

# **Accurate detection of aneuploidies in array CGH and gene expression microarray data**

Chad L. Myers<sup>1,3</sup>, Maitreya J. Dunham<sup>1</sup>, S.Y. Kung<sup>2</sup>, Olga G. Troyanskaya<sup>1,3\*</sup>

<sup>1</sup>Princeton University, Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory

<sup>2</sup>Princeton University, Department of Electrical Engineering, <sup>3</sup>Princeton University, Department of Computer Science Princeton, NJ 08544

\*To whom correspondence should be addressed.

## ABSTRACT

**Motivation:** Chromosomal copy number changes (aneuploidies) are common in cell populations that undergo multiple cell divisions including yeast strains, cell lines, and tumor cells. Identification of aneuploidies is critical in evolutionary studies, where changes in copy number serve an adaptive purpose, as well as in cancer studies, where amplifications and deletions of chromosomal regions have been identified as a major pathogenetic mechanism. Aneuploidies can be studied on whole-genome level using array CGH (a microarray-based method that measures DNA content), but their presence also affects gene expression. In gene expression microarray analysis, identification of copy number changes is especially important in preventing aberrant biological conclusions based on spurious gene expression correlation or masked phenotypes that arise due to aneuploidies. Previously suggested approaches for aneuploidy detection from microarray data mostly focus on array CGH, address only whole-chromosome or whole-arm copy number changes, and rely on thresholds or other heuristics, making them unsuitable for fully automated general application to gene expression data sets. There is a need for a general and robust method for identification of aneuploidies of any size from both array CGH and gene expression microarray data.

**Results:** We present ChARM (Chromosomal Aberration Region Miner), a robust and accurate expectation-maximization based method for identification of segmental aneuploidies (partial chromosome changes) from gene expression and array CGH microarray data. Systematic evaluation of the algorithm on synthetic and biological data shows that the method is robust to noise, aneuploidal segment size, and p-value cutoff. Using our approach, we identify known chromosomal changes and predict novel potential segmental aneuploidies in commonly used yeast deletion strains and in breast cancer. ChARM can be routinely used to identify aneuploidies in array CGH data sets and to screen gene expression data for aneuploidies or array biases. Our methodology is sensitive enough to detect statistically significant and biologically relevant aneuploidies even when expression or DNA content changes are subtle as in mixed populations of cells.

**Availability:** Matlab code available by request from the authors and on web supplement at <http://function.cs.princeton.edu/ChARM/>.

**Contact:** ogt@cs.princeton.edu

## INTRODUCTION

Chromosomal amplifications, deletions, and rearrangements are thought to play important evolutionary roles in speciation (Fischer *et al.*, 2000) and adaptive mutation in yeast and microbial populations (Hendrickson *et al.*, 2002; Dunham *et al.*, 2002), and constitute a key mechanism in cancer progression (Cahill *et al.*, 1999; Phillips *et al.*, 2001). Aneuploidies are especially common in cell populations that undergo multiple cell divisions such as laboratory strains or cell lines, and presence of amplifications or deletions of whole chromosomes or their parts (segmental aneuploidies) can have substantial effects on gene expression (Fritz *et al.*, 2002; Haddad *et al.*, 2002; Hughes, Roberts *et al.*, 2000). Thus, identification of aneuploidies is important in cancer pathogenesis and molecular evolution studies, as well as in every genome-scale gene expression microarray experiment because copy number changes can alter expression profiles and result in spurious correlations of functionally unrelated genes.

Recent developments in microarray technology have enabled genome-wide investigations of copy-number changes through array-based comparative genomic hybridization (array CGH), where differentially labeled sample and reference DNA are hybridized to DNA microarrays (Pinkel *et al.*, 1998; Pollack *et al.*, 1999). This technology has proven effective in identifying aneuploidies in tumor cells (Gray and Collins, 2000; Phillips *et al.*, 2001; Wilhelm *et al.*, 2002; Linn *et al.*, 2003), experimental evolution studies (Dunham *et al.*, 2002), and in yeast strains (Hughes *et al.*, 2000; Pérez-Ortín *et al.*, 2002). Routine application of array CGH to every strain or tissue used in gene expression studies is unfortunately not feasible. However, several studies have demonstrated that chromosomal abnormalities correlate with spatial biases in gene expression along chromosomes (Pollack *et al.*, 2002; Fritz *et al.*, 2002; Haddad *et al.*, 2002; Hughes, Roberts *et al.*, 2000; Linn *et al.*, 2003; Mukasa *et al.*, 2002; *et al.*, 2003; Phillips *et al.*, 2001; Virtaneva *et al.*, 2001). For example, Pollack *et al.* estimate that 62% of highly amplified genes in 37 breast cancer tumors demonstrate moderately or highly elevated expression. Thus, aneuploidies can be detected in gene expression or array CGH microarray data, and it is necessary to develop analysis methods that can accurately identify chromosomal abnormalities based on either.

Accurate identification of aneuploidies from thousands of array CGH or gene expression measurements requires robust computational methods. Most array CGH data analyses involve heuristics and threshold-based

methods (Dunham *et al.*, 2002; Hughes, Roberts *et al.*, 2000; Pollack *et al.*, 2002). Recently, Autio *et al.* (2003) presented a dynamic-programming-based approach to identifying copy-number changes from array CGH data, which addressed the problem algorithmically for CGH data but lacked significance analysis. Accurate identification of potential copy number changes based on gene expression data is even more challenging because of mRNA expression levels reflect transcriptional regulation as well as DNA copy number. Previous approaches for aneuploidy detection from gene expression data focus only on whole-chromosome or chromosomal-arm copy number changes, and most methods are based on heuristics or dataset-specific thresholds. In the most sophisticated method to date, Crawley *et al.* (2002) employ a sign test for detecting whole chromosome (or whole arm) expression biases. Hughes and Roberts *et al.* (2000) use a simpler error-weighted mean approach for whole-chromosome aneuploidy detection and a heuristic scanning method that identifies adjacent occurrences of 4 over or under-expressed genes as potential segmental aneuploidies. A visualization-based imbalance detection scheme for identifying biases common in cancer specimens as compared to normal samples is proposed by Kano *et al.* (2003). These methods address the problem of whole chromosome or chromosomal arm copy changes, but the issue of robust identification of segmental aneuploidies remains open.

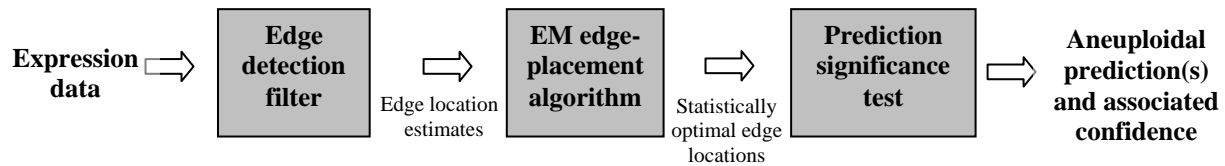
Here we present ChARM, a robust and accurate statistical method for identification of segmental aneuploidies from gene expression or array CGH microarray data. Our technique provides three key improvements over previously suggested approaches. First, nearly all current aneuploidy detection schemes for expression data rely on thresholds for defining significant over- and under-expression levels (some requiring up to a 1.7-1.8 fold change). Recent studies suggest, however, that expression level changes do not always directly reflect copy change proportions, and thresholds determined for one data set often will not generalize to others (Phillips *et al.*, 2001). Our method is statistical, and therefore generalizes to different datasets, microarray platforms, and organisms. Second, we focus on the problem of detecting segmental aneuploidy, which is generally more difficult than detecting whole-chromosome aneuploidy for which the methods developed by Hughes and Roberts *et al.* (2000) or Crawley *et al.* (2002) are effective. Third, our method is general and performs well with both gene expression and array CGH data.

ChARM employs an edge detection filter that identifies potentially aneuploid regions, an EM algorithm that finds maximum likelihood breakpoints based on a local search in these potential regions, and a statistical analysis that determines which predicted aneuploidies correspond to statistically significant biases as opposed to experimental noise. Our scheme can accurately identify known aneuploidies in biological gene expression or array CGH data (Hughes and Roberts *et al.*, 2000), and rigorous performance analysis with synthetic data demonstrates that the method is robust to noise and aneuploidy size and thus can generalize to other microarray data sets. Applying ChARM to 300 gene expression profiles of laboratory yeast strains, we identify multiple previously unknown aneuploidies, most of which are supported by current biological knowledge of yeast chromosomal rearrangement mechanisms. Our analysis of breast cancer array CGH and gene expression microarray data identifies both known and novel areas of chromosomal instability and reveals two groups of immune system genes on different chromosomes that are overexpressed and often amplified in a subset of breast tumors. This novel result may, upon experimental verification, contribute to understanding of how cancers escape immune response.

## **METHODS**

ChARM is composed of three sub-systems: an edge detection filter that identifies points on chromosomes where potential aneuploidies start or end, an EM-based edge-placement algorithm that statistically optimizes these start and end locations, and a window significance test that determines whether predicted amplifications and deletions are statistically significant or are artifacts of noise (Figure 1). The EM algorithm has a well-known tendency to find local rather than global maxima, but this three-stage structure is useful in setting initial conditions that ensure meaningful convergence. All three stages assume input in the form of array CGH or gene expression log ratios arranged in the order in which the corresponding genes appear along a single chromosome.

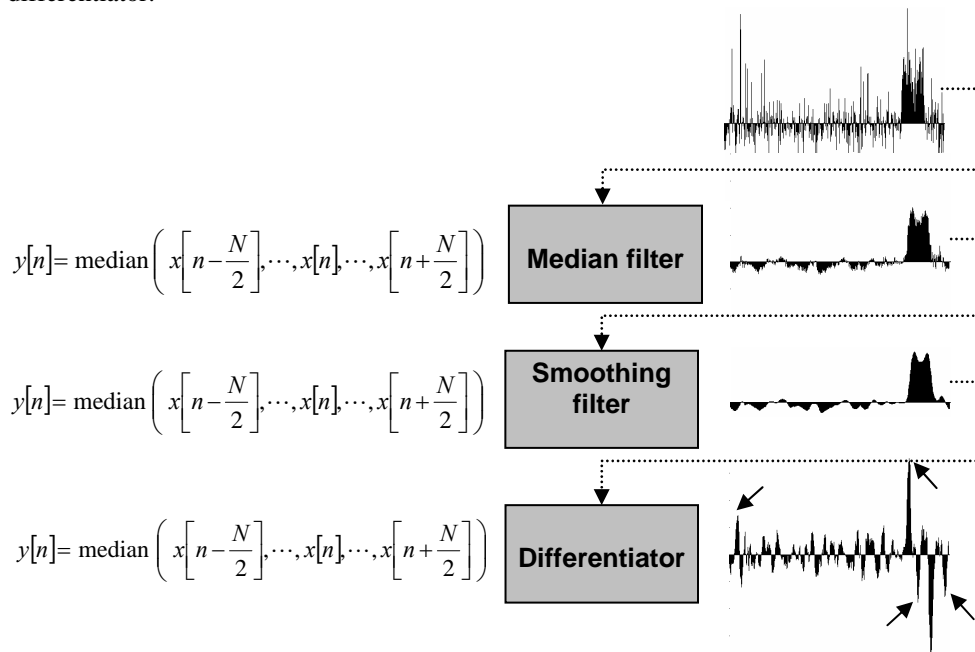
**Figure 1.** Three-stage segmental aneuploidy detection scheme. The edge detection filter estimates edge coordinates, which are then refined by the EM edge-placement algorithm. The resulting edges serve as input to the prediction significance test that analyzes statistical significance of spatial biases.



### Edge Detection Filter

The edge detection filter estimates locations along the chromosome where abrupt changes in gene expression occur. This is accomplished by a simple cascade of a non-linear median filter, a linear smoothing filter, and a linear differentiator (Figure 2). The median filter functions as a high-level smoother, removing outliers, which are common in microarray data, and preserving only sustained changes in the input sequence. Finer smoothing, which is a necessary pre-processing step for the differentiator, is accomplished by a linear averaging filter with a smaller window size. The differentiator effectively computes the derivative over a short window flagging any substantial changes with large peaks. These peaks and the corresponding chromosomal locations serve as the input to the more precise EM algorithm.

**Figure 2.** Preliminary edge detection filtering process illustrated on gene expression data positioned along the chromosome. Bars above the coordinate axis represent overexpression, bars below represent underexpression. The input-output relation for each of the filters is given on the left.  $y[n]$  is the output as a function of  $x[n]$  where  $n$  refers to gene index on the chromosome and  $N$  is the window size of each filter. Significant peaks are marked at the output of the differentiator.



### Expectation-Maximization Edge-Placement Algorithm

The purpose of the EM edge-placement algorithm is to provide fine adjustments to the edge estimates from the previous filter. To facilitate convergence to statistically optimal gene indices, each edge is surrounded by a "radius of influence" (ROI), which includes an equal-length set of adjacent genes on either side that is allowed to affect the placement at a given iteration. Furthermore, each edge is associated with two distributions, one for each of the two distinct regions (left and right) it is potentially separating. Each iteration of the algorithm consists of two stages: a typical EM clustering stage for learning the maximum likelihood parameters of the two distributions for each ROI (see E-step, M-step 1 below) and an edge-placement stage which adjusts the edge

position optimally given the learned parameters (see M-step 2 below). Before each edge adjustment, every pair of adjacent windows<sup>1</sup> is tested for similarity to ensure that the edge between these windows actually separates chromosomal regions of different copy number. The algorithm converges when all edge positions are fixed for several iterations. Each of these steps is described in detail below.

### Update membership (E-step)

Soft (fuzzy) memberships are computed for all genes in the radius of influence of an edge and are proportional to the probability of observing the gene given the left and right distributions associated with that edge. Let  $G_i = [g_i, l_i]$  represent the log-transformed ratio (array CGH or expression) and location of gene  $i$ ,  $e_j^{(t)}$  denote edge  $j$ , and  $\theta_{j,1}^{(t)}$  and  $\theta_{j,2}^{(t)}$  the left and right edge distributions at iteration  $t$  of the EM algorithm. Also, let  $r_{inf}$  denote the radius of influence. Here, we assume that the set of genes in the ROI lie in two normal distributions, i.e.  $\theta_{j,k}^{(t)}$  is parameterized<sup>2</sup> by  $[\mu_{j,k}^{(t)}, \sigma_{j,k}^{(t)}]$ . Then, the conditional probability of observing gene  $i$  given the distribution  $\theta_{j,k}^{(t)}$  is:

$$P(G_i | \theta_{j,k}^{(t)}) = \begin{cases} N(g_i; \mu_{j,k}^{(t)}, \sigma_{j,k}^{(t)}) & \text{for } l_i \in [e_j^{(t)} - r_{inf}, e_j^{(t)} + r_{inf}] \\ 0 & \text{otherwise} \end{cases}$$

which allows us to compute the posterior probability of  $\theta_{j,k}^{(t)}$  given gene  $i$  as:

$$P(\theta_{j,k}^{(t)} | G_i) = \frac{P(G_i | \theta_{j,k}^{(t)}) P(\theta_{j,k}^{(t-1)})}{\sum_{m=1,2} P(G_i | \theta_{j,m}^{(t)}) P(\theta_{j,m}^{(t-1)})}$$

where  $P(\theta_{j,k}^{(t-1)}) = \frac{1}{n_g} \sum_{i=1}^{n_g} P(\theta_{j,k}^{(t-1)} | G_i)$  and  $n_g$  is the number of genes on the chromosome of interest.

### Mean and variance computation (M-step 1)

Based on the membership  $P(\theta_{j,k}^{(t)} | G_i)$  determined in the E-step, the maximum likelihood mean and variance parameters for the next iteration ( $t+1$ ) are computed as follows (Dempster *et al.*, 1976):

$$\mu_{j,k}^{(t+1)} = \frac{\sum_{i=1}^{n_g} P(\theta_{j,k}^{(t)} | G_i) g_i}{\sum_{i=1}^{n_g} P(\theta_{j,k}^{(t)} | G_i)} \quad \sigma_{j,k}^2{}^{(t+1)} = \frac{\sum_{i=1}^{n_g} (x_i - \mu_{j,k}^{(t+1)})^2 P(\theta_{j,k}^{(t)} | G_i)}{\sum_{i=1}^{n_g} P(\theta_{j,k}^{(t)} | G_i)} \quad \text{when}$$

$$G_i \sim N(\mu_j^{(t)}, \sigma_j^{(t)}) .$$

### Edge adjustment (M-step 2)

For edge adjustment, we use the information theoretic notion of surprise (i.e. the amount of information learned from observing a probabilistic event). At each iteration, we restrict the possible edge locations to only the set of indices included in the current ROI. Each placement implies a different clustering of the genes around the edge into the left or right edge distributions. Each gene's placement in the implied cluster is treated as the observation of a random variable whose probability distribution is the gene's posterior probability of being associated with that cluster. For instance, if  $G_i$  falls in  $\theta_{j,1}^{(t)}$  for a particular placement of the edge  $e_j^{(t)}$ , the

<sup>1</sup> We refer to the regions between any two adjacent edges or between an edge and a chromosome end as "windows".

<sup>2</sup> Note that in our implementation, we use normally distributed  $g_i$ 's. Empirically, this has demonstrated adequate performance, but this approach can be generalized to other, more accurate models as well.

surprise of this event is  $S(G_i) = -\log\left(P\left(\theta_{j,k}^{(t)} \mid G_i\right)\right)$ . Then, the “minimum surprise” edge placement is given by:

$$e_j^{(t+1)} = \arg \min_i \left[ \sum_{k=1}^{i-1} \log\left(P\left(\theta_{j,1}^{(t)} \mid G_k\right)\right) + \sum_{k=i}^{2r_{inf}+1} \log\left(P\left(\theta_{j,2}^{(t)} \mid G_k\right)\right) \right]$$

where the indices  $1 \dots (2r_{inf}+1)$  refer to those genes in the ROI. Upon adjusting the edge placement for each window, the window parameters are updated accordingly (i.e.

$$e_j^{(t)} \rightarrow e_j^{(t+1)}, \theta_{j,k}^{(t+1)} = \left[ \mu_{j,k}^{(t+1)}, \sigma_{j,k}^2{}^{(t+1)} \right].$$

### Window similarity test

The window similarity test is needed at each iteration to ensure that edges about to be adjusted actually separate different windows with distinct chromosomal biases (separate aneuploidy predictions). The difference between left and right windows on either side of an edge must exceed a minimum signal-to-noise threshold or the edge is removed. As noted earlier, a window that extends beyond the ROI includes all genes up to the next edge or chromosome end. We have evaluated several parametric and non-parametric statistical metrics for measuring the difference between two sets of samples including t-test, non-parametric t-tests, rank-sum test, Kolmogorov-Smirnov test. Empirically, the ratio of the difference in medians between two adjacent windows and the pooled absolute deviation from the median has demonstrated the best performance. Thus, we impose the following criterion on this modified signal-to-noise ratio (SNR) for removing an edge ( $e_j^{(t)}$ ) at iteration  $t$ :

$$SNR_{i,j} = \frac{|med_{j,1} - med_{j,2}|}{\frac{1}{n_{j,1} + n_{j,2}} \left( \sum_{k \in w_{j,1}} |g_k - med_{j,1}| + \sum_{k \in w_{j,2}} |g_k - med_{j,2}| \right)} < SNR_{\text{thresh}} \left( \bar{\delta}_e \right)$$

for  $med_{j,k} = \text{median}(w_{j,k})$  where  $w_{j,1}$  and  $w_{j,2}$  include all the genes in the adjacent windows with sizes  $n_{j,1}$  and  $n_{j,2}$  respectively.  $SNR_{\text{thresh}}$  is a threshold dependent on the current convergence behavior

measured by  $\bar{\delta}_e$ , the average edge position change (in gene indices) from one iteration to the next. We raise the minimum SNR threshold as the edge positions begin to converge so that adjacent windows must be “more different” to remain separate as edges approach their final estimates.

### Window Significance Analysis

Once the EM algorithm obtains precise window positions, the significance analysis scheme determines if each window represents a statistically significant spatial bias in DNA content or expression. We consider three statistical tests for assessing the significance of windows identified by the EM algorithm: a one-sample sign test, a mean permutation test, and a coefficient of variance permutation test, as well as combinations of the mean and sign tests and the variance and sign tests. The sign test is that reported by Crawley *et al.* (2002) with the modification that the threshold is chosen dynamically for each chromosome to allow for identification of biased regions exhibiting lower degrees of over or under-expression than the 1.7-1.8 fold threshold used by others (Crawley *et al.*, 2002). Both permutation tests require performing approximately 5,000 random permutations of the genes on the chromosome and comparing the statistic (mean or variance) obtained on the actual arrangement with the most significant statistic for the same window size on each random permutation. We use the Bonferroni method to correct for multiple hypothesis tests on the same chromosome. Our permutation tests are designed specifically for the segmental aneuploidy problem, while other methods such as the sign test or the error-weighted mean approach proposed in (Hughes and Roberts *et al.*, 2000) are more appropriate for chromosome-wide bias detection.

## EVALUATION

To systematically assess ChARM's accuracy and robustness, we evaluate it using a synthetic microarray measurement error model described below. Using this model, we assess which window significance test yields the best performance for aneuploidy detection and thoroughly evaluate the robustness of our scheme. We further evaluate our scheme on biological data (see Application to Biological Data).

### Synthetic data model

We generate synthetic two-color microarray data according to the model proposed by Rocke and Durbin (2001). Under this two-component model, reference ( $y_R$ ) and test ( $y_T$ ) intensity values are simulated as:

$$y_R = \alpha_R + \mu_R e^{\eta_S + \eta_R} + \varepsilon_S + \varepsilon_R \quad y_T = \alpha_T + \mu_T e^{\eta_S + \eta_T} + \varepsilon_S + \varepsilon_T,$$

where  $\alpha$  is the mean background intensity,  $\mu$  is the intensity contributed by the quantity of interest, and

$$\begin{aligned} \eta_S &\sim N(0, \sigma_{\eta_S}), \quad \eta_R \sim N(0, \sigma_{\eta_R}), \quad \eta_T \sim N(0, \sigma_{\eta_T}) \\ \varepsilon_S &\sim N(0, \sigma_{\varepsilon_S}), \quad \varepsilon_R \sim N(0, \sigma_{\varepsilon_R}), \quad \varepsilon_T \sim N(0, \sigma_{\varepsilon_T}). \end{aligned}$$

This model was originally proposed for gene expression microarrays, but it is also appropriate for array CGH experiments with the modification that  $\mu_R$  and  $\mu_T$  are amounts of reference and test genomic DNA rather than mRNA. The parameters denoted by the subscript "s" are characteristics of the microarray spot and common to both reference and test samples. The mean background intensities ( $\alpha$ ) are typically estimated by microarray image analysis software and used to compute estimates of test and reference signal intensities,  $x_T$  and  $x_R$ , as follows:

$$x_R = y_R - \hat{\alpha}_R \quad x_T = y_T - \hat{\alpha}_T.$$

We model the error in this background estimation,  $\hat{\alpha}$ , as an additional normally distributed error term,  $\varepsilon_{est}$ , so that the pre-log-ratio intensities are generated as:

$$x_R = \mu_R e^{\eta_S + \eta_R} + \varepsilon_S + \varepsilon_R + \varepsilon_{est} \quad x_T = \mu_T e^{\eta_S + \eta_T} + \varepsilon_S + \varepsilon_T + \varepsilon_{est}$$

**Table 1.** Estimated parameters for array CGH and expression human breast cancer data. Parameters were estimated as suggested by Rocke and Durbin (2001).

Parameter	Microarray type	
	Array CGH	Expression
$\hat{\alpha}_T, \hat{\alpha}_R$	59.2, 45.9	399, 238
$\hat{\mu}_T, \hat{\mu}_R$	111, 113	3980, 4130
$\hat{\sigma}_{\eta_S}, \hat{\sigma}_{\eta_T}, \hat{\sigma}_{\eta_R}$	.63, .059, .090	.53, .17, .13
$\hat{\sigma}_{\varepsilon_S}, \hat{\sigma}_{\varepsilon_T}, \hat{\sigma}_{\varepsilon_R}$	25, 11, 0	137, 54, 94

Parameters for this model are estimated as suggested by Rocke and Durbin (2001) for biological array CGH and gene expression experiments (Table 1). Prior to noise addition, test and reference intensities across each synthetic chromosome for all simulations are drawn from a normal distribution with  $\mu \sim N(3980, 800)$ , and the mean background

intensity is assumed to be 400 for test and reference samples with  $\varepsilon_{est} \sim N(0, 40)$ . Regions of aneuploidy are

synthetically produced by setting all affected genes' test-to-reference ratio  $\left(\frac{\mu_T}{\mu_R}\right)$  to  $1.5^3$  (prior to noise

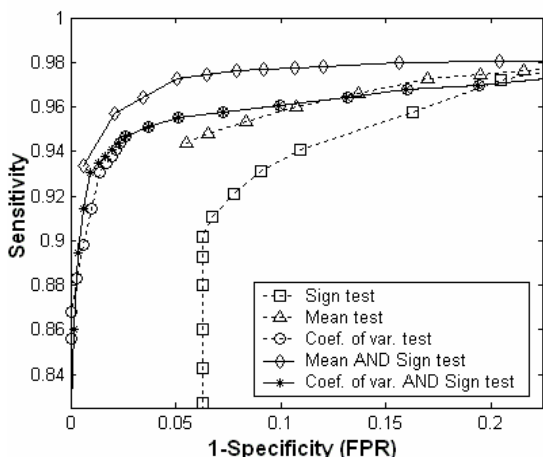
effects). Furthermore, to model expression scenarios realistically, 10% of the genes outside of aneuploidal regions are randomly set to over- or under-expressed with no spatial correlation.

<sup>3</sup> As gene expression changes do not directly reflect DNA copy number, the test-to-reference ratio for a gene that has been duplicated will not necessarily be 2. We chose to set these ratios to 1.5 to provide a conservative evaluation of our method.

### Choice and performance of window significance test

We first address the question of choosing the window significance test for our framework. We consider three window significance tests (sign test, mean test, coefficient of variance test) and evaluate their performance on simulated 50-gene aneuploidies under varying p-value cutoffs (Figure 3). Under all conditions tested, the mean and coefficient of variance permutation tests perform overwhelmingly better than the one-sample sign test, which is used by Crawley *et al.* (2002) and Haddad *et al.* (2002). However, when an aneuploidy is located on the end of a chromosome, the mean test, which is generally very specific, can falsely report the region spanning the rest of the chromosome as significant based on the permutations. This shortcoming of the permutation-based approach can be overcome by combination with the simpler sign test. This combined mean permutation and sign test scheme performs best both in terms of specificity and sensitivity, and is thus used in the rest of evaluation experiments. A similar combination of the coefficient of variance test and the sign test is less effective because the variance-based test yields lower sensitivity due to the noisy characteristics of microarray data.

**Figure 3.** Receiver operating characteristic (ROC) curves for sign test, mean test, coefficient of variance, and combined tests with p-value cutoffs between  $10^{-6}$  and .4. Performance was evaluated on synthetic data with simulated 50-gene aneuploidies and generated with  $\sigma_{\eta_R}, \sigma_{\eta_T} = .25; \sigma_{\eta_S} = .15; \frac{\sigma_{\epsilon_T}}{\alpha_T}, \frac{\sigma_{\epsilon_S}}{\alpha_R}, \frac{\sigma_{\epsilon_R}}{.5(\alpha_T + \alpha_R)} = .2$ . A combined mean and sign test shows the highest sensitivity at every false positive rate (FPR) tested.



### Robustness evaluation

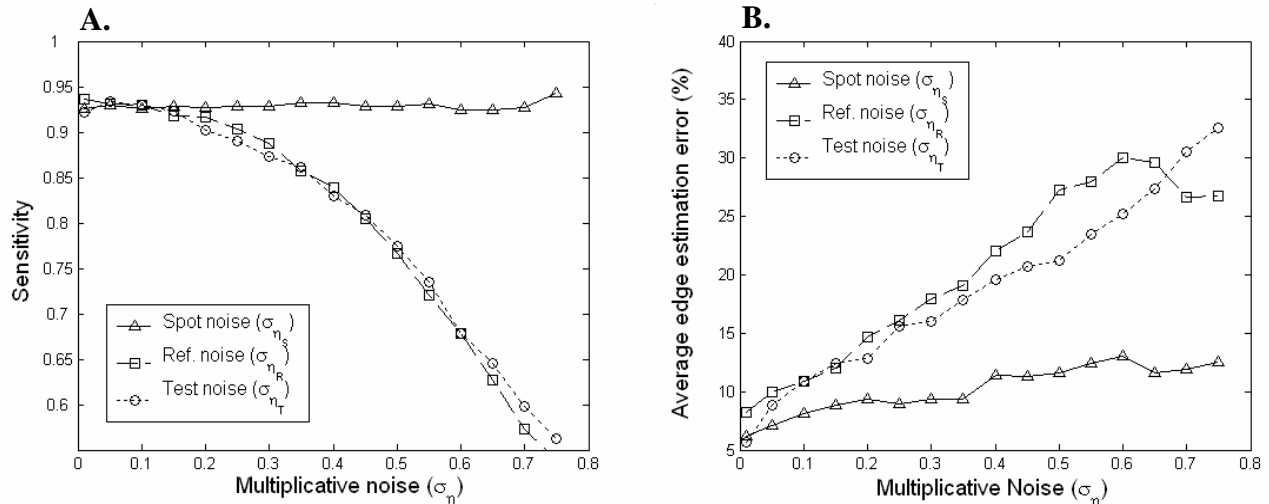
We also examine the performance of ChARM under varying noise conditions. The performance of the method is only minimally affected by additive noise ( $\epsilon$  parameters) (data not shown). The effect of multiplicative error ( $\eta$ ) in test and reference samples is shown in Figure 4. The sensitivity of the algorithm is robust ( $\geq .9$ ) to noise levels well above the biological range (Figure 4A, Table 1), and the specificity ranges from 1 to .94 for all noise parameters (data not shown). Our method provides accurate edge placement at biologically realistic noise levels (average edge coordinate error  $< 8\%$ ) (Figure 4B). Edge coordinate error is defined as

$$\Delta = \frac{\sum_i (|\hat{e}_{i,1} - e_{i,1}| + |\hat{e}_{i,2} - e_{i,2}|)}{\# \text{ of identified aneuploidies}}, \text{ where parameters } \hat{e}_{i,1} \text{ and } \hat{e}_{i,2} \text{ refer to the edge estimates of the } i^{\text{th}} \text{ prediction,}$$

and  $e_{i,1}$  and  $e_{i,2}$  are the known edge locations of the synthetic aneuploidy. Both sensitivity and edge placement error are more sensitive to multiplicative reference and test noise than to shared spot noise.

To test for bias in our method's performance toward particular aneuploidal segment sizes, we perform a similar noise analysis across a range of typical lengths (results not shown). At moderate biological noise levels (0.1), the algorithm identifies even small segments ( $< 10$  genes) of copy-number change with very high specificity ( $> .95$ ). Under severe noise conditions the sensitivity of the detection algorithm degrades quite noticeably for very small aneuploidies (much less than 100 genes in length). However, the algorithm is able to detect larger copy number changes ( $> 100$  genes) even under high noise conditions ( $\sigma_{\eta}$  10 times greater than typical biological noise) with relatively high sensitivity. The edge coordinate errors behave similarly, although with less degradation. Both effects are due to the fact that separating signal from noise becomes more difficult as the length of spatial correlation decreases. Therefore our scheme is robust to noise and can accurately identify aneuploidy regions even under high noise conditions.

**Figure 4.** Effect of multiplicative noise on **A.** sensitivity and **B.** errors in edge coordinates (as % of total window size). Performance of the scheme in identifying a 50 gene aneuploidal segment was evaluated under varying degrees of noise.  $\sigma_{\eta_S}$  was varied while the remaining terms were fixed at .1. Similarly,  $\sigma_{\eta_T}, \sigma_{\eta_R}$  were varied with  $\sigma_{\eta_S} = .5$ . Biological noise is typically under 0.65 for  $\sigma_{\eta_S}$  and under 0.2 for  $\sigma_{\eta_T}, \sigma_{\eta_R}$  (Table 1). P-value cutoffs were set at  $10^{-3}$  and  $10^{-2}$  for the sign and mean permutation tests respectively, and the tests were combined as previously described. The detection scheme with the combined mean and sign window significance test identifies most windows (>90%) with high accuracy in placement of edge coordinates (error < 0.1%) and is robust to high levels of spot, test, and reference noise (substantially higher than noise levels common in biological data shown in Table 1).



## APPLICATIONS TO BIOLOGICAL DATA

We applied ChARM to the yeast deletion mutants' gene expression data set of Hughes and Marton *et al.* (2000) and to gene expression and array CGH data for breast cancer patients from (Pollack *et al.*, 2002). The results, presented below, demonstrate that our method can be successfully applied to both gene expression and array CGH biological data for different organisms. We outline known amplifications and deletions that ChARM identifies and present some novel aneuploidies we find as well.

### Segmental aneuploidies in *S. cerevisiae* deletion mutants

We applied our method to the compendium of expression profiles of 300 *S. cerevisiae* deletion mutants and drug-treated strains developed and previously analyzed for aneuploidies by Hughes and Roberts *et al.* (2000). The analysis by Hughes *et al.* emphasizes whole-chromosome copy number changes, and they identify based on gene expression data and confirm by array CGH only two segmental aneuploidies<sup>4</sup>. Our method identifies these confirmed segmental aneuploidies (*rpl20aΔ/rpl20aΔ* and *rad27Δ/rad27Δ* strains) with high confidence (*rad27Δ/rad27Δ*: sign test p-value of  $10^{-5}$ , mean permutation test p-value of  $<10^{-4}$ ; *rpl20aΔ/rpl20aΔ*: sign test p-value of  $10^{-7}$ , mean permutation test p-value of  $<10^{-4}$ ).

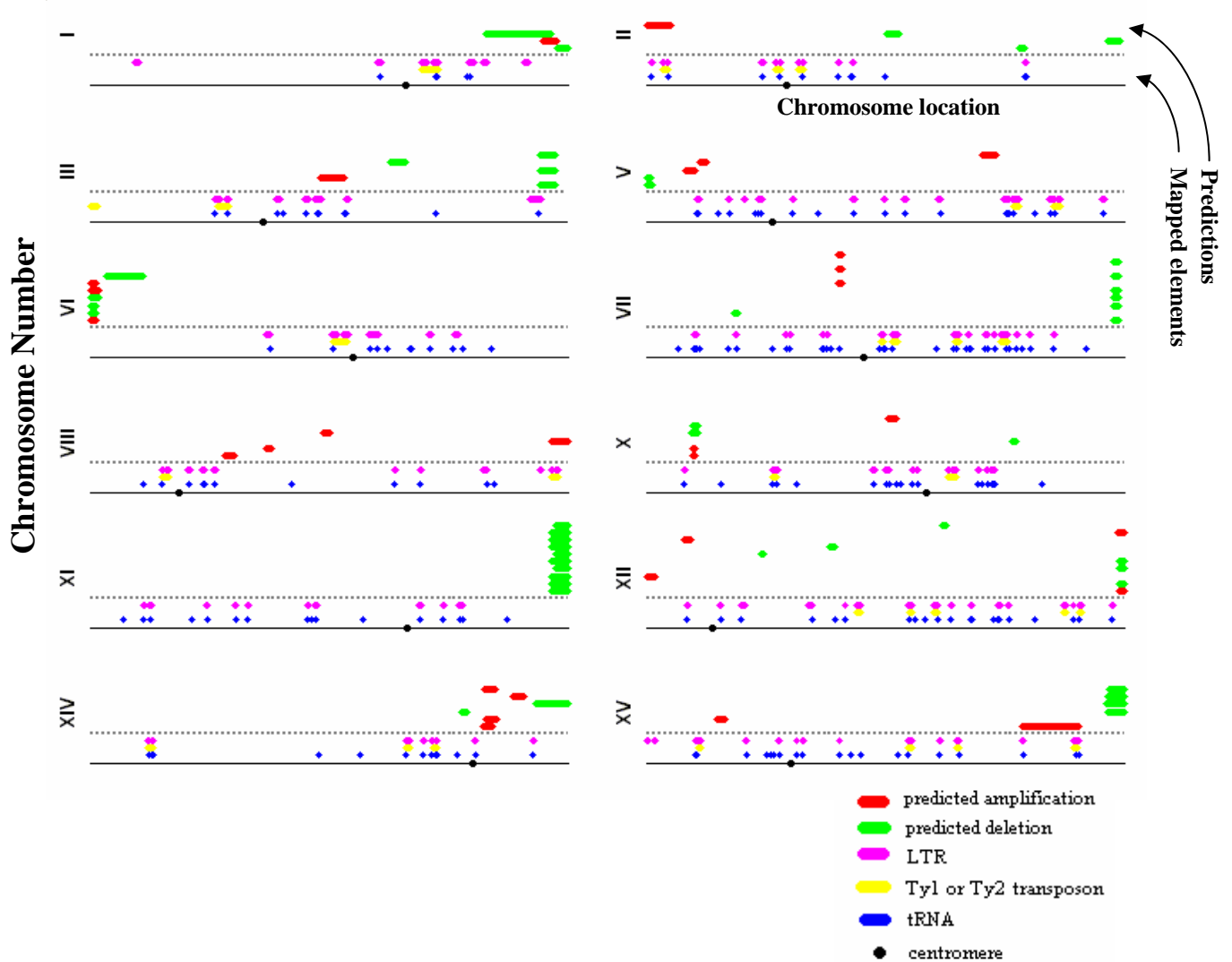
In addition to confirming the segmental aneuploidies identified by Hughes *et al.*, we identify a number of previously unknown potential aneuploidal regions<sup>5</sup>, the top 100 (sign test p-values of  $<10^{-3}$  and mean permutation test p-values of  $<10^{-2}$ ) of which are pictured in Figure 5, and expression profiles of two are displayed in Figure 6. To assess the biological significance of these results, we use biological models of mechanisms of chromosomal breakage and aneuploidy formation in yeast. Chromosomal amplifications and deletions in yeast are thought to arise through ectopic recombination between homologous sequences, such as Ty transposons, transposon-related long terminal repeats (LTRs), or tRNA sequences (Infante *et al.*, 2003).

<sup>4</sup> Hughes *et al.* identified one additional segmental aneuploidy (in *top3Δ*) based on array CGH. This aneuploidy is not reflected in the gene expression data and thus cannot be identified by any gene expression analysis method.

<sup>5</sup> Predictions that represented two adjacent occurrences of Ty transposons or included centromeric regions were excluded from further analysis due to the potential of cross-hybridization artifacts.

Thus, presence of transposons, LTRs, or tRNA sequences near the edges of a predicted aneuploidy region can serve as biological evidence that the region in question truly contains an amplification or deletion. In addition, increased chromosomal breakage may be observed in the conserved Y' areas at the ends of the yeast chromosomes (Chan and Tye, 1983). Our analysis reveals that 73% of predictions presented in Figure 5 are significantly ( $p$ -value  $< 0.1$ ) closer to such homologous sequences than expected by chance or are located in the Y' regions. These predictions likely correspond to novel segmental aneuploidies, while other predictions may represent array artifacts or aneuploidies that arose through an alternative molecular mechanism.

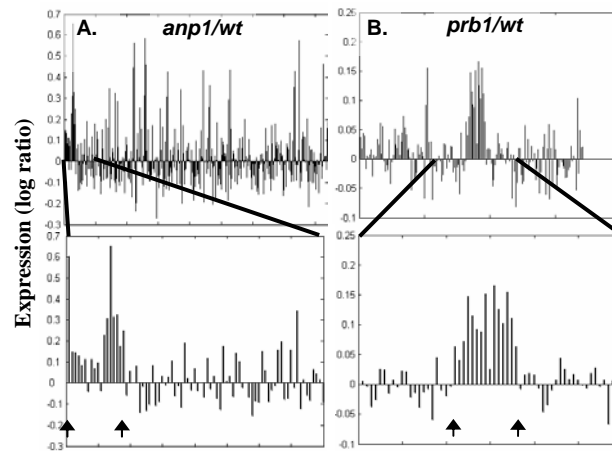
**Figure 5.** Chromosomal maps showing a subset of predicted aneuploidies (sign test  $p$ -values of  $< 10^{-3}$  and mean permutation test  $p$ -values of  $< 10^{-2}$ ) and biologically relevant mapped chromosomal elements. Aneuploidies are color-coded: red indicates amplification and green indicates deletion. Predictions shown in different rows on the same chromosome correspond to different yeast strains (e.g. Chr II), and multiple predictions at the same chromosomal coordinate represent identical aneuploidies found in multiple strains (e.g. Chr XI). Proximity of predictions to LTR, transposon, and tRNA elements was evaluated through 10,000 random placements of same-sized regions on the chromosomal map and through finding the proportion of random regions with shorter distance ( $d_{rand}$ ) to homologous elements than real predictions ( $d_{obs}$ )  $\left[ p = \frac{\text{count}(d_{rand} < d_{obs})}{\text{count}(\text{rand placements})} \right]$ .



In yeast deletion mutant strains undergoing multiple divisions, an aneuploidy that compensates for or masks the deleted gene's phenotype could confer a selective advantage (Dunham *et al.*, 2002). For example, growth

defects (Saccharomyces Genome Database, 2004) caused by the deletion of *anp1* (Figure 6A), an endoplasmic reticulum (ER) protein with a role in retention of glycosyltransferases in the Golgi (Jungmann and Munro, 1998), may be alleviated by the amplification of the region on chromosome II that includes *SFT2*, a gene involved in ER-Golgi transport (Conchon *et al.*, 1999). The *hdf1* deletion mutant also exhibits a compensatory mechanism. Hdf1 protein functions as a heterodimer with the Ku protein in maintaining normal telomere length and structure, but cells can maintain telomeres in the absence of telomerase through a recombination-dependent “survivor” pathway that replicates Y’ regions of chromosomes (Lendvay *et al.*, 1996). Indeed, we identify amplifications in the Y’ region of chromosomes II, VI, and XII in this *hdf1*Δ*hdf1*Δ strain.

**Figure 6.** Gene expression levels plotted by chromosomal location in example segmental aneuploidies: **A.** *anp1* (chromosome II, sign test p-value of  $< 10^{-10}$ , mean permutation test p-value of  $10^{-5}$ ) and **B.** *prb1* (chromosome III, sign test p-value of  $< 10^{-10}$ , mean permutation test p-value of  $< 10^{-4}$ ) heterozygous deletion mutants. Aneuploidies predicted by our method are identified by arrows and correspond to spatial expression biases.



### Identification of aneuploidies in breast cancer gene expression and array CGH data

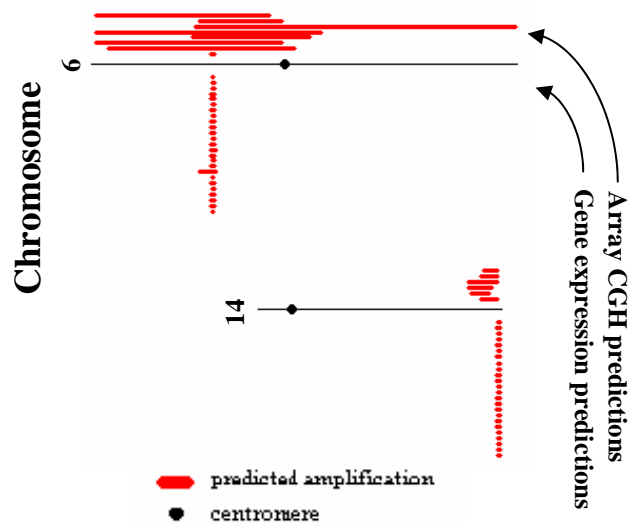
Genomic instability is thought to play a major role in oncogenesis, and breast tumors specifically are known to harbor multiple aneuploidies (Gollin, 2004; Pollack *et al.*, 2002). Using ChARM, we analyzed array CGH (Pollack *et al.*, 2002) data for 44 breast tumors and the corresponding gene expression studies for 37 of these sample (Sorlie *et al.*, 2003). Our method identifies the known “hot spots” of amplifications and deletions in breast cancer (Hyman *et al.*, 2002; Pollack *et al.*, 2002), including multiple cases of deletions on 13q that include tumor suppressor protein Rb1 and on 17p that span tumor suppressor protein Tp53. Deletion of either Rb1 or Tp53 is known to cause chromosomal instability, and we do identify multiple additional aneuploidies in tumors with predicted Rb1 or Tp53 deletion (Lentini *et al.*, 2002). We also identify a known 17q amplification that includes proto-oncogene ERBB2/HER2 (Menard *et al.*, 2000).

One advantage of our method is the ability to make predictions based independently on array CGH or gene expression data. Overlaps in these independent predictions can be used to focus on potentially functionally relevant segmental aneuploidies. The two most striking overlap regions both include immune system proteins: genes that encode class II major histocompatibility complex proteins (MHCII) on chromosome 6, and immunoglobulin heavy chain genes on chromosome 14 (Figure 7). It is surprising to find such expression levels of these immune proteins in the tumor samples. One concern is that the data reflect the presence of a lymphocytic infiltrate in tumor tissue, however in such a case one would not expect correlated amplification data. Immune system effects on tumor progression are relatively poorly understood; a key question is why some tumors are recognized and destroyed by the immune system while others successfully proliferate.

Immunoglobulins, also known as antibodies, are secretable proteins produced by mature B lymphocytes. These molecules play an essential part in the adaptive immune system by binding and neutralizing foreign particles. As immunoglobulin gene expression typically occurs only in B lymphocytes after directed germline rearrangement, immunoglobulin heavy chain overexpression and amplification of the corresponding region is potentially an important finding, but requires further investigation into the functional status of the transcripts. MHCII is another key component of adaptive immune response – it is a membrane protein whose primary role is

the presentation of protein fragments for immune recognition. However, MHCII presentation of foreign proteins activates a response optimally in the presence of other costimulatory molecules, and MHCII overexpression outside of this immune context may lead to immune tolerance, a condition when tumors do not activate immune response (Hardwick, 1998; Hendrickson *et al.*, 2002). One theory is that malignant tumors may induce tolerance with out-of-context immune stimuli, thereby evading immune response, which allows them to grow and proliferate (Mapara and Sykes, 2004). No definitive evidence for this theory exists, but these effects have been observed in model systems (Ostrand-Rosenberg *et al.*, 1996; Byrne and Halliday, 2003) and MHCII overexpression has been associated with poor prognosis in melanomas (Brocker *et al.*, 1985). Experimental verification of our findings may provide novel evidence of induction of immune tolerance in tumors.

**Figure 7.** Overlapping amplification predictions in array CGH and gene expression microarray data for breast cancer. Amplifications predicted from gene expression data are shown below the chromosomal map, those predicted from array CGH data are shown above the map.



## CONCLUSIONS

We have demonstrated that segmental aneuploidies can be identified based on array CGH or gene expression microarray data and have presented a robust statistical method that can accurately locate aneuploidies in biological data. Evaluations on synthetic and biological data show that our method is robust to experimental noise and aneuploidy size and thus is appropriate for general and automated application to microarray data sets. ChARM allows routine screening of gene expression data for aneuploidies and is sensitive enough to detect small statistically significant signal biases in mixed populations of cells. It is important to note that gene expression does not always reflect copy number and, furthermore, algorithms based on gene expression data alone cannot discriminate between spatial expression biases that arise from DNA abnormalities and biases that are the result of spatial coregulation or array artifacts. Our method can identify spatial expression biases due to either aneuploidies or technology artifacts and thus can be used as a general screening tool for gene expression microarray data. In cases when ChARM is used to screen for aneuploidies only, gene expression microarray data should be normalized for special artifacts prior to applying ChARM (Yang *et al.*, 2002). Applying ChARM to biological data, we have identified multiple previously unknown aneuploidies in public yeast gene expression data, several of which are supported by biological evidence, and potential amplification and overexpression of immune genes in breast cancer. These predictions should be further evaluated through targeted laboratory investigation.

## ACKNOWLEDGEMENTS

We would like to thank Peter Kasson, Mitchell Garber, Kai Li, Kara Dolinski, and David Botstein for valuable discussions and help in analyzing biological results. CLM is supported by Program in Integrative Information, Computer and Application Sciences funded by the NSF.

## REFERENCES

- Autio, R., Hautaniemi, S., Kauraniemi, P., Yli-Harja, O., Astola, J., *et al.* (2003) CGH-Plotter: Matlab toolbox for CGH-data analysis. *Bioinformatics*, **19**, 1714-1715.
- Baskar, S., Clements, V.K., Glimcher, L.H., Nabavi, N., Ostrand-Rosenberg, S. (1996) Rejection of MHC class II-transfected tumor cells requires induction of tumor-encoded B7-1 and/or B7-2 costimulatory molecules. *J. Immunol.* **156**, 3821-7.
- Brocker, E.B., Suter, L., Bruggen, J., Ruiters, D.J., Macher, E., Sorg, C. (1985) Phenotypic dynamics of tumor progression in human malignant melanoma. *Int. J. Cancer.* **36**, 29-35.
- Byrne, S.N., Halliday G.M. (2003) High levels of Fas ligand and MHC class II in the absence of CD80 or CD86 expression and a decreased CD4+ T cell Infiltration, enables murine skin tumours to progress. *Cancer Immunol Immunother.* **52**, 396-402.
- Cahill, D.P., Kinzler, K.W., Vogelstein, B., Lengauer, C. (1999) Genetic instability and Darwinian selection in tumours. *Trends Cell Biol.*, **9**, M57-M60.
- Chan, C.S., Tye, B.K. (1983) Organization of DNA sequences and replication origins at yeast telomeres. *Cell.* 1983, **33**, 563-73.
- Conchon, S., Cao, X., Barlowe, C., Pelham, H.R.B. (1999) Got1p and Sft2p: membrane proteins involved in traffic to the Golgi complex. *The EMBO Journal*, **18**, 3934-3946.
- Crawley, J.J., Furge, K.A. (2002) Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Gen. Biol.*, **3**, 0075.1-0075.8.
- Dempster, A.P., Laird, N.M., Rubin, D.B. (1976) Maximum likelihood from incomplete data via the EM algorithm. *Journ. Royal Statistical Society*, **B39**, 1-39.
- Dolinski, K., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., *et al.* Saccharomyces Genome Database. <http://db.yeastgenome.org/cgi-bin/SGD/locus.pl?locus=anp1>. (May 27, 2004)
- Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O., Rosenzweig, F., Botstein, D. (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *PNAS*, **99**, 16144-16149.
- Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G., Louis, E.J. (2000) Chromosomal evolution in *Saccharomyces*. *Nature*, **405**, 451-454.
- Fritz, B., Schubert, F., Wrobel, G., Schwaenen, C., Wessendorf, S., *et al.* (2002) Microarray-based copy number and expression profiling in dedifferentiated pleomorphic liposarcoma. *Cancer Research*, **62**, 2993-2998.
- Gray, J., Collins, C., (2000) Genome changes and gene expression in human solid tumors. *Carcinogenesis*, **21**, 443-452.
- Gollin, S.M. (2004) Chromosomal instability. *Curr Opin Oncol.* **16**, 25-31.
- Haddad, R., Furge, K.A., Miller, J., Haab, B.B., Schoumans, J., *et al.* (2002) Genomic profiling and cDNA microarray analysis of human colon adenocarcinoma and associated intraperitoneal metastases reveals consistent cytogenetic and transcriptional aberrations associated with progression of multiple metastases. *App. Gen. and Prot.*, **1**, 123-134.
- Hardwick, K.G. (1998) The spindle checkpoint. *Trends Genet.*, **1**, 1-4.
- Hendrickson, H., Slechta, E.S., Bergthorsson, U., Andersson, D.I., Roth, J.R. (2002) Amplification-mutagenesis: Evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification. *PNAS*, **99**, 2164-2169.
- Hoyt, M.A., Totis, L., Roberts, B.T. (1991) *S. cerevisiae* genes required for cell cycle arrest in response to loss of microtubule function. *Cell*, **66**, 507-517 (1991).
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Sloughton, R., *et al.* (2000) Function discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.
- Hughes, T.R., Roberts, C.J., Dai, H., Jones, A.R., Meyer, M.R., *et al.* (2000) Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature*, **25**, 333-337.

- Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S. (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, **62**, 6240-6245.
- Infante, J., Dombek, K., Rebordinos, L., Cantoral, J., Young, E. (2003) Genome-wide amplifications caused by chromosomal rearrangements play a major role in the adaptive evolution of natural yeast. *Genetics*, **165**, 1745-59.
- Jungmann, J., Munro, S. (1998) Multi-protein complexes in the cis Golgi of *Saccharomyces cerevisiae* with alpha-1,6-mannosyltransferase activity. *The EMBO Journal*, **17**, 423-34.
- Kano, M., Nishimura, K., Ishikawa, S., Tsutsumi, S., Hirota, K., *et al.* (2003) Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions. *Physiol Genomics*, **13**, 31-46.
- Lendvay, T. S., Morris, D. K., Sah, J., Balasubramanian, B., Lundblad, V. (1996) Senescence mutants of *Saccharomyces cerevisiae* with a defect in telomere replication identify three additional EST genes. *Genetics*, **144**, 1399-1412.
- Lentini, L., Pipitone, L., Di Leonardo, A. (2002) Functional inactivation of pRB results in aneuploid mammalian cells after release from a mitotic block. *Neoplasia*, **4**, 380-7.
- Linn, S.C., West, R.B., Pollack, J.R., Zhu, S., Hernandez-Boussard, T., *et al.* (2003) Gene expression patterns and gene copy number changes in Dermatofibrosarcoma protuberans. *Am. Journ. of Path.*, **163**, 2383-2395.
- Mapara, M.Y., Sykes, M. (2004) Tolerance and cancer: mechanisms of tumor evasion and strategies for breaking tolerance. *J Clin Oncol*, **22**, 1136-51.
- Menard, S., Tagliabue, E., Campiglio, M., Pupa, S.M. (2000) Role of HER2 gene overexpression in breast carcinoma. *J Cell Physiol*, **182**, 150-62.
- Mukasa, A., Ueki, K., Matsumoto, S., Tsutsumi S., Nishikawa, R., *et al.* (2002) Distinction in gene expression profiles of oligodendrogliomas with and without allelic loss of 1p. *Oncogene*, **21**, 3961-3968.
- Ostrand-Rosenberg, S., Baskar, S., Patterson, N., Clements, V.K. (1996) Expression of MHC Class II and B7-1 and B7-2 costimulatory molecules accompanies tumor rejection and reduces the metastatic potential of tumor cells. *Tissue Antigens*, **47**, 414-21.
- Pérez-Ortín, J.E., García Martínez, J., Alberola, T.M. (2002) DNA chips for yeast biotechnology. the case of wine yeasts. *Journ. of Biotechnology*, **98**, 227-241.
- Phillips, J.L., Hayward, S.W., Wang, Y., Vasselli, J., Pavlovich, C., *et al.* (2001) The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. *Cancer Research*, **61**, 8143-8149.
- Pinkel, D., Seagraves, R., Sudar, D., Clark, S., Poole, I., *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207-211.
- Pollack, J., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., *et al.* (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature*, **23**, 41-46.
- Pollack, J., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *PNAS*, **99**, 12963-12968.
- Rice, J.A. (1995) *Mathematical Statistics and Data Analysis*, 2<sup>nd</sup> edn., Duxbury Press, CA.
- Roche, D.M., Durbin, B. (2001) A model for measurement error for gene expression arrays. *Journ. of Comp. Biol.*, **8**, 557-569.
- Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*, **100**, 8418-8423.
- Tan P., Anasetti C., Hansen J.A., Melrose J., Brunvand M., *et al.* (1993) Induction of alloantigen-specific hyporesponsiveness in human T lymphocytes by blocking interaction of CD28 with its natural ligand B7/BB1. *J Exp Med*, **177**, 165-73.
- Virtaneva, K., Wright, F.A., Tanner, S.M., Yuan, B., Lemon, *et al.* (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *PNAS*, **98**, 1124-1129.
- Wilhelm, M., Veltman, J.A., Olshen, A.B., Jain, A.N., Moore, D.H., *et al.* (2002) Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Research*, **62**, 957-960.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D., Peng, V., *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**, 15.