Exploring genetic regulatory networks in metazoan development: methods and models*

MARC S. HALFON AND ALAN M. MICHELSON
Division of Genetics, Department of Medicine, Brigham and Women’s Hospital and
Howard Hughes Medical Institute, Boston, Massachusetts 02115

Received 7 June 2002; accepted in final form 16 July 2002

Halfon, Marc S., and Alan M. Michelson. Exploring genetic regulatory networks in metazoan development: methods and models. Physiol Genomics 10: 131–143, 2002; 10.1152/physiolgenomics.00072.2002.—One of the foremost challenges of 21st century biological research will be to decipher the complex genetic regulatory networks responsible for embryonic development. The recent explosion of whole genome sequence data and of genome-wide transcriptional profiling methods, such as microarrays, coupled with the development of sophisticated computational tools for exploiting and analyzing genomic data, provide a significant starting point for regulatory network analysis. In this article we review some of the main methodological issues surrounding genome annotation, transcriptional profiling, and computational prediction of cis-regulatory elements and discuss how the power of model genetic organisms can be used to experimentally verify and extend the results of genomic research.

genomics; cis-regulation; microarrays; computational biology

THE RECENT AVAILABILITY of whole genome model organism sequences, genome-wide transcriptional profiling methods, and computational tools for exploring these rapidly accumulating new data has set the stage for the study of developmental genetic regulatory networks to come into its own as a vital component of 21st century experimental biology. While studies of gene regulatory networks in yeast (reviewed in Ref. 111) illustrate the power of combining genomic techniques with genetic analysis, delineating the networks that mediate metazoan development, with their more complex transcriptional regulation and multiple interacting intercellular signaling pathways, presents additional challenges. In this article, we will review some of the recent methods that are now available for characterizing developmental regulatory networks with a focus on how model organism genetics can provide a framework for applying and empirically validating the data from these approaches.

Metazoan development occurs by the progressive determination of cells from a pluripotent undifferentiated state through successive states of more and more restricted developmental potential, until the full complement of mature terminally differentiated cells have been specified. While various epigenetic and environmental factors can influence this developmental program, in essence its course is driven by a series of genetic regulatory networks set into motion at the beginning of embryogenesis (see Refs. 20, 25). These networks receive input in the form of intercellular signals and output instructions in the form of the regulated expression of specific genes. The particular combination of signals received by a given cell comprises a form of “signaling code,” which is transduced to the nucleus to create a “transcriptional code” composed of both signal-induced and tissue-specific transcription factors (TFs). The linchpins of the regulatory networks are the cis-regulatory elements that directly

*This review was submitted in conjunction with the meeting entitled “Physiological Genomics of Cardiovascular Disease: from Technology to Physiology,” in San Francisco, CA, February 2002, sponsored by the American Physiological Society.

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: A. M. Michelson, Brigham and Women’s Hospital, Thorn 10th floor, 20 Shattuck St., Boston, MA 02115 (E-mail: michelson@rascal.med.harvard.edu).
control gene expression through interpretation of the transcriptional code and thus act as sites of integration for the combinatorial action of multiple signal transduction pathways and tissue-specific selector proteins (Fig. 1).

Although the notion of genetic regulatory networks is not a new one (13), elucidating them has proven a difficult task; surprisingly few have been defined in any detail, with perhaps only the network involved in sea urchin endomesoderm specification described comprehensively (26). The techniques that have worked so well for investigating the functions of individual genes and essentially linear signaling pathways (e.g., forward and reverse genetics, targeted gene knockouts, and gene misexpression studies) turn out to be not nearly as well-suited for exploring complex networks (although these methods continue to play a critical role in testing and validating new hypotheses about network components, as discussed below). A major limitation here is a paucity of knowledge regarding what genes are at play, and what their functions are, in any given process. One clear lesson from the sequencing of the worm, fly, and human genomes is just how many previously unidentified genes await characterization. Genetics has been a powerful tool for identifying individual network members, but extensive pleiotropy and redundancy limits the amount of information that is likely to be gathered by genetic screens alone. Moreover, these methods cannot reveal the detailed molecular information about cis-regulatory elements that is a critical component of comprehending regulatory networks. Fortunately, the explosion of new genomic technologies is making the study of these networks an ever more tractable problem. Fully sequenced genomes, microarray platforms that allow for genome-wide transcriptional profiling, and the development of computational tools for identifying TF binding sites and cis-regulatory modules are beginning to provide a wealth of data that can be combined with more traditional genetic analyses in forming an integrated understanding of embryonic development.

**Genome Annotation**

The rapidly growing availability of whole genome and expressed sequence tag (EST) sequences is an invaluable resource for the study of regulatory networks. The number of eukaryotic organisms undergoing sequencing is approaching 200 (see [http://wit.integratedgenomics.com/GOLD/eukaryagenomes.html](http://wit.integratedgenomics.com/GOLD/eukaryagenomes.html)) with the inclusion of a substantial number of metazoans. Essentially complete are the genomic sequences of model organisms such as *Caenorhabditis elegans* (23) and *Drosophila melanogaster* (1), and a number of vertebrate genomes have been completed or are near to completion, including human (64, 103), mouse ([http://www.ensembl.org/Mus_musculus/](http://www.ensembl.org/Mus_musculus/)), and puffer fish (*Fugu rubripes*; [http://genome.jgi-psf.org/fugu3/fugu3.home.html](http://genome.jgi-psf.org/fugu3/fugu3.home.html)).

A large part of the utility of genomic sequence data lies in the accuracy of the corresponding gene annotation; correct knowledge of the genes encoded by the genome is critical for constructing tools such as microarrays to assess the full complement of genes expressed at any time or place and to identify the components of each functioning network. Genome annotation requires careful curation to integrate data for genes known from traditional molecular cloning studies and EST sequencing with computational predictions based on homology searching and ab initio gene finding programs. When used in combination, these methods function surprisingly well in terms of identifying genes, although they are significantly weaker in their ability to correctly predict intron/exon structure and to detect 5' noncoding exons and transcription start sites (22, 84). Continued refinement of annotation and gene prediction methods is clearly required. For example, a variety of systematic studies conducted since the initial annotation of the *Drosophila* genome (1) have identified a large number of unannotated or incorrectly annotated genes (2, 42, 57, 88). Homology-based searching using sensitive methods...
such as PSI-BLAST and extensive sequencing of tissue-specific EST libraries (which are more likely to contain rare transcripts not widely expressed in other tissues) are two methods that have yielded good results (2, 42), providing a strong argument in favor of continued EST sequencing even for organisms for which genomic sequence is available. The ability to compare sequences of related species should also be of significant help here, as coding regions tend to be highly conserved among sister species (88). To this end, it is gratifying that a number of sister species or close relatives to already-sequenced organisms are currently undergoing or slated to undergo sequencing, including Caenorhabditis briggsae, Drosophila pseudoobscura, rat and mouse, and the mosquitoes Anopheles gambiae and Aedes egypti. Interspecific comparisons will also be of enormous benefit in identifying regulatory regions (see below).

An additional challenge to genome annotation efforts lies in the identification of noncoding RNAs (29). A large variety of noncoding RNAs are known to exist, including the rRNA and tRNA genes, the snRNAs that function in the spliceosome, the Xist and rOX RNAs that function in X chromosome inactivation and dosage compensation in mammals and flies, respectively, the small nucleolar RNAs (snoRNAs), and the “micro” RNAs that appear to play an important role in mRNA regulation (30, 87). Noncoding RNAs tend to score poorly in gene-finding programs due to their lack of an open reading frame, and their diverse nature has made it difficult to develop computational methods for their detection. Homology searches have been among the most effective ways of finding new noncoding RNAs, but a lack of strong sequence conservation means that closely related genomes are required for comparison. Moreover, it is difficult to distinguish noncoding RNAs from conserved regulatory regions. Extensive Northern analysis or hybridization to microarrays containing tiled whole genome sequence (56, 79, 92) may be required to distinguish between these possibilities, but these are labor- and resource-intensive methods that will not quickly be applied to large, whole genomes. A full and accurate accounting of all the varied noncoding RNAs, let alone their functions, in eukaryotic genomes may thus be a long time in coming.

Unraveling the Transcriptome

With sufficient genome data in hand, it becomes possible to move on to full-scale analysis of the transcriptome, the population of genes expressed at a particular time or in a specific tissue, or in response to a defined set of experimental conditions. Given the well-established ubiquitous role of transcriptional regulation in biological systems (25), virtually any developmental process is amenable to investigation by gene expression profiling. The major limitation of this approach in any given context is the accessibility of suitable material for analysis and the ability to perturb the system by genetic or other means. The most thorough way of transcriptional profiling may be serial analysis of gene expression (SAGE; 102), wherein the sensitivity of the assay is limited mainly by the extent of DNA sequencing the investigator is able to undertake. In practice, however, this can become an expensive and labor-intensive endeavor, and microarray technology is emerging as the method of choice for transcriptional analysis.

Although microarrays, both cDNA or long oligonucleotide arrays and high-density Affymetrix GeneChip arrays, continue to gain in popularity and are rapidly becoming a mainstream experimental technique, consensus is still lacking as to the best methods to use in evaluating the data obtained from microarray experiments. This is a nontrivial issue, as experiments in which 15,000 or more genes are interrogated simultaneously build up data with great rapidity and demand sophisticated computational and statistical analytical techniques unfamiliar to many traditionally trained biologists. A common metric used for analysis of microarray data has been a simple “fold-change” cutoff, in which an arbitrary ratio (frequently 2:1) of gene expression values is used to define a change in expression level. However, such a cutoff carries little statistical authority; genes with high variability in expression in replicate experiments can exhibit a large apparent fold change but little statistical significance, and conversely, genes showing a small fold change can never-

![Fig. 2. The importance of statistical evaluation of microarray data. Illustrated are expression data showing the average relative activation of two genes ("A" and "B") compared with a control condition or reference sample from 3 replications of a microarray experiment (Choe S and Halfon MS, unpublished data). Using a simple fold change cutoff of 2-fold, gene B would be called upregulated due to its apparent 2.4-fold activation, whereas gene A would be considered unchanged due to its 1.7-fold increase. However, gene B displays substantial variation and might not satisfy statistically valid criteria for being considered activated. Gene A, on the other hand, is consistently upregulated by over 1.5-fold and may thus represent a true activation despite the smaller magnitude of its change.](http://physiolgenomics.physiology.org/看他:10.220.33.3)
theless be shown to change with high significance due to their low variance among replicates (Fig. 2).

Fortunately, the importance of replication and proper statistical analysis is becoming widely recognized as a critical component of microarray analysis. A compelling case has been made by a number of laboratories for the use of well-proven analysis of variance (ANOVA) models for use in both normalization of arrays between experiments and evaluation of the significance of microarray results (28, 58–60, 109, 112). Used in conjunction with thoughtful attention to experimental design, these methods can reduce the number of microarrays required for an experiment while simultaneously improving the statistical power of the analysis (21, 59). For example, the popular reference sample design for cDNA arrays in which each array is simultaneously hybridized with an experimental RNA and a common “reference” RNA is not the most efficient method, as one-half of the data is devoted to the reference, rather than to meaningful experimental samples. Kerr and Churchill (59) propose a “loop” design in which each sample is compared directly with every other, i.e., A→B, B→C, C→A, allowing more data to be gathered about each experimental condition. In this example, for instance, one obtains twice as much data as the reference sample method would allow, without requiring more arrays to be used. However, loop designs can become unwieldy when large numbers of experimental conditions are to be compared, and loop designs also suffer from being “closed”; that is, with such a design, it is difficult to come back later to add one new sample for comparison. Devising other experimental designs that overcome some of the limitations of both the reference-sample and loop designs should thus remain an area of priority in the microarray field.

Another important contribution has come from Tusher et al. (100), whose SAM (“statistical analysis of microarrays”) program introduces a practical solution to the multiple hypothesis testing problem (i.e., the special statistical considerations that come into play when testing numerous hypotheses simultaneously, as when looking at expression changes in thousands of genes) with their implementation of a false discovery rate (FDR) controlling procedure. Traditional approaches to multiple hypothesis testing, such as the Bonferroni correction (94), address the probability of having any false-positive results (type I errors). However, these corrections are often so stringent as to allow few if any genes to pass a significance score cutoff (e.g., for a P < 0.05 level on a 15,000 gene array, a classic Bonferroni correction would require a P value of 0.05/15,000 = 3.3 × 10⁻⁶). On the other hand, the FDR procedure sets an upper bound on the number of potential false-positive results, rather than ruling them out completely. This is a fair tradeoff in many microarray experiments, where a small number of false-positive results might be tolerable among the entire data set. For Affymetrix GeneChip arrays, the Affymetrix Microarray Suite version 5 incorporates measures of statistical significance, a substantial improvement over earlier, empirically based versions of the software.

An attractive alternative to the Affymetrix software is the independently developed dChip software of Li and Wong (67), which uses statistical models to perform normalization and a variety of downstream analyses of GeneChip data. Numerous other approaches have been suggested for analysis of both cDNA and Affymetrix arrays based on t-statistics, nonparametric (e.g., Wilcoxon) tests, regression analysis, and Bayesian methods (inter alia, 31, 55, 77, 99).

This growing multitude of proposed analysis methods leaves the investigator with the dilemma of choosing from among many possible approaches. Unfortunately, little critical assessment of the various methods has yet been performed. Although comparisons have been made among the different strategies (77), these have typically been done using existing data sets in which the “correct” patterns of expression are not known. Thus differences in how the methods perform can be seen, but it is not known which results most accurately reflect true changes in gene expression. Analysis of a microarray experiment in which a large number of RNAs of known relative concentration are used could provide invaluable insight into microarray sensitivity and the relative effectiveness of the different existing methods for statistical analysis. A step in this direction has been taken by Irizarry et al. (51), who have compared a number of methods for analysis of Affymetrix data based on a set of spiked-in control RNAs. More extensive testing along these lines using a larger set of controlled RNAs will be a welcome contribution to the microarray field, both for evaluating existing algorithms and for testing new ones, as will continued community-wide assessment efforts such as those sponsored by the CAMDA competition (Critical Assessment Of Microarray Data Analysis; http://www.camda.duke.edu/).

Although statistical methods are necessary for reliably assessing what genes are differentially regulated in microarray experiments, useful overviews of the data can be obtained from clustering algorithms (24, 83, 91). Clustering is an important tool for developing hypotheses based on perceived patterns of gene expression and is particularly well-suited for issues of class discrimination, such as differentiating tumor types from transcriptional profiling of biopsy samples (81). Clustering can also provide hints as to what genes may be acting in common pathways, based on the assumption that such genes will have well-correlated expression. Commonly used clustering algorithms include both hierarchical methods, in which clusters are linked together in successively larger clusters, and nonhierarchical methods, such as k-means clustering and self-organizing maps, which partition the data into a discrete number of groups. Vector algebra approaches, such as principal component analysis, also provide a flexible and useful way of visualizing and sorting out microarray data and are particularly good at reducing the complexity of the data (63, 83). The latter is a major goal of all clustering methods, as microarray experiments contain a great deal of noise in the form of “background” gene expression unrelated to the partic-
ular condition being assessed, which can often obscure subtle but important actual changes in expression. Filtering genes prior to clustering using statistical criteria, as discussed above, also helps to ensure that only the most relevant genes are used in the analysis.

It is important to note that there is no “right” clustering approach to use for microarray analysis. Clustering outcomes will vary considerably depending on the algorithm used and the manner in which the data are processed prior to clustering, and it is often worthwhile to cluster the data in a number of different ways to see which method provides the most useful view for the questions being addressed. For instance, one clustering algorithm might provide a good grouping of genes based on high vs. low levels of expression, whereas a different clustering of the same data might identify periodic fluctuations in expression such as those that are associated with genes subject to cell cycle or circadian regulation. Genes involved in regulating numerous cellular processes provide a particular challenge to cluster analysis as they could properly fit into many biologically relevant clusters. Ultimately, it lies with the investigator to decide what views of the data provide the most useful biological picture and generate the most reasonable hypotheses for further testing.

Transcriptional profiling experiments in metazoans bring with them additional challenges not encountered with single-celled organisms such as yeast and bacteria. Large among these is the heterogeneity of cell and tissue types within the animal, which makes it difficult to isolate pure cell populations for profiling. This is particularly disadvantageous when the tissue of interest might constitute only a small fraction of the experimental sample; in this case, a gene that changes dramatically in that tissue but is expressed at a high but constant level in the contaminating cells will be difficult to detect. One solution is to mark individual cell populations using a fluorescent protein reporter construct and to isolate the marked cells by flow cytometry (14); another is to isolate cells by laser-capture microdissection (32). However, the difficulty of obtaining large amounts of tissue by these methods carries its own complications, as an RNA amplification method must then be used to generate enough RNA for microarray experiments. Both PCR-based and linear amplification methods can be used for this purpose (7, 69, 101), but it is still unclear to what extent such manipulations might skew the RNA population and affect the accuracy of the transcriptional profiling results.

Motifs and Modules

Transcriptional profiling can reveal the genes that are expressed in a tissue, but for an understanding of the network architecture underlying this expression we must also turn to a characterization of the cis-regulatory mechanisms that modulate it. Traditionally, analysis of cis-regulatory elements (i.e., enhancers and promoters) has been a laborious, gene-by-gene, empirically based approach that entailed the isolation and testing of the 5′ and 3′ flanking DNA and intronic sequences of the gene of interest in reporter gene assays. The growth of genomic sequence data, however, promises to greatly accelerate the task of regulatory module identification and characterization.

A key step in defining cis-regulatory elements is the identification of individual TF binding sites (motifs). A substantial number of binding site motifs are known from years of study of individual TFs and enhancers, and many of these have been collected into searchable databases such as TRANSFAC (108). When TFs are known, in vitro binding site specificities can be found by methods such as SELEX (12); the development of rapid, high-throughput methods such as protein binding to DNA microarrays (15, 16) promises to greatly accelerate the pace of such determination. In vitro binding affinities do not always accurately reflect where TFs are bound in vivo, however, as clearly illustrated by promising new methods in which chromatin immunoprecipitation (ChiP) is combined with DNA microarray analysis to generate a genomic profile of TF binding (11, 52, 68, 85). Taking full advantage of ChiP methods will require the generation of microarrays with whole genome noncoding sequence. This is no small task with respect to the higher eukaryotes, although smaller-scale efforts based on sequencing of immunoprecipitated DNAs or using microarrays targeted at specific sequences of particular interest are immediately feasible. As even in vivo binding does not necessarily always correlate with regulatory function (11, 52, 68, 85), a combination of ChiP and transcriptional profiling should prove to be an important and powerful combination of analyses (9).

Computational approaches, although unable to assign specific TFs to their cognate binding sites, can be of tremendous aid in identifying relevant binding site motifs (76, 80). The most successful of these methods to date appear to be those that rely on gene expression profiles. Genes are clustered based on common expression (see discussion above), and their flanking genomic regions are aligned and analyzed for the presence of common sequence motifs using any of a number of algorithms developed for this purpose. Among the more commonly used of these programs are CONSENSUS (47), MEME (4), AlignACE (48, 86), and the Gibbs sampler (65) (for a review of the basic theory behind these approaches, see Ref. 96). In yeast, where regulatory elements tend to be promoter-proximal, these methods have worked well (86, 98). Extension to the higher eukaryotes, where regulatory regions are more extensive and can lie many kilobases in either direction from the promoter or within introns (3), introduces significant complexity into such approaches, and false-positive rates can be high (see below). Identifying motifs that occur in pairs should help in terms of both specificity and sensitivity (43). However, not all motifs are consistently paired with others, and developing appropriate measures to evaluate the significance of motifs discovered through multiple alignment remains an important priority.
An attractive variation on the multiple alignment theme inverts the process by beginning with identified motifs and then determining whether they are associated with coexpressed genes. The motifs can be drawn from known TF binding sites or from computational selection based on statistically overrepresented “words” in genomic sequence (17, 54, 82). Motifs that are associated with the same expression profiles are good candidates for motifs that act together to jointly regulate gene expression. Pipel et al. (82) introduce a series of tools for constructing “combinograms,” which aid in the visualization of relationships between motifs and expression profiles. Again, however, challenges lie in the fact that regulatory elements can be distant and to either side of a gene, making it difficult to be certain which motifs match individual expression profiles.

Although identifying TF binding sites is a necessary step, the true key to cis-regulation lies in the modular nature of cis-regulatory elements, each containing a number of binding sites for an assortment of TFs and directing gene expression in a particular spatiotemporal pattern. Identifying not just individual binding sites but rather entire regulatory modules is thus a critical component of defining regulatory networks. Again, a variety of computational techniques have been developed to aid in this task (33, 76, 80).

One way to bypass the difficulties inherent in the fact that metazoan cis-regulatory modules can be far from the genes they regulate is to consider only promoter sequences. Since these sequences lie close to the transcription start site, the sequence space that must be analyzed to find regulatory modules and their associated binding site motifs is vastly reduced, provided, of course, that the transcriptional start can be accurately located. The presence in many promoters of the TATA box provides some help in this regard, although the low complexity of this site means that false-positive rates will be high when it is used as the main criterion. In mammals, the presence of CpG islands near many promoter sequences allows for better discrimination. A recent algorithm by Davuluri et al. (27), which incorporates information on CpG islands, RNA splice-donor sites, and models built from known promoter and first exon sequences, displays a strong ability to recognize transcription start sites associated with CpG islands, but a much lower ability to discern those that are not CpG related. A database of transcriptional start sites for human genes has been created (97) and can thus provide a rich source of raw data for undertaking searches for promoter-proximal regulatory modules.

What of more distant regulatory elements? One powerful method for locating these lies in comparative genomics, wherein homologous regions from related organisms are aligned and then searched for conserved blocks of sequence (also referred to as “phylogenetic footprinting”). This approach has proven valuable not only on a gene-by-gene scale, but also on a genomic basis (46), and several useful tools have been developed to produce graphical views of aligned genomic sequences complete with gene and, in some cases, TF binding site, annotation (53, 70, 72, 89, 113). Nevertheless, phylogenetic footprinting alone is usually insufficient for defining the regulatory elements involved in a particular developmental network as it is not possible to tell, merely from alignment data, the functional role of an identified conserved sequence block. Again, as conserved elements may be found in intergenic regions far from any coding sequence, it is not always clear what gene is subject to regulation by the element. Moreover, the amount of divergence between the two sequences being compared has a significant impact on how useful the comparison will be (46); in some cases when comparing mouse and human sequences, for example, the conservation is so extensive that individual regulatory elements cannot be distinguished. In these instances, comparison with a more diverged vertebrate sequence will be of higher value. Thus, despite the potential power of comparative sequence-based approaches, they will have limited utility for conducting genome-wide searches for regulatory sites until sufficient whole genome sequencing of related organisms has been conducted.

A number of model-based approaches to cis-regulatory element prediction have been developed that rely on the prior characterization of several regulatory modules of a similar functional class; models generated from known vertebrate muscle gene enhancers and retroviral long terminal repeats (LTRs) have been effectively applied (36, 107). These methods use a number of similar well-characterized enhancers to formulate a model that can serve as the basis for a genomic search for other related elements. The models can be highly specific, requiring a number of motifs in a particular order with specific spacing (36, 40, 61), or more general, seeking only a common occurrence of a number of specified motifs (62, 107). One of the most promising methods described so far is the logistic regression analysis used by Wasserman and colleagues (62, 107) to search for elements regulating muscle and liver gene expression, which has been combined with a phylogenetic footprinting approach to better distinguish true regulatory modules from false-positive results. A drawback to most of the model-based approaches is that they require knowledge of a number of regulatory elements of a similar type, which is available for only a limited number of gene expression patterns. In addition, the false-positive rate for all of these methods appears to be high, exceeding 50%.

A third type of approach has relied on the identification of dense clusters of known TF binding sites as the basis for computational prediction of regulatory elements (10, 37, 71, 78). Clustering of binding sites is a known characteristic of cis-regulatory modules (25), and this approach has led to the identification of at least two previously unidentified enhancers (10, 71). Here again, the false-positive rates are high. Furthermore, as with the model-based approaches, a substantial number of known elements appear to be required to select an appropriate site-density cutoff (10) or to train a probabilistic model (37). Methods based on a single TF (71) may not be widely applicable, as not all important TFs are represented in large numbers in...
regulatory elements (for example, two enhancers we have characterized in detail contain only a single site for the Wnt/Wg-responsive TF Tcf) (44, 45). Also, these methods are highly sensitive to the length and extent of degeneracy of the motif consensus sequence (i.e., a motif’s information content). Shorter and/or more degenerate sites, for example, the E-box (CANNTG) sequence to which bHLH-family TFs bind (75), will occur by chance more frequently, and therefore cluster more frequently, than sites with high information content, leading to a higher false-positive rate. These considerations must become an important component of any predictive strategy that relies on site number or density. Use of a combinatorial strategy, in which several motifs are chosen to reflect a particular transcriptional code, the selectivity of the approach, especially when the TFs are known to direct expression in a specific temporal-spatial pattern (45). The use of multiple concurrent strategies, such as motif number, interspecific sequence conservation, and information based on existing model enhancers should also be of significant aid in reducing false-positive rates.

The importance of empirical validation of predicted elements cannot be overstated. The most commonly used criterion in assessing the success of regulatory element prediction algorithms has been to determine whether the predicted element maps near to a gene with the expected expression pattern; those that map in proximity to such a gene are considered true-positive results, while those that do not are deemed false positives. However, in vivo testing of predictions that we have made demonstrates that this criterion can be misleading: although predicted elements can lie close to genes with the expected expression pattern, they do not always turn out to be bona fide regulatory elements when tested using reporter gene assays (45). Unfortunately, actual testing of predicted elements has not played a large role in the studies so far reported, and false-positive rates may be considerably higher than generally appreciated.

**The Power of Genetics**

Fundamentally, understanding developmental regulatory systems is a job for hypothesis-driven experimental biology, a fact sometimes overlooked in the trend toward high-throughput genomics and computational approaches. Despite the power of these techniques and their ability to vastly accelerate the pace of discovery, to understand the workings of a regulatory network we must know more than only what genes are expressed in a tissue and the composition of their cis-regulatory elements. We must also know the individual inputs into the system at each level, in terms of signaling pathways and transcriptional effectors, and how perturbations at each level of the system affect the system as a whole. In the system that has been most thoroughly explored to date, sea urchin endomesoderm development, a combination of experimental techniques were used to dissect each component of the system, including gene disruption via antisense RNA injection, use of dominant-negative proteins, overexpression of signaling molecules, site-directed mutagenesis of specific TF binding sites and cDNA microarrays (26). This testing has revealed important and in some cases surprising results, such as the considerable role played by repression of gene expression and the existence of extensive negative feedback loops, for example in the repressive “subnetworks” that both prevent endomesodermal genes from being downregulated in the endomesoderm and from being induced outside of the endomesoderm (26). In turning to genetic model systems for analysis, an expanded range of experimental tools become available including whole genome sequence, whole genome microarrays, and both gain- and loss-of-function genetic analysis. In yeast, a combination of genetic, genomic, and proteomic methods has been used to define the galactose signaling pathways (49). In *Drosophila*, fluorescent embryo sorting has been used to accumulate homozygous mutant embryos for transcriptional profiling using microarrays, and individual cell populations can be isolated as well (14, 38, 39). Recent work using adult *Drosophila* provides an elegant demonstration of how a combination of genetics, cell culture, and double-stranded RNA interference (RNAi) can be used together with whole genome microarray analysis to address specific hypotheses regarding the various signal transduction pathways that comprise the innate immune response (M. Boutros, H. Agaisse, and N. Perrimon, unpublished observations).

**Computer Simulation and Mathematical Modeling**

As our identification of components of genetic regulatory networks grows, keeping track of the many different regulatory interactions within the network and testing network models will become increasingly complicated. As network models are built up through the experimental techniques outlined above, therefore, computer simulation will play a useful and perhaps necessary role in organizing and testing the models and in highlighting missing interactions and erroneously assembled network connections. The simulation data will serve as a guide to the experimental biologist in correcting flawed hypotheses and in suggesting new hypotheses and directions of investigation. Already, modeling and simulation have proven useful in understanding such diverse situations as segmentation in *Drosophila* (105), lateral inhibition during neurogenesis (73), receptor tyrosine kinase signaling (93), and sea urchin endomesoderm development (26). In many instances, these approaches are limited by the difficulties inherent in obtaining the types of accurate quantitative data describing the particular biological system that are necessary for formulating a specific mathematical model. Nevertheless, computer modeling and simulation are clearly areas ripe for further development, and as computer scientists and mathematicians begin to take an increasing interest in analyzing biological systems, these techniques are likely to
Deconstructing a Network: an Example From Drosophila

As an example of how the many lines of investigation discussed in this review can be used to gradually build up a comprehensive picture of a regulatory network, we focus in this section on the specification of a set of cardiac and somatic muscle progenitor cells in the Drosophila dorsal mesoderm. These cells, which can be distinguished by their expression of the TF even skipped (eve), provide an excellent entry point into a large-scale developmental regulatory network analysis (8, 34, 41). The long history of developmental genetics in Drosophila provides a context for this analysis: we can approach it already possessing a detailed understanding of the initial events leading to specification of the dorsal mesoderm, viz., establishment of dorsal fates stemming from the maternal gradient of the morphogen dorsal, and segmental patterning of the mesoderm deriving from the early activities of the gap, pair-rule, and segment polarity genes (Fig. 3A) (66).

Directly resulting from these early patterning events are orthogonally intersecting domains, in each embryonic hemisegment, of the activities of the Wingless (Wg) and Decapentaplegic (Dpp) signaling pathways (Fig. 3B). These signals provide competence for the specification of the Eve-expressing progenitors by inducing the expression of the mesodermal selector proteins Twist (Twi) and Tinman (Tin). They also induce localized activity of the Ras/MPK pathway, which acts through the TF PointedP2 (Pnt), by virtue of their regulation of receptor tyrosine kinase signaling pathways (6, 19, 35, 44, 95, 110). Furthermore, they act directly to induce eve expression via their downstream TFs, “mothers against Dpp” (Mad) and dTcf (Pangolin) (44). The key to this network is the integration of these events at the level of cis-regulation of the eve gene itself, through a 300-bp regulatory module, the “muscle and heart enhancer” (MHE) (Fig. 3, C and D) (44).

What of additional members of this network? A variety of genomic approaches are now helping to address this question. Microarray analysis is beginning to define additional genes that respond to mesodermal activation of the Wg, Dpp, and Ras/MPK signaling pathways (Fig. 3F; Halfon MS, unpublished results). A computational search for regulatory modules sharing characteristics of the MHE has identified the mesodermal regulatory sequences for the heartbroken gene (50, 74, 104), a member of the FGF-receptor signaling cascade known from genetic studies to be involved in specification of the Eve mesodermal progenitors (Fig. 3E) (45). Like the MHE, the hbr regulatory module appears to integrate signaling from the three described pathways through the binding of the TFs dTcf, Mad, and Pnt, as well as from the mesodermal selector proteins Twi and Tin. Furthermore, it drives reporter gene expression in a pattern overlapping that driven by the MHE, and its sequence is conserved in an evolutionarily distant species. How is it that a member of a receptor tyrosine kinase signaling cascade, which acts via the Ras/MPK pathway, has regulatory sequences that show it is regulated by this same pathway? This finding is consistent with genetic (18) and microarray (Halfon MS, unpublished results) data showing positive feedback in the FGF-receptor pathway, providing a clear example of how each of the many threads of the network analysis (genetic, computational, and microarray) can provide cross-validation and reinforcement of the results from each.

Fig. 3. Defining regulatory networks in the Drosophila dorsal mesoderm. A: a long history of Drosophila developmental genetics has described the initial pattern-formation events that lead to specification of the mesoderm (indicated by Twist expression, arrow in left embryo) and segmentation (Even skipped expression, blue stripes in left embryo; Wingless expression, brown stripes in right embryo) and thus provides a context in which to place an analysis of the immediate events surrounding induction of cardiac and somatic muscle progenitors (Eve expression, arrowheads in right embryo). B: detailed genetic analysis coupled with gene expression studies using immunohistochemistry and whole mount in situ hybridization to RNA has revealed the roles of the orthogonally intersecting domains of Wg and Dpp expression and localized activation of the Ras/MPK pathway, as providing a signaling code for specification of the Eve-positive progenitors (19). C and D: additional genetic studies along with reporter gene analysis in transgenic embryos provided a more detailed picture of the actions of this signaling code, including the presence of positive feedback loops, lateral inhibition, and cross talk between the signaling pathways, as well as an understanding of how the signaling code leads to a transcriptional code consisting of the TFs Mad, dTcf, Tin, Twi, and Pnt; these TFs are integrated at the Eve “muscle and heart enhancer” (MHE) cis-regulatory module to activate expression of eve in the progenitors (18, 44). Direct regulation is indicated with solid black arrows; dashed gray arrows represent regulation which is not yet known to be direct or indirect. E: a computation-based genome-wide search for additional cis-regulatory modules containing binding sites for the five TFs that act at the MHE led to the discovery that the hbr gene is regulated in a manner similar to eve (45); additional putative elements for other genes are still under investigation. F: microarray experiments, currently in progress, are being used as another method of discovering target genes downstream of the Wg-Dpp-Ras signaling code. G: clustering and motif-extraction algorithms can then be used to discern possible TF binding sites associated with the genes found in the microarray experiments. The AlignACE motif-finding program (48, 86) has been used to discover common motifs in the elements identified in the computation search; at least one of these has been experimentally determined to represent a repressor binding site (45). Yeast two-hybrid analysis and similar methods can now be used to determine the TFs that bind to motifs discovered in this way or to sequences in the regulatory elements that show evolutionary conservation (not pictured). In this way, genetic, molecular, and computational techniques are all being used together to uncover the details of the genetic regulatory networks underlying dorsal mesoderm development.

Physiol Genomics • VOL 10 • www.physiolgenomics.org
INVITED REVIEW: GENETIC REGULATORY NETWORKS IN DEVELOPMENT

A Embryonic pattern formation

B Signaling code for Eve progenitors

C Transcriptional code, feedback loops, and lateral inhibition

D Cis-regulation: transcriptional integration at Eve MHE

E Computational search for MHE-like regulatory modules/phylogenetic footprinting

F Discovery of new respondents to signaling code via microarray analysis

G Computational discovery of new binding site motifs
Additional analysis of the computational search results using the AlignACE program (48, 86) has revealed a number of previously unidentified TF binding sites in the MHE (Fig. 3G), one of which has been experimentally determined to be involved in transcriptional repression (45). Yeast one-hybrid assays and candidate gene approaches guided by genetic data can be used in an effort to identify the factors that bind these sites. These results, too, can be checked with respect to the microarray analysis and by in situ hybridization of mRNA to whole embryos, which will confirm the spatial expression patterns of the genes. Genetic experiments have revealed the existence of cooperative non-autonomous inhibition, and mutually antagonistic cross-talk. As the mechanisms underlying these interactions become understood in greater molecular detail, they too can be added to our picture of the network. In this fashion we are slowly building up a map of the entire regulatory network, starting with the initial inductive signals responsible for progenitor specification and moving through the genes regulated by these signals, the additional cooperating mesodermal cofactors, the cis-regulatory apparatus that integrates this information, and the downstream output of genes involved in mesodermal progenitor development. The latter will in turn provide a basis for understanding the later steps of organogenesis that lead to the foundation of a mature heart and body wall musculature.

Concluding Remarks

What other developmental systems might be suitable for undertaking a detailed regulatory analysis? An ideal system would be in an organism having a fully sequenced, annotated genome; sequenced sister species; available whole transcriptome microarrays, or better, whole genome microarrays; the ability to undergo genetic manipulation; a range of additional experimental techniques for perturbing gene expression such as regulated ectopic gene expression, construction of dominant-negative proteins, and RNAi or antisense RNA; defined cell culture assays; and a wide range of available cell-specific markers so that experimental outcomes could be assessed in detail. The system itself should be one for which the utilized signal transduction pathways and main tissue-specific TFs have been defined and in which at least one well-characterized transcriptional enhancer exists that can be used as a starting point for cis-element discovery. Most of these characteristics are present or are soon to be present in the genetic model organisms currently favored by developmental biologists and in the organ systems that have been their focus of study. The Drosophila eye, the C. elegans vulva, and the mammalian pituitary gland are just three among the many well-studied systems that seem well-poised to serve as foci for regulatory network discovery (5, 90, 106).

Metazoan development is a complex and highly orchestrated process in which multiple cell movements, signaling interactions, and changes in gene expression must be spatially and temporally coordinated to ensure that embryogenesis proceeds correctly. Understanding in detail each of these many interrelated processes is a daunting task. Posttranscriptional processing of RNAs and proteins plays an important role in development, and our tools for comprehensively investigating these processes remain limited. With the rise of genomic technologies, however, the necessary tools for exploring transcriptional regulation on a system-level scale are finally becoming available. By combining genetic analysis, microarray and other transcriptional profiling methods, and computational approaches to identifying and characterizing cis-regulatory sequences, we can begin not just to unravel complex genetic networks, but also to consider them holistically as part of a single integrated process, embryonic development.

We thank M. Boutros and Y. Grad for helpful comments on the manuscript and S. Choe for sharing the data in Fig. 2.

M. S. Halfon is supported by the Charles A. King Trust of The Medical Foundation. A. M. Michelson is an Associate Investigator of the Howard Hughes Medical Institute.

REFERENCES


56. Kapranov P, Cawley SE, Drenkov J, Bekiranov S, Strauberg RL, Fodor SP, and Gingeras TR. Large-scale