

Improved statistical analysis of budding yeast TAG microarrays revealed by defined spike-in pools

Brian D. Peyser^{1,3}, Rafael A. Irizarry², Carol W. Tiffany¹, Ou Chen¹, Daniel S. Yuan³,
Jef D. Boeke³ and Forrest A. Spencer^{1,3,*}

¹McKusick-Nathans Institute of Genetic Medicine, School of Medicine, ²Department of Biostatistics, Bloomberg School of Public Health and ³Department of Molecular Biology and Genetics, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Received April 11, 2005; Revised June 24, 2005; Accepted August 26, 2005

ABSTRACT

***Saccharomyces cerevisiae* knockout collection TAG microarrays are an emergent platform for rapid, genome-wide functional characterization of yeast genes. TAG arrays report abundance of unique oligonucleotide 'TAG' sequences incorporated into each deletion mutation of the yeast knockout collection, allowing measurement of relative strain representation across experimental conditions for all knockout mutants simultaneously. One application of TAG arrays is to perform genome-wide synthetic lethality screens, known as synthetic lethality analyzed by microarray (SLAM). We designed a fully defined spike-in pool to resemble typical SLAM experiments and performed TAG microarray hybridizations. We describe a method for analyzing two-color array data to efficiently measure the differential knockout strain representation across two experimental conditions, and use the spike-in pool to show that the sensitivity and specificity of this method exceed typical current approaches.**

INTRODUCTION

Introduction of the yeast knockout collections, containing arrayed strains harboring deletion mutations for >95% of predicted open reading frames (ORF), allows systematic genome-wide screens for various phenotypes to be readily accomplished (1–3). One aspect of the knockout collections that facilitates rapid screens is the pair of unique 20mer TAGs within each deletion mutation (4). A gene in each yeast knockout strain (YKO) is replaced with a selectable marker flanked by two TAGs, termed UPTAG and DNTAG. All UPTAGs or

all DNTAGs in a sample can be amplified in a PCR using universal primers. Individual YKO representation is subsequently interrogated by hybridizing labeled TAGs to microarrays and observing changes in signal intensity between experimental conditions. This approach has been applied to genome-wide screens for mutant phenotypes (3,5), synthetic genetic interactions (6–10) and synthetic-chemical-genetic interactions (10). TAG microarray approaches are rapid and comprehensive, but systematic optimization of analysis methods is lacking. We address this need here.

The most common application of TAG arrays is comparison of YKO representation in two samples. Typically, samples are co-hybridized on one array using complementary fluorescent labels (Cy5 and Cy3). Various general approaches for two-color arrays have been proposed for quality assessment, background adjustment (11,12) and normalization (13–15). However, in analysis of TAG arrays, each YKO has UPTAG and DNTAG probes; four measurements corresponding to each YKO are obtained. Finding the best way to summarize this information in one quantity reflecting differential YKO representation is not trivial (16). Here we demonstrate the utility of a spike-in experiment by evaluating a simple quality assessment procedure and a novel strategy for combining the UPTAG and DNTAG information in a way that is robust to problematic TAGs. Our results show that implementing these two data procedures can greatly improve the utility of TAG arrays.

MATERIALS AND METHODS

Preparation of spike-in pools. Heterozygous YKOs (Research Genetics) were grown on solid media and combined, with the exception of YKOs from plate 259 which were separately mixed in subpools, as shown in Figure 1, before incorporation into pool A or B at appropriate representation levels. Genomic

*To whom correspondence should be addressed. Tel: +1 011 410 614 2536; Fax: +1 011 410 614 8600; Email: fspencer@jhmi.edu
Correspondence may also be addressed to Rafael A. Irizarry, Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, 615 N Wolfe Street E3620, Baltimore, MD 21205, USA. Email: rafa@jhu.edu

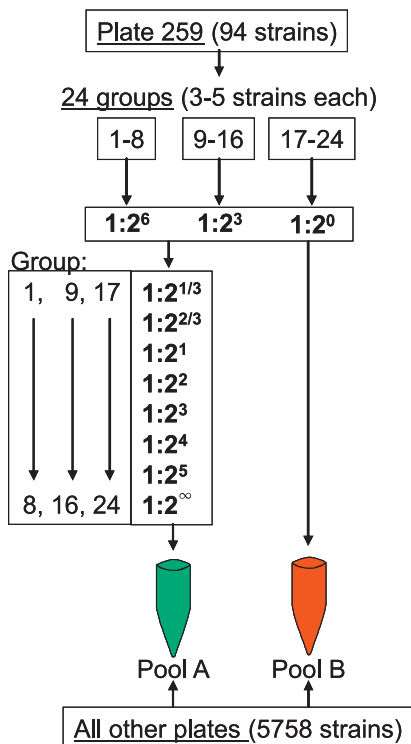


Figure 1. Design of spike-in pools. Two pools were created ('A' and 'B') such that 94 strains were differentially represented between the two pools. The 94 differentially represented strains were diluted 1:1, 1:8 or 1:64 (High, Medium or Low representation), then added to Pool B. Each strain was then diluted again from 1:2^{1/3} to 1:2⁵ and added to Pool A. One set of strains from each representation group was not added to Pool A (dilution 1:2[∞]).

DNA was extracted from samples of each spike-in pool using the Masterpure Yeast DNA kit (Epicentre). Pool A and B TAGs were labeled with Cy3 and Cy5, respectively, and TAG microarrays were hybridized, washed and scanned as described (17).

Analyses were performed using custom scripts written in R, an open-source statistical language (18). GenePix local background intensities were not used for correction because, as suggested by Yang *et al.* (12), subtracting these severely increases noise (data not shown). Normalization was performed using a procedure similar to the one previously proposed (14). Alternate normalization methods did not impact results (data not shown).

The GEO accession number for microarray data is GSE2832. Data and code necessary to reproduce all the results and figures are available upon request.

RESULTS

To evaluate statistical procedures for TAG microarray data, we tailored defined spike-in pools to resemble expected results in a typical synthetic lethality analyzed by microarray (SLAM) experiment (6). Synthetic lethality is defined as inviability of cells containing two mutations which are individually not required for growth. In a SLAM experiment, viable YKOs in pooled form are compared under two conditions: absence versus presence of a specific second mutation (the 'query').

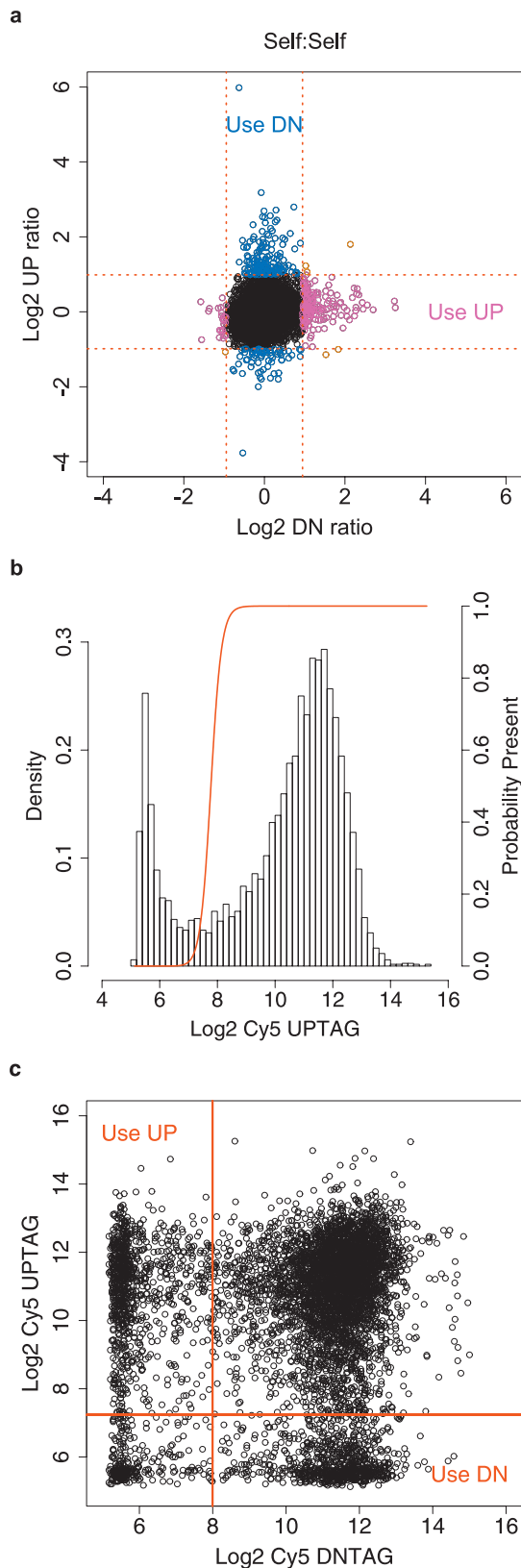
The average number of genetic interactions expected in a genome-wide screen has been estimated to be ~35, although several query mutations with interactions exceeding 100 have been analyzed (10,19). Therefore, we designed a pair of pools ('A' and 'B') with 5758 YKOs at equivalent representation, and a set of 94 YKOs with known differential representation ranging from 1:2^{1/3} to 1:2⁵ and 1:infinity (Figure 1 and Supplementary Table 1). Additionally, certain YKOs grow slowly, and representation of these in the control SLAM sample is expected to be lower than YKOs with wild-type growth rate. To examine our ability to address these mutants, we designed three representation levels in the control (B) pool: high (about equal to all other strains), medium (8-fold dilute) and low (64-fold dilute). TAGs from pools A and B were amplified with Cy3- and Cy5-labeled primers, respectively. These samples were mixed at equal ratio, such that most TAGs should exhibit equal hybridization, while Cy5: Cy3 ratios that deviate from one are expected for the few differentially represented TAGs. This design allows discovery of the best method to produce a measure of differential representation from hybridization results.

Before addressing differential representation, we document the utility of two filtering steps in data pre-processing. First, we noted TAG-specific hybridization artifacts evident in self-hybridizations performed to examine the noise distribution. Pool A DNA served as template for preparation of labeled TAGs with both Cy5 and Cy3. Thus, all TAGs were present at equal amounts between channels. Figure 2a shows a scatter-plot of normalized log₂ Cy5: Cy3 ratios for corresponding UPTAG and DNTAG probes. Because all these values should be zero plus measurement error, we expect these to be uncorrelated and centered at zero. Figure 2a confirms this except for a few YKOs with extreme values for one TAG. These outliers may have a negative impact on specificity. They are likely to be due to individual tag templates that enter the labeling PCR as contaminants, which are detectable even at very low levels (17).

We determined that these artifacts are consistent across experiments performed with a single batch of labeled primer, but not between different primer batches (Supplementary Figure 1). To create a useful filter, we assumed that the data follow a bivariate normal distribution and defined outliers as TAGs with log ratios three SDs away from zero, using a robust estimate of the SD. If the log ratio data follow a normal distribution, excluding outliers, we expect to inappropriately remove only ~0.5% of the data (32 TAGs). We applied this filter (Figure 2a, red lines) independently to UP- and DNTAGs. Fortunately, the YKOs were designed with two TAGs per gene (except for 192 strains lacking DNTAGs), greatly improving chances that at least one TAG performs adequately. Because non-outlier UP/DNTAGs appear to provide independent measurements, the chance of inappropriately removing both TAGs for the same YKO is less than 0.00001. Using this procedure we defined 193 DNTAGs (purple circles) and 244 UPTAGs (blue circles) as primer-batch specific outliers. Six YKOs had both UP and DN ratios filtered (orange circles).

Next, we considered the effect of TAG-specific hybridization behavior resulting from the presence of nucleotide mutations found in some of the TAGs and universal priming sites (20). This is important because sensitivity will be markedly

affected when TAGs fail to provide a meaningful measure of strain representation. Histograms of \log_2 signal intensity display a bimodal distribution (Figure 2b and data not shown) for UP- and DNTAGs whether Cy3 or Cy5 labeling is used. The



lower peak is close to background intensities and contains non-functional TAGs with absent or inefficient hybridization. While TAG sequence discrepancies have been characterized (20), knowledge of the presence and nature of mutations was insufficient to fully predict hybridization behavior (17,20).

The naïve approach to summarizing UP- and DNTAG information is to average their observed log ratios. This solution will yield suboptimal measures when one of the TAGs is non-functional. We propose a procedure exploiting the bimodal distribution of TAG intensities to improve on simple averaging. To determine if a TAG is non-functional we fit a mixture model, as in Irizarry *et al.* (16), to the log intensity data for the control sample. The model fits two normal distributions to the Cy5 data, one for the lower mode and one for the upper mode. The 'blank' (YQL) features (17) define the location and width (mean and SD) of the lower distribution. With this fitted model in place, we can predict the probability that each TAG is 'present' (Figure 2b). We consider a DN/UPTAG non-functional when it is predicted absent while the complementary UP/DNTAG is present. We define a weighted average = $w * UP + (1 - w) * DN$, where $w = 0.5 + [P(UP \text{ present}) - P(DN \text{ present})]/2$. Thus, when UP is present ($P_{UP} = 1$) and DN is absent ($P_{DN} = 0$), $w = 1$ and only UPTAG is used (Figure 2c). We describe a less complex procedure in Supplementary Note 1 that uses binary absent or present values ($P = 0$ or 1) and performs similarly (data not shown). Researchers using unsophisticated analysis software such as spreadsheet applications may prefer the simple procedure.

We compared the performance of these two strategies and use of UP- or DNTAGs alone with Receiver Operating Characteristic (ROC) curves based on the spike-in experiment. For this analysis, nominal ratios below 2-fold were excluded from the list of True Positives. This choice is appropriate because 2-fold representation difference corresponds to a subtle growth defect at the margin of detection in colony measurement (1.25-fold colony diameter difference is predicted by hemispherical colony volume = $2 \pi r^3/3$). Supplementary Figures 2–3 present ROC curves with varying stringencies for inclusion as 'True Positive', including every spiked-in YKO (1.26-fold or higher). ROC curves in the range of false positives likely to be acceptable (Figure 3a and Supplementary Figure 2) demonstrate that the artifact filtering process has a significant effect on specificity (Supplementary Figure 4 shows the full ROC curves). Additional filtering of non-functional TAGs by the weighted average improves results further (Figure 3b and Supplementary Figure 3).

Figure 2. Development of TAG filters. (a) Self-self hybridizations. Pool A gDNA was used as template for TAG labeling reactions with each primer set (UP/DNTAG Cy5/Cy3). Median values across three experiments were displayed. Each point represents a single YKO. Red dotted lines are three SDs. Blue circles, artifacts specific to UP ratio; purple circles, artifacts in both TAGs; orange circles, artifacts in both TAGs. (b) Histogram of \log_2 UPTAG Cy5 signal values from pool A versus B hybridization. Results from DNTAG and Cy3 are similar. Red line, probability each feature belongs to the upper (right-hand) distribution. (c) UPTAG Cy5 versus DNTAG Cy5. Each point represents a single YKO. Therefore, for each point the numbers of UP- and DNTAGs in the sample are identical. Red lines, values at which absent/present probabilities equal 0.5. In the weighted average method, DN ratio was weighted higher for YKOs in the upper left quadrant, and UP ratio was weighted higher for YKOs in the lower right quadrant.

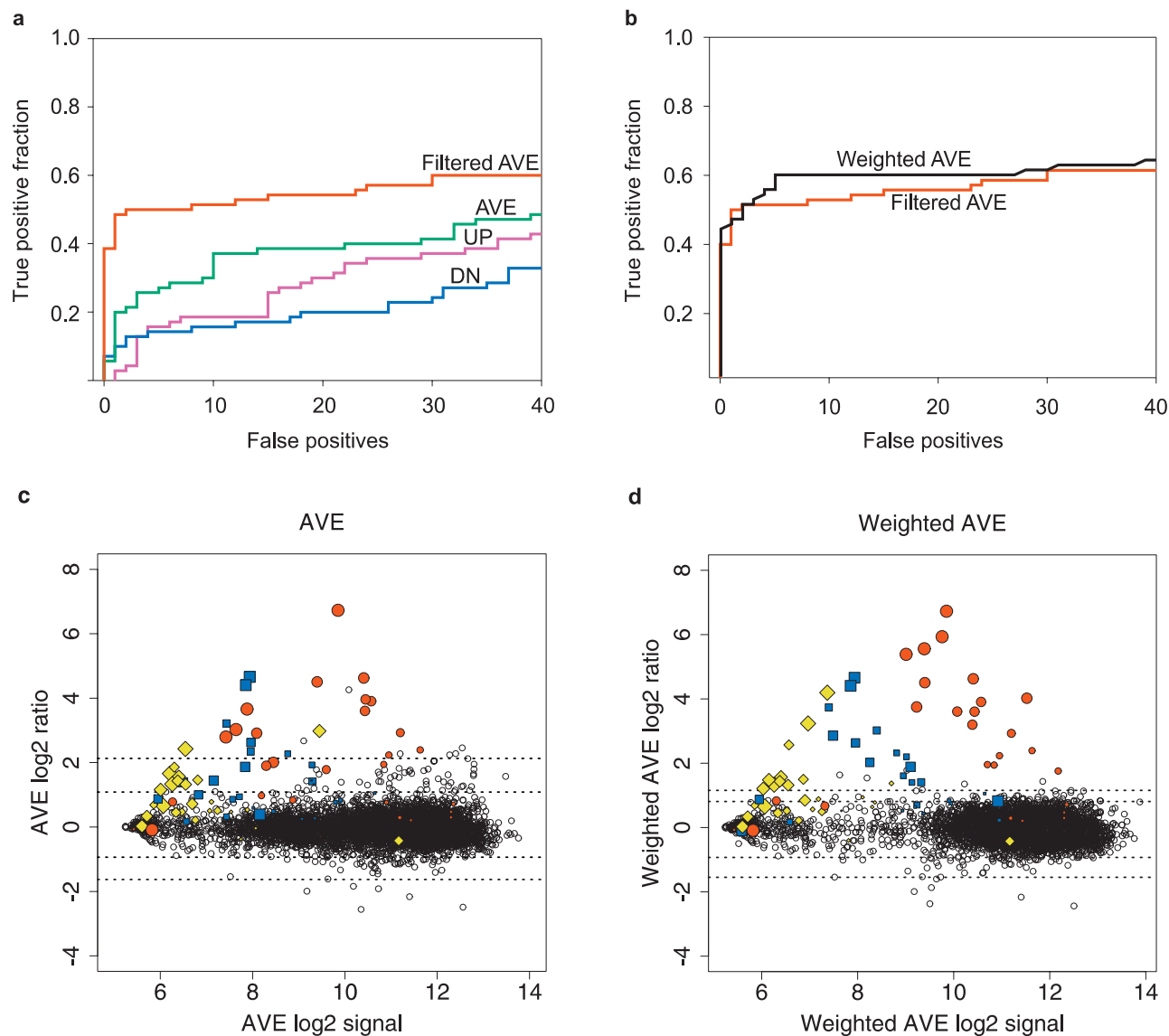


Figure 3. Measures of sensitivity/specificity. (a and b) ROC for several methods of calculating relative YKO representation. True Positive is defined as any YKO with known pool B:A ratio ≥ 2 . False Positive is defined as any YKO with known B:A ratio of 1. ROC curves for UP ratio alone (purple line), DN ratio (blue line), Average ratio (green line), Filtered average ratio (red line) and weighted average ratio (black line). (c and d) Ratio-intensity plots using simple averaging of UP and DN ratios or weighted average ratios. Black open circles, YKOs with known B:A ratio of 1; filled circles, high representation YKOs; squares, medium representation YKOs; diamonds, low representation YKOs. Point size is related to known B:A ratio: from small to large, $2^{1/4}$, $2^{1/2}$, 2 , 2^2 , 2^3 , 2^4 , 2^5 and 2^{∞} . Black dotted lines are 0.1, 1, 99 and 99.9 percentiles for YKOs with known B:A ratio = 1.

The effect of two filters, one which removes the systematic artifacts and a second which removes non-hybridizing TAGs, is demonstrated by ratio-intensity plots. A naive approach to analysis would average UP and DN log ratio to produce a measure M for relative strain representation. By filtering systematic artifacts, noise is significantly reduced (Figure 3c and d, open circles and Supplementary Figure 5). Additionally, combining UP and DN selectively provides increased sensitivity for a number of spiked-in strains (filled shapes).

DISCUSSION

In summary, we present a spike-in pool design that allows evaluation of various methods for generating measures of

differential strain representation. Using this experiment, we determined that the largest factor affecting specificity is the presence of primer batch-specific artifacts, evident in control self-self hybridizations. These artifacts may result from extremely low levels of contaminating TAG sequence template introduced before the labeling PCR. Accidental introduction of contaminants may occur at multiple steps, including the high-performance liquid chromatography (HPLC) column purification of Cy5- and Cy3-labeled primers at their manufacture as well as laboratory manipulation of primer batches during initial stock and aliquot preparation. Because the artifacts are consistent only within batches of primer sets, contamination must occur at initial preparation or during manufacture. Yuan *et al.* (17) discuss the unusually large dilutions required to prevent contamination in TAG labeling

reactions. While the source of these artifacts is uncertain, there are several options for minimizing their effect. The approach we present uses a control hybridization of one DNA sample labeled with both primer sets, such that every TAG is present at equal amounts in the two labeling reactions. Deviations from expected 1:1 ratio can be recognized and filtering is applied.

The methods we describe improve detection of true signal difference between samples; however, they are not perfect. Once primer-batch specific artifacts are removed, noise is increased slightly with the weighted average method compared to averaging (see Supplementary Figure 5d and e). Additionally, the weighted average could cause decreased sensitivity when cross-hybridization occurs for one TAG from a low represented strain. The TAG that accurately reflects the low representation of the YKO may be discounted while the cross-hybridizing TAG is emphasized. These problems could be minimized by improving the criteria for selecting a TAG as non-hybridizing, perhaps by examining behavior across many microarrays. The advantage of this method is that it requires as few as two microarray hybridizations (self-self and experiment) to perform well. We have tested these methods to provide optimal results from SLAM experiments, where YKOs that decrease in representation from control to experimental samples are sought. However, appropriately applied, other TAG microarray experiments should benefit from the procedures we describe.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Z. Wu, X. Pan, J. Bader and colleagues from our laboratories for stimulating discussions. This work was supported by NIH HG02432 to J.B., R.I. and F.S. which also funded the Open Access publication charges for this article.

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., Chu, A.M. *et al.* (1999) Functional characterization of the *S.cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
2. Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.
3. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
4. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittman, M. and Davis, R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.*, **14**, 450–456.
5. Ooi, S.-L., Shoemaker, D. and Boeke, J.D. (2001) A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science*, **294**, 2552–2556.
6. Ooi, S.-L., Shoemaker, D. and Boeke, J.D. (2003) DNA helicase interaction network defined using synthetic lethality analyzed by microarray. *Nat. Genet.*, **35**, 277–286.
7. Warren, C.D., Eckley, D.M., Lee, M.S., Hanna, J.S., Hughes, A., Peysner, B., Jie, C., Irizarry, R. and Spencer, F.A. (2004) S-phase checkpoint genes safeguard high-fidelity sister chromatid cohesion. *Mol. Biol. Cell*, **15**, 1724–1735.
8. Lee, M.S. and Spencer, F.A. (2004) Bipolar orientation of chromosomes in *Saccharomyces cerevisiae* is monitored by Mad1 and Mad2, but not by Mad3. *Proc. Natl Acad. Sci. USA*, **101**, 10655–10660.
9. Arevalo-Rodriguez, M., Pan, X., Boeke, J.D. and Heitman, J. (2004) FKBP12 controls aspartate pathway flux in *Saccharomyces cerevisiae* to prevent toxic intermediate accumulation. *Eukaryot. Cell*, **3**, 1287–1296.
10. 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.
11. Kooperberg, C., Fazzio, T.G., Delrow, J.J. and Tsukiyama, T. (2002) Improved background correction for spotted DNA microarrays. *J. Comput. Biol.*, **9**, 55–66.
12. Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P. (2002) Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Stat.*, **11**, 108–136.
13. Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
14. Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.
15. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–104.
16. Irizarry, R., Ooi, S.-L., Wu, Z. and Boeke, J.D. (2003) Use of mixture models in a microarray-based screening procedure for detecting differentially represented yeast mutants. *Stat. Appl. Genet. Mol. Biol.*, **2**, 1–17.
17. Yuan, D.S., Pan, X., Ooi, S.L., Peysner, B.D., Spencer, F.A., Irizarry, R.A. and Boeke, J.B. (2005) Improved microarray methods for profiling the Yeast Knockout strain collection. *Nucleic Acids Res.*, **33**, e103.
18. Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
19. Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.
20. 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.