Mapping the Chromosome through a Novel Use of GIS and Spatial Analysis

M

Jane L. Garb
Baystate Health, USA

D. Joseph Jerry
University of Massachusetts, USA & Pioneer Valley Life Sciences Institute, USA

Mary J. Hagen
University of Massachusetts, USA

Jennifer Friderici
Baystate Health, USA

INTRODUCTION

This article merges three disciplines**:** *molecular biology*, *Geographic Information Systems* (GIS) and spatial statistics. GIS is the hardware and software for storing, managing and visualizing data on geographic location (traditionally referring to location on the earth's surface)-- at the *scale* of continents, countries, states, towns, neighborhoods, blocks, etc. The application of GIS and spatial analysis to cells and the *chromosomes* within their nuclei, which collectively form the *genome*, represents a vastly different scale of exploration. The genome is distributed across different chromosomes and distinct regions are organized into more than 20,000 genes in mammals.

The expression of *genes* and their relationship to one another is the target of this application. Genes are linear sequences of base pairs (Adenine, Cytosine, Guanine, Thymine) which compose the genetic code. Genes are functional units that are transcribed into *RNA*, which provides the template for making proteins. This process is called *gene expression*, and the amount of RNA synthesized indicates the activity level of a biochemical pathway. *Microarray* technology measures RNA levels of the 20,000+ genes in the genome at one point in time. Gene expression on a microarray is typically visualized as a "heat map," with color-coded RNA expression levels (e.g., red to green from high to low). Figure 1 shows a microarray heat map for two experimental groups of mice described later in this article. The arrangement of genes on the microarray does not correspond to their location on chromosomes. However, now that the exact location along the chromosome is known for most genes for humans and many other species (Chou, et al., 2004; Hanin, et al., 2009; Ishii, et al., 2000; Jurata, et al., 2004; Van de Wiel, et al., 2005) we may visualize gene expression spatially with technologies like GIS. Spatial analysis can provide insight into how each gene interacts with its neighbors. This is important to understanding how expression of more than 20,000 genes is coordinated to allow the development of dramatically different tissues, such as muscle and brain.

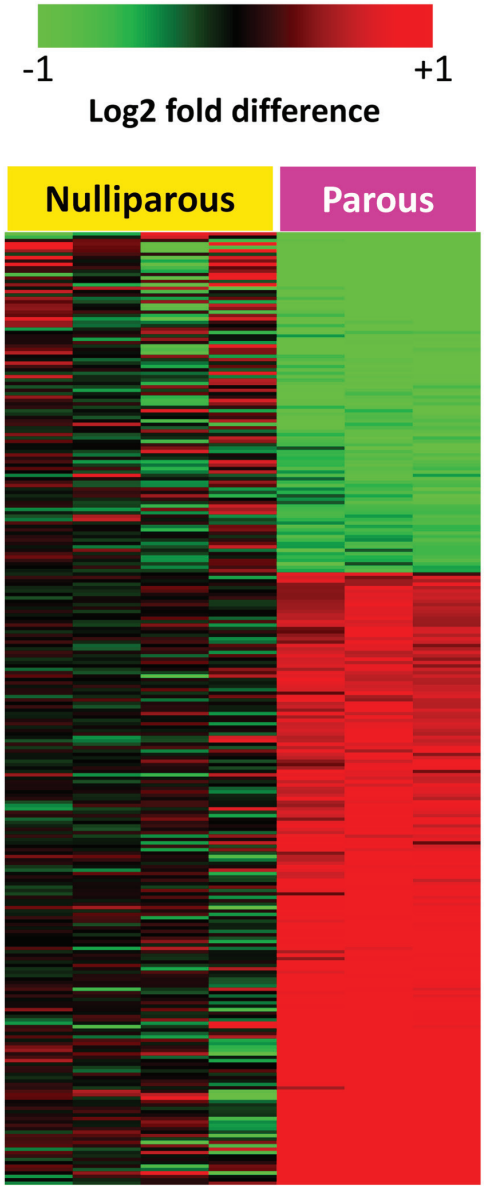
The aim of this article is to provide proof-of-concept for a novel application of spatial statistics and GIS to enhance our understanding of mechanisms underlying gene regulation. The methodological advantages and shortcomings in the analysis of gene expression by previous methods are contrasted with those of our proposed method.

BACKGROUND

Various methods have been used to analyze gene expression patterns in microarray data. Hahn (Hahn, 2006) used time series to look for *coordination* of gene expression (implying that genes act together) in the *Drosophila* (fruit fly). He treated the chromosome as one-dimensional space with data measured at

DOI: 10.4018/978-1-4666-5888-2.ch550

Figure 1. Heat map of gene expression. A total of 279 genes were differentially expressed in mammary tissues from nulliparous mice compared to parous mice. Each row represents a unique gene. The heat map shows the levels of mRNA expression for each mouse relative to the mean value of the nulliparous group. Green indicates decreased levels and red increased levels using a log2 scale where -1 represents a 2-fold decrease and +1 represents a 2-fold increase relative to the mean of the nulliparous group.



equally-spaced intervals along its length, rather than at more realistic irregularly-spaced intervals. Along with others (De Iorio & Verzilli, 2007; Guanghua, et al., 2013), he considered spatial *autocorrelation* (clustering) a statistical nuisance to be corrected, rather than a phenomenon to be studied.

Other methods have focused on finding spatial autocorrelation as a goal rather than as a nuisance factor. These studies looked for local variations of gene expression and clustering of similar expression levels in neighboring genes. To identify clustering, Reyal et al. (Reyal, et al., 2005) examined gene expression patterns in 130 human breast cancer samples. They calculated a "transcriptome correlation score" for each gene which measured the correlation of its expression level and that of 20 neighboring genes. They did not consider the actual distances between genes, which can vary greatly. They then used hierarchical clustering methods to find groups of genes whose expression levels were highly correlated. Kovanen (Kovanen, et al., 2005) used a small fixed scanning window of 2-10 genes to look for clusters of gene expression in immune response. Other statistical methods used to examine gene expression patterns include Bayesian modeling (Caron, et al., 2001; Guanghua, et al., 2013; Xiao, et al., 2009) and wavelet transform (Turkheimer, 2006). Hanin (Hanin, et al., 2009) used spectral analysis to look for repeating patterns of gene expression across the entire chromosome.

Mice provide a convenient model to study spatially-associated gene expression to identify mechanisms for tumor suppression in humans. An example is pregnancy, which reduces the lifetime risk of breast cancer in humans by up to 50% (Albrektsen, et al., 2005). Hormonal exposure during pregnancy also renders the mammary glands of rodents resistant to tumors (Russo & Russo, 1978). Pregnancy imparts persistent changes in gene expression patterns in the mammary gland which presumably mediate the protective effect of pregnancy (Blakely, et al., 2006). However, it is unknown whether these changes are due to coordinated regulation of genes within specific regions of the genome.