Transcriptomics: Open Access
Editorial Open Access

The Utilization of Online Gene Expression Data Repositories to Generate Testable Hypotheses in the Laboratory

Brendan D Stamper*

School of Pharmacy, Pacific University, Hillsboro, USA

Since their invention in the late 1980s, microarrays have revolutionized our understanding of gene expression patterns and the important role transcription plays in human health and disease. As the popularity of high-throughput hybridization arrays has increased over the years, the public's demand for access to original array datasets has increased as well. In response to this demand, the National Center for Biotechnology Information (NCBI) established the Gene Expression Omnibus (GEO) in 2000, which can be accessed at http://www.ncbi.nlm.nih.gov/geo [1]. Two years later, the European Bioinformatics Institute (EBI) established ArrayExpress, which can be accessed at http://www.ebi.ac.uk/Arrayexpress [2]. Currently, these repositories hold over 1 million samples combined. Recent updates on both the GEO and ArrayExpress repositories were published earlier this year [3,4].

Many journals and funding sources require manuscript authors to deposit their microarray datasets into one of these large-scale, publicly accessible repositories with minimum information about a microarray experiment (MIAME)-compliant content [5]. Direct access to these datasets allows for independent researchers to mine and interpret these resources in efficient and creative ways. For example, Atul Butte's lab published an association study last year consisting of 130 independent gene expression array experiments that identified CD44 as a key biomarker associated with type 2 diabetes [6]. Based on this finding, they proceeded to design experiments in the laboratory to examine the contribution of CD44 in the molecular pathogenesis of type 2 diabetes. They found CD44 expression was enhanced in obese adipose tissue and that it played a role in adipose tissue inflammation and insulin resistance. Additionally, they found that anti-CD44 antibody treatment decreased adipose tissue inflammation and caused blood glucose levels to drop. Using available datasets as a discovery and hypothesis-generating tool, Dr. Butte's lab was able to identify a novel molecular target associated with type 2 diabetes.

Similarly, in our most recent publication, we used the correlation between gene-gene pairs to identify two subtypes (termed A and B) from primary cell lines derived from craniosynostosis patients [7]. To investigate the potential consequences of unique gene-gene correlation structures in these craniosynostotic subtypes, we mined expression data from over 4,500 studies and identified forty-two studies in which our subtype gene expression patterns were highly active. One of these studies (GSE12264) [8], suggested that during mineralization subtype A gene expression would be reduced, whereas subtype B gene expression would be enhanced. Based on this data-driven prediction, we hypothesized that these two subtypes would differ in their mineralization ability. Indeed, we found that representative primary cell lines within subtype A demonstrated significantly lower levels of mineralization compared to subtype B lines by Alizarin Red staining. Taken together, these results indicate mineralization ability varies within unique craniosynostosis subtypes, and that dysregulation of the mineralization process may play an important role in the pathogenesis of the disease.

These are two examples of how online hybridization array repositories can be used to generate testable hypotheses in the laboratory using a completely data-driven approach. The potential for numerous comparisons exists though the utilization of information currently available in GEO or Array Express. Oftentimes microarray experiments are designed to answer a specific question posed by the original investigator. However, other researchers are capable of repurposing these datasets from their original intent to ask novel questions from unique perspectives. Furthermore, the vast number of samples that are currently available to analyze allows investigators to leverage multiple studies across multiple species and platforms to inform their research. These techniques are cost-effective methods for identifying more robust biomarkers and generating stronger, more testable hypotheses with a high probability for success.

References

1. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
2. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. (2003) Array Express-a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31: 68-71.
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 41: 991–995.
4. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, et al. (2013). Array Express update-trends in database growth and links to data analysis tools. Nucleic Acids Res 41: 987-990.
5. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365-371.
6. Kodama K, Horikoshi M, Toda K, Yamada S, Hara K, et al. (2012) Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes. Proc Natl Acad Sci USA 109: 7049-7054.
7. Stamper BD, Mecham B, Park SS, Wilkerson H, Farin FM, et al. (2012) Transcriptome correlation analysis identifies two unique craniosynostosis subtypes associated with IRS1 activation. Physiol Genomics 44: 1154-1163.
8. Granchi D, Ochoa G, Leonardi E, Devescovi V, Baglìo SR, et al. (2010) Gene expression patterns related to osteogenic differentiation of bone marrow-derived mesenchymal stem cells during ex vivo expansion. Tissue Eng Part C Methods 16: 511-524.

***Corresponding author:** Brendan D Stamper, School of Pharmacy, Pacific University, 222 S.E. 8th Avenue #451, Hillsboro, OR 97123, Tel: 503-352-7287; Fax: 503-352-7270; E-mail: stamperb@pacificu.edu

Received July 15, 2013; **Accepted** September 12, 2013; **Published** September 16, 2013

Citation: Stamper BD (2013) The Utilization of Online Gene Expression Data Repositories to Generate Testable Hypotheses in the Laboratory. Transcriptomics 1: e104. doi:10.4172/2329-8936.1000e104