensure a homogeneous melting temperature of duplex DNA. We have shown that this array design universally increases the sensitivity of DNA probes, enabling accurate SNP detection in haploid and heterozygous DNA samples.

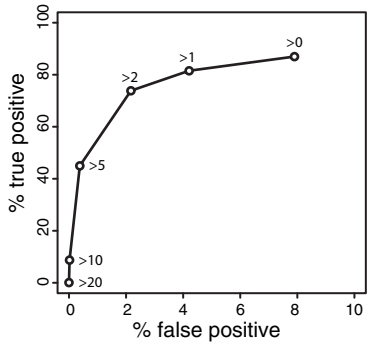

Fig. 5. The SNPScanner algorithm is able to correctly predict the presence of 83% heterozygous SNPs, with a 7.2% false positive rate. The number of false positives rapidly decreases with increasing $\log_{10}$ likelihood cutoffs (numerical values adjacent to points).

Our study is not the first to make use of isothermal probe designs. Previously, mutation detection using isothermal microarrays has been attempted in *Helicobacter pylori* (5), *Escherichia coli* (21), and *Plasmodium falciparum* (22); however, these studies have been characterized by high false negative rates (22, 23). The findings from this study make clear why that is the case; first, for comprehensive mutation detection it is essential that probes overlap, as the ability to detect mismatches that occur at the termini of probes is poor, and second, it is essential to consider the relationship between probe T_m and hybridization temperature. Previous studies (22) using "isothermal" probe designs have allowed for a far greater distribution of probe T_ms (60–80 °C), and hybridization has been performed at a temperature (42 °C) far from the optimal relationship discovered in this study.

The thermodynamics of duplex formation are affected by other factors, including salt concentration and the presence of denaturants such as formamide or urea. Therefore, it should be noted that the details of the relationship identified in our study might only apply to the specific composition of hybridization solution used. We expect, however, that the general relationship should hold, although its refinement of the optimal conditions may have to be empirically determined for other buffer compositions.

As well as identifying the appropriate relationship between probe T_m and hybridization temperature, we discovered asymmetries in the effect of mismatches. The most extreme of these is the effect of mismatches with C, which is greatest when the C occurs in the probe. This discovery has practical implications for designing microarrays that interrogate double-stranded DNA for applications such as genotyping: Where possible, a probe containing a C should be preferred over a probe containing a G.

Our isothermal microarray was designed with a target T_m of 57 °C and provided best discrimination when DNA samples were hybridized at 60 °C. Although it remains untested, it seems probable that designing microarrays with a higher target probe T_m and hybridizing at temperatures 2–5 °C higher should provide equal sensitivity. This has the advantage of allowing the design of longer probes, which increases their specificity in a genomic context. Hybridization temperatures above 65 °C are generally avoided due to limitations of standard hybridization ovens and, thus, this is likely to be the upper bound. Further enhancements of stringency by using denaturants may make it possible to increase sensitivity without increasing the hybridization temperature.

Design Guidelines for Isothermal Microarrays. The design and experimental guidelines derived from this study that are relevant to either genotyping or mutation detection microarrays can be summarized as follows:

(*i*) Design probes with a target T_m of 57 °C and perform hybridization experiments at 60–62 °C.
(*ii*) Exclude probes that are shorter than 19 bp or longer than 30 bp.
(*iii*) When assaying double-stranded DNA for genotyping or other applications, use the relevant strand such that:
  a. C, not G, occurs in the probe.
  b. An A-C mismatch is formed instead of T-G.
  c. A T-C mismatch is formed instead of A-G.
(*iv*) For mutation detection arrays, overlap probes so that every nucleotide position falls within the inner 70th percentile of at least one probe.

These rules should be employed in conjunction with standard probe design rules including the use of unique sequences and an

absence of repetitive sequences and sequences with predicted secondary structures (11). Clearly, for genotyping microarrays, inclusion of isothermal probes perfectly complementary to each allele for a given SNP will ensure high-confidence genotyping. Using this approach, cross-hybridization of the two alleles will be minimized, enabling accurate determination of the proportion of each allele in the sample for applications such as bulk-segregant mapping and allele-specific expression. It is possible that using these design guidelines will also improve the accuracy of quantification of copy-number variation and gene expression, as cross-hybridization to off-target DNA should be greatly reduced.

Comprehensive mutation detection using microarrays enables the global analysis of large numbers of samples to study intraspecific variation (24), the products of evolution experiments (25), and genetic selections (26). Dense SNP genotyping has enabled high-resolution global studies of recombination (27) and allele-specific expression (7). Although it is conventional to believe that high-throughput sequencing will overtake DNA microarrays for all applications (28, 29), we believe that optimized microarrays designed following our guidelines will find many applications. One reason may be cost, but others include the possibility of accurate determination of allele frequencies in mixed-DNA samples for applications such as bulk-segregant mapping (8) and allele-specific expression. These methods and others that require quantitative allele-specific information should be greatly enhanced by the optimized design parameters identified in this study.

Methods

Microarray Design and Manufacture. Probes for DNA microarrays were designed complementary to genomic loci containing the 24,549 SNPs that differ between the S288c (reference) and RM11-1a (nonreference) genomes and are spaced at least 25 nucleotides apart. Probes were tiled across each SNP with their position relative to the SNP systematically varied. For each SNP, two flanking probes were designed that flank the SNP. To design isothermal microarrays, we used custom scripts to calculate T_ms using NN parameters (17). Stilts of either 6 or 10 monomeric dT were added to each probe.

Hybridization Conditions. Genomic DNA was fragmented using a sonicator and labeled with Cy3 or Cy5 using random primed Klenow enzyme labeling at 24 °C, resulting in labeled fragments of ≈100 bases. For initial experiments using test microarrays we used 2000 ng, but all subsequent isothermal array hybridizations were performed using 200 ng of each labeled DNA sample corresponding to ≈50 amoles of DNA. Each microarray feature is about 60 amoles. As there are approximately five features competing for each target, the molar ratio of probe:target is ≈3:1. Samples were cohybridized in Agilent Hi-rpm 2× hybridization buffer, with a final concentration of 750 mM Li^+. Microarrays were hybridized at the specified temperature for 16 h. Arrays were washed with a low-stringency buffer followed by a high-stringency buffer and finally by immersion in acetonitrile. Microarrays were scanned using an Agilent DNA microarray scanner at 5 μm pixel size using the XDR setting.

Probe Melting Temperature Calculations. T_m was calculated using the relationship $T_m = \Delta H^\circ \times 1000/(\Delta S^\circ + R \times \ln(C_T/x)) - 273.15$. For enthalpic calculations, we used the NN parameters of ref. 18 and then computed T_m using R = 1.9872 cal/Kmol, x = 4, and a strand concentration of 0.6×10^{-12} M.

Data Processing. Microarrays were normalized using the set of ∼48,000 probes that targeted identical sequences in the reference and nonreference genomes. A linear-lowess normalization method implemented in the Agilent Feature Extractor software was used.

SNPScanner Algorithm. The SNPScanner algorithm was implemented in R. For each probe length, we modeled the $\log_2$ ratio for a SNP at each site in each probe as $\mu_p = \alpha + \beta_0(GC) + \beta_1(\text{nucleotide}) + \varepsilon$.

The coefficients are the position of the mismatch in the probe (α), the %GC of the probe, and the identity of the base in the probe (A, C, T, or G). These parameters differ from those used in our original implementation of the SNPScanner algorithm, in which we included the triplet sequence at each site and the intensity measure at the corresponding mismatched probe on the Affymetrix tiling microarray (4). We partitioned the data according to probe length and applied this same model for each subset of data. We did not include interaction terms.

ACKNOWLEDGMENTS. Research was supported by an NIH research grant (GM046406) and the NIGMS Center for Quantitative Biology (GM071508). All microarray data have been deposited in GEO under series number GSE19319.

1. Gresham D, Dunham MJ, Botstein D (2008) Comparing whole genomes using DNA microarrays. *Nat Rev Genet* 9:291–302.
2. Lipshutz RJ, et al. (1995) Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 19:442–447.
3. Winzeler EA, et al. (1998) Direct allelic variation scanning of the yeast genome. *Science* 281:1194–1197.
4. Gresham D, et al. (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* 311:1932–1936.
5. Albert TJ, et al. (2005) Mutation discovery in bacterial genomes: Metronidazole resistance in *Helicobacter pylori*. *Nat Methods* 2:951–953.
6. Wong CW, et al. (2004) Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res* 14:398–405.
7. Gagneur J, et al. (2009) Genome-wide allele- and strand-specific expression profiling. *Mol Syst Biol* 5:274.
8. Borevitz JO, et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res* 13:513–523.
9. Segrè AV, Murray AW, Leu JY (2006) High-resolution mutation mapping reveals parallel experimental evolution in yeast. *PLoS Biol* 4:e256.
10. Brauer MJ, Christianson CM, Pai DA, Dunham MJ (2006) Mapping novel traits by array-assisted bulk segregant analysis in *Saccharomyces cerevisiae*. *Genetics* 173: 1813–1816.
11. Hu G, Llinás M, Li J, Preiser PR, Bozdech Z (2007) Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. *BMC Bioinformatics* 8:350.
12. Xu Q, Schlabach MR, Hannon GJ, Elledge SJ (2009) Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc Natl Acad Sci USA* 106:2289–2294.
13. Alemayehu S, et al. (2009) Influence of buffer species on the thermodynamics of short DNA duplex melting: Sodium phosphate versus sodium cacodylate. *J Phys Chem B* 113:2578–2586.
14. Fish DJ, et al. (2007) DNA multiplex hybridization on microarrays and thermodynamic stability in solution: A direct comparison. *Nucleic Acids Res* 35:7197–7208.
15. Fish DJ, Horne MT, Searles RP, Brewood GP, Benight AS (2007) Multiplex SNP discrimination. *Biophys J* 92:L89–L91.
16. Horne MT, Fish DJ, Benight AS (2006) Statistical thermodynamics and kinetics of DNA multiplex hybridization reactions. *Biophys J* 91:4133–4153.
17. SantaLucia J, Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95:1460–1465.
18. SantaLucia J, Jr, Allawi HT, Seneviratne PA (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35:3555–3562.
19. SantaLucia J, Jr, Hicks D (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33:415–440.
20. Ronald J, et al. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res* 15: 284–291.
21. Herring CD, et al. (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* 38: 1406–1412.
22. Tan JC, et al. (2009) Optimizing comparative genomic hybridization probes for genotyping and SNP detection in *Plasmodium falciparum*. *Genomics* 93:543–550.
23. Herring CD, Palsson BØ (2007) An evaluation of comparative genome sequencing (CGS) by comparing two previously-sequenced bacterial genomes. *BMC Genomics* 8: 274.
24. Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 458:342–345.
25. Gresham D, et al. (2008) The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet* 4:e1000303.
26. Ho CH, et al. (2009) A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol* 27:369–377.
27. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.
28. Kahvejian A, Quackenbush J, Thompson JF (2008) What would you do if you could sequence everything? *Nat Biotechnol* 26:1125–1133.
29. Shendure J (2008) The beginning of the end for microarrays? *Nat Methods* 5:585–587.