Arch Oncol 2005;13(3-4):124-5.
DOI: 10.2298/AOO0503124T

Microarrays: Design, analysis, and applications in functional genomics

Miloš Tanurdžić[1], Rebecca Doerge[2], Robert Martienssen[1]

[1]Cold Spring Harbor Laboratory and Watson School of Biological Sciences, Cold Spring Harbor, NY, [2]Purdue University, Department of Statistics, West Lafayette, IN, USA, Address correspondence to: Miloš Tanurdžić, Cold Spring Harbor Laboratory and Watson School of Biological Sciences, Cold Spring Harbor, NY11724, USA, E-mail: milos@cshl.edu, The manuscript was received: 15. 09. 2005, Accepted for publication: 31. 10. 2005

ABSTRACT

The last decade provided biologists with wealth of DNA sequence information. This rich resource is now being utilized to create microarrays as a platform for investigating gene expression profiles as well as epigenetic DNA modifications and DNA-protein interactions on the genome-wide scale. An overview of the microarray technology and its applications in basic and applied life sciences will be presented as well as the statistical and bioinformatic approaches taken to analyze large amounts of quantitative information produced during a typical microarray experiment.

KEY WORDS: *DNA; Gene Expresion Profiling; Genomics; Computational Biology*

MICROARRAYS: TECHNOLOGY

BDNA microarray technology has become a ubiquitous tool of functional genomics, allowing researchers to obtain data about hybridization signal intensity for thousands of DNA probes in a single experiment. In its original incarnation, the goal of a microarray experiment was to measure the fold change in signals that come from the two fluorescently labeled mRNAs from different treatment conditions, or from differentially labeled genomic DNA. This experimental approach uses arrays of thousands of DNA elements (either oligonucleotides or longer fragments of DNA, usually cDNA clones or PCR products) deposited onto a carrier surface in an ordered fashion. These DNA probes can be either spotted onto a solid support, usually glass slides, or in the case of oligonucleotides, they are most commonly synthesized in situ, using proprietary technologies, which often leads to higher cost for this type of microarrays. After hybridization, reactions on microarrays reach their equilibria the arrays are then scanned and the amount of fluorescence that corresponds to each of the used dyes can be quantified for each of the probes.

Measuring the steady-state levels of mRNA molecules as a direct measurement of transcription was only the first microarray application albeit still the most widely utilized microarray experiment. Comparative genome hybridization, which followed suit, is used to measure differences in gene copy number between two genomes and has been an extremely valuable tool in cancer genomics (1,2). In addition to these, several novel applications have been developed, such as sequencing by hybridization and SNP genotyping as well as a combination of chromatin imunoprecipitation and microarray analysis referred to as ChIP on chip used in discovery of transcription factor DNA binding sites throughout a genome.

All of these approaches have also had their counterparts developed for high throughput analyses in the filed of epigenetics, and nowadays microarrays are being used to identify genomic regions that have undergone cytosine methylation or, to identify regions of chromosomal DNA associated with histones carrying particular modifications (ChIP on chip). We have taken advantage of the availability of the complete genome sequence of Arabidopsis thaliana to construct a tiling microarray of chromosome 4. This microarray contains ~22000 1kb DNA fragments which cover most of the chromosomal sequence in a tiled (overlapping) fashion, except for most of the centromeric region of this chromosome. We have used this resource to ask questions about the distribution of DNA and histone methylation along this chromosome (Figure 1), their co-dependence as well the extent of natural variation and heritability of cytosine methylation in higher eukaryotes. In addition, we have used this platform to investigate the roles of major chromatin modifying factors in gene regulation and genome biology by measuring the amount and distribution of modified cytosines and histones in corresponding mutant backgrounds (3).

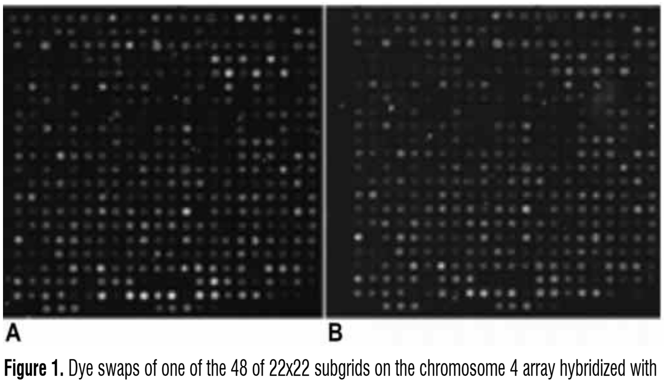

Figure 1. Dye swaps of one of the 48 of 22x22 subgrids on the chromosome 4 array hybridized with unmethylated DNA in the green channel (A) and the red channel (B). The contrast signals can be observed (green in one is red in the other microarray)

MICROARRAYS: STATISTICAL SIGNIFICANCE

An obvious statistical hurdle encountered during microarray analysis is the question of whether an observed fold change is "real" (i.e., significant), or whether it is a result of inherent variability in the experimental procedures. The estimation of this variability, therefore, is a key component in any statistical analysis. A number of experimental factors contributes to the total variability that is of interest. In order to determine whether observed intensity fold changes actually represent real changes in gene expression between treatment conditions, it is necessary to separate unwanted experimental effects from the signals of interest. Generally, the raw intensity readings are background subtracted with the additional correction made via a data normalization procedure, which attempts to standardize the data to allow comparisons between channels (fluors) on an array, as well across multiple arrays.

In an effort to overcome the difficulties of data normalization and analysis, a linear model based approach to the design and analysis of microarray experiments has been developed (4). This approach advocates identifying experimental factors, which introduce variability into the data, and designing array experiments, which allow the statistical estimation of these factors. In brief, our approach has identified the need for replicated spots on the arrays, as well the need for technical and biological replicates in order to establish statistically significant fold changes using either per array or per gene variances. By incorporating these potential sources of variation into an analysis of variance (ANOVA) model, data analysis can be performed without having to normalize the data a priori. The use of such models in conjunction with appropriate experimental designs maximizes the amount of information that can be extracted from the microarray experiment.

Following robust LOWESS normalization when needed, we applied a simple one-way ANOVA linear model in which a control set of 340 tiles on the array known to have not undergone any cytosine methylation from prior pilot experiments are used to establish the baseline and following a simple linear regression model:

$$\tilde{y} = \mu + A_i + D_j + T_k + G_g + AG_{ig} + DG_{jg} + TG_{kg} + \varepsilon_{ijkg}$$

with the null hypothesis:

$$H_0 \div \left(T_1 + TG_{1g}\right) - \left(T_2 + TG_{2g}\right) = 0$$

yielded information about statistically significantly methylated tiles along chromosome 4. Given the multiple testing conundrum created by having to perform ~22000 t-tests on this dataset we routinely adjust our significant p-values by applying Holm's and FDR multiple testing corrections. As an illustration, Figure 2 shows the distribution and magnitude of statistically significant methylated features along the chromosome 4 with the two multiple testing corrections and using either per array or per gene variance.

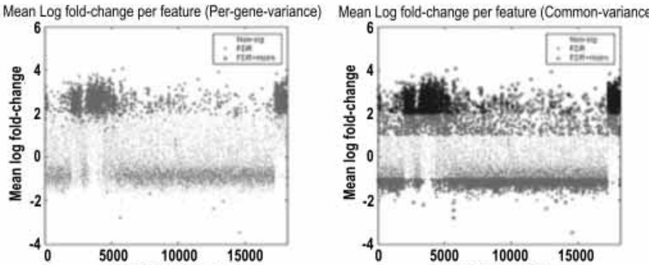

Figure 2. The distribution of statistically significant methylation along chromosome 4 of Arabidopsis. Note the dense methylation around position 3000, which corresponds to a heterochromatic knob *hk4S* known to harbor heavily methylated transposable elements (3)

MICROARRAYS: BIOINFORMATICS

Data manipulation, visualization, and proper data storage appropriate for remote access and storage represent another set of challenges for microarray experiments and data analysis. In recent years many software packages have been developed with this purpose in mind, often relying heavily on community-curated sequence databases. While these options provide users with easy access to their data and a plethora of analysis options, they are often rigid in their implementations allowing very little input into the finer details of data analysis. In addition, they are often hard to implement in cross-platform settings with several different types of data needed to be visualized. With that in mind, a collaborative effort was first spearheaded by the *C. elegans* genomics community and later expanded into a generic genome browser developed by Stein et al. (5). The major benefit of using the GB is the ease with which different data can be visualized and queried within the browser all within the framework provided by the genome sequence of the chosen organism. In our projects, the GB is used to bring together microarray data from expression studies, DNA and histone methylation profiling, small RNA profiling and comparative genome hybridizations (3). All of these data are anchored onto the genome sequence and presented as different data tracks, allowing for easy and intuitive visual inspection. In addition, the GB allows simultaneous representation of not only the qualitative component of the dataset but also the quantitative one, showing intensities and statistical significance of gene expression of methylation status, for example (Figure 3).

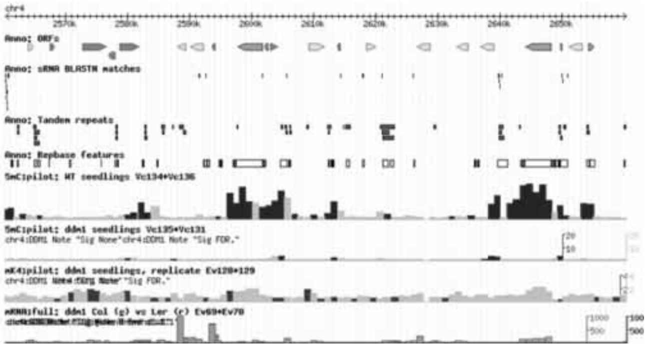
chr4
257k
258k
259k
260k
261k
262k
263k
264k
265k
Anno: ORFs
Anno: sRNA BLASTN matches
Anno: Tandem repeats
Anno: Repbase features
5mC1pilot: WT seedlings Vc134+Vc136
5mC1pilot: ddm1 seedlings Vc135+Vc131
chr4:DDM1 Note "Sig None"chr4:DDM1 Note "Sig FDR."
20
10
mK4ipilot: ddm1 seedlings, replicate Ev128+129
chr4:DDM1 Note "Sig FDR."
mRNA:full: ddm1 Col (g) vs Ler (r) Ev69+Ev70
1000
500
100
500

Figure 3. An example of the GB view of 100kb of chromosome 4 sequence. In addition to genes and transposons annotated along this sequence, the GB is capable of presenting data regarding multiple microarray experiments as separate tracks (5mC track present significantly methylated tiles in maroon, while the next track down shows that in ddm1 mutant most of cytosine methylation is gone; mRNA track shows RNA expression levels along the chromosome in the ddm1 mutant) as well as various sequence features, such as tandem repeats and small RNAs.

Acknowledgments

The work described here has been funded by NSF and USDA and was performed in collaboration with Z. Lippman, M. Vaughn, and H. Jiang.

REFERENCES

1. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. Applications of DNA tilling arrays for whole-genome analysis. Genomics 2005;85(1):1-15.
2. Russo G, Zegar C, Giordano A. Advantages and limitations of microarray technology in human cancer. Oncogene 2003;22:6497-507.
3. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature 2004;430(6998):471-6.
4. Craig BA, Vitek O, Black MA, Tanurdzic M, Doerge RW. Designing Microarrays. In: Milliken GA, editor. Proceedings of the 2001 Kansas State University Conference on Applied Statistics in Agriculture. Manhattan, KS: Department of Statistics; 2002. p. 159-82.
5. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, et al. The generic genome browser: a building block for a model organism system database. Genome Res 2002;12(10):1599-610.