Published in final edited form as:
Int Rev Neurobiol. 2012 ; 104: 91–133. doi:10.1016/B978-0-12-398323-7.00005-7.

Using Genome-Wide Expression Profiling to Define Gene Networks Relevant to the Study of Complex Traits: From RNA Integrity to Network Topology

M.A. O'Brien[*,†,1], **B.N. Costin**[*,†,1], and **M.F. Miles**[*,†,2]

[*]Department of Pharmacology and Toxicology, Virginia Commonwealth University, Richmond, Virginia, USA

[†]The VCU Alcohol Research Center, Virginia Commonwealth University, Richmond, Virginia, USA

Abstract

Postgenomic studies of the function of genes and their role in disease have now become an area of intense study since efforts to define the raw sequence material of the genome have largely been completed. The use of whole-genome approaches such as microarray expression profiling and, more recently, RNA-sequence analysis of transcript abundance has allowed an unprecedented look at the workings of the genome. However, the accurate derivation of such high-throughput data and their analysis in terms of biological function has been critical to truly leveraging the postgenomic revolution. This chapter will describe an approach that focuses on the use of gene networks to both organize and interpret genomic expression data. Such networks, derived from statistical analysis of large genomic datasets and the application of multiple bioinformatics data resources, potentially allow the identification of key control elements for networks associated with human disease, and thus may lead to derivation of novel therapeutic approaches. However, as discussed in this chapter, the leveraging of such networks cannot occur without a thorough understanding of the technical and statistical factors influencing the derivation of genomic expression data. Thus, while the catch phrase may be “it's the network … stupid,” the understanding of factors extending from RNA isolation to genomic profiling technique, multivariate statistics, and bioinformatics are all critical to defining fully useful gene networks for study of complex biology.

1. INTRODUCTION

Complex trait analysis describes an area of biology that is extremely crucial to our understanding of most prevalent human diseases, such as cancer, heart disease and Alzheimer’s, among many more. The “complex” part of this biology implies what now seems obvious, that both health and disease occur through a very intricate interaction between environment and our genome—with the interaction levels including organ, cellular, and molecular systems. Variation in a single gene only rarely causes disease, and even in

[2]Corresponding author: address: mfmiles@vcu.edu.
[1]These authors contributed equally to this work.

those instances, the phenotypic expression of the disease is modulated by multiple other genes and environmental factors.

Given such complexity, how is modern biomedical science ever expect to identify genes modulating complex disease, much less produce hypothesis-driven development of new treatments? Indeed, the absolute revolution in the production of high-dimensional data for DNA polymorphisms (e.g., genome-wide association studies, GWAS), DNA or chromatin modifications (epigenomics), gene expression (genomics; e.g., microarrays or RNA sequencing (RNA-Seq)), or protein (proteomics) and metabolite (metabolomics) abundance has threatened to actually impair hypothesis-driven research by their proclivity for producing hypothesis generation without causality. The answers to this riddle lie perhaps in the use of new tools for data warehousing, organization, and analysis. In particular, the organization of genomic-level data into networks and analysis of such modules across multiple experimental conditions has recently allowed the generation of testable hypotheses for novel intervention in complex traits (Zhu et al., 2004, 2008).

This chapter will provide an overview of the use of mRNA expression profiling, predominantly through use of DNA microarrays, together with complex gene set meta-analysis and network analysis for the study of complex traits. We introduce elementary concepts central to the successful performance of such studies and provide at least an introduction to the elegant complexity of modern data analysis that is possible for such high-dimensional data. The overall goal of our effort is to identify hopeful directions for future studies that will truly realize the promise of postgenomic studies in the understanding of complex biology and treatment of human disease.

2. FUNDAMENTALS OF GENE EXPRESSION ANALYSIS

The central dogma of molecular biology can be expanded to say that at any given time, the state of a cell is governed by the selection of genes undergoing transcription and translation, influenced by cellular function and environmental factors. This concept is the basis for gene expression profiling, which allows us to study the steady-state level of RNA, known as the transcriptome, under specific biological conditions. During the epoch of genomic sequencing, we have taken enormous strides in identifying genes, and as we now move forward in the postgenomic era, we can characterize not only individual gene functions, but how genes work in aggregate to produce more complex phenomenon, such as behavior.

2.1. Experimental design

Prior to defining a particular platform for performing genomic studies or conducting detailed bioinformatics analysis of such results, choosing optimal experimental design is absolutely critical. As the saying goes, “garbage in … garbage out.” This is particularly true for whole-genome expression analyses. A variety of factors involved in experimental design are discussed throughout this chapter. Additionally, extensive statistical discussions of experimental design issues for genomic studies have been published (Dudoit, Yang, Callow, & Speed, 2000; Yang & Speed, 2002) and are outside the purview of this chapter. However, put simply, perhaps the most important factors in using genomic profiling for identification of gene networks related to complex traits is the prevention of systematic bias either through

technical or environmental factors (Chesler, Wilson, Lariviere, Rodriguez-Zas, & Mogil, 2002; Kerr & Churchill, 2007; Verdugo, Deschepper, Munoz, Pomp, & Churchill, 2009; Yang & Speed, 2002).

A wide number of techniques have been used to measure steady-state levels of mRNA. Traditional methods, such as PCR, nuclease protection assay, and Northern blotting, only permit analysis of one or a few genes at a time. Over the past two decades, high-throughput methods have been developed which enable the monitoring of thousands of genes or sequences simultaneously. This allows researchers to take a holistic perspective when working to understand the function of a disease state at the level of systems biology. While methods such as differential display (Liang & Pardee, 1992), suppression subtractive hybridization (Diatchenko et al., 1996; Gurskaya et al., 1996), expressed sequence tag (EST) sequencing, and serial analysis of gene expression (SAGE) have been used to study expressed genomic sequences, microarrays have predominated in the literature as the standard genomic expression analysis platform over the past 15 years. Technological advances in sequencing methodologies (NextGen sequencing) have allowed for “deep-sequencing” of the entire genome and ushers in a possible new revolution in gene expression analysis (Wang, Gerstein, & Snyder, 2009). Such RNA-Seq analysis will be discussed in further detail later in this chapter. However, the overall approach and methodology applied to microarray-derived expression data will be discussed as a template for functional genomics analysis.

There are numerous array fabrications (Chee et al., 1996; Lockhart et al., 1996; Schena, Shalon, Davis, & Brown, 1995) (the specifics of which are in excess of what can be described within this chapter), but despite their design differences, all microarray platforms function according to the same basic molecular principles. Total RNA, from the biological samples under investigation, is processed to produce fluorescently labeled probes. Nucleic acid hybridization is then performed between the labeled molecules and complementary probes immobilized on a solid surface, allowing for the determination of what genes are expressed and their relative or absolute abundance. Fluorescent signal is detected by a scanner, and intensity correlates directly with the abundance of mRNA present in the sample. In addition to mRNA expression analysis, different microarray platforms have been developed for numerous applications. Arrays now exist to assay genomic sequence, microRNA abundance, DNA–protein interactions (ChIP on chip), single nucleotide polymorphisms (SNPs), alternative splicing (exon arrays), and DNA methylation. Many of these applications have more recently been approached using next-generation sequencing (NGS).

As previously mentioned, gene expression measures assess the steady-state levels of transcripts, which are regulated by a balance between transcription, degradation, and modification. For optimum results one should be highly concerned with the treatment of tissue and the subsequent quality of RNA, as this is of the upmost importance to ensure reflection of the state of experimental conditions, and not technical manipulations. Although more related to overall quality control (QC) analysis prior to genomic studies, the determination of the integrity of the input RNA is one of the most important design features for conducting a successful genome-wide expression analysis. One of the most effective

tools for characterizing RNA integrity is microfluidics capillary electrophoresis. These automated platforms detect RNA degradation by gauging alterations in 28S/18S rRNA signal ratio, as calculated from software generated electropherographs. This metric has been used as the benchmark of RNA integrity based on the assumption that rRNA quality and quantity reflect that of the underlying mRNA population. However, it has also been shown that simply using 28S/18S ratio is insufficient for distinguishing quality RNA from degraded RNA for subsequent gene expression analyses (Copois et al., 2007; Imbeaud et al., 2005). To overcome the limitations of the 28S/18S ratio, other quality assessment methods have been described (Auer et al., 2003; Copois et al., 2007; Schroeder et al., 2006). The degradometer (Auer et al., 2003) is an open-source software application which calculates a degradation factor from the ratio of 18S peak height relative to the average signal of degradation products present on the chromatograph. RIN software, designed specifically for the Aglilent 2100 Bioanalyzer (Schroeder et al., 2006), has been developed to classify a total RNA sample based on predefined critical features extracted from the electrophoretic trace. Bio-Rad has also developed a RNA Quality Indicator for the Experion™ Electrophoresis system, which quantitates RNA integrity by comparing the sample's electrophoretic trace to that of a series of standardized degraded RNA samples (Denisov, Strong, Walder, Gingrich, & Wintz, 2008). RNA integrity can sometimes be ascertained after sample analysis. For instance, certain Affymetrix chips utilize the 3′/5′ ratio of certain "housekeeping genes" as a measure of the degradation in the original samples. It must be considered, however, that this metric reflects not only original quality of the RNA but also the accuracy of sample processing (Popova, Mennerich, Weith, & Quast, 2008).

In addition to traditional experimental design factors, such as statistical power and control of undesired independent variables, genome-wide analyses require particular effort be afforded during the planning of gene expression studies so that technical bias is not inadvertently introduced into the data. Systematic batch effects, or technical sources of variation due to grouping samples throughout the protocol, can result in perceived nonbiological changes in gene expression. While normalization, discussed in greater detail later in the chapter, and more extensive batch correction methods exist (Benito et al., 2004; Johnson, Li, & Rabinovic, 2007) to identify and remove systematic error, it is preferable to avoid these cofounds as much as possible in the initial planning stages. Supervised randomization at every step during sample processing can be a proactive way to minimize the effect systematic bias has in the gene expression data (Chesler et al., 2002; Wolen et al., 2012).

Neurobehavioral gene expression studies are further complicated by the complexity of the nervous system. Differences between distinct brain regions and the presence of multiple cell types that work as an amalgam to function as a system can prove challenging when designing an experiment. Indeed, expression profiling has been used to study regional and cellular function within the nervous system (Oldham et al., 2008). This heterogeneity can diminish an investigator's ability to detect changes in low-abundant genes. Or, genes in which small alterations in expression level can result in significant changes in function. Additional concerns when dealing with postmortem human tissue include postmortem interval, manner of death, and antemortem exposure to drugs and toxins; all of which can affect RNA quality and gene expression (Lipska et al., 2006).

2.2. Low-level analysis of gene expression data

Low-level analysis, also known as preprocessing, refers to steps that convert raw physical data of a gene expression experiment to information that can be processed in a way that will produce optimal downstream statistical and bio-informatics analysis of the data. In this section, we will focus on some specific low-level analyses applicable to microarray data analysis. The initial steps for preprocessing of microarrays, including image analysis and methods whereby the raw luminescence intensities, indicative of hybridization events, are transformed to numerical values are generally platform specific and beyond the scope of this chapter. Typically, this entails locating probes on the chip/bead by superimposing a predesigned grid on to the scanned image and extracting pixel level data to compute probe mean intensity values. The second stage of low-level analysis concerns conversion of raw hybridization intensity/annotation information into defined expression values for each probe or gene. Again, methods for these are highly platform specific and have been evaluated or reviewed extensively in other publications (Mieczkowski, Tyburczy, Dabrowski, & Pokarowski, 2010).

Preprocessing of gene expression data additionally involves the identification and removal of systematic sources of variation that could obscure the true biological changes present in the data. This is accomplished through data normalization. A number of the aforementioned summarization algorithms involve a normalization of the data to reduce these potential biases (Reimers, 2005). The first decision in determining a normalization scheme is to choose which genes the normalization will be based upon. One option is to use all genes represented on the chip to calculate a normalization factor. Normalizing against thousands of genes at once will provide a more stable and robust measure of the variation in the data, but is only advisable if there is a large number of equally expressed genes between conditions. In most experiments, the proportion of differentially expressed genes is low, but heterogeneity of samples must be considered before taking this approach (Irizarry et al., 2003). If a global normalization is used, it has been found that specificity and sensitivity of arrays are improved if such normalization is calculated with attention to both constant and intensity-dependent sources of error. Thus, Lowess normalization is typically used for two-color cDNA or oligonucleotide arrays and is based on the assumption that biases are intensity dependent, and separate factors should be used to normalize high- and low-expressing genes (Smyth & Speed, 2003). Robust multi-array analysis (RMA) employs quantile normalization which forces the distributions of probe intensities between chips to have the same statistical parameters (Irizarry et al., 2003).

An alternative is to use a small set of “housekeeping genes” which *a priori* are expected to have equivalent levels of expression in all compared samples. One major caveat of this method is the circular logic that occurs when attempting to define a set of housekeeping genes as their validity must be determined by comparing their expression to another predetermined reliable measure. An experiment prior to the gene expression analysis may need to be performed to estimate the expression levels of potential reference genes in response to the experimental conditions. In another scheme, spike-in control mRNAs from an organism other than the one being tested can be added to the samples in known concentrations. The concentration range of the controls should span the expected level of

expression by high- and low-expressing genes, allowing for adequate normalization at the extremes. The main limitation of this method is ensuring that variation throughout the entire sample preparation is accounted for. The earlier in the process control mRNAs are added, the more technical error that can be factored into the normalization.

Another preprocessing step applicable to particular microarray platforms is probe signal summarization. For instance, Affymetrix oligonucleotide arrays utilize multiple probes, termed a probe set, to interrogate the expression level for each gene. Often, these probe sets consist of 8–15 “perfect match” or PM oligonucleotides which are perfectly complementary to the sequence they interrogate, paired with “mismatch” or MM probes containing a single-base substation aimed at disrupting hybridization. Most recent Affymetrix array versions have eliminated the MM probes. Therefore, to be able to analyze data obtained from such arrays requires an algorithm to compute a single numeric intensity value based on the expression values for each probe. Numerous methods have been described as how to produce the most accurate estimate of gene expression from the individual probe data (Affymetrix, 2002, 2005; Irizarry et al., 2003; Li & Wong, 2001). A more recent method, PLIER, (Probe Logarithmic Intensity ERror) uses mixed-model error estimation and a multiplicative model to summarize signal data (Affymetrix, 2005). Irizarry et al. (2003) describes an RMA which assumes that intensity values for each individual probe are affected not only by random error but also by the probe’s unique affinity for its target. The algorithm also excludes MM values from the calculation after it was determined that MM hybridizations were highly correlative to PM, and potentially detecting signal beyond that of nonspecific binding. Therefore, including MM probes into the computation could not improve upon the results based on PM values alone. Summarizing the PMs of a probe set to a single expression value using RMA involves fitting a linear-model to the log-transformed probe intensities. An alternative approach developed in our laboratory also takes into consideration the concept that probes in a set have different affinities for the mRNA molecule they target, and can be treated as individual reactions. The S-score algorithm compares individual probe intensities between two arrays. These relative changes in individual probe pairs are then combined to compute a single variate measure of the significance of change for each gene, referred to as the S-score (Kennedy, Archer, & Miles, 2006; Kennedy, Kerns, Kong, Archer, & Miles, 2006; Zhang, Wang, Ravindranathan, & Miles, 2002). The S-score algorithm has been previously used by our laboratory to characterize differential gene expression changes in response to acute ethanol exposure that differ between C57BL/6J and DBA2/J mice across various brain regions (Kerns et al., 2005).

Following hybridization, scanning, image processing, normalization, and low-level analysis, arrays must undergo a rigorous QC evaluation. Such QC testing should typically be done both before and after low-level analysis. Arrays that fail to meet QC standards could confound the downstream analysis and those samples should either be reprocessed or discarded (Kerns & Miles, 2008). Each microarray platform includes unique QC metrics, but there are approaches that can be used to assess the quality of data from any array platform. Scatter plots that graph the log transformation of probe or probe set intensities of two arrays against each other can reveal abnormalities in data across chips. Within a single experiment, pairwise plots between all arrays should appear fairly linear with high Pearson correlation

(*R*). This is expected, since in reality, the experimental condition being investigated should alter few genes relative to the scale of the entire genome. Plots that exhibit nonlinear data at the extreme intensity values or excessive scatter indicate problems in one or more of the arrays. An alternative approach is an M versus A plot, which graphs average log intensity of a probe in both arrays versus the log ratio of the intensities between arrays. No change in expression of a particular probe would result in a ratio near 1 and a log value of 0. Once again when comparing two arrays from the same experiment, one would expect few differences; and therefore, would predict a cigar-shaped scattering of data points around 0 on the y-axis. Another descriptive statistical approach is to graph the spread of data from each array. Box plots allow visualization of the median, first and third quartiles, and any extreme probe intensity values. This can be used to evaluate the overall distribution of probe intensities on an array, and could help in identifying outlier samples that may need to be removed before subsequent analysis. After the gene expression data has been corrected, normalized, summarized, and validated by QC analysis, the results are ready to be statistically probed to discover any biologically reproducible changes in expression.

2.3. Utilizing statistical inference for determining significant differential expression

Statistical inference is used to examine gene expression data across biological replicates to isolate significant changes, beyond what would be expected by random chance. Multiple reviews have addressed issues of statistical analysis of microarray data (Kerr & Churchill, 2007; Kim, Lee, & Sohn, 2006; Reimers, 2005). Customary statistical analyses, such as the *t*-test or ANOVA, simply tests whether the mean expression level of a gene between treatment groups is significantly different, when taking variance of measurement into consideration. A *p*-value is calculated to assess the probability of obtaining a test statistic as extreme as the one observed and is compared to a predefined significance level, α. These statistical approaches become problematic when we apply them to the field of gene expression analysis, due to the large number of genes being tested in parallel. With multiple comparisons occurring simultaneously, a significance level deemed acceptable for testing of a single gene, may result in an unacceptable number of false positives. Consider comparing the mean expression level of 1000 genes at one time. If the common significance level, $\alpha = 0.05$, was chosen for each test, one would expect, just by random chance, for a possible 50 genes that falsely rejected the null hypothesis to come through the analysis. Since gene expression analyses survey the entire transcriptome, they present an extreme multiple testing issue.

One commonly used method to balance significance and power in statistical analyses is to set an acceptable level for the expected proportion of false positives among the genes declared as differential, also known as a false discovery rate (FDR) (Storey & Tibshirani, 2003). Each hypothesis test can then be associated with a *q*-value, which is the minimum FDR at which the particular test may be called significant. A popular method for statistical filtering of data that utilizes FDR is the significance analysis of microarrays (Tusher, Tibshirani, & Chu, 2001). This method takes into consideration that expression of genes correlate in an unknown manner. An empirical distribution can be created by permuting, or randomizing the data, multiple times and determine how many genes come through as differentially expressed by chance. This will provide an estimate of the FDR for the genes

reported to be differentially expressed, put into context of the actual data. The genes that come through the statistical filtering may prove to be influential in mediating the neurobiological process being examined.

3. GENE EXPRESSION DATABASES

Genome-wide expression analysis allows a outstanding opportunity for meta-analysis across datasets to leverage the power of individual studies (Ball et al., 2004; Geschwind, 2001). Access to multiple genomic datasets allows a comparative genomics strategy, whereby candidate genes, or genes which when mutated may be responsible for a particular disease state, may be functionally conserved across species or conditions (Bhandari et al., 2012; Ewart-Toland & Balmain, 2004; Gorgels et al., 2005; Moore, 1999; Phillips et al., 2002; Rosen, Chesler, Manly, & Williams, 2007; van Bokhoven et al., 2000; Young, 2001). Some journals and public funding agencies make public availability of high-throughput data a condition of publication or funding (Ball et al., 2004; Goodman, 2001). Here, we identify several open-source databases that can both aid in identification of candidate genes for further study and serve as data repositories. These include Gene Expression Omnibus (GEO), ArrayExpress, GeneWeaver, Genenetwork, and Phenogen.

In 2002, *Nature* journals began requiring authors to deposit microarray data in either GEO (http://www.ncbi.nlm.nih.gov/geo) or ArrayExpress Archive of Functional Genomics Data (http://www.ebi.ac.uk/arrayexpress) (Parkinson et al., 2009). Both databases are subject to minimal information about microarray experiment (MIAME) reporting guidelines, which outline the minimum information that must be included when describing a microarray experiment to ensure the data can be interpreted by others (Brazma et al., 2001). GEO was originally created in response to the increasing demand for a public repository for high-throughput gene expression data, but it has adapted and evolved to store various types of datasets and information (Barrett et al., 2011; Edgar, Domrachev, & Lash, 2002).

As of 2011, GEO archived roughly 20,000 studies for over 1300 organisms submitted by 8000 laboratories from around the world supporting data from over 10,000 published works (Barrett et al., 2011). GEO data types currently stored include among others cDNA and oligonucleotide microarrays, SAGE, massively parallel signature sequencing (MPSS), and tandem mass spectrometry (Barrett & Edgar, 2006; Barrett et al., 2011). In addition, in more recent years, applications that go beyond analyzing gene expression levels have been developed. Such applications include studies analyzing genome SNPs and copy number variations known as array comparative genomic hybridization studies, genome-protein binding studies or chromatin-immunoprecipitation on arrays (ChIP-chip) studies, and NGS applications (Barrett et al., 2009, 2011). Non-expression data are housed in GEO under a division known as “Omix”—this name describes a mixture of “omic data” (Barrett et al., 2009). Information provided with datasets in GEO includes primary database information comprising submitter-supplied data regarding the platform, sample and series records (Barrett & Edgar, 2006; Barrett et al., 2011). Platform records include a summary of the array or sequencer and the array template (Barrett et al., 2011). Sample records describe the biological material and series records describe the overall study aim and design (Barrett et al., 2011). In addition, secondary database information including data extracted from the

submitter-supplied records and organized into a GEO dataset is also supplied in the database (Barrett & Edgar, 2006; Barrett et al., 2011). The dataset provides two differing views of the data: a synopsis of the experiment and a gene-centered profile that presents quantitative gene expression measurements for individual genes across a dataset (Barrett & Edgar, 2006). Genes in GEO profiles are periodically re-annotated (Barrett et al., 2011). All data uploaded to the GEO site undergo review and validation by a site curator (Barrett et al., 2009). Although GEO does not allow for dataset analysis, complete GEO records and raw data files are available for download from the GEO site in multiple formats: text tab-delimited tables, plain text, XML, and raw data are provided in native format where possible (Barrett et al., 2011; Bhave et al., 2007). GEO does have features allowing for search retrievals and the database provides graphical tools that allow the visualization and interpretation of data (Barrett et al., 2009). Data from GEO is also available in ArrayExpress for further analysis (Parkinson et al., 2009).

ArrayExpress is a major international repository for functional genomics high-throughput data and includes data generated by array-based or sequencing technologies (Parkinson et al., 2011). In addition to serving as a data repository, ArrayExpress allows for the normalization, filtering and basic statistical analysis of data (Bhave et al., 2007). The database was launched in 2002 and as of 2011 ArrayExpress contained 15,000 experiments with 425,000 assays and represented over 200 different organisms (Parkinson et al., 2007, 2009, 2011). Datasets stored in ArrayExpress include gene expression datasets, protein arrays, ChIP-chip datasets, genotyping datasets, and sequencing data (Parkinson et al., 2009). The database consists of three elements: the ArrayExpress Repository of Microarray and Transcriptomics Data, the ArrayExpress warehouse, and the ArrayExpress Atlas (Parkinson et al., 2009). The ArrayExpress Repository is a public well-annotated archive of genomics experiments and supporting data (Parkinson et al., 2009). The ArrayExpress warehouse allows users to evaluate gene expression profiles by gene name and properties and the ArrayExpress Atlas allows users to search for specific genes or conditions and evaluate gene expression levels across many data sets (Parkinson et al., 2009; Rustici et al., 2008). In addition to the databases, ArrayExpress also includes an online data analysis tool called Expression Profiler, which allows for the exploration, mining, analysis, and visualization of data from ArrayExpress or data uploaded from other sources (Kapushesky et al., 2004; Parkinson et al., 2007; Rustici et al., 2008). Curated GEO data satisfying MIAME compliance and quality metrics and are included in the ArrayExpress Data Warehouse and made available following Atlas searches (Parkinson et al., 2009). In the future, ArrayExpress will be closely integrated with a new BioSample Database at the European Bioinformatics Institute (Parkinson et al., 2011). The BioSample Database (http://www.ebi.ac.uk/biosamples) stores information about biological samples used in molecular experiments including sequencing, gene expression, or proteomics experiments (Gostev et al., 2012).

GeneWeaver (http://geneweaver.org/), described in Chapter 1, is a free, Internet accessible resource for the storage, sharing and analysis of genomic data sets across species and model systems (Baker, Jay, Bubier, Langston, & Chesler, 2012; Baker et al., 2009). The database aims to identify and incorporate quality public research data and allows large numbers of independently derived and published genomic results to be deposited, re-analyzed, and

organized into new frameworks (Baker et al., 2009, 2012). GeneWeaver currently contains more than 48,000 gene sets from seven different species derived from gene expression microarray experiments, RNA-Seq experiments, quantitative trait loci (QTL) mapping, and GWAS (Baker et al., 2009, 2012). The database also integrates additional data sources: NCBI, ENSEMBL, various model organism databases (i.e., The Mouse Genome Database (Blake, Bult, Kadin, Richardson, & Eppig, 2011), Rat Genome Database (Twigger, Shimoyama, Bromberg, Kwitek, & Jacob, 2007), HUGO Gene Nomenclature Committee (Bruford et al., 2008), Saccharomyes Genome Database (Cherry et al., 1998), WormBase (Harris et al., 2010), FlyBase (Tweedie et al., 2009), and the Zebrafish Model Organism Database (Sprague et al., 2008)), the drug related gene database of the Neuroscience Information Framework (Gardner et al., 2008), GeneNetwork (discussed below) (Chesler, Lu, Wang, Williams, & Manly, 2004; Wang, Williams, & Manly, 2003), and the Comparative Toxicogenomics Database (Davis et al., 2011). While GeneWeaver is not subject to MIAME reporting guidelines, gene sets are stored with descriptive data and all data on the GeneWeaver site is subject to curator review (Baker et al., 2012).

GeneWeaver's tools include gene set graphing, phenome graphing, anchored bicliques of biomolecular associates (ABBA), gene set similarity, Boolean gene set functions and geneset similarity clustering (Baker et al., 2009, 2012). The gene set graph tool allows users to identify highly connected genes and gene sets (Baker et al., 2012). The phenome graph displays a hierarchical network of gene set interactions in which similar gene sets are joined (Baker et al., 2012). The ABBA tool allows users to find genes with similar functional associations to a particular gene or set of genes. In addition to identifying gene sets, the ABBA tool also returns a list of similar genes that are enriched among the same gene sets as the input gene or genes (Baker et al., 2012). Gene set similarity uses the Jaccard Similarity index and the Hypergeometric test to produce a matrix of similarity statistics and a matrix of Venn diagrams for all gene sets in the analysis (Baker et al., 2012). The Boolean Gene Set Logic feature can allow users to reduce large numbers of gene sets to smaller numbers of gene sets that focus on specific criteria (Baker et al., 2012). The gene set similarity clustering tool can organize large gene sets into hierarchical clusters that can be used to eliminate redundant inputs (Baker et al., 2012).

GeneWeaver is advantageous as it allows for cross-species data integration and the analysis of sets of related biological processes through the use of data sets derived through various experimental methods and models (Baker et al., 2009, 2012). For example, a GeneWeaver query for nicotine retrieves 98 Gene Sets including studies in human cell lines (GS18839) (GS18841) (GS15464), mice (GS14889) (GS14888), various brain regions (GS14888) (GS87149), mouse strains (GS14888) (GS14889), and GWAS (GS14907) (GS14908). This feature is useful as combining results from human and animal models can potentially overcome inherent limitations of each model. Gene expression studies in animal models can identify groups of genes that change together on a homogeneous genetic background, with the signal not masked by the noise generated from the variable genetic background present in human studies (Kerns et al., 2005; Le-Niculescu et al., 2007). In addition, endophenotypes of the disorder can be deliberately mimicked in animal models with pharmacological approaches (Kerns et al., 2005; Ogden et al., 2004; Wolstenholme et al., 2011) or observed in genetic mutants (Le-Niculescu et al., 2007; Roybal et al., 2007).

Finally, gene–environment interactions can be minimized in animal models (Kerns et al., 2005; Le-Niculescu et al., 2007; Wolstenholme et al., 2011). GeneWeaver is also advantageous in that inclusion of diverse data resources allows the database to promote further elucidation of the function of poorly annotated genes. In the future, GeneWeaver hopes to abandon the current necessary assignment of gene symbols to all data allowing for the submission of RNA, SNP, or methylation data to the database (Baker et al., 2012).

The PhenoGen informatics Web site (http://phenogen.ucdenver.edu/) serves as a resource for storing, analyzing, and interpreting microarray, genotype, and phenotype data (Bhave et al., 2007; Hoffman et al., 2011). Data in the PhenoGen database is subject to the MIAME reporting standards (Bhave et al., 2007; Brazma et al., 2001; Geschwind, 2001). Registered users of the PhenoGen site can upload their data and/or download data from the site or use data stored on the site to perform "*in silico*" experiments (Bennett et al., 2011; Bhave et al., 2007; Hoffman et al., 2011). The PhenoGen site is a useful tool as each array on the site represents mRNA generated from a single animal; data on 4–7 biological replicates is available; recombinant inbred (RI) and inbred datasets have been normalized and non-unique probes or probes with SNPs have been removed; several options for data normalization, filtering, and statistical analysis are available; and users can search the PhenoGen database for individual transcripts of interest (Bennett et al., 2011; Bhave et al., 2007; Hoffman et al., 2011). Additional site tools include annotation tools, promoter analysis tools, and literature search options (Bhave et al., 2007; Hoffman et al., 2011). PhenoGen does allow users to perform their own QC measures and perform individual statistical tests comparing gene expression and phenotype-gene expression correlations (Hoffman et al., 2011). Registered users of the PhenoGen Web site have access to data that is classified as "open access" and they need not obtain permission from the data curator; however, users can only access "semi-public" data when granted permission by the curator of the data (Bhave et al., 2007).

In particular, PhenoGen is useful for combining QTL and microarray data to identify candidate genes contributing to complex traits (Bhave et al., 2007; Hoffman et al., 2011). Using a genetic reference population (GRP) of RI strains, relative transcript levels are treated as quantitative traits allowing the chromosomal regions regulating transcript levels to be mapped as expression quantitative trait loci (eQTLs) (Bennett et al., 2011; Chesler et al., 2005; Hoffman et al., 2011; Li et al., 2001; Mogil et al., 2003; Shirley, Walter, Reilly, Fehr, & Buck, 2004). PhenoGen allows for candidate gene identification through the analysis of the correlations between gene expression levels and phenotypes (Bhave et al., 2007; Hoffman et al., 2011; Hovatta et al., 2005; Korostynski, Kaminska-Chowaniec, Piechota, & Przewlocki, 2006; Nadler et al., 2006; Saba et al., 2006). Diverse phenotypes can be collected from the GRP and overlap of eQTLs with phenotypic QTLs can be a powerful tool for candidate gene identification (Bennett et al., 2011; Chesler et al., 2005; Hoffman et al., 2011; Li et al., 2001; Mogil et al., 2003; Shirley et al., 2004; Chesler et al., 2003). PhenoGen allows access to data from 32 strains in the BXD RI mouse panel, the HXB/BXH RI rat panel, and 20 common inbred mouse strains (Bennett et al., 2011). PhenoGen shares a focus with another database, GeneNetwork, correlating behavioral phenotypes with gene

expression levels in RI and inbred panels of rats and mice and these two sites can be used to perform complementary analyses (Hoffman et al., 2011).

GeneNetwork (www.genenetwork.org), described in Chapter 6, is a suite of data sets and bioinformatics tools that stores, analyzes, and displays phenotypes as well as large gene expression data sets for several species (human, monkey, mouse, rat, fly, barley, tomato, and Arabidopsis) (Durrant et al., 2012; Hoffman et al., 2011; Rosen et al., 2007). GeneNetwork users can take advantage of a systems genetics approach (Rosen et al., 2003, 2007). While the candidate gene approach asks which one gene mutation causes a particular disease, the systems genetics approach explores which phenotypes and diseases result from diverse sets of genetic and molecular markers (Rosen et al., 2003, 2007). The majority of data sets in GeneNetwork are collected from GRPs consisting of hundreds of diverse, inbred strains of mice and rats (Rosen et al., 2007). GeneNetwork includes phenotype data both submitted by users and from the literature and the data in GeneNetwork have already undergone normalization and QC measures (Hoffman et al., 2011). GeneNetwork provides users with useful background information regarding their gene or genes of interest including the trait identifier, gene symbol, chromosomal location, and megabase position of the gene. In addition to this, GeneNetwork can be used to study correlations between traits and to perform data mining in genomic regions containing candidates for quantitative trait genes (Hoffman et al., 2011). All datasets in GeneNetwork are linked to a materials and methods information page that summarizes experimental details relating to the dataset.

Databases within GeneNetwork include the transcriptome database, the BXD published phenotypes database, the genotype database, and the SNP database (Rosen et al., 2007). The transcriptome database provides microarray estimated mRNA levels of tissue from various GRPs including the BXD RI line (Rosen et al., 2007). The BXD RI mapping panel was first generated at the Jackson laboratory in the mid-1970s and was more recently extended to 80 strains (Peirce et al., 2004). The two parental strains, C57BL/6J and DBA/2J, are sequenced allowing for efficient positional candidate gene evaluation (Chesler et al., 2005; Taylor et al., 1999). The BXD published phenotypes database provides user submitted data as well as data obtained through a search of all PubMed-indexed journals where GRPs were used (Chesler et al., 2003; Rosen et al., 2007). The Genotype Database provides a summary of the genetic makeup of each individual GRP strain (Rosen et al., 2007). For example, in the case of the BXD strains, the files indicate whether a strain inherited both copies of its' gene from the C57BL/6J and DBA/2J parental strain and tell approximately where segments inherited from one parent or the other start and stop (Rosen et al., 2007). The SNP database identifies SNPs between parental strains throughout the genome (Rosen et al., 2007). In addition, the WebQTL module of GeneNetwork can map QTLs and tell where in the genome genetic variation is located and which parental strain (C57BL/6J or DBA/2J) is associated with higher gene expression (Hoffman et al., 2011; Rosen et al., 2003, 2007). GeneNetwork also provides links to the following external databases: NCBI Entrez Gene, Summary from on Mendelian Inheritance in Man (OMIM), GenBank, HomoloGene, UCSC Genome Browser (UCSC), BioGPS, STRING, PANTHER, Gemma, the brain synapse database, and the Allen Brain Atlas.

4. BIOINFORMATICS APPROACHES IN BEHAVIORAL NEUROSCIENCE

Despite the various high-throughput technologies employed and plat-forms available to perform expression analysis, a unifying consequence is the generation of large-scale expression datasets. The communal goal and challenge of researchers is to elucidate the biological implications of the data; relating the enormous wealth of acquired knowledge to the biological phenomenon under investigation. Tools are necessary to organize and prioritize the substantial data obtained from the high-throughput genomic methods. Statistical differential expression analysis will produce gene lists, which in and of themselves will not advance understanding of the phenotype under investigation. Rather, further interrogation of the list of significant genes, as a whole, will allow the investigator to assess interactions amongst genes. When applied to the field of neuroscience, this can reveal biologically relevant meaning and render novel insights into the molecular mechanisms that govern behavior. Focusing on these interactions and the gene networks that emerge capitalize on the unbiased investigational methods imparted in whole-genome analysis. Moreover, due to the complexity of neurobehavioral traits, it may be more relevant and informative to correlate the function of a network of genes with a phenotype, rather than an individual gene.

4.1. Functional overrepresentation analysis

One strategy for extracting meaning from an experimentally derived gene list involves revealing any shared biological function. Overrepresentation of genes associated with any particular biological classification, beyond that expected by chance alone, suggests involvement of a coherent biological mechanism within the data set. This approach is reliant on proper annotation of genes and gene products into previously defined categorizations. For instance, Gene Ontology (http://www.geneontology.org) is an initiative to curate genes based on three broad categories—cellular component, molecular function, and biological process—allowing for a common vocabulary to describe gene attributes. A variety of applications utilize these functional categories for the basis of enrichment analysis. The Database for Annotation, Visualization, and Integrated Discovery (DAVID) is one popular Web based bioinformatics resource that assesses overrepresentation based on 40 well-known publically available annotation categories, including gene ontologies (Huang, Sherman, & Lempicki, 2009). DAVID offers multiple analytic modules that allow investigators to explore the underlying biological themes present it their data. DAVID's gene functional classification and functional annotation clustering modules cluster individual genes or enriched biological terms, utilizing fuzzy clustering algorithms to allow elements to belong to more than one cluster. This permits a view of the overall enrichment that is more reflective of the biology, where genes or gene products are often involved in more than one system or pathway.

Another useful Web-based tool is ToppGene Suite (http://topgene.cchmc.org), which can provide information regarding gene list functional enrichment, functional annotation-based candidate gene prioritization and identification, and prioritization of novel disease candidate genes in the protein interactome (Chen, Bardes, Aronow, & Jegga, 2009). Currently, most computational disease candidate gene prioritization methods rely on functional annotations

and gene expression data (Adie, Adams, Evans, Porteous, & Pickard, 2005; Aerts et al., 2006; Chen, Xu, Aronow, & Jegga, 2007; Chen et al., 2009; Freudenberg & Propping, 2002; Thornblad, Elliott, Jowett, & Visscher, 2007; Tiffin et al., 2005; Turner, Clutterbuck, & Semple, 2003; Zhu & Zhao, 2007). In addition to this, ToppGene also evaluates candidate genes based on protein–protein interaction networks. ToppGene consists of four separate applications: ToppFun, ToppGene, ToppNet, and ToppGeNet. ToppFun evaluates candidate gene lists, detecting enrichment for 14 annotation categories, including proteome interactions, gene and pathway ontologies, transcription factor binding sites, microRNAs, phenotypes, drug–gene associations, and literature citations (Chen et al., 2009). The ToppGene module performs candidate gene prioritization through the comparison of a representative “training” gene set, whose enrichment profile is derived from ToppFun, to a “test” gene set (Chen et al., 2009). Genes in the test set can be differentially expressed due to a specific disease or treatment or candidate genes from linkage analysis studies. A similarity score for each gene in the test set is derived by comparison to the training profile, essentially assigning a prioritization to the genes. ToppNet ranks candidate “training” set genes based on their protein–protein interactions with a “test” gene set in a protein–protein interaction network (Chen et al., 2009). ToppGeNet identifies and prioritizes the neighboring genes in the protein–protein interaction network based on their functional similarity to the original “seed” list. After removal of overlapping genes in the test or training set, remaining genes in the interactome can be prioritized using ToppGene or ToppNet (Chen et al., 2009). The limitations of ToppGene’s capabilities are the same as any functional annotation-based prioritization, in that the enrichment analyses are only as good as the sources from which the annotations are retrieved (Chen et al., 2009).

One additional caveat to the use of overrepresentation analysis concerns the actual statistical methodology used to determine whether a given gene set is overrepresented with members of a previously defined functional gene group (mechanistically or empirically defined). The current methods generally utilize family-wise statistics calculated by permutation of data or FDR approaches to produce statistics for ranking or significance for potentially overrepresented groups (Irizarry, Wang, Zhou, & Speed, 2009).

4.2. Literature association analysis

A complementary technique for detecting key biological significance from genome-wide expression data involves identifying connections between genes that are derived from previously published biomedical literature. Relationships such as physical protein–protein interactions, protein phosphorylation, and transcription factor binding can then be visualized as a network of direct and indirect literature associations. A number of resources are available to extract informative interactions from the literature, utilizing algorithms based on two fundamental approaches: co-occurrence and natural language processing (Jensen, Saric, & Bork, 2006). Co-occurrence methods simply indicate that two entities are mentioned together in an abstract or sentence while natural language processing is able to apply syntax and semantics to the process, facilitating the extraction of biological meaning. For instance, a relationship between two gene products can be deemed directional, where one protein acts upon another as is in the case of an enzyme and a substrate. Indirect relationships can also be inferred by algorithms that establish connections across multiple publications.

Ingenuity Pathway Analysis, IPA (http://www.ingenuity.com), is one such commercial application that derives connections between genes based on literature associations. IPA amasses extracted biological interactions and functional annotations in their Knowledge Base, which can be used for the exploration of experimental results in a manner that will stimulate and guide subsequent research avenues. Information in the Knowledge Base comes from in-depth reviews of the literature, either manually extracted and curated by experts, or automatically extracted by natural language processing algorithms which are then manually reviewed. Various other programs that utilize literature associations as the basis for data exploration include Chilibot (http://www.chilibot.net), BiblioSphere (http://www.genomatix.de), Agilent Literature plug-in for Cytoscape (http://www.cytoscape.org), and GeneGo (http://www.genego.com).

The hypothesis generation capabilities of these programs are facilitated by text-mining functions that are capable of drawing connections from the literature that are not explicit. By cross-referencing various published resources, predicted associations between genes can be integrated with genomics and proteomics data, unveiling potentially novel paths of investigation. STRING (http://string.embl.de/) is a protein–protein interaction database which aims to not only store experimentally derived evidence of protein associations and categorize them into functional pathways, but uses computational prediction techniques for predicting protein–protein relationships *de novo* (Szklarczyk et al., 2011; von Mering et al., 2005). STRING also uses sequence similarity to map orthologous proteins across species, allowing associations to extend beyond the organism originally described (von Mering et al., 2005). A database which may prove to be specifically beneficial for the field of neuroscience and investigations into the molecular mechanisms that underlie behavioral adaptation is G2C: Genes to Cognition database (http://www.genes2cognition.org). G2C catalogs genes and proteins experimentally determined to be present at synapses and uses text-mining and expert curation to extract information from published neurobiological studies. This allows an investigator to integrate coordinated synaptic complexes with their own data, providing a useful tool in the study of plasticity, behavior, and brain pathologies (Croning, Marshall, McLaren, Armstrong, & Grant, 2009).

4.3. Expression correlation analysis

While genome-wide expression analysis can yield the identification of lists of candidate genes or even overrepresented gene groups for the system under investigation, the greatest utility of these studies arise from the ability to study thousands of genes in tandem. Taking a more holistic approach to viewing the data provides investigators with the ability to observe the effects of the experimental perturbation on entire cellular systems or pathways. One inherent way to organize the data is to group genes together that share similar expression patterns. The clustering of genes together on the basis of similarity of expression in genome-wide expression experiments has been shown to efficiently segregate genes into groups of similar function (Eisen, Spellman, Brown, & Botstein, 1998; Hughes et al., 2000). These methods are termed multivariate, as they bestow the ability to focus on multiple genes at once, in attempt to study how multiple loci interact, contributing to a complex trait.

TIGR MultiExperiment Viewer (TMeV) is a downloadable software application that provides a comprehensive, versatile collection of tools for the visualization, statistical analysis, and clustering of expression data (www.tm4.org/mev/). Using TMeV, genes that come through statistical analyses as significant can be subjected to hierarchical or k-means clustering. Clustering genes in a hierarchical manner involves segregating observations into an increasing number of nested sets, resulting in a dendrogram where genes that have greater similarity in expression are more closely linked. The k-means clustering algorithm partitions genes into a fixed number of clusters, each gene segregating into the cluster that has an average expression vector most similar to its own (Quackenbush, 2001). The number of predefined clusters can be determined by either a principle component or figures of merit analysis, which will estimate the number of distinct sources of variation in the data. Pavlidis template matching (PTM) provides an additional avenue for organizing genes based on relative expression patterns, and also calculates a measure of statistical significance, indicating the likelihood that the pattern is simply due to chance (Pavlidis & Noble, 2001). Using PTM, the investigator specifies a template expression profile and the Pearson Correlation of each gene's expression pattern across experimental samples is calculated. This allows the data to be queried by asking specific questions regarding particular patterns of gene expression that the investigator may be interested in.

Coexpression analysis has benefited from application of the rapidly developing field of network or graph theory. Coexpression networks have been found to display a scale-free topology (van Noort, Snel, & Huynen, 2004), where a few highly connected nodes, or hubs, are linked to the rest of the less connected nodes of the system. The edges between gene nodes define the expression pattern relationship across experimental conditions, often as a Pearson's correlation coefficient. Network based methods of co-expression analysis have proven useful in identifying evolutionarily conserved gene and protein interactions (Stuart, Segal, Koller, & Kim, 2003), revealing highly connected hub genes that are crucial for survival (Carter, Brechbuhler, Griffin, & Bond, 2004), and detecting cell-type specific networks, even amongst heterogeneous populations such as the nervous system (Oldham et al., 2008). Weighted gene correlation network analysis (WGCNA) is one such network analysis technique that has proved notably befitting for identifying hub genes in highly correlated modules, and examining the relationship between gene modules and the experimental trait (Langfelder & Horvath, 2008; Zhang & Horvath, 2005). WGCNA assigns a connection weight between pairs of genes within the network, based on biologically motivated criterion and attempts to identify highly relevant modules by applying a "soft" threshold to correlations between pairs of genes within a network (Zhang & Horvath, 2005). WGCNA has proved effective in gene expression analyses for numerous neurobehavioral phenomena. For instance, WGCNA has been used to identify modules related to disease status in schizophrenia (Torkamani, Dean, Schork, & Thomas, 2010), reveal underlying epigenetic modifications in alcohol dependence (Ponomarev, Wang, Zhang, Harris, & Mayfield, 2012), and validated biological pathways previously identified in vocal learning (Hillard, Miller, Fraley, Horvath, & White, 2012). Other approaches similar to WGCNA have also been described, such as defining paracliques of interconnected genes (Perkins & Langston, 2009). Our laboratory has recently used such a paraclique analysis to study

genetic regulation of gene networks responding to acute ethanol treatment in mouse brain prefrontal cortex (Wolen et al., 2012).

As mentioned, the result of large-scale genome-wide analyses tend to be large lists of potential gene candidates and interconnected networks that attempt to represent the data at a systems level. It is predicted that the complex, multigenic traits that predominate in the field of neuroscience will only be truly be understood by studying alterations to the networks of genes that mediate behavior. It has been posited that one can study the effect of a network on a trait by targeting one or a few of these hub genes, whose wide reaching downstream effects could alter the entire gene network's action.

5. FUTURE DIRECTIONS AND CONCLUSIONS

The transcriptome includes all transcripts produced in a given cell or tissue population and transcriptome research strives to characterize these transcripts and the mechanisms driving their expression (Harbers & Carninci, 2005; Ruan, Le Ber, Ng, & Liu, 2004). Technologies, including hybridization and sequence-based approaches, have been developed to quantify and characterize the transcriptome (Wang et al., 2009). Here, we will discuss various hybridization and sequencing methods used to quantify the transcriptome and the advantages and disadvantages of these methods. This information is summarized in Tables 5.1 and 5.2.

Hybridization-based approaches typically involve incubating fluorescently labeled cDNA with microarrays and can be broadly separated into tiling arrays, high-density, nonbiased whole-genome arrays and non-tiling arrays, biased annotation-dependent arrays (Wang et al., 2009). Since their advent in 1995, non-tiling microarray technology changed the face of transciptome research by allowing research to move from a gene-by-gene approach to genome-wide studies (Coppee, 2008; Schena et al., 1995). Non-tiling arrays rely on prior annotations to investigate a particular subset of genomic features (Mockler et al., 2005). They are high throughput, relatively inexpensive (Wang et al., 2009), and they can help to recognize previously identified low abundance or rare transcripts (Mockler et al., 2005). But, various cDNA isoforms may have the same annotation across platforms (Marshall, 2004) and comparing expression levels across experiments can require complex normalization methods (Marshall, 2004; Wang et al., 2009). Tiling microarrays cover the entire genome—non-overlapping or partially overlapping probes may be regularly spaced throughout the genome or tiled to cover the entire genome—and allow for the discovery of new genes and exons (Gregory & Belostotsky, 2009; Mockler et al., 2005; Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008). In addition to the quantification of gene expression, tiling arrays allow for the analysis of various genomic features including alternative splicing, RNA-binding protein transcript identification, methylome analysis, polymorphism analysis, and complete resequencing of the genome (Mockler et al., 2005). Like non-tiling arrays, tiling arrays are high throughput (Wang et al., 2009), and they can also help to identify low level, rare transcripts, or yet to be annotated transcripts (Cawley et al., 2004; Euskirchen et al., 2007; Kampa et al., 2004; Kapranov et al., 2002; Mockler et al., 2005). The results of high-density oligonucleotide tiling arrays used to evaluate human chromosomes 21 and 22 suggest that a larger portion of the genome is transcribed than previously acknowledged (Cawley et al., 2004; Kampa et al., 2004; Kapranov et al., 2002).

Forty-nine percent of observed transcription on chromosomes 21 and 22 was outside of any known annotation (Kampa et al., 2004). The novel transcripts identified had lower, less fickle expression than the recognized, well-characterized genes (Kampa et al., 2004). Currently, limitations exist regarding tiling arrays in the number of unique probe features that can included on a chip and therefore how many chips are required to cover an entire genome (Mockler et al., 2005). Additional limitations regarding tilling arrays include the large amount of input RNA required and the challenge of organizing and analyzing the large amount of data generated from a single array experiment (Mockler et al., 2005; Mortazavi et al., 2008). In addition, hybridization methods are subject to a limited dynamic range of detection due to saturation of signals and high background levels due to cross hybridization, as RNAs sharing greater than 75% sequence identity to the probe can cross hybridize (Kane et al., 2000; Wang et al., 2009).

In contrast to hybridization methods, sequence-based approaches directly determine expressed sequences. Sequencing approaches began with conventional dideoxy-based (Sanger & Coulson, 1975) sequencing of cDNA or EST libraries (Sanger & Coulson, 1975). Although Sanger sequencing has excellent accuracy and good read length, it is low throughput, expensive, and nonquantitative (Wang et al., 2009; Zhang, Chiodini, Badr, & Zhang, 2011). Tag-based methods, including SAGE, cap analysis of gene expression (CAGE), and MPSS, were developed to overcome limitations in original Sanger sequencing methods (Harbers & Carninci, 2005; Velculescu, Zhang, Vogelstein, & Kinzler, 1995). SAGE allows for quantitative estimation of mRNA expression levels by measuring short (10–14 mer) sequences of transcribed messages, and using these to infer the identity of specific transcripts (Bertone, Gerstein, & Snyder, 2005). SAGE provides expression values without the need for probe design, but it is not suitable for promoter identification or full-length cDNA cloning (Harbers & Carninci, 2005; Shiraki et al., 2003). CAGE allows for the profiling of transcriptional start sites by sequencing short sequence tags originating from the 5′ end of full-length mRNA (de Hoon & Hayashizaki, 2008). MPSS is a highly sensitive technique that captures data by *in vitro* cloning of DNA templates on microbeads to monitor biochemical reactions by fluorescent probes (Brenner et al., 2000; Reinartz et al., 2002). Sequencing methods produce digital counts of transcript abundance in contrast to the intensity signals provided by florescent dye-based microarrays (Mortazavi et al., 2008). Despite their advantages, tag-based methods are based on expensive Sanger sequencing technology, short tags cannot always be mapped to the reference genome, only a portion of the transcript is analyzed, and isoforms are sometimes indistinguishable from one another (Mortazavi et al., 2008; Wang et al., 2009). In addition, some sequencing approaches may not be able to identify transcripts that are low abundance, expressed in rare cell types or in response to specific stimuli (Mockler et al., 2005; Mortazavi et al., 2008).

More recently, RNA-Seq, a novel high-throughput DNA sequencing method, has provided a new, simple, and possibly more comprehensive method for mapping and quantifying transcriptomes (Mortazavi et al., 2008). RNA-Seq allows for the identification of exons and introns and the mapping of their boundaries (Nagalakshmi, Waern, & Snyder, 2010). It also allows the identification of transcription start sites (Tsuchihara et al., 2009), the identification of splice variants, and the monitoring of allele expression and the

quantification of exon expression and splicing variants (Cloonan et al., 2008; Marguerat, Wilhelm, & Bahler, 2008; Morin et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Shendure, 2008; Wang et al., 2009; Wilhelm et al., 2008). RNA-Seq studies are performed by sequencing sheared double-stranded cDNA libraries (strandless RNA-Seq) (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Sultan et al., 2008; Wilhelm et al., 2008), or by sequencing cDNA libraries prepared using adaptor-tagged random hexamers (Cloonan et al., 2008), or serial ligation of adapters to fragmented RNA populations (stranded RNA-Seq) (Lister et al., 2008). Following sequencing, the resulting short reads are aligned against an appropriate reference genome, or assembled without the use of a reference sequence to provide transcriptome structure and the number of reads from each known exon, splice event or new candidate gene is also provided (Cloonan & Grimmond, 2008; Mortazavi et al., 2008; Wang et al., 2009).

RNA-Seq offers many distinct advantages over prior methods for mapping and quantifying the transcriptome. It provides expression values with-out the need for probe design, it can reveal the precise location of transcriptional boundaries, it has low background and no upper limit of quantification, it is highly reproducible, and it can be used for polymorphism analysis (Cloonan & Grimmond, 2008; Cloonan et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2009). Wang et al. compared two studies, one using tiling arrays (David et al., 2006) and the other using RNA-Seq (Nagalakshmi et al., 2008), examining gene expression levels in *Saccharromyces cerevisiae* cells grown in nutrient-rich media (Wang et al., 2009). Wang et al. found that the two methods correlate highly with one another ($r = 0.059$) for genes with medium expression levels, but correlation levels are much less robust for genes with low expression levels ($r = 0.099$) and high expression levels ($r = 0.177$) (Wang et al., 2009). This analysis suggests that RNA-Seq provides a more accurate quantification of genes expressed at very low or high levels due to the limited dynamic range of tiling microarrays (Wang et al., 2009).

Like other methods, RNA-Seq also has its own unique set of challenges. Larger RNA molecules must be fragmented into smaller molecules prior to sequencing. In strandless RNA-Seq (Cloonan & Grimmond, 2008), RNAs are converted into cDNA libraries through DNA or RNA fragmentation. Each of these methods can bias the outcome of sequencing differently (Wang et al., 2009). RNA fragmentation depletes the transcript ends whereas DNA fragmentation provides a strong bias toward the 3′ transcript ends (Wang et al., 2009). When total RNAs or mRNAs are fragmented and converted into cDNA libraries, poly(A) selection or removal of rRNA contamination can be used to remove ribosomal RNA and abundant transcripts during library construction (Martin & Wang, 2011). Poly(A) selection effectively enriches mRNAs in eukaryotes, but it depletes populations of non-coding RNAs and mRNAs lacking a poly(A) tail (Martin & Wang, 2011). To avoid these issues, ribosomal RNA (rRNA) contamination can be removed using hybridization-based depletion methods (Chen & Duan, 2011; He et al., 2010). Poly(A) selection and removal of rRNA increase the likelihood of the detection and assembly of rare transcripts, but these depletion methods may bias the quantification of highly abundant transcripts (Christodoulou, Gorham, Herman, & Seidman, 2011; Martin & Wang, 2011). If quantification is the goal of the study, then sequencing of nondepleted libraries is necessary. In addition, it may be necessary to

eliminate PCR amplification by using an amplification-free protocol (Kozarewa et al., 2009; Martin & Wang, 2011). PCR amplification results in low sequencing coverage for transcripts with a high-GC content leading to gaps in the assembled transcripts and missing transcripts (Kozarewa et al., 2009; Mamanova et al., 2010). In using double-stranded cDNA libraries, followed by the addition of adaptors for NGS, the directionality of the cDNA fragment is lost (Cloonan & Grimmond, 2008). To account for this loss in directionality, studies have prepared strand-specific libraries (Cloonan et al., 2008; Levin et al., 2010; Lister et al., 2008), but strand-specific libraries are laborious to produce (Cloonan et al., 2008) and direct RNA–RNA ligation is inefficient (Lister et al., 2008). Because RNA-Seq is based on resequencing, it is currently a more useful technology for organisms that already have quality sequenced reference genomes (Cloonan & Grimmond, 2008). While the yeast and *Arabidopsis thaliana* transcriptomes have been successfully profiled using RNA-Seq, the human and mouse genomes are larger and more complex and thus provide exclusive challenges (Mortazavi et al., 2008). For example, large genomes usually have large numbers of similar genes, which means these genes could map equally well to multiple genomic locations (Cloonan & Grimmond, 2008; Mortazavi et al., 2008). In addition, many identical short reads can be obtained from amplified cDNA libraries—these could be PCR artifacts or genuine measures of abundant RNA products (Wang et al., 2009). Finally, RNA-Seq faces challenges in the development of efficient methods to process, retrieve, and store large amounts of data (Wang et al., 2009).

High-throughput sequencing using NGS technologies is used for RNA-Seq. Following library preparation, templates can be amplified via clonal amplification techniques including emulsion PCR (emPCR) (Dressman, Yan, Traverso, Kinzler, & Vogelstein, 2003) and solid-phase amplification/bridge PCR (Fedurco, Romieu, Williams, Lawrence, & Turcatti, 2006; Metzker, 2010; Shendure & Ji, 2008) or via the preparation of single-molecule templates (Metzker, 2010). In emPCR, an adaptor-flanked library is PCR amplified in the context of a water-in-oil emulsion. A PCR primer is tethered to the surface of micron sized beads included in the reaction. Most bead-containing compartments have zero-one template molecules present and PCR amplicons are captured to the surface of the bead. After breaking the emulsion, beads bearing amplification products can be selectively enriched. Each clonally amplified bead will bear on its surface PCR products corresponding to amplification of a single molecule from the template library (Dressman et al., 2003; Metzker, 2010; Shendure & Ji, 2008). Solid-phase amplification can amplify templates on a glass slide. Solid-phase amplification relies on bridge PCR. Briefly, an adaptor-flanked library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5′ ends by a flexible linker. Thus, amplification products originating from an individual template remain tethered near the point of origin. Following PCR, each clonal cluster contains approximately 1000 copies of a member of the template library (Adessi et al., 2000; Fedurco et al., 2006; Metzker, 2010; Shendure & Ji, 2008). emPCR and solid-phase amplification can require large amounts of genomic DNA (3–20 μg) (Harris et al., 2008); PCR amplification can create mutations in templates that can mistakenly be considered sequence variants (Harris et al., 2008; Sjoblom et al., 2006); and AT and GC-rich sequences may show amplification bias in product yield resulting in their underrepresentation in genome assemblies (Harris et al., 2008; Metzker, 2010). In addition,

strand-specific protocols can be used to assemble and quantify overlapping transcripts from opposite strands of the genome (Cloonan et al., 2008; Levin et al., 2010; Lister et al., 2008).

The preparation of single-molecule templates can overcome some of the limitations of clonal amplification. Preparing single-molecule templates requires less starting material (<1 μg), it does not require PCR, and it can be used with larger DNA molecules (Metzker, 2010). The library preparation process for single-molecule templates results in single-stranded, poly(dA)-tailed templates. To capture the poly(dA)-tailed templates, poly(dT) oligonucleotides are covalently anchored to glass cover slips at random positions. The oligomers then serve as either a primer for the template-directed primer extension that forms the basis of the sequence reading or they can aide in template replication prior to sequencing. Up to 224 sequencing cycles can be performed. Each cycle consists of adding the polymerase and a labeled nucleotide mixture containing one of the four bases, rinsing, imaging, and cleaving the dye labels (Eid et al., 2009; Harris et al., 2008). In the preparation of single-molecule templates, multiple nucleotide or probe additions can occur in a cycle; deletion errors can occur due to quenching effects between adjacent dye molecules; and a nucleotide or probe can lack a label (Erlich, Mitra, de la Bastide, McCombie, & Hannon, 2008; Metzker, 2010).

Sequencing and imaging strategies are used for both clonally amplified and single molecule templates. Some current NGS sequencing methods are summarized in Table 5.2. Cyclic reversible termination (CRT) is used by the Illumina/Solexa's GA and Helicos BioSciences HeliScope (Braslavsky, Hebert, Kartalov, & Quake, 2003; Metzker, 2010). With the Illumina system, DNA polymerase bound to the primed template adds a fluorescently modified nucleotide complementary to the template base. These nucleotides are "reversible terminators," a chemically cleavable moiety at the 3′ hydroxyl position allows only a single-base incorporation to occur per cycle. In addition, one of four fluorescent labels corresponds to the identity of each nucleotide (four-color CRT, Table 5.2). Following nucleotide addition, the remaining unincorporated nucleotides are washed away, imaging is performed and cleavage occurs followed by additional washing (Metzker, 2005, 2010; Shendure & Ji, 2008). With the Heliscope in each cycle, DNA polymerase and a single species of fluorescently labeled nucleotide are added resulting in template extension (one-color CRT, Table 5.2). Following images of the full array, chemical cleavage, and release of the fluorescent label allow further cycles of extension and imaging (Shendure & Ji, 2008). Life/APG's SOLiD 3 and the Polonator G.007 both use sequencing by ligation (SBL) (Metzker, 2010; Valouev et al., 2008). SBL is driven by a DNA ligase, rather than a polymerase. A fluorescently labeled probe hybridizes to its complementary sequence adjacent to the primed template and DNA ligase joins the dye-labeled probe to the primer (Metzker, 2010; Shendure & Ji, 2008; Tomkinson, Vijayakumar, Pascal, & Ellenberger, 2006). Nonligated probes are washed away and florescent imaging is performed to determine the identity of the ligated probe (Landegren, Kaiser, Sanders, & Hood, 1988). The cycle can be repeated by removing and hybridizing a new primer to the template or by using cleavable probes to remove the fluorescent dye and regenerate a 5′ phosphate (Metzker, 2010). The Roche/454's GS FLX Titanium sequences via pyrosequencing. Pyrosequencing measures the release of inorganic pyrophosphate by converting it into visible light via a series of enzymatic reactions (Metzker, 2010). The order and intensity of the peaks of light distinguish the

underlying DNA sequence (Ronaghi, Karamohamed, Pettersson, Uhlen, & Nyren, 1996; Ronaghi, Uhlen, & Nyren, 1998). The latest revisions of Next-Gen sequencing real-time sequencing (Metzker, 2010). Real-time sequencing involves imaging the incorporation of dye-labeled nucleotides during DNA synthesis (Metzker, 2009). Currently, real-time sequencing has the highest error rates of an technology in the field (Metzker, 2010). Following the generation of NGS reads they can then be aligned to a known reference sequence or assembled *de novo* (Chaisson, Brinza, & Pevzner, 2009; Pop & Salzberg, 2008; Trapnell & Salzberg, 2009). The application and analysis of such mapping algorithms and downstream analysis approaches for RNA-Seq are topics too detailed for discussion here.

In addition to potentially having a greater dynamic range and sensitivity compared to microarray-based approaches for genomic analysis, RNA-Seq is suggesting new and revised transcriptome models. Thus far in all genomes surveyed including mouse (Cloonan et al., 2008; Mortazavi et al., 2008), human (Morin et al., 2008), yeast (Nagalakshmi et al., 2008), *Schizosaccharomyces pombe* (Wilhelm et al., 2008), and *A. thaliana* (Lister et al., 2008), RNA-Seq results suggest the existence of additional transcribed regions and allows the study of alternative splicing, as well as transcription of micro-RNA and other regulatory, noncoding RNA species (Wang et al., 2009).

While RNA-Seq findings suggest the transcriptome is more complex than previously acknowledged, this does not mean hybridization technology is becoming obsolete. It seems that sequencing and hybridization approaches can serve complimentary, rather than mutually exclusive roles (Liu, Lin, Jiang, Wang, & Xing, 2011). A comparison between ChIP-Chip and ChIP-PET technologies in mapping the transcription factor STAT1 binding regions in mammalian cells revealed a strong correlation between the two technologies for the most highly expressed genes (Coppee, 2008; Euskirchen et al., 2007). The two technologies showed less similarity in identifying more lowly expressed targets, but each method detected validated targets that were missed by the other method (Coppee, 2008; Euskirchen et al., 2007). Wilhlem et al. also reports that analyzing the yeast genome using both RNA-Seq and tiling arrays proved beneficial as each method is able to overcome weaknesses of the other method. Overall the two methods showed a Pearson's correlation of $r = 0.68$ (Wilhelm et al., 2008). As mentioned earlier, large numbers of overlapping transcripts means that genes can map equally well to multiple genomic locations when using RNA-Seq. In addition decreasing read-numbers at the 5′ ends because of olido(dT) priming can make determining the precise length of long genes a challenge with some applications of RNA-Seq technology. Hybridization data distinguished transcriptional direction and did not show 5′ bias, but the tiling arrays proved less sensitive due to background noise and a limited dynamic range (Wilhelm et al., 2008).

After nearly two decades of genome-wide transcriptional analysis since the invention of the microarray (Schena, 1996), there have been many advances in analysis, quantitation, and data synthesis from differential expression to networks, but consistency across technological platforms and predicting function from transcriptional profiles remain elusive goals. Thus, it seems the future of transcriptome mapping may include using multiple approaches to overcome weaknesses inherent in each individual approach. Finally, focusing consistent attention on some of the experimental design and technical issues raised in this chapter,

together with emphasis on network analysis combined across datasets, likely offer best principles for productive genomic analysis.

Acknowledgments

Work supported by NIAAA Grants: R01AA014717, U01AA016667, P20AA017828 (M. F. M.); 1F31AA021035-01 (M. A. O.); and 1F31AA020141-01 (B. N. C.).

REFERENCES

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, et al. Solid phase DNA amplification: Characterisation of primer attachment and amplification mechanisms. Nucleic Acids Research. 2000; 28(20):E87. [PubMed: 11024189]

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics. 2005; 6:55. [PubMed: 15766383]

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nature Biotechnology. 2006; 24(5):537–544.

Affymetrix. Technical note 701903, Revision 1: Guide to probe logarithmic intensity error (PLIER) estimation. Santa Clara, CA 95051 USA: Affymetrix, Inc.; 2005.

Affymetrix. Technical Note 701137, Revision 3: Statistical algorithms description document. Santa Clara, CA 95051 USA: Affymetrix, Inc.; 2002.

Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci G, Kornacker K. Chipping away at the chip bias: RNA degradation in microarray analysis. Nature Genetics. 2003; 35(4):292–293. [PubMed: 14647279]

Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ. GeneWeaver: A web-based system for integrative functional genomics. Nucleic Acids Research. 2012; 40(Database issue):1067–D1076.

Baker EJ, Jay JJ, Philip VM, Zhang Y, Li Z, Kirova R, et al. Ontological Discovery Environment: A system for integrating gene-phenotype associations. Genomics. 2009; 94(6):377–387. [PubMed: 19733230]

Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, et al. Submission of microarray data to public repositories. PLoS Biology. 2004; 2(9):E317. [PubMed: 15340489]

Barrett T, Edgar R. Gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. Methods in Enzymology. 2006; 411:352–369. [PubMed: 16939800]

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: Archive for functional genomics data sets–10 years on. Nucleic Acids Research. 2011; 39(Database issue):D1005–D1010. [PubMed: 21097893]

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: Archive for high-throughput functional genomic data. Nucleic Acids Research. 2009; 37(Database issue):D885–D890. [PubMed: 18940857]

Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. Bioinformatics. 2004; 20(1):105–114. [PubMed: 14693816]

Bennett B, Saba LM, Hornbaker CK, Kechris KJ, Hoffman P, Tabakoff B. Genetical genomic analysis of complex phenotypes using the PhenoGen website. Behavior Genetics. 2011; 41(4):625–628. [PubMed: 21184165]

Bertone P, Gerstein M, Snyder M. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology. 2005; 13(3): 259–274.

Bhandari P, Hill JS, Farris SP, Costin B, Martin I, Chan CL, et al. Chlo-ride intracellular channels modulate acute ethanol behaviors in Drosophila, Caenorhabditis elegans and mice. Genes, Brain, and Behavior. 2012; 11(4):387–397.

Bhave SV, Hornbaker C, Phang TL, Saba L, Lapadat R, Kechris K, et al. The PhenoGen informatics website: Tools for analyses of complex traits. BMC Genetics. 2007; 8:59. [PubMed: 17760997]

Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT. The Mouse Genome Database (MGD): Premier model organism resource for mammalian genomics and genetics. Nucleic Acids Research. 2011; 39(Database issue):D842–D848. [PubMed: 21051359]

Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(7):3960–3964. [PubMed: 12651960]

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genetics. 2001; 29(4):365–371. [PubMed: 11726920]

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on micro-bead arrays. Nature Biotechnology. 2000; 18(6):630–634.

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. The HGNC Database in 2008: A resource for the human genome. Nucleic Acids Research. 2008; 36(Database issue):D445–D448. [PubMed: 17984084]

Carter SL, Brechbuhler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 2004; 20(14):2242–2250. [PubMed: 15130938]

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, et al. Unbiased mapping of transcription factor binding sites along human chromo-somes 21 and 22 points to widespread regulation of noncoding RNAs. Cell. 2004; 116(4):499–509. [PubMed: 14980218]

Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Research. 2009; 19(2):336–346. [PubMed: 19056694]

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. Science. 1996; 274(5287):610–614. [PubMed: 8849452]

Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research. 2009; 37(Web Server Issue):W305–W311. [PubMed: 19465376]

Chen Z, Duan X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. Methods in Molecular Biology. 2011; 733:93–103. [PubMed: 21431765]

Chen J, Xu H, Aronow BJ, Jegga AG. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics. 2007; 8:392. [PubMed: 17939863]

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, et al. SGD: Saccharomyces Genome Database. Nucleic Acids Research. 1998; 26(1):73–79. [PubMed: 9399804]

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nature Genetics. 2005; 37(3):233–242. [PubMed: 15711545]

Chesler EJ, Lu L, Wang J, Williams RW, Manly KF. WebQTL: Rapid exploratory analysis of gene expression and genetic networks for brain and behavior. Nature Neuroscience. 2004; 7(5):485–486.

Chesler EJ, Wang J, Lu L, Qu Y, Manly KF, Williams RW. Genetic correlates of gene expression in recombinant inbred strains: A relational model system to explore neurobehavioral phenotypes. Neuroinformatics. 2003; 1(4):343–357. [PubMed: 15043220]

Chesler EJ, Wilson SG, Lariviere WR, Rodriguez-Zas SL, Mogil JS. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. Neuroscience and Biobehavioral Reviews. 2002; 26(8):907–923. [PubMed: 12667496]

Christodoulou DC, Gorham JM, Herman DS, Seidman JG. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease/edited by Frederick M. Ausubel … [et al.]. chapter 4 unit 4 1. Current Protocols in Molecular Biology. 2011

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. 2008; 5(7):613–619. [PubMed: 18516046]

Cloonan N, Grimmond SM. Transcriptome content and dynamics at single-nucleotide resolution. Genome Biology. 2008; 9(9):234. [PubMed: 18828881]

Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, Bareil C, et al. Impact of RNA degradation on gene expression profiles: Assessment of different methods to reliably determine RNA quality. Journal of Biotechnology. 2007; 127(4):549–559. [PubMed: 16945445]

Coppee JY. Do DNA microarrays have their future behind them? Microbes and Infection/Institut Pasteur. 2008; 10(9):1067–1071. [PubMed: 18662797]

Croning MD, Marshall MC, McLaren P, Armstrong JD, Grant SG. G2Cdb: The Genes to Cognition database. Nucleic Acids Research. 2009; 37(Database issue):D846–D851. [PubMed: 18984621]

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, et al. A high-resolution map of transcription in the yeast genome. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(14):5320–5325. [PubMed: 16569694]

Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, et al. The Comparative Toxicogenomics Database: Update 2011. Nucleic Acids Research. 2011; 39(Database issue):D1067–D1072. [PubMed: 20864448]

de Hoon M, Hayashizaki Y. Deep cap analysis gene expression (CAGE): Genome-wide identification of promoters, quantification of their expression, and network inference. BioTechniques. 2008; 44(5):627–628. 630, 632. [PubMed: 18474037]

Denisov, V.; Strong, W.; Walder, M.; Gingrich, J.; Wintz, H. Development and validation of RQI: An RNA Quality Indicator for the Experion (TM) Automated Electrophoresis System. B.-R. Laboratories. , editor. 2008.

Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, et al. Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proceedings of the National Academy of Sciences of the United States of America. 1996; 93(12):6025–6030. [PubMed: 8650213]

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(15): 8817–8822. [PubMed: 12857956]

Dudoit, S.; Yang, YH.; Callow, MJ.; Speed, TJ. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stanford, CA: Stanford University School of Medicine; 2000.

Durrant C, Swertz MA, Alberts R, Arends D, Moller S, Mott R, et al. Bio-informatics tools and database resources for systems genetics analysis in mice—A short review and an evaluation of future needs. Briefings in Bioinformatics. 2012; 13(2):135–142. [PubMed: 22396485]

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002; 30(1):207–210. [PubMed: 11752295]

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323(5910):133–138. [PubMed: 19023044]

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95(25):14863–14868. [PubMed: 9843981]

Erlich Y, Mitra PP, de la Bastide M, McCombie WR, Hannon GJ. Alta-Cyclic: A self-optimizing base caller for next-generation sequencing. Nature Methods. 2008; 5(8):679–682. [PubMed: 18604217]

Euskirchen GM, Rozowsky JS, Wei CL, Lee WH, Zhang ZD, Hartman S, et al. Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. Genome Research. 2007; 17(6):898–909. [PubMed: 17568005]

Ewart-Toland A, Balmain A. The genetics of cancer susceptibility: From mouse to man. Toxicologic Pathology. 2004; 32(Suppl. 1):26–30. [PubMed: 15209400]

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Research. 2006; 34(3):e22. [PubMed: 16473845]

Freudenberg J, Propping P. A similarity-based method for genome-wide predic-tion of disease-relevant human genes. Bioinformatics. 2002; 18(Suppl 2):S110–S115. [PubMed: 12385992]

Gardner D, Akil H, Ascoli GA, Bowden DM, Bug W, Donohue DE, et al. The neuroscience information framework: A data and knowledge environment for neuroscience. Neuroinformatics. 2008; 6(3):149–160. [PubMed: 18946742]

Geschwind DH. Sharing gene expression data: An array of options. Nature Reviews. Neuroscience. 2001; 2(6):435–438.

Goodman L. Unlimited access-limitless success. Genome Research. 2001; 11(5):637–638. [PubMed: 11337460]

Gorgels TG, Hu X, Scheffer GL, van der Wal AC, Toonstra J, de Jong PT, et al. Disruption of Abcc6 in the mouse: Novel insight in the pathogenesis of pseudoxanthoma elasticum. Human Molecular Genetics. 2005; 14(13):1763–1773. [PubMed: 15888484]

Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A, et al. The BioSample Database (BioSD) at the European Bioinformatics Institute. Nucleic Acids Research. 2012; 40(Database issue):D64–D70. [PubMed: 22096232]

Gregory BD, Belostotsky DA. Whole-genome microarrays: Applications and technical issues. Methods in Molecular Biology. 2009; 553:39–56. [PubMed: 19588100]

Gurskaya NG, Diatchenko L, Chenchik A, Siebert PD, Khaspekov GL, Lukyanov KA, et al. Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-myristate 13-acetate. Analytical Biochemistry. 1996; 240(1):90–97. [PubMed: 8811883]

Harbers M, Carninci P. Tag-based approaches for transcriptome research and genome annotation. Nature Methods. 2005; 2(7):495–502. [PubMed: 15973418]

Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: A comprehensive resource for nematode research. Nucleic Acids Research. 2010; 38(Database issue):D463–D467. [PubMed: 19910365]

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008; 320(5872):106–109. [PubMed: 18388294]

He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nature Methods. 2010; 7(10):807–812. [PubMed: 20852648]

Hillard AT, Miller JE, Fraley ER, Horvath S, White SA. Molecular microcircuitry underlies functional specification in a basal ganglia circuit dedicated to vocal learning. Neuron. 2012; 73:537–552. [PubMed: 22325205]

Hoffman PL, Bennett B, Saba LM, Bhave SV, Carosone-Link PJ, Hornbaker CK, et al. Using the Phenogen website for 'in silico' analysis of morphine-induced analgesia: Identifying candidate genes. Addiction Biology. 2011; 16(3):393–404. [PubMed: 21054686]

Hovatta I, Tennant RS, Helton R, Marr RA, Singer O, Redwine JM, et al. Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature. 2005; 438(7068):662–666. [PubMed: 16244648]

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009; 4(1):44–57.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. Cell. 2000; 102(1):109–126. [PubMed: 10929718]

Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E, et al. Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. Nucleic Acids Research. 2005; 33(6):e56. [PubMed: 15800207]

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research. 2003; 31(4):e15. [PubMed: 12582260]

Irizarry RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Statistical Methods in Medical Research. 2009; 18(6):565–575. [PubMed: 20048385]

Jensen LJ, Saric J, Bork P. Literature mining for the biologist: From information retrieval to biological discovery. Nature Reviews. Genetics. 2006; 7(2):119–129.

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray ex-pression data using empirical Bayes methods. Biostatistics. 2007; 8(1):118–127. [PubMed: 16632515]

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Research. 2004; 14(3):331–342. [PubMed: 14993201]

Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Research. 2000; 28(22):4552–4557. [PubMed: 11071945]

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science. 2002; 296(5569):916–919. [PubMed: 11988577]

Kapushesky M, Kemmeren P, Culhane AC, Durinck S, Ihmels J, Korner C, et al. Expression Profiler: Next generation—An online platform for analysis of microarray data. Nucleic Acids Research. 2004; 32(Web Server Issue):W465–W470. [PubMed: 15215431]

Kennedy RE, Archer KJ, Miles MF. Empirical validation of the S-score algorithm in the analysis of gene expression data. BMC Bioinformatics. 2006; 7(1):154. [PubMed: 16545131]

Kennedy RE, Kerns RT, Kong X, Archer KJ, Miles MF. SScore: An R package for detecting differential gene expression without gene expression summaries. Bioinformatics. 2006; 22(10): 1272–1274. [PubMed: 16574698]

Kerns RT, Miles MF. Microarray analysis of ethanol-induced changes in gene expression. Methods in Molecular Biology. 2008; 447:395–410. [PubMed: 18369932]

Kerns RT, Ravindranathan A, Hassan S, Cage MP, York T, Sikela JM, et al. Ethanol-responsive brain region expression networks: Implications for behav-ioral responses to acute ethanol in DBA/2J versus C57BL/6J mice. The Journal of Neuroscience. 2005; 25(9):2255–2266. [PubMed: 15745951]

Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genetical Research. 2007; 89(5–6):509–514. [PubMed: 18976541]

Kim SY, Lee JW, Sohn IS. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Statistical Methods in Medical Research. 2006; 15(1):3–20. [PubMed: 16477945]

Korostynski M, Kaminska-Chowaniec D, Piechota M, Przewlocki R. Gene expression profiling in the striatum of inbred mouse strains with distinct opioid-related phenotypes. BMC Genomics. 2006; 7:146. [PubMed: 16772024]

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature Methods. 2009; 6(4):291–295. [PubMed: 19287394]

Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science. 1988; 241(4869):1077–1080. [PubMed: 3413476]

Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. [PubMed: 19114008]

Le-Niculescu H, McFarland MJ, Mamidipalli S, Ogden CA, Kuczenski R, Kurian SM, et al. Convergent Functional Genomics of bipolar disorder: From animal model pharmacogenomics to human genetics and biomarkers. Neuroscience and Biobehavioral Reviews. 2007; 31(6):897–903. [PubMed: 17614132]

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Methods. 2010; 7(9): 709–715. [PubMed: 20711195]

Li X, Inoue M, Reed DR, Huque T, Puchalski RB, Tordoff MG, et al. High-resolution genetic mapping of the saccharin preference locus (Sac) and the putative sweet taste receptor (T1R1) gene (Gpr70) to mouse distal Chromosome 4. Mammalian Genome: Official Journal of the International Mammalian Genome Society. 2001; (1):12. 13–16.

Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(1):31–36. [PubMed: 11134512]

Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science. 1992; 257(5072):967–971. [PubMed: 1354393]

Lipska BK, Deep-Soboslay A, Weickert CS, Hyde TM, Martin CE, Herman MM, et al. Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. Biological Psychiatry. 2006; 60(6):650–658. [PubMed: 16997002]

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Ara-bidopsis. Cell. 2008; 133(3):523–536. [PubMed: 18423832]

Liu S, Lin L, Jiang P, Wang D, Xing Y. A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Research. 2011; 39:578–588. [PubMed: 20864445]

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology. 1996; 14(13):1675–1680.

Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, et al. FRT-Seq: Amplification-free, strand-specific transcriptome sequencing. Nature Methods. 2010; 7(2):130–132. [PubMed: 20081834]

Marguerat S, Wilhelm BT, Bahler J. Next-generation sequencing: Applications beyond genomes. Biochemical Society Transactions. 2008; 36(Pt 5):1091–1096. [PubMed: 18793195]

Marshall E. Getting the noise out of gene arrays. Science. 2004; 306(5696):630–631. [PubMed: 15499004]

Martin JA, Wang Z. Next-generation transcriptome assembly. Nature Reviews. Genetics. 2011; 12(10):671–682.

Metzker ML. Emerging technologies in DNA sequencing. Genome Research. 2005; 15(12):1767–1776. [PubMed: 16339375]

Metzker ML. Sequencing in real time. Nature Biotechnology. 2009; 27(2):150–151.

Metzker ML. Sequencing technologies—The next generation. Nature Reviews. Genetics. 2010; 11(1): 31–46.

Mieczkowski J, Tyburczy ME, Dabrowski M, Pokarowski P. Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements. BMC Bioinformatics. 2010; 11:104. [PubMed: 20181266]

Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. Applications of DNA tiling arrays for whole-genome analysis. Genomics. 2005; 85(1):1–15. [PubMed: 15607417]

Mogil JS, Wilson SG, Chesler EJ, Rankin AL, Nemmani KV, Lariviere WR, et al. The melanocortin-1 receptor gene mediates female-specific mechanisms of analgesia in mice and humans. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(8): 4867–4872. [PubMed: 12663858]

Moore KJ. Utilization of mouse models in the discovery of human disease genes. Drug Discovery Today. 1999; 4(3):123–128. [PubMed: 10322264]

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques. 2008; 45(1):81–94. [PubMed: 18611170]

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008; 5(7):621–628. [PubMed: 18516045]

Nadler JJ, Zou F, Huang H, Moy SS, Lauder J, Crawley JN, et al. Large-scale gene expression differences across brain regions and inbred strains correlate with a behavioral phenotype. Genetics. 2006; 174(3):1229–1236. [PubMed: 16980393]

Nagalakshmi U, Waern K, Snyder M. RNA-Seq: A method for comprehensive transcriptome analysis. chapter 4 p. Unit 4 11. Current Protocols in Molecular Biology / edited by Frederick M. Ausubel … [et al.]. 2010:1–13.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; (5881):320. 1344–1349. [PubMed: 18420917]

Ogden CA, Rich ME, Schork NJ, Paulus MP, Geyer MA, Lohr JB, et al. Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: An expanded convergent functional genomics approach. Molecular Psy-chiatry. 2004; 9(11):1007–1029.

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, et al. Functional organization of the transcriptome in human brain. Nature Neuroscience. 2008; 11(11):1271–1282.

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, et al. ArrayExpress update—From an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Research. 2009; 37(Database issue):D868–D872. [PubMed: 19015125]

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress—A public database of microarray experiments and gene expression profiles. Nucleic Acids Research. 2007; 35(Database issue):D747–D750. [PubMed: 17132828]

Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, et al. ArrayExpress update—An archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Research. 2011; 39(Database issue):D1002–D1004. [PubMed: 21071405]

Pavlidis P, Noble WS. Analysis of strain and regional variation in gene expression in mouse brain. Genome Biology. 2001; 2(10) RESEARCH0042.

Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD re-combinant inbred lines from advanced intercross populations in mice. BMC Genetics. 2004; 5:7. [PubMed: 15117419]

Perkins AD, Langston MA. Threshold selection in gene co-expression networks using spectral graph theory techniques. BMC Bioinformatics. 2009; 10(Suppl 11):S4. [PubMed: 19811688]

Phillips TJ, Belknap JK, Hitzemann RJ, Buck KJ, Cunningham CL, Crabbe JC. Harnessing the mouse to unravel the genetics of human disease. Genes, Brain, and Behavior. 2002; 1(1):14–26.

Ponomarev I, Wang S, Zhang L, Harris RA, Mayfield RD. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. The Journal of Neuroscience. 2012; 32(5):1884–1897. [PubMed: 22302827]

Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends in Genetics: TIG. 2008; 24(3):142–149. [PubMed: 18262676]

Popova T, Mennerich D, Weith A, Quast K. Effect of RNA quality on transcript intensity levels in microarray analysis of human post-mortem brain tissues. BMC Genomics. 2008; 9:91. [PubMed: 18298816]

Quackenbush J. Computational analysis of microarray data. Nature Reviews. Genetics. 2001; 2(6): 418–427.

Reimers M. Statistical analysis of microarray data. Addiction Biology. 2005; 10(1):23–35. [PubMed: 15849016]

Reinartz J, Bruyns E, Lin JZ, Burcham T, Brenner S, Bowen B, et al. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Briefings in Functional Genomics & Proteomics. 2002; 1(1):95–104. [PubMed: 15251069]

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. Analytical Biochemistry. 1996; (1):242. 84–89.

Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science. 1998; 281(5375):363–365. [PubMed: 9705713]

Rosen GD, Chesler EJ, Manly KF, Williams RW. An informatics approach to systems neurogenetics. Methods in Molecular Biology. 2007; 401:287–303. [PubMed: 18368372]

Rosen GD, La Porte NT, Diechtiareff B, Pung CJ, Nissanov J, Gustafson C, et al. Informatics center for mouse genomics: The dissection of complex traits of the nervous system. Neuroinformatics. 2003; 1(4):327–342. [PubMed: 15043219]

Roybal K, Theobold D, Graham A, DiNieri JA, Russo SJ, Krishnan V, et al. Mania-like behavior induced by disruption of CLOCK. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(15):6406–6411. [PubMed: 17379666]

Ruan Y, Le Ber P, Ng HH, Liu ET. Interrogating the transcriptome. Trends in Biotechnology. 2004; 22(1):23–30. [PubMed: 14690619]

Rustici G, Kapushesky M, Kolesnikov N, Parkinson H, Sarkans U, Brazma A. Data storage and analysis in ArrayExpress and Expression Profiler. chapter 7 unit 7 1. Current Protocols in Bioinformatics/editoral board, Andreas D. Baxevanis … [et al.]. 2008

Saba L, Bhave SV, Grahame N, Bice P, Lapadat R, Belknap J, et al. Candidate genes and their regulatory elements: Alcohol preference and tolerance. Mammalian Genome: Official Journal of the International Mammalian Genome Society. 2006; 17(6):669–688. [PubMed: 16783646]

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. 1975; 94(3):441–448. [PubMed: 1100841]

Schena M. Genome analysis with gene expression microarrays. BioEssays. 1996; 18(5):427–431. [PubMed: 8639166]

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270(5235):467–470. [PubMed: 7569999]

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: An RNA integrity number for assigning integrity values to RNA measurements. BMC Molecular Biology. 2006; 7:3. [PubMed: 16448564]

Shendure J. The beginning of the end for microarrays? Nature Methods. 2008; 5(7):585–587. [PubMed: 18587314]

Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008; (10):26. 1135–1145.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(26):15776–15781. [PubMed: 14663149]

Shirley RL, Walter NA, Reilly MT, Fehr C, Buck KJ. Mpdz is a quantitative trait gene for drug withdrawal seizures. Nature Neuroscience. 2004; 7(7):699–700.

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314(5797):268–274. [PubMed: 16959974]

Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003; 31(4):265–273. [PubMed: 14597310]

Sprague J, Bayraktaroglu L, Bradford Y, Conlin T, Dunn N, Fashena D, et al. The Zebrafish Information Network: The zebrafish model organism database provides expanded support for genotypes and phenotypes. Nucleic Acids Research. 2008; 36(Database issue):D768–D772. [PubMed: 17991680]

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(16):9440–9445. [PubMed: 12883005]

Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003; 302(5643):249–255. [PubMed: 12934013]

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321(5891):956–960. [PubMed: 18599741]

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research. 2011; (Database issue):39. D561–D568.

Taylor BA, Wnek C, Kotlus BS, Roemer N, MacTaggart T, Phillips SJ. Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. Mammalian Genome: Official Journal of the International Mammalian Genome Society. 1999; 10(4):335–348. [PubMed: 10087289]

Thornblad TA, Elliott KS, Jowett J, Visscher PM. Prioritization of positional candidate genes using multiple web-based software tools. Twin Research and Human Genetics. 2007; 10(6):861–870. [PubMed: 18179399]

Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Research. 2005; 33(5): 1544–1552. [PubMed: 15767279]

Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T. DNA ligases: Structure, reaction mechanism, and function. Chemical Reviews. 2006; 106(2):687–699. [PubMed: 16464020]

Torkamani A, Dean B, Schork NJ, Thomas EA. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Research. 2010; 20(4):403–412. [PubMed: 20197298]

Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nature Biotechnology. 2009; 27(5):455–457.

Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S, et al. Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Research. 2009; 37(7):2249–2263. [PubMed: 19237398]

Turner FS, Clutterbuck DR, Semple CA. POCUS: Mining genomic sequence annotation to predict disease genes. Genome Biology. 2003; 4(11):R75. [PubMed: 14611661]

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(9):5116–5121. [PubMed: 11309499]

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, et al. FlyBase: Enhancing Drosophila Gene Ontology annotations. Nucleic Acids Research. 2009; 37(Database issue):D555–D559. [PubMed: 18948289]

Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ. The Rat Genome Database, update 2007—Easing the path from disease to data and back again. Nucleic Acids Research. 2007; 35(Database issue):D658–D662. [PubMed: 17151068]

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Research. 2008; 18(7):1051–1063. [PubMed: 18477713]

van Bokhoven H, Celli J, Kayserili H, van Beusekom E, Balci S, Brussel W, et al. Mutation of the gene encoding the ROR2 tyrosine kinase causes autosomal recessive Robinow syndrome. Nature Genetics. 2000; 25(4):423–426. [PubMed: 10932187]

van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Reports. 2004; 5(3):280–284. [PubMed: 14968131]

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995; 270(5235):484–487. [PubMed: 7570003]

Verdugo RA, Deschepper CF, Munoz G, Pomp D, Churchill GA. Importance of randomization in microarray experimental designs with Illumina platforms. Nucleic Acids Research. 2009; 37(17): 5610–5618. [PubMed: 19617374]

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research. 2005; 33(Database issue):D433–D437. [PubMed: 15608232]

Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews. Genetics. 2009; 10(1):57–63.

Wang J, Williams RW, Manly KF. WebQTL: Web-based complex trait analysis. Neuroinformatics. 2003; 1(4):299–308. [PubMed: 15043217]

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453(7199): 1239–1243. [PubMed: 18488015]

Wolen AR, Phillips CA, Langston MA, Putman AH, Vorster PJ, Bruce NA, et al. Genetic dissection of acute ethanol responsive gene networks in prefrontal cortex: Functional and mechanistic implications. PloS One. 2012; 7(4):e33575. [PubMed: 22511924]

Wolstenholme JT, Warner JA, Capparuccini MI, Archer KJ, Shelton KL, Miles MF. Genomic analysis of individual differences in ethanol drinking: Evidence for non-genetic factors in C57BL/6 mice. PloS One. 2011; 6(6):e21100. [PubMed: 21698166]

Yang YH, Speed T. Design issues for cDNA microarray experiments. Nature Reviews. Genetics. 2002; 3(8):579–588.

Young LJ. Oxytocin and vasopressin as candidate genes for psychiatric disorders: Lessons from animal models. American Journal of Medical Genetics. 2001; 105(1):53–54. [PubMed: 11424998]

Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics. 2011; 38(3):95–109. [PubMed: 21477781]

Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005; 4 Article 17.

Zhang L, Wang L, Ravindranathan A, Miles MF. A new algorithm for analysis of oligonucleotide arrays: Application to expression profiling in mouse brain regions. Journal of Molecular Biology. 2002; 317:225–235. [PubMed: 11902839]

Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, et al. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenetic and Genome Research. 2004; 105(2–4):363–374. [PubMed: 15237224]

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics. 2008; 40(7):854–861. [PubMed: 18552845]

Zhu M, Zhao S. Candidate gene identification approach: Progress and challenges. International Journal of Biological Sciences. 2007; 3(7):420–727. [PubMed: 17998950]

Table 5.1

Hybridization Versus Sequencing Methods for Transcriptome Quantification

Method	Description	Advantages	Disadvantages
Microarray	Non-tiling arrays (biased annotation-dependent arrays)	Relatively inexpensive and high throughput; can help to identify low abundance or rare transcripts	Do not cover the entire genome; rely upon existing knowledge regarding genome sequences; comparing expression levels across experiments can require complex normalization; high background levels; limited dynamic range of detection due to background and saturation of signals
	Tiling arrays (non-biased whole-genome arrays)	Can discover new genes and exons; high throughput; cover the entire genome; can help to identify low abundance or rare transcripts; allow for the analysis of various genomic features including alternative splicing, RNA-binding protein transcript target identification, methylome analysis, polymorphism analysis and complete resequencing of a genome	Limited in the number of probe features included on a chip and how many chips are required to cover an entire genome; techniques for organizing and analyzing the large amount of data obtained are lacking; high background levels; limited dynamic range of detection due to background and saturation of signals; require large amounts of input RNA
Tag-based sequencing	Serial analysis of gene expression (SAGE)	Provides digital counts of transcript abundance; assigns expression values without the need for probe design; data analysis is well established	Based on expensive Sanger sequencing technology; short tags cannot be mapped to the reference genome; only a portion of the transcript is analyzed; isoforms are indistinguishable from one another; may not be able to identify transcripts that are low abundance, expressed in rare cell types or in response to specific stimuli; MPSS-relies on sophisticated instrumentation limiting its' general use
	Cap analysis of gene expression (CAGE)	Provides digital counts of transcript abundance; Maps transcriptional start points	
	Massively parallel signature sequencing (MPSS)	Provides digital counts of transcript abundance; no requirement that genes be characterized prior to the experiment; very sensitive; datasets are in a digital format that simplifies data management and analysis	
RNA-Seq		Provides expression values without the need for probe design, can reveal the precise location of transcriptional boundaries, can be used for polymorphism analysis, low background; no upper limit for quantification; highly accuracy; highly reproducible; does not require sophisticated normalization methods across experiments; capable of *de novo* annotation	Can require large amounts of RNA; provides short reads that may map to many regions of the genome; library preparation can bias the outcome of sequencing; the directionality of the cDNA fragment is lost; faces challenges in the development of efficient methods to process, retrieve, and store large amounts of data

Table 5.2

Platforms used in next-generation sequencing technologies

Platform	Template preparation	Sequencing method	Advantages	Disadvantages
Roche/454's GS FLX Titanium	emPCR	PS	Longest short reads among all NGS platforms (400–600 bp reads); fast run times	High reagent costs; high error rates in homopolymer repeats
Illumina/HiSeq 2000	Solid-phase	Four-color CRT	Currently, the most widely used platform in the field; most adaptable, easiest to use sequencing platform	Substitutions are the most common error type; underrepresentation of AT-rich and GC-rich regions
Life/APG's support oligonucleotide ligation detection (SOLiD) 3	emPCR	Cleavable probe SBL	Two-base encoding provides inherent error correct so the technology is highly accurate; emPCR can achieve high data densities	DNA library preparation procedures can be tedious and time consuming; Underrepresentation of AT-rich and GC-rich regions; long run times; emPCR can be cumbersome and challenging
Polonator G.007	emPCR	Noncleavable probe SBL	Least expensive platform; programmable instrument enabling user innovation; emPCR can achieve high data densities	Shortest NGS read lengths; emPCR can be cumbersome and challenging
Helicos BioSciences HeliScope	Single molecule	One-color CRT	Can directly sequence single DNA molecules without amplification; fast sequencing speed; can read extremely long sequences	High error rates
Pacific Biosciences (not yet available)	Single molecule	Real-time sequencing	Has potential for reads greater than 1 kb; does not require PCR amplification prior to sequencing	Highest error rates compared with other NGS technologies

emPCR, emulsion PCR; PS, pyrosequencing; CRT, cyclic reversible termination; SBL, sequencing by ligation.