The Journal of Experimental Biology 209, 1602-1606
Published by The Company of Biologists 2007
doi:10.1242/jeb.002402

# Comparative environmental genomics in non-model species: using heterologous hybridization to DNA-based microarrays

Bradley A. Buckley
*Department of Biology, Portland State University, Portland, OR 97201, USA*
e-mail: bbuckley@pdx.edu

*Accepted 13 March 2007*

## Summary

**The emerging field of comparative environmental genomics involves the cross-species comparison of broad-scale patterns of gene expression. Often, the goal is to elucidate the evolutionary basis or ecological implications of genomic responses to environmental stimuli. DNA-based microarrays represent powerful means with which to investigate gene expression, and the application of genomic tools to studies on non-model species is becoming increasingly feasible. The use of a microarray generated from one species to probe gene expression in another, a method termed 'heterologous hybridization', eliminates the need to fabricate novel microarray platforms for every new species of interest. In this review, recent advances in heterologous hybridization are reviewed, and the technical caveats of this approach are discussed.**

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1602/DC1

Key words: DNA microarray, comparative environmental genomics, gene expression, heterologous hybridization.

## Introduction

Understanding how evolutionary history and genotype are linked to phenotypic responses to the environment is a central goal of modern comparative physiology. In the post-genomic era, this goal is being approached in novel ways through the application of resources, such as DNA-based microarrays, to the investigation of environmental regulation of gene expression. Of the growing number of applications of such resources, the use of transcriptomic profiling to characterize gene expression in organisms undergoing environmental perturbation is perhaps the most mature (Hoheisel, 2006). Transcriptional profiling has been used widely in both genetic model species [e.g. in yeast (Gasch et al., 2000; Gasch et al., 2001; Chen et al., 2003), in *Drosophila* (Girardot et al., 2004), in human cells (Murray et al., 2004) and in *E. coli* (Riehle et al., 2005)] and non-model species, with studies on fishes representing a dominant part of the literature in this field (Gracey et al., 2001; Oleksiak et al., 2002; Feder and Mitchell-Olds, 2003; Gracey et al., 2004; Podrabsky and Somero, 2004; Krasnov et al., 2005; Sneddon et al., 2005; Vornanen et al., 2005; Buckley et al., 2006). These types of studies can identify both phylogenetically conserved patterns of environmentally responsive gene expression as well as taxa-specific responses. In a similar way, differences in general and stressor-specific responses can be characterized. While a detailed treatment of the myriad applications of DNA-based microarrays is outside the scope of the current review, several recent syntheses are available that examine the revolutionary advances that these new genomic tools have enabled in various fields, including medicine, ecology, evolution, ecophysiology and ecotoxicology (Gracey and Cossins, 2003; Hoheisel, 2006; Lettieri, 2006).

The lack of available genomic sequence information for species outside the traditional genetic models is no longer an impediment to using genomic tools to investigate patterns of gene expression in these organisms. It is becoming increasingly clear that within related phylogenetic groups, adequate sequence identity exists for many genes to allow for a genomic platform developed for one species in the group to be applied to its other members. In the case of cross-species comparisons of gene expression using DNA-based microarrays designed for a single species, this approach has been termed 'heterologous' hybridization (Renn et al., 2004). In this review, I discuss the recent applications of heterologous hybridization to DNA microarrays, highlighting its strengths and weaknesses. Various factors are considered that may affect the efficacy of this approach, including such variables as the phylogenetic distance between the species involved, the nature and length of the DNA probes affixed to the microarray platform and the experimental design employed.

## Measuring gene expression with DNA-based microarrays

In the following discussion, a 'microarray' refers to any platform (usually glass microscope slides or nylon membranes) to which oligonucleotide or cDNA probes are permanently affixed and to which samples of fluorescently labeled target cDNAs, generated through reverse transcription of expressed mRNA, are hybridized. The competitive hybridization of two samples, each labeled with a unique fluorescent dye, allows for the direct comparison of the mRNA levels from each sample. The results are generally reported as a set of ratios representing the fold-difference in expression between the two samples for each spot or 'feature' on the microarray. In the majority of microarray studies, the DNA probes affixed to the platform are generated from the same species that provide the fluorescently labeled samples (i.e. 'single-species' hybridization). A growing number of studies are utilizing probes generated from one species to examine gene expression in other species. The obvious advantage of this heterologous hybridization is that it avoids the expense in time and money associated with the production of cDNA clones and expressed sequence tag (EST) data when fabricating novel microarrays for every new species of interest, provided a platform for a related extant species is accessible. Often, the primary question then becomes: to what degree must one species be related to another for microarray technology to be transferable across taxa?

## The effect of sequence divergence on microarray analyses

The primary technical challenge presented by heterologous hybridization is the problem of sequence divergence between the species for which the microarray was constructed and the species providing the sample to which it will be hybridized. Sequence divergence influences hybridization kinetics and therefore it is important to differentiate between differences in detection intensity that are due to actual differential gene expression and those that may be due to sequence mismatches. The endeavor is complicated by the fact that gene-by-gene divergence rates will differ from other metrics of phylogenetic distance such as species' evolutionary divergence time or average genome-wide divergence rates.

The competitive hybridization of genomic DNA from multiple species to a single-species array can be helpful in providing a quantitative assessment of the impact of sequence divergence on overall hybridization efficiency. For example, in a study on different species of *Drosophila* (Ranz et al., 2003), genomic DNA from *D. melanogaster* displayed an average of 4.2% greater hybridization to a *D. melanogaster* array than did genomic DNA from *D. simulans*. This disparity in hybridization strength was in broad agreement with the known sequence divergence between these two species (3.8% different at the nucleotide level). Especially in cases where the degree of sequence divergence between two species is not known, the relative binding of genomic DNA from the two species will provide an idea as to the effect of evolutionary distance on hybridization efficiency.

The problem of comparative differences in expressed isoforms creating false positives is also likely to increase with evolutionary distance. This is of particular concern for members of large gene families with many isoforms and/or variants of ancestral genes. As species diverge, it becomes increasingly difficult to discern specific patterns of expression in such families where multiple cDNAs may bind to a single probe bearing a conserved region shared by all isoforms or variants.

Owing solely to sequence divergence, the number of features that a single-species microarray can detect in targets from another species is expected to decrease with increasing phylogenetic divergence. This appears to generally hold true, although not to the extent that one might initially suppose. In a study employing a 16 006-gene salmonid microarray, those features generated from Atlantic salmon (*Salmo salar*) or rainbow trout (*Oncorhynchus mykiss*) were equally able to detect target cDNAs from either species, despite the 8–20 million years of divergence time between these two species (von Schalburg et al., 2005). In another study, Rise et al. (Rise et al., 2004) tested the ability of a 7356-feature cDNA microarray, generated from ESTs from rainbow trout and Atlantic salmon, to detect target cDNAs from lake whitefish (*Coreogonus clupeaformis*) and smelt (*Osmerus mordax*). As expected, hybridization performance did rank according to evolutionary relationships, with the lowest number of features being detected in the most diverged species (smelt). However, 38% of the Atlantic salmon features on the microarray detected smelt target cDNAs, compared with 70% of Atlantic salmon targets. While hybridization performance decreased by approximately half in smelt, this nevertheless resulted in nearly 2500 features being successfully detected. In a sense, then, this approach merely reduces the effective size of a given microarray. With current technology allowing for the dense spotting of many thousands of features onto glass slides, the detecting power of even a numerically diminished microarray still remains considerable.

It is important to bear in mind however that 'number of detected spots' does not translate into information on changes in expression level. The ability of a given platform to detect changes in gene expression, particularly in poorly detected features, would be expected to diminish with increasing phylogenetic distance as sequence mismatches begin to create variation in hybridization strength, even for features that pass the detection threshold. However, these challenges may be mitigated by the choice of experimental design, as discussed in a later section.

## Short oligonucleotides *versus* full-length cDNAs

Heterologous hybridization efficiency may also depend upon the nature and length of the probes employed. Short oligonucleotide probes, such as those on Affymetrix GeneChips® (Affymetrix, Inc., Santa Clara, CA, USA), are likely to be more sensitive to sequence mismatches than are longer probes, such as full-length cDNAs. In one of the first studies to employ heterologous hybridization (Enard et al., 2002), arrays generated from either human oligonucleotide sequences (Affymetric U95A arrays) or longer cDNAs (~1000 bp) were used to characterize quantitative differences in gene expression among several primate species. The authors acknowledge the likelihood that sequence differences between species may have affected the outcome of the experiments using short oligonucleotide probes. They assert, however, that with the use of longer cDNA probes, the 0.8% nucleotide sequence difference between human and chimpanzee was not expected to affect the results significantly and that variation in the data due to sequence divergence was smaller than that due to experimental error. By using longer probes and by maintaining high stringency in hybridization conditions (e.g. keeping hybridization temperature at or close to 65°C for all hybridizations and using high-stringency washing procedures), non-specific binding of mismatched targets can be kept to a minimum.

Where the use of short oligonucleotides is desired, advantage can be gained from the fact that the complete sequences of the spotted features from probes and targets are often known and the effect of sequence identity on detection strength can be determined directly. Methods of masking the poorly hybridized probes *in silico* can then restrict further analysis to those probes that possess sufficient sequence homology to detect target cDNAs across taxa. This approach has been used to compare gene expression among mammals as evolutionarily diverged as dogs, pigs, cows and humans (Ji et al., 2004). As demonstrated in a recent study on *Xenopus*, it is also possible in such circumstances to remove the hybridization bias from interspecific comparisons by calculating and applying correction factors to expression data (Sartor et al., 2006).

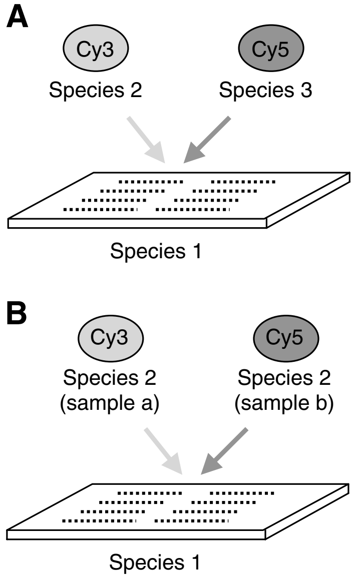

Fig. 1. Two general experimental designs often employed by studies using heterologous hybridization to microarrays. In the first design (A), samples from two different species (species 2 and 3) are competitively hybridized against one another to a microarray generated from oligonucleotides or cDNAs from a single species (species 1). Note that, in some cases, species 1 may be the same as either 2 or 3. In this design, the sequence distance between species 1 and 2 will differ, to some degree, from that between 1 and 3; if this difference is too great, it may affect hybridization kinetics, which may in turn artificially affect the generated gene expression values. Under the second design (B), the two hybridized samples are always from the same species, and the two samples generally differ in another variable, e.g. treatment, time point or tissue. With this design, the only sequence divergence factor is that between species 1 and 2, and this factor should affect both hybridized samples equally.

## ‘Apples to apples’: the importance of experimental design

The choice of experimental design can also help mitigate the challenges presented by between-species sequence divergence. Heterologous hybridization studies have generally employed one of two basic experimental designs (Fig. 1). In some cases, a microarray containing probes directed against species 1 is used to compare gene expression profiles in two different species (species 2 and 3; Fig. 1A). With this design, there are two divergence factors to consider; namely, the percentage of sequence homology between species 1 and 2 and that between species 1 and 3 (note that in some cases, species 1 may be the same as 2 or 3, in which case one of the percentages of sequence identity will be 100%).

For very closely related species, this design may nevertheless be effective and has been used, for example, to explore patterns of gender-biased gene expression in different species of *Drosophila* (Ranz et al., 2003; Meiklejohn et al., 2003). However, in a study on primates that employed both single- and multi-species microarrays to directly test the limits of inter-specific competitive hybridization (Gilad et al., 2006), it was demonstrated that the difference in sequence homology between humans and chimpanzees was sufficient to affect the resulting gene expression values when target cDNAs from each species were directly compared first on a human-based microarray and then on a chimpanzee microarray. Even the use of relatively long cDNA probes apparently did not eliminate the problem, which was especially significant for instances when the differences in gene expression between species were subtle (e.g. ~1–2 fold).

The use of an alternative experimental design (Fig. 1B) avoids the problem of phylogenetic distance between the two samples being competitively hybridized by always comparing two different samples from the same species (i.e. comparing ‘apples to apples’ rather than ‘apples to oranges’). With this design, there is only a single divergence factor to consider (species 1 *vs* 2; Fig. 1B) and it applies equally to both samples, allowing for accurate measurements of their relative levels of specific mRNAs. The two samples could differ in any experimental variable, such as treatment, time point, collection site, developmental stage or tissue.

This design was recently used to demonstrate the efficacy of heterologous hybridization in measuring biologically meaningful differences in gene expression for several species of fish, using a ~4500-feature cDNA microarray that was generated from brain tissue of an African cichlid, *Astatotilapia burtoni* (Renn et al., 2004). Target cDNA samples from brain and mixed muscle from this species were competitively hybridized against one another on the microarray to establish a set of 804 'reference' genes that were expressed differentially between these two tissues. Subsequently, similar hybridizations were performed comparing muscle and brain samples from seven other fish species. These species included three other members of the order Perciformes, as well as more distantly related species, such as the zebrafish *Danio rerio* (diverged from *A. burtoni* by ~200 million years). As expected, the total number of features detected decreased with phylogenetic distance, although the decrease was surprisingly moderate. In even the most diverged species, Renn et al. found that 3000–4000 spots out of 4500 were detected by the *A. burtoni* microarray. Hybridization efficiency was particularly high among the perciform fishes, even though this order spans over 65 million years of divergence time.

Another important finding of this study was that, of the 804 reference spots whose expression differed between tissues in *A. burtoni*, nearly 80% also differed in the other perciform species. This number did decrease significantly, however, in comparisons of more highly diverged species. For instance, only ~20% of the reference spots displayed changes in expression in zebrafish, the most phylogenetically distant species examined. This underscores the inverse relationship between sequence divergence and the conservation of gene regulatory patterns, even for features that are well detected by a given array. Nevertheless, these results support the ability of heterologous hybridization to reveal conserved patterns of biologically relevant gene expression across considerable taxonomic spans such as those encompassing the perciform fishes.

In my laboratory, similar success has been achieved using a 9200-feature cDNA microarray generated from ESTs from the eurythermal goby *Gillichthys mirabilis* to characterize the responses to heat stress in the cold-adapted (and evolutionarily distant) fish species of the Antarctic (B.A.B., unpublished data). In keeping with the findings above, the number of spots detected using heterologous targets tends to decline with evolutionary distance, but not significantly. Interestingly, the fold-changes in expression measured in the heterologous hybridizations were lower than those measured in hybridizations using the homologous targets [a similar phenomenon was observed among fish species by Renn et al. (Renn et al., 2004)]. Whether this represents a reduced ability of the Antarctic fish to up- and down-regulate gene expression or is an artifact of heterologous hybridization remains to be determined.

## Conclusions

Based on the first round of studies exploring cross-species microarray analysis, it appears that single-species platforms present a promising means by which to explore genomic responses to the environment across related species, even in non-model organisms. Prudent measures should always be employed to ensure that poorly detected features are excluded from analysis, even though this may reduce the effective size of a given microarray. While successful detection of numerous features in cross-species analyses is encouraging, the impact of sequence divergence on the conservation of gene regulatory patterns is significant. As with any microarray experiment, 'spot-checking' of selected expression data with routine methods of mRNA quantification such as quantitative real-time PCR (qPCR) and northern blotting can also provide another layer of quality control and help strengthen the results obtained by heterologous hybridization to DNA-based microarrays.

This manuscript was much improved by the comments of Dr George Somero and those of two anonymous reviewers.

## References

**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.

**Chen, D., Toone, W. M., Mata, J., Lyne, R., Burns, G., Kivinen, K., Brazma, A., Jones, N. and Bähler, J.** (2003). Global transcriptional responses of fission yeast to environmental stress. *Mol. Cell. Biol.* **14**, 214-229.

**Enard, W., Khaltovich, P., Klose, J., Zollner, S., Heisseg, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varkl, A., Ravid, R. et al.** (2002). Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340-343.

**Feder, M. E. and Mitchell-Olds, T.** (2003). Evolutionary and ecological functional genomics. *Nat. Rev. Genet.* **4**, 649-655.

**Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. and Brown, P. O.** (2000). Genomic expression programs in the response of yeast cells to environmental change. *Mol. Biol. Cell* **11**, 4241-4257.

**Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J. and Brown, P. O.** (2001). Genomic expression responses to DNA-damaging agent and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell* **12**, 2987-3003.

**Gilad, Y., Rifkin, S. A., Bertone, P., Gerstein, M. and White, K. P.** (2006). Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.* **15**, 674-680.

**Girardot, F., Monnier, F. and Tricoire, H.** (2004). Genome wide analysis of common and specific stress responses in adult *Drosophila melanogaster*. *BMC Genomics* doi: 10.1186/1471-2164-5-74.

**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.

**Hoheisel, J. D.** (2006). Microarray technology: beyond transcript profiling and genotype profiling. *Nat. Rev. Genet.* **7**, 200-210.

**Ji, W., Zhou, W., Gregg, K., Yu, N., Davis, S. and Davis, S.** (2004). A method for cross-species gene expression analysis with high-density oligonucleotide arrays. *Nucl. Acids Res.* doi: 10.1093/nar/guh084.

**Krasnov, A., Koskinen, H., Pehkonen, P., Rexroad, C. E., III, Afanasyev, S. and Molsa, H.** (2005). Gene expression in the brain and kidney of rainbow trout in response to handling stress. *BMC Genomics* doi: 10.1186/1471-2164-6-3.

**Lettieri, T.** (2006). Recent applications of DNA microarray technology to toxicology and ecotoxicology. *Env. Health Perspec.* **114**, 4-9.

**Meiklejohn, C. D., Parsh, J., Ranz, J. M. and Hartl, D. L.** (2003). Rapid evolution of male-biased gene expression in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **100**, 9894-9899.

**Murray, J. I., Whitfield, M. L., Trinklein, N. D., Myers, R. M., Brown, P. O. and Botstein, D.** (2004). Diverse and specific gene expression responses to stresses in cultured human cells. *Mol. Biol Cell*. **15**, 2361-2374.

**Oleksiak, M. F., Churchill, G. and Crawford, D. L.** (2002). Variation in gene expression within and among natural populations. *Nature Genet.* **32**, 261-266.

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

**Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D. and Hartl, D. L.** (2003). Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300**, 1742-1745.

**Renn, S. C. P., Aubin-Horth, N. and Hofmann, H. A.** (2004). Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics* doi: 10.1186/1471-2164-5-42.

**Riehle, M. M., Bennett, A. F. and Long, A. D.** (2005). Changes in gene expression following high-temperature adaptation in experimentally evolved populations of *E. coli*. *Physiol. Biochem. Zool*. **78**, 299-315.

**Rise, M. L., von Schalburg, K. R., Brown, G. D., Mawer, M. A., Devlin, R. H., Kuipers, N., Busby, M., Beetz-Sargent, M., Alberto, R., Gibbs, A. R. et al.** (2004). Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res*. **14**, 478-490.

**Sartor, M. A., Zorn, A. M., Schwanekamp, J. A., Halbleib, D., Karyala, S., Howell, M. L., Dean, G. E., Medvedovic, M. and Tomlinson, C. R.** (2006). A new method to remove hybridization bias for interspecies comparison of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in *Xenopus*. *Nucl. Acids Res*. **34**, 185-200.

**Sneddon, L. U., Margaretto, J. and Cossins, A. R.** (2005). The use of transcriptomics to address questions in behaviour: production of a suppression subtractive hybridisation library from dominance hierarchies of rainbow trout. *Physiol. Biochem. Zool*. **78**, 695-705.

**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* doi: 10.1186/1471-2164-6-126.

**Vornanen, M., Hassinen, M., Koskinen, H. and Krasnov, A.** (2005). Steady-state effects of temperature acclimation on the transcriptome of the rainbow trout heart. *Am. J. Physiol. Regul. Integr. Comp. Physiol*. **289**, R1177-R1184.

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.