Review TRENDS in Ecology and Evolution Vol.21 No.1 January 2006

Uncovering evolutionary patterns of gene expression using microarrays

José M. Ranz[1] and Carlos A. Machado[2]

[1]Department of Genetics, University of Cambridge, Downing Street Cambridge, UK, CB2 3EH
[2]Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, AZ 85721, USA

The advent of microarray technology is providing new insights into fundamental questions in evolutionary biology. Here, we review the recent literature on the use of microarrays to study the evolution of genome-wide patterns of gene expression within and between species. Large levels of variation in gene expression patterns have been observed at the intra and inter-specific level, and a substantial fraction of transcriptional variation has a genetic component that is contributed by changes in both *cis*-acting and *trans*-acting regulatory elements. We argue that there is solid evidence to show that the temporal dynamics of transcriptional variation is largely determined by natural selection, with the fraction of the transcriptome more closely related to sex and reproduction evolving more rapidly.

Having a detailed knowledge of the patterns and mechanisms of evolution at the structural and regulatory level is fundamental for understanding the genetic basis of evolutionary change. Much of the attention of evolutionary geneticists in the past decade has been devoted to understanding the evolution of the structural elements of the genome (i.e. DNA sequences, chromosomes and repetitive elements). By contrast, our knowledge of the patterns, rates and mechanisms of change at the regulatory level is still in its infancy in spite of increasing evidence indicating that regulatory changes can have extraordinary evolutionary consequences [1,2]. These gaps in our understanding of regulatory evolution are partly the result of difficulties in conducting comparative evolutionary studies of regulatory variation using traditional molecular techniques. However, our ability to conduct comparative studies of regulatory evolution is changing dramatically owing to the development of several techniques that enable the analysis of gene expression patterns on a genomic scale (Box 1).

We are currently witnessing a revolution in molecular biology that could change the way in which we approach basic questions in evolutionary genetics. Today, one can measure a basic molecular phenotype (gene expression level) without regard to how gene expression relates to genetic variation or to particular phenotypes, and do so across most or all of the genes in a genome using microarrays (Box 2). Although most microarray platforms, both commercial and in house, have been mainly developed for model organisms, such as *Drosophila*, they have begun to be used in comparative studies of gene expression evolution, providing us with useful insights into several fundamental questions in evolutionary biology.

Here, we review the recent literature on the use of high-throughput technologies to study the evolution of gene expression patterns within and among species, focusing on the use of microarrays and their applications in comparative studies of transcriptome evolution. Based on evidence accumulated over the past few years, we argue

Box 1. Transcriptome evolution

The transcriptome is the complete set of transcribed elements of the genome and includes all types of RNA from the cell: mRNAs (with all their spliced forms), tRNAs, rRNAs, and non-coding RNAs involved in RNA-based regulation (antisense RNA, microRNAs, and non-coding RNAs [58]). The transcriptome is dynamic and is time, environment, tissue and cell specific. Transcriptome changes among species or individuals are due to changes both in *cis* and in *trans*-acting regulatory elements that are scattered across the genome. The mechanisms affecting transcriptome evolution act at two interconnected levels: at the level of the sequence of the regulatory elements and at the level of transcript abundance. There is currently solid evidence that natural selection has a major role in the dynamics of change of transcript abundance. By contrast, and although there is some evidence that natural selection also has a role in the evolution of regulatory sequences (e.g. [54,60]), a better understanding of its role is hampered by the difficulty in identifying those regulatory sequences as well as identifying functionally important nucleotide changes in those sequences.

The suggestion that changes in the time, level and location of gene expression (i.e. transcriptome changes) are fundamental for generating evolutionary change and have a major role in the adaptation process, has a relatively long history in molecular evolutionary studies [1,2] (see [61] for a recent discussion and expanded references). In a classic molecular evolution study, Wilson and colleagues [62] observed that rates of morphological evolution are poorly correlated with rates of protein evolution, suggesting that morphological evolution is mainly the result of changes in the patterns of gene expression rather than in the alteration of protein coding sequences. This suggestion has been empirically investigated mainly in the context of the evolution of development, both at micro and macroevolutionary scales, focusing on particular genes or suites of genes involved in specific developmental pathways [1,2]. However, in spite of significant progress achieved during the past decade, our understanding of the connection between gene expression changes and evolution is still confined mainly to a few regulatory pathways from a handful of model organisms. However, genomic techniques such as microarrays are providing the opportunity to address questions about regulatory evolution from a different perspective and at different scales by changing the focus from a few genes and a single regulatory pathway to the whole genome and multiple regulatory networks.

Corresponding author: Ranz, J.M. (j.ranz@gen.cam.ac.uk).
Available online 21 September 2005

 0169-5347/$ - see front matter doi:10.1016/j.tree.2005.09.002

Box 2. Genomic methodologies to study gene expression in different species

The past decade has seen the development of new techniques to study the transcriptome. The most popular are: (i) serial analysis of gene expression (SAGE) [63]; (ii) sequencing of expressed sequence tags (ESTs) [64]; (iii) subtractive hybridization [65]; (iv) differential display [66]; and (v) microarrays [67].

SAGE and ESTs are sequencing-based methodologies that have had important roles in gene and exon discovery efforts, but are not suitable for routine comparative studies mainly because of the large expense inherent to sequencing thousands of clones or PCR products. Subtractive hybridization methods are useful for isolating up- or down-regulated transcripts using reassociation kinetics, but are technically demanding. Differential display is a straightforward modification of RAPDs methods using RNA as template, but the downstream procedures to isolate cDNAs of interest normally lead to large numbers of false positives. Subtractive hybridization and differential display can both be used to isolate transcripts showing large differences in abundance between two samples, but are not useful for quantifying precisely the relative amounts of differentially expressed genes on a genomic scale.

The microarray technology is the most useful high-throughput tool with which to conduct transcriptome analyses. Microarray methods are an extension of standard nucleic-acid hybridization procedures (Northern or Southern blots), which enable the simultaneous monitoring of mRNA levels for thousands of genes. The rationale behind microarray technology is to perform a hybridization reaction between a sample of mRNA species (represented as labeled cDNA or cRNA) and a large set of gridded DNA reporters attached to a solid support. The DNA reporters normally used are either short (25-mer, Affymetrix chips) or long (70-mer) oligonucleotides or PCR products (>200 bp) that usually represent distinct transcribed protein-coding genes. The thousands of parallel hybridization reactions occur under stringent experimental conditions that enable preferential binding of each reporter with its homologous DNA sequence. The level of hybridization between each reporter and its homologous mRNA species is measured, and the magnitude of hybridization detected is assumed to be proportional to the amount of that mRNA species in the sample assayed (Figure I). A compilation of the most common protocols to conduct microarrays experiments can be found at http://www.microarrays.org/index.html.

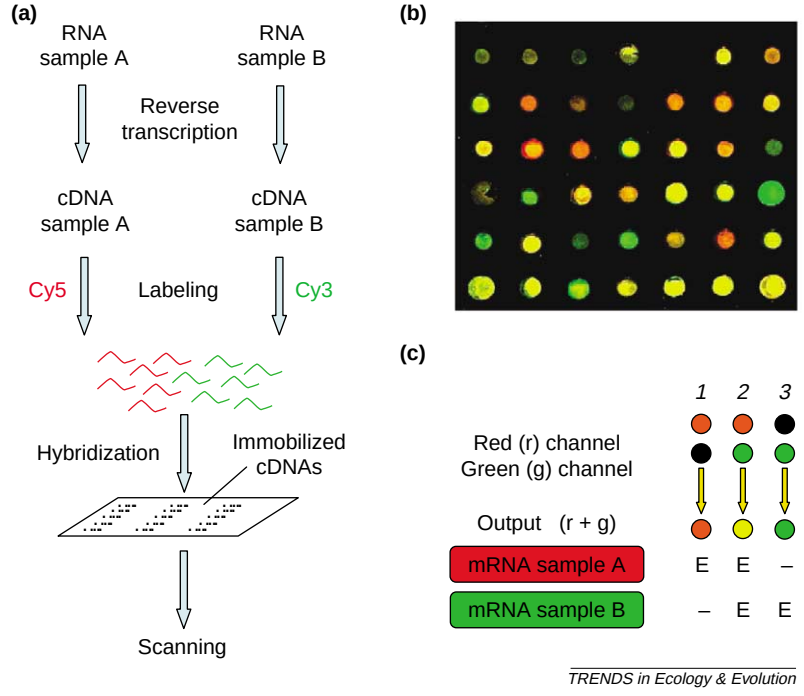

Figure I. Experimental principles of a two-color experiment using cDNA microarrays. **(a)** The level of expression of ideally the whole gene complement of a particular organism in two samples of interest is to be estimated. To this end, the two samples under study, which can comprise either total RNA or mRNA, must be reverse-transcribed into cDNA, fluorescently labeled with two dyes that differ in their absorption and emission wavelengths, and forced to undergo competitive hybridization with their respective reporters. In this case, the reporters are a collection of cDNAs that have been spotted and immobilized at high density onto a glass surface. The two different types of fluorescence are detected with a commercial confocal laser. **(b)** Typical visualization of a competitive fluorescent hybridization onto a cDNA microarray. A partial region of a microarray from [21] is shown. Each spot represents a different gene of *D. melanogaster*. The different colors and intensities at each spot are the result of similar (yellow) or different (red or green) gene expression levels in the two samples compared. **(c)** Interpretation of the outcome for three genes (E, expressed). In the example, this single experiment shows evidence of lack of expression of gene 1 in sample A and gene 3 in sample B. Gene 2 appears to be expressed equally in both samples. In a real experiment, inferring significant differences in mRNA abundance requires: (i) performing multiple replicates; (ii) incorporating all the relevant sources of variation; and (iii) analyzing the data within a suitable statistical framework. Validation of the results with other techniques, such as northern blot or quantitative real-time PCR, is also advisable.

that patterns of transcriptome variation are largely a consequence of natural selection, with stabilizing selection having a major role in constraining transcriptome divergence.

Intraspecific transcriptome variation

In their earliest application, microarrays were used to compare transcription profiles (see Glossary) of samples from cells or tissues from individuals of the same species differing in a given phenotype of interest (e.g. different tissue or cell type, infection by a pathogen, or response to drugs). This technique was then adopted to survey levels of gene expression variation in natural populations, showing that there is significant intraspecific variation in transcript abundance for a large fraction of the genome [3–9]. For example, in *Drosophila*, at least 10% of

Glossary

Additive genetic model: standard quantitative genetic model in which the underlying genetic basis of a continuous phenotypic trait is constituted by many genes of small effect, and the phenotypic value of the trait is the result of the sum of the small effects of each of the genes. Because transcription level is a molecular phenotype with a continuous range of values, there is interest in determining whether transcriptional variation is mostly due to additive genetic effects.

***cis*-Acting changes:** changes in the sequence of regulatory elements that influence the transcript abundance of a nearby gene by affecting its mode of transcription or the stability of its transcript.

Epistasis: in the context of gene expression, represents the interaction between alleles of two or more genes that has, as a consequence, altered one or more gene expression aspects (e.g. abundance or timing) in at least one of the genes involved.

Haldane's rule: the tendency of hybrid sterility and inviability to first appear in the heterogametic sex of interspecific hybrids.

Mutation accumulation line: one of the multiple highly inbred lines derived from the same group of genetically homogeneous individuals. Owing to the small population size of each line, natural or artificial selection is relaxed, enabling mutations to accumulate at a rate similar to the spontaneous mutation rate.

Mutation–drift equilibrium model: genetic model in which, in the absence of selection, every generation new genetic variants are introduced in the population by mutation, whereas other preexisting variants are eliminated by random genetic drift. Over time, the magnitude of the newly-introduced genetic variation and the magnitude of eliminated genetic variation become similar, thus reaching an equilibrium.

Pleiotropy: condition whereby a mutation on a particular gene has effects on several phenotypic traits of the organism. These traits can present different degrees of dependence or relationship. At the level of regulation of gene expression, pleiotropy can be seen, for example, as the widespread effect that a change in the concentration of transcription factor or a non-functionally equivalent amino acid change in its DNA binding domain can have downstream in the transcriptional program of the organism.

Retroposition: mechanism whereby a processed mRNA (i.e. an mRNA whose introns have been spliced out) is reverse-transcribed into cDNA and then inserted elsewhere in the genome. The new gene copy must either recruit adjacent or *de novo* evolved regulatory sequences to avoid becoming non-functional. This mechanism is important for the diversification of gene functions over time.

***trans*-Acting changes:** genetic changes that affect the expression of distant genes. An amino acid replacement in the DNA binding domain of a transcription factor, or a change that modifies the level of expression of one of the necessary cofactors in the transcriptional complex are examples of *trans*-acting changes.

Transcription factory: discrete sites in the nucleus where genes that need to be transcribed and the different components of the transcriptional machinery concentrate.

Transcription profile: characterization of the set of genes and their levels of expression on a genomic scale in a particular group of cells, at one determined moment, and under specific experimental conditions.

the surveyed fraction of the genome shows significant variation among genotypes [4,9–11]. In the most extreme case, significant differences in 94% of the genes among individuals of the same population of killifish *Fundulus heteroclitus* were observed when gene expression was surveyed in heart tissue [12].

In several cases, genome-wide transcript variation is correlated with phenotypic differences among the surveyed strains [4] or populations [5]. Importantly, several studies in a few model species show that this variation is heritable [7,8,10,11,13,14]. One study in *Drosophila melanogaster* showed that a large fraction of gene expression variation has a non-additive genetic component [10], posing important questions about the use of additive genetic models to model phenotypic evolution.

The levels of intraspecific gene expression polymorphism appear to be greater than those observed for proteins or DNA sequences. The main reason seems to be the epistatic and pleiotropic nature of the molecular mechanisms underlying gene expression. For instance, genes involved in the same pathway will show correlated responses at the transcriptional level, and some genes can participate in more than one pathway. Therefore, only a few regulatory changes can affect the patterns of gene expression of many different genes, which might represent a considerable fraction of the transcriptome. This has been well documented in experiments of artificial selection on odor-guided behaviour and mating speed in *D. melanogaster* [15,16].

Tempo and mode of transcriptome evolution

The logical extension from intraspecific to interspecific gene expression surveys has enabled us to obtain a better picture of the temporal dynamics of change that the transcriptome experiences. The main limitation that has accompanied the comparison of expression profiles from different species is that few microarray platforms have been developed (mostly for model organisms). For this reason, the same microarray platform has been used to compare expression profiles of closely or distantly related species. In this type of experiment, potential sequence mismatches between the probe and its corresponding reporter can appear as artifactual gene expression changes. Thus, different measures have been proposed to account for this type of bias: (i) consider only genes for which all their different probes on the oligonucleotide array provide consistent hybridization results regardless of the species [9,17,18]; (ii) discard those genes for which there is not a perfect match in the sequence of the species compared [19]; (iii) avoid directly hybridizing the RNA of different species [20]; and (iv) carry out competitive hybridizations using genomic DNA to perform a genome-wide assessment of the discordance in hybridization efficiency between the samples, and according to that eliminate highly diverged genes from the analyses and/or decrease hybridization temperatures to ameliorate the effect [21].

Here, we review recent microarray studies that have shed light on the issue of how changes in gene expression accumulate over time, on the role that different evolutionary mechanisms have had in driving transcriptome evolution, on the importance of sex and developmental stage in transcriptome divergence, and on the link between transcriptome evolution and the formation of new species.

Rates of transcriptome divergence

As with DNA sequences, an obvious question to address using comparative microarray data is whether transcriptome divergence increases linearly with time. Analyses of the magnitude of gene expression divergence among *Drosophila yakuba*, *Drosophila simulans* and four strains of *D. melanogaster* during early metamorphosis [20] indicated that 27% of the genes assayed exhibited significant changes in expression between at least two strains or species, and the magnitude of change was in good agreement with the phylogenetic relationships of the compared lineages. Interestingly, genes acting at the top of the transcriptional hierarchy (i.e. transcription factors and signal transducers) during the analyzed developmental transition were less prone to evolve changes in

their mRNA abundance than were genes encoding enzymes and structural proteins [20]. This is a trend that is consistent with the higher constraint expected to be imposed on genes that have crucial functions early in ontogeny, as well as on those encoding proteins with widespread pleiotropic effects, as also observed in *C. elegans* [22].

A similar increase in changes in gene expression with time has been found after comparing the expression profile of the prefrontal cortex of *Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus* and *Macaca mulatta* [19], and in reanalyses [17,23] of microarray data [6] from the left prefrontal lobe of the first three species. Interestingly, the rate of change in gene expression appears to be accelerated in the human lineage as compared to the chimpanzee lineage, which could be due to an increase in the level of gene expression in humans [6,17,23]. These comparisons among primates have been encouraged by the idea that cognitive and behavioral differences between humans and our closest relatives are associated mainly with differences in gene expression in the brain rather than to differences in the coding sequences themselves. A quantification of interspecific differences in gene expression in different organs indicates that liver [6,17,18,23] and heart [18] exhibit more expression differences than does the brain. However, unlike the differences in non-neural tissues, which have become fixed at approximately the same rate between the human and chimpanzee lineages, involving a similar proportion of up-regulated and down-regulated genes, the interspecific differences in the brain have occurred at a higher rate in the human lineage, usually involving increased rather than decreased expression [17,18,23].

Comparison of the gene expression profile among primates [6,17,18,23] suggests that a simple metric of phenotypic differentiation could be developed from microarray data. However, distances based on percents of differentially expressed genes have to be interpreted with caution owing to epistasis and pleiotropy, which influence the expression of multiple genes in the genome. Future knowledge of the precise interactions among genes within the expression network will help distinguish between independent and correlated transcriptional changes, enabling a more precise assessment of the degree of differentiation.

Mechanisms governing transcriptome divergence

What fraction of the observed transcriptome divergence is due to natural selection? This question has been addressed using different approaches, leading to inconsistent results across different taxa. Rifkin *et al.* [20] developed a quantitative method for addressing this question using microarray data, with additional analytical approaches suggested by Khaitovich *et al.* [19] and Nuzhdin *et al.* [9]. The basic rationale of these methods is based on earlier theoretical studies of phenotypic divergence [24,25]. The expectation is that genes with low intraspecific variation in expression and low divergence between species should be under stabilizing selection; genes under directional selection should show little intraspecific variation but large interspecific divergence; and genes under balancing selection should have large intraspecific variation but low interspecific divergence.

Studies of the *D. melanogaster* subgroup [20] suggested that, in a group of 6742 genes that changed during development, 67% are conserved and thus have low mutational variance or are under strong stabilizing selection; 22% show evidence of lineage-specific selection, and only 7% show patterns consistent with a mutation–drift equilibrium model. Using an identical approach, the reanalysis of data on expression profiles in the brain among primates [17] indicated that only a small fraction of the assayed genes have a interspecific:intraspecific divergence ratio that is consistent with the action of positive selection, although results were extremely sensitive to parameter values used in the model.

In *C. elegans*, the comparison between the transcriptional mutational variance and the transcriptional genetic variance is also consistent with the pervasive role of stabilizing selection on gene expression levels [22] (Box 3). Different implementations of neutral models of phenotypic evolution are also in good agreement with this notion [26].

However, a different analytical approach in which expression profiles of expressed pseudogenes are used to detect departures from a pure neutral model suggests that most evolutionary changes in the primate brain transcriptome are neutral [19]. A controversial point of this study, however, is the use of expressed pseudogenes as indicators of neutral patterns of gene expression evolution. This is because several expressed pseudogenes show patterns of evolution that are atypical of neutral sequences (e.g. codon usage bias and low rates of change in nonsynonymous sites) and are also involved in the regulation of gene expression of their functional paralogs [27,28]. More refined models will help to more accurately evaluate the relative importance of different evolutionary mechanisms in shaping the evolution of the transcriptome.

Sex-biased gene expression

Coding sequences and morphological characters related to sex and reproduction appear to evolve faster than those that are primarily devoted to survival, an observation that is in agreement with sexual selection theories and with the idea of two gene pools in the genome that evolve at different rates [29,30]. The prediction that gene expression, as well as many other characters, can have a different optimal level depending on the sex of an individual has been confirmed for many genes in *Drosophila* [4,31]. Most transcriptome changes (83%) detected between *D. melanogaster* and *D. simulans* species are observed in genes with sex-biased expression [21]. Furthermore, genes with male-biased expression displayed the largest differences when compared with female- and non-sex-biased genes. Subsequent comparison of the expression profile in males of eight strains of *D. melanogaster* [32] showed that male-biased genes were also over-represented among those genes that exhibit intraspecific changes in the level of expression. Among genes with somatic expression, genes with sex-biased expression appear to evolve faster than those with monomorphic expression [32].

Box 3. A case study in transcriptome evolution

Denver *et al.* [22] recently addressed the relative importance of mutation and natural selection in shaping transcriptome evolution in *C. elegans*. Transcript profiles were compared among five natural strains (NI) and four mutation accumulation (MA) lines that had been propagated for 280 generations. Selection is relaxed in MA lines, thus enabling the fixation of most mutations (except those that are severely deleterious). When genome-wide levels of gene expression were surveyed, Denver and colleagues found significant differences in transcript level in 9% and 2% of the genes among the MA and NI lines, respectively. This difference is remarkable considering that the NI lines have been separated for thousands of generations, and suggests the importance of stabilizing selection in shaping transcriptome variation in *C. elegans*.

Denver and colleagues used a co-expression map of *C. elegans* [68] to infer the effect of *trans*-acting mutations on the observed transcriptional differences (Figure I). Co-expression maps are based on the notion that genes with correlated expression across manifold experimental conditions are likely to be co-regulated and contribute jointly to particular functions. Of the differentially expressed genes among the MA lines, 67.7% (447/660) were associated with the seven co-expression gene clusters (mounts) that contained a significant overrepresentation of differentially expressed genes. However only 31.4% (37/118) of the genes were associated with four co-expression clusters in the NI lines. This significant difference between the MA and NI lines, and the tendency of the genes of particular mounts (especially number 4 and 8) of having expression changes in the same direction within a particular line upholds the idea that most of the observed changes in the MA lines can be the result of mutations with *trans*-acting effects. One can thus infer that correlated expression changes in the NI lines are less common because *trans*-acting changes are eliminated by stabilizing selection. Comparisons of these data to expectations from neutral models also suggest that stabilizing selection is the major force driving transcriptome evolution in this species.

Interestingly, sperm genes were over-represented among the differentially expressed genes in the MA lines (266 out of 447), and were all located in a single expression mount (number 4). This result might indicate that the large variation in gene expression observed in reproductive-related genes in other studies [21,32,43] is largely the result of a part of the expression network that is more sensitive to *trans*-acting changes.

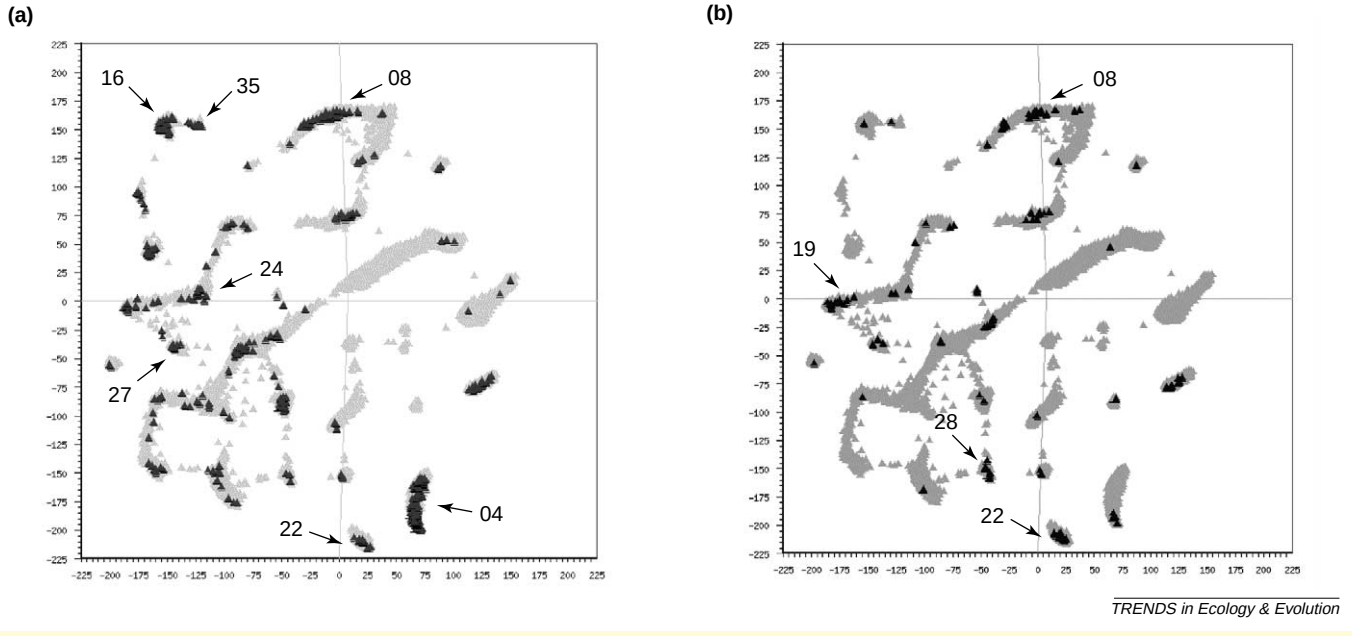

Figure I. Interplay between the co-expression map of *C. elegans* and the differential expression profile among MA lines **(a)** and NI lines **(b)**. The gray areas, which are called mounts, represent zones of the co-expression map of *C. elegans* with a high number of genes related throughout their expression profile within specific biological processes. Denver *et al.* [22] inspected how the differentially expressed genes (black triangles) in each type of line are related to the co-expression map of *C. elegans*. Mounts with statistically significant enrichment for differentially expressed genes are indicated with a number, and the x–y dimensions are arbitrary and simply facilitate to spatially organize groups of genes that are co-expressed. Co-expression mounts are related with: sperm genes (mount 4); intestine genes (mount 8); muscle and collagen genes (mount 16); amino acid and lipid metabolism genes and cytochrome P450 genes (mount 19); collagen genes (mount 22); amino acid and lipid metabolism (mount 24); amino acid metabolism genes (mount 27); unknown function (mount 28); and a different group of collagen genes (mount 35). Reproduced with permission from [22].

Consistent with observations reported elsewhere [31], the X chromosome of *D. melanogaster* and *D. simulans* was found to be depleted of male-biased genes and enriched with female-biased genes [21] (Figure 1). This unequal genomic distribution of genes with sex-biased expression has also been documented in worms and mice [33,34]. Proposed explanations include the different time that a X chromosome will spend in males (1/3) versus females (2/3), the meiotic X inactivation during early male germ line development, and the different probability of fixation of sexually antagonistic alleles on the X chromosome and autosomes depending on their degree of dominance [35,36].

One of the molecular mechanisms underlying this striking distribution of sex-biased genes is the preferential retroposition from the X chromosome to the autosomes of genes whose activity is of relevance during male spermatogenesis [37]. This pattern appears to be common in mammals and *Drosophila* [38,39]. Overall, the non-random distribution of sex-biased genes among the X chromosome and the autosomes represents one of the ways in which genes seem to redistribute across the genome based on their expression profile (Box 4). Nevertheless, these general patterns can hide more subtle tendencies. For instance, genes expressed early in mice spermatogenesis, when germinal cells are still

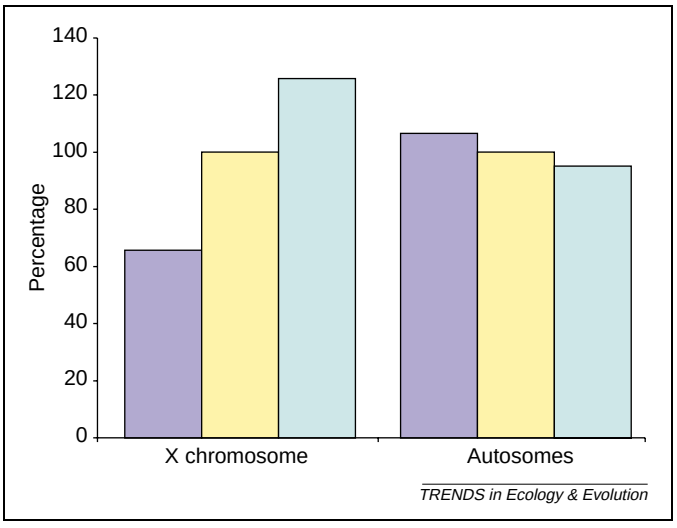

Figure 1. Relative proportion of genes with sex-biased expression in relation to the random expectation in the X chromosome and autosomes of two *Drosophila* species [21]. Blue bars refer to genes that are preferentially expressed in males compared with females; green bars indicate genes that are preferentially expressed in females compared with males. According to a random distribution across the genome, both male- and female-biased genes in expression should be equally represented on the X chromosome and the autosomes. Over-representation and under-representation of male and female-biased genes on the X and autosomes in relation to that random genomic distribution appear in the chart as departures from a level of 100 (yellow bars). For the genes present on the array, reliable measures of the level of expression were obtained for 4700 genes, 753 of which were located on the X chromosome and 3947 on the autosomes. Differences in gene-expression between the sexes were statistically significant for 2090 genes at $P<0.01$ for both *D. melanogaster* and *D. simulans*. For the X chromosome, 83 and 262 genes were deemed male-biased and female-biased, respectively. For the autosomes, 706 and 1039 genes were deemed male-biased and female-biased, respectively. Male-biased genes in expression are depleted on the X chromosome and enriched on the autosomes ($G_{adj}=19.76$; 1 d.f.; $P=8.8\times10^{-6}$), whereas female-biased genes exhibit the opposite pattern ($G_{adj}=15.38$; 1 d.f.; $P=8.8\times10^{-5}$). Another analysis, where a larger fraction of the *Drosophila* genome was interrogated with the same purpose, found identical results for *D. melanogaster* [31]. Data from [21].

diploid and mitotic, appear to be in excess in the X chromosome [34,40].

Transcriptome divergence and speciation

Understanding the genetic basis of speciation has been one of the most crucial and elusive goals in evolutionary genetics. After more than 70 years of research in this area, only three genes in *Drosophila* and one in the platyfish *Xiphophorus* have been shown to be involved in the fitness reduction of interspecific hybrids [41]. Microarray technology can provide a means to identify additional candidate genes based on the expectation that gene regulatory incompatibilities might be the cause of lower fitness in species hybrids [42], as exemplified by two recent microarray studies in *Drosophila* [43,44].

A study of gene expression dysfunction in *D. simulans*–*D. mauritiana* hybrid sterile males identified a suite of loci primarily related to spermatogenesis, suggesting the potential involvement of these genes in the male sterility phenotype [43]. Follow-up experiments for five of those genes discarded a potentially spurious link between the candidate genes and the observed phenotype of sterility, and mapped the upstream factor in the transcriptional hierarchy responsible for the coordinated regulation of the candidate genes [45]. A second study in a more distant species pair (*D. melanogaster*–*D. simulans*) [44] showed that 69% of the assayed transcripts appeared to be either over- or under-represented in the hybrid females, with a large underexpression of genes with a female-biased pattern of expression accompanying gonadal atrophy. The study also showed that genes preferentially expressed in males relative to females are more commonly misexpressed than are genes with no sex bias, a pattern that might indicate the faster evolution of mechanisms that regulate male-biased genes in expression when present in a female background. These results [43,44], and the patterns of variation of gene expression in males within and between species [21,32] provide some support to the faster-male theory, one of the proposed causes of Haldane's rule [41].

However, novel patterns of gene expression in hybrids are not always associated with fitness reduction as epitomized by those observed after polyploidization events in plants [46]. Furthermore, first steps have been taken to begin to understand intraspecific transcriptome variation in relation to other important stages of reproductive isolation, such as mating success, among genotypically distinct *D. melanogaster* males in competition experiments [47].

Box 4. Non-random genomic location of co-expressed genes in eukaryotes

Clear patterns of non-random gene organization as a function of their expression profile have emerged in eukaryotes. Such clustering of co-expressed genes in the same genomic region can be due to organization of genes in operons (only reported in nematodes and trypanosomes) or to the more common local grouping of genes with similar expression profiles [69]. Transcriptome analyses of 12 human tissue types [70] revealed the existence of alternated genomic regions with elevated and reduced levels of expression. In *D. melanogaster*, genes expressed in testes, head and embryo are usually organized in groups of three or more on the autosomes [71]. This phenomenon of gene clustering could be explained by the presence of regional enhancers controlling the expression of multiple genes [72], in agreement with a gene organization that facilitates a coordinated type of expression [73], or by the indirect effect of the local opening of the chromatin during transcription [74].

Could this non-random gene clustering be the result of natural selection acting on genome organization? The fact that in two yeast species highly co-expressed genes are physically close at twice the average rate, or that clusters of co-expressed genes in human and mouse tend to accumulate fewer chromosomal breakpoints than the random expectation, have been interpreted as evidence of the role of selection in preventing their separation [75,76].

Nevertheless, a generalized role of selection to explain gene clustering should be treated with caution. First, it has been demonstrated that genes physically apart in the genome can co-localize in the same transcription factory and, therefore, the facilitating role of clustering might not be that important [77]. Second, there is evidence that clustering of co-expressed genes is preferentially lineage specific. For instance, clustering of metabolic genes is observed in *S. cerevisiae* but not in *D. melanogaster* [78]. Furthermore, the comparison of the identity of genes included in regions of correlated expression patterns of humans and mouse reveals that after correcting for duplication events, most of those regions are not conserved in gene composition, which might reflect the physiological differences of both species [79]. Even in the case of identical clustering between the lineages compared, it is complex to distinguish between phylogenetic inertia and actual selective forces. Engineered disruptions of clusters with correlated expression should help distinguish between those two scenarios.

The genetic basis of gene expression differences

Although microarrays provide information about genes that are differentially expressed, they do not provide information about the actual genetic changes that are responsible for that variation. Therefore, there is an increasing interest in understanding the genetic architecture of gene expression variation on a genomic scale using microarrays and high-throughput genotyping methods [48]. Given that mRNA expression level is a molecular phenotype, classic mapping and quantitative genetic approaches are being used to elucidate its genetic basis. We currently know that a substantial fraction of the variation in gene expression has a heritable genetic component with different contributions of *cis*-acting and *trans*-acting changes [7,8,13,14,49].

Expression QTLs (eQTLs)

Genome-wide linkage analysis and microarray experiments of segregating populations have been combined to map genetic regions that are linked to gene expression phenotypes in several model organisms [7,8,14]. This approach, originally termed 'genetical genomics' [48], consists of conducting standard linkage analyses on genome-wide expression profiles. Hundreds to thousands of markers distributed across the genome are used for quantitative trait loci (QTL) mapping, and the expression profile is also determined for hundreds of individuals using microarrays. Linkage analyses are then conducted for each gene in the microarray that shows variation in expression levels or for any suite of genes of interest. In one example [8], eQTLs were mapped using >100 immortalized human B cells lines and 2756 SNPs. eQTLs were mapped for 984 genes, most of which were *trans*-acting and very few were *cis* acting. The authors found genomic regions containing 'master regulators' of transcription that influenced the expression of multiple genes. However, this apparently high density of eQTLs in particular genomic regions could be more a statistical artifact than a fundamental pattern of biological organization, because differences in the power of QTL detection, which are associated with variation in recombination rates across the genome, were not taken into account. In addition, differences in sample size, microarray platform and experimental design can influence the power of QTL detection among different studies. Overall, and in spite its appeal, the generalized use of this approach in evolutionary biology is some way off, given the large expense involved. Alternatively, microarrays can complement QTL analyses to identify candidate genes for complex traits in the large genomic regions pinpointed by QTL mapping, as exemplified by a study in *D. melanogaster* [50].

cis *versus* trans *regulatory changes*

The relative importance of *cis* or *trans* regulatory variation continues to vex researchers and different approaches have been developed to assess it either on a genomic scale [7,8,13,14,49] or for large numbers of genes [51,52]. Experiments performed in yeast showed that 100–200 *trans*-acting loci controlled the variation in gene expression of up to 1716 genes [14]. Surprisingly, the identified loci were not only transcription factors, but were also comprised of multiple functional classes of genes that were pervasively distributed across regulatory networks. Other studies in humans [8,49], flies [11] and worms [22] also found a preponderance of *trans*-acting effects in gene expression across the genome. Interestingly, contrary evidence supporting a role of *cis*-acting factors has been obtained in flies [52], human [53], mice [7,51], and maize [7]. At least in the case of genome-wide association studies, these contradictory results can be explained by the different statistical cutoffs at which eQTLs are detected. At less stringent cutoffs, *trans*-acting factors surpass *cis*-acting factors, whereas the opposite pattern is found at more stringent cutoffs. This is possibly the result of *trans*-acting factors that have a moderate subtle effect on large sets of genes, as expected from the pleiotropic nature of *trans*-acting changes [7].

cis-Acting and *trans*-acting changes do not necessarily alter gene expression [54], but when they do, they can have major effects on fitness. For instance, abnormal gene expression accompanies some diseases in humans [53] whereas in other cases, new patterns of gene expression appear to be associated with functional or morphological innovations during evolution. For example, the loss of expression of the gene *Pitx1* in the pelvic region and caudal fin is coupled with the pelvic reduction in the fish *Gasterosteus aculeatus* [55] and a new pattern of expression of the gene *yellow* is involved in differences of wing pigmentation among related *Drosophila* species [56].

Concluding remarks and future directions

As a result of the advent of microarray technology, evolutionary genetics has redirected part of its attention to analyses of genome-wide expression profiles. Multiple studies support natural selection as the main mechanism governing transcriptome variation and evolution. Furthermore, as observed for other sex-related traits, the fraction of the transcriptome that is more closely related to sex and reproduction has evolved rapidly. The epistatic and pleiotropic nature of the molecular mechanisms underlying gene expression is largely responsible for the variation in the level of transcript abundance that has been observed within and between species.

Microarray technology holds great promise for providing a different and more efficient avenue for identifying genes in non-model organisms that affect their ecological and evolutionary success [5]. The imminent completion of multiple genome sequences will also expand the phylogenetic range in which the changes in gene expression that have accompanied the process of species diversification can be investigated.

Although remarkable advances have occurred in the statistical analysis of genomic data, new theoretical approaches will be necessary to clarify the relative importance of the diverse evolutionary mechanisms that govern evolution at the level of gene expression as well as in connection to central biological processes, such as speciation. In addition, evolutionary geneticists will have to pay attention to aspects of the regulation of gene expression, such as RNA stability and its translational control by specific RNA molecules [57,58], as well as to the contribution of multiple alleles to the observed variation

in gene expression in diploid organisms [59]. Incorporating all these new levels of information in comparative studies of transcriptome evolution on a genomic scale will be challenging but fundamental if we aim to achieve a more complete view of the evolutionary process at the regulatory level.

Acknowledgements

We apologize to all colleagues whose work has not been cited because of space limitations. We thank Sergey Nuzhdin, Luciano Matzkin and three referees for constructive comments, and Dee Denver for allowing us to reproduce his graphical material. J.M.R. is funded by a long-term EMBO fellowship and a BBSRC grant (BBS/B/07705). C.A.M. is funded by start-up research funds from the University of Arizona and by the National Science Foundation (DEB-0108475).

References

1 Carroll, S.B. *et al*. (2001) *From DNA to Diversity*, Blackwell Science
2 Wilkins, A.S. (2002) *The Evolution of Developmental Pathways*, Sinauer Associates
3 Cavalieri, D. *et al*. (2000) Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12369–12374
4 Jin, W. *et al*. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* 29, 389–395
5 Oleksiak, M.F. *et al*. (2002) Variation in gene expression within and among natural populations. *Nat. Genet.* 32, 261–266
6 Enard, W. *et al*. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science* 296, 340–343
7 Schadt, E.E. *et al*. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302
8 Morley, M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747
9 Nuzhdin, S.V. *et al*. (2004) Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol. Biol. Evol.* 21, 1308–1317
10 Gibson, G. *et al*. (2004) Extensive sex-specific non-additivity of gene expression in *Drosophila melanogaster*. *Genetics* 167, 1791–1799
11 Wayne, M.L. *et al*. (2004) Additivity and trans-acting effects on gene expression in male *Drosophila simulans*. *Genetics* 168, 1413–1420
12 Oleksiak, M.F. *et al*. (2005) Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. *Nat. Genet.* 37, 67–72
13 Cheung, V.G. *et al*. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33, 422–425
14 Yvert, G. *et al*. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* 35, 57–64
15 Anholt, R.R. *et al*. (2003) The genetic architecture of odor-guided behavior in *Drosophila*: epistasis and the transcriptome. *Nat. Genet.* 35, 180–184
16 Mackay, T.F. *et al*. (2005) Genetics and genomics of *Drosophila* mating behavior. *Proc. Natl. Acad. Sci. U. S. A.* 102 (Suppl 1), 6622–6629
17 Hsieh, W.P. *et al*. (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165, 747–757
18 Caceres, M. *et al*. (2003) Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13030–13035
19 Khaitovich, P. *et al*. (2004) A neutral model of transcriptome evolution. *PLoS Biol.* 2, E132
20 Rifkin, S.A. *et al*. (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 33, 138–144
21 Ranz, J.M. *et al*. (2003) Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* 300, 1742–1745
22 Denver, D.R. *et al*. (2005) The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat. Genet.* 37, 544–548
23 Gu, J. and Gu, X. (2003) Induced gene expression in human brain after the split from chimpanzee. *Trends Genet.* 19, 63–65
24 Lande, R. (1976) Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30, 314–334
25 Lynch, M. and Hill, W.G. (1986) Phenotypic evolution by neutral mutation. *Evolution* 40, 915–935
26 Lemos, B. *et al*. (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* 59, 126–137
27 Kylsten, P. *et al*. (1990) The cecropin locus in *Drosophila*; a compact gene cluster involved in the response to infection. *EMBO J.* 9, 217–224
28 Korneev, S.A. *et al*. (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.* 19, 7711–7720
29 Carson, H.L. (1985) Unification of speciation theory in plants and animals. *Syst. Bot.* 10, 380–390
30 Singh, R.S. (2000) Toward a unified theory of speciation. In *Evolutionary Genetics from Molecules to Morphology* (Singh, R.S. and Krimbas, C.B., eds), pp. 570–604, Cambridge University Press
31 Parisi, M. *et al*. (2003) Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science* 299, 697–700
32 Meiklejohn, C.D. *et al*. (2003) Rapid evolution of male-biased gene expression in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9894–9899
33 Reinke, V. *et al*. (2000) A global profile of germline gene expression in *C. elegans*. *Mol. Cell* 6, 605–616
34 Khil, P.P. *et al*. (2004) The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat. Genet.* 36, 642–646
35 Rogers, D.W. *et al*. (2003) Male genes: X-pelled or X-cluded? *Bioessays* 25, 739–741
36 Oliver, B. and Parisi, M. (2004) Battle of the Xs. *Bioessays* 26, 543–548
37 Hendriksen, P.J. (1999) Do X and Y spermatozoa differ in proteins? *Theriogenology* 52, 1295–1307
38 Betran, E. *et al*. (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12, 1854–1859
39 Emerson, J.J. *et al*. (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303, 537–540
40 Wang, P.J. *et al*. (2001) An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* 27, 422–426
41 Coyne, J.A. and Orr, H.A. (2004) *Speciation*, Sinauer Associates
42 Johnson, N.A. and Porter, A.H. (2000) Rapid speciation via parallel, directional selection on regulatory genetic pathways. *J. Theor. Biol.* 205, 527–542
43 Michalak, P. and Noor, M.A. (2003) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. *Mol. Biol. Evol.* 20, 1070–1076
44 Ranz, J.M. *et al*. (2004) Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* 14, 373–379
45 Michalak, P. and Noor, M.A. (2004) Association of misexpression with sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. *J. Mol. Evol.* 59, 277–282
46 Adams, K.L. *et al*. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4649–4654
47 Drnevich, J.M. *et al*. (2004) Quantitative evolutionary genomics: differential gene expression and male reproductive success in *Drosophila melanogaster*. *Proc. Biol. Sci.* 271, 2267–2273
48 Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.* 17, 388–391
49 Monks, S.A. *et al*. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* 75, 1094–1105
50 Wayne, M.L. and McIntyre, L.M. (2002) Combining mapping and arraying: An approach to candidate gene identification. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14903–14906
51 Cowles, C.R. *et al*. (2002) Detection of regulatory variation in mouse genes. *Nat. Genet.* 32, 432–437
52 Wittkopp, P.J. *et al*. (2004) Evolutionary changes *in cis* and trans gene regulation. *Nature* 430, 85–88
53 Yan, H. *et al*. (2002) Allelic variation in human gene expression. *Science* 297, 1143
54 Ludwig, M.Z. *et al*. (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567

55 Shapiro, M.D. *et al*. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428, 717–723
56 Gompel, N. *et al*. (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433, 481–487
57 Yang, E. *et al*. (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* 13, 1863–1872
58 Mattick, J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet.* 5, 316–323
59 Lo, H.S. *et al*. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res.* 13, 1855–1862
60 Rockman, M.V. *et al*. (2004) Positive selection on MMP3 regulation has shaped heart disease risk. *Curr. Biol.* 14, 1531–1539
61 Carroll, S.B. (2005) Evolution at two levels: on genes and form. *PLoS Biol.* 3, e245
62 Wilson, A.C. *et al*. (1974) Two types of molecular evolution. Evidence from studies of interspecific hybridization. *Proc. Natl. Acad. Sci. U. S. A.* 71, 2843–2847
63 Velculescu, V.E. *et al*. (1995) Serial analysis of gene expression. *Science* 270, 484–487
64 Adams, M.D. *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656
65 Diatchenko, L. *et al*. (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. U. S. A.* 93, 6025–6030
66 Liang, P. and Pardee, A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 257, 967–971
67 Schena, M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
68 Kim, S.K. *et al*. (2001) A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087–2092
69 Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* 20, 480–487
70 Caron, H. *et al*. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292
71 Boutanaev, A.M. *et al*. (2002) Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* 420, 666–669
72 Li, Q. *et al*. (2002) Locus control regions. *Blood* 100, 3077–3086
73 Kosak, S.T. and Groudine, M. (2004) Form follows function: the genomic organization of cellular differentiation. *Genes Dev.* 18, 1371–1384
74 Spellman, P.T. and Rubin, G.M. (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* 1, 5
75 Hurst, L.D. *et al*. (2002) Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* 18, 604–606
76 Singer, G.A. *et al*. (2005) Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* 22, 767–775
77 Osborne, C.S. *et al*. (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* 36, 1065–1071
78 Lee, J.M. and Sonnhammer, E.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13, 875–882
79 Su, A.I. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067