Microarrays and beyond: What potential do current and future genomics tools have for breeders?[1]

B. Walsh*[2] and D. Henderson†

*Departments of Ecology and Evolutionary Biology, *Molecular and Cellular Biology, *Plant Science,
†Animal Science, and †Epidemiology and Biostatistics, University of Arizona, Tucson 85721

ABSTRACT: One of the most exciting tools from genomics is the ability to obtain a whole-genome snapshot of gene expression. This is typically called a *microarray analysis*, because probes for the genes of interest, which can run into the thousands, are spotted in a very small array on a glass slide or some other substrate. The resulting array is often called a *gene chip*, or simply a *chip*, in the case of short oligo arrays, or slides in the case of cDNA or long oligo arrays. Microarrays offer the awesome potential of simultaneously examining the level of expression, where expression is intended to measure the standing amount of mRNA, for all of the genes in a genome. Given this potential, it is not surprising that microarrays have attracted a great deal of attention from animal geneticists and breeders. The purpose of this review is to provide a brief, yet critical, overview of some of the potential uses of such whole-genome expression studies in applied animal breeding and to speculate about what additional forthcoming tools might be of use.

J. Anim. Sci. 2004. 82(E. Suppl.):E292–E299

Introduction: Biotechnology and Animal Breeding

This review arose from a symposium focusing on applications of biotechnology in animal breeding, in which Dekkers (2004) examined the current successes (and failures) of marker-assisted selection (MAS) whereas Pomp et al. (2004) reviewed the use of mouse as a model system for moving beyond QTL. Here we build on material presented in both these papers, showing how MAS and analysis from model systems interface with structural genomics and, in particular, expression and pathway analysis may affect the breeder of both today and tomorrow. Many of our comments by their very nature are highly speculative, and hence in the future are likely to (at best) be regarded as somewhat ignorant or (at worst) be downright misleading. Our main purpose here is to stimulate critical discussion of how helpful detailed genomic information currently is to breeding programs and how helpful it may become in the future. Although genomics is revolutionizing how we address a great number of problems in biology, how useful this technology will ultimately prove to breeders is yet unclear. Without question, genomic information will provide vital and critical insights into how biological systems work, and such knowledge is obviously very helpful to any applied biologist. What (at least in our minds) remains uncertain is whether genomics will fundamentally transform future breeding programs or simply be another tool (akin to BLUP) for use by the sophisticated breeder.

A Pico-Primer on Microarrays

A few very brief comments on microarray technology are in order to provide the appropriate background for subsequent comments. Detailed reviews can be found in Knudsen (2002) and Causton et al. (2003). An excellent bibliographic source for the most-current material can be found at the Microarray Data Analysis website, http://www.nslij-genetics.org/microarray/bar.html.

An Overview of the Technology

The basic concept behind microarrays is hybridization, wherein one extracts mRNA from a collection of cells of interest and then hybridizes the mRNA into a series of probes for target genes of interest. Each individual hybridization reaction on an array is referred to as a *spot* or a *feature*, and a typical array may contain thousands of spots. In theory, at one extreme all of the roughly 30,000 genes within a mammalian genome can be spotted onto a single microarray. In

[1] This article was presented at the 2003 ADSA-ASAS-AMPA meeting as part of the breeding and genetics symposium Molecular Genetics: Lessons from the Past/New Directions.

[2] Correspondence: Dept. of Ecology and Evolutionary Biology, BioSciences West (phone: 520-621-1915; fax: 520 621-9190; e-mail: jbwalsh@u.arizona.edu).

Received November 18, 2003.
Accepted January 25, 2004.

practice, however, one portion of good experimental design technique includes multiple technical replicates of each gene on the same slide, in which case scoring an entire genome then requires several microarrays. The (sample-adjusted) hybridization levels for any particular gene are then compared across two (or more) samples (treatments), which could be different cell types, time points, or individuals. In this fashion, any set of genes of interest, up to the entire genome, can be simultaneously compared for differences in relative expression across the treatments of interest.

Two different approaches have been used for generating the probes in an array. With synthetic oligonucleotide arrays, one uses sequence information to chemically synthesize the probe sequence (oligonucleotide), often directly onto the slide/chip/membrane, for example, by using photolithography, as is the approach of Affymetrix (Affymetrix, Inc., Santa Clara, CA) and Agilent chips (Agilent Technologies, Inc., Palo Alto, CA). The other approach is to use PCR to generate cDNA probes, which are then spotted or printed on the array using a high-precision robot. We will focus on spotted microarrays, although our comments equally apply to synthetic oligonucleotide arrays.

Following extraction, mRNA from two treatments under comparison are reverse-transcribed into cDNA for hybridization on the arrays, using two fluorescent dyes (Cy3 and Cy5) to mark the cDNA from the two treatments. The cDNA generated from one treatment fluoresces green and the cDNA from the other fluoresces red under different wavelengths of light. The total cDNA from both samples is then run over the array, resulting in a grid of small green, yellow, and red spots. Genes with a yellow spot (feature) indicate roughly equal levels of mRNA in both samples for that feature, whereas those that appear green or red have the majority of mRNA from just one sample. Formally, the levels of fluorescent intensity are measured on both fluorescent wavelengths (or channels) for the two dyes. A digital image is produced for each channel and an estimate of the relative intensities of both colors is generated for each spot.

Analysis of Microarray Data

As one might imagine, there are a number of image-processing and standardization issues in correcting both samples to allow for a proper comparison. Parmigiani et al. (2003) provides an excellent review of the various issues, which we will regard as being (largely) solved for our discussion, although this is still an area of active research.

After the preparation and image processing stages, the critical issue of assessing statistical significance of observed differences in expression arises. Shockingly, many of the initial microarrays were run with absolutely no biological replication. Although the molecular technology was at the very cutting edge, the initial experimental designs were certainly "Stone Age." It is now appreciated that at least some replication must occur. An ideal design involves multiple spots for the same gene within an array (i.e., technical replication), and several replicates of each complete array representing different biological samples (i.e., biological replication). Further, it has been shown that there are dye-gene interactions, so that one can obtain different levels of expression for a gene simply from which dye is used. As a result, a dye-swapping design, or a loop design (Kerr and Churchill, 2001), is good practice, in which one replicates the comparisons, swapping the allocation of the Cy3 and Cy5 dyes over the two treatments.

The resulting data can easily be handled as simple linear models not unlike those used by plant breeders in multiregional field testing of new cultivators (e.g., Gauch, 1992) . For example, consider gene i (the index i may run into the thousands) from a sample of type j (the treatment; for simplicity, we consider two types). In a well-designed experiment, there are multiple arrays and replicate spots of each gene within an array. The resulting level of expression y_{ijkl} for replicate l of gene i in treatment j on array k is just

$$y_{ijkl} = u + A_k + R_{l(k)} + T_i + G_j + TG_{ij} + e_{lkij} \quad [1]$$

where y is the log of the spot intensity. The A and R terms control for array and location within array effects, T is the average treatment effect, and G the average gene effect. Of interest are the within-gene TG contrast terms, corresponding to the gene × treatment interaction, indicating changes in the level of expression of the gene over treatments.

The final, and perhaps most vexing issue, is that of clustering and classification. Clustering, detecting groups of coexpressed genes (in an attempt to draw some inference about the underlying regulatory networks or response mechanisms) is largely done in an ad-hoc fashion. For example, one standard approach is k-means clustering (Tavazoie et al., 1999), wherein one specifies the number of clusters k in advance, and after which least squares, or other approaches, are used to find the optimal classifiers for the clusters. Given that a typical microarray experiment involves measurement of hundreds to thousands of genes with very few (typically far less than a dozen) replicates, the very nature of these data will result in clusters. This additional noise complicates attempts to deduce any underlying biological clustering. Other common approaches to clustering are self-organizing maps (Tamayo et al., 1999), model-based clustering (Yeung et al., 2001), and hierarchical clustering (Eisen et al., 1998).

The related problem of classification, finding those genes at which changes in mRNA-expression level predict phenotype, rests on a far firmer statistical foundation. For example, one could use logistic regression to estimate those genes contributing the most information in classifying the identity of a binary phenotype (Lee et al., 2003). Classification is the issue that is of more immediate concern to a breeder because we would like to use microarrays to find genes of interest for selection and their subsequent incorporation into elite lines.

Microarray Analysis Is Best Regarded as an Exploratory Data Analysis Approach

Given that the number of tests of significance (for nonzero gene × treatment interactions) greatly exceeds the number of data vectors, there are very serious issues with multiple tests. A further complication is that we expect many of the tests to be correlated as gene expression, by its very nature, is often highly correlated. Standard approaches for controlling the experiment-wide *P*-value are often extremely overly conservative, mostly ignoring the problem of highly correlated tests, although Benjamini and Yekutieli (2001) address the treatment of dependence in multiple hypothesis testing. Further, the reason for performing a microarray experiment is that we indeed expect a large number, but not large in relation to the total number of contrasts, of the T × G interactions (the TG terms in Eq. [1]) to be significant.

A more reasonable approach is to consider a microarray experiment as exploratory data analysis, with the goal of using the results from one experiment to produce a reduced set of genes for future consideration. For a chip testing 10,000 genes, extracting a set of 200 potential candidates is a very significant reduction. In such a setting, instead of controlling the experiment-wide *P*-value, we should instead aim to control either the false discover rate or the proportion of false-positives (Storey and Tibshirani, 2003a,b; Fernando et al., 2004). Empirical Bayesian approaches (Efron et al., 2001; Newton and Kendziorski, 2003) offer yet another avenue.

An unfortunate and egregious idea common among early practitioners of microarrays is that exploratory data analysis is free of statistical rigor. This is far from the truth, as many exploratory methods based on sound statistical theory exist (Hastie et al., 2001; Cook et al., 2004). Additionally, inferential statistical methods, such as ANOVA, are commonly used as exploratory tools—where the false discovery rate is used to report and control Type I error rates in detecting differentially expressed genes with adjusted *P*-value thresholds often set at rather liberal limits (e.g., $P < 0.20$). Last, to facilitate explorations by others, scientists often curate data in publicly available databases, where researchers without ties to the experimental labs that generated the data are free to explore in search of new sets of genes suitable for new experiments (http://genome-www5.stanford.edu/).

Problems (and Pitfalls) of Gene Discovery via Microarray Analysis

With a complete genome sequence in hand, arrays can be constructed to monitor levels of mRNA expression at conceivably every protein-coding gene. Given that the majority of genes discovered in genomic sequence scans are of unknown function, microarrays obviously offer a powerful approach for detecting potential candidates genes. A suitably designed microarray experiment can generate a list of genes whose expression is significantly different in high vs. low lines (or individuals), over different environments, or in different tissues. This is an exciting opportunity and, hence, it is not surprising that breeders have been very interested in the potential of microarrays.

But just how informative are microarrays for candidate loci? There are several critical limitations. The first concern is that the wrong treatments may be considered. The difference in some economic trait may be due to differential expression of genes in tissues very different from the target tissue in which the trait is apparently expressed. A good (conceptual) example is to think of grain yield in plants. An exhaustive microarray study focusing on flower and seed expression would completely miss changes in expression in the roots that might result in more resources for the plant and hence greater yield. One can imagine comparable situations in animal systems.

A second caveat is that microarrays simply score mRNA levels, and other levels for the control of gene expression may result in the differences between trait values, for example post-mRNA processing or post-translation processing of proteins. Microarray systems that use a series of small nonoverlapping probes within each gene (e.g., Affymetrix chips) may catch differences in mRNA processing. Spotted cDNA or long oligo microarrays, on the other hand, use much larger probes, and, therefore, a variety of differentially processed mRNA from the same gene may yield the same expression signal. Probe size is critical as with a longer probe (70 to 100 nucleotides for long oligo arrays and >1,000 nucleotides for cDNA arrays), a single-base-pair change typically will not result in a change in hybridization pattern. In such cases, two alleles showing the same level of mRNA expression but producing different protein products can still show the same amount of hybridization and, hence, be missed by a microarray screen of candidate loci based on differential expression.

A far greater complication is that an observed change in the expression level at gene X may actually be caused by a *trans*-acting factor from gene Y. Indeed, it seems that the majority of observed changes in experiments using recombinant inbred lines of mice for QTL mapping of microarray changes, wherein the expression level at each gene is treated as the trait of interest for mapping, are due to *trans*-acting factors (D. Threadgill, University of North Carolina, personal communication). Analysis of such experiments often shows QTL for expression level in a particular gene mapping to that gene itself (a *cis*-acting control factor). However, there very often are QTL for expression that map to locations very different from the target gene (*trans*-acting factors). One especially interesting class of *trans*-acting factors that appears in joint QTL-microarray analyses are so-called master controllers, or global regulators, wherein a single genomic location influences expression levels for a very large number of genes. An interesting graphical way to display information from a joint QTL-microarray experiment is given by Pomp et al. (2004). Here, one plots the genomic location for the gene whose expression is being considered on one axis and genomic locations for any QTL for expression of this gene on the other axis. Points on the diagonal indicate *cis*-acting factors, presumably either within the gene itself or very tightly linked to it. Off-diagonal elements indicate *trans*-acting factors, wherein the gene influencing levels of expression at the target is some distance from the target. Master-control genes are indicated by either a horizontal or vertical run of spots, depending on whether QTL position is plotted on the horizontal or vertical axes, respectively.

The nature of regulation that is causing the change in expression in the target gene in a microarray is a key and usually unknown feature for breeders. If the change in expression is entirely *cis*-acting, then one can treat the target as a candidate gene and a potential target for MAS. If the change in expression is due to one (or more) *trans*-acting factors, although correlations between target expression and phenotype may be very high, the only way for a breeder to exploit the existing variation is to perform a QTL mapping experiment in order to find markers for MAS on the *trans*-acting factor. If the change in expression is in part due to a master controller, this has potentially significant implications for correlated response in other, perhaps unwanted, traits.

Exploiting Expression Data for Breeding

Keeping in mind the foregoing important, serious caveats, what is the best approach to use expression data in an applied breeding setting? The most obvious approach is to use genes demonstrating changes in expression that correlate with desirable phenotypes. A standard Lande-Thompson (1990) or other modified MAS index can be used with markers linked to the candidates. Given that change in expression level can often be due to *trans*-acting factors, this may only result in a modest (or even no) improvement in selection response, and it is unlikely to be cost-efficient in many settings.

Although there may be a mixed (or even bleak) picture for using many expression-suggested target genes for improving selection response, there are two areas where microarrays may indeed prove useful for breeders. The first is in genotype × environment interaction. Ideally, one could screen gene expression levels over several environments (such as heat or water stress). Although the obvious focus is on genes that show significant changes in expression over environments, more interesting targets are those candidate genes that show reduced levels of variance in gene expression across environments. This may suggest loci that are well buffered across environments, leading to more stable phenotypes over a range of environments.

A second potentially interesting use of a genomewide expression analysis is in dealing with undesirable correlated effects from selection response. Recalling that master controllers effecting the expression of a large number of genes are seemingly not uncommon, it is highly desirable when using MAS to purposely avoid using any such controlling genes. The collection of genes that serve as master controllers are likely to be somewhat conserved over related species. If this assertion is correct, then experiments from mice and rats will suggest at least a subset of candidate master controllers for most domesticated animals. With this set in hand, one can look for marker-trait associations involving the trait of interest and variation at these candidate controllers. A restricted selection index (e.g., Kempthorne and Nordskog, 1959; Tallis, 1962) can then be used to maximize change in the trait of interest while constraining the increase in frequency of any markers associated with variation in the expression of master control genes. Although the ideal, and usually unrealized, setting is when one has QTL underlying expression changes, one can also use estimated gene clusters from a microarray. Suppose we have two potential candidates, one of which is clustered with a number of other expression changes, the other apparently not an obvious member of any cluster. There is a good chance that the first candidate is influenced by a change in a master controller gene, and hence selection on this target is likely to influence expression at a number of other sites, and potentially other traits. On the other hand, the second candidate may be more localized solely to the trait of interest, and selection on this target may result in less-undesirable correlated responses.

Beyond Microarrays: Pathway Analysis

We can obviously do a better job exploiting the information from microarrays if we know the actual genes influencing expression at target sites, as opposed to simply selecting on the target sites alone. This hints at the next phase of functional genomics, namely the estimation of the topology of gene networks and (ideally) the dynamic control parameters of such networks. Surely, with such information in hand, we can dramatically improve selection response. In reality, however, the improvement may be rather modest, as we detail below.

Microarrays provide a snapshot of the global pattern of mRNA expression, and, even though this is certainly quite helpful, it is by no means definitive in the estimation of genetic networks. A variety of other tools from molecular biology are available for looking at how products within a cell interact. One approach is to use two-hybrid screens, originally developed in yeast (Fields and Song, 1989). In a two-hybrid screen, a reporter gene is expressed if and only if the two proteins of interest come into direct physical contact. In both yeast and the nematode, tests of all pairwise interaction among all known genes have generated two-hybrid interaction maps, detailing which proteins appear to interact with one another within the cell (e.g., Wagner, 2000). Another tool is the two-dimensional protein gel, the protein equivalent of a microarray, giving a snapshot of all proteins within a cell and providing a rough quantification of their concentrations. Other approaches—such as fluorescent energy resistance transfer, or FRET, and fluorescence recovery after photobleaching, or FRAP—have been developed for estimating how close two proteins come into contact and for following individual protein products around in the living cell (reviewed by Lippincott-Schwartz et al., 2001). Coupled with microarrays, these and other genomics tools offer the very real promise of a far greater elucidation of genetic pathways than we currently possess. Ultimately, breeders may have access to a detailed wiring diagram of the genetic regulation (i.e., the genetic networks) for traits of interest, such as milk production or fat and muscle deposition. These pathways will have been worked out in model organisms, such as the mouse and rat, and this information will then be available for breeders to potentially exploit in their breeds of interest. What level of such information will be helpful and how can it best be exploited?

Analysis and Exploitation of Gene and Metabolic Networks

A genetic or metabolic network involves two components: a topology and a set of dynamic rules that act on the topology. The topology of a network—namely, elucidation (for each gene) of the set of other genes that it immediately interacts with—can be expressed as a matrix and conveniently visualized as a graph. The nodes of the graph are the elements in the pathways, whereas the edges (lines) of the graph show all of the elements that any particular gene directly influences. In matrix form, the topology matrix M contains only zeros and ones. If the ij^{th} element of A is 1, this implies that gene i influences gene j. Note that influences can be asymmetric, with $M_{ij} = 1$, whereas $M_{ji} = 0$. Nonzero elements in M^k denote elements that influence a gene after k-steps through a pathway. When the dynamic rules of a pathway are known, the simple matrix of zeros and ones is replaced by a vector-valued function, taking in the current state space at each gene and returning the new state of the system. Obviously, estimating the topology is the first (and easiest) step in completely describing a network.

Genomic tools, such as two-hybrid screens, offer a glimpse of the underlying topology of protein-protein interactions. A major quest among functional genomicists is to use the static snapshot of the entire genome offered by microarray and two-dimensional protein gels to at least partially reconstruct the topology of cellular networks. One standard approach is to follow expression over time, and another is to exploit perturbations in the mRNA or protein levels of particular genes to see how the network responds. Although progress has been slow using such approaches, the inventiveness of molecular biologists will undoubtedly provide additional tools for estimation of network topologies. We note that the statistical issues of such a reconstruction are still largely open, although tools from phylogenetics, such as bootstrapping (Felsenstein, 1983), may prove useful. Thus, in the near future, breeders will have access to at least partial (perhaps very noisy) topologies of traits of interest, showing which genetic members influence each other.

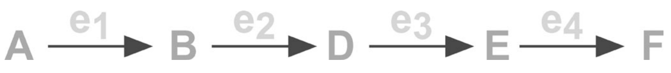

Figure 1. A very simple (linear) gene network. A through F correspond to products in the pathway; e1 through e4 are genes or enzymes involved in the transition from one step to the next. Note that this is a very simplistic version, and a typical gene network is expected to be very convoluted, with loops, branches, and feedback cycles.

Just what information can a breeder exploit from such a topology? Consider the simple linear network given in Figure 1. Suppose the ultimate goal is to increase the production of the product F. The flux of a pathway is the (steady state) rate at which a product is produced. The gene products e1 through e4 (typically proteins, although these could also be structural RNA) move inputs through the pathway to produce F. Hence, if Figure 1 represents the topology of the gene network for the trait of interest in a model system, it immediately suggests strong candidate loci (the loci coding for e1 through e4) to identify association between variation at these loci and the trait of interest. Of course, as suggested from microarray studies, any number of other genes may influence the expression levels of the genes responsible for e1 through e4. We can go a step further and ask how best to increase the flux through the system, although this moves us from simple issues of the topology to the even more complex issues of the dynamics of the network. In particular, if we could engineer an up- (or down-)regulated gene in this pathway, which gene should be the target? One choice would be to up-regulate the gene(s) responsible for the production of e1, front-loading the pathway. However, one could also argue that up-regulation of e4 may have a greater impact on the flux. The point here is that although the topology is certainly useful, by itself it does not immediately suggest the rate-limiting step(s) in the pathway. Deduction of the key limiting steps in a pathway requires not only the topology, but additional information as well. Fortunately, there is a large body of theoretical literature on the control of metabolism (e.g., Fell, 1997; Hofmeyr and Westerhoff, 2001) and the results from such metabolic control theory provide some critical insights and tools for exploiting pathway information.

Kascer-Burns Sensitivity Analysis

In a landmark paper, Kacser and Burns (1973) developed the basic framework for the study of flux through biochemical pathways. Before Kacser and Burns (1973), there was the widespread notion that most pathways were limited by one or more rate-limiting steps. Kacser and Burns (1973) quantified the actual limitation (or, equivalently, the regulation) imposed at each step in a pathway by introducing the concept of a flux control coefficient. The control coefficient is denoted by C_j^i, for the flux at step i (F_i) in a pathway due to enzyme j (E_j). Roughly speaking, C_j^i is the percent change in flux (through i) divided by the percent change from a small change in the activity (or amount) of j:

$$C_i^j = (\partial F_i/\partial E_j)(E_j/F_i) = \partial \ln F_i/\partial \ln E_j \qquad [2]$$

Control coefficients provide a quantitative description of the critical control points within a pathway (and indeed, for any part of the pathway). The largest change in flux is generated by increasing the amount (or activity) of the element (gene, gene product, or protein) with the largest control coefficient. A remarkable feature of control coefficients is the Kacser-Burns summation theorem, which states that the sum of the control coefficients over all elements contributing to the flux at step j sums to 1, namely,

$$\textstyle\sum_i C_i^j = 1 \qquad [3]$$

Because control coefficients are generally positive (see below), the total control of flux through a pathway is distributed over the components, and large C values (i.e., close to 1) are expected to be rare. Rather, the maximal value of a control coefficient for any element within a pathway is likely modest at best, resulting in rate-limiting steps being rather rare. In general, the control of the flux through a pathway or system is shared by all members of that pathway or system. In a quantitative-genetics framework, genes of modest effect on flux are expected to be much more common than major genes (those with large control coefficients). If pathways are highly branched or negative control is exerted by one (or more) members within a network, negative flux coefficients can occur (wherein an increase in a gene product reduces the flux).

The second feature following from the summation theorem is that if a particular control coefficient is greatly increased in value, this decreases the values of other control coefficients in the system as flux control is shared by all members of the pathway. Hence, control coefficients are not intrinsic properties of an enzyme (or gene) but rather properties of a (local) system. This is much akin to the average effects of an allele, which are also population-dependent. The interested reader is referred to Kacser and Burns (1981) for a description of the relationship between metabolic control theory and allelic forms. Likewise, control coefficients evolve, again akin to the average effect of an allele changing as the population evolves.

The control coefficients can be related to the factor by which flux can be increased by the Small-Kacser (1993) theorem: an r-fold increase in activity (or amount) of gene E results in an f-fold increase in the flux through element j, where

$$f = 1/[1 - C_E^J(r - 1)/r] \rightarrow 1/(1 - C_E^J) \text{ as } r \rightarrow \infty$$

Thus, even an infinite increase in the activity/amount of E results only in a twofold increase in flux if the control coefficient is 0.5 (which is a rather large value). A fourfold increase in flux requires a control coefficient of 0.75, and a tenfold increase, a coefficient of 0.9.

Just how extendable metabolic control theory will prove to be for gene networks remains unclear, as it makes a key assumption that the system is in steady state. Nonsteady states may turn out to be the norm in many gene networks, yet metabolic control theory has been shown valid for nearly steady states (Liao and Delgado et al., 1997). Despite this (and other) limitations, something similar to metabolic control theory will be developed for dissecting gene networks, and analogs to control coefficients will likely be key components in such a theory. Hence, although we view metabolic control theory as a useful cartoon for the initial analysis of pathways, it is helpful to consider what use the breeder would have with such information.

Selection on Gene Networks

For point of discussion, suppose that we not only know the topology of the gene network underlying the trait(s) we wish to select on, but also have some information on the control coefficients as well. Obviously, this is an ideal situation, but just how would we use this information?

First, the genes with the largest C values are the obvious initial candidates for selection. For example, one could screen for variation in expression level or activity at such loci and incorporate this information into a selection index, or more generally in a BLUP framework, along with other genetic and phenotypic information. A much more interesting use of an estimated gene network would be in trying to reduce undesirable correlated responses. Consider the network in Figure 2, in which our goal is to increase the flux to F (the trait we wish to improve) while keeping the flux through H unchanged (a trait where any change would be undesirable). With knowledge of just the topology of this network, one might search for variation in the expression or activity of products e3 or e4, as these occur after the pathways from A to both F and H split from each other. The problem with this argument is that "pulling" more of a flux toward F might result in less product going to H, creating a correlated response. With control coefficients in hand, one can look at the effects of each element in the network on the fluxes

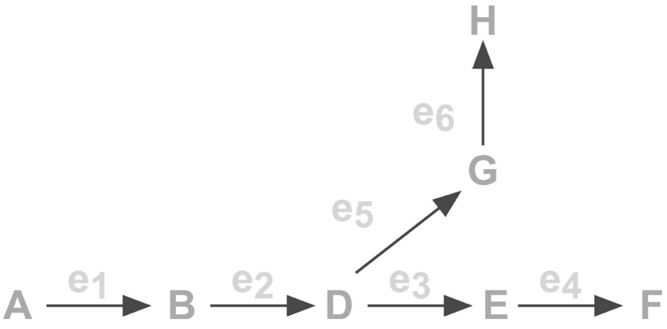

Figure 2. A very simple branched gene network, leading to two final products, F and H, from a common precursor A. The letters A through H correspond to products in the pathway; e1 through e6 are genes or enzymes involved in the transition from one step to the next.

through F and H, and base selection decisions on the genes suggested by such an analysis. For example, a Kempthorne and Nordskog (1959) or Tallis (1962) restricted selection index can be used to select for genes increasing the flux through F while keeping the flux through H as constant as possible (since flux is shared, any change in the flux toward F will likely result in a change in the flux through H).

Another interesting possibility is to select for variants that tend to smooth out the changes in flux as inputs vary, increasing the stability of the character. With both a topology in hand and information on which pathway components are influenced by environmental factors (for example, heat or water stress) it might be possible to select for networks where the flux control coefficients are reduced for those elements that interact with the initial environmental signals. This would have the effect of damping the effects of environmental signals on the final flux through the pathway.

Implications

Modern molecular biology and genomics have certainly forever changed the face of biology, and animal breeding is no exception. It is imperative, however, to not simply rush blindly forward with the implementation of new technologies but rather to critically assess which tools are likely to be of significant benefit for improving breeding. The lessons from the initial heady days of quantitative trait loci mapping are quite informative, namely that although the promise of these approaches is extremely exciting, their application is very often more problematic than initially envisioned. The use of expression arrays and the potential future use of information on gene networks should heed these lessons. Selection based solely on phenotypic information still performs favorably, and any movement toward a more genomics-based approach for any particular selection problem should first consider a serious cost-benefit analysis.

Literature Cited

Benjamini, Y., and D. Yekutlieli. 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Statistics. 29:1165–1188.

Causton, H. C., J. Quackenbush, and A. Brazma. 2003. A Beginner's Guide to Microarray Gene Expression Data. Blackwell, Oxford.

Cook, D., B. Nikolau, and E. Wurtele. 2004. Visual methods for data from two factor single replicate gene expression studies. J. Comp. Graph. Stat. (In press).

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82:(E. Suppl.):E313–E328.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. PNAS 95:14863–14868.

Efron, B., J. Storey, R. Tibshirani, and V. Tusher. 2001. Empirical Bayes analysis of a microarray experiment. JASA 96:1151–1160.

Felsenstein, J. 1983. Statistical inference of phylogenies (with discussion). J. Royal Stat. Soc. A, 146:246–272.

Fell, D. 1997. Understanding the Control of Metabolism. Portland Press, London.

Fernando, R. L., D. Nettleton, B. Southey, J. C. M. Dekkers, M. F. Rothschild, and M. Soller. 2004. Controlling the proportion of false positives (PFP) in multiple dependent tests. Genetics 166:611–619.

Fields, S., and O. Song. 1989. A novel genetic system to detect protein-protein interactions. Nature 340:245–246.

Gauch, H. G., Jr. 1992. Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier, Amsterdam.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, New York.

Hofmeyr, J.-H. S., and H. V. Westerhoff. 2001. Building the cellular puzzle. Control in multi-level reaction networks. J. Theor. Biol. 208:261–285.

Kacser, H., and J. A. Burns. 1973. The control of flux. Symp. Soc. Exp. Biol. 27:65–104.

Kacser, H., and J. A. Burns. 1981. The molecular basis of dominance. Genetics 97:639–666.

Kempthorne, O., and A. W. Nordskog. 1959. Restricted selection indices. Biometrics 15:10–19.

Kerr, M. K., and G. A. Churchill. 2001. Analysis of variance for gene expression microarray data. J. Comp. Biol. 7:819–837.

Knudsen, S. 2002. A Biologist's Guide to Analysis of DNA Microarray Data. Wiley, New York.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Lee, K. E., N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick. 2003. Gene selection: A Bayesian variable selection approach. Bioinformatics 19:90–97.

Liao, J. C., and J. Delgado. 1997. Flux calculation using control constraints. Biotechnol. Prog. 14:554–560.

Lippincott-Schwartz, J., E. Snapp, and A. Kenworthy. 2001. Studying protein dynamics in living cells. Nat. Rev. Mol. Cell Biol. 2: 444–456.

Newton, R. A., and C. Kendziorski. 2003. Parametric empirical Bayes methods for microarrays. Pages 254–270 in The Analysis of Gene Expression Data. G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, ed. Springer, New York.

Parmigiani, G., E. S. Garrett, R. A. Irizarry, and S. L. Zeger. 2003. The Analysis of Gene Expression Data. Springer, New York.

Pomp, D., M. Allan, and S. Wesolowski. 2004. Quantitative genomics: Exploring the genetic architecture of complex trait predisposition. J. Anim. Sci. 82:(E. Suppl.):E300–E312.

Small, J. R., and H. Kacser. 1993. Responses of metabolic systems to large changes in enzyme activities and effectors. 1. The linear treatment of unbranched chains. Eur. J. Biochem. 213:613–624.

Storey, J. D., and R. Tibshirani. 2003a. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. Pages 272–290 in The Analysis of Gene Expression Data. G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, ed. Springer, New York.

Storey, J. D., and R. Tibshirani. 2003b. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U.S.A. 100:9440–9445.

Tallis, G. M. 1962. A selection index for optimum genotype. Biometrics 18:120–122.

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. U.S.A. 96:2907–2912.

Tavazoie, S., D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. 1999. Systematic determination of genetic network architecture. Nat. Genet. 22:281–285.

Wagner, A. 2000. Mutational robustness in genetic networks of yeast. Nat. Genet. 24, 355–361.

Yeung, K. Y., C. S. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. 2001. Model based clustering and data transformations for gene expression data. Bioinformatics 17:10:977–987.