Gene expression

Assessing the need for sequence-based normalization in tiling microarray experiments

Thomas E. Royce[1], Joel S. Rozowsky[2] and Mark B. Gerstein[1,2,3,*]

[1]Interdepartmental Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, [3]Department of Computer Science. Yale University, New Haven, CT 06520, USA

ABSTRACT

Motivation: Increases in microarray feature density allow the construction of so–called tiling microarrays. These arrays, or sets of arrays, contain probes targeting regions of sequenced genomes at regular genomic intervals. The unbiased nature of this approach allows for the identification of novel transcribed sequences, the localization of transcription factor binding sites (ChIP–chip), and high resolution comparative genomic hybridization, among other uses. These applications are quickly growing in popularity as tiling microarrays become more affordable. To reach maximum utility, the tiling microarray platform needs be developed to the point that 1nt resolutions are achieved and that we have confidence in individual measurements taken at this fine of resolution. Any biases in tiling array signals must be systematically removed to achieve this goal.

Results: Towards this end, we investigated the importance of probe sequence composition on the efficacy of tiling microarrays for identifying novel transcription and transcription factor binding sites. We found that intensities are highly sequence dependent and can greatly influence results. We developed three metrics for assessing this sequence dependence and use them in evaluating existing sequence–based normalizations from the tiling microarray literature. In addition, we applied three new techniques for addressing this problem; one method, adapted from similar work on GeneChip brand microarrays, is based on modeling array signal as a linear function of probe sequence, the second method extends this approach by iterative weighting and re–fitting of the model, and the third technique extrapolates the popular quantile normalization algorithm for between-array normalization to probe sequence space. These three methods perform favorably to existing strategies, based on the metrics defined here.

Availability: http://tiling.gersteinlab.org/sequence_effects/

1 INTRODUCTION

1.1 Motivation

Following any genome sequencing project comes the desire for identifying the functional elements therein (ENCODE Consortium, 2004). These elements include, but are not limited to, protein coding regions, regulatory regions and methylation sites. In addition to defining functional elements, it is also of great interest to understand variability within the genome sequence itself. This variability may be present in mutation hot spots or in single nucleotide and copy number polymorphisms, for example. Luckily, the DNA microarray technology (Chee *et al.*, 1996; Schena *et al.*, 1995) has evolved beyond the targeting of known mRNA transcripts to the unbiased targeting of any genomic target with the advent of high density genome tiling microarrays (Selinger *et al.*, 2000). All of the post–genome investigations listed here are enabled in a high–throughput fashion by the hybridization of labeled nucleic acids to this emerging microarray technology.

As reviewed in (Mockler *et al.*, 2005) and (Johnson *et al.*, 2005), tiling microarrays contain hundreds of thousands to millions of features, each containing probes that target some short (~25–1,000nt) genomic region, or tile. Their construction involves either printing PCR products (Rinn *et al.*, 2003), oligonucleotide inkjet deposition (Shoemaker *et al.*, 2001), or photolithographic in situ synthesis on a solid substrate (Kapranov *et al.*, 2002). This last construction yields the greatest feature densities and is therefore best suited to tiling even the very large human genome (Bertone *et al.*, 2004) and is therefore the focus of our study. A perfect tiling of a target genome contains one feature representing every *k*mer therein but current large–scale tiling designs typically leave short gaps (~5–50nt) between tile start positions to achieve greater coverage. Nevertheless, feature densities are ever–increasing (not unlike integrated circuits' transistor densities) and we may soon witness the manufacture of a comprehensive 1nt resolution human genome tiling microarray. Clearly, understanding the tiling microarray technology will become fundamental to our understanding of genome biology.

One challenge in developing the microarray platform to this level is that while genome tiling enables massively parallel experimentation, each individual experiment is not an optimized one. To clarify this point, each of these experiments relies on nucleic acid hybridization and both the sensitivity and specificity of these hybridizations are highly sequence dependent. For a given microarray probe, there exists an ideal set of experimental conditions (determined largely by its nucleotide sequence) that maximizes its ability to form a duplex with its intended target relative to that of its non–targets. Every microarray feature contains probes with a different nucleic acid sequence, so the hybridization conditions for a microarray experiment are necessarily a compromise. The degree to which this compromise is detrimental can be lessened a great deal in gene–centric microarrays by selecting each genes' representative probe(s) such that its optimum target hybridization condition lies somewhere near the pre–selected conditions for the microarray experiment. This luxury disappears when we move to tiling microarrays since probe selection is more limited and goes to zero as tiling resolution approaches 1nt with greater feature densities. Therefore, other solutions are needed. In the current work, we investigated the degree to which probe sequence-based normalization can alleviate this problem.

1.2 Previous work

Affymetrix GeneChip brand microarrays are a popular platform for studying genes' mRNA expression levels and are manufactured analogously to the tiling microarrays studied in this work (Lipshutz *et al.*, 1999). In the GeneChip platform, each assayed transcript is targeted by a probeset which consists of ~16-20 unique features. Each of these features contains probes that target the same transcript but the probes in any feature are not the same as those in any other feature. It was noted early on that features targeting the same transcript can yield signals that vary by orders of magnitude (Li & Wong, 2001). When the arrays' sequences became publicly available, it became clear that these differences were chiefly due to differences in probe nucleotide composition (Naef & Magnasco, 2003). It is reasonable to expect that similar sequence effects are present within the results of tiling microarray hybridizations but the effect's presence and prevalence has not yet been measured and documented. It is furthermore expected that the effects in tiling microarray experiments will be similar, but not necessarily identical to those observed in GeneChips. Two reasons for this are that (1) the majority of tiling microarray features should not exhibit signal whereas the majority of gene-centric arrays' features do, and (2) tiling microarray experiments usually involve hybridization of labeled cDNA to the microarrays as opposed to labeled cRNA used in GeneChip experiments.

The methodology initially adopted by Affymetrix for coping with sequence biases in their GeneChip platform involves the so–called mismatch probe control. For every feature with probes perfectly matching (PM) the target, a mismatch (MM) feature is provided. Each MM feature has probes identical to its corresponding PM features, save the middle nucleotide. The idea is that non–targets will bind to the MM features' probes with affinities similar to those that they have for the paired PM features' probes but that the affinity for the PM's target is greatly reduced. Thus, subtracting MM signal from PM signal theoretically yields the amount of observed PM signal due to target–specific binding.

The details of this solution have proven unsatisfactory (Irizarry *et al.*, 2003). Therefore, considerable effort has been put into developing Affymetrix GeneChip analysis methodologies that do not utilize the MM features (this has the added benefit of requiring half the number of features to quantify transcript abundances). One general approach is to model a feature's signal as the product of its target's expression level and a feature-specific 'affinity' (Irizarry *et al.*, 2003; Li & Wong, 2001). Given a number of independent array hybridizations and multiple features in a probe set, the probe affinities and the transcript levels corresponding to each hybridization can be reliably obtained.

The discovery that features' affinities for a transcript can be predicted by their probes' sequences (Naef & Magnasco, 2003) motivates models in which neither MM features nor multiple arrays are necessary to estimate target concentrations. In one model (Zhang *et al.*, 2003), a feature's signal is decomposed into specific and nonspecific parts, each of which contains concentration and probe sequence–related parameters. The transcript concentration is specific to a probeset but the sequence parameters are universal to the whole array. Another model (Hekstra *et al.*, 2003) of this type fits microarray signals to a Langmuir adsorption model that has parameters estimable from probe sequence.

Practicalities of tiling microarray experiments make applying many of the aforementioned GeneChip methods difficult, if not impossible. Using mismatch probes is straightforward to implement and has been applied to tiling microarrays (eg (Kapranov *et al.*, 2002)). The downside to this approach is that tiling density and/or coverage must be sacrificed. Methods which require hybridizations under multiple cellular conditions (Irizarry *et al.*, 2003; Li & Wong, 2001; Wu & Irizarry, 2005) are not always practical either. Given the current expense of conducting whole genome tiling microarray experiments, typically just a singular condition is analyzed. In addition, to apply these methods or the method of (Zhang *et al.*, 2003) would require analyzing multiple neighboring probes in a sliding window. Sliding windows are currently part of the standard approach for analyzing tiling arrays, but we would ideally like to move away from this resolution-decreasing technique and be able to obtain reliable measurements at the resolution of a single tile. To this end, analysis techniques which estimate affinities based solely on sequence composition (Hekstra *et al.*, 2003)hold promise for tiling microarrays, but parameters of these models need to be estimated from spike-in datasets such as the Affymetrix latin square studies (www.affymetrix.com). Parameters cannot be taken directly from these GeneChip studies either, since the experiments (1) focus on hybridization within known transcripts primarily, and (2) utilize cRNA spike-ins whereas most tiling array experiments make use of cDNA targets. The differences between cRNA and cDNA hybridization are significant (Eklund *et al.*, 2006).

Currently, two methods for estimating sequence effects have been employed in the tiling array literature, besides the PM-MM approach. The first divides each feature's signal by the median signal of all features having identical GC content on the same array (Samanta *et al.*, 2006). The second uses control hybridizations of genomic DNA to estimate relative binding strengths. The latter approach is used extensively in ChIP-chip and aCGH applications and has been recently applied to transcript identification as well (David *et al.*, 2006; Huber *et al.*, 2006).

In addition to these two methods, we developed three sequence normalization techniques for tiling microarrays: one based on the GeneChip analysis of (Naef & Magnasco, 2003), another which extends their approach by iterative re–weighting and re–fitting of their model, and a third technique based on quantile normalization, which we extended to multivariate probe sequence space. Overall, we find that these corrections perform well at removing sequence biases based on metrics that we have defined. These corrections and metrics are useful tools for studying and improving the tiling microarray platform.

2 METHODS

2.1 Definitions

Before proceeding, we explicitly define a few commonly used microarray terms as they are sometimes used differently elsewhere. We define a *microarray*, or an *array* for brevity, as a substrate on which there are present numerous serially addressable features. Each *feature* contains many oligonucleotide *probes*. Within a feature, these probes all have identical sequence. *Hybridization* occurs when a labeled nucleic acid population is introduced to a microarray and allowed to seek and anneal their reverse–complement probes. Those labeled nucleic acids are called *targets* herein.

2.2 Microarray data

We utilized four microarray data sets, each briefly described below. Within each data set we first applied between–array quantile normalization (Bolstad *et al.*, 2003), removing any possible array–specific effects. Following normalization, we selected a single array representative of each data set. To achieve this, we first calculated pairwise correlation coefficients between all arrays in an experiment and then selected the array having the highest minimum correlation with all other arrays.

The first data set (Emanuelsson *et al.*, 2006) we used employs Affymetrix tiling microarrays hybridized with cDNA reverse transcribed from total RNA derived from human NB4 cells. This data set comprises four biological replicates, each replicated once, yielding eight arrays worth of data. The experiment's microarray design calls for features having twenty–five nucleotide probes representing genomic sequences that are, on average, twenty–one genomic base pairs apart. For each PM feature, there is a paired MM control feature. The array design has 737,680 such feature pairs and incorporates one strand of non–repetitive sequence from the entire ENCODE region (ENCODE Consortium, 2004).

Second, we utilized microarray data (Emanuelsson *et al.*, 2006) generated by hybridizing cDNA obtained via reverse transcribing NB4 total RNA to Nimblegen ENCODE tiling microarrays. These arrays contain 372,078 perfect–match features, each containing probes thirty–six nucleotides long. Their design targets both strands of non–repetitive sequence from ENCODE regions ENm001 through ENm011 at an average density of one feature per thirty–six genomic bases. Within this data set are three biological replicates, each technically replicated once.

The third dataset (David *et al.*, 2006) we investigated uses Affymetrix tiling microarrays targeting the whole of the *S. cerevisiae* genome. These arrays contain 3,276,800 feature pairs and were also used for transcript mapping. We focused here on polyadenylated transcripts. This is also the first transcript mapping experiment to include a control genomic hybridization.

To investigate our algorithms' utility in another experimental system, we applied them to an Affymetrix ChIP–chip dataset which investigates Sp1 binding across human chromosomes 21 and 22 (Cawley *et al.*, 2004). We used 'chip C' which examines binding within chromosome 22 and of which there are six replicates. Importantly, ChIP–chip data usually comes with a genomic control hybridization (Horak *et al.*, 2002). This control is present here as well.

2.3 Quantification of position–specific sequence effects

Of practical use is a scalar metric that can quantify any sequence effects observed in the previously described tiling datasets. Let m be the size, in nucleotides, of an array's probes. For each nucleotide position $k = 1 \ldots m$, calculate the Kruskal–Wallis statistic,

$$K_k = \frac{12\sum_{j\in\{A,C,G,T\}} C_{j,k}\left(\bar{r}_{j,k} - \frac{N}{2}\right)^2}{N(N+1)}, \qquad (1)$$

where N is the total number of features on the array, $C_{j,k}$ is the number of features having nucleotide j at position k in their probes, and $\bar{r}_{j,k}$ denotes the average rank of intensities from features having nucleotide j at probe position k. The scalar metric quantifying position–specific sequence effects, which we denote here by γ, is then the average over K_k,

$$\gamma = \frac{1}{m}\sum_{k=1}^{m} K_k. \qquad (2)$$

2.4 Assessment of tiling array performance

The scalar γ, by itself, does not suffice to quantify the quality of a data set with respect to sequence effects caused by ubiquitous background hybridization. This is because both data with low ubiquitous hybridization and randomized data would yield low γ values. Therefore, in addition to γ and in transcript detection experiments, we investigated the enrichment of features targeting known genes relative to features having identical GC content.

Specifically, each data set's probe sequences were compared against the latest version of Refseq (using BLAT (Kent, 2002)) to identify those features perfectly targeting a known transcript. These features' GC content were computed and used to select a set of non–Refseq 'control' features having probe–wise GC content identical to the Refseq features. The enrichment of both the Refseq and control features' signals in the top of the entire signal distributions was investigated by simply computing the percentage of features observed above the entire distribution's median intensity. Values greater than 50% indicated enrichment. In the yeast transcription dataset, we first identified those features outside of known ORFs and then sampled ORF–specific features with identical GC content. This modification was necessary due to the high percentage of coding DNA in the yeast genome.

To understand performance of ChIP–chip at the individual feature level, we followed an analogous strategy. We first identified those features whose probes target known promoter sequences (500bp upstream of Refseq annotated transcription start site). We also identified those features having probes with identical GC content as the promoter–targeting features. Enrichment in the ChIP–chip signal was then computed similarly as enrichment of known genes in the transcription experiments. We note that that within-promoter binding is a weak indicator of performance. This is required due to the paucity of known binding sites for any given transcription factor.

Theoretically, different tiling microarray platforms targeting the same regions should yield similar results. This is not always the case (Emanuelsson *et al.*, 2006; Johnson *et al.*, 2005). So, we decided to use platform concordance between the Affymetrix and Nimblegen NB4 datasets as an additional tiling microarray quality metric. To do this, we used BLAT (Kent, 2002) to identify those features from the Affymetrix design having probe sequences that lie completely within a features probe sequences from the Nimblegen design. (Supplemental Figure 1). We isolated 40,729 such feature pairs between the two designs. We used the Spearman correlation coefficient of these feature pairs' signals to assess platform concordance at the feature level. Note that correlation is not an absolute indicator of concordance for tiling microarrays because the majority of the features are merely generating noise. Therefore, correlations will always be fairly low. However, an increase in platform correlation following a normalization step would still represent an increase in concordance because, presumably, at least one source of error has been removed.

To assess agreement at the gene level, we first identified those Refseq genes represented by both platforms. Then, the features from the Affymetrix design were isolated along with their signal intensities. The signals were summarized for each gene with the pseudomedian, a commonly used summary statistic in the tiling array literature (Kampa *et al.*, 2004). Each gene's pseudomedian was computed from the Nimblegen data as well and the correlation of these pseudomedians was assessed with the Spearman correlation coefficient.

We realized that high correlation coefficients could be a GC content–related effect. If high GC content leads to high intensities in both experiments (sequence–specific ubiquitous hybridization is present in both), then we might expect significant correlations simply due to the fact that feature pairs have similar GC content. Therefore, we performed the above concordance studies with feature pairs having the same GC content as the previously identified feature pairs but that do not necessarily overlap with one another or with known genes.

3 RESULTS

3.1 Ubiquitous hybridization on tiling arrays

It is known that microarray features targeting the same transcript can yield significant intensity differences in GeneChip experiments (Li & Wong, 2001). This phenomenon has been identified in tiling microarrays as well (Royce *et al.*, 2005) and would prevent accurate estimation of nucleic acid abundance at the desired single feature resolution. One hypothesis is that the differences are at least partly due to differences in features affinities for their bound target. It is widely believed that these affinities are sequence dependent. To investigate the sequence dependence of feature intensities, we constructed position-specific quantile plots (Figure 1).

The plots' motivation came from previous work where linear models are fit to measured intensities with position-specific nucleotide content as regressors (Naef & Magnasco, 2003). Instead of fitting a regression explicitly, we calculated the qth percentile of signal intensities for features having an A, C, G or a T at position k in their probes. This was done for each nucleotide position $k = 1 \ldots m$ where m is the nucleotide length of each probe. These plots primarily show that sequence effects are present in both of the tiling microarray platforms investigated. Such effects are known to occur in GeneChips where cRNA constitutes the labeled target. Here, we

have demonstrated that such effects are also present when cDNA is used in place of cRNA. Interestingly, the effects are markedly different for the Affymetrix and Nimblegen experiments we studied. Specifically, cytosines appear to lend the largest contributions to signal in the Nimblegen experiment whereas guanines have this role in the Affymetrix experiment. Most importantly, these plots demonstrate that the effects are present for the lowest–intensity features on these arrays. The implication of sequence effects being present at low intensities in *H. sapiens* transcriptional tiling array data is that non–specific, ubiquitous binding is present at every feature since we do not expect specific binding (due to transcription) to be present for the entirety of the human genome (Harrow *et al.*, 2006). One strategy that we investigated for removal of these biases is to perform a control hybridization of genomic DNA and use this data to normalize the signal of interest. This is the common practice in ChIP-chip investigations and was recently applied in an *S. cerevisiae* transcript mapping experiment (David *et al.*, 2006). As we demonstrate in Figure 2 this approach may need some additional consideration, at least in ChIP-chip experiments.

3.2 Consequences of ubiquitous hybridization

Ubiquitous hybridization influences intensity distributions such that features with GC rich probes tend to have higher intensities than those with AT rich probes. When we compared Refseq targeting features' intensities to their array's median intensity, we found that 79% (Binomial test, $p < 10^{-15}$) and 68% ($p < 10^{-15}$) are greater than the median intensity for the Affymetrix and Nimblegen experiments, respectively (Table 1). By themselves, these numbers are reassuring. However, when we did the same computation for control features, we still found significant enrichment, albeit a bit less. Sixty–eight percent of the Affymetrix GC control probes and sixty–four percent of the Nimblegen control probes exhibited intensities above their slide median. Clearly, GC content is a main determinant of signal intensity; much more so than the targeting of known genes. This point is illustrated in Figure 3.

3.3 Corrective algorithms

In this subsection, we present algorithms for removing the sequence effects identified above. Following the algorithms' descriptions, we will report their performance with respect to metrics defined in Methods.

One approach (Samanta *et al.*, 2006) for dealing with these issues is to scale all of a microarray's intensities by each feature's GC content. To do this, all features must first be binned by their probes' GC content. Then, the median intensity is calculated within each GC bin. Finally, the features' intensities can be divided by the median intensity of features having identical GC content.

While the GC scaling approach may remove some of the problematic sequence biases observed in tiling array data, it only uses a summary of sequence content (%GC) and does not incorporate position–specific effects. To incorporate more sequence information, and to utilize knowledge of positional effects, we next adopted a model of background hybridization from the GeneChip literature in an attempt to more greatly reduce the observed sequence biases. This model is due to (Naef & Magnasco, 2003) and can be summarized as

$$\hat{S}_i = \sum_{k=1}^{m} \bar{S}_{i,k} + \bar{S} \tag{3}$$

where $\hat{S}_i$ indicates feature i's predicted log intensity due to ubiquitous hybridization, $\bar{S}_{i,k}$ denotes the mean logged intensity of features having the same nucleotide as feature i's probes at nucleotide index k, and $\bar{S}$ is the overall average logged feature intensities. This model has been suggested for tiling array analysis independently in (Munch *et al.*, 2006).

Another model for background hybridization on GeneChips (Zhang *et al.*, 2003) could have been applied here. However, this model is much more difficult to fit, would require sliding window estimation, and has been shown to less accurately predict nonspecific hybridization in Affymetrix GeneChips (Wu & Irizarry, 2005).

We extended the algorithm for computing Naef's affinities by fitting the same multiple linear regression to probe sequence in a more robust way. Following the initial fit (Equation 3), we down–weighted those features disagreeing with the model (e.g. exhibit large residuals) and re–fit the regression. This process of fitting and down–weighting was iterated until convergence as in the standard robust least squares regression model (Beaton, A.E., Tukey. J.W., 1974).

Formally, the procedure was to first predict logged signal intensity as a function of its m nucleotides following Equation 3. Once the predictions were computed, they were used to compute residuals,

$$\hat{\varepsilon}_i = S_i - \hat{S}_i \tag{4}$$

where S_i is the ith feature's logged intensity. The residuals were used to compute feature–specific weights such that features with high residuals receive low weight,

$$\begin{aligned} w_i &= \left(1 - x_i^2\right)^2, \quad \text{for} \quad |x| < 1 \\ &= 0, \quad \text{for} \quad |x| \geq 1 \end{aligned} \tag{5}$$

where

$$x_i = \frac{\hat{\varepsilon}_i}{C}. \tag{6}$$

C is a constant which controls the balance between iterations until convergence and overfitting. We set C to be six times the median of the residuals, following (Cleveland. W.S., 1979). Once weights were computed, the above steps were iterated until the $\hat{S}_i$ converged. Residuals from the final fit were taken to be the procedure's normalized values, $\tilde{S}_i$.

A nonparametric between–array normalization technique for Affymetrix GeneChip data is the so–called quantile normalization (Bolstad *et al.*, 2003). Briefly, this algorithm first computes a 'meta–array' by calculating either the mean or median signal for each feature across all microarrays being normalized. The meta–array's signal distribution is then used as the distribution for each array being normalized. This is achieved by replacing the signal of each feature having signal rank r with the signal having rank r within the meta–array. This nonparametric approach performs between–array normalization very well. In fact, nonparametric methods, in general, have been useful for microarray data analysis due to microarray data's lack of reproducible distributional form and their abundant outliers. For these reasons, we sought to apply the concept of quantile normalization to probe sequence space (Supplemental Figure 2).

However, to apply this technique to our task is non–trivial as our problem is a multivariate regression whereas the algorithm's original domain is inherently univariate. Our approach was to first force

the four distributions of signals coming from features having either an A, C, G or a T at nucleotide position 1 to be the same. Without loss of generality, we forced a uniform distribution between zero and one for each nucleotide group. The resulting normalized signals were stored as $\hat{S}_{i,1}$ where i indexes the ith array feature. The same computation was applied to each of the m positions, recording each $\hat{S}_{i,k}$, for $k = 1 \ldots m$. The element–wise mean over all m $\hat{S}$ vectors was then taken to be the normalized signal, $\tilde{S}_i$. This procedure did not explicitly normalize for nucleotide composition over all positions simultaneously since the sequence bias is more or less severe for different positions (Figure 1) and we had simply taken the feature–wise mean of the m individual corrections. Therefore, we iterated the previous procedure until convergence, setting the signals $S_i = \tilde{S}_i$ after each iteration. The effect is that at each iteration, the position with the strongest biases influenced the averaging more than the other positions. That is, the worst offending position will have the greatest influence on any iteration.

Formally stating the above procedure, we began by counting the number of probes $C_{j,k}$ having nucleotide $j = \{A,C,G,T\}$ at position k. This quantity was then used in computing the normalized rank intensities relative to position k,

$$\hat{S}_{i,k} = \frac{r_{i,j,k}}{C_{j,k}}, \tag{7}$$

where j is the nucleotide at position k within feature is probes, and $r_{i,j,k}$ is the magnitude rank of S_i relative to all other features having nucleotide j at position k. The normalized intensity, $\tilde{S}_i$, was then computed as the average of $\hat{S}_{i,k}$ over all positions k:

$$\tilde{S}_i = \frac{1}{m} \sum_{k=1}^{m} \hat{S}_{i,k}. \tag{8}$$

This procedure was iterated until convergence.

3.4 Position specific effects

We quantified position–specific sequence effects following Equation 2. The results of these calculations are summarized in Table 1 for the Affymetrix and Nimblegen NB4 transcriptional data. For both datasets, multivariate quantile normalization completely removed sequence biases as defined by Equation 2. Its performance was followed by corrections that used Naef's affinities, robust least squares, and finally by GC scaling. Out of these corrections, and for the Affymetrix data, the quantile normalization was the only method that outperformed the standard PM-MM approach. We found similar results in the ChIP–chip and yeast transcriptional data (data not shown).

3.5 Enrichment of Refseq Genes

As noted earlier, Equation 2 cannot be used as the sole determinant of tiling microarray performance with respect to their sequence biases. It is also important to demonstrate biological significance. To do this for transcriptional data, we computed the percentage of Refseq–targeting features' signals appearing in the top half of their signal distributions. Since gene annotations can have their own sequence biases which can confound this analysis, we also computed enrichment of features having identical GC content as the Refseq features. We provide these percentages in Table 2 for the Affymetrix and Nimblegen NB4 datasets. A generalization of this analysis is illustrated in Figure 4 where the percentages are plotted for one-hundred different evenly-spaced thresholds. In this figure, we have defined known positive features as those whose probes exactly match a Refseq gene and known negatives as features whose probes do not match any of Refseq but have the same GC content as the known positives.

The GC–richness of Refseq is immediately apparent in the first line of Table 2. Seventy–eight and seventy–three percent of Refseq targeting features are above slide median intensities in the Affymetrix and Nimblegen experiments, respectively. However, the control features show very high enrichment as well. Clearly, corrections are needed and the ones we described here all performed roughly equivalently at reducing the enrichment of control features while retaining enrichment of Refseq features. Interestingly, applying any of these corrections to yeast transcriptional data resulted in just minimal improvements over raw data and performed similarly to using a genomic DNA control hybridization as described in (Huber *et al.*, 2006) (Supplemental Figure 3).

We found that sequence normalizations do have an effect on selective enrichment of promoter–targeting features in ChIP–chip data, however. Table 3 indicates that Naef affinities, robust least squares, and multivariate quantile normalizations perform roughly equivalently at enriching for known promoters while diminishing enrichment of our computed control features. This table also demonstrates that utilizing the genomic control is important for ChIP–chip data and that its importance is enhanced by sequence normalizations.

3.6 Platform concordance

In performing the described sequence normalizations, it is important to achieve biological relevance but it also important to achieve platform concordance. Without platform concordance, it can become very difficult to reproduce other labs' results and skepticism about the technology can justifiably arise (Johnson *et al.*, 2005). We therefore performed two platform concordance analyses and summarized the results in Table 4. Probe–wise correlation between the two platforms' raw data was very low relative to GC content controls. The agreement appeared even worse when we compared Affymetrix PM-MM to Nimblegen's raw data. The disagreement is at least in part due to the low number of features exhibiting biological signal in these experiments. The degree to which this is causative of the low correlations is as yet unclear.

When using Naef's affinities to correct for sequence biases, we achieved the best results with respect to platform concordance at both the probe and gene levels. Robust least squares and multivariate quantile normalization performed nearly as well, with robust least squares apparently leaving more residual correlation in the gene–wise control. This residual correlation was also seen when we applied Naef's affinities or multivariate quantile normalization.

A final observation is that gene–wise correlations are always better than their probe–wise counterparts. This implicates the practical benefit of utilizing robust statistics within moving windows when scoring tiling array data (Kampa *et al.*, 2004), although this technique reduces our effective resolution.

4 DISCUSSION AND CONCLUSIONS

We have introduced here the problem of sequence biases caused by ubiquitous nonspecific hybridization in tiling microarray experiments. The effects were found to be strong and were much larger than the differences in intensity observed between Refseq targeting features' and non–Refseq targeting features' signals in *Homo sapiens* transcript mapping experiments (Figure 3). Furthermore, the effects are different for different platforms and can confound the study of platform concordance – a serious problem if results between experiments are to be integrated in downstream analyses. If the tiling

microarray technology is to eventually reach its goal of 1nt resolution these are issues that need to be resolved.

To this end, we investigated a number of approaches for mitigating the observed sequence biases. We found that these biases can be lessened by any of the methods employed. One method in particular, the multivariate quantile normalization, completely removed sequence effects present in tiling microarray data. Most importantly, the removal of these sequence biases did not come at the cost of removing biological realities from the data. In particular, in the Affymetrix system, we found that using approaches such as those presented here can allow for the removal of mismatch probes from the experimental design. In tiling microarrays, improvements in efficiency such as these allow for greater regions of DNA sequence to be interrogated. Benefits were found when we applied the algorithms to both transcript mapping and ChIP–chip data. The one surprise was that the algorithms provided little benefit to a recently published *S. cerevisiae* transcript mapping experiment. This is possibly due to the lessened complexity of the system being studied. With a much smaller transcriptome, there are fewer sequences able to bind to every probe. Therefore, the fraction of bound specific targets to bound off–targets at each feature is expected to be much higher than in the human transcript mapping experiments.

One of goals here was to improve the platform concordance between the Affymetrix and Nimblegen tiling microarray platforms. We have achieved this to some degree. After applying sequence normalizations described here, signal correlations within known genes can achieve Spearman's ρ of 0.64. However, correlations at the probe level remain very low (Spearman's ρ = 0.22) albeit much higher than the correlation observed without any corrections (ρ = 0.07). One possible source for this remaining disagreement, which has not been resolved, is the use of different probe lengths (Affymetrix' 25nt vs. Nimblegen's 36nt probes in this study). An experiment in which 25mers are synthesized with the Nimblegen technology might be able to address this question. One other possibility is differences in hybridization protocols which again could be targeted by a Nimblegen experiment which exactly mimics an Affymetrix study, using 25mers and following Affymetrix hybridization protocols. These studies require additional experimentation. However, we have addressed here the role that sequence effects can have on platform concordance – one of several factors that must be systematically studied.

The tiling microarray promises a wide spectrum of genome–scale experiments. For these experiments to be as useful as the genome sequences that enabled them, a deeper understanding of the technology itself is needed. One aspect of this understanding is the behavior of ubiquitous hybridization which we have begun to address here. Moving forward, the algorithms we provide should help researchers in recovering biologically useful information from tiling microarrays – both from transcript mapping experiments, and from ChIP–chip experiments.

ACKNOWLEDGMENTS

Many calculations in this work were made possible by the Yale Center for High Performance Computation in Biology and Biomedicine and NIH grant: RR19895-02, which funded the instrumentation. This work was also supported by National Institutes of Health grant P50 HG02357-01.

REFERENCES

Beaton, A.E., Tukey. J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, **16**(2), 147-185.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S. et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**(5705), 2242-2246.

Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185-193.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J. et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**(4), 499-509.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S. & Fodor, S.P. (1996). Accessing genetic information with high-density DNA arrays. *Science*, **274**(5287), 610-614.

Cleveland. W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J Am Statist Assoc*, **74**(368), 829-836.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W. & Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A*, **103**(14), 5320-5325.

Eklund, A.C., Turner, L.R., Chen, P., Jensen, R.V., deFeo, G., Kopf-Sill, A.R. & Szallasi, Z. (2006). Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat Biotechnol*, **24**(9), 1071-1073.

Emanuelsson, O., Nagalakshmi, U., Zheng, D., Rozowsky, J., Urban, A., Du, J., Lian, Z., Stolc, V., Weissman, S., Snyder, M. et al. (2006). Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res*, (), .

ENCODE Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**(5696), 636-640.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D. et al. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol*, **7 Suppl 1**(), S4.1-9.

Hekstra, D., Taussig, A.R., Magnasco, M. & Naef, F. (2003). Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res*, **31**(7), 1962-1968.

Horak, C.E., Luscombe, N.M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M. & Snyder, M. (2002). Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev*, **16**(23), 3017-3033.

Huber, W., Toedling, J. & Steinmetz, L.M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **22**(16), 1963-1970.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. & Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, **31**(4), e15.

Johnson, J.M., Edwards, S., Shoemaker, D. & Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*, **21**(2), 93-102.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G. et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*, **14**(3), 331-342.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A. & Gingeras, T.R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**(5569), 916-919.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res*, **12**(4), 656-664.

Li, C. & Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, **98**(1), 31-36.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet*, **21**(1 Suppl), 20-24.

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E. & Ecker, J.R. (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**(1), 1-15.

Munch, K., Gardner, P.P., Arctander, P. & Krogh, A. (2006). A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7**(), 239.

Naef, F. & Magnasco, M.O. (2003). Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys*, **68**(1 Pt 1), 011906.

Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M. et al.

(2003). The transcriptional activity of human Chromosome 22. *Genes Dev*, **17**(4), 529-540.

Royce, T.E., Rozowsky, J.S., Bertone, P., Samanta, M., Stolc, V., Weissman, S., Snyder, M. & Gerstein, M. (2005). Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet*, **21**(8), 466-475.

Samanta, M.P., Tongprasit, W., Sethi, H., Chin, C. & Stolc, V. (2006). Global identification of noncoding RNAs in Saccharomyces cerevisiae by modulating an essential RNA processing pathway. *Proc Natl Acad Sci U S A*, **103**(11), 4192-4197.

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235), 467-470.

Selinger, D.W., Cheung, K.J., Mei, R., Johansson, E.M., Richmond, C.S., Blattner, F.R., Lockhart, D.J. & Church, G.M. (2000). RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nat Biotechnol*, **18**(12), 1262-1268.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G. et al. (2001). Experimental annotation of the human genome using microarray technology. *Nature*, **409**(6822), 922-927.

Wu, Z. & Irizarry, R.A. (2005). Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol*, **12**(6), 882-893.

Zhang, L., Miles, M.F. & Aldape, K.D. (2003). A model of molecular interactions on short oligonucleotide microarrays. *Nat Biotechnol*, **21**(7), 818-821.

	Affymetrix NB4	Nimblegen NB4
No Correction	33465.07	14959.68
PM-MM	469.81	NA
GC Scaling	73230.20	9319.38
Robust Least Squares	880.08	301.82
Naef & Magnasco, Munch	595.10	44.63
Quantile Normalization	< 0.00	< 0.00

Table 1: γ was calculated for array data following each of the normalization methods employed. The PM-MM correction is not applicable for Nimblegen data as this platform provides no MM probes.

	Affymetrix		Nimblegen	
	Refseq	Control	Refseq	Control
No Correction	78.0%	68.4%	72.9%	64.8%
PM-MM	64.1%	51.0%	NA	NA
GC Scaling	61.7%	48.4%	59.9%	49.4%
Robust Least Squares	66.0%	53.2%	56.9%	50.9%
Naef & Magnasco, Munch	63.4%	49.7%	56.5%	50.6%
Quantile Normalization	63.2%	49.8%	56.4%	50.3%

Table 2: Percentage of features exhibiting signals greater than their distributions' median signal. PM-MM is not applicable for Nimblegen data because no MM probes are present in the array design.

	Sp1 Channel		Log(Sp1/Genomic)	
	Promoters	Control	Promoters	Control
No Correction	73.0%	67.8%	73.4%	56.7%
PM-MM	61.5%	53.8%	67.8%	50.6%
GC Scaling	59.6%	54.2%	66.2%	47.5%
Robust Least Squares	53.7%	48.7%	72.9%	56.0%
Naef & Magnasco, Munch	54.4%	48.8%	71.9%	53.9%
Quantile Normalization	54.8%	49.4%	72.2%	53.0%

Table 3: Percentage of features exhibiting signals greater than their distributions' median log ratio. The two left–hand columns refer to Sp1 ChIP data. The right–hand columns are with respect to logged Sp1/genomic ratios. For the ratios, PM-MM includes only those features for which PM-MM is positive in both channels and is applied to the signals before taking the log ratio. All other normalizations are performed on the log ratio directly.

	Probe-Wise		Gene-Wise	
	Matching	Control	Matching	Control
No Correction	0.49	0.42	0.69	0.74
PM-MM	0.07	0.03	0.39	0.08
GC Scaling	0.11	-0.01	0.46	0.03
Robust Least Squares	0.22	0.01	0.64	0.19
Naef & Magnasco, Munch	0.22	0.01	0.64	0.11
Quantile Normalization	0.19	-0.01	0.64	0.10

Table 4: Correlation coefficients were computed between signals from features targeting identical sequences (first column) and between signals from feature pairs having identical GC content as the original pairs but otherwise having no significant sequence similarity (second column). The pseudomedian of signals from features targeting the same gene were computed for each platform and presented under the Gene-Wise column header. Specifically, a correlation coefficient was calculated between genes' pseudomedians for genes represented in both array designs (third column). Features with identical GC content to these were then substituted into the pseudomedian calculation and correlation coefficients were again computed (fourth column). All coefficients are Spearman's ρ.

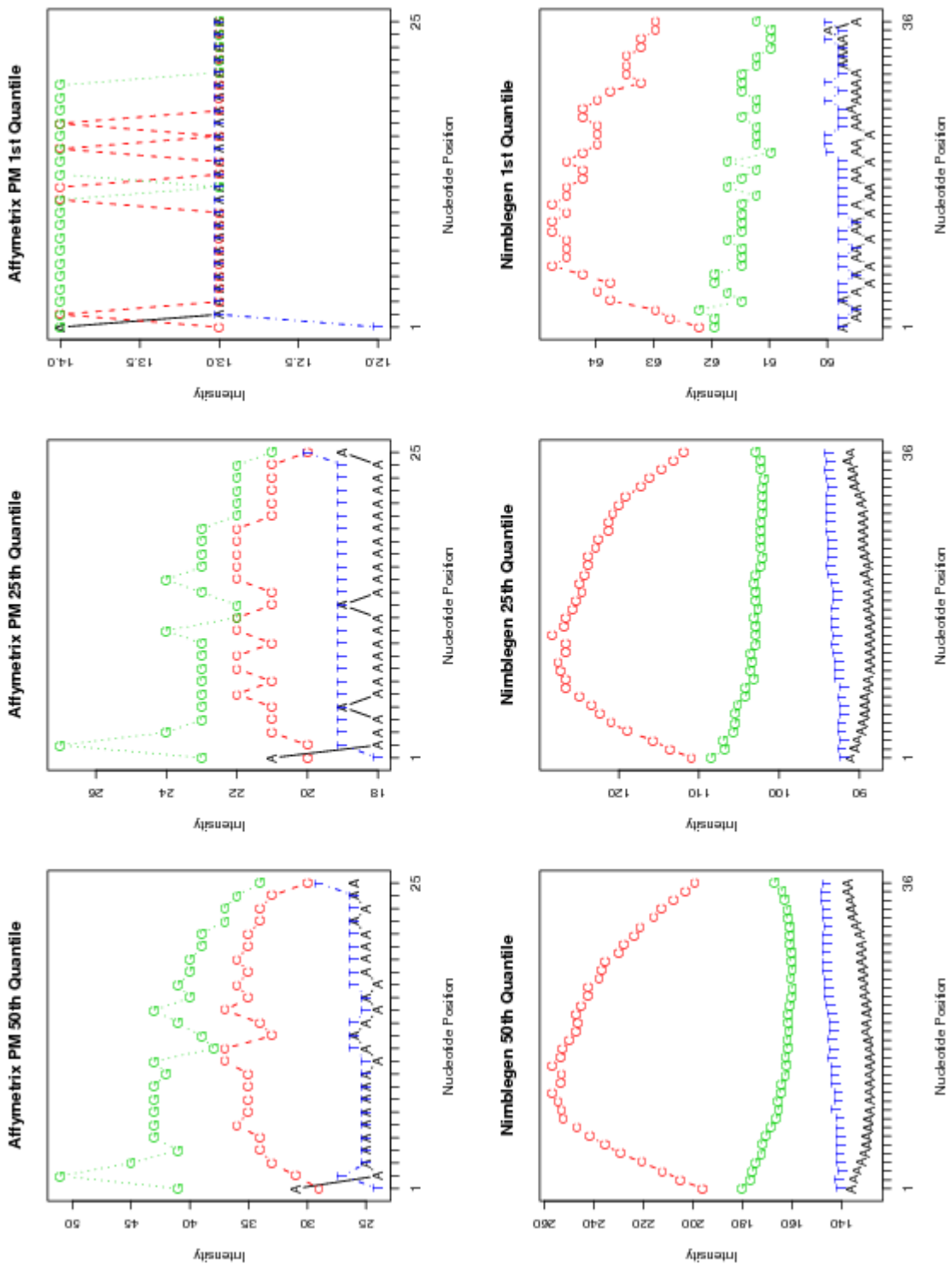

Figure 1: Position-specific quantile plots for Affymetrix and Nimblegen NB4 transcription data. On each *x*-axis are nucleotide indices. The *y*-axes are signal intensities. In a given plot, the *q*th percentile is computed for probes having an A, C, G or T at each nucleotide position. These percentiles are plotted for q=0.5, q=0.25 and q=0.01.

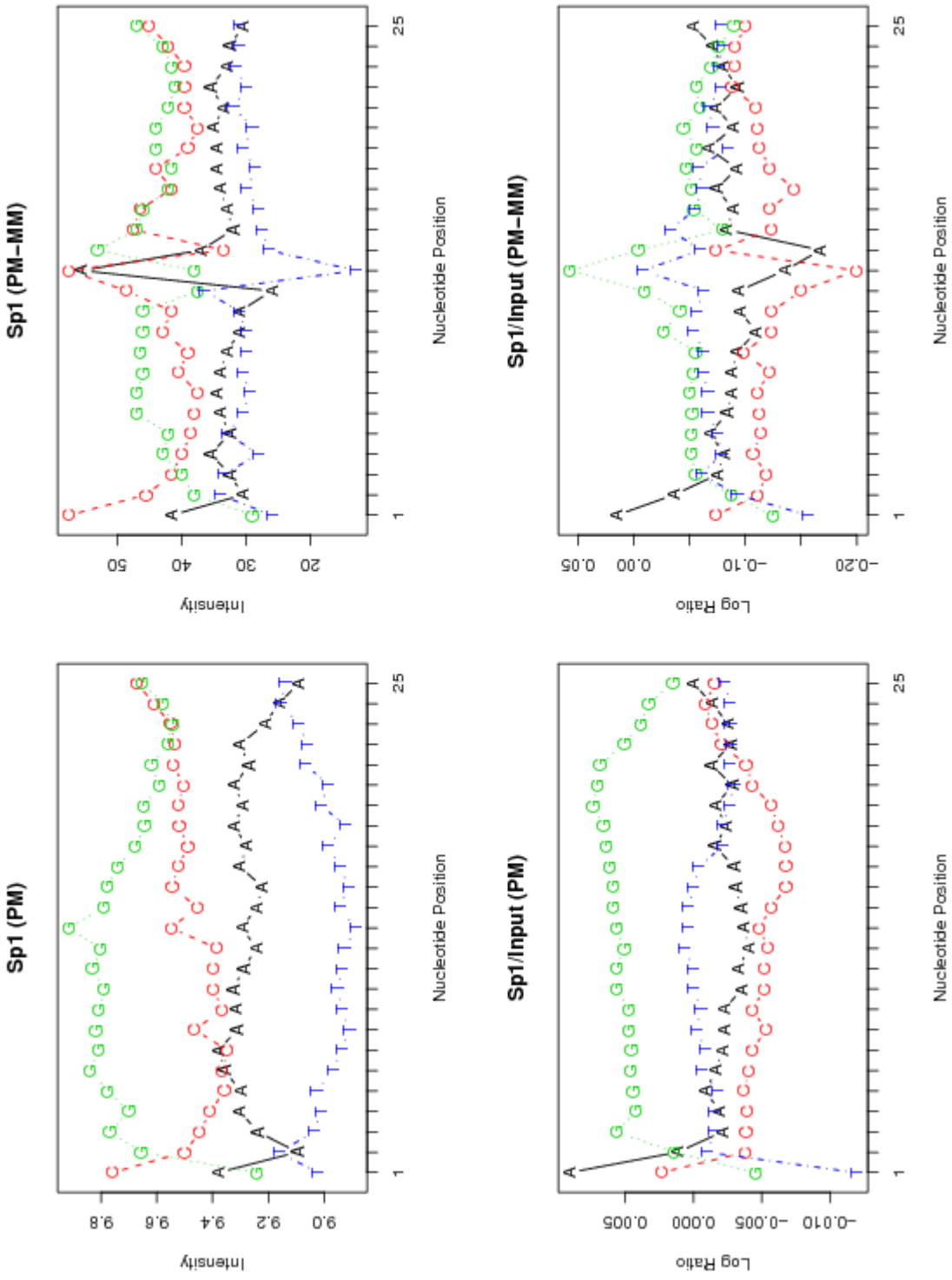

Figure 2: Position–specific quantile plots for Affymetrix ChIP–chip data.

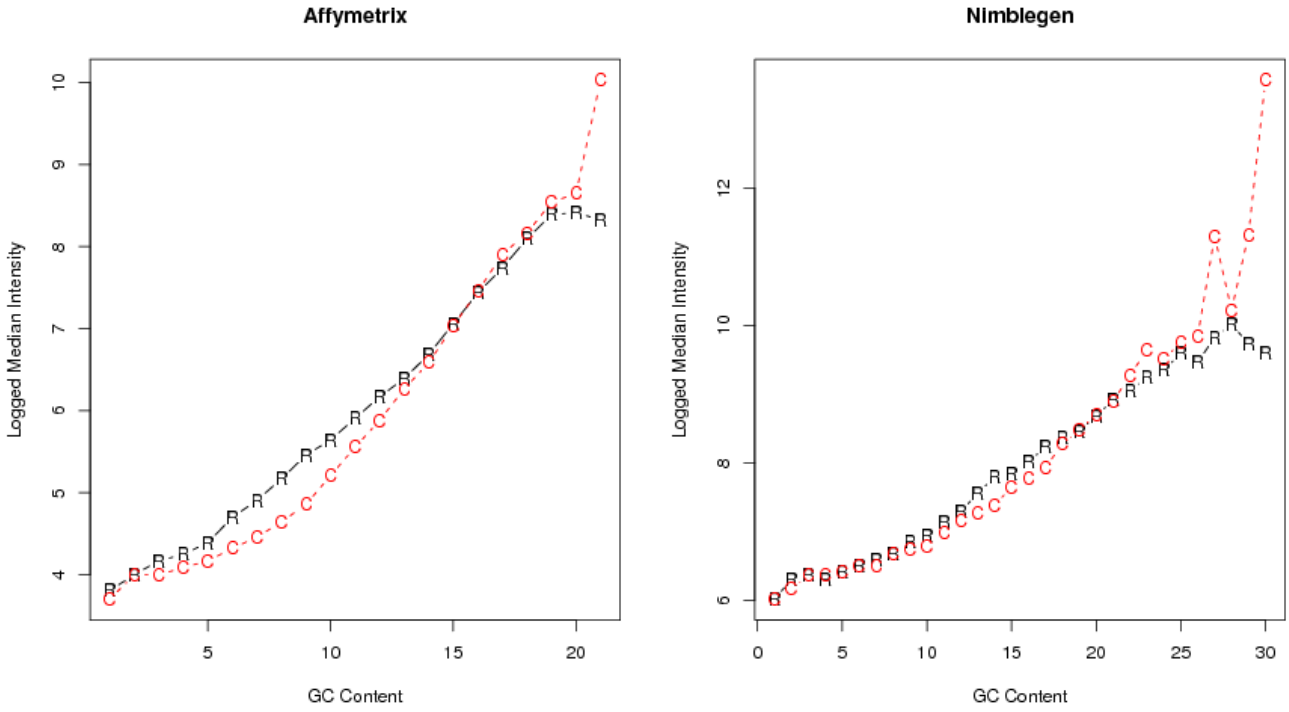

Figure 3: Probes targeting Refseq genes were binned by their GC content. The log of the median signal intensity was computed for each GC bin and plotted. Probes having identical GC content to the Refseq probes were isolated and plotted in the same fashion. The series labeled with Rs represents the Refseq probes while the series labeled with Cs represents control probes. Results for Affymetrix and Nimblegen NB4 transcription experiments are plotted.

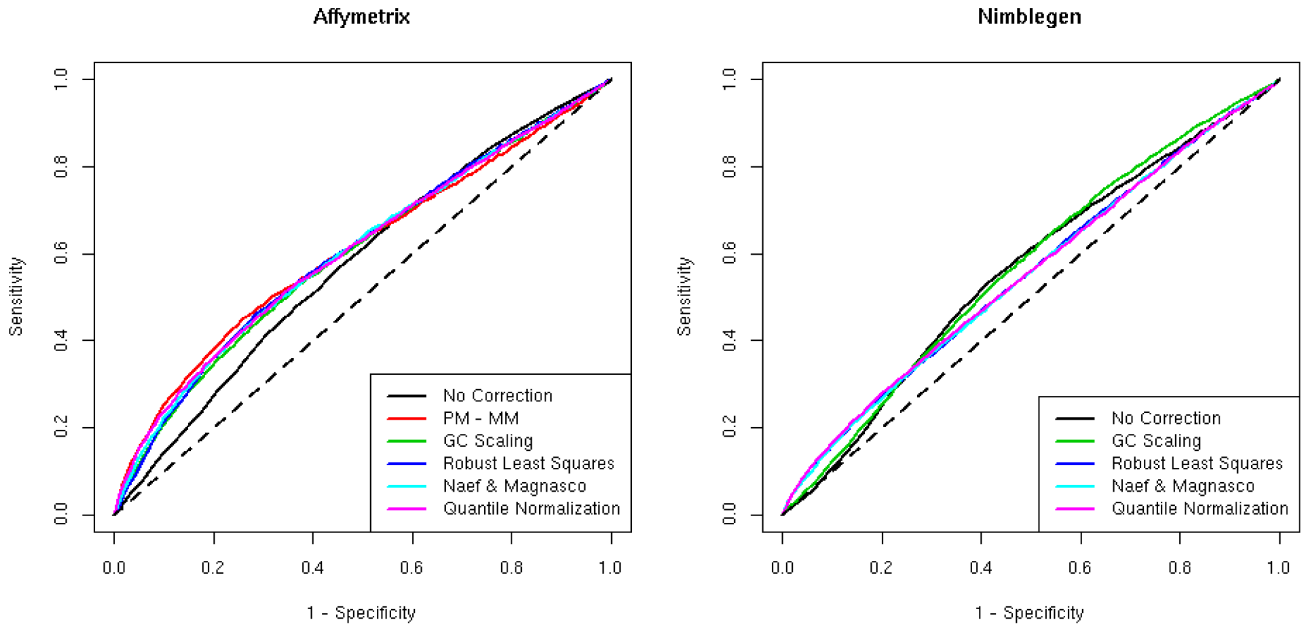

Figure 4: Sensitivity versus specificity plots for Affymetrix and Nimblegen tiling array data at the probe level. Known positives are taken to be features whose probes exactly match a Refseq sequence. Known negatives are features with the same GC content as the known positives but do not match a Refseq gene. For each normalization method, positives and negatives are obtained by simple thresholding. Sensitivity is the number of known positives with signals above a set threshold divided by the total number of known positives. Specificity is number of known negatives below the threshold divided by the total number of known negatives.