Published in final edited form as:

Adv Genet. 2008 ; 60: 571–601. doi:10.1016/S0065-2660(07)00420-8.

Integrating Global Gene Expression Analysis and Genetics

Charles R. Farber and Aldons J. Lusis

Dept. of Medicine, Dept. of Microbiology, Immunology and Molecular Genetics, and Dept. of Human Genetics, University of California, Los Angeles, CA 900095

I. INTRODUCTION

Biological systems can be thought of as a series of stages (commonly referred to using “-omics” nomenclature) that can be interrogated using specific technologies (Figure 1). These stages include DNA (genome), RNA (transcriptome), proteins (proteome), metabolites (metabolome) and phenotypes (phenome), among many others. Although each stage can be considered individually a great deal of crosstalk between them is required for proper cellular and physiological function. As mentioned above, classical genetic studies link genes (genome) to disease (phenome) without considering other stages. However, spurred by technological advances in the ability to perform bioassays in a massively parallel fashion, the sequencing of the human genome, and the development of statistical methodologies, researchers now have the capacity to leverage information from other levels of the system to better understand the role of genetic perturbations in disease. Currently, the transcriptome has proven the most accessible with regards to high-throughput analysis. The transcriptome is most commonly viewed as the full complement of mRNA species present in a given cell type or tissue at a defined time in development. However, recent data suggests other RNA species such as noncoding RNAs (microRNAs, snRNAs, etc.) are important information carriers which can have profound effects on quantitative traits [1,2].

The decoding of the human (and other model organisms such as the mouse and rat) blueprint represents an astonishing scientific achievement and has provided a comprehensive view of the first stage of the human biological system [3-6]. One immediate application of this “genetic parts list” was the development of DNA microarrays, which are now the most widely used tool for global gene expression profiling. DNA microarrays with the capacity to profile the entire transcriptome (at least the part we have correctly identified as transcribed) now exist and have been used in a plethora of applications. To illustrate their growing utility a PubMed search at the National Center for Biotechnology Information (NCBI) using the search string “microarray AND gene AND expression” returned 14,331 articles, 926 (6.5%) of which were published within 90 days of this search (April 16, 2007).

Probably the most significant applications of expression array profiling to common disease are in the area of cancer. Expression signatures of cancers have been used to subdivide cancers and to predict survival and responses to specific drugs. Recently, Golub and colleagues [7], have proposed the development of a resource they term the “connectivity map”. They propose to use mRNA expression assayed on DNA microarrays to determine genomic signatures that describe all biologic states – physiologic, disease, or those induced with chemicals or genetic constructs. The connectivity map would be a large public database of such signatures along with tools to determine pattern matching of similarities among these signatures.

The last decade has seen a paradigm shift in our ability to confront disease. The tools now exist to transition from “one gene at a time” to more global “systems-level” approaches which promise an unprecedented understanding of affected and normal states. Global snapshots of the transcriptome can now be linked to both disease status and genetic polymorphisms, significantly increasing our ability to pinpoint master disease regulators. This transition will certainly lead to more creative and effective therapeutic intervention programs that are designed to confront head on instead of sidestepping the complexity of disease. The purpose of this chapter is to describe one aspect of this transition; the use of gene expression analysis to the context of common disease. The discussion begins with the platform for change – DNA microarrays. Our aim is to highlight technical and data analysis issues pertaining to their use in genetic studies. Our discussion then shifts to ways in which microarray technologies have and can be used to prioritize candidate genes based on potential relevance to disease. The last sections will discuss recent advances in the integration of gene expression and genetics, as well as novel analytical approaches in the development of gene co-expression networks.

II. TECHNICAL AND EXPERIMENTAL DESIGN ISSUES FOR MICROARRAYS

Approaches in systems biology rely on the collection of highly-parallel information from different biological levels which can be used to infer system function in the face of genetic and environmental perturbations. The two levels which are the most amenable to comprehensive screening are the genome and transcriptome. This is due to their “relative” lack of complexity and the complementary nature of nucleic acids. In contrast, technological challenges remain for the interrogation of many levels, such as the proteome which is not only comprised of components (individual proteins) but also many regulatory relationships (posttranslational modifications and protein-protein interactions).

Several different technologies exist for whole transcriptome profiling and detection of differentially expressed genes, including serial analysis of gene expression (SAGE) [8], massively parallel signature sequencing (MPSS) [9], differential display [10], cDNA representational difference analysis [11] and DNA microarrays [12,13]. Although each is useful in certain applications, DNA microarrays are by far the most widely used. Similar to Northern blotting, the basis of microarrays is hybridization between complementary nucleic acids. In a Northern blot, a labeled probe is hybridized to a membrane containing an RNA sample [14] and the amount of probe that binds its complementary RNA is used to compare gene expression across samples. In essence, DNA microarrays simultaneously perform Northern blots for every gene in the genome.

In general terms, a DNA microarray is a collection of DNA sequences covalently attached to a stable substrate such as a glass slide, silicon wafer or silica beads. Spots of DNA (referred to as probes and typically consisting of cDNAs or oligonucleotides) represent specific genes and are arrayed in a grid-like pattern across the solid surface. In the context of gene expression analysis, the target is comprised of a population of cDNA or cRNA copies of mRNAs, that are labeled and applied directly to the microarray. On the array, complementary probe-target pairs bind through hybridization. After hybridization microarrays are scanned and signal intensity is quantified for each spot or feature. This signal is proportional to the amount of target present in the starting RNA sample and is used as a proxy for the actual mRNA levels either in relative or absolute terms, depending on microarray platform. Although DNA microarrays are most commonly used for gene profiling they can also be used for a plethora of other applications such as comparative genomic hybridization (CGH), genome wide chromatin immunoprecipitation (ChIP-chip), genomic re-sequencing, and single nucleotide polymorphism (SNP) genotyping [15,16].

1. DNA microarray platforms

Two general types of microarray platforms are currently in use, one- and two-color [17]. The most significant difference between one- and two-color microarrays is the type of hybridization. Two-color arrays are simultaneously hybridized using two samples (control and experimental) each tagged with a different label. Cyanine (Cy3 and Cy5) labeled deoxynucleotide triphosphate incorporated into cDNA is the most common fluorescent label used in two-color systems [18]. After hybridization a scanner is used to measure the amount of fluorescent target bound to each probe. If the ratio of experimental to control intensity for a gene is significantly more or less than one, the transcript level in the experimental sample, is up- or down-regulated, respectively. In contrast, a single sample is hybridized to a one-color array and unlike two-color systems several different types of target and target labeling protocols exist. In general the signal intensity for each probe is a direct readout of gene expression in absolute terms. A hypothetical experiment using one-color microarrays is illustrated in Figure 2.

In the following sections we discuss details for the most widely used commercial platforms. It should be noted, however, that many researchers use “homemade” arrays. These are almost always of the two-color version and are made using printing devices which deposit spots of DNA onto glass slides [19]. In addition there are technologies which are still in early stages of commercialization, but are worth noting. These include NimbleGen and CombiMatrix which have developed novel *in situ* synthesis (synthesis of the probe directly on the slide) methods; digitally controlled micromirrors and electrode-directed synthesis, respectively [16]. Both platforms offer significant advantages for generating custom microarrays.

a. Affymetrix—The Affymetrix GeneChip array was one of the first commercially available whole genome expression profiling technologies and is still in widespread use today. One advantage of the GeneChip array is the extremely high feature density, in excess of 1 million features/chip, relative to other platforms [20]. This density is possible because of photolithography, a unique method of *in situ* synthesis [21]. The process of manufacturing a GeneChip begins by adhering linker molecules with photolabile protecting groups to the surface of a silica wafer. A photolithographic mask is applied and light is introduced, removing the protecting groups at defined positions depending on the predetermined sequence and location of the oligonucleotide probes to be synthesized. Protected deoxynucleosides are added, which covalently attach to the unprotected linker, and this process is repeated with new masks until all 25-mer oligonucleotide probe sequences are fully synthesized [20].

Another unique attribute of the GeneChip arrays is the inclusion of both perfect match (PM) and mismatch (MM) probes. The PM component of the probe pair is identical to a complementary sequence in the target sample, whereas the MM probe contains a mismatch at the central nucleotide. In the most common array design 11 probe pairs (11 PM and 11 corresponding MM probes) per gene are designed within the 600 bp most proximal to the polyadenylation site. In theory, signal intensity originating from the MM probes should represent background noise and can be used to correct the raw intensities of PM probes. During a gene expression experiment biotinylated cRNA is hybridized to the array and stained with a fluorescent streptavidin-phycoerythrin conjugate which binds biotin. The GeneChip is scanned and the intensity of each probe is determined. A number of software packages as well as libraries for the Bioconductor software implement algorithms for calculating signal intensities from GeneChip arrays [22,23].

b. Illumina—Illumina Universal BeadArrays represent a novel approach to genomic applications including gene expression profiling. There are two general types of BeadArrays, the Sentrix Array Matrix (SAM) and Sentrix BeadChip. The SAM is used for the analysis a specific gene sets (on the order of 1500 genes per sample) whereas BeadChips are used for whole genome profiling. For the purpose of our discussion we will focus on details of the BeadChip, although SAM arrays are identical in many technical aspects.

The Sentrix BeadChip consists of a silicon coated chip with millions of microscopic wells etched in a regular pattern along its surface [24]. Each well is approximately 3 μm in diameter and is designed to capture and hold a signal bead. BeadChip beads are impregnated with approximately 700,000 covalently attached two-part oligonucleotide probes. The first part or sequence closest to the bead is a unique 29-mer address sequence used for array decoding and the second part is a gene specific sequence [25]. A pool of all bead types is applied to each array and individual beads become randomly seated in microwells.

Due to the randomness of bead placement, BeadChip arrays are decoded to discern the identity of each bead type [26]. This is accomplished using the 29-mer address sequence. In the decoding process, decoder oligo pools are constructed with a set of fluorescently labeled oligonucleotides complementary to the address sequences for a subset of all bead types. Decoder pools are hybridized and the fluorescence intensity is measured for all beads across the array. In the second stage, the BeadChip is stripped and a different decoder pool is hybridized. This process is repeated for the number of stages needed to decode all possible bead types and at the end of this process a unique signature for each bead is generated. This signature provides the sequence identity of each bead on the array [26].

One of the advantages of the BeadChip is its extremely high feature density [25]. This high density allows for the processing of multiple samples per BeadChip on a substrate the size of a typical microscope slide, significantly decreasing cost per sample. For human and mouse, two different platforms are commercially available. The first is a six sample format which quantitates the expression of over 40,000 transcripts, and the second is an eight sample format which analyzes over 20,000 genes. In addition, there is on average a 30-fold redundancy per bead type present on each array.

The target sample for each decoded BeadChip array is generated and labeled in a process similar to that described above for GeneChips. For data analysis a BeadStudio analysis software package is available which is capable of data normalization and analysis. In addition two libraries for the Bioconductor software, Beadarray and BeadExplorer (www.bioconductor.org) [22], have been developed to assist in the analysis of BeadChip data.

c. Other platforms—A number of other commercial platforms exist, including Agilent, Applied Biosystems and Eppendorf (Table 1). Recently, these platforms have been compared as part of the MicroArray Quality Control (MAQC) Project [27]. For this project, expression data on four titration pools from two distinct reference RNA samples were generated at multiple test sites using a variety of microarray-based platforms. This paper provides a reference to an investigator by which inter-platform consistency and inter-platform concordance can be evaluated. For example the study showed that, in these samples, the differentially expressed genes averaged approximately 89% overlap between test sites using the same platform, and approximately 74% across one-color microarray platforms. Significant differences in various dimensions of performance between microarray platforms were noted.

2. Microarray data analysis

A detailed description of procedures for analysis of microarray data is beyond the scope of this review. However, the reader is referred to a number of excellent reviews and volumes dealing with the subject [28-30].

The first step in the analysis of microarray data involves image analysis to convert the numerous number of pixels into expression values for each gene. Image analysis includes filtering to “clean” images, gridding and segmentation to define the region to be quantitated, and quantification of the fluorescence intensity.

The second step involves normalization, the process of removing systematic bias as a result of experimental artifacts, from the data. One way of doing this is the analysis of variance (ANOVA) method, but a problem with ANOVA is that it is computationally intensive. Generally, microarrays are analyzed using a sequential approach, in which the normalization is done before any further analysis. Microarray data are usually scaled using a logarithmic scale. The logarithmic function can be done without loss of information regarding the original signal and it is the most natural scale to describe fold changes. Generally, microarrays are normalized with each other by assuming an average overall intensity for each array. Microarrays are also frequently equipped with control spots that can also be used for normalizing data. After normalization, two-color data are generally reported as the logarithm of the expression ratio and one color data as the logarithm of the intensity. Other important aspects of normalization exist, depending on the array platform, including background correction and spatial normalization.

After normalization, one typically identifies genes that are differentially expressed. The ability to distinguish such differences depends on the variance of the data and on the number of arrays analyzed per sample. A common preference for the analysis of differential expression is to use a t-test or a threshold fold-change. The problem of multiple comparisons is best addressed by analysis of “false discovery rates” [31].

Another common goal is to identify genes that show similar patterns of expression, using statistical methods generally referred to as “cluster analyses”. “Similarity” of expression patterns is mathematically defined using an “expression vector” for each gene that represents its location in “expression space”. In such an analysis, each experiment represents a separate entry in space and the $\log_2$ intensity or $\log_2$ ratio measured for a gene represents its geometric coordinate. For instance, in a study with three experiments, the $\log_2$ expression for a given gene in Experiment 1 is its X coordinate, the $\log_2$ expression in Experiment 2 is its Y coordinate, and the $\log_2$ expression in Experiment 3 is its Z coordinate. Thus, one can represent all the information about a gene by a point in the X, Y and Z expression space. Another gene, with similar $\log_2$ expression values for each experiment will be represented by a spatially nearby point. The most widely used cluster analyses are hierarchical, with an increasing number of nested classes. Non-hierarchical clustering techniques which simply partitions genes into different clusters can also be applied. Advantages of hierarchical clustering includes its relative simplicity and straightforward visualization [29,30].

III. IDENTIFYING DISEASE CANDIDATES USING DNA MICROARRAYS

Expression array profiling can provide clues which aid in the identification of genes underlying complex traits. For example the expression of the phenotype in different tissues can be compared with the expression of genes in different tissues. Additionally the differential expression of genes between diseased individuals and normal individuals can be determined. In the sections below we outline ways in which expression profiling can be used to better understand the genetic factors which regulate disease.

1. Gene expression catalogs

Complex human disease is typically not confined to alterations in one tissue or cell type. In contrast, most disease states are the result of multiple perturbations involving an array of tissue and organ systems. However, in almost all cases a relatively small set of relevant tissues can be considered the most likely drivers of disease. For example, a major cause of osteoporosis is an increase in the bone remodeling rate [32]. The bone remodeling rate is the relative activity of two cell populations, the osteoclasts which resorb old bone and osteoblasts which form new bone [33]. Although changes in the bone remodeling rate are inherent to bone, gene expression changes in other tissues, such as the bone marrow and adipose, are known to influence this process. Therefore, a comprehensive catalog of gene expression of every gene in the genome, in bone, bone marrow and adipose, would constitute a valuable filter for prioritizing candidate osteoporosis regulators.

The earliest versions of human gene expression catalogs were collections of expressed sequence tags (ESTs). ESTs are generated by sequencing clones from a complementary DNA (cDNA) library and represent a fragment of a specific mRNA. cDNA libraries are commonly generated from a single tissue and thus, relative frequency of a particular EST roughly corresponds to its expression level *in vivo*. In some cases, ESTs have been generated from both normal and diseased tissues. These datasets along with the Digital Differential Display tool (now referred to as the Digital Gene Expression Displayer <http://cgap.nci.nih.gov/Tissues/GXS>) available via the Cancer Genome Anatomy Project (CGAP) at NCBI, have recently been used to identify candidate cancer genes whose EST sequences were overrepresented in malignant or normal tissue [34]. Two of the candidates were subsequently validated as true cancer genes [35]. Various online sources in addition to NCBI can be used to query EST collections, such as TIGR (<http://www.tigr.org/index.shtml>) and RIKEN (<http://read.gsc.riken.go.jp/>) [36].

The EST approach has limitations and can be biased depending on depth of sequencing. Recently, more quantitative assessments of whole genome expression have been made in human, mouse and rat tissues. The Genomics Institute of the Novartis Research Foundation (GNF) has generated DNA microarray expression profiles in a diverse set of human and rodent tissues [37]. These data are publicly available through the GNF SymAtlas online database (<http://symatlas.gnf.org/SymAtlas/>). In addition, the Allen Institute for Brain Science has generated the Allen Brain Atlas (<http://www.brain-map.org/welcome.do?sessionId=AE1089DFC3CB220BF3CB842D84129287>) which measured the expression of over 20,000 genes in the mouse brain using *in situ* hybridization [38]. These data provide tremendous insight into the regional expression of a gene in the mouse brain.

2. Differential gene expression in disease

The use of expression profiling has been widely used in animal models to discover disease genes. In many studies microarrays are used to interrogate regions previously found to harbor a gene(s) affecting a complex trait (referred to as quantitative trait loci or QTL). Genes are prioritized based on differential expression dependent on QTL genotype. The resulting hypothesis is that one of the differentially expressed genes controls the phenotypic difference. If the list is short or contains biological relevant candidates then subsequent experiments can be used to determine which gene(s) regulates the disease. One of the first studies to demonstrate the feasibility of this approach identified the *Cd36* fatty acid translocase as the gene responsible for several metabolic defects, including insulin resistance, in the spontaneously hypertensive rat (SHR) [39]. In a cross between two rat strains (SHR and Wistar Kyoto) a QTL was identified on chromosome 4. The metabolic disturbances observed in the SHR strain were corrected in a chromosome 4 congenic strain

(congenics contain a chromosomal segment from one strain introgressed onto the genetic background of a second strain). The analysis of adipose tissue gene expression between the congenic and control using two-color spotted cDNA arrays revealed a 90% reduction in the levels of *Cd36* mRNA in congenic rats. To prove the reduction in *Cd36* was causative, the authors identified multiple sequence variations in SHR *Cd36* and demonstrated that transgenic mice overexpressing *Cd36* had reduced triglycerides.

To identify genes contributing to asthma, Affymetrix GeneChip arrays were used to detect differentially expressed genes in the lungs of A/J (highly susceptible to allergen induced airway hyperresponsiveness (AHR)) and C3H/HeJ (highly resistant) mouse strains and a limited number of A/J X C3H/HeJ F1 and F1 X A/J backcross mice [40]. Of 21 differentially expressed genes, the complement factor 5 (C5) gene was located near the *Abhr2* (allergen-induced bronchial hyperresponsiveness 2) QTL and its expression was negatively correlated with AHR. It was previously known that A/J mice have a 2-bp deletion in exon 5 which eliminates C5 mRNA and protein, while C3H possess normal levels and activity of C5 [41]. The combination of microarray gene expression data in addition to functional studies strongly suggested a role for C5 in allergic asthma. In a subsequent study polymorphisms in the human C5 gene were associated with bronchial asthma in a Japanese population [42].

Osteoporosis is one of the most common diseases associated with aging and is under strong genetic control. A QTL controlling bone mineral density (BMD), a major predictor of osteoporotic fracture risk in humans, was identified on mouse chromosome 11 between the DBA/2J and C57BL/6J strains [43]. The locus was captured in a congenic strain and DNA microarray analysis between in kidney tissue identified a 20 fold reduction in the expression of the 12/15-lipoxygenase gene (*Alox15*) in C57BL/6J mice. *Alox15* knockout mice and mice treated with a pharmacological inhibitor of 12/15 lipoxygenase had higher BMD, validating the role of *Alox15* gene expression in the acquisition of bone mass.

3. Functional annotation of gene expression patterns

In many cases the underlying biological theme of a specific group of genes altered by disease is not immediately clear and requires functional annotation. The basis for nearly all supervised annotation is the Gene Ontology (GO). The GO is a controlled vocabulary designed to annotate the biological process, molecular function and cellular component of all eukaryotic genes and gene products [44]. A number of annotation tools have been developed which use GO annotation for biological interpretation of an otherwise anonymous gene set (<http://www.geneontology.org/GO.tools.shtml>). Of particular use is the Database for Annotation, Visualization and Integrated Discovery (DAVID) suite of annotation and visualization tools [45]. DAVID allows one to identify biological themes which are enriched in a particular gene list, visualize genes in well known biological pathways such as KEGG and BioCarta, and cluster redundant annotation terms among a group of genes. The Expression Analysis Systematic Explorer (EASE) software, developed by the DAVID bioinformatics group, also has the capacity to identify the biological theme of a gene list, and can be downloaded as a stand alone program [46].

In many group comparisons only a small number of genes with statistically significant changes in gene expression are identified. This is often due to low statistical power, which in most experiments is a function of small sample sizes and a large number of statistical tests. An alternative biological explanation is that some disease is caused by or elicits subtle coordinate changes in the expression of gene pathways. Small changes in pathway expression can be expected to have biologically significant effects on metabolite flux, the induction of transcriptional cascades and, ultimately, disease. Recently, an analytical tool termed Gene Set Enrichment Analysis (GSEA) was developed to increase the statistical

power of microarray experiments by identifying known biological pathways enriched for differentially co-regulated genes [47,48]. GSEA takes an input set of genes, such as all genes expressed in a tissue, and ranks them based on a standard metric of differential expression between two groups. Next, a running cumulative enrichment score (ES) is calculated for each biological pathway or functionally related gene set. An example would be all genes known to be involved in atherosclerosis or inflammation. If a pathway is enriched for genes either positively or negatively correlated with disease status, a high mean ES (MES) will be assigned to that pathway. The statistical significance of the MES is assessed using permutations of the disease status label. In the seminal GSEA study, transcriptome profiles were generated from muscle biopsies collected from normal glucose tolerant, impaired glucose tolerant and type 2 diabetic patients [47]. Using traditional statistical techniques no significant changes in gene expression were observed in any of the possible pairwise group comparisons. However, using GSEA a set of genes involved in oxidative phosphorylation possessed the highest MES. Interestingly, 89% of all genes in this pathway displayed a modest 20% reduction in expression in diabetic versus normal patients. GSEA has also been used in a mouse intercross to analyze liver gene expression profiles [49]. In this study, GSEA was integrated with genetics to identify metabolic pathways and regulatory loci controlling obesity. GSEA is a powerful tool to detect subtle changes in the expression of a pathway, which would not be identified using the standard differential expression paradigm. However, it should be noted that this analysis relies on predefined biological pathways and will miss important changes in unannotated genes and in novel pathways.

4. Identification of disease biomarkers

The discovery of disease biomarkers and prediction of disease subtypes are promising applications for expression profiling. Both are critical to the early detection and proper treatment of many diseases and recently a number of studies have demonstrated the feasibility of microarrays for both applications. In addition, biomarkers can be used to group patients in clinical trials based on observed or predicted drug responses. This may improve the clinical success of drugs with limited efficacy in the population as a whole, but which are highly efficacious for a subset of the population.

Examples of using DNA microarrays in this context have been numerous. Highlights include the work by Seo and colleagues [50] who recently identified a set of signature genes whose expression in human aorta was predictive of atherosclerosis burden. Using the expression of this gene set the authors were successful in classifying new aortic sections as diseased or normal over 93% of the time. Other success stories include a recent series of studies identifying distinct breast cancer subtypes using expression profiles from cancerous and normal breast samples [51-53].

IV. INTEGRATION OF GENETICS AND GENOMICS

Genome-wide transcript levels can be considered as intermediate phenotypes or “endophenotypes” for a disease. A powerful way to integrate genetics and genomics is to define the genetic control of transcript levels and at the same time, the genetics of disease phenotypes. In such analyses, transcript levels can be treated as other quantitative traits and the loci controlling them can be mapped using classical linkage and association approaches. As summarized in Figure 3, such combined genetic and genomic data can then be used to identify positional candidate genes; to identify known pathways involved in the disease; to model casual interactions involved in the disease; and to model gene networks and relate those to the disease. As yet, most studies have been performed using animal models [49,54-57], where the analyses are greatly simplified by the ability to control the environment, design crosses, perform invasive procedures, and sample tissues. Although

likely an order of magnitude more difficult, the same approaches appear feasible in human populations.

1. Mapping gene expression quantitative trait loci (eQTL)

Genomic regions harboring variation affecting a quantitative trait are referred to as quantitative trait loci (QTL) [58]. QTL identification has been used extensively in humans and model organisms to identify regions containing key disease regulators. A QTL can be composed of a single gene or as recent data indicate a cluster of genes whose cumulative effects are represented as one locus. Statistical strategies for identifying QTL can be quite mathematically rigorous and many different types of analyses have been developed. However, correlating genotype with phenotype is the basis of all approaches. QTL mapping is crucial to any study integrating genetics and gene expression and Figure 4 illustrates a simple example for a gene expression trait. Although beyond the scope of this chapter a more detailed description of statistical methodologies for QTL mapping are the focus of prior chapters and can be found in recent reviews [58-61].

The first genetical genomics experiment using global gene expression profiles was published in yeast [62]. In this work, the authors described two general classes of QTL controlling gene expression, *cis* and *trans*-acting. Both terms refer to the location of the genetic variation giving rise to the expression QTL (eQTL). For example, if the expression of gene X is regulated by a *cis*-eQTL, the variation is located in or near the gene X locus. If the expression of gene X is controlled by a *trans*-eQTL the variation is located outside the gene X locus. *Cis*-eQTL can be due to variation such as polymorphism in a promoter or other regulatory region, whereas a *trans*-eQTL could be a polymorphism which alters the activity of a transcription factor which in turn modulates expression. Recently, Rockman and Kruglyak [63] have proposed a more general lexicon to describe *cis* and *trans* affects. They propose using the terms local and distant linkage instead of *cis* and *trans*, respectively. Using this terminology, local linkage can be due to *cis*-acting variants affecting allele specific transcription rates in addition to the effects of neighboring genes, autoregulation and feedback loops. The concept of local and distant linkage is illustrated in Figure 5.

One of the first applications of this approach was the investigation of the genetic architecture of gene expression. It allowed questions to be asked such as how many QTL regulate the expression of a given gene and what fraction of variance in expression is explained by genetics? Although definitive answers are still elusive, clear trends have emerged, including the realization that expression phenotypes are relatively complex despite a direct relationship between DNA and the mRNAs it encodes. In general many more distant linkages are observed relative to local linkages and in some cases expression phenotypes are controlled by many eQTL. Additionally, evidence for epistasis regulating a significant fraction of gene expression traits has been reported in yeast [64,65]. Despite this surprising complexity the average eQTL in humans and mice explain approximately 25% of the variation in expression which is significantly larger than the average clinical trait QTL [57,66].

2. Prioritizing candidate genes

A list of all genes with local eQTL is valuable in prioritizing candidate genes at a locus harboring clinical trait QTL and this information can be combined with genetic fine mapping of the region. This approach has recently led to the positional cloning of the ATP-binding cassette, sub-family C (CFTR/MRP), member 6 (*Abcc6*) gene responsible for *Dyscalc1*, a major determinant of dystrophic cardiac calcification in mice [67]. The *Dyscalc1* QTL was first narrowed to an 840-kb region. The *Abcc6* gene, located in this region, was found to have a very strong local eQTL controlling its expression. The authors

proved *Abcc6* was responsible for *Dyscalc1* using a transgenic model which recapitulated the resistance phenotype. Therefore, genes with local eQTL coincident with clinical trait QTL are excellent positional candidates and these data can be useful as a screening tool especially when combined with additional genetic data. As yet, the list of known human eQTL is very small but this is expected to increase greatly with larger population and family studies.

Related to this, gene expression databases may help prioritize genes for diseases that display sexual dimorphism. However, until recently it was unclear the extent of sexual differences in global gene expression. In a study by Wang and colleagues [68], significant sex X QTL interactions were demonstrated for thousands of mouse liver eQTL. More importantly, obesity also differed between the sexes and many transcripts were identified that correlated with fat mass in a sex dependent manner. A second study further demonstrated the importance of sex by showing that the expression of thousands of genes in multiple tissues in the mouse were sexually dimorphic [69]. Moreover, numerous tissue-specific chromosomal hotspots were identified for eQTL controlling the expression of sexually dimorphic genes. Together these studies indicate a strong role for gender in the control of male and female transcriptomes and the importance of sex dependent expression in the context of disease.

Combining genetics and genomics also allows the prioritization of candidate pathways. The GSEA approach described above is an example of this. Moreover, known causal genes can be linked to known pathways by testing for significant correlations between the two. The study of dystrophic cardiac calcification discussed above is a good example. The function of *Abcc6* and how it contributed to calcification was entirely unknown, in fact the substrate for this transporter has yet to be identified [67]. To examine which processes might involve *Abcc6*, correlations between *Abcc6* transcript levels and other transcripts in the mouse cross were determined. Interestingly, *Abcc6* transcripts were found to be significantly correlated with a *Wnt* signaling pathway previously proposed to contribute to calcification, suggesting testable hypotheses for the role of *Abcc6* [67].

3. Modeling causal interactions

Orthogonal data sets such as genotypes, gene expression profiles and disease status provide the data necessary to infer causality. Causality can be predicted for any gene expression – clinical trait pair by evaluating the relative likelihood of a casual, reactive and independent model. In a causal model a genetic variant (assayed in the population using a tightly linked genetic marker) elicits a change in gene expression that pleiotropically affects the clinical trait. In a reactive model the genetic variant produces a change in the clinical, that in turn alters gene expression (gene expression is reacting to the perturbed phenotype) and in an independent model the mutation affects both the gene expression and clinical trait independently. Likelihoods for each model can be calculated based on conditional probabilities and used to assess the most probable scenario for a given gene.

Recently, Schadt *et al.* [56] developed and applied causality modeling algorithms to a mouse intercross to predict key drivers of obesity. In that study, genes whose transcript levels correlated with adiposity were identified, and then this set was intersected with the set of genes whose eQTL overlapped with adiposity QTL (cQTL) in the cross. Several genes were predicted as casual and in this and ongoing studies a number have been validated using transgenic mice. Almost all the validated targets were novel obesity genes, illustrating the enormous power of this approach. A simplified example of causality modeling is presented in Figure 6.

4. Gene co-expression networks

Genes do not function in isolation, but instead are members of gene groups or biological pathways which work in concert to perform particular functions. This coordinated action is due in part to transcriptional regulation. Consider the peroxisome proliferator-activated receptor (PPAR) family of transcription factors. PPARs respond to extracellular stimuli (either endogenous or exogenous) by increasing or decreasing the expression of hundreds of genes belonging to a highly diverse set of biological pathways. This concordant transcriptional regulation allows a cell to quickly respond to changing conditions. Thus, genes whose expression is concordantly regulated over a set of differing conditions are likely to be functionally related.

Recently, much focus has been placed on developing biological networks using datasets such as gene expression, protein-protein interactions and literature citations. A network is defined by a collection of nodes and edges, and in the case of gene co-expression networks the nodes are genes and the edges represent a measure of expression similarity. In an unweighted co-expression network a connection (edge) exists between two genes (nodes) only if their expression is correlated above a certain threshold. In a weighted network all nodes are connected but the edges differ based on the strength of the relationship. Much of the theory behind the generation of biological networks comes from the work of Barabasi and colleagues who discovered that most networks exhibit a scale-free topology. Scale-free networks consist of a small number of highly connected nodes with many edges and a large number of nodes with few edges [70].

In the context of gene expression, the purpose of network analysis is the identification of “modules”, or groups of genes which share a highly similar pattern of expression. Network modules are created by grouping co-regulated genes together based on a measure of similarity. An integral component of network construction is calculation of gene connectivity. In weighted gene co-expression networks the connectivity of a gene is the sum of its connection strengths with all other genes, and connection strengths are typically measured using the absolute value of the correlation coefficient between two genes [71]. If a gene is highly connected its expression will be correlated with the expression of many other genes. Highly connected genes are referred to as network “hubs”.

Gene co-expression networks have been generated in both human and mice as a tool to identify modules involved in specific cellular processes, to characterize unannotated genes and as a tool to model the relationship between gene expression and disease. This procedure is summarized in Figure 7. Gargalovic *et al.* [72] examined a relatively small number of primary human endothelial cells for responses to oxidized phospholipids, a trait relevant to atherosclerosis. In this study, the clinical status of individuals from which the cells were derived was unknown, but the co-expression modules identified were significantly enriched in known pathways. One module was enriched for genes involved in the unfolded protein response (UPR) and also contained interleukin-8 (IL-8), an inflammatory stimulus important in atherosclerosis. Importantly, it was shown the UPR pathway contributed to the transcriptional regulation of IL-8. In the mouse, Ghazalpour *et al.* [55] developed a weighted gene co-expression network using liver expression profiles from F2 mice. Several modules were identified, one of which contained genes highly correlated with body weight. The authors demonstrated that a model accounting for genetic information on the location of key drivers of module gene expression and network properties of module genes (namely, connectivity) was an excellent predictor of the relationship between module gene expression and adiposity.

5. Genetical genomics in human studies

Several general surveys of the genetics of gene expression in humans have now appeared [66,73,74]. The genetical genomic studies reported in humans thus far are in their infancy and essentially represent surveys with no attempts to connect gene expression to disease. These studies have also been relatively underpowered and so a small number of clear expression QTL have been identified. Also, most of the reported studies have utilized tissue culture cells, primarily Epstein Barr virus transformed lymphoblastoid cells, that may have significant alterations in genomic content as compared to the individuals from which they were derived. Clearly, however, the results indicate that it is possible to map loci contributing to transcript levels in humans using both linkage analysis and association. There is every reason to believe that, with larger sample numbers, databases of hundreds or thousands of genes commonly varying in transcript levels can be constructed. These will then serve to identify variations that will help prioritize the identification of genes underlying common disease. Moreover, it should be possible to correlate gene expression traits with clinical traits, as has been done in animal models, to identify potential causal genes and to begin to construct networks relevant to disease.

VI. CONCLUSIONS

We have discussed a number of ways in which DNA microarray expression profiling can be used to investigate the genetic basis of disease. Our ability to predict and treat disease will only increase as novel approaches for using DNA microarrays are developed and technologies for quantifying different biological levels mature. Until recently, attempts to identify genes and pathways involved in common diseases were rarely successful. A few successful examples were primarily restricted to candidate genes that were previously identified by biochemical studies, such as apolipoprotein E and Alzheimer disease. However, with the development of relatively inexpensive high throughput genotyping methods, including genome-wide association, and the assembly of large family-based or population-based study samples, the number of genes identified for common disease is ever increasing. The primary challenge, then, will be not to identify the underlying genes, but rather to understand pathways perturbed by genetic variation, the interactions between genes and between genes and environment and the most suitable targets for therapeutic intervention. Global analysis of transcript levels offers an important bridge between genetic variation at the level of DNA and phenotypic variation.

References

1. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LS, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW. Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science*. 2005; 310(5746):317–20. [PubMed: 16224024]
2. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibe B, Bouix J, Caiment F, Elsen JM, Eychenne F, Larzul C, Laville E, Meish F, Milenkovic D, Tobin J, Charlier C, Georges M. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet*. 2006; 38(7):813–8. [PubMed: 16751773]
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaanty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyraas E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS,

Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapozhnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002; 420(6915):520–62. [PubMed: 12466850]

4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science*. 2001; 291(5507):1304–51. [PubMed: 11181995]
5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczy J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N,

Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaanty KD, Miner TL, Delehaanty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korfi I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrino A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860–921. [PubMed: 11237011]

6. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodward C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyraes E, Searle SM, Cooper GM, Batzoglu S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 2004; 428(6982):493–521. [PubMed: 15057822]

7. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006; 313(5795):1929–35. [PubMed: 17008526]
8. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995; 270(5235):484–7. [PubMed: 7570003]
9. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*. 2000; 18(6): 630–4. [PubMed: 10835600]
10. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*. 1992; 257(5072):967–71. [PubMed: 1354393]
11. Hubank M, Schatz DG. Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Res*. 1994; 22(25):5640–8. [PubMed: 7838717]
12. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet*. 1996; 14(4):457–60. [PubMed: 8944026]
13. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995; 270(5235):467–70. [PubMed: 7569999]
14. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A*. 1977; 74(12):5350–4. [PubMed: 414220]
15. Koczan D, Thiesen HJ. Survey of microarray technologies suitable to elucidate transcriptional networks as exemplified by studying KRAB zinc finger gene families. *Proteomics*. 2006; 6(17): 4704–15. [PubMed: 16933337]
16. Stoughton RB. Applications of DNA microarrays in biology. *Annu Rev Biochem*. 2005; 74:53–82. [PubMed: 15952881]
17. Jaluria P, Konstantopoulos K, Betenbaugh M, Shiloach J. A perspective on microarrays: current applications, pitfalls, and potential uses. *Microb Cell Fact*. 2007; 6:4. [PubMed: 17254338]
18. Ehrenreich A. DNA microarray technology for the microbiologist: an overview. *Appl Microbiol Biotechnol*. 2006; 73(2):255–73. [PubMed: 17043830]
19. Hager J. Making and using spotted DNA microarrays in an academic core laboratory. *Methods Enzymol*. 2006; 410:135–68. [PubMed: 16938550]
20. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: an overview. *Methods Enzymol*. 2006; 410:3–28. [PubMed: 16938544]
21. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science*. 1991; 251(4995):767–73. [PubMed: 1990438]
22. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10):R80. [PubMed: 15461798]
23. Konradi C. Gene expression microarray studies in polygenic psychiatric disorders: applications and data analysis. *Brain Res Brain Res Rev*. 2005; 50(1):142–55. [PubMed: 15964635]
24. Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, Lebruska LL, Laurent M, Shen R, Barker D. Illumina universal bead arrays. *Methods Enzymol*. 2006; 410:57–73. [PubMed: 16938546]
25. Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, Bennett H, Rigault P, Barker D, McDaniel TK, Chee MS. A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res*. 2004; 14(11):2347–56. [PubMed: 15520296]
26. Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A,

- Fan JB, Barnard S, Chee MS. Decoding randomly ordered DNA arrays. *Genome Res.* 2004; 14(5): 870–7. [PubMed: 15078854]
27. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, Fan XH, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, LeClerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, Philips KL, Pine PS, Pusztai L, Qian F, Ren H, Rosen M, Rosenzweig BA, Samaha RR, Schena M, Schroth GP, Shchegrova S, Smith DD, Staedtler F, Su Z, Sun H, Szallasi Z, Tezak Z, Thierry-Mieg D, Thompson KL, Tikhonova I, Turpaz Y, Vallanat B, Van C, Walker SJ, Wang SJ, Wang Y, Wolfinger R, Wong A, Wu J, Xiao C, Xie Q, Xu J, Yang W, Zhang L, Zhong S, Zong Y, Slikker W Jr. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol.* 2006; 24(9):1151–61. [PubMed: 16964229]
 28. Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov.* 2002; 1(12):951–60. [PubMed: 12461517]
 29. Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet.* 2001; 2(6):418–27. [PubMed: 11389458]
 30. Wit, E.; McClure, J. *Statistics for Microarrays: Design, Analysis and Inference.* Wiley; 2004. p. 278
 31. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001; 98(9):5116–21. [PubMed: 11309499]
 32. Jilka RL. Biology of the basic multicellular unit and the pathophysiology of osteoporosis. *Med Pediatr Oncol.* 2003; 41(3):182–5. [PubMed: 12868116]
 33. Seeman E, Delmas PD. Bone quality--the material and structural basis of bone strength and fragility. *N Engl J Med.* 2006; 354(21):2250–61. [PubMed: 16723616]
 34. Scheurle D, DeYoung MP, Binnering DM, Page H, Jahanzeb M, Narayanan R. Cancer gene discovery using digital differential display. *Cancer Res.* 2000; 60(15):4037–43. [PubMed: 10945605]
 35. Narayanan R. Bioinformatics approaches to cancer gene discovery. *Methods Mol Biol.* 2007; 360:13–31. [PubMed: 17172723]
 36. Walker JR, Wiltshire T. Databases of free expression. *Mamm Genome.* 2006; 17(12):1141–6. [PubMed: 17143588]
 37. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A.* 2002; 99(7):4465–70. [PubMed: 11904358]
 38. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, Chen L, Chen L, Chen TM, Chin MC, Chong J, Crook BE, Czaplinska A, Dang CN, Datta S, Dee NR, Desaki AL, Desta T, Diep E, Dolbeare TA, Donelan MJ, Dong HW, Dougherty JG, Duncan BJ, Ebbert AJ, Eichele G, Estin LK, Faber C, Facer BA, Fields R, Fischer SR, Fliss TP, Frensley C, Gates SN, Glattfelder KJ, Halverson KR, Hart MR, Hohmann JG, Howell MP, Jeung DP, Johnson RA, Karr PT, Kawal R, Kidney JM, Knapik RH, Kuan CL, Lake JH, Laramie AR, Larsen KD, Lau C, Lemon TA, Liang AJ, Liu Y, Luong LT, Michaels J, Morgan JJ, Morgan RJ, Mortrud MT, Mosqueda NF, Ng LL, Ng R, Orta GJ, Overly CC, Pak TH, Parry SE, Pathak SD, Pearson OC, Puchalski RB, Riley ZL, Rockett HR, Rowland SA, Royall JJ, Ruiz MJ, Sarno NR, Schaffnit K, Shapovalova NV, Sivasay T, Slaughterbeck CR, Smith SC, Smith KA, Smith BI, Sodt AJ, Stewart NN, Stumpf KR, Sunkin SM, Sutram M, Tam A, Teemer

- CD, Thaller C, Thompson CL, Varnam LR, Visel A, Whitlock RM, Wohnoutka PE, Wolkey CK, Wong VY, Wood M, Yaylaoglu MB, Young RC, Youngstrom BL, Yuan XF, Zhang B, Zwingman TA, Jones AR. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. 2007; 445(7124):168–76. [PubMed: 17151600]
39. Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PJ, Wahid FN, Al-Majali KM, Trembling PM, Mann CJ, Shoulders CC, Graf D, St Lezin E, Kurtz TW, Kren V, Pravenec M, Ibrahimi A, Abumrad NA, Stanton LW, Scott J. Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nat Genet*. 1999; 21(1):76–83. [PubMed: 9916795]
 40. Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, Cuomo PJ, Kohl J, Wahl L, Kuperman D, Germer S, Aud D, Peltz G, Wills-Karp M. Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immunol*. 2000; 1(3):221–6. [PubMed: 10973279]
 41. Wetsel RA, Fleischer DT, Haviland DL. Deficiency of the murine fifth complement component (C5). A 2-base pair gene deletion in a 5'-exon. *J Biol Chem*. 1990; 265(5):2435–40. [PubMed: 2303408]
 42. Hasegawa K, Tamari M, Shao C, Shimizu M, Takahashi N, Mao XQ, Yamasaki A, Kamada F, Doi S, Fujiwara H, Miyatake A, Fujita K, Tamura G, Matsubara Y, Shirakawa T, Suzuki Y. Variations in the C3, C3a receptor, and C5 genes affect susceptibility to bronchial asthma. *Hum Genet*. 2004; 115(4):295–301. [PubMed: 15278436]
 43. Klein RF, Allard J, Avnur Z, Nikolcheva T, Rotstein D, Carlos AS, Shea M, Waters RV, Belknap JK, Peltz G, Orwoll ES. Regulation of bone mass in mice by the lipoxygenase gene Alox15. *Science*. 2004; 303(5655):229–32. [PubMed: 14716014]
 44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; 25(1):25–9. [PubMed: 10802651]
 45. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 2003; 4(5):P3. [PubMed: 12734009]
 46. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biol*. 2003; 4(10):R70. [PubMed: 14519205]
 47. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003; 34(3):267–73. [PubMed: 12808457]
 48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102(43):15545–50. [PubMed: 16199517]
 49. Ghazalpour A, Doss S, Sheth SS, Ingram-Drake LA, Schadt EE, Lusi AJ, Drake TA. Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol*. 2005; 6(7):R59. [PubMed: 15998448]
 50. Seo D, Wang T, Dressman H, Herderick EE, Iversen ES, Dong C, Vata K, Milano CA, Rigat F, Pittman J, Nevins JR, West M, Goldschmidt-Clermont PJ. Gene expression phenotypes of atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2004; 24(10):1922–7. [PubMed: 15297278]
 51. Kapp AV, Jeffrey SS, Langerod A, Borresen-Dale AL, Han W, Noh DY, Bukholm IR, Nicolau M, Brown PO, Tibshirani R. Discovery and validation of breast cancer subtypes. *BMC Genomics*. 2006; 7:231. [PubMed: 16965636]
 52. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747–52. [PubMed: 10963602]

53. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001; 98(19):10869–74. [PubMed: 11553815]
54. Doss S, Schadt EE, Drake TA, Lusis AJ. Cis-acting expression quantitative trait loci in mice. *Genome Res*. 2005; 15(5):681–91. [PubMed: 15837804]
55. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*. 2006; 2(8):e130. [PubMed: 16934000]
56. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 2005; 37(7):710–7. [PubMed: 15965475]
57. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. Genetics of gene expression surveyed in maize, mouse and man. *Nature*. 2003; 422(6929):297–302. [PubMed: 12646919]
58. Falconer, DS.; Mackay, TFC. *Introduction to Quantitative Genetics*. 4. Prentice Hall; 1996. p. 480
59. Darvasi A. Experimental strategies for the genetic dissection of complex traits in animal models. *Nat Genet*. 1998; 18(1):19–24. [PubMed: 9425894]
60. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet*. 2002; 3(1):43–52. [PubMed: 11823790]
61. Flint J, Valdar W, Shifman S, Mott R. Strategies for mapping and cloning quantitative trait genes in rodents. *Nat Rev Genet*. 2005; 6(4):271–86. [PubMed: 15803197]
62. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science*. 2002; 296(5568):752–5. [PubMed: 11923494]
63. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet*. 2006; 7(11):862–72. [PubMed: 17047685]
64. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*. 2005; 102(5):1572–7. [PubMed: 15659551]
65. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*. 2005; 436(7051):701–3. [PubMed: 16079846]
66. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature*. 2004; 430(7001):743–7. [PubMed: 15269782]
67. Meng H, Vera I, Che N, Wang X, Wang SS, Ingram-Drake L, Schadt EE, Drake TA, Lusis AJ. Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics. *Proc Natl Acad Sci U S A*. 2007; 104(11):4530–5. [PubMed: 17360558]
68. Wang S, Yehya N, Schadt EE, Wang H, Drake TA, Lusis AJ. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet*. 2006; 2(2):e15. [PubMed: 16462940]
69. Yang X, Schadt EE, Wang S, Wang H, Arnold AP, Ingram-Drake L, Drake TA, Lusis AJ. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res*. 2006; 16(8):995–1004. [PubMed: 16825664]
70. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999; 286(5439):509–12. [PubMed: 10521342]
71. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005; 4:Article17. [PubMed: 16646834]
72. Gargalovic PS, Imura M, Zhang B, Gharavi NM, Clark MJ, Pagnon J, Yang WP, He A, Truong A, Patel S, Nelson SF, Horvath S, Berliner JA, Kirchgessner TG, Lusis AJ. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc Natl Acad Sci U S A*. 2006; 103(34):12741–6. [PubMed: 16912112]

73. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*. 2005; 437(7063): 1365–9. [PubMed: 16251966]
74. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*. 2004; 75(6):1094–105. [PubMed: 15514893]

Figure 1. Biological systems can be viewed as being comprised of discrete stages including the genome, transcriptome, proteome, metabolome and phenome.

Figure 2.

Description of a hypothetical one-color microarray analysis between affected and unaffected muscle biopsy samples. In this example global gene profiles are generated from diabetic and normal muscle biopsies. First, mRNA is isolated from both samples and labeled cRNA (in most cases a biotin labeled RNA copy) is hybridized to the array. A fluorescent reporter is used to determine signal intensities at each feature on the array. For simplicity we have focused our attention on nine features, each containing a gene-specific probe sequence. In reality most one-color arrays contain millions of features. The expression of Gene X is determined using the feature outlined in red. A dark spot indicates low or no signal and white represents high expression. Signal intensities are quantitated, processed and normalized, yielding the level of expression of each gene within the arrayed target. The bar chart at the bottom represents the expression of gene X in both samples and indicates that it is down-regulated in diabetic muscle. In a real experiment the average of multiple biological replicates would be used to represent the expression of a gene in the affected and unaffected states.

Figure 3.

Schema for combining genetics and genomics to investigate human disease. The approach begins by collecting clinical, global gene expression and genotype data from family or population based samples. The gene expression data can be used to identify differentially expressed genes, for biomarker discovery and Gene Set Enrichment Analysis by subdividing the population into groups based on disease status or genotype. QTL or association analysis, depending on the population type, can be used to identify correlations between genotype and clinical/gene expression traits. These data can then be used to prioritize genes based on coincidence between gene expression and clinical QTL or associations and causality modeling. Additionally, network data on highly connected genes or genes belonging to a module correlated with a clinical trait can also be used to screen candidates. High priority genes and pathways can be validated in large populations using association analysis and/or in animal models using transgenic mice.

Figure 4.

The genetics of gene expression. The example illustrates the principles of mapping expression QTL. A) Global gene expression profiles and genotypes are collected from a mouse F2 intercross between parental strains A and B. B) QTL analysis is performed by correlating the levels of an individual gene (gene X) with the presence of strain A and B alleles at markers spaced across the genome. In this example the expression of gene X is regulated by a strong local eQTL (see text and Figure 4 for description of local and distant eQTL) on chromosome 4. The position of gene X on chromosome 4 is denoted by the red bar. Other distant linkages such as ones on chromosomes 11 and 18 also influence the expression of gene X. C) A closer look at gene X expression reveals the basis of the local eQTL. Gene X is highly expressed in parental strain A, intermediate in F1 mice and lowly expressed in strain B. F2's, homozygous for strain A alleles at markers on chromosome 4, express gene X at high levels and those inheriting strain B alleles at the same markers express gene X at a lower level, thereby explaining the strong correlation between chromosome 4 markers and the expression of gene X.

Figure 5.

Local versus distant expression QTL. A) Explanation of different types of variation leading to local eQTL. For each example dark green boxes represent the coding region of functional genes whereas the light orange colored boxes represent upstream regulatory sequences. Variation can occur in a neighboring gene whose protein product regulates the expression of a nearby gene. *Cis*-regulatory variation can also affect allele-specific transcription rates giving rise to an eQTL. Autoregulatory feedback and feedback with unknown intermediates can also produce local eQTLs. A real example is presented of a local eQTL regulating the expression of the monoacylglycerol O-acyltransferase 1 (*Mogat1*) gene in a large mouse F2 intercross. The red bar denotes the location of the *Mogat1* gene in the mouse. These data are not sufficient to indicate the exact nature of the local eQTL. B) Explanation of different types of variation leading to distant eQTL. Variation in individual genes can alter the expression of a single unlinked gene, variation in one gene can alter the expression of many genes or the expression of many genes can regulate the transcription of a single gene. A real example is presented for three distant eQTL regulating the expression of the myeloperoxidase (*Mpo*) gene in a large mouse F2 intercross. The red bar denotes the location of the *Mpo* gene in the mouse.

Modeling causal relationships

Figure 6.

Modeling casual relationships between gene expression and clinical traits. Causality between gene expression and clinical traits can be modeled by determining the likelihoods of independent, casual and reactive models. Additionally, information on multiple clinical trait and eQTL can be incorporated to further strengthen casual predictions. See text for details of causality modeling.

Data from a series of arrays is used to quantify gene expression changes due to a specific perturbation

Coexpressed genes share similar profiles of expression across samples

Coexpression relationships are quantified using correlation

	G1	G2	G3	G4	G5	G6	G7	G8
G1	1	0.8	0.9	0.1	0.3	0.2	0.5	0.7
G2	0.8	1	0.2	0.3	0.6	0.9	0.1	0.1
G3	0.9	0.2	1	0.9	0.8	0.2	0.1	0.6
G4	0.1	0.3	0.9	1	0.2	0.3	0.7	0.6
G5	0.3	0.6	0.8	0.2	1	0.3	0.2	0.1
G6	0.2	0.9	0.2	0.3	0.3	1	0.6	0.9
G7	0.5	0.1	0.1	0.7	0.2	0.6	1	0.8
G8	0.7	0.1	0.6	0.6	0.1	0.9	0.8	1

Gene coexpression networks are constructed by clustering correlated genes and can be visualized in a number of ways

2-D heatmap representation

Multidimensional network representation

Figure 7.

Generating gene co-expression networks with global expression profiles. Co-expression networks rely on the collection of global gene expression profiles sampled across a series of perturbations such as differing genotypes. Within the collection of profiles sets of genes will demonstrate similarity in expression patterns due to transcriptional co-regulation. The co-expression relationships between genes can be quantified using correlation coefficients. Sets of correlated genes are then clustered using standard algorithms to identify “modules” of co-expressed genes. Co-expression networks can be visualized in a number of ways such as a 2-D heatmap or in a multidimensional spherical space. In both plots distinct modules are labeled with different colors.

Table 1

Commerically available DNA microarray platforms commonly used for expression profiling.

Platform	Highlights of each platform
Affymetrix GeneChips	Whole genome expression profiling for a large number of species – proven platform with exceptional annotation
Illumina BeadArray	Whole genome expression profiling in humans and rodents – ultra-high feature density decreases cost per sample
Agilent	Whole genome expression profiling for a large number of species
Applied Biosystems	Whole genome expression profiling in humans and rodents – platforms provide comprehensive genome coverage
Eppendorf DualChips	Pathway focused content
CombiMatrix CustomArray	Custom content arrays – Electrode-directed synthesis allows quick developments of custom arrays
NimbleGen	Provides complete array services with standard arrays for a nearly all species with genome sequence – platform is also flexible for designing custom content arrays