Nucleic Acids Research, 2005, Vol. 33, No. 12 ***e103***
doi:10.1093/nar/gni105

Improved microarray methods for profiling the yeast knockout strain collection

Daniel S. Yuan[1,2,*], Xuewen Pan[1], Siew Loon Ooi[1], Brian D. Peyser[1], Forrest A. Spencer[1], Rafael A. Irizarry[2] and Jef D. Boeke[1]

[1]Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA and [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

Received April 8, 2005; Revised and Accepted June 14, 2005

ABSTRACT

A remarkable feature of the Yeast Knockout strain collection is the presence of two unique 20mer TAG sequences in almost every strain. In principle, the relative abundances of strains in a complex mixture can be profiled swiftly and quantitatively by amplifying these sequences and hybridizing them to microarrays, but TAG microarrays have not been widely used. Here, we introduce a TAG microarray design with sophisticated controls and describe a robust method for hybridizing high concentrations of dye-labeled TAGs in single-stranded form. We also highlight the importance of avoiding PCR contamination and provide procedures for detection and eradication. Validation experiments using these methods yielded false positive (FP) and false negative (FN) rates for individual TAG detection of 3–6% and 15–18%, respectively. Analysis demonstrated that cross-hybridization was the chief source of FPs, while TAG amplification defects were the main cause of FNs. The materials, protocols, data and associated software described here comprise a suite of experimental resources that should facilitate the use of TAG microarrays for a wide variety of genetic screens.

INTRODUCTION

The Yeast Knockout (YKO) strain collection was designed and created by an international consortium of yeast geneticists (1) in the wake of the sequencing of the *Saccharomyces cerevisiae* genome in 1996. Each YKO strain contains a precisely defined null deletion of its specified open reading frame (ORF). Each ORF is substituted by a kanamycin drug resistance cassette. The value of this collection for functional genomics lies in the ease with which each strain can be systematically tested for how the loss of each gene affects physiological function (2).

All strains in the YKO collection include unique sequence TAGs linked to each YKO mutation (3). Specifically, a 56 bp cassette comprised of a unique 20mer 'TAG', flanked by shared ('universal') primer sites for amplification by PCR, is located immediately 5′ of every YKO mutation. In all but 192 knockouts, a second cassette with similar structure but a different pair of universal primer sites is situated immediately 3′. The TAGs in these upstream and downstream cassettes are termed 'UPTAGs' and 'DNTAGs', respectively. Thus, PCR products obtained using fluorescently labeled universal primers can be synthesized in a single PCR tube and hybridized to microarrays of complementary TAG sequences to profile the relative representation of strains within a complex pool (3). Such a massively parallel technology avoids much of the labor involved in conducting a conventional yeast genetic screen (1–3).

Despite their potential as a high-throughput technology, TAG microarrays have yet to become established in the mainstream of yeast genetics research. We are aware of relatively few applications [e.g. (4–9)]. While we do not necessarily speak for all yeast researchers, our own experience has been that TAG microarrays have either been unavailable as catalog items from commercial sources, or have been sold under provisions that restrict access to the underlying oligonucleotide sequences.

This paper describes a TAG microarray designed to analyze genetic interactions on a large scale. In addition to explaining the features of this design, we describe upgrades to existing TAG PCR and hybridization protocols and analyze per-TAG false positive (FP) and false negative (FN) rates observed with these protocols. It is important to note that since most strains

*To whom correspondence should be addressed. Tel: +1 410 502 1877; Fax: +1 410 502 1872; Email: dyuan@jhmi.edu
Present address:
Siew Loon Ooi, Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

carry two TAGs, these FP and FN rates do not equate to FP and FN rates for overall TAG array performance. Illustrative uses of our microarrays in a wide variety of genetic screens have already been described (9).

MATERIALS AND METHODS

Preparation of genomic DNA

Genomic DNA was prepared from the pooled YKO strains by disruption with glass beads in phenol (10), but with two extra phenol/chloroform and chloroform extractions after RNAse treatment so that spectrophotometry of the product yielded OD_{260}/OD_{280} ratios >1.80.

Precautions against Cy-dye degradation

All dye-containing stock solutions and PCR samples were overlaid with a filtered inert gas (1,1,1,2-tetrafluoroethane, Dust-Pro, Sigma). Atmospheric ozone has been shown to degrade Cy-dyes, especially Cy5 (11).

TAG PCR

Primers were purified by high-performance liquid chromatography (HPLC). Reaction mixes (20 µl/tube) for UPTAG or DNTAG PCR contained 10 µl of 2× ExTaq Premix (Takara), 0.5 µM primer 1 (U1 or D1), 5.0 µM primer 2 (U2c or D2c, 5′-labeled with either Cy3 or Cy5 dye), and at least 200 ng genomic DNA (to minimize sampling artifacts; 200 ng ~1000 diploid yeast genomes, assuming that genomic DNA and mitochondrial DNA copurify). PCR cycling parameters were as follows: 94°C 3 min; 95°C 10 s, 50°C 20 s, 72°C 20 s for 50 cycles; 4°C. Use of a terminal 72°C 7 min step resulted in a ladder of artifactual products, so this step was deleted. After PCR, blocking oligonucleotides (U1 and U2.3, or D1 and D2.3) were added to 10 µM, without further denaturation. EDTA (10 mM) was also added as a precaution against residual polymerase-associated exonuclease activity. Precautions against PCR contamination are detailed in Supplementary Note 1 online.

Gel electrophoresis of TAG PCR products

Raw PCR product (9 µl) was mixed with 1 µl of 40% glycerol/0.05% bromophenol blue and subjected to agarose gel electrophoresis at 12 V/cm for 15 min. Gels contained 3% MetaPhor agarose (Cambrex Biosciences) and ethidium bromide 0.5 µg/ml. Early experiments used a conventional 1× Tris/acetate/EDTA (TAE) buffer; later experiments used 5 mM sodium tetraborate for improved resolution (12). Ethidium fluorescence appeared to be quenched by Cy5-labeled samples (data not shown).

Microarrays

The 'Hopkins Tag Array' microarray design was implemented as Agilent Technologies '22k' microarrays and were manufactured on a custom basis. All non-proprietary sequences are available in the Gene Expression Omnibus (GEO) microarray database (http://www.ncbi.nlm.nih.gov/geo/) under accession number GPL1444. Contact information for procurement is described at our website (http://slam.bs.jhmi.edu).

Microarray hybridization

Hybridization buffer (0.5% Triton X-100, 1 M NaCl and 100 mM Tris–HCl, pH 7.5) (4) was prepared with colorless batches of Triton X-100 as a precaution against excessive background fluorescence in hybridized slides. Triton-salt solutions were not microwaved to avoid emulsion formation. The solution was passed through a 0.2 µm nitrocellulose filter, degassed under vacuum to 400 mTorr, overlaid with tetrafluoroethane gas and supplemented with 1 mM DTT just before use. For the dye-flip experiment, microarray slides were incubated within plastic slide holders filled with silanized glass balls (to reduce dead space) and jacketed in beakers of water maintained at 42°C. For the heterozygous diploid pool experiment, slides were incubated in 5 ml volumes of hybridization buffer in six-compartment polypropylene boxes (Alpha Rho Inc., Fitchberg MA, USA) and rocked in a 42°C hybridization oven (Robbins 400, Robbins Scientific); homogeneous mixing was achieved by setting the box at an angle and minimizing the eccentricity of the rocker. After a 30 min incubation in buffer, with oligonucleotides U1c, U2c, D1c and D2c added to 0.25 µM each to block spurious binding sites, the solution was replaced with buffer at 42°C containing a 1:500 dilution of each TAG PCR product. (For a typical two-color experiment, four such products are combined, corresponding to UPTAGs and DNTAGs amplified with primers labeled with each of the two dyes.) Incubations were at 42°C for 16 h. Compartments were carefully gassed to avoid splashing, the box was enclosed in a zippered bag that was gassed before sealing, and the oven was shielded from light.

Microarray washing

Slides were removed to a Coplin staining jar that was filled with wash buffer at 23°C. The buffer consisted of 6× SSPE + 0.5% Triton X-100 (4), and was filtered, degassed, purged and supplemented with 1 mM DTT as with the hybridization buffer. After 10 cycles of slow withdrawal and reimmersion >10 min, slides were treated the same way in a second Coplin jar identical to the first except for the use of 1× SSPE. Slides were then ready for scanning. A scan of duplicate slides revealed that about half the signal intensity was retained after storage for three days in the 1× SSPE solution.

Microarray scanning

Images were acquired at 10 µm resolution using a GenePix 4000B scanner (Axon Instruments) and analyzed using GenePix Pro 3.0 or 5.1 software (Axon Instruments), with laser power and photomultiplier tube voltages adjusted manually based on preliminary scans. The scanner chamber was gassed with tetrafluoroethane before scanning.

Microarray stripping

Slides were reimmersed in the 1× SSPE wash solution after scanning. To remove bound oligonucleotides, the slides were incubated in stripping solution (1% SDS and 10 mM EDTA) at 65°C for 15 min, soaked in 0.06× SSPE at 23°C for 10 min, slowly withdrawn from the wash solution without retaining residual droplets on the slide, and stored in a plastic slide box. Solutions were passed through a 0.2 µm nitrocellulose filter before use.

Data processing

GenePix result files (marked by a '.gpr' extension) were analyzed using *hoptag*, a software package specially developed for our microarray design. *hoptag* is a customization and extension of the widely used *marray* microarray data analysis package (13), which provided the formal object-oriented model for microarray data used in *hoptag*. *hoptag* was designed to systematize and simplify primitive operations such as data import and the extraction of subsets of microarray features germane to our array design, e.g. UPTAGs, DNTAGs, negative controls and replicates, as well as arbitrary sets of genes identified by their *Saccharomyces* Genome Database (SGD) ID or ORF name. *marray* is part of the BioConductor Project (http://www.bioconductor.org) (14). In turn, BioConductor is the bioinformatics arm of R, a multi-platform, Open Source scripting language and data processing environment freely available under the GNU Public License (http://www.r-project.org). The *hoptag* package and accompanying *hoptagInfo* package of gene annotations (derived from SGD) are available as Supplementary Notes 2 and 3 online in '.tar.gz' format (http://www.gzip.org/) and are released under version 2 of the GNU General Public License (http://www.gnu.org/licenses/gpl.txt). An archive of all the source code and datafiles used in this paper is available as Supplementary Note 4 online.

RESULTS

Design of a TAG microarray

Practical considerations led us to choose a commercial custom microarray fabrication service (Agilent Technologies). The 22k arrays available on this platform contained 21 939 user-defined features. The Yeast Knockout project specified 6018 ORFs as knockout candidates. We were thus able to include all available TAGs in our microarray design. Because TAGs are amplified in separate PCR tubes and fluorescent-labeled extract yields may vary, features for the two TAG types were segregated in separate halves of the microarray (Figure 1a). We developed an elaborate set of experimental controls with two goals in mind: to facilitate assessment of hybridization efficiency, and to promote the statistical analysis and possible future normalization of spatial artifacts.

A key preparatory step was to develop a list of 'artificial' TAGs that would resemble the existing TAGs but would ideally exhibit minimal cross-hybridization. Because the detailed algorithms used to design the existing TAGs were not publicly available, we subjected the existing TAGs to statistical analysis. We found that they satisfied three constraints: (i) All were 20mers with 9–11 G/C bases; (ii) none contained four or more tandem repeats of the same base; and (iii) none had 10 or more bases of exact homology with other TAGs on the same strand. These constraints were used to filter a randomly generated list of 40 000 20mers, with additional constraints to minimize other types of spurious hybridization (homology with antisense sequences or with other artificial TAGs; universal primer sequences were not included in this search) (see the 'SyntheticTags' subdirectory archived in Supplementary Note 4 online). This computation yielded 1099 new TAGs, of which 193 were assigned as aliases for those ORFs lacking DNTAGs (one ORF was listed twice). The remaining TAGs were sequentially assigned, for consistency of nomenclature, as UPTAG and DNTAG pairs corresponding to ORFs on a non-existent 'chromosome 17' (YQL002C, YQL003C, etc). (See Methods for sequence accession information.)

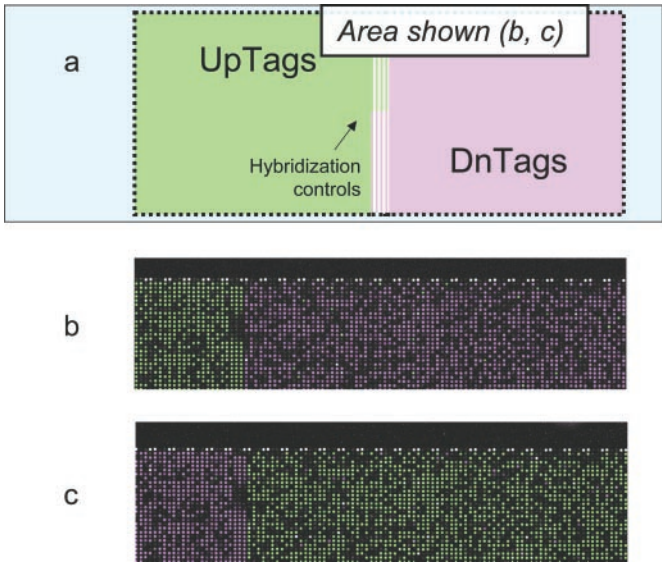

Figure 1. Detection of UPTAGs and DNTAGs in a dye-flip microarray experiment. (**a**) Schematic of the microarray layout (not to scale). (**b**) False-color ratio image of Cy3-labeled UPTAGs and Cy5-labeled DNTAGs hybridized to a microarray. Green and magenta colors signify Cy3- and Cy5-predominant signals, respectively. Image intensities increase as the square root of average signal in the two signal channels but have been enhanced towards saturation for clarity. (**c**) Same as part (b) but with Cy5-labeled UPTAGs and Cy3-labeled DNTAGs hybridized to the same microarray after removing bound signal.

To assess hybridization efficiency in intact and intentionally and systematically 'mutated' TAGs, we chose an UPTAG and DNTAG corresponding to each of three YQL ORFs, and assigned them as microarray features in a bar-shaped domain of the array that separated the UPTAG and DNTAG halves of the array. For each of the six TAGs, we included 17 random substitutions, 17 single-base deletions and 17 truncations of varying lengths from either end. TAGs foreshortened by deletions or truncations were lengthened to their original 20 bases by adding T's at the 3′ end (towards the glass substrate of the array).

The other controls consisted of a set of 959 pairs of 5-fold replicates. The number 5 was chosen as a compromise between maximizing the statistical power of a *t*-test (which decreases dramatically when the number of degrees of freedom is small), and maximizing the number of TAGs with replicates. Eight hundred pairs of TAGs corresponding to actual genes were chosen for replication, including all genes with genetic interactions recorded at the Munich Information Center for Protein Sequences (MIPS) as of June 2002 (http://mips.gsf.de), an *ad hoc* set of genes of special interest to our laboratory, and a random subset of genes implicated in protein–protein interactions as recorded at MIPS. In addition, 159 pairs of YQL TAGs were also chosen for replication. While the systematic set of existing TAGs was distributed in a rectangular lattice pattern, the replicates were dispersed at random to the remaining features. (The remaining seven features were assigned as additional YQL features.) The 800 replicated genes are listed in Supplementary Table 1 online.

Each TAG was synthesized *in situ* on a hydrophobic glass substrate using a contact-free inkjet printing process (15) (Agilent Technologies). A T_{10} spacer separated the 3′ end of the TAG sequence from the glass substrate. Proprietary negative and positive control sequences were added by the manufacturer at each feature on the edge of the array.

TAG amplification by asymmetric PCR

Existing protocols for TAG PCR (1–4) yield labeled products that are predominantly double-stranded. The two strands must be separated by denaturation before hybridization of the sense strand to the microarray can occur. The universal primer sites are blocked with an excess of complementary oligonucleotides to reduce artifactual hybridization (4,5). To avoid the potential loss of sensitivity caused by reannealing of the two strands, we used asymmetric PCR (16) to generate products that were predominantly single-stranded. Since PCR amplification was observed to saturate after ~26 cycles, we increased the concentration of labeled primer by 10-fold and increased the number of PCR cycles from 30 to 50 to consume most of the labeled primers. Native gel electrophoresis revealed a product that had a faster electrophoretic mobility than the usual double-stranded product and comigrated with a specific single-stranded 56mer (see Supplementary Figure 1 online). Gels using known quantities of synthetic oligonucleotides demonstrated that the efficiency of ethidium bromide staining of a single-stranded 56mer was ~10-fold less than its double-stranded counterpart (data not shown), leading us to estimate a final typical single-stranded product concentration of ~2 μM. Calculations and experiments confirmed that PCR product yields were limited by primer concentrations rather than nucleotide concentrations (see Supplementary Figure 2 online).

Methods for dealing with reagent contamination by TAG PCR products

Contamination of PCRs by TAG sequences was an unexpectedly severe problem. Although it is common knowledge that contamination can confound PCR experiments (17), this issue has not been emphasized in the TAG microarray literature, to our knowledge. The risk of contamination is exacerbated by the high concentrations of PCR product generated during asymmetric TAG PCR (~2 μM), coupled with the small product size (56 bases). These considerations may make it more difficult to remove or degrade DNA adsorbed to surfaces. We estimate the dilution factor necessary for decontamination to be $\sim 10^{12}$.

We identified sources of contamination by constructing a series of negative controls (containing primers but no added template). These controls were validated by confirming that they were able to generate a product of the expected size only after adding trace concentrations of TAG sequences (see Supplementary Figure 1 online). Suggested procedures for decontamination and isolation are outlined in Supplementary Note 1 online.

A dye-flip experiment for microarray diagnostics

Microarrays are fundamentally an assay methodology for detecting specific nucleic acids. A foremost concern of any assay is to determine its sensitivity and specificity. This is a potentially complex problem in the case of microarrays, not only because of the multiplexed nature of the assay but also because of the diverse range of experimental variables that come into play during the assay procedure. As a control for these variables, we performed a 'dye-flip' experiment (18) in which matched samples were hybridized on consecutive days by the same person to the same slides using the same reagents, but with labels that were in reversed order on the second day. The genomic DNA was prepared from the *MATa* haploid YKO strain collection (4), and this was taken as the positive control. The set of essential genes was taken as a negative control, since by definition, these genes are excluded from the haploid strain collection.

UPTAG and DNTAG sequences were amplified from their templates using primers labeled with either Cy3 or Cy5 dyes. The four possible combinations were prepared in separate but concurrent PCRs. Cy3-UPTAGs and Cy5-DNTAGs were hybridized together to duplicate microarrays, omitting the other two combinations so that the Cy3 would be visualized only on the UPTAG (left) half of the array and the Cy5 would be only on the right. After scanning these microarrays, the slides were stripped, hybridized with the two PCR products not already used, and then processed as before. The green and magenta separation seen in two-color ratio images (Figure 1b) provided our first indication that the methodology had high sensitivity and specificity, and this impression was reinforced by the color reversal observed after the second hybridization (Figure 1c).

Characterization of signal intensity and dynamic range

To quantify these findings, we displayed the raw signal intensities for the negative and positive controls as density plots on a logarithmic scale (Figure 2). One of each duplicate slide was analyzed since variability between duplicate slides was much less than that between dye-flip slides. The only microarray features included in the analysis were those that uniquely represented a single gene, as defined by the September 2004 revision of the SGD (http://www.yeastgenome.org/). Duplicate and obsolete entries were excluded, and the YQL TAGs that were substituted as aliases for missing DNTAGs (to maintain symmetry between the two halves of the microarray) were excluded as well.

The overall range of intensities spanned about three orders of magnitude. With the same microarray scanner settings as those customarily used at the time, signal saturation was much more common, despite the fact that the concentration of PCR product (vol/vol) had been reduced 10-fold. This suggested that at least some of our changes to the methodology had been effective.

Characterization of false positives

Our negative controls consisted of either the essential gene YKOs absent from the constructed pool (Figure 2, blue lines) or the 'artificial TAGs' that do not correspond to existing yeast strains (red lines). Both exhibited narrow intensity distributions. Median intensities were about three times background, where background was measured outside each microarray feature by the imaging software (data not shown). However, a long tail of outliers was consistently observed, indicating the presence of FPs. The positive controls (black lines)

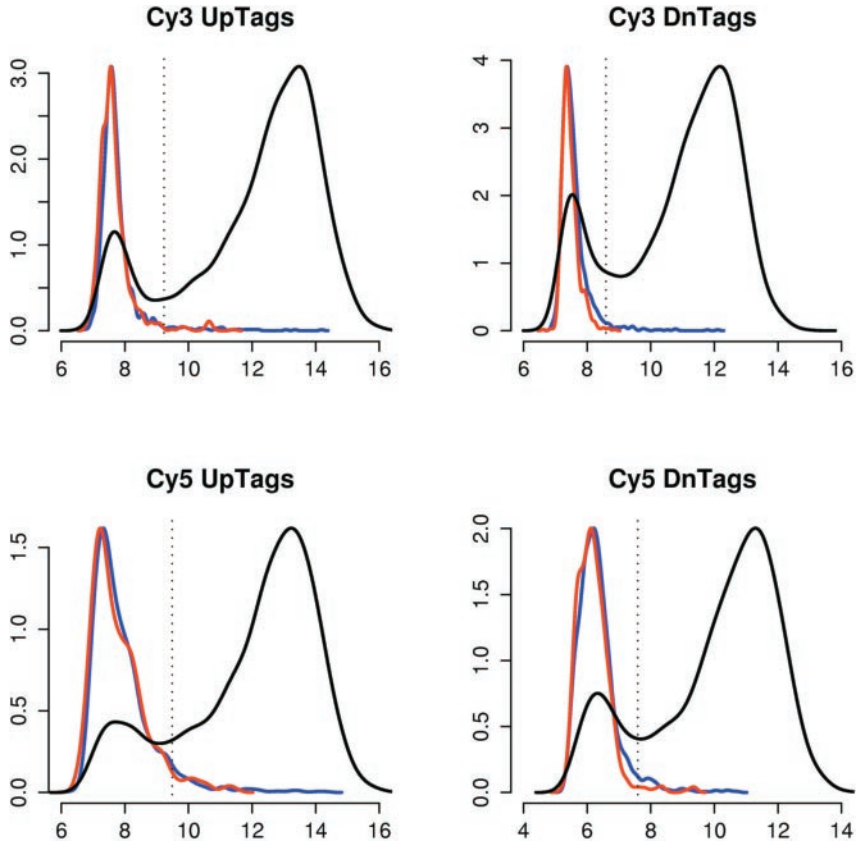

Figure 2. Density plots for positive and negative control features. Each plot depicts one of the four UPTAG/DNTAG, Cy3/Cy5 combinations in the experiment of Figure 1. Relative probability densities are plotted on the *y*-axis against the logarithm (base 2) of raw signal intensities on the *x*-axis. Positive controls (solid lines) were TAGs in haploid YKO strains; negative controls were either TAGs in essential genes (blue lines) or 'YQL' TAGs specially created for this microarray and not associated with any YKO strain (red lines). Intensity thresholds separating the negative and positive controls (dotted lines) were calculated by logistic regression. Density plots were generated with a bandwidth of 0.3 log units.

exhibited a strongly bimodal distribution. The first peak coincided with the negative controls and thus represented FPs, while the second peak was broader and spanned a 60-fold range of intensities beyond the first peak, thereby defining the dynamic range available for signal detection. A dividing line calculated by logistic regression using the negative and positive controls together effectively separated the two peaks. The FP and FN rates for all TAGs ranged between 3 and 6% and 15 and 18%, respectively. Of note, the non-spike-in YQL replicates had consistently fewer and weaker FPs compared to those for the essential genes, consistent with the more stringent procedures used to select the YQL TAG sequences.

To characterize the FPs, we observed the effects of a dye-flip on the FPs. Of 36 Cy3-UPTAGs in slide 1 and 63 Cy5-UPTAGs in slide 2 called as FPs, 32 were shared. For DNTAGs, the counts were 35, 41 and 27. This sharing made it unlikely that contamination of dye-specific PCR primers was responsible for the FPs. Rather, it suggested that most of the FP behavior was due to intrinsic features of those oligonucleotide sequences in combination with this sample.

In addition to microarray features that were deemed FPs because of their unexpectedly high signal intensities, we also found a few features with the wrong colors (Figure 1b and c). After a dye-flip, the colors were still wrong (e.g. Figure 3a compares Cy5-UPTAGs in slide 1 with Cy3-UPTAGs in slide 2). This correlation was missing in comparisons that did not reflect the dye-flip (Figure 3b–d). Because the wrong-color FPs were definitely not synthesized—the particular labeled primer pair needed to make these products was not used—they represent a class of FPs distinct from TAG contamination that is likely due to cross-hybridization.

Characterization of false negatives

Like the FPs, the FNs were highly correlated after a dye-flip. Of 631 Cy3-UPTAG FNs in slide 1 and 744 Cy5-UPTAG FNs in slide 2, 595 were shared. For DNTAGs, the counts were 802, 732 and 697. These results indicated that FNs were also a property of the sequences rather than a hybridization artifact.

We had already learned from sequencing DNTAGs in 63 haploid strains that almost half of these harbored mutations. These included single-base substitutions, single-base deletions, larger deletions (up to seven bases), or some combination of these. The average mutation rate was 1.2% per base. These observations led us to design microarray features that surveyed each of these classes of mutations. The oligonucleotides needed for this mutation survey were included as additional YQL spike-in controls in the dye-flip experiment. Surprisingly, none of the single-base mutations affected signal

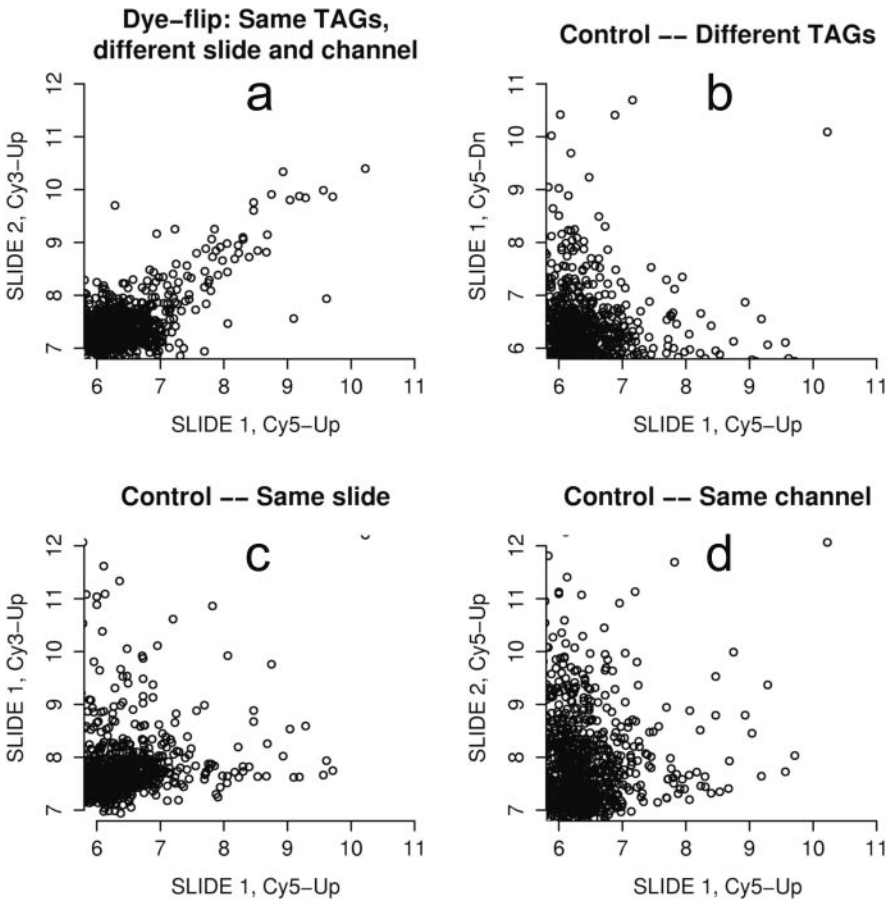

Figure 3. Correlation plots of FPs in the dye-flip experiment. Slide 1 and slide 2 are as defined in Figure 1, i.e. they are dye-flip replicates. TAGs of essential genes, detected as Cy5-labeled UPTAGs in slide 1, are plotted on the *x*-axis against each of four other sets of FPs on the *y*-axis, as shown. The *x*- and *y*-axes denote raw intensity values plotted on a log (base 2) scale.

intensity, and block deletions up to six bases long were well-tolerated (Figure 4). Thus, single point mutations in the TAGs are unlikely to account for a significant portion of FNs under these hybridization conditions.

We thus performed a third hybridization experiment using the pooled heterozygous diploid YKO mutants. The rationale for this choice was the publication of experimentally derived TAG-associated sequences for this collection of mutants (19). Of 98 FNs with the lowest signal intensities, 77 had confirmed sequence data. Examination of these sequences revealed two distinct classes of TAG-associated mutations that together accounted for 60 of the 77 FNs (78%) (see the 'FalseNegatives' subdirectory archived in Supplementary Note 4 online). The largest class (46/60) had mutations within the seven terminal (3′) bases of a universal primer. The second class (14/60) had multiple mutations, either within a TAG or distributed between the TAG and its associated primers. The remaining 17 FNs (22%) were unexplained in that they either had no mutation or had single-base TAG mutations that should not have affected signal intensity (cf. Figure 4). Thus, TAG amplification problems were the leading cause of FNs in this experiment, but mutations did not fully account for all FNs. These observations follow those of Eason *et al*. (19), although 37 of the above 98 FNs are missing from the FN lists in Eason *et al*. Differences in normalization procedures appear to account for most of these discrepancies (see the 'Eason' and 'Giaever' (20) subdirectories archived in Supplementary Note 4 online).

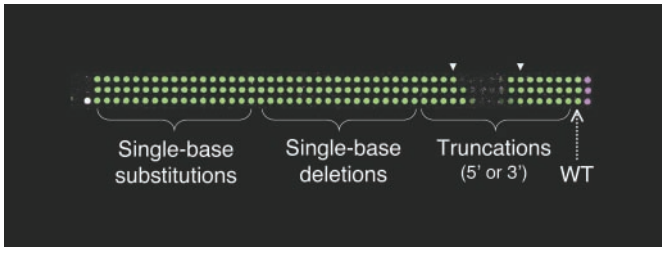

Figure 4. Tolerance of microarray hybridization to single-base TAG mutations. 'Candle' hybridization controls from slide 2 of the dye-flip experiment are shown as false-color ratio images. Each row of features represents a mutation series for a different YQL TAG, amplified here by DNTAG universal primers. The two arrowheads signify deletion of the last six bases and the first seven bases, respectively. The white feature is from the edge of the array; magenta features are part of the set of hybridization controls amplified by UPTAG universal primers. See legend of Figure 1 for details of image generation.

Validation of 'spike-in' positive controls

The YQL replicate features were also designed to serve as 'spike-in' positive controls, simply by supplementing the PCR templates with appropriate concentrations of the requisite oligonucleotides. In this way, they might be used to calibrate

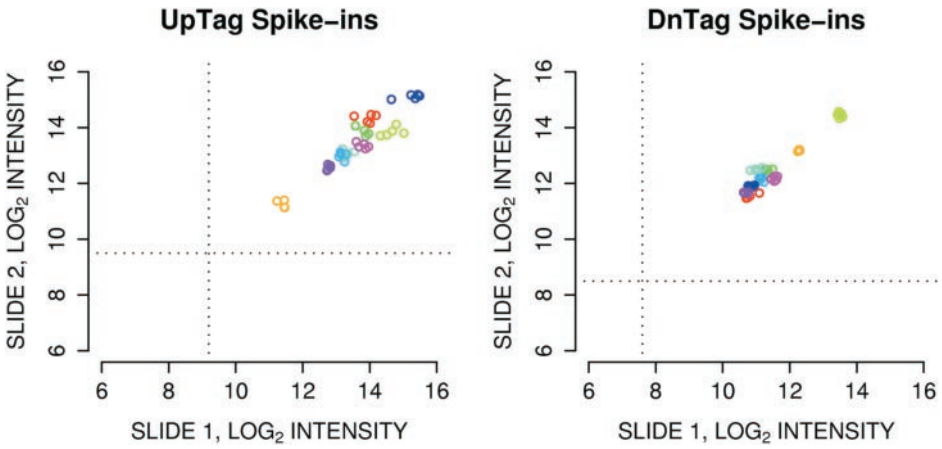

Figure 5. Detection of spike-in positive controls. YQL TAGs encoded on appropriately diluted spike-in oligonucleotides were amplified along with TAGs from genomic DNA in the same dye-flip experiment as described in Figure 1. Raw signal intensities from slides 1 and 2 are depicted as a scatterplot on the same log (base 2) scales as the corresponding axes in Figure 2, along with the same detection thresholds (dotted lines). Each TAG is represented by five replicate measurements plotted in the same color; nine colors are shown.

hybridization reactions. The fact that genomic DNA and TAG oligonucleotides are amplified equally efficiently (i.e. one million-fold amplification in 20 ± 1 cycles) was established in preliminary experiments (see Supplementary Figures 3 and 4 online). As a feasibility test, 18 YQL oligonucleotides (nine for each TAG type) were synthesized without purification. Universal primer sequences were included for amplifying the YQL TAGs concurrently with the TAGs in the genomic DNA. After combining nominally equal concentrations of the stock solutions, the resulting cocktail was precisely diluted in four stages to match the molar concentration of yeast genomes in a PCR ($\sim$ 1 fM) after addition to the reaction. Hybridization to the same microarrays used in the dye-flip experiment demonstrated that all nine TAGs were easily detected (Figure 5). Reproducibility was excellent, both within each set of replicates and across the two dye-flip slides, suggesting that use of the YQL TAGs as calibration standards is indeed feasible.

DISCUSSION

We have described a TAG microarray with elaborate controls and have characterized its performance in terms of FP and FN rates. We showed that most of the FPs can be attributed to cross-hybridization, at least in the experiments described here. We also found that about half of the FNs can be understood in terms of TAG-associated mutations.

Of all the causes of FPs, reagent contamination by traces of amplified TAGs is the most serious. Contamination reveals itself as FPs (or 'hits' in a genetic screen) that later fail to be validated. Results may masquerade as findings that are 'reproducible' for many weeks if the contamination spreads to stock solutions or the general laboratory environment. Detection is best carried out by sham (template-free) PCR controls, provided reagents can be prepared and handled without contamination. An alternative is to carry out surveillance hybridizations with sham PCR products prepared without genomic DNA, or to perform dye-flip experiments as described here, although debugging by such methods is expensive. Fortunately, there was little evidence of contamination by TAGs in the results presented here. A different kind of contamination, e.g. the presence of diploid strains in the haploid strain collection [see Supplementary Figure 3 in (4)], could also potentially account for some of the FPs observed here, but do not explain the 'wrong-color' FPs (Figure 3).

Cross-hybridization is a concern shared by all hybridization technologies. Considering the difficulty of designing a large set of oligonucleotide sequences with minimal mutual cross-hybridization, the 3–6% FP rate we found is fairly low and probably more than adequate for many screening applications, especially when confirmatory screening is available. However, for whole-genome profiling applications, the 300 FPs identified with this FP rate will confound interpretation. Even though most FPs are relatively weak, a few are quite strong (Figure 2), and conventional statistical criteria (such as the '2 SD' cutoff for P-values < 0.05) are questionable in this setting. The best approach for reducing the FP rate may be to optimize hybridization and washing parameters, but this would have to be achieved without inflating the FN rate. Identifying FPs in advance is an alternative solution, but it is likely that many of the FPs associated with any given subset of hybridized TAGs will overlap with signals from other subsets of TAGs, shielding them from identification. To address this problem, we have partitioned the heterozygous diploid strain collection into a series of subpools. Hybridization of TAGs derived from each subpool to separate microarrays should yield more definitive information on which microarray features are most prone to cross-hybridization. Until then, FPs are best interpreted probabilistically and should be considered in any inferences made with these microarrays.

We made a concerted effort to lower the FN rate through procedures aimed at improving signal strength, e.g. asymmetric PCR and precautions against oxidation. It was not practical to perform the extensive range of control experiments that would be needed to define which if any of these modifications were in fact effective. Nevertheless, our experience has been that with the new protocols, the signal intensities are stronger and more reliable. Even so, TAG-associated mutations impose a lower bound on the FN rate. Assuming that these

mutations will not be remediated in the near future, the best approach may be to leverage the most trustworthy information from both UPTAGs and DNTAGs into a single statistic. We and others have compared different ways to accomplish this and have devised empirical procedures that maximize predictive value (B.D. Peyser, R.A. Irizarry, C. Tiffany, O. Chen, D.S. Yuan, J.D. Boeke and F.A. Spencer, manuscript submitted). Assuming that UPTAG and DNTAG data are statistically independent, overall FN rates will be 2.2–3.2%, corresponding to a lack of informative data for 140–200 genes. This estimate is consistent with that obtained by Eason *et al*. (19).

Knowledge of the FPs and FNs of a microarray is essential for informed data interpretation. However, TAG microarrays will typically be used in two-color experiments in which the labeled samples are derived from two comparable pools of yeast strains. The focus of such experiments is the log ratio of the two signal intensities, and knowledge of the distribution of these log ratios is key to identifying the log ratios that are statistically significantly different. We have learned that although FP and FN errors contribute to this distribution, random variability between the two pools is probably more important. This variability depends strongly on how the pools were sampled for measurement as well as on various sources of noise within the signal intensities themselves. This information can only be obtained from dedicated control experiments that closely reproduce typical experimental conditions. For comparisons between a control pool and an experimental pool with just a few missing strains, pilot studies that compare a control pool with a 'drop-out' experimental pool (the complement of a 'spike-in' pool) may be ideal for this purpose (B.D. Peyser, R.A. Irizarry, C. Tiffany, O. Chen, D.S. Yuan, J.D. Boeke and F.A. Spencer, manuscript submitted).

We add for completeness that the 5-fold replicate features designed into our microarray have a novel application beyond their use as negative or positive (YQL) controls (cf. Figures 2 and 5). The novelty lies not so much in the paradigm of calculating standard errors from '$n = 5$' replicates, as in the more powerful idea of estimating and correcting systematic errors that take the form of irregular biases over the surface of the microarray. Such biases have many potential causes, ranging from manufacturing defects to fingerprints to temperature gradients. Because our replicates are intimately co-mingled with the systematic set of TAGs and yet are in random order, they are well-suited to serve as probes of these biases. We have recently developed software to estimate these biases from replicate data and have found that they account almost entirely for the spatially correlated errors in these microarrays. A statistical analysis of these errors will be presented elsewhere.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank C. Tiffany, C. Hollender and S. Sookhai-Mahadeo for preparing pools of the YKO strain collection. Supported by fellowships from the Burroughs-Wellcome Center for Computational Biology at Johns Hopkins (D.S.Y.), NHGRI (D.S.Y.), the Leukemia and Lymphoma Society (X.P.), and by grant HG02432 from the NIH (J.D.B.). Funding to pay the Open Access publication charges for this article was provided by NHGRI.

Conflict of interest statement. None declared.

REFERENCES

1. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al*. (1999) Functional characterization of the *S.cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
2. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Veronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., Andre,B. *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
3. Shoemaker,D.D., Lashkari,D.A., Morris,D., Mittmann,M. and Davis,R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet*., **14**, 450–456.
4. Ooi,S.L., Shoemaker,D.D. and Boeke,J.D. (2001) A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science*, **294**, 2552–2556.
5. Hanway,D., Chin,J.K., Xia,G., Oshiro,G., Winzeler,E.A. and Romesberg,F.E. (2002) Previously uncharacterized genes in the UV- and MMS-induced DNA damage response in yeast. *Proc. Natl Acad. Sci. USA*, **99**, 10605–10610.
6. Ooi,S.L., Shoemaker,D. and Boeke,J.D. (2003) DNA helicase interaction network defined using synthetic lethality analyzed by microarray. *Nature Genet*., **35**, 277–286.
7. Birrell,G.W., Giaever,G., Chu,A.M., Davis,R.W. and Brown,J.M. (2001) A genome-wide screen in *Saccharomyces cerevisiae* for genes affecting UV radiation sensitivity. *Proc. Natl Acad. Sci. USA*, **98**, 12608–12613.
8. Lum,P.Y., Armour,C.D., Stepaniants,S.B., Cavet,G., Wolf,M.K., Butler,J.S., Hinshaw,J.C., Garnier,P., Prestwich,G.D., Leonardson,A. *et al*. (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, **116**, 121–137.
9. Pan,X., Yuan,D.S., Xiang,D., Wang,X., Sookhai-Mahadeo,S., Bader,J.S., Hieter,P., Spencer,F. and Boeke,J.D. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol. Cell*, **16**, 487–496.
10. Ausubel,F.M., Brent,R., Kingston,R., Moore,D.D., Seidman,J.G., Smith,J.A. and Struhl,K. (1995) *Short Protocols in Molecular Biology*. 3rd edn. Wiley, NY, pp. 13–46.
11. Fare,T.L., Coffey,E.M., Dai,H., He,Y.D., Kessler,D.A., Kilian,K.A., Koch,J.E., LeProust,E., Marton,M.J., Meyer,M.R. *et al*. (2003) Effects of atmospheric ozone on microarray data quality. *Anal. Chem*., **75**, 4672–4675.
12. Brody,J.R. and Kern,S.E. (2004) Sodium boric acid: a Tris-free, cooler conductive medium for DNA electrophoresis. *Biotechniques*, **36**, 214–216.
13. Dudoit,S. and Yang,Y.H. (2002) Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In Parmigiani,G., Garrett,E.S., Irizarry,R.A. and Zeger,S.L. (eds), *The Analysis of Gene Expression Data: Methods and Software*. Springer, NY, pp. 73–101.
14. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
15. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol*., **19**, 342–347.
16. Gyllensten,U.B. and Erlich,H.A. (1988) Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus. *Proc. Natl Acad. Sci. USA*, **85**, 7652–7656.

17. Kwok,S. and Higuchi,R. (1989) Avoiding false positives with PCR. *Nature*, **339**, 237–238.
18. Nadon,R. and Shoemaker,J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.*, **18**, 265–271.
19. Eason,R.G., Pourmand,N., Tongprasit,W., Herman,Z.S., Anthony,K., Jejelowo,O., Davis,R.W. and Stolc,V. (2004) Characterization of synthetic DNA bar codes in *Saccharomyces cerevisiae* gene-deletion strains. *Proc. Natl Acad. Sci. USA*, **101**, 11046–11051.
20. Giaever,G., Flaherty,P., Kumm,J., Proctor,M., Nislow,C., Jaramillo,D.F., Chu,A.M., Jordan,M.I., Arkin,A.P. and Davis,R.W. (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc. Natl Acad. Sci. USA*, **101**, 793–798.