

Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2009 - 21st Annual Conference Proceedings

STATISTICAL METHODS FOR AFFYMETRIX TILING ARRAY DATA

Gayla Olbricht
olbrichtg@mst.edu

Nagesh Sardesai


Stanton B. Gelvin

Bruce A. Craig

R. W. Doerge

See next page for additional authors

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

 Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Olbricht, Gayla; Sardesai, Nagesh; Gelvin, Stanton B.; Craig, Bruce A.; and Doerge, R. W. (2009). "STATISTICAL METHODS FOR AFFYMETRIX TILING ARRAY DATA," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1080>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

Author Information

Gayla Olbricht, Nagesh Sardesai, Stanton B. Gelvin, Bruce A. Craig, and R. W. Doerge

STATISTICAL METHODS FOR AFFYMETRIX TILING ARRAY DATA

Gayla R. Olbricht¹, Nagesh Sardesai², Stanton B. Gelvin², Bruce A. Craig¹, and R.W. Doerge¹

¹Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2066 USA; ²Department of Biological Sciences, Purdue University, 915 W. State Street, West Lafayette, IN 47907 USA

Abstract

Tiling arrays are a microarray technology currently being used for a variety of genomic and epigenomic applications, such as the mapping of transcription, DNA methylation, and histone modifications. Tiling arrays provide high-density coverage of a genome, or a genomic region, through the systematic and sequential placement of probes without regard to genome annotation. In this paper we compare the Affymetrix tiling array to the Affymetrix GeneChip® 3' expression array and propose methods that address statistical and bioinformatic issues that accompany gene expression data that are generated from Affymetrix tiling arrays. Real data from the model organism *Arabidopsis thaliana* motivate this work and application.

Keywords: Microarray; tiling array; differential expression; ANOVA model

1. Introduction

Microarray technology is a powerful tool for studying large genomic regions, often a whole genome, in a single experiment. Different types of microarrays have been designed for a variety of applications and are commonly used to study gene expression. One type of microarray, the tiling array, offers the opportunity to study many different biological phenomena (e.g., differential expression, methylation status, etc.) using the same array design.

Tiling arrays are designed to cover entire genomic regions of interest (e.g., chromosomes) through the systematic selection of probes from one end of the region to the other. Probe selection is performed without reference to the genome annotation, as probes are not chosen within a certain type of genomic element (e.g., genes), but rather covering the entire region. It is the nature of their design that allows tiling arrays to enjoy a broad range of applicability. For example, they can be used for epigenomic applications to study DNA methylation and histone modifications, both of which may occur anywhere in the genome. They can also be used to identify transcription factor binding sites, investigate alternative splicing, and study gene expression (Mockler & Ecker, 2005). An understanding of both the biological aspects of the particular application and the design of the technology are essential for implementation of a meaningful statistical analysis of tiling array data. Here, we focus on the study of gene expression using an oligonucleotide tiling array commercially produced by Affymetrix.

The most common goal of gene expression studies is to measure the transcription level of annotated genes. Specifically, gene expression levels are then compared between different conditions of interest (e.g., treatment vs. control) to obtain a set of genes that are differentially expressed. Gene expression microarrays were developed for this purpose by selecting probes in regions of known genes. While gene expression arrays have been used to study differential gene

expression for many years, relatively few studies have focused on differential expression using tiling arrays (Naouar et al., 2009; Zeller et al., 2009; Ghosh et al., 2007). In the studies that do use tiling arrays to measure gene expression many focus on transcript mapping, where regions of transcription are identified through statistical models (e.g., Kapranov et al., 2002; Bertone et al., 2004; Huber et al., 2006). Tiling arrays are well-suited for this purpose since they offer dense genomic coverage in both annotated and un-annotated regions. This can lead to the identification of novel transcripts and can improve genome annotation.

Statistical issues inherent to data from gene expression arrays have been thoroughly investigated and many analysis methods (e.g., Kerr et al., 2000; Wolfinger et al., 2001; Bolstad et al., 2003; Irizarry et al., 2003; Smyth, 2004) are available through statistical packages such as R/Bioconductor (R Development Core Team, 2009; Gentleman et al., 2004). Therefore when using tiling arrays to study differential gene expression, it is important to learn from gene expression array methodology, while keeping design differences between microarrays and tiling arrays in mind. In this work, we compare the Affymetrix tiling array to the Affymetrix GeneChip® 3' expression array, a popular oligonucleotide gene expression array.

Arabidopsis thaliana is the model organism for all plants. Using *Arabidopsis thaliana* we investigate gene expression changes between a wild-type *Arabidopsis* (i.e., a control) and an over-expressing line of *Arabidopsis* by hybridizing the same mRNA samples to both Affymetrix GeneChip® *Arabidopsis* ATH1 Genome (3' expression) arrays and Affymetrix GeneChip® *Arabidopsis* Tiling 1.0R arrays. Differential expression analysis for tiling arrays is proposed through an initial bioinformatic step which allows the same statistical model to be used for both the tiling and gene expression arrays. A review and comparison of potential advantages and disadvantages of both technologies is given and differential expression results are compared.

2. Technology Overview

2.1 Gene Expression Review

The Central Dogma of molecular biology describes the process by which information contained in deoxyribonucleic acid (DNA) is used to produce proteins, which are the fundamental unit of cellular function. The Central Dogma states that DNA is transcribed to ribonucleic acid (RNA), and RNA is translated to protein (Crick, 1970). Specifically, messenger RNA (mRNA) is a special class of RNA that is responsible for encoding proteins (Griffiths et al., 2008). Microarrays can be used to measure mRNA transcription levels of genes through the hybridization of an mRNA sample with probes that are selected from a reference genome and placed as targets on the array. Thus the mRNA transcription levels measured on an array indicate which genes are active in making proteins.

A gene is comprised of exons and introns. Introns are regions within a gene that are removed in a process called RNA splicing. The remaining exon sequences are then joined together to form the mature mRNA. Sometimes variation in the splicing process results in different forms of an mRNA from the same gene. This phenomenon is known as alternative splicing, and can occur when an exon gets removed or an intron does not get removed in the splicing process. This

aside, mRNA typically arises from exons of genes (Griffiths et al., 2008). Because mRNA is the genetic material that is hybridized to microarrays in a gene expression study, it is imperative that there exist probes on the array that correspond to the exons of genes, as this is where cross-hybridization is expected to occur. A detailed look at the probe selection process is essential for understanding which probes are relevant to the study of gene expression.

2.2 Array Design: *Arabidopsis* ATH1 Array vs. Tiling 1.0R Array

Understanding the design of the Affymetrix ATH1 (3' expression) array and the Affymetrix tiling array is essential for developing statistical methods that test for differential expression. Both arrays utilize 25 base oligonucleotide probes. Each genomic sequence is represented by a probe pair which consists of a perfect match (PM) probe and a mismatch (MM) probe which differs only at the 13th base pair (Technical Note: GeneChip® Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis).

ATH1 and other Affymetrix 3' expression arrays are specifically designed to measure gene expression by selecting probes that cover exons of genes from the 3' end of transcripts (Figure 1A). Each gene is typically represented by 11-20 probes (called a probe set) that are chosen for their optimal hybridization quality (Technical Note: Array Design for the GeneChip® Human Genome U133 Set). Affymetrix provides a chip definition file (CDF) that connects probes to their corresponding probe sets. Probe sets can later be matched to the genes they represent, noting that some probe sets represent more than one gene. There are 22,810 probe sets represented by 251,078 probe pairs on the ATH1 array. Differential expression between two conditions is assessed for each probe set.

Testing for differential expression at each probe set results in thousands of hypotheses tests that are conducted simultaneously in a single experiment. For a single test, the probability of a Type I error (i.e., a false positive declaring a gene is differentially expressed when it truly is not) is controlled by setting the significance level (α). However, when all tests are considered together, the chance of at least one false positive increases with the number of independent tests being performed. This issue is known as the multiple testing problem and several procedures have been developed to control different variations of the Type I error rate for a set of simultaneous tests while also considering the power of the tests. Dudoit et al. (2003) offer a review of methods developed to address the multiple testing problem in the context of microarray experiments and Farcomeni (2008) gives a general extensive review of the issue.

Recall that tiling arrays are designed to cover an entire genomic region by systematically selecting probes from one end of the region to the other. Tiling array probes are not specifically designed to optimize the study of gene expression, but rather to provide dense, unbiased genomic coverage. The Affymetrix tiling array 1.0R for *Arabidopsis* covers the whole genome by placing probes along non-repetitive regions with an average gap of 10 base pairs between probes (Figure 1B) (Package Insert: GeneChip® *Arabidopsis* Tiling 1.0R Array). There are 3,039,991 million probe pairs which cover the five *Arabidopsis* chromosomes.

For tiling arrays, Affymetrix provides a binary probe mapping file (BPMAP) that identifies the genomic position and sequence of all probes. Unfortunately, the file does not indicate the corresponding genomic annotation of probes (i.e., which probes belong to which genes). Thus, it is unknown which probes correspond to genes or, specifically, which probes correspond to exons and introns (Figure 2A). Without this information, testing differential expression between two conditions is limited to testing each of the ~3 million probes individually. This is problematic since knowing whether or not an individual probe is differentially expressed does not give researchers the level of information they need. Furthermore, many of the probes that are tested are not of primary interest, since they correspond to introns or intergenic regions. Also, testing at a probe level basis increases the magnitude of the number of tests (~3 million probes vs. 22,810 probe sets for the ATH1 array) and this greatly affects the multiple testing problem. Connecting probes to their genomic annotation is crucial to conducting biologically relevant tests for differential expression in tiling arrays.

2.3 Annotation of Tiling Array

The probes on tiling array can be mapped to their genomic annotation using data from The *Arabidopsis* Information Resource (TAIR) website. Specifically, each probe is mapped to an exon, intron, or intergenic region of the TAIR8 genome (The Arabidopsis Information Resource, 2008). A large percentage (54.6%) of probes on the tiling array correspond to introns or intergenic regions. While these regions may contain useful information for studying alternative splicing or novel transcription, the focus of this work is to investigate differential gene expression in coding regions of annotated genes. Therefore, using only the probes in exons of genes (45.4% of probes) as probe sets (Figure 2B), differential expression can be assessed for each gene. In turn, testing exon probe sets greatly reduces the number of tests and leads to more biologically relevant results than probe level tests.

There are 31,391 genes that are represented by probes in exons on the Arabidopsis tiling array, covering 95% of TAIR8 genes (Figure 3). On average, there are 44 tiling array probes per gene. In comparison, the ATH1 Array has 22,810 probe sets with some of these corresponding to more than one gene. A total of 23,087 genes are represented on the ATH1 array, covering 70% of TAIR8 genes (Figure 3). On average, there are 11 ATH1 array probes in each probe set. There are 22,850 genes that are common to both arrays.

2.4 Summary: *Arabidopsis* ATH1 Array vs. Tiling 1.0R Array

When using microarrays to study gene expression, it is important to keep in mind that mRNA transcript accumulation is typically expected to occur in exons. While new regions of transcription or transcript variants will continue to be found, making use of the current genome annotation to obtain differential expression results for known genes is a common practical need for researchers. Since both ATH1 arrays and tiling arrays can be used for this purpose, a brief summary of their potential advantages and disadvantages is merited.

ATH1 arrays are specifically designed to study gene expression through the selection of probes with optimal hybridization quality/ability within exons of genes. In contrast, tiling array probes are selected to provide unbiased, dense genomic coverage. Due to this discrepancy in coverage

density, tiling arrays have on average 4 times more probes per gene than ATH1 arrays. Thus, there is a trade-off between probe hybridization quality/ability and amount of coverage per gene on the two arrays.

One convenient feature of ATH1 arrays is the availability of the previously mentioned CDF file that connects probes on the array to their corresponding probe sets. The TAIR website provides additional data that links probe sets to genes in the current genome annotation version (The Arabidopsis Information Resource, 2008). However, ATH1 arrays are designed by selecting probes in known genes available as of December 2001 in The Institute for Genome Research (TIGR) database (Data Sheet: GeneChip® Arabidopsis ATH1 Genome Array). Any subsequent new information from more recent genome versions is not incorporated into the probe design. Therefore, if a new gene is discovered in the *Arabidopsis* genome, it will not have probes representing it on the ATH1 array. Alternatively, tiling array probes are based on the sequence of the TIGR5 genome version which was completed in 2004 (Package Insert: GeneChip® Arabidopsis Tiling 1.0R Array). Since probes are selected to cover the whole genome, it is possible that newly discovered genes will be represented on the array by probes that were previously thought to be intergenic. However, since Affymetrix does not provide a file that connects probes on the tiling array to their current genomic annotation, this connection must be completed by using data from the TAIR website (The Arabidopsis Information Resource, 2008). Completion of this annotation reveals that the tiling array represents 25% more TAIR8 genes than does the ATH1 array. Thus there is an additional trade-off between ease of obtaining annotation information and gene coverage on the two arrays.

3. Differential Expression Analysis of Affymetrix Tiling Array Data

3.1 Gene Level Model for Differential Expression

If we consider a differential expression study based on annotated genes, one goal is to determine for each gene whether or not there is a significant difference in expression levels between conditions (e.g., treatment vs. control). Whereas this is a common application of ATH1 arrays, it is only with the availability of genomic annotation for the tiling array that it is possible to conduct such gene level tests for differential expression. Recall that the first step in conducting such an analysis for tiling arrays is to identify the probes that are biologically relevant. This is accomplished by filtering out probes inside introns and intergenic regions, while retaining probes covering exons (Figure 2B). This gives probe sets corresponding to 31,391 genes for the tiling array. Because both ATH1 and tiling arrays now have data for each gene in the form of probe sets, the same statistical model can be applied to both array types.

Drawing from methodology in the 3' expression array literature (for a review see Craig et al., 2003), the following differential expression analysis is conducted for both expression arrays and tiling arrays. The arrays are first pre-processed by performing a background correction and normalization of variation across arrays. Specifically, a robust multi-array analysis (RMA) background correction (Irizarry et al., 2003) and quantile normalization (Bolstad et al., 2003) are performed on the PM intensities, setting the distribution of all arrays to be the same. An analysis of variance (ANOVA) model is employed to detect probe sets which are differentially expressed

between two treatment groups using the natural log of the background corrected, normalized data as the gene expression level. The following ANOVA model (1) is fit for each probe set and is similar to the two-step approach employed by Wolfinger et al. (2001) and extended by Chu et al. (2002) for Affymetrix arrays:

$$y_{ijk} = \mu + T_i + P_j + (TP)_{ij} + \varepsilon_{ijk}; i = 1, 2; j = 1, \dots, p; k = 1, \dots, n \quad (1)$$

where y_{ijk} is the gene expression level for the k^{th} replicate of probe P_j under treatment T_i , μ is the average gene expression level over all probes, treatments and replicates, T and P are the treatment and probe main effects, TP is the interaction between treatment and probe, and ε_{ijk} are independent errors which are normally distributed with mean 0 and variance σ^2 .

To determine if there is a statistically significant difference in expression between two (treatment) groups, the following hypotheses, based on the treatment effect, are tested for each probe set:

$$H_o : T_1 - T_2 = 0 \quad \text{vs.} \quad H_a : T_1 - T_2 \neq 0 \quad (2)$$

The test statistic is:
$$\frac{\bar{Y}_{1..} - \bar{Y}_{2..}}{\sqrt{\frac{2 * MSE}{3p}}} \sim t_{4p} \quad \text{under } H_o \quad (3)$$

where the mean squared error (MSE) and number of probes (p) from model (1) will differ for each probe set.

Two approaches are employed to adjust for multiple testing. The Holm adjustment controls the familywise error rate, which is the probability of making at least one false discovery among the probe set level tests (Holm, 1979). Benjamini and Hochberg's method controls the false discovery rate (FDR), which bounds the expected rate of false discoveries (Benjamini & Hochberg, 1995). Holm's procedure is more conservative than that of the FDR approach.

3.2 Application to *Arabidopsis thaliana* data

Data from an *Arabidopsis thaliana* study are used to demonstrate the application of tiling arrays for studying differential expression, as well as to compare tiling and ATH1 array results. In this study, the consequences of over-expression of a myb transcription factor (MTF) gene are investigated. Certain mutations to the MTF gene increase the plant's susceptibility to Agrobacterium-mediated transformation (i.e., allowing the transfer of foreign DNA from Agrobacterium to the plant; Gelvin, 2003) in *Arabidopsis thaliana*. Two different MTF mutants (*hat3* and *mtf2*), a MTF over-expressing line (Myb4), and wild-type Columbia (Col-0) are studied. Gene expression is measured via hybridizing samples of mRNA from the root tissue of the four different sample types to both Affymetrix GeneChip® *Arabidopsis* Tiling 1.0R Arrays and Affymetrix GeneChip® *Arabidopsis* ATH1 Genome Arrays. The same mRNA samples are hybridized to both types of arrays. Two of these sample types, Col-0 and Myb4, are examined here for illustration purposes. Three biological replicates of each of the two sample types are

measured, yielding a total of 6 arrays of each type. The goal is to identify differentially expressed genes between Col-0 and Myb4.

Since the same biological samples are hybridized to both array types, variation in results should be due to the technological differences between the arrays or other experimental factors, such as RNA degradation, rather than due to biological differences in the samples. Thus, a comparison of the results can help reveal similarities and differences between the two types of arrays. The same ANOVA model (1) is applied to data from both array types where the treatment effect is the sample type (Col-0 or Myb4). The hypotheses (2) are tested via the test statistic (3) for each probe set. This will test for differential expression between Col-0 and Myb4 at each probe set.

On the ATH1 array, the FDR and Holm's procedures identified 4228 and 660 significant differentially expressed probe sets, respectively, at $\alpha=0.05$ (Figure 4A). On the tiling array, 2285 and 510 probe sets showed significant differential expression using FDR and Holm's (Figure 4B). Figure 4(A & B) shows the average log fold change of each probe set for both arrays, with probe sets that are not significant in grey, probe sets significant using the FDR procedure in blue, and probe sets significant with both FDR and Holm's in red. A positive log fold change indicates up-regulation (higher expression) in Col-0 than in Myb4 and a negative log fold change is indicative of down-regulation (lower expression) in the Col-0 sample. Note that the probe sets in the ATH1 graph (Figure 4A) are not ordered since some probe sets correspond to more than one gene; whereas each probe set on the tiling array (Figure 4B) corresponds to one gene and can be ordered by the gene's position on the chromosome. The tiling array identified almost half as many differentially expressed probe sets using the FDR procedure as the ATH1 array, with the majority of significant probe sets demonstrating up-regulation and a loss of significant down-regulation compared to the ATH1 results (Figure 4 A & B). Note the presence of gaps in significant up-regulation accompanied by significant down-regulation in centromeric regions of each chromosome in the tiling array results (Figure 4B).

To compare the results of differential expression in terms of genes rather than probe sets, the 22,850 genes (Figure 3) that are present on both arrays are investigated. Figure 5A shows a comparison of the number of significant differentially expressed genes found with both array types, using genes represented on both arrays. While many of the same significant genes are identified using both array types (1046 with FDR; 199 with Holm's), there are also many genes uniquely identified as significant by one of the arrays but not the other. Finally, since some genes are represented on one array but not the other, it is important to note that a number of these genes are also found significant (Figure 5B).

To compare the similarity of the two arrays in terms of hybridization intensities, the average log fold change for genes represented on both arrays is examined. If both arrays are performing similarly at the gene level, it is expected that the average log fold change for a particular gene will be similar on both arrays and thus follow a 45° line if plotted against each other (Figure 6). Results using the FDR procedure are highlighted in different colors (Figure 6). Significant genes on both arrays (blue points) are clearly further from zero and tend to follow the 45° line, with a noticeable larger number of genes in the upper right quadrant than the lower left quadrant,

meaning that these genes have a positive log fold change and are up-regulated in Col-0 on both arrays. However, points in the upper left and lower right quadrants are genes that differ in the sign of their log fold change between arrays. For example, genes in the lower right quadrant have a positive log fold change in the tiling array, but have a negative log fold change in the ATH1 array. Observing the (orange) points in that quadrant which are genes identified as significant (with a negative log fold change) on the ATH1 array only, as well as several (green) points which are genes identified as significant (with a positive log fold change) on the tiling array only, can help explain why many more down-regulated genes are identified in the ATH1 analysis than in the tiling analysis.

In addition to the average log fold change, there are also two other quantities that affect the significance of a gene. The number of probes (p) and mean squared error (MSE) per probe set also affect the test statistic (3) for differential expression. Recall that the tiling array has an average of 44 probes per probe set, giving it an advantage over the ATH1 array which has an average of 11 probes per probe set. However, the ATH1 array has a much smaller MSE per probe set on average (0.256) than the tiling array (0.884). This comparative reduction in variation in the ATH1 array may be due to the ATH1 probe selection process for optimal hybridization quality.

In summary, several genes are identified as differentially expressed using both arrays and may be of interest for further study. However, even though the same biological samples are hybridized to both array types, there are many discrepancies in results. Some differences are expected due to the design differences in the two arrays. However, it is difficult to say which array gives more accurate results, since neither gives a perfect measure of gene expression. What can be said is that tiling arrays typically have more error degrees of freedom while ATH1 arrays have less variation.

4. Summary and Future Work

Tiling arrays are a flexible type of microarray that can be used for many different applications. In this work, we focus on one application (differential expression analysis in annotated genes) in one type of tiling array (Affymetrix GeneChip® *Arabidopsis* Tiling 1.0R Array). Gene level results are a practical need for researchers using tiling arrays to study differential expression. To this end, the Affymetrix *Arabidopsis* tiling array is annotated to the TAIR8 genome by identifying whether each probe corresponds to an exon or intron within a gene or to an intergenic region. This annotation enables the selection of biologically relevant probes (exons in genes) for gene level differential expression tests.

Real data are presented where the same biological samples are hybridized to both the Affymetrix *Arabidopsis* tiling array and the ATH1 (3' expression) array, which has been widely used to study differential expression in the past. The same pre-processing techniques and ANOVA model are applied to data of both array types and results between two different sample types are compared. While many genes are found to show significant differential expression with both arrays, the ATH1 array identified almost twice as many significant genes using a false discovery rate correction than the tiling array. While this study alone cannot offer conclusive evidence to

explain the discrepancy of these results, it is important to consider a few things when deciding whether to use ATH1 or tiling arrays to study differential expression in *Arabidopsis*.

ATH1 arrays are designed specifically for the purpose of studying gene expression, while tiling arrays have broad applications. More work is needed to better understand the use of tiling arrays for studying differential expression of known genes, and, until then, many researchers may find ATH1 arrays more suitable for their needs. However, using tiling arrays to study differential expression may be useful if the researcher is interested in studying gene(s) that are not represented on the ATH1 array, or if combining results with other biological phenomena (e.g., DNA methylation) is of interest. Ultimately, the researcher must examine the goals of the study and consider these factors when deciding on which type of array to use for studying differential expression.

Finally, once the tiling array probe sets are found by mapping probes to the TAIR8 genome, the statistical model presented here for gene level differential expression tests is relatively simple. Possible model improvements that are currently being explored include the incorporation of a fixed exon effect and a random array effect, as well as the implementation of a variance pooling method. However, more work needs to be done to investigate statistical issues present in the data. This said, since tiling arrays are used in many different applications, statisticians have ample opportunities to contribute to the modeling of tiling array data.

Acknowledgments

We thank the RWD research group for providing helpful feedback and the system administrators (Doug Crabill and My Truong) in the Department of Statistics, Purdue University for their invaluable computing assistance. This work was funded by a NSF Plant Genome grant to RWD (DBI-0733857)

References

(2008, May). Retrieved July 2008, from The Arabidopsis Information Resource:
<http://www.arabidopsis.org/>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society (Series B)*, 57, 289-300.

Bertone, P., Stolc, V., Royce, T., Rozowsky, J., Urban, A., Zhu, X., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306, 2242-2246.

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

Chu, T., Weir, B., & Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences*, 176, 35-51.

Craig, B., Black, M., & Doerge, R. (2003). Gene expression data: The technology and statistical analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, 8, 1-28.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227, 561-563.

Data Sheet: GeneChip® Arabidopsis ATH1 Genome Array. (n.d.). Retrieved 2009, from Affymetrix: http://www.affymetrix.com/support/technical/datasheets/arab_datasheet.pdf

Dudoit, S., Shaffer, J.P., & Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18, 71-103.

Facromeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17, 347-388.

Gelvin, S.B. (2003). *Agrobacterium* and plant transformation: The biology behind the “gene-jockeying” tool. *Microbiology and Molecular Biology Reviews*, 67, 16-37.

Gentleman, R.C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.

Ghosh, S., Hirsch, H. A., Sekinger, E. A., Kapranov, P., Struhl, K., & Gingeras, T. R. (2007). Differential analysis for high density tiling microarray data. *BMC Bioinformatics*, 8, 359.

Griffiths, A., Wessler, S., Lewontin, R., & Carroll, S. (2008). *Introduction to Genetic Analysis*. W.H. Freeman and Company.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

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

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 249-264.

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

Kerr, M., Martin, M., & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7, 819-837.

Mockler, T. C., & Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85, 1-15.

Naouar, N., Vandepoele, K., Lammens, T., Casneuf, T., Zeller, G., van Hummelen, P., et al. (2009). Quantitative RNA expression analysis with Affymetrix tiling 1.0R arrays identifies new E2F target genes. *The Plant Journal*, 57, 184-194.

Package Insert: GeneChip® Arabidopsis Tiling 1.0R Array. (n.d.). Retrieved 2009, from Affymetrix:
https://www.affymetrix.com/support/downloads/package_inserts/tiling_arabidopsis_insert.pdf

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org>

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3, Article 3.

Technical Note: Array Design for the GeneChip® Human Genome U133 Set. (n.d.). Retrieved 2009, from Affymetrix:
http://www.affymetrix.com/support/technical/technotes/hgu133_design_technote.pdf

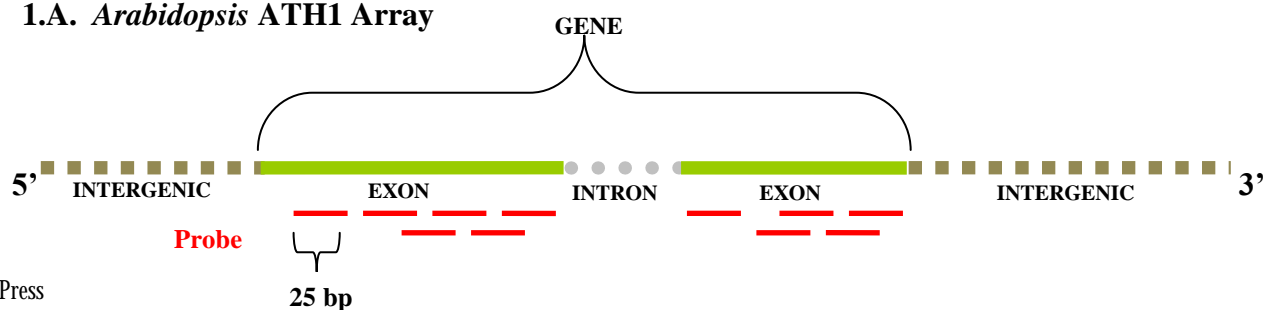
Technical Note: GeneChip® Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis. (n.d.). Retrieved 2009, from Affymetrix:
http://www.affymetrix.com/support/technical/technotes/25mer_technote.pdf

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8, 625-637.

Zeller, G., Henz, S. R., Widmer, C. K., Sachsenberg, T., Ratsch, G., Weigel, D., et al. (2009). Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. *The Plant Journal*, 58, 1068-1082

Figures

1.A. Arabidopsis ATH1 Array



1.B. *Arabidopsis* Tiling Array

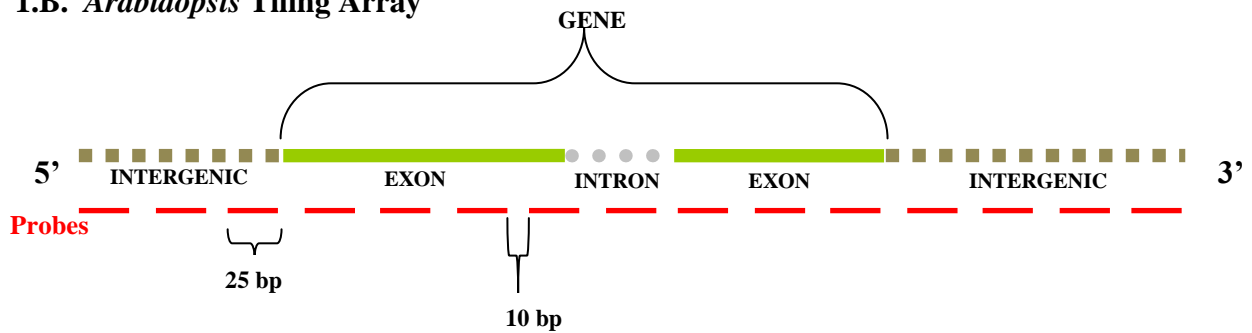
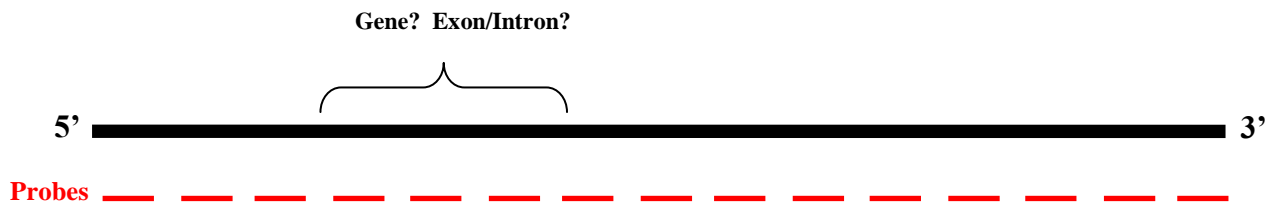


Figure 1. A. Example of probes covering a gene on the Affymetrix *Arabidopsis* ATH1 (3' expression) array. The 25 base pair probes only cover the exons of genes and can be overlapping. **B.** Example of probes covering a genomic region on an Affymetrix *Arabidopsis* tiling array. The 25 base pair probes cover exons, introns, and intergenic regions with an average gap of 10 base pairs between probes.

2.A. Tiling array without annotation



2.B. Tiling array without annotation

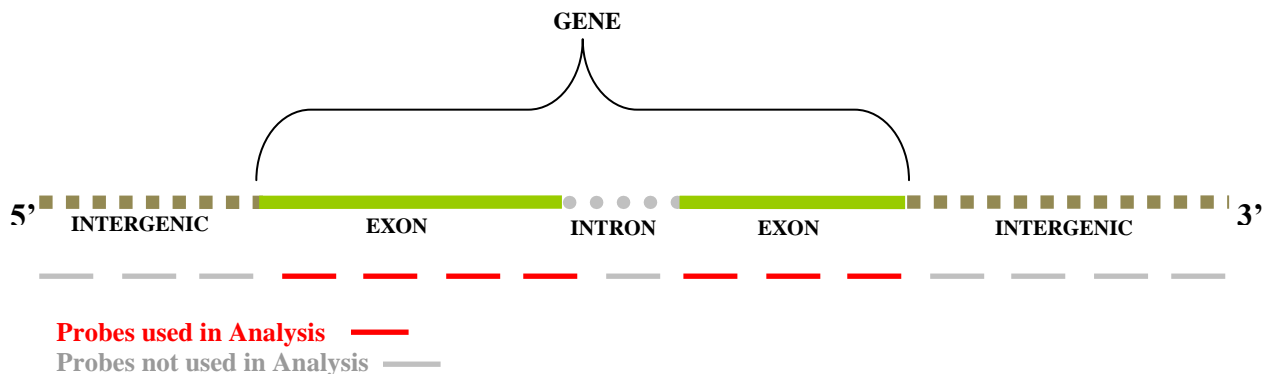


Figure 2. A. Without annotating the tiling array, it is known where probes are in the genome, but information about whether probes correspond to exons or introns within genes or to an intergenic region is not available. **B.** Once the tiling array has been annotated, biologically relevant probes in a genomic region can be identified and used for differential expression analysis. Red probes correspond to exons and are used in further analysis; whereas light grey probes correspond to introns and intergenic regions and are not retained in the analysis. The set of red probes for this gene are considered to be a probe set.

Number of Genes Represented on Array

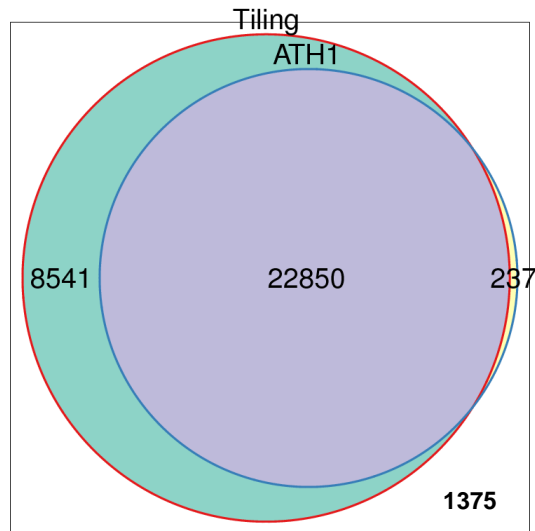
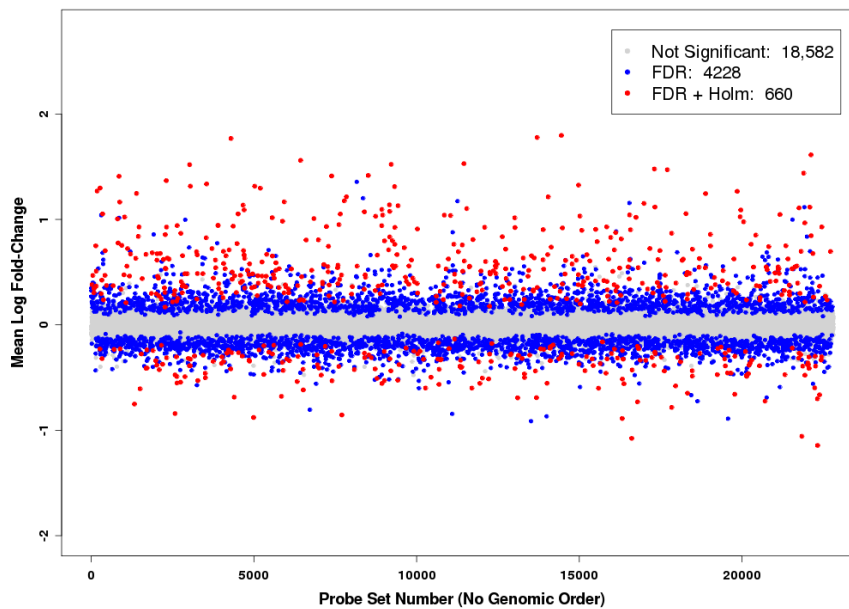


Figure 3. Proportional Venn diagram comparing the number of genes represented on the Affymetrix *Arabidopsis* Tiling and ATH1 arrays. The area of each region is proportional to the number of genes in the set. There are 22,850 genes that are covered on both arrays; 8,541 that are only covered on the tiling array; 237 that are only covered on the ATH1 array; and 1,375 that are not covered on either array.

4.A. ATH1 Array Results



4.B. Tiling Array Results

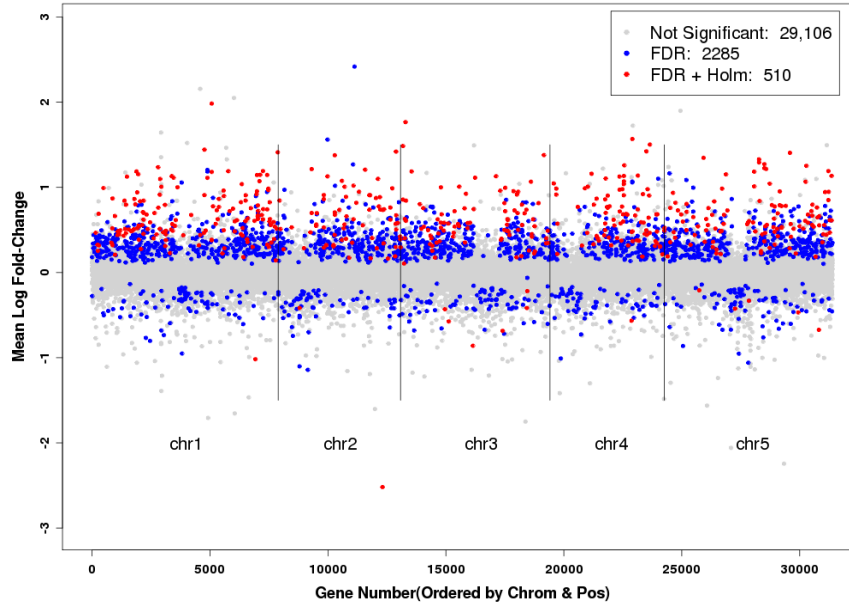
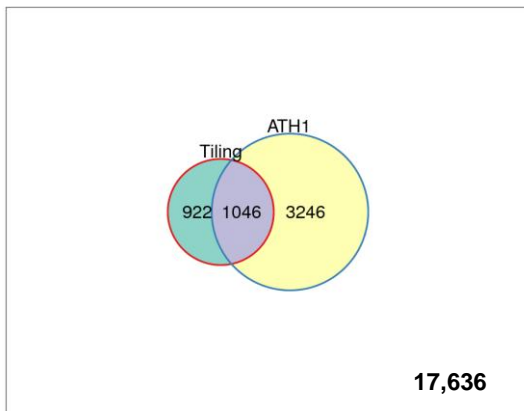


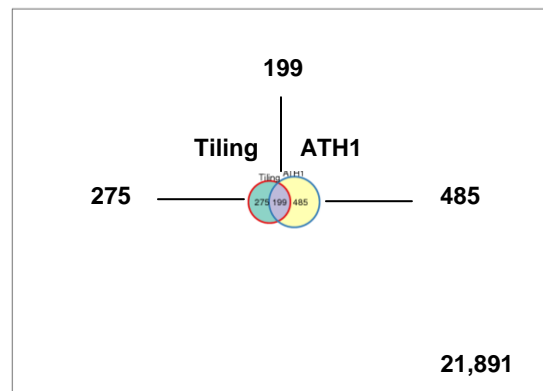
Figure 4. A. Mean log fold change vs. probe set number for the ATH1 array. Note that the probe set numbers do not correspond to genomic order since some probe sets correspond to more than one gene. **B.** Mean log fold change vs. gene number for the tiling array. The gene numbers are ordered by chromosomal position. For both graphs, probe sets that are not significant are shown in grey, probe sets significant with FDR only are in blue, and probe sets significant with both FDR and Holm’s are in red. The numbers in the legend correspond to the number of probe sets which correspond to each of those groups.

5.A.

FDR Significant Genes



Holm’s Significant Genes



5.B.

Col-0 vs. Myb4	FDR	Holm
Tiling Only	317	36
ATH1 Only	37	5

Figure 5. A. Proportional Venn diagrams comparing the number of significant genes found by the tiling and ATH1 arrays, using FDR and Holm’s procedures at $\alpha=0.05$. The area of each region is proportional to the number of genes in the set. For the FDR results, there are 1046 genes that are identified as significant using both arrays; 922 that are only significant on the tiling array; 3246 that are only significant on the ATH1 array; and 17,636 that are not significant on either array. The Holm’s results can be interpreted likewise with numbers enlarged for readability since the area outside the circles is large due to the number of genes that were not significant on either array. **B.** Number of significant differentially expressed genes represented only on the tiling or ATH1 array.

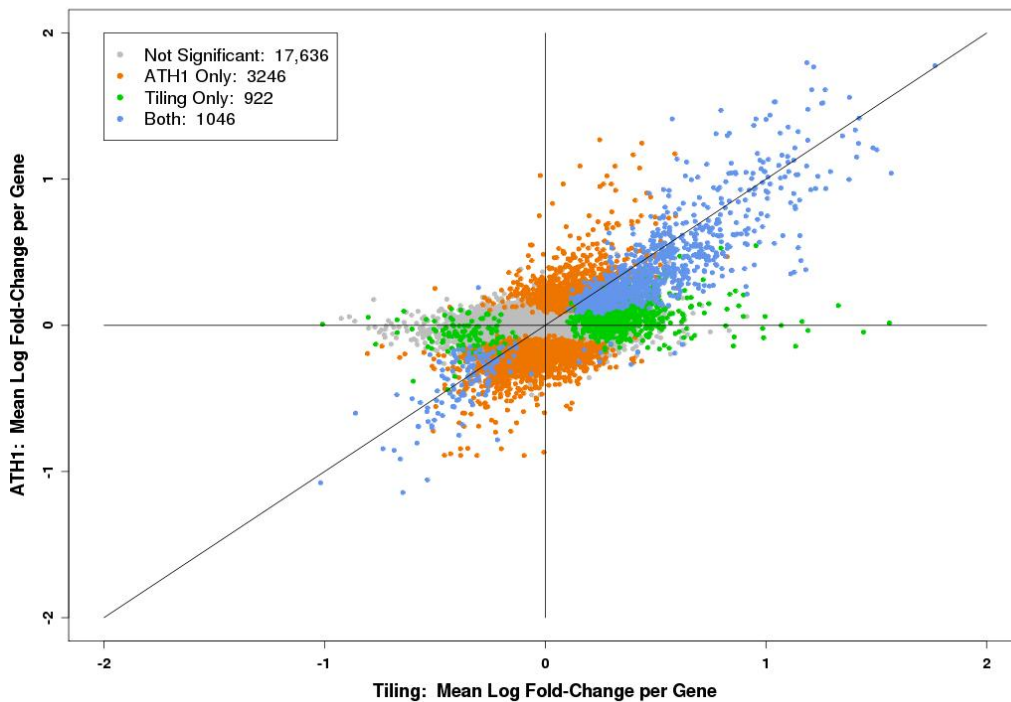


Figure 6. Mean log fold change per gene for genes represented on both the ATH1 and tiling arrays. FDR results are shown as grey (non-significant), orange (significant in ATH1 only), green (significant in tiling only), and blue (significant in both) points. The 45° line is for comparison purposes.