The Journal of Experimental Biology 209, 1584-1592
Published by The Company of Biologists 2007
doi:10.1242/jeb.004333

Interpreting physiological responses to environmental change through gene expression profiling

Andrew Y. Gracey

Marine Environmental Biology, University of Southern California, 3616 Trousdale Parkway, Los Angeles, CA 90089, USA

e-mail: gracey@usc.edu

Accepted 12 March 2007

Summary

Identification of differentially expressed genes in response to environmental change offers insights into the roles of the transcriptome in the regulation of physiological responses. A variety of methods are now available to implement large-scale gene expression screens, and each method has specific advantages and disadvantages. Construction of custom cDNA microarrays remains the most popular route to implement expression screens in the non-model organisms favored by comparative physiologists, and we highlight some factors that should be considered when embarking along this path. Using a carp cDNA microarray, we have undertaken a broad, system-wide gene expression screen to investigate the physiological mechanisms underlying cold and hypoxia acclimation. This dataset provides a starting point from which to explore a range of specific mechanistic hypotheses at all levels of organization, from individual biochemical pathways to the level of the whole organism. We demonstrate the utility of two data analysis methods, Gene Ontology profiling and rank-based statistical methods, to summarize the probable physiological function of acclimation-induced gene expression changes, and to prioritize specific genes as candidates for further study.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1584/DC1

Key words: acclimation, adaptation, gene expression, microarray.

Introduction

The phenotype of an organism is determined by the combined activities of thousands of genes that are coordinated both temporally and spatially. A goal of so-called '-omic' approaches is to understand complex biological systems by modeling the relationship between multiple measured attributes of biomolecular organization and the phenotype of the organism. Development of high-throughput methods means that large amounts of information can now be gathered at distinct levels of biological organization, allowing genotype, mRNA expression, protein, metabolic and physiological data, to be gathered for a particular organism under a specified set of conditions. Collecting these data is becoming increasingly simple and a rich resource of molecular information is available for the common laboratory model organisms (Bieri et al., 2007; Crosby et al., 2007; Nash et al., 2007).

Since phenotype results ultimately from the expression of genes and gene complexes, understanding patterns of gene expression evoked during changes in physiological state, or in response to environmental change, yields insights regarding the molecular basis of phenotype from the cellular to the whole organism level. A key tool deployed in this research is the measurement of mRNA transcript levels by microarray hybridization (Gracey and Cossins, 2003). The microarray-based approach monitors the expression of many thousands of genes simultaneously, providing a broad view of the transcriptional changes that accompany alterations in physiological state.

The role of the transcriptome in physiological regulation

A significant challenge is how to decipher the large amounts of '-omic' data and then relate them to the phenotype of the study organism. To meet this challenge, a new approach to understanding complex biological systems, termed 'systems biology', has been proposed (Ideker et al., 2001a). A precise definition of systems biology is difficult, but generally speaking systems-based approaches aim to measure several sources of molecular information during genetic or environmental perturbations of a biological system, integrate these data, and then build predictions as to how the system might respond to a different perturbation. Thus, the better the

understanding of, and ability to model, a biological system, the better the predictions will agree with the experimental observations.

Systems-based interpretations of biological processes are revealing new insights into the role of the transcriptome in the regulation of phenotype. A common theme in most systems-based investigations is an effort to link gene or protein expression data with protein–protein and protein–DNA interaction data (Ideker et al., 2001b). The rationale behind this approach is that genes do not function in isolation and instead are components of wider networks of interacting molecules. As components of a wider system, the consequences of gene expression should be interpreted within the context of the behavior of the other molecules that participate in the biology of the organism. For example, systems-based analysis of gene expression has begun to explain why genes that are strongly differentially expressed in yeast stress experiments often turn out to have no discernable effect on stress sensitivity in deletion mutants (Birrell et al., 2002; Giaever et al., 2002). Using the yeast response to arsenic as a model, the analysis of deletion knockouts revealed that the genes that conferred the most sensitivity to arsenic were in pathways upstream of the arsenic detoxification pathways, while expression profiling identified genes that were members of downstream pathways that ultimately protect against toxicity but which share redundant functions, explaining why they have no phenotypic effect with deletion (Haugen et al., 2004). A similar conclusion was reached upon a system-based analysis of the DNA-damage response of yeast (Workman et al., 2006). So interpretation of gene expression data within the context of regulatory and metabolic networks suggests that gene expression profiling tends to interrogate the downstream effectors of biological responses.

A frequently asked question is whether changes in mRNA levels are a useful proxy for inferring changes in protein abundance? The scientific literature is replete with studies that have tried to address this question, most often by applying a combination of microarray and proteomic techniques to explore the concordance between mRNA and protein levels (Tian et al., 2004). A recent study provides the most definitive exploration of this relationship, and supersedes others by using an extremely accurate method to directly measure protein abundance in living cells (Newman et al., 2006). Using a collection of yeast strains in which each gene is expressed as a GFP-tagged fusion protein, GFP fluorescence was measured to profile the expression of each protein under different environmental perturbations and correlated with changes in the corresponding mRNA abundance. Use of fluorescence provided unprecedented resolution of protein abundance and revealed that mRNA abundance in 87% of genes changed >twofold, showing correlated changes in protein abundance. In a minority of cases, changes in protein abundance were observed in the absence of a change in mRNA level. The conclusion that can be drawn from these results is that mRNA expression profiling is an effective method to identify genes whose protein expression is regulated at the transcriptional level, with the obvious caveat that proteomic techniques are required to identify post-translationally regulated genes. Indeed, a recent estimate assigns 73% of protein variance in yeast to transcriptional regulation (Lu et al., 2007), and so gene expression screens will not provide insights into the regulation of at least 25% of the proteome.

Constructing microarrays for non-model organisms

Array-based technologies are still the main platforms for undertaking large-scale gene expression screens. Arrays can be produced for any organism for which DNA sequence or nucleic acid material is available, and so in theory can be applied to any given organism. In my laboratory this has included constructing custom arrays for a variety of species including common carp (Gracey et al., 2004), the goby *Gillichthys mirabilis* (Gracey et al., 2001), golden-mantled ground squirrel (Williams et al., 2005) and, recently, for the colonial ascidian *Botryllus schlosseri* and the California ribbed mussel *Mytilus californianus*. Probably the greatest challenge to producing these species-specific ‘boutique’ or ‘bespoke’ arrays is generating the DNA probes that will be physically deposited on the microarray. Two major sources of probes are either PCR products or oligonucleotides. PCR products are generated either by targeted amplification of specific genes based on their DNA sequence, or alternatively by amplification of cDNAs that have been cloned into plasmids. As an alternative to PCR products, long oligonucleotides can be spotted on the array but their application is restricted to genes for which a DNA sequence is available.

Comprehensive sequence data are often limited for most non-model organisms, precluding the design of gene-specific oligonucleotides, and so cDNA microarrays produced in-house will remain the most viable option for most laboratories in the short-term. In this format, cDNA libraries provide a source of cDNAs, which are then amplified by PCR and the products spotted onto the array. Because the primers employed in the PCR are based on the vector sequences that flank the cloned cDNAs, this approach can be employed without prior knowledge of the sequence of the cloned cDNA. This means that sets of PCR-amplified cDNAs can be quickly and affordably created for almost any species.

The procedure for preparing PCR products from cloned cDNAs is simple and within the capacity of most molecular biology laboratories. Briefly, a cDNA library cloned in plasmids is transformed into *E. coli*, plated onto Luria-Bertani agar plates, yielding bacterial colonies that each represent a single cDNA clone. A small portion of each colony is then picked into either 96- or 384-well microtiter plates containing Luria-Bertani broth and propagated. Picking is done in a random fashion and thus the sequence and identity of the cDNA clone present in the host *E. coli* is unknown. The microtiter plates become the picked cDNA library with each coordinate in the plate representing the location of a discrete cDNA clone. Microtiter plates of clones can be copied and frozen, allowing the picked library to be stored indefinitely. Accurate tracking

of the plates of clones throughout the picking and archiving process is an essential step (Konno et al., 2001), since these plates will be the source for the next steps of PCR amplification, arraying and sequencing.

Ideally, we would like to be able to curate a set of cDNA clones that represents all the transcripts encoded by the genome of the organism we wish to study. An array fabricated with this clone-set, a so-called whole transcriptome microarray, would be invaluable since it would offer a global overview of how the expression of every gene in the organism is orchestrated under particular physiological or environmental conditions. However, creating comprehensive cDNA collections that cover the entire genome has proved a challenge. For most of the model organisms, laboratories around the world are collaborating in efforts to create complete cDNA collections, yet after years of work many cDNAs have remained elusive and the collections are still incomplete. The project to identify and sequence every transcript in the mouse genome is particularly noteworthy and an impressive range of strategies and tools has been deployed in this effort (Carninci et al., 2003; Okazaki et al., 2002; Carninci, 2007). With these problems in mind, it is important to consider the most effective strategy to create comprehensive cDNA collections and arrays for new species.

Normalized cDNA libraries

A number of strategies exist for the isolation and curation of cDNA clones for array fabrication. Since the expression pattern of the arrayed genes will be the guide for the interpretation of complex biological response, it is important that genes linked to the physiological process are well represented on the array. The first step to achieving this is to prepare the cDNA library using RNA isolated from the specific tissue(s) and physiological condition that will be the subject of the study. This greatly increases the likelihood that genes that are expressed under these conditions are present as cDNA clones in the library. For example, arrays developed to study the transcriptional response of gobies to hypoxia (Gracey et al., 2001), and killifish to thermal cycling (Podrabsky and Somero, 2004), were prepared from RNA isolated from animals exposed to these respective conditions. Still, the frequency with which the potentially interesting cDNAs are encountered in the library will depend on their abundance in the RNA sample, with highly transcribed genes being more likely to be picked from the library, whilst rare transcripts will be encountered with less frequency. For this reason every effort is made to ensure that rare genes are isolated during the picking of clones from the library. Two important procedures improve this situation, namely normalization and serial-subtraction. Normalization reduces redundancy within the library, bringing the frequency with which each gene is present in the library to within a narrow range, while subtraction enriches for genes specific to a particular environmental treatment, and serial subtraction increases the probability that new genes will be added to the clone-set (Carninci et al., 2000). In our experience, several rounds of serial subtraction is the most effective method of creating cDNA libraries of low redundancy, and while these are time-consuming steps, the results justify the cost.

A universal goal of all of the major cDNA collection and annotation projects for model organisms is to isolate and sequence full-length cDNAs clones (Imanishi et al., 2004; Strausberg et al., 2002). A full-length cDNA clone is one in which the entire coding sequence is present together with the flanking 5′ and 3′ untranslated regions (UTRs). The contribution of full-length clones to understanding gene function cannot be understated. First, the complete cDNA sequence is useful for interpreting genomic sequences, where exons are interspersed with non-coding introns, and each gene may give rise to range of transcripts based on differential splicing. Thus, a full-length cDNA is evidence of the genomic sequence that was transcribed and of alternative splicing events (Zavolan et al., 2003). Second, identification of the translation initiation codon and the 5′ UTR indicates the location of the promoter sequence of the gene, thus helping to direct exploration of the gene's regulatory elements. Third, and most importantly, knowing the sequence of the complete open reading frame of a gene greatly facilitates its functional annotation. In the first instance, it improves homology searches against the public databases and increases the likelihood that its putative function can be assigned based on homology. For cDNAs without recognizable homology to a known gene, characterization of the functional motifs of the protein may help predict its function (Okazaki et al., 2002). Furthermore, knowing the complete amino acid sequence of the encoded protein is important for comparative analyses that aim to understand differences in functional properties of orthologous proteins. For all these reasons, full-length cDNA clones are a valuable foundation upon which further investigations can be constructed.

One caveat with respect to using full-length clones as microarray probes is that their sequence may contain regions that share homology with other genes or contain repetitive sequence. These elements may lead to cross-hybridization between the cDNA and transcripts other than the target transcript. The degree to which these non-specific hybridization signals compromise array analysis is still unclear, but it appears that increasing the stringency of both hybridization and wash conditions can alleviate these problems (Drobyshev et al., 2003; Korkola et al., 2003). On the other hand, the tolerance of long cDNAs to base-pair mismatches can be turned into an advantage, since it allows nucleic acids from related taxa to be hybridized heterologously to a single species array (Renn et al., 2004; von Schalburg et al., 2005).

Subtracted cDNA libraries

Subtracted cDNA libraries are an alternative source of cDNAs for preparing array probes. Subtractive hybridization approaches are used to compare two RNA samples and yield populations of cDNAs enriched for genes that are specifically expressed at higher levels in one sample more than the other (Sagerstrom et al., 1997). Their main advantage is that they

purposefully enrich for cDNAs that are differentially expressed, so fewer cDNAs have to be screened to identify interesting genes. Screening fewer genes on an array improves statistical power since it reduces the number of type I errors that occur when multiple statistical comparisons are made. Subtraction also enriches for genes that are present at low abundance and may not have been discovered in the early stages of randomly picking clones from normalized libraries. In addition the cloned cDNAs tend to be biased towards the most unique gene-specific sequences making them ideal array probes in terms of their specificity. So arrays produced from subtracted libraries can offer some advantages but also present some unique problems with regard to analysis. Most microarray data normalization protocols are based on the assumption that the majority of genes on the array are not differentially expressed and that approximately equal numbers of genes are up- and downregulated (Quackenbush, 2001), but these assumptions may fail for arrays produced from subtracted libraries since spots may show a biased direction of differential expression (Oshlack et al., 2007). Therefore, subtracted libraries should be complemented with cDNAs from non-subtracted sources for microarray construction to alleviate this bias. Another problem is that most methods of cDNA subtraction generate cDNA fragments rather than intact complete cDNAs (Diatchenko et al., 1996), and identification of fragments by sequence homology is problematic for unsequenced organisms (Gracey et al., 2001), and converting fragments into full-length clones is time-consuming.

Oligonucleotide-based arrays and standardization

A problem that has plagued cross-laboratory and cross-platform comparisons of microarray datasets has been the comparison of the probe content of different arrays. Typically, the arrayed probes are annotated using gene names but the assignment of names to DNA sequence is imprecise and strongly dependent upon the sequence database that is used as the reference for the annotation. So as these databases grow and more sequence data come to light, the putative identity assigned to genes evolves and can be subject to revision. This ambiguity leads to real problems when studies have sought to extract matching probes sets across platforms, leading to greater than expected discordance between data derived from different platforms (Tan et al., 2003). Seeking to resolve this problem, recent work has revealed that gene expression data show much greater cross-platform and between-laboratory consistency if the array elements are treated in a sequence-orientated *versus* gene annotation-centered manner (Kuo et al., 2006). This suggests that standardization will only be achieved if probes are matched by DNA sequence instead of by gene name, and this will necessitate a switch to oligonucleotide-based probes for those organisms for which either complete or extensive DNA sequence is available. While arrayed cDNAs offer a cheap and accessible route to gene expression profiling, they suffer the problem that the arrayed cDNAs are often incompletely sequenced, and instead represented by a 5′ or 3′ expressed sequence tag (EST), which prevents the adoption of sequence-orientated interpretation of the data. For this reason, array platforms developed for non-model organisms that support a large community of interested researchers are expected to gravitate towards the oligonucleotide array format to provide a degree of standardization across laboratories in the community.

Oligonucleotide probes provide additional advantages beyond that of standardization. Most importantly, oligonucleotide probes can be designed to distinguish between genes with high degrees of sequence similarity (Hughes et al., 2001). This is particularly important given the complexity of the transcriptome, which may include transcripts that are alternatively spliced (Zavolan et al., 2003), antisense (Kiyosawa et al., 2003), allele-specific (Yan et al., 2002) or non-coding (Okazaki et al., 2002). The ability to explore the function of these transcripts initially through an understanding of when and where they are expressed will be dependent on the discriminatory power offered by oligonucleotide arrays.

Cross-laboratory and cross-platform standardization is only relevant if gene expression datasets are shared and made available in public databases. Most journals, including *The Journal of Experimental Biology*, require that gene expression data are submitted to one of the two public databases, either ArrayExpress (Parkinson et al., 2007), or NCBI GEO (Barrett et al., 2007). In the past we have found that submitting data to both these repositories was cumbersome and required informatics support in order to organize massive amounts of data into the format required for compliance. Realising that the complexity of the submission process was an obstacle to submission and full disclosure of expression data, the public databases have recently introduced a spreadsheet-based submission format (Rayner et al., 2006). This simple tabular format is similar to the one used by most gene expression analysis software packages, meaning that submission of expression data should be within the capability of any research group with the ability to generate and analyse microarray expression data. Removal of this impediment should streamline the submission and publication of expression data generated for non-model organisms, opening up the field to laboratories with limited informatics capacity. Accordingly, it is hoped that submitting new expression data to public databases will become as routine a step in microarray investigations as preparing high quality RNA.

High-throughput sequencing

A number of approaches have been developed recently that interrogate the transcriptome through very high-throughput sequencing, simultaneously facilitating gene discovery as well as generating a comprehensive view of transcript abundance without prior knowledge of gene sequence. In principle, these approaches should have considerable utility for transcriptome-based investigations into non-model organisms. In ‘massively-parallel signature sequencing’ (MPSS), the most-established of these high-throughput methods (Brenner et al., 2000), hundreds

of thousands of short gene-specific signature sequences are generated from a huge array of cDNAs that are bounds to beads, and the frequency with which each sequence is detected provides a measure of the abundance of each gene in the transcriptome (Hoth et al., 2003; Jongeneel et al., 2003). MPSS is analogous to the established sequencing-based technique of Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995), but MPSS offers much greater sensitivity because it interrogates many more sequences per sample (typically $>1\times10^6$ sequences in MPSS *versus* $<1\times10^5$ in SAGE) in a fraction of the time. This depth of sequencing means that transcripts that are present at extremely low levels can be detected and quantified in theory, if sufficient numbers of sequences are collected. In model organisms, the short signature sequences can then be mapped onto the genomic sequence, but in unsequenced organisms the genes must still be identified on a gene-by-gene basis using PCR primers based on the signature sequences. Despite this shortcoming, MPSS was used to investigate gene expression patterns underlying growth heterosis in oysters whose genome has not been sequenced, revealing that ribosomal proteins are differentially expressed, and suggesting an important role for protein metabolism in heterosis (Hedgecock et al., 2007). Finally, access to most of these technologies is only available through commercial companies, meaning that it may prove financially restrictive to use them for gene expression profiling. Instead, it is anticipated that high-throughput sequencing approaches will fulfill a pivotal role in gene discovery, by providing a benchmark of transcript abundance in selected RNA samples and the identification candidate genes whose expression can then be profiled using more affordable methods.

Gene expression data interpretation

A significant challenge arising from large-scale gene expression profiling experiments is how to interpret the resulting data. Opinions on how a particular dataset should be examined differ widely and there is unlikely to be a single perfect approach to extract all the useful insights from a large dataset. Testament to this is that there are many more publications describing the analysis of expression datasets than there are papers presenting original expression data. So as data analysis methods evolve and previous protocols are superceded, favorite expression datasets are mined repeatedly for more biological insights.

In the course of our investigations into the transcriptional response of carp to environmental cold (Gracey et al., 2004) and hypoxia (Fraser et al., 2006), we have adopted two different computational approaches to extract novel insights into the physiological response of carp to these perturbations. Our choice of methods was driven by the data: cold acclimation caused a substantial proportion of the transcriptome to be differentially expressed and this complexity prompted us to adopt a holistic interpretation method with an emphasis on the detection of biological themes within the expression data. In contrast, the transcriptional response to hypoxia was far more constrained in terms of the number of genes that were expressed, and so we adopted a simple statistical method to identify genes that exhibited large changes in expression as candidates for follow-up studies.

Metabolic reprogramming during cold-acclimation

To generate new insights into the molecular processes that underlie cold-acclimation, we subjected common carp to a stepwise cooling regime from 30°C down to 10°C over 3 days and differential gene expression patterns across seven tissues were determined by hybridization to a custom carp microarray (Gracey et al., 2004). Interpreting the functional role of the complex changes in gene expression that accompany cold-acclimation response is a challenging task, and requires computational methods that transcend from individual genes to interpretations based on biological processes. Recent trends in gene expression analysis have indicated that the interpretation of expression data can be greatly improved if new data, instead of being analyzed in isolation, are interpreted within the context of a larger collection of curated sets of genes. In this approach, sets of functionally related genes are compiled as already characterized ‘genesets’, allowing expression data to be represented in terms of the behavior of these known sets of genes. The most common method of grouping genes is according to the Gene Ontology terms with which they are annotated: Gene Ontology, or GO, is a structured vocabulary that seeks to assign a consistent annotation to genes according to the biological function they perform, the biological process in which they involved, and their cellular or extracellular location (Ashburner et al., 2000). A variety of computational methods are then applied to test whether there are more genes associated with a GO term or biological pathway within a list of differentially expressed genes than would be expected by chance, to yield a heuristic measure of enrichment.

This approach has the ability to detect coherent changes in the expression of sets of genes that may be hidden when considering the expression patterns of individual genes in isolation. For example, analysis of the expression signatures of muscle samples from diabetics identified that oxidative phosphorylation genes were coordinately down-regulated in diabetes, despite the fact that the decrease in their expression was modest, just 20% (Mootha et al., 2003). Since the same ontology is applied universally to genes from all organisms, it allows gene expression datasets from different species to be compared from a thematic rather than from an orthologous gene perspective. For example, shared metabolic programs associated with aging were identified in *Drosophila* and *C. elegans* by identifying over-represented GO terms in the lists of genes that were differentially expressed during the onset of aging in each organism (McCarroll et al., 2004).

For the analysis of the carp cold-acclimation dataset presented here, one (Subramanian et al., 2005) of the many methods (Dennis et al., 2003; Zeeberg et al., 2003; Zhong et al., 2004) for detecting biological themes in sets of differentially expressed genes was applied to the data to

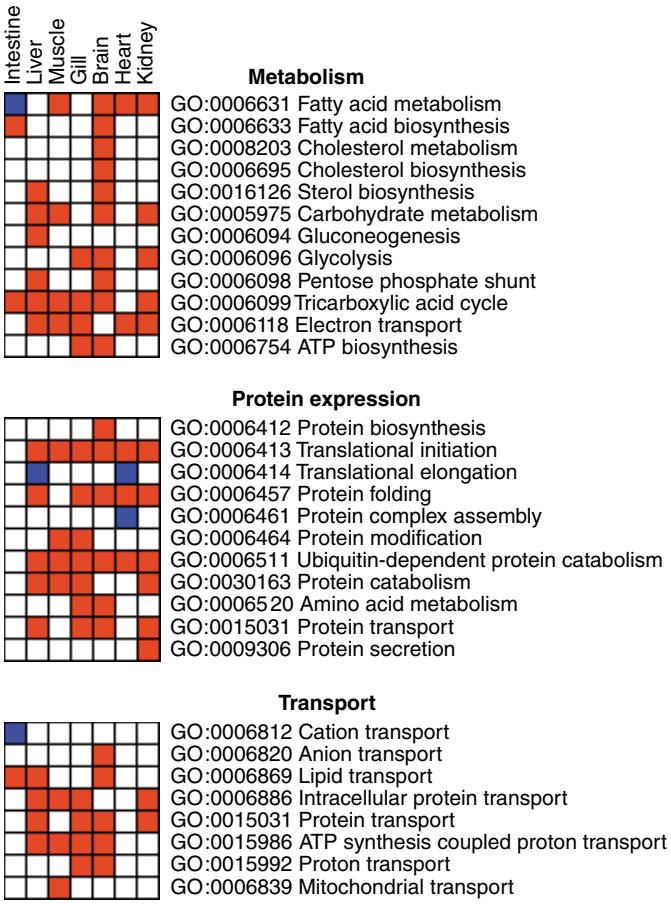

Fig. 1. Gene Ontology based functional classes of genes that show statistically significant (P<0.05) upregulation (red) or downregulation (blue) of gene expression across seven tissues of carp subjected to cooling.

identify GO-defined biological processes that were up- or downregulated with cooling in each tissue. The data are visualized as a heatmap matrix that describes the pattern of changes in metabolism, protein expression and transport in each tissue (Fig. 1). Inspection of the biological processes associated with metabolism reveals that the brain is the most active tissue with respect to the number of metabolic processes that are upregulated during cold-acclimation. The matrix highlights the most distinctive features of each tissue's metabolic adjustments to cold. For example, changes in cholesterol, fatty acid and sterol metabolism were evident in brain tissue, consistent with the importance of lipids for proper function of the central nervous system. Genes involved in the TCA cycle are elevated in six of the seven tissues, consistent with other studies that have shown that mitochondrial enzymes activities are elevated in cold-adapted fish populations (Lucassen et al., 2006). Cold-acclimation is also associated with the differential expression of genes that function in regulating protein expression. The process of translation initiation appears to be upregulated in all tissues with the exception of the intestinal mucosa, consistent with evidence from prokaryotes that suggests that cold compromises the translational apparatus (Thieringer et al., 1998). In general, the intestinal mucosa exhibits a muted response to cold, with fewer biological processes showing transcriptional evidence of acclimation. An explanation for this might be that the mechanisms of cellular acclimation might be different in this tissue since it experiences extremely high cellular turnover.

This thematic analysis suggested a specific role for 'protein folding' genes during cold-acclimation, since this GO term was highly enriched in the list of differentially expressed genes in five of the seven tissues. The specific genes contributing to this enrichment were identified and are listed in Table 1. The list includes all the subunits of the T-complex chaperonin that assist in the folding of proteins upon ATP hydrolysis. Isomerization of proline residues has a large temperature-dependence and can be a rate-limiting step in protein folding at low temperatures (Stoller et al., 1995), and so the inclusion in the gene list of three genes involved in proline *cis–trans* isomerization (FK506-binding proteins) may indicate that this step is a target of compensatory regulation in the cold. Recent analysis has identified two distinct sets of protein chaperone genes in yeast, one with a role in refolding stress-induced damaged proteins and another involved in protein synthesis (Albanese et al., 2006). The yeast orthologues of the T-complex chaperonin, prefoldin, and FK506-binding proteins, all fall into the latter category, suggesting that cold compromises the folding of nascent proteins rather than directly causing protein denaturation. Genes involved in 'ubiquitin-dependent protein catabolism' are also highly enriched in the cold-acclimating tissues, and it remains to be seen if the expression of these genes (mainly components of the proteasome), is a consequence of the production in the cold of aberrantly folded proteins that require degradation.

Identification of novel hypoxia-responsive genes

In an effort to identify candidate genes involved in promoting hypoxia tolerance in fish, we subjected carp that had been acclimated to either 30°C or 17°C to low levels of dissolved oxygen (0.3 mg l^{-1}) and profiled changes in their transcriptome by hybridization to our carp microarray (Fraser et al., 2006). A popular signal-to-noise based statistic was used to rank genes by differential expression between control and hypoxia-treated carp liver (Tusher et al., 2001). We were surprised to observe that microarray spots corresponding to different myoglobin cDNA clones were ranked consistently high as transcripts that were strongly upregulated by hypoxia treatment at both acclimation temperatures (Table 2). Given that myoglobin is not found normally in non-muscle tissues, we went on to use 2D protein gel electrophoresis and mass spectrometry to demonstrate that myoglobin protein was indeed present in liver tissue and that myoglobin protein level increased in hypoxic carp (Fraser et al., 2006). The role of myoglobin in muscle tissues is to facilitate oxygen transport and we speculate that the hypoxic-induction of myogobin in liver may assist in oxygen delivery to this tissue and account

Table 1. *Genes that contribute most significantly to the upregulation of overall 'protein folding' gene expression in cooled carp tissues*

GO:0006457 Protein folding	Function
FK506-binding protein 1A	cis-trans prolyl isomerases
FK506-binding protein 3	
FK506 binding protein 9 precursor	
GrpE protein homolog 1, mitochondrial precursor	Molecular chaperones and facilitators of protein
GrpE protein homolog 2, mitochondrial precursor	folding and membrane translocation
Prefoldin subunit 2	Binds nascent polypeptides and promotes folding
ATP-dependent Clp protease ATP-binding subunit ClpX-like	Protein chaperone
DnaJ homolog subfamily A member 3, mitochondrial precursor	Prevents protein aggregation
Heat shock protein HSP 90-beta	Molecular chaperone
T-complex protein 1, alpha subunit	
T-complex protein 1, beta subunit	Subunits of the chaperonin containing
T-complex protein 1, eta subunit	TCP1 complexthat assists in the folding
T-complex protein 1, gamma subunit	of newly translated proteins in cytosol
T-complex protein 1, zeta subunit	

for the profound hypoxia tolerance exhibited by carp. It is noteworthy that myoglobin protein was detected in the liver of fish living under well-oxygenated conditions, suggesting that myoglobin has a functional role in liver that goes beyond that of its putative role in hypoxia-adaptation. The detection and then hypoxia-induction of myoglobin was a completely unexpected discovery and exemplifies the use of gene expression screens in combination with gene-prioritizing statistical approaches to uncover novel aspects of physiological responses to environmental change.

Applying rank-based statistical methods to prioritize genes for further study has proved useful in other investigations of physiology. For example, gene expression profiling of leptin-regulated genes in the liver of obese *ob*/*ob* mice (Cohen et al., 2002) identified stearoyl-CoA desaturase as the top ranked gene under regulation, leading to follow-up studies that demonstrated that the desaturase has a key role in fat deposition and that mutations in desaturase promote weight loss (Ntambi et al., 2002). Using a custom microarray prepared for Darwin's ground finches, calmodulin was identified as the regulatory gene whose expression was most correlated with peak length across species (Abzhanov et al., 2006). This discovery prompted further studies, first to validate that there was more calmodulin protein expressed in the beaks of finches with elongated beak morphology, and then to demonstrate that manipulating calmodulin expression in the beaks of chickens could modify morphology. These two studies demonstrate the usefulness of using rank-based criteria for prioritizing specific genes identified in gene expression screens for further physiological or evolutionary investigations.

Conclusions

Microarray-based gene expression profiling is just one of a growing number of '-omic' approaches but it has already had a significant impact on investigations into the physiological responses to environmental change. A particular strength of the technique is that it is relatively simple to construct a microarray for any organism, and access to the equipment needed to make and screen arrays is now available at most university core facilities. Studies to date have indicated that properly designed and executed gene expression screens can reveal new insights into physiological responses and can highlight aspects of biological responses that have perhaps been overlooked. Increasingly, gene expression data are being complemented with metabolic, proteomic and molecular interaction data, which together present an unprecedented holistic view of the machinery of the cell across several layers of complexity. The

Table 2. *Ranked list of statistically significant genes induced by hypoxia in carp liver*

Gene rank	Day 5 hypoxia at 30°C	Day 5 hypoxia at 17°C
1	Myoglobin	Hemopexin precursor
2	Myoglobin	Differentially regulated trout protein 1
3	Unclassifiable EST	Vitelline membrane layer protein 1
4	Cofilin	Unclassifiable EST
5	Warm-temperature acclimation protein	FLJ00246 protein
6	Hemopexin precursor	GTP-binding protein hflX
7	Unclassifiable EST	Fatty acid binding protein A
8	TFIIH basal transcription factor	Myoglobin
9	Death effector domain-containing 1	Cofilin
10	Unclassifiable EST	Unclassifiable EST

challenge of integrating these diverse data sources will be met by new computational tools that seek to understand and model biological systems and ultimately to predict how they will respond to environmental and physiological perturbations.

I would like to thank Andrew Cossins and The Laboratory for Environmental Gene Regulation at The University of Liverpool for making this work possible. This work was supported by a grant from Natural Environmental Research Council (UK).

References

Abzhanov, A., Kuo, W. P., Hartmann, C., Grant, B. R., Grant, P. R. and Tabin, C. J. (2006). The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* **442**, 563-567.

Albanese, V., Yam, A. Y., Baughman, J., Parnot, C. and Frydman, J. (2006). Systems analyses reveal two chaperone networks with distinct functions in eukaryotic cells. *Cell* **124**, 75-88.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* **35**, D760-D765.

Bieri, T., Blasiar, D., Ozersky, P., Antoshechkin, I., Bastiani, C., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P. et al. (2007). WormBase: new content and better access. *Nucleic Acids Res.* **35**, D506-D510.

Birrell, G. W., Brown, J. A., Wu, H. I., Giaever, G., Chu, A. M., Davis, R. W. and Brown, J. M. (2002). Transcriptional response of *Saccharomyces cerevisiae* to DNA-damaging agents does not identify the genes that protect against these agents. *Proc. Natl. Acad. Sci. USA* **99**, 8778-8783.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630-634.

Carninci, P. (2007). Constructing the landscape of the mammalian transcriptome. *J. Exp. Biol.* **210**, 1497-1506.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617-1630.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D. et al. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* **13**, 1273-1289.

Cohen, P., Miyazaki, M., Socci, N. D., Hagge-Greenberg, A., Liedtke, W., Soukas, A. A., Sharma, R., Hudgins, L. C., Ntambi, J. M. and Friedman, J. M. (2002). Role for stearoyl-CoA desaturase-1 in leptin-mediated weight loss. *Science* **297**, 240-243.

Crosby, M. A., Goodman, J. L., Strelets, V. B., Zhang, P. and Gelbart, W. M. (2007). FlyBase: genomes by the dozen. *Nucleic Acids Res.* **35**, D486-D491.

Dennis, G., Jr, Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, P3.

Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. et al. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. USA* **93**, 6025-6030.

Drobyshev, A. L., Machka, C., Horsch, M., Seltmann, M., Liebscher, V., Hrabe de Angelis, M. and Beckers, J. (2003). Specificity assessment from fractionation experiments (SAFE): a novel method to evaluate microarray probe specificity based on hybridisation stringencies. *Nucleic Acids Res.* **31**, E1.

Fraser, J., de Mello, L. V., Ward, D., Rees, H. H., Williams, D. R., Fang, Y., Brass, A., Gracey, A. Y. and Cossins, A. R. (2006). Hypoxia-inducible myoglobin expression in nonmuscle tissues. *Proc. Natl. Acad. Sci. USA* **103**, 2977-2981.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B. et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-391.

Gracey, A. Y. and Cossins, A. R. (2003). Application of microarray technology in environmental and comparative physiology. *Annu. Rev. Physiol.* **65**, 231-259.

Gracey, A. Y., Troll, J. V. and Somero, G. N. (2001). Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis*. *Proc. Natl. Acad. Sci. USA* **98**, 1993-1998.

Gracey, A. Y., Fraser, E. J., Li, W., Fang, Y., Taylor, R. R., Rogers, J., Brass, A. and Cossins, A. R. (2004). Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. *Proc. Natl. Acad. Sci. USA* **101**, 16970-16975.

Haugen, A. C., Kelley, R., Collins, J. B., Tucker, C. J., Deng, C., Afshari, C. A., Brown, J. M., Ideker, T. and Van Houten, B. (2004). Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Biol.* **5**, R95.

Hedgecock, D., Lin, J. Z., Decola, S., Haudenschild, C. D., Meyer, E., Manahan, D. T. and Bowen, B. (2007). Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). *Proc. Natl. Acad. Sci. USA* **104**, 2313-2318.

Hoth, S., Ikeda, Y., Morgante, M., Wang, X., Zuo, J., Hanafey, M. K., Gaasterland, T., Tingey, S. V. and Chua, N. H. (2003). Monitoring genome-wide changes in gene expression in response to endogenous cytokinin reveals targets in *Arabidopsis thaliana*. *FEBS Lett.* **554**, 373-380.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R. et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342-347.

Ideker, T., Galitski, T. and Hood, L. (2001a). A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343-372.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. and Hood, L. (2001b). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-934.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M. et al. (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2**, E162.

Jongeneel, C. V., Iseli, C., Stevenson, B. J., Riggins, G. J., Lal, A., Mackay, A., Harris, R. A., O'Hare, M. J., Neville, A. M., Simpson, A. J. et al. (2003). Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc. Natl. Acad. Sci. USA* **100**, 4702-4705.

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. and Hayashizaki, Y. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13**, 1324-1334.

Konno, H., Fukunishi, Y., Shibata, K., Itoh, M., Carninci, P., Sugahara, Y. and Hayashizaki, Y. (2001). Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library. *Genome Res.* **11**, 281-289.

Korkola, J. E., Estep, A. L., Pejavar, S., DeVries, S., Jensen, R. and Waldman, F. M. (2003). Optimizing stringency for expression microarrays. *Biotechniques* **35**, 828-835.

Kuo, W. P., Liu, F., Trimarchi, J., Punzo, C., Lombardi, M., Sarang, J., Whipple, M. E., Maysuria, M., Serikawa, K., Lee, S. Y. et al. (2006). A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.* **24**, 832-840.

Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117-124.

Lucassen, M., Koschnick, N., Eckerle, L. G. and Portner, H. O. (2006). Mitochondrial mechanisms of cold adaptation in cod (*Gadus morhua* L.) populations from different climatic zones. *J. Exp. Biol.* **209**, 2462-2471.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C. S., Jan, Y. N., Kenyon, C., Bargmann, C. I. and Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* **36**, 197-204.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E.

et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267-273.

Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E. et al. (2007). Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.* **35**, D468-D471.

Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L. and Weissman, J. S. (2006). Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840-846.

Ntambi, J. M., Miyazaki, M., Stoehr, J. P., Lan, H., Kendziorski, C. M., Yandell, B. S., Song, Y., Cohen, P., Friedman, J. M. and Attie, A. D. (2002). Loss of stearoyl-CoA desaturase-1 function protects mice against adiposity. *Proc. Natl. Acad. Sci. USA* **99**, 11482-11486.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-573.

Oshlack, A., Emslie, D., Corcoran, L. and Smyth, G. K. (2007). Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biol.* **8**, R2.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M. et al. (2007). ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**, D747-D750.

Podrabsky, J. E. and Somero, G. N. (2004). Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. *J. Exp. Biol.* **207**, 2237-2254.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418-427.

Rayner, T. F., Rocca-Serra, P., Spellman, P. T., Causton, H. C., Farne, A., Holloway, E., Irizarry, R. A., Liu, J., Maier, D. S., Miller, M. et al. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* **7**, 489.

Renn, S. C., Aubin-Horth, N. and Hofmann, H. A. (2004). Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics* **5**, 42.

Sagerstrom, C. G., Sun, B. I. and Sive, H. L. (1997). Subtractive cloning: Past; present; and future. *Annu. Rev. Biochem.* **66**, 751-783.

Stoller, G., Rucknagel, K. P., Nierhaus, K. H., Schmid, F. X., Fischer, G. and Rahfeld, J. U. (1995). A ribosome-associated peptidyl-prolyl cis/trans isomerase identified as the trigger factor. *EMBO J.* **14**, 4939-4948.

Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., Wagner, L., Shenmen, C. M., Schuler, G. D., Altschul, S. F. et al. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA* **99**, 16899-16903.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545-15550.

Tan, P. K., Downey, T. J., Spitznagel, E. L., Jr, Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676-5684.

Thieringer, H. A., Jones, P. G. and Inouye, M. (1998). Cold shock and adaptation. *BioEssays* **20**, 49-57.

Tian, Q., Stepaniants, S. B., Mao, M., Weng, L., Feetham, M. C., Doyle, M. J., Yi, E. C., Dai, H., Thorsson, V., Eng, J. et al. (2004). Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol. Cell. Proteomics* **3**, 960-969.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116-5121.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* **270**, 484-487.

von Schalburg, K. R., Rise, M. L., Cooper, G. A., Brown, G. D., Gibbs, A. R., Nelson, C. C., Davidson, W. S. and Koop, B. F. (2005). Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray. *BMC Genomics* **6**, 126.

Williams, D. R., Epperson, L. E., Li, W., Hughes, M. A., Taylor, R., Rogers, J., Martin, S. L., Cossins, A. R. and Gracey, A. Y. (2005). Seasonally hibernating phenotype assessed through transcript screening. *Physiol. Genomics* **24**, 13-22.

Workman, C. T., Mak, H. C., McCuine, S., Tagne, J. B., Agarwal, M., Ozier, O., Begley, T. J., Samson, L. D. and Ideker, T. (2006). A systems approach to mapping DNA damage response pathways. *Science* **312**, 1054-1059.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. and Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science* **297**, 1143.

Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D. A., Hayashizaki, Y. and Gaasterland, T. (2003). Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* **13**, 1290-1300.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S. et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28.

Zhong, S., Storch, K. F., Lipan, O., Kao, M. C., Weitz, C. J. and Wong, W. H. (2004). GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl. Bioinformatics* **3**, 261-264.

The Journal of Experimental Biology 210, 1492-1496
Published by The Company of Biologists 2007
doi:10.1242/jeb.02783

Glossary of terms

This section is designed to help readers adapt to the complex terminology associated with contemporary molecular genetics, genomics and systems biology. Fuller descriptions of these terms are available at http://www.wikipedia.org/

Ab initio prediction	methods used to predict the potential genes encoded in the genome, which are trained on datasets made of known genes, and used computationally to predict coding regions out of genome without the aid of cDNA sequence. Although their performance is improving, these algorithms perform very poorly on non-protein coding genes.
Annotation	as applied to proteins, DNA sequences or genes. The storage of data describing these entities (protein/gene identities, DNA motifs, gene ontology categorisation, etc.) within a biological database. Active projects include FlyBase and WormBase. See Gene ontology.
Assembly	the process of aligning sequenced fragments of DNA into their correct positions within the chromosome or transcript.
cDNA	complementary DNA. This is DNA synthesised from a mature mRNA template by the enzyme reverse transcriptase. cDNA is frequently used as an early part of gene cloning procedures, since it is more robust and less subject to degradation than the mRNA itself.
ChIP	chromatin immunoprecipitation assay used to determine which segments of genomic DNA are bound to chromatin proteins, mainly including transcription factors.
Chip	see Microarray.
ChIP-on-chip	use of a DNA microarray to analyse the DNA generated from chromatin immunoprecipitation experiments (see ChIP).
cis-acting	a molecule is described as *cis*-acting when it affects other genes that are physically adjacent, on the same chromosome, or are genetically linked or in close proximity (for mRNA expression, typically a promoter).
Collision-induced dissociation	a mechanism by which molecules (e.g. proteins) are fragmented to form molecular ions in the gas phase. These fragments are then analysed within a mass spectrometer to provide mass determination.
Connectivity	a term from graph theory, which indicates the number of connections between nodes or vertices in a network. Greater connectedness between nodes is generally used as a measure of robustness of a network.
CpG islands	regions that show high density of ‘C followed by G’ dinucleotides and are generally associated with promoter elements; in particular, stretches of DNA of at least 200 bp with a C–G content of 50% and an observed CpG/expected CpG in excess of 0.6. The cytosine residues can be methylated, generally to repress transcription, while demethylated CpGs are a hallmark of transcription. CpG dinucleotides are under-represented outside regulatory regions, such as promoters, because methylated C mutates into T by deamination.
Edge	as in networks. Connects two nodes (or vertices) within a system. These concepts arise from graph theory.
Enhancer	a short segment of genomic DNA that may be located remotely and that, on binding particular proteins (*trans*-acting factors), increases the rate of transcription of a specific gene or gene cluster.
Epistasis	a phenomenon when the properties of one gene are modified by one or more genes at other loci. Otherwise known as a genetic interaction, but epistasis refers to the statistical properties of the phenomenon.

eQTL	the combination of conventional QTL analysis with gene expression profiling, typically using microarrays. eQTLs describe regulatory elements controlling the expression of genes involved in specific traits.
EST	expressed sequence tag. A short DNA sequence determined for a cloned cDNA representing portions of an expressed gene. The sequence is generally several hundred base pairs from one or both ends of the cloned insert.
Exaptation	a biological adaptation where the current function is not that which was originally evolved. Thus, the defining (derived) function might replace or persist with the earlier, evolved adaptation.
Exon	any region of DNA that is transcribed to the final (spliced) mRNA molecule. Exons interleave with segments of non-coding DNA (introns) that are removed (spliced out) during processing after transcription.
Gene forests	genomic regions for which RNA transcripts, produced from either DNA strand, have been identified without gaps (non-transcribed genomic regions). Conversely, regions in which no transcripts have ever been detected are called ‘gene deserts’.
Gene interaction network	a network of functional interactions between genes. Functional interactions can be inferred from many different data types, including protein–protein interactions, genetic interactions, co-expression relationships, the co-inheritance of genes across genomes and the arrangement of genes in bacterial genomes. The interactions can be represented using network diagrams, with lines connecting the interacting elements, and can be modelled using differential equations.
Gene ontology (GO)	an ontology is a controlled vocabulary of terms that have logical relationships with each other and that are amenable to computerised manipulation. The Gene Ontology project has devised terms in three domains: biological process, molecular function and cell compartment. Each gene or DNA sequence can be associated with these annotation terms from each domain, and this enables analysis of microarray data on groups of genes based on descriptive terms so provided. See http://www.geneontology.org
Gene set enrichment analysis	a computational method that determines whether a defined set of genes, usually based on their common involvement in a biological process, shows statistically significant differences in transcript expression between two biological states.
Gene silencing	the switching-off of a gene by an epigenetic mechanism at the transcriptional or post-transcriptional levels. Includes the mechanism of RNAi.
Genetic interaction (network)	a genetic interaction between two genes occurs when the phenotypic consequences of a mutation in one gene are modified by the mutational status at a second locus. Genetic interactions can be aggravating (enhancing) or alleviating (suppressing). To date, most high-throughput studies have focussed on systematically identifying synthetic lethal or sick (aggravating) interactions, which can then be visualised as a network of functional interactions (edges) between genes (nodes).
Genome	a portmanteau of gene and chromosome, the entire hereditary information for an organism that is embedded in the DNA (or, for some viruses, in RNA). Includes protein-coding and non-coding sequences.
Heritability	phenotypic variation within a population is attributable to the genetic variation between individuals and to environmental factors. Heritability is the proportion due to genetic variation usually expressed as a percentage.
Heterologous hybridization	the use of a cDNA or oligonucleotide microarray of probes designed for one species with target cRNA/cDNAs from a different species.
Homeotic	the transformation of one body part to another due to mutation of specific developmentally related genes, notably the *Hox* genes in animals and *MADS-box* genes in plants.
Hub	as in networks. A node with high connectivity, and thus which interacts with many other nodes in the network. A hub protein interacts with many other proteins in a cell.

Hybridisation	the process of joining (annealing) two complementary single-stranded DNAs into a single double-stranded molecule. In microarray analysis, the target RNA/DNA from the subject under investigation is denatured and hybridised to probes that are immobilised on a solid phase (i.e. glass microscope slide).
Hypomorph	in genetics, a loss-of-function mutation in a gene, but which shows only a partial reduction in the activity it influences rather than a complete loss (cf. hypermorph, antimorph, neomorph, etc).
Imprinting	a phenomenon where two inherited copies of a gene are regulated in opposite ways, one being expressed and the other being repressed.
Indel	insertion and deletion of DNA, referring to two types of genetic mutation. To be distinguished from a 'point mutation', which refers to the substitution of a single base.
Interactome	a more or less comprehensive set of interactions between elements within cells. Usually applied to genes or proteins as defined by transcriptomic, proteomic or protein–protein interaction data.
Intron	see Exon.
KEGG	The Kyoto Encyclopedia of Genes and Genomes is a database of metabolic and other pathways collected from a variety of organisms. See http://www.genome.jp/kegg
Metabolomics	the systematic qualitative and quantitative analysis of small chemical metabolite profiles. The metabolome represents the collection of metabolites within a biological sample.
Metagenomics	the application of genomic techniques to characterise complex communities of microbial organisms obtained directly from environmental samples. Typically, genomic tags are sequence characterised as markers of each species to inform on the range and abundance of species in the community.
Microarray	an arrayed set of probes for detecting molecularly specific analytes or targets. Typically, the probes are composed of DNA segments that are immobilised onto the solid surface, each of which can hybridise with a specific DNA present in the target preparation. DNA microarrays are used for profiling of gene transcripts.
Model species	a species used to study particular biological phenomena, the outcome offering insights into the workings of other species. Usually, the selection is based on experimental tractability, particularly ease of genetic manipulation. For the geneticist, it is an organism with inbred lines where sibs will be >98% identical (i.e. *Drosophila*, *Caenorhabditis elegans* and mice). For genomic science, it refers to a species for which the genomic DNA has been sequenced.
miRNA	a category of novel, very short, non-coding RNAs, generated by the cleavage of larger precursors (pri-miRNA). These short RNAs are included in the RNA-induced silencing complex (RISC) and pair to the 3′ ends of target RNA, blocking its translation into proteins (in animals) or promoting RNA cleavage and degradation (in plants).
mRNA	a protein-coding mRNA containing a protein-coding region (CDS), preceded by a 5′ and followed by a 3′ untranslated region (5′ UTR and 3′ UTR). The UTRs contain regulatory elements. A full-length cDNA contains the complete sequence of the original mRNA, including both UTRs. However, it is often difficult to assign the starting–termination positions for protein synthesis unambiguously. A cDNA containing the entire CDS is often considered acceptable for bioinformatic and experimental studies requiring full-length cDNAs.
ncRNA	non-coding RNA is any RNA molecule with no obvious protein-coding potential for at least 80 or 100 amino acids, as determined by scanning full-length cDNA sequences. It includes ribosomal (rRNA) and transfer RNAs (tRNA) and is now known to include various sub-classes of RNA, including snoRNA, siRNA and piRNA. Just like the coding mRNAs, a large proportion of ncRNAs are transcribed by RNA polymerase II and are large transcripts. A description of the many forms of ncRNA can be found at http://en.wikipedia.org/wiki/Non-coding_RNA.

Node	as in networks. Objects linked by edges to create a network.
PCR	polymerase chain reaction. A molecular biology technique for replicating DNA *in vitro*. The DNA is thus amplified, sometimes from very small amounts. PCR can be adapted to perform a wide variety of genetic manipulations.
piRNA	Piwi-interacting RNA. A class of RNA molecules (29–30 nt long) that complex with Piwi proteins (a class of the Argonaute family of proteins) and are involved in transcriptional gene silencing.
PMF	peptide mass fingerprinting. An analytical technique for protein identification in which a protein is fragmented using proteases. The resulting peptides are analysed by mass spectrometry and these masses compared against a database of predicted or measured masses to generate a protein identity.
Polyadenylation	the covalent addition of multiple A bases to the 3′ tail of an mRNA molecule. This occurs during the processing of transcripts to form the mature, spliced molecule and is important for regulation of turnover, trafficking and translation.
Post-source decay	in mass spectrometry. The fragmentation of precursor molecular ions as they accelerate away from the ionisation source of the mass spectrometer. All precursor ions leaving the ion source have approximately the same kinetic energy, but fragmentation results in smaller product ions that can be distinguished from precursor ions using a ‘reflectron’ by virtue of their lower kinetic energies.
Post-translational modification	the chemical modification of a protein after synthesis through translation. Some modifications, notably phosphorylation, affect the properties of the protein, offering a means of regulating function.
Principal component analysis (PCA)	a technique for simplifying complex, multi-dimensional datasets to a reduced number of dimensions, the principal components. This procedure retains those characteristics of the data that relate to its variance.
Promoter	a regulatory DNA sequence, generally lying upstream of an expressed gene, which in concert with other often distant regulatory elements directs the transcription of a given gene.
Proteome	the entire protein complement of an organism, tissue or cell culture at a given time.
Quantitative trait	inheritance of a phenotypic property or characteristic that varies continuously between extreme states and can be attributed to interactions between multiple genes and their environment.
qPCR	quantitative real-time PCR, sometimes called real-time PCR. A more quantitative form of RT-PCR in which the quantity of amplified product is estimated after each round of amplification.
QTL	quantitative trait loci. A region of DNA that contains those genes contributing to the trait under study.
RISC	RNA-induced silencing complex. A protein complex that mediates the double-stranded RNA-induced destruction of homologous mRNA.
RNAi	RNA interference or RNA-mediated interference. The process by which double-stranded RNA triggers the destruction of homologous mRNA in eukaryotic cells by the RISC.
RT-PCR	reverse transcription–polymerase chain reaction. A technique for amplifying a defined piece of RNA that has been converted to its complementary DNA form by the enzyme reverse transcriptase. See qPCR.
siRNA	small interfering RNA, or silencing RNA. A class of short (20–25 nt), double-stranded RNA molecules. It is involved in the RNA interference pathway, which alters RNA stability and thus affects RNA concentration and thereby suppresses the normal expression of specific genes. Widely used in biomedical research to ablate specific genes.

snoRNA	small nucleolar RNA. A sub-class of RNA molecules involved in guiding chemical modification of ribosomal RNA and other RNA genes as part of the regulation of gene expression.
SNP	single nucleotide polymorphism. A single base-pair mutation at a specific locus, usually consisting of two alleles. Because SNPs are conserved over evolution, they are frequently used in QTL analysis and in association studies in place of microsatellites, and in genetic fingerprinting analyses.
SSH	suppressive subtractive hybridisation. A powerful protocol for enriching cDNA libraries for genes that differ in representation between two or more conditions. It combines normalisation and subtraction in a single procedure and allows the detection of low-abundance, differentially expressed transcripts, such as those involved in signalling and signal transduction.
Structural RNAs	a class of non-coding RNA, long known to have a structural role (for instance, the ribosomal RNAs), transcribed by RNA polymerase I or III.
Systems biology	treatment of biological entities as systems composed of defined elements interacting in defined ways to enable the observed function and behaviour of that system. The properties of the systems are embedded in a quantitative model that guides further tests of systems behaviour.
TATA-boxes	sequences in promoter regions constituted by TATAAA, or similar variants, which were considered the hallmark of Promoters. Recent data show that they are present only in the minority of promoters, where they direct transcription at a single well-defined location some 30 bp downstream of this element.
trans-acting	a factor or gene that acts on another unlinked gene, a gene on a separate chromosome or genetically unlinked usually through some diffusible protein product (for mRNA expression, typically a transcription factor).
Transcript	an RNA product produced by the action of RNA polymerase reading the sequence of bases in the genomic DNA. Originally limited to protein-coding sequences with flanking UTRs but now known to include large numbers of products that do not code for a protein product.
Transcriptome	the full set of mRNA molecules (transcripts) produced by the system under observation. Whilst the genome is fixed for a given organism, the transcriptome varies with context (i.e. tissue source, ontogeny, external conditions or experimental treatment).
Transgene	a gene or genetic material that has been transferred between species or between organisms using one of several genetic engineering techniques.
Transinduction	generation of transcripts from intergenic regions. At least some such products do not relate to a definable promoter or transcriptional start site.
Transposon	sequences of DNA able to move to new positions within the genome of a single cell. This event might cause mutation at the site of insertion. Also called ‘mobile genetic elements’ or ‘jumping genes’.
Transvection	an epigenetic phenomenon arising from the interaction between one allele and the corresponding allele on the homologous chromosome, leading to gene regulation.
TUs	transcriptional units. Used to group all of the overlapping RNA transcripts that are transcribed from the same genomic strand and share exonic sequences.
UTR	untranslated region. Regions of the mRNA that lie at either the 3′ or 5′ flanking ends of the molecule (i.e. 3′ UTR and 5′ UTR). They bracket the protein-coding region and contain signals and binding sites that are important for the regulation of both protein translation and RNA degradation.