The Journal of Experimental Biology 210, 1593-1601
Published by The Company of Biologists 2007
doi:10.1242/jeb.000141

Functional genomics and proteomics of the cellular osmotic stress response in 'non-model' organisms

Dietmar Kültz[1,*], Diego Fiol[1], Nelly Valkova[1], Silvia Gomez-Jimenez[2], Stephanie Y. Chan[1] and Jinoo Lee[1]

[1]*Physiological Genomics Group, Department of Animal Science, One Shields Avenue, University of California, Davis, CA 95616, USA and* [2]*Laboratorio de Fisiología de Invertebrados Marinos, CIAD, A.C. Carr. a la Victoria Km. 0.6, CP 83000, Hermosillo, Sonora, México*

*Author for correspondence (e-mail: dkueltz@ucdavis.edu)

Accepted 3 January 2007

Summary

All organisms are adapted to well-defined extracellular salinity ranges. Osmoregulatory mechanisms spanning all levels of biological organization, from molecules to behavior, are central to salinity adaptation. Functional genomics and proteomics approaches represent powerful tools for gaining insight into the molecular basis of salinity adaptation and euryhalinity in animals. In this review, we discuss our experience in applying such tools to so-called 'non-model' species, including euryhaline animals that are well-suited for studies of salinity adaptation. Suppression subtractive hybridization, RACE-PCR and mass spectrometry-driven proteomics can be used to identify genes and proteins involved in salinity adaptation or other environmental stress responses in tilapia, sharks and sponges. For protein identification in non-model species, algorithms based on sequence homology searches such as MSBLASTP2 are most powerful. Subsequent gene ontology and pathway analysis can then utilize sets of identified genes and proteins for modeling molecular mechanisms of environmental adaptation. Current limitations for proteomics in non-model species can be overcome by improving sequence coverage, N- and C-terminal sequencing and analysis of intact proteins. Dependence on information about biochemical pathways and gene ontology databases for model species represents a more severe barrier for work with non-model species. To minimize such dependence, focusing on a single biological process (rather than attempting to describe the system as a whole) is key when applying 'omics' approaches to non-model organisms.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1593/DC1

Key words: salinity adaptation, osmotic stress, systems biology, euryhaline fish, proteomics.

Systems biology approaches in traditional comparative biology

Major advances in high-throughput technologies for the detection and quantification of nucleic acids, proteins and metabolites have led to a paradigm shift in biological research. The new field of systems biology attempts to integrate the complex data sets generated by high-throughput approaches to develop a holistic understanding of complex biological structures, their dynamics and their responsiveness to external stimuli such as salinity stress. The level of complexity of structures modeled by systems biology ranges from molecular networks *via* cells and organs to whole organisms (Kitano, 2002). Systems biology organizes data obtained by genomic, transcriptomic, proteomic and metabolomic approaches to attempt to build descriptive and mechanistic models of integrative biological phenomena such as development or interactions of organisms with their environment.

The ultimate goal of this approach is to generate a mathematical model that describes the biological system and has predictive power (Aggarwal and Lee, 2003). Achieving this tremendously ambitious goal depends on in-depth knowledge about each element constituting the system of interest. For instance, experimental data about the expression, regulation, function, compartmentation, interaction, modification and stability of individual RNAs and proteins have to be collected and integrated for each state of the system that is described by the model. Systems biology approaches have largely been applied to organisms whose genomes have been sequenced (so-called 'model' organisms) because many of the available bioinformatics tools are based on prior genome sequencing and annotation of the encoded transcriptome and proteome. However, most high-throughput technologies for analyzing the transcriptome, proteome and metabolome do not strictly

depend on prior knowledge of genomic sequence. Therefore, these approaches can also be applied to 'non-model' organisms for which few if any genome sequence data are available. Such organisms account for the majority of species used in traditional comparative biology.

We utilized high-throughput transcriptomics and proteomics approaches for identifying key molecular constituents associated with osmoregulatory functions in euryhaline tilapia and other animals. Our work shows that these data can be used for bioinformatics analysis to generate models about biological processes, cellular pathways and molecular functions associated with osmotic stress responses. Nevertheless, many obstacles are encountered when attempting a systems biology approach with non-model species. In the following, we briefly review our efforts to apply high-throughput transcriptomics and proteomics methods to euryhaline tilapia, dogfish shark and an intertidal sponge to pave the way towards a systems biology approach for studying osmoregulation.

Identification of an immediate-early gene network involved in osmoregulation

Approaches for studying environmental regulation of the transcriptome

Organisms have osmoregulatory mechanisms that span all levels of biological organization from molecules to behavior. Therefore, understanding responses of fishes and other animals to salinity stress requires knowledge of the biological system as a whole, starting with its molecular components. A very popular approach for gaining insight into genes that are regulated by environmental stimuli is based on the use of DNA microarray chips. However, to harness the full power of this approach, sequence knowledge of the entire transcriptome of the species of interest is required. Although DNA microarrays can also be used very effectively for non-model species (Buckley, 2007; Gracey, 2007), prior sequence knowledge greatly facilitates their production and use and maximizes coverage of the transcriptome during the analysis. Serial analysis of gene expression (SAGE) and recent modifications of SAGE represent alternative approaches to DNA microarrays that identify very short sequence tags that can then be matched against existing sequences (Maillard et al., 2005). In addition to SAGE and DNA microarrays, the identification of a subset of RNAs whose levels are regulated by salinity can be achieved by methods that are based on differential analysis of RNA levels and less dependent on prior sequence knowledge of the transcriptome. Such methods include differential display (Stein and Liang, 2002), representational difference analysis (Hubank and Schatz, 1999) and suppression subtractive hybridization (SSH) (Diatchenko et al., 1999).

Suppression subtractive hybridization of salinity-responsive tilapia genes

Because SSH works optimally even in the absence of prior sequence knowledge, we have used this technique for enriching a subset of genes that are rapidly upregulated by salinity stress in gill epithelial cells of tilapia. Tilapia are strongly euryhaline teleosts that are capable of adaptation to a wide range of environmental salinity. Therefore, mechanisms and molecules involved in salinity adaptation are very prominent in this fish and can be effectively studied.

SSH was performed with mRNA from control fish transferred for 4 h from freshwater (FW) to FW (as the Driver sample) and the corresponding mRNA from fish transferred for 4 h from FW to seawater (SW) (as the Tester sample). The final resulting PCR products were cloned into pGEM-Teasy vector (Promega, Madison, WI, USA) to generate a subtracted cDNA library. This library was screened by colony PCR to yield 650 individual clones. Only 72 clones were selected for sequencing because many of the 650 clones had similar lengths and likely represented the same sequence. Thirty-two unique cDNAs were represented by the 72 clones that were sequenced. Increased mRNA abundance after 4 h SW transfer was confirmed for 22 of these 32 cDNAs by quantitative real-time PCR (qPCR). This result demonstrates that the rate of false-positive identification by SSH is low (less than one-third of the analyzed clones were false positives). Full-length coding sequences for most of the 22 cDNAs were obtained by RACE-PCR and degenerate primer PCR, starting with original SSH clones, many of which represented 3′ terminal fragments of cDNAs (Fiol and Kültz, 2005; Fiol et al., 2006a). Based on amino acid homology in the coding sequences, we were able to identify more than 90% of the SSH genes cloned from tilapia (Table 1).

Induction kinetics suggests that tilapia SSH cDNAs are immediate-early genes

Hyperosmotic induction of tilapia SSH genes was confirmed and the kinetics of their induction analyzed by real-time qPCR. Most SSH genes show a rapid and transient increase in mRNA abundance, with peak levels observed between 2 and 8 h after SW transfer. An example of a time course depicting the salinity-induced induction of a tilapia SSH gene (protein phosphatase 2A, catalytic subunit) is shown in Fig. 1. Most of the SSH genes differ only slightly in the kinetics of their induction, suggesting that they participate in a common cellular network as part of an overall essential and coordinated adaptive response that involves changes in metabolism, transport, damage repair, etc. (Fiol and Kültz, 2005; Fiol et al., 2006a). A notable single exception is a cDNA clone (SSH#7) that displays a robust sustained induction in response to SW transfer. This cDNA increases as rapidly as the other SSH genes but it remains high even after a prolonged stay of tilapia in SW, and the degree of its increase is much greater than for any other SSH gene, exceeding five orders of magnitude (Fiol et al., 2006a). Unfortunately, we were not able to match the sequence of the SSH#7 cDNA to any known sequence. The lack of an open reading frame (ORF), even after extensive attempts to extend the sequence by RACE-PCR, may suggest that SSH#7 represents a non-protein-coding RNA (ncRNA). Because ncRNAs are involved in transcriptional and translational regulation, modulation of protein function, and regulation of RNA and protein localization, SSH#7 may have an

Table 1. *cDNA clones that have been identified by suppression subtractive hybridization (SSH) as immediate early genes involved in hyperosmotic stress signaling in tilapia gill cells*

SSH clone	Symbol	Protein name	GenBank no.
SSH#2	OSTF1	Osmotic stress transcription factor 1	AAT84345
SSH#10	TFIIB	Basal transcription factor IIB	AAT84346
SSH#40	BMPR	Bone morphogenetic protein receptor type II	DQ466076
SSH#81	PPP2CB	Protein phsophatase-2A catalytic subunit	DQ465382
SSH#82	MIG-6	Mitogen-inducible gene 6 protein	DQ465383
SSH#888	MAP3K7IP2	Mitogen-activated protein kinase kinase kinase 7 interacting protein 2/TAK 1 binding protein 2/TAB 2	DQ465390
SSH#4	PAD2	Protein arginine deiminase type II	DQ465373
SSH#30	RFP128	Ring finger protein 128/Goliath/ubiquitin E3 ligase	DQ465380
SSH#848	ELAV	Embryonic Lethal, Abnormal Vision/Hu antigen R	DQ465389
SSH#5	ENO1	Enolase 1 (alpha)	DQ465374
SSH#509	IDH2	Isocitrate dehydrogenase 2 ($NADP^+$), mitochondrial	DQ465384
SSH#9	FABP6	Fatty acid-binding protein 6/gastrotropin	DQ465392
SSH#58	APOD	Apolipoprotein D	DQ465391
SSH#8	GSN	Gelsolin	DQ465376
SSH#602	LGALS4	Galectin-4/ L-36 lactose binding protein	DQ465385
SSH#740	ANXA11	Annexin A11b	DQ465387
SSH#16	SA1	Amino acid system A transporter	DQ465378
SSH#160	INO1	Myo-inositol-1 phosphate synthase	DQ465381
SSH#7		Unknown	DQ465375
SSH#29	CHST6	Carbohydrate (*N*-acetylglucosamine 6-O) sulfotransferase 6	DQ465379
SSH#718	Mit16S	Mitochondrial 16S RNA	DQ465386
SSH#828		Unknown	DQ465388

important role for salinity adaptation even if it is an ncRNA (Goodrich and Kugel, 2006). For one of the SSH genes that encodes a protein [osmotic stress transcription factor 1 (OSTF1)], we have already performed a follow-up study to analyze the mechanisms of its regulation in more depth. The results of this study show that hypertonicity *per se*, rather than indirect systemic factors, is responsible for the salinity-induced OSTF1 increase, that the mechanism of increase is based on transient mRNA stabilization and that a return to normal OSTF1 levels at later time points is promoted by systemic factors (Fiol et al., 2006b).

The rapid and transient induction kinetics of most SSH cDNAs during hyperosmotic stress is characteristic of immediate-early genes (IEGs). IEGs were originally identified in neuronal tissue and encode many inducible transcription factors but also other proteins such as scaffolding proteins, signaling proteins and proteases (Lanahan and Worley, 1998). IEGs are key regulators during environmental adaptation because they control the transcriptional regulation of effector genes (delayed response genes) necessary for adaptation. IEGs represent the first set of genes that are induced in response to environmental signals and encode mRNAs that do not require *de novo* protein synthesis for their induction (Morgan and Curran, 1989). Induction of many IEGs is often mediated by mitogen-activated protein kinases (MAPKs) (Thomson et al., 1999), which we have shown to be activated by salinity stress in euryhaline teleosts (Kültz and Avila, 2001). Moreover, IEGs function within a network of other, constitutively expressed, proteins to control cellular regulation in response to environmental cues (Lanahan and Worley, 1998). Thus, we utilized the gene set identified by SSH for a systems biology approach that is based on bioinformatics tools to gain insight into the regulatory network that underlies salinity adaptation in tilapia.

Gene ontology and pathway analysis using the tilapia SSH gene set

To gain insight into the nature of regulatory networks that are activated early in response to salinity stress in tilapia gill

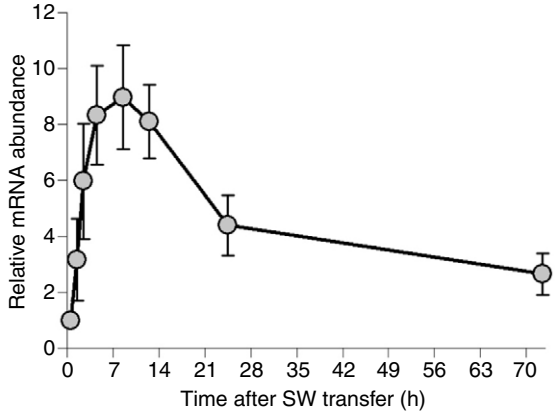

Fig. 1. Example of a time course of mRNA expression of SSH clones in response to salinity stress (SSH#81, protein phsophatase-2A catalytic subunit, is shown). mRNA expression was analyzed at the indicated times after transfer of tilapia from freshwater to seawater (SW). Expression levels were quantified by qPCR in gill epithelial cells. Values are means ± s.e.m. (*N*=4).

cells, we used bioinformatics tools that extract and synthesize most available information about any given gene. The SSH gene set was compared to templates for known biochemical pathways, molecular functions and biological processes. Using Pathway Studio software (Ariadne Genomics, Inc., Rockville, MD, USA), we distilled all published information about biological relationships of mammalian orthologs of the salinity-induced tilapia genes and modeled the interaction of the SSH genes within a common stress response signal transduction network (Fiol et al., 2006a). This network contains 13 of the 22 identified tilapia SSH genes (*OSTF1*, *TFIIB*, *BMPR2*, *PPP2CB*, *MIG-6*, *MAP3K7IP2*, *ENO1*, *IDH2*, *FABP6*, *APOD*, *GSN*, *LGALS4*, *ANXA11*) (see Table 1). Furthermore, a different type of bioinformatics approach, PANTHER (Mi et al., 2005), revealed that mammalian homologs of 11 of the 22 identified tilapia SSH genes participate in apoptotic and cell cycle regulatory pathways (*OSTF1*, *TFIIB*, *BMPR2*, *PPP2CB*, *MIG-6*, *MAP3K7IP2*, *RFP128*, *PAD2*, *GSN*, *LGALS4*, *ANXA11*) (see Table 1). Apoptosis and cell cycle regulation are key processes of the cellular stress response, indicating that these biological processes are targeted during osmotic stress in tilapia gills (Kültz, 2005). In addition to apoptosis and cell cycle signaling, cellular functions associated with the 22 tilapia SSH genes are organic osmolyte accumulation, energy metabolism, lipid transport and membrane protection, modulation of actin-based cytoskeleton dynamics, and control of mRNA and protein stability (Fiol et al., 2006a). Regulation of these processes reflects the need for extensive remodeling and rapid cellular turnover in tilapia gill epithelium during salinity stress.

Although the pathway analysis described above yielded remarkably consistent and meaningful results, some potential pitfalls need to be considered. First, comparisons are made based on data that are heavily biased towards a few model organisms such as human, mouse and rat. However, the involvement of a particular protein in cellular functions, molecular pathways and biological processes in non-model species may differ from that in model species. This first aspect represents a significant disadvantage for work with non-model species. Second, when working with gene sets that are based on altered levels of mRNA abundance, corresponding changes in protein abundance are often assumed but frequently not observed (Hack, 2004). Thus, further analyses of proteins encoded by genes that have been identified or independent proteome analysis are logical next steps.

Identification of protein networks involved in osmoregulation

Targeted confirmation of genes induced by salinity stress at the protein level

We have started confirming the upregulation of the immediate-early osmotic stress response genes identified by the SSH approach. The most practical approach for achieving this task is by using antibodies that specifically recognize the proteins of interest. Antibodies are much more readily available for model species, and currently projects are underway to generate specific antibodies for every human and every mouse protein. Therefore, this approach is less work-intensive for model species. We have used two strategies for antibody-based quantification of salinity-induced tilapia proteins by western immunodetection. First, we have tested available antibodies that were made against mammalian TFIIB for cross-reactivity and specificity with tilapia TFIIB. After successful identification of a suitable antibody that is cross-species reactive and specific (detection of a single band of expected molecular mass on western blots), we used it to confirm that tilapia TFIIB is also salinity-induced at the protein level (Fiol and Kültz, 2005). Second, we have generated a new antibody against tilapia OSTF1 based on immunization of rabbits with an immunogenic peptide. This antibody was used to confirm the induction of OSTF1 protein in gill cells of tilapia after transfer from FW to SW (Fiol and Kültz, 2005). Currently, we are working on generating additional antibodies for more tilapia SSH genes. However, this approach is work-intensive, expensive and time-consuming and, therefore, not well-suited for the high-throughput data generation needed for systems biology approaches that utilize unbiased discovery-driven screens. Nevertheless, such a more traditional approach will greatly advance our knowledge of osmotic stress signaling in tilapia because it is firmly focused on prior knowledge of a limited set of elements involved in a specific biological process.

Discovery-driven proteomics using osmoregulatory tissues

An alternative approach for protein analysis that is much faster, less time-consuming and more economical is based on comparing protein patterns of two different samples. This can be achieved by separating proteins *via* two-dimensional gel electrophoresis (2DGE) or two-dimensional liquid chromatography (2DLC) followed by subsequent mass spectrometry (MS) analysis. Both techniques (2DGE and 2DLC) are capable of resolving in the order of 1500 proteins and they are complementary, with different advantages and drawbacks (Wagner et al., 2000; Ishihama, 2005; Hu et al., 2005; Bunai and Yamane, 2005; Okano et al., 2006). A common limitation of both techniques is their bias towards high-abundance proteins, which is a result of the huge dynamic range of protein abundance in biological samples. This limitation can be addressed by extensive sample pre-fractionation, but such an approach requires a lot of starting material and greatly increases the cost, time and work load for the analysis. High-throughput proteomics using protein chips represents an alternative approach that is capable of detecting low-abundance proteins. However, protein arrays are expensive and depend on very high-quality antibodies that are often not available for a large number of proteins from non-model organisms.

Keeping these limitations in mind, we have compared the proteomes of four osmoregulatory tissues (rectal gland, kidney, gill epithelium, intestinal epithelium) with heart and brain from dogfish shark (*Squalus acanthias*) by 2DGE and MS (Lee et

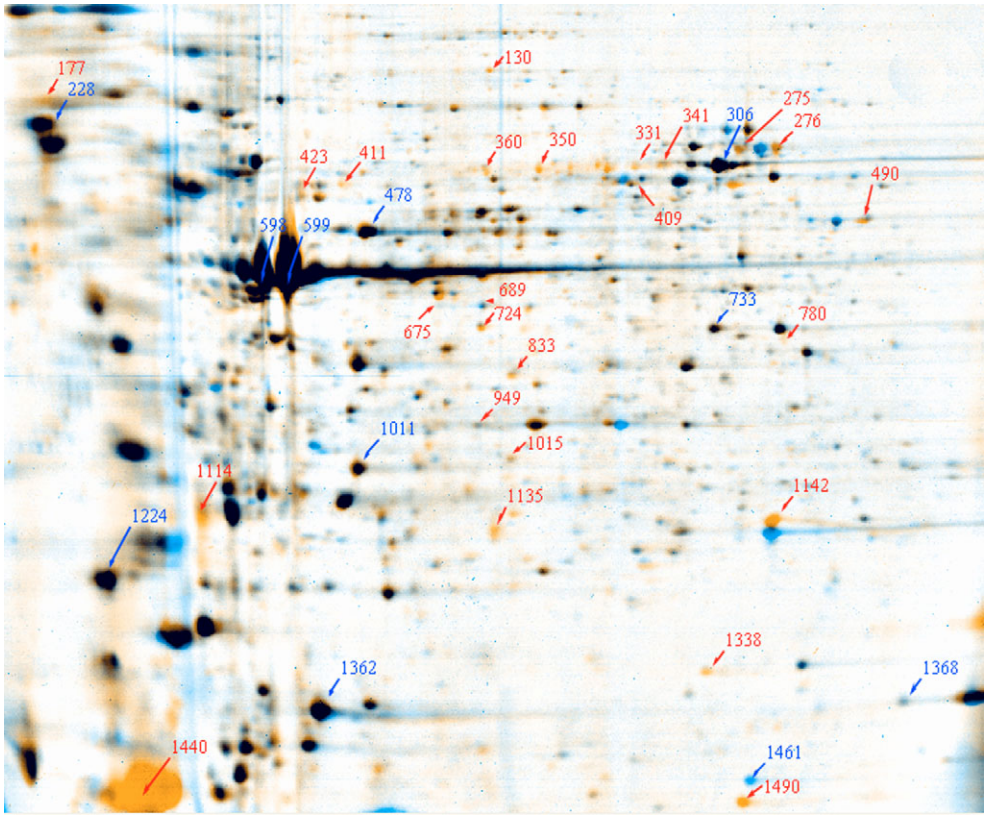

Fig. 2. *Tetilla mutabilis* proteome map, with proteins that were picked for mass spectrometric analysis labeled. Red labels denote proteins that are upregulated during emersion or post-emersion stress, and blue labels denote proteins that are unchanged by emersion stress. The vertical dimension of the 2-D gel represents the molecular mass scale (10–120 kDa on 11% uniform polyacrylamide gels), and the horizontal dimension represents the isoelectric point scale (pH 3–10 on non-linear 24 cm IPG strips).

al., 2006). The rationale of this proteomics experiment was to determine whether highly abundant proteins that are common to osmoregulatory tissues can be identified and whether they are indicative of specific signaling pathways and biological functions in osmoregulatory tissues. The number of consistently resolved protein spots ranged from 984 in brain to 1230 in intestine, and overall 1465 unique protein spots were detected in the six tissues. Thirty-six percent of these protein spots (535 spots) were present at comparable abundance in all tissues. Two hundred and seventy protein spots that were significantly over-represented in one specific tissue or common to osmoregulatory tissues were identified using the Delta2D software package (Decodon GmBH, Greifswald, Germany). To identify such proteins we applied a robust and accurate method of protein quantitation that is based on 10 internal standard spots that are used for normalization of gel-to-gel loading and staining differences (Valkova and Kültz, 2006; Lee et al., 2006). Gel-based protein quantitation procedures are of course not without pitfalls but in our experience this method is more accurate than isotope-labeling or other approaches that rely on MS quantitation. The 270 identified protein spots were cut out from the gel, in-gel digested with trypsin, and analyzed by MS to determine their identity (Lee et al., 2006).

Identification of shark proteins by MS

Of the 270 selected shark proteins, we were able to identify 62 proteins (=23%) with high confidence (Lee et al., 2006). Fifty-four of these 62 proteins were correctly identified using the MS-BLASTP2 algorithm (=87%), compared to 43 that were identified with the Mascot search engine (=69%) at the same mass accuracy limit of 50 p.p.m. High-quality spectra and *de novo* sequence information were obtained for many of the 270 analyzed protein spots, but a large portion was not identified by Mascot or MSBLASTP2, probably because of significant amino acid sequence variation in the shark proteins compared to their orthologs in model species (Lee et al., 2006). The genome of dogfish sharks has not been sequenced and few cDNA/protein sequences of this species are available. Therefore, identification of shark proteins by mass spectrometry poses unique challenges that are not encountered when working on model species. In particular, approaches based on peptide mass fingerprinting (Mascot, etc.), which generally work very well for model organisms (exceptions are proteins that are extensively post-translationally modified), are less successful in non-model species. Even if homologous peptide sequences differ by only a single amino acid, identification of the corresponding protein is prevented because peptide masses will not match. Nevertheless, fish proteins with highly homologous sequence stretches can still be identified using peptide mass fingerprinting (Smith et al., 2005; Monti, G. et al., 2005; Zupanc et al., 2006; Lee et al., 2006). Success rates of protein identification in non-model species can be increased using an MSBLAST approach that is exclusively based on sequence similarity and does not require matching peptide masses (Lee et al., 2006; Kjaersgard et al., 2006; Russell et al., 2006). Success rates of protein identification can be further increased by combining different methods of peptide

Table 2. *Identification of Tetilla mutabilis protein spots after 2DGE and tryptic in-gel digestion by MALDI-TOF/TOF MS*

Spot ID	Protein name	Mascot score (PSD, CID)	MSBLASTP2 score (PSD, CID)
130	–	–	–
177	Heat shock protein 70	–, 53	–, 106
228	Calreticulin	74, 62	191, 216
240	–	–	–
275	–	–	–
276	Hypothetical protein	–, –	134, –
306	Catalase	91, 88	321, 294
331	Flagellar basal body rod protein	–, –	–, 65
341	Catalase	–, –	–, 209
350	–	–	–
360	–	–	–
409	Catalase	68, 56	–, 273
411	Putative thiolase	–, 63	–, 59
423	HSP70-like protein	–, –	106, –
478	Gelsolin	89, 82	143, 157
490	Serine hydroxymethyl transferase	94, 88	181, 69
598	Actin	360, 316	618, 669
599	Actin	521, 618	607, 615
675	Actin	–, 248	348, 341
689	–	–	–
724	Actin	524, 508	–, 457
733	Glyceraldehyde-3-phosphate dehydrogenase	153, 150	107, 108
780	–	–	–
833	Catalase	75, –	222, 186
949	–	–	–
1011	Actin	227, 203	285, 264
1015	Polyubiquitin	–, 67	–, 67
1114	Hypothetical protein	–, –	123, 105
1135	–	–	–
1142	Peroxiredoxin	108, 73	155, 189
1224	Tropomyosin	53, –	205, 106
1362	–	–	–
1368	Peptidyl prolyl isomerase A	192, 136	116, –
1440	–	–	–
1461	Nucleoside diphosphate kinase	58, –	65, 65
1490	Nucleoside diphosphate kinase	–, –	133, 171

Spot ID values correspond to the labels of protein spots in Fig. 2. Scores are based on searches against non-redundant NCBI databases and a 50 p.p.m. threshold for mass accuracy.

fragmentation for MS/MS [e.g. collision-induced dissociation (CID) and post-source decay (PSD)] to generate a greater number of informative daughter ions (Lee et al., 2006).

Identification of sponge proteins by MS

Even invertebrate proteins for which close homologs from a related and fully sequenced reference organism are currently unavailable can be identified by MS. For instance, using MSBLASTP2 and a combination of CID and PSD for MS/MS, we identified many proteins from the evolutionarily primitive sponge *Tetilla mutabilis* (Fig. 2). In this experiment, we were interested in proteins that are upregulated during emersion and post-emersion stress (Fig. 2; Table 2). Such stress is experienced by this intertidal sponge during ebb-tide in its natural habitat at Estero La Cruz in the Gulf of California. The nature of stressors represents a mixture of hyperosmotic, heat, UV and oxidative stress. Sponges were collected in their natural habitat before emersion, after being emersed for 3 h and after being re-submersed for 2 h following 3 h emersion. Twenty-seven upregulated proteins and nine other proteins were picked for MS analysis, in which 25 of these 36 proteins (69%) were identified. This high success rate seems surprising compared with the 25% success rate with shark proteins (see above). However, for shark proteins we only considered proteins successfully identified if at least three out of the four search methods (PSD–Mascot, CID–Mascot, PSD–MSBLASTP2, CID–MSBLASTP2) yielded consistent results (Lee et al., 2006). If the same criteria are applied to the sponge samples, then the success rate drops to 44%. Again, the MSBLASTP2 approach identified five additional proteins (14% more) compared with the Mascot approach (Table 2), illustrating that database searches based on *de novo* peptide

sequence alignments are more successful in non-model species than PMF approaches.

To overcome current limitations for proteomics in non-model species and to increase the success rate of protein identification, a number of technical improvements can be made. In particular, efforts directed at increasing protein sequence coverage in mass spectra are promising. Increased sequence coverage can be achieved by combining multiple proteases with different sequence-specific cutting patterns (Biringer et al., 2006). Likewise, top-down mass spectrometry approaches increase sequence coverage but are currently limited by the low efficiency of elution of intact proteins from polyacrylamide (PAG) gels (Bogdanov and Smith, 2005). Thus, elution of whole proteins from PAG gels is another area of research that should be addressed to improve the versatility of proteomics approaches. The development of new methods for N- and C-terminal protein sequencing by MS will also increase success rates for protein identification based on cross-species comparison (McDonald et al., 2005). If terminal protein sequences are known, then BLAST searches can be greatly narrowed and degenerate primers for full-length cloning and sequencing of the corresponding cDNA can be designed. Finally, improvements in bioinformatics tools such as MSBLASTP2 will increase success rates of protein identification approaches that are based on sequence similarity searches (Shevchenko et al., 2001).

Pathway analysis using shark protein sets

Once protein sets have been identified, they can be used for pathway analysis as described for the SSH gene set above. We have performed such analysis for the set of shark proteins that are common to osmoregulatory tissues and unique to a single tissue. Even though the number of shark proteins used for the analysis (62) was relatively small, it was sufficient to provide some insight into molecular functions associated with osmoregulatory tissues in shark (Lee et al., 2006). Our analysis of molecular functions, cellular pathways and biological processes using PANTHER revealed that stress response proteins such as molecular chaperones and oxidoreductases are enriched and highly abundant in shark osmoregulatory tissues. In addition, we identified the Rho-GTPase pathway of cytoskeletal regulation (PANTHER pathway 00016) as being significantly over-represented in shark osmoregulatory tissues (Lee et al., 2006). Thus, the bioinformatics analysis of protein sets identified by proteomics technology generates specific targets for hypothesis-driven, focused and in-depth follow-up research in non-model organisms.

An important limitation of using bioinformatics approaches for pathway analysis based on identified protein sets in non-model organisms is the use of pathway models and other information from model species as templates for non-model species (see discussion above for SSH pathway analysis). In addition, virtually all proteins have multiple functions that depend on their subcellular compartmentation, posttranslational state and interaction with other cellular constituents. However, currently, functions of proteins stored in databases often do not take into account the aforementioned parameters and, even if they do, information about these parameters is often not available for protein sets of interest. This problem is exacerbated by mixing information about protein functions across species boundaries, which is inevitable when using high-throughput approaches with non-model species. These pitfalls illustrate that, despite the rapid growth of sequencing data, there is still an enormous need for the experimental analysis of proteomes and metabolomes. Even for a single-celled organism, experimentally determining all possible states of posttranslational modification, subcellular compartmentaion and interaction of proteins and correlating those with environmentally induced and developmental cell states may not be feasible. Thus, a narrow focus on a few model species seems warranted for collecting as many data as possible to achieve the ultimate goal of systems biology. Currently, this need for focusing on a handful of model species appears to outweigh the classical advantage of traditional comparative biology, as formulated elegantly by the August Krogh principle: '*For many problems there is an animal on which it can be most conveniently studied*' (Krebs, 1975). For instance, using stenohaline zebrafish for studying mechanisms of osmotic stress adaptation and euryhalinity in fishes may be counterintuitive. Medaka is better suited for such studies, but extremely euryhaline fishes such as tilapia, killifish or desert pupfish are the organisms of choice from a physiological perspective even if they are considered non-model species from a genomic perspective. Future breakthroughs in high-throughput technologies and bioinformatics, in combination with a reductionist focus on a single problem (or biological process) that lends itself exceptionally well for study in a non-model species, may provide a solution for optimally combining the strengths of traditional comparative biology and systems biology.

Analysis of posttranslational modification, interaction and compartmentation of osmotic stress response proteins

A significant advantage of identifying sets of genes and proteins associated with a particular biological process (e.g. salinity adaptation or osmoregulation) in non-model species is that such sets represent well-focused and manageable arrays of elements that can be studied in-depth in the context of a specific biological process. Such follow-up studies may include reverse genetics assays (using siRNA, overexpression, etc.), biochemical assays and imaging methodologies that provide information about the regulation and function of each gene/protein involved in osmotic stress adaptation or any other biological process of interest. In the following, we briefly outline selected examples of biochemical and histological approaches for characterizing proteins associated with salinity adaptation.

Posttranslational protein modification in response to osmotic stress can be studied by using phospho-specific or other state-specific antibodies (Kültz et al., 1997; Kültz and Avila, 2001) or MS following protein separation by 2DGE or 2DLC (Dihazi et al., 2005; Salih, 2005; Valkova and Kültz, 2006). For the

latter approach, protein abundance is a limiting factor. If protein sets have previously been identified by a proteomics approach, this is not a problem because all identified proteins exceed the detection threshold for 2DGE or 2DLC. However, when further studying gene sets identified by transcriptomic approaches, this limitation may represent a serious obstacle.

Another property of proteins that critically controls protein regulation and function is their association with other proteins, nucleic acids and micromolecules (inorganic ions, organic osmolytes, metabolic intermediates). The interactome (sum of bound macro- and micromolecules) can be determined for each of the proteins previously identified to be involved in salinity adaptation. Protein–protein interactions can be revealed using genetic or biochemical approaches (Cusick et al., 2005; Monti, M. et al., 2005; Suter et al., 2006). Genetic approaches require a cell line and knowledge about strong promoters that may not be available for many non-model organisms. Therefore, biochemical approaches may be easier to adapt to non-model organisms. However, most of them rely on high-quality antibodies. If not already available, such antibodies can be generated for a limited set of proteins of interest. Once an antibody is available, co-immunoprecipitation of proteins followed by separation of protein complexes by 2DGE or 2DLC and MS can be used to identify interacting proteins. Chromatin immunoprecipitation and variations of this method can be used to identify nucleic acid sequences that bind a protein of interest. Finally, using these approaches, salinity-dependent changes in the interactomes of osmotic stress response proteins can be identified.

The tissue-specific, cellular and subcellular compartmentation of osmotic stress response proteins can also be studied using genetic and biochemical approaches (Giltnane and Rimm, 2004; Suter et al., 2006). Again, biochemical/histological approaches for studying protein localization seem more amendable to non-model organisms than genetic approaches. However, like biochemical methods used for studying the interactome, they also rely to a great extent on high-quality antibodies. In particular, high-throughput immunohistochemical techniques such as tissue microarrays are promising tools for large-scale studies of protein compartmentation. We have recently used this technique to localize Na^+/K^+-ATPase to chloride cells in gills of FW- and SW-adapted tilapia (Lima and Kültz, 2004). This approach can be extended to other proteins involved in salinity adaptation once appropriate antibodies are available.

Conclusions and perspective

High-throughput transcriptomics and proteomics technologies are powerful tools for identifying sets of genes and proteins associated with salinity adaptation and the osmotic stress response. They can be effectively used in ‘non-model’ organisms, although some technical improvements are required to further increase success rates of protein identification. These more holistic and global approaches have the advantage of being less biased and more informative with regard to the biological process of interest when compared to a reductionist single-gene/protein approach that requires prior knowledge about which genes/proteins are involved.

Sets of genes or proteins associated with salinity adaptation can be effectively identified in non-model organisms as illustrated by the SSH and proteomics examples outlined above and by recent cDNA microarray experiments that are based on normalized cDNA libraries from non-model species or on cross-species hybridization (Gracey et al., 2001; Podrabsky and Somero, 2004; Buckley et al., 2006). A more severe barrier for work with non-model species is the reliance on models of biochemical pathways, gene ontology databases and other information that is largely based on data obtained with model species. Although many biochemical pathways and biological processes are evolutionarily highly conserved, a great amount of plasticity is inherent in cellular metabolic and signal transduction networks. We would expect such plasticity to be particularly pronounced when a non-model organism differs in its properties, with regard to the biological process under study, from model organisms. For instance, how valid are models of biochemical pathways/networks based on sets of osmotic stress response proteins for a euryhaline fish when modeled against templates based on information available for stenohaline zebrafish? Even though the individual elements (proteins) may be evolutionarily highly conserved, their posttranslational modification, interaction, compartmentation and function during salinity stress may differ substantially. Currently, there are no alternatives to this approach, but this area undoubtedly requires further study.

In our opinion, the most promising and feasible avenue for merging strengths of traditional comparative biology with strengths of systems biology approaches is to select a single biological process and focus on it (rather than attempting to describe the system as a whole) using an organism selected according to the August Krogh principle. Gene and protein sets associated with the biological process of interest (e.g. salinity adaptation) can then be identified. Next, tools for studying the regulation (including posttranslational modification, interaction and compartmentation) and function of these proteins (e.g. antibodies, cell lines, siRNA) should be generated for a manageable set of proteins. Then, the regulation and function of these proteins can be analyzed in detail using such tools. Finally, models describing the biological process of interest can be generated based on the resulting data. This avenue of research depends on extensive collaboration and, therefore, represents an exciting path for the future.

This project was supported by grant number IOB-0542755 from the National Science Foundation (NSF), UCMexus-CONACYT, an NIA from the MDIBL, and grant number 5 P42 ES004699 from the National Institute of Environmental Health Sciences (NIEHS), NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NSF or NIEHS, NIH.

References

Aggarwal, K. and Lee, K. H. (2003). Functional genomics and proteomics as a foundation for systems biology. *Brief. Funct. Genomic. Proteomic.* **2**, 175-184.

Biringer, R. G., Amato, H., Harrington, M. G., Fonteh, A. N., Riggins, J. N. and Huhmer, A. F. (2006). Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS. *Brief. Funct. Genomic. Proteomic.* **5**, 144-153.

Bogdanov, B. and Smith, R. D. (2005). Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom. Rev.* **24**, 168-200.

Buckley, B. A. (2007). Comparative environmental genomics in non-model species: using heterologous hybridization to DNA-based microarrays. *J. Exp. Biol.* **210**, 1602-1606.

Buckley, B. A., Gracey, A. Y. and Somero, G. N. (2006). The cellular response to heat stress in the goby Gillichthys mirabilis: a cDNA microarray and protein-level analysis. *J. Exp. Biol.* **209**, 2660-2677.

Bunai, K. and Yamane, K. (2005). Effectiveness and limitation of two-dimensional gel electrophoresis in bacterial membrane protein proteomics and perspectives. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **815**, 227-236.

Cusick, M. E., Klitgord, N., Vidal, M. and Hill, D. E. (2005). Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14**, R171-R181.

Diatchenko, L., Lukyanov, S., Lau, Y. F. C. and Siebert, P. D. (1999). Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. *Meth. Enzymol.* **303**, 349-380.

Dihazi, H., Kessler, R., Muller, G. A. and Eschrich, K. (2005). Lysine 3 acetylation regulates the phosphorylation of yeast 6-phosphofructo-2-kinase under hypo-osmotic stress. *Biol. Chem.* **386**, 895-900.

Fiol, D. F. and Kültz, D. (2005). Rapid hyperosmotic coinduction of two tilapia (*Oreochromis mossambicus*) transcription factors in gill cells. *Proc. Natl. Acad. Sci. USA* **102**, 927-932.

Fiol, D. F., Chan, S. Y. and Kültz, D. (2006a). Identification and pathway analysis of immediate hyperosmotic stress responsive molecular mechanisms in tilapia (*Oreochromis mossambicus*) gill. *Comp. Biochem. Physiol.* **1D**, 344-356.

Fiol, D. F., Chan, S. Y. and Kültz, D. (2006b). Regulation of osmotic stress transcription factor 1 (Ostf1) in tilapia (Oreochromis mossambicus) gill epithelium during salinity stress. *J. Exp. Biol.* **209**, 3257-3265.

Giltnane, J. M. and Rimm, D. L. (2004). Technology insight: identification of biomarkers with tissue microarray technology. *Nat. Clin. Pract. Oncol.* **1**, 104-111.

Goodrich, J. A. and Kugel, J. F. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* **7**, 612-616.

Gracey, A. Y. (2007). Interpreting physiological responses to environmental change through gene expression profiling. *J. Exp. Biol.* **210**, 1584-1592.

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.

Hack, C. J. (2004). Integrated transcriptome and proteome data: the challenges ahead. *Brief. Funct. Genomic. Proteomic.* **3**, 212-219.

Hu, S., Xie, Y., Ramachandran, P., Ogorzalek Loo, R. R., Li, Y., Loo, J. A. and Wong, D. T. (2005). Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. *Proteomics* **5**, 1714-1728.

Hubank, M. and Schatz, D. G. (1999). cDNA representational difference analysis: a sensitive and flexible method for identification of differentially expressed genes. *Meth. Enzymol.* **303**, 325-349.

Ishihama, Y. (2005). Proteomic LC-MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *J. Chromatogr. A* **1067**, 73-83.

Kitano, H. (2002). Systems biology: a brief overview. *Science* **295**, 1662-1664.

Kjaersgard, I. V., Norrelykke, M. R. and Jessen, F. (2006). Changes in cod muscle proteins during frozen storage revealed by proteome analysis and multivariate data analysis. *Proteomics* **6**, 1606-1618.

Krebs, H. A. (1975). The August Krogh Principle: "For many problems there is an animal on which it can be most conveniently studied". *J. Exp. Zool.* **194**, 221-226.

Kültz, D. (2005). Molecular and evolutionary basis of the cellular stress response. *Annu. Rev. Physiol.* **67**, 225-257.

Kültz, D. and Avila, K. (2001). Mitogen-activated protein kinases are in vivo transducers of osmosensory signals in fish gill cells. *Comp. Biochem. Physiol.* **129B**, 821-829.

Kültz, D., Garcia-Perez, A., Ferraris, J. D. and Burg, M. B. (1997). Distinct regulation of osmoprotective genes in yeast and mammals. Aldose reductase osmotic response element is induced independent of p38 and stress-activated protein kinase/Jun N-terminal kinase in rabbit kidney cells. *J. Biol. Chem.* **272**, 13165-13170.

Lanahan, A. and Worley, P. (1998). Immediate-early genes and synaptic function. *Neurobiol. Learn. Mem.* **70**, 37-43.

Lee, J., Valkova, N., White, M. P. and Kültz, D. (2006). Proteomic identification of processes and pathways characteristic of osmoregulatory tissues in spiny dogfish shark (*Squalus acanthias*). *Comp. Biochem. Physiol.* **1D**, 328-343.

Lima, R. N. and Kültz, D. (2004). Laser scanning cytometry and tissue microarray analysis of salinity effects on killifish chloride cells. *J. Exp. Biol.* **207**, 1729-1739.

Maillard, J. C., Berthier, D., Thevenon, S., Piquemal, D., Chantal, I. and Marti, J. (2005). Efficiency and limits of the Serial Analysis of Gene Expression (SAGE) method: discussions based on first results in bovine trypanotolerance. *Vet. Immunol. Immunopathol.* **108**, 59-69.

McDonald, L., Robertson, D. H., Hurst, J. L. and Beynon, R. J. (2005). Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat. Methods* **2**, 955-957.

Mi, H. Y., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J. et. al. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284-D288.

Monti, G., De Napoli, L., Mainolfi, P., Barone, R., Guida, M., Marino, G. and Amoresano, A. (2005). Monitoring food quality by microfluidic electrophoresis, gas chromatography, and mass spectrometry techniques: effects of aquaculture on the sea bass (*Dicentrarchus labrax*). *Anal. Chem.* **77**, 2587-2594.

Monti, M., Orru, S., Pagnozzi, D. and Pucci, P. (2005). Interaction proteomics. *Biosci. Rep.* **25**, 45-56.

Morgan, J. I. and Curran, T. (1989). Stimulus-transcription coupling in neurons – role of cellular immediate-early genes. *Trends Neurosci.* **12**, 459-462.

Okano, T., Kondo, T., Kakisaka, T., Fujii, K., Yamada, M., Kato, H., Nishimura, T., Gemma, A., Kudoh, S. and Hirohashi, S. (2006). Plasma proteomics of lung cancer by a linkage of multi-dimensional liquid chromatography and two-dimensional difference gel electrophoresis. *Proteomics* **6**, 3938-3948.

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.

Russell, S., Hayes, M. A., Simko, E. and Lumsden, J. S. (2006). Plasma proteomic analysis of the acute phase response of rainbow trout (*Oncorhynchus mykiss*) to intraperitoneal inflammation and LPS injection. *Dev. Comp. Immunol.* **30**, 393-406.

Salih, E. (2005). Phosphoproteomics by mass spectrometry and classical protein chemistry approaches. *Mass Spectrom. Rev.* **24**, 828-846.

Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W. and Standing, K. G. (2001). Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73**, 1917-1926.

Smith, R. W., Wood, C. M., Cash, P., Diao, L. and Part, P. (2005). Apolipoprotein AI could be a significant determinant of epithelial integrity in rainbow trout gill cell cultures: a study in functional proteomics. *Biochim. Biophys. Acta* **1749**, 81-93.

Stein, J. and Liang, P. (2002). Differential display technology: a general guide. *Cell. Mol. Life Sci.* **59**, 1235-1240.

Suter, B., Auerbach, D. and Stagljar, I. (2006). Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. *Biotechniques* **40**, 625-644.

Thomson, S., Mahadevan, L. C. and Clayton, A. L. (1999). MAP kinase-mediated signalling to nucleosomes and immediate-early gene induction. *Semin. Cell Dev. Biol.* **10**, 205-214.

Valkova, N. and Kültz, D. (2006). Constitutive and inducible stress proteins dominate the proteome of the murine inner medullary collecting duct-3 (mIMCD3) cell line. *Biochim. Biophys. Acta* **1764**, 1007-1020.

Wagner, K., Racaityte, K., Unger, K. K., Miliotis, T., Edholm, L. E., Bischoff, R. and Marko-Varga, G. (2000). Protein mapping by two-dimensional high performance liquid chromatography. *J. Chromatogr. A* **893**, 293-305.

Zupanc, M. M., Wellbrock, U. M. and Zupanc, G. K. (2006). Proteome analysis identifies novel protein candidates involved in regeneration of the cerebellum of teleost fish. *Proteomics* **6**, 677-696.

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.