Available online at www.sciencedirect.com

SCIENCE DIRECT

GENOMICS

Genomics 84 (2004) 449–457

www.elsevier.com/locate/ygeno

Review

The genomics of gene expression

John A. Stamatoyannopoulos*

Department of Molecular Biology, Regulome, Canal View Building, 551 N. 34th Street, Seattle, WA, 98103, USA

Received 6 May 2004; accepted 7 May 2004
Available online 2 July 2004

Abstract

The study of gene regulation on a genomic scale has been constrained by the modest pace with which new *trans*-regulatory factors have been identified and by the fact that *cis*-regulatory sequences have to date been described even in part for only a small fraction of vertebrate genes. An indirect approach for assessing the significance of *cis*- and *trans*-regulatory mechanisms on a global scale is to utilize gene expression as a surrogate for transcriptional regulation and to combine genome-scale transcriptional profiling with studies of genetic variation, classical genetic techniques such as linkage analysis, and examination of allelic expression patterns that reveal *cis*-regulatory variability. A number of recent studies employing these methods provide insight into questions of central importance to our understanding of the larger role of transcriptional regulation in the organization of the human and other complex genomes.

The study of gene regulation in complex organisms has been carried out largely at the level of individual genes or, less frequently, clusters of related genes. The classical paradigm relies heavily on the identification and analysis of *cis*-regulatory sequences, which play a determinative role in the tissue and developmental specificity of gene expression. Such sequences may be further decomposed to reveal sequence motifs that signal the role of specific *trans*-acting factors, thus providing insight into networks of potentially coregulated genes [1,2]. However, the extensibility of this "bottom–up" paradigm to a genomic scale has been severely constrained by the modest pace with which new *trans*-regulatory factors have been identified and by the fact that *cis*-regulatory sequences have to date been described even in part for only a small fraction of vertebrate genes and more exhaustively for only a few. A variety of computational [3–7], phylogenetic [8–13], and molecular methods [14,15] have attempted to address this deficit, identification of regulatory sequences with high accuracy over genomic space remains an elusive goal.

An alternative "top–down" approach to the study of gene regulation on a genomic scale is to utilize gene expression as a surrogate for transcriptional regulation. Since measurement of the former may be influenced by factors relating to mRNA integrity or turnover, the correspondence is not perfect. However, by combining genome-scale transcriptional profiling with studies of genetic variation and classical genetic techniques such as linkage analysis [16], a number of recent studies have provided insight into questions of central importance to our understanding of the larger role of transcriptional regulation in the organization of the human and other complex genomes.

One such question concerns the biological variability of gene expression in the context of natural populations and in particular the extent to which interindividual variability in tissue-specific patterns of gene activity reflects heritable determinants of gene regulation versus the action of environmental factors. Another question relates to the interplay of cis-acting sequences and trans-acting factors in establishing patterns of expression, a topic of long-standing interest in the study of gene regulation. Since cis-acting variation has been proposed to be a major determinant of quantitatively varying traits, this question is of particular significance for our understanding of the genetic basis of common diseases.

This review examines these questions in the context of recent studies and discusses their implications for our understanding of the genomic architecture of gene regulation and for human disease.

* Fax: (206) 267-1094.
E-mail address: jstam@regulome.com.

0888-7543/$ - see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.ygeno.2004.05.002

Natural variation in gene expression

Evolution requires variation, and it has long been hypothesized that natural selection may be more dependent on variability in gene expression than on variation in protein coding sequences [17]. The following three related questions are considered below: (i) What proportion of genes exhibit natural variation in expression? (ii) How heritable are patterns of gene expression? And (iii) what proportion of such patterns can be accounted for by cis-acting versus trans-acting components?

Numerous recent surveys [18–30] now suggest that interindividual variability in gene expression is a regular feature of eukaryotic genomes, providing a rich substrate for evolutionary selection. Quantifying the extent of such variability and determining its source are of considerable importance for our understanding of the genetic basis of common diseases in which quantitative traits comprise defining features.

It is useful to divide studies of interindividual variation in gene expression into two classes according to their underlying methodologies and objectives. Studies in the first class [18–25] employ conventional microarray expression profiling. This approach provides a global view, though potentially at the expense of sensitivity in the context of nonhaploid genomes in which allelic (and hence *cis*-acting) variation in gene expression is necessarily averaged. The second class of studies [26–30] takes advantage of the large number of coding sequence variants now available to effect direct measurement of allelic expression patterns in a controlled fashion and thereby expose directly *cis*-acting effects on gene regulation.

Microarray analysis of individualized gene expression

Microarrays have been heavily utilized to analyze the expression patterns of large numbers of genes across different tissues or within the same tissue under a variety of experimental conditions or even between species. Several groups have now taken advantage of this platform to perform global analyses of the variability of gene expression between naturally occurring populations of the same species and between individuals within populations.

Differences between populations of the same species

One of the first studies to examine differences between related populations was carried out by Jin et al. in *Drosophila melanogaster* [18]. Of 3931 genes surveyed between two strains, sexes, and age time points, Jin et al. found genotype to be the major determinant of variability in expression for 267 (7%) genes, and estimated further that approximately 25% of the transcriptome was affected by genotypic factors in at least one sex. Sandberg et al. [19] found proportionally similar results in a comparison of brain expression profiles between two mouse strains, in which both regional and strain-specific differences were noted. An additional feature of this study was that some of the differentially expressed genes mapped to chromosomal regions that have been linked to complex quantitative phenotypes, including alcohol consumption and seizure susceptibility. In a study of budding yeast, Brem et al. [20] compared expression profiles of two strains of *Saccharomyces cerevisiae* using cDNA microarrays that contained more than 6000 open reading frames from the yeast genome. They found that 1528 genes were expressed differentially instead of the 23 genes expected by chance ($p < 0.005$).

The aforementioned studies examined strain-specific population differences versus differences between individuals within a population. Indeed, in the case of yeast and *Drosophila*, such measurements are not feasible for individual organisms, necessitating the use of pooled samples. However, truly individualized tissue samples are readily obtainable from larger organisms.

Differences between individuals

Several studies have now demonstrated more clearly the existence of interindividual variability in gene expression in natural populations. In a study of *Fundulus* populations, Oleksiak et al. [21] used an elegant experimental design and quantitative techniques (both adapted from Jin et al.) to reveal that approximately 18% of the genome is differentially expressed in individual fish grown under controlled environmental conditions. Schadt et al. [22] provide comparable figures for mouse, maize, and human genes. Of 23,574 murine genes profiled in hepatic tissue across two inbred strains, 7861 (33%) were found to be differentially expressed in the parental strains or among at least 11 F2 progeny (of which a total of 111 were examined). In a parallel survey of 76 F3 progeny constructed from two inbred lines of maize, 77% (18,805 of 24,473) of genes surveyed in ear leaf tissue were found to be differentially expressed across at least 10 individuals. When Schadt et al. examined 24,479 human genes in lymphoid cells, they found 2726 (11%) to be differentially expressed among 16 CEPH/Utah pedigree founders. These studies provide direct evidence for the prevalence of individualized gene expression patterns in a variety of organisms and implicate underlying genetic causes.

Results from another study in humans were more ambiguous. Whitney et al. [23] examined patterns of gene expression in whole blood collected from 75 healthy donors. Of ~18,000 genes, 600 (3.3%) varied by more than 2.5-fold between donors, but only 340 (1.9%) could be ascribed to genotypically related factors. These results should be interpreted with caution since this study was confounded by the intrinsic complexity of the tissue sample and various environmental variables related to its collection, resulting in limited power to detect individualized variation reliably.

Taken together, the aforementioned studies suggest that interindividual variability in expression of a large number of genes is a natural feature of higher eukaryotic genomes and likely of all eukaryotes. However, the mechanism(s) contributing to such variability need not be the same in each species.

Heritability of gene expression patterns

What proportion of genes differentially expressed between individuals can be attributed to heritable factors? Several studies have now addressed this question and provide direct evidence for a significant heritable component of individual variation in gene expression.

Brem et al. [20] took advantage of the yeast system to effect further dissection of observed interstrain variation in expression of 1528 genes. Testing the expression levels of these genes in 40 haploid segregants from a cross between the two parental strains revealed the established expression phenotypes to be highly heritable traits, with the median proportion of the genetic component of observed variation estimated to be 84%.

Following the paradigm of Brem et al., Schadt et al. [22] analyzed 40 individual descendents of 16 pedigree founders, revealing that 29% of 2726 genes differentially expressed among founders exhibited heritable expression phenotypes.

In another study of humans, Cheung et al. [24] used cDNA microarrays to examine variation in expression of 813 genes among 35 unrelated individuals. This analysis focused on only the top 5% variant genes, and the data provided do not permit estimation of the overall proportion of variant genes for a given variance ratio threshold. However, 5 genes exhibiting high interindividual variability were analyzed further using quantitative PCR in 49 unrelated individuals (including the 35 originally studied by microarray), 41 siblings, and 10 sets of monozygotic twins. Clear evidence for familial aggregation of expression phenotypes was found, with the greatest variability between unrelated individuals, intermediate variability among siblings, and only minor—but reproducibly detectable—variance between twins. The last finding in genetically identical individuals suggests an appreciable, though quantitatively less significant, role for epigenetic phenomena.

Collectively, the studies discussed above have established that a significant component of observed interindividual variability in gene expression is genetic, implying heritable variation in *cis*-acting sequences, *trans*-acting factors, or both.

Cis vs trans effects

A central question arising from the aforementioned studies relates to the relative contribution of *cis*-acting (i.e., gene-proximal) vs *trans*-acting determinants of heritable variation in gene expression. To a certain degree, these effects may not be clearly separable since many *trans* factors exert their gene-specific transcriptional effects through cognate *cis* sequences and vice versa. Preliminary insight into this question has been provided by combining expression analysis with classical genetic techniques such as marker-based chromosomal linkage analysis. In such studies, gene expression is treated as a quantitative trait. Linkage analyses are then carried out using a genome-wide genetic marker panel to identify chromosomal loci that influence (positively or negatively) the pattern of expression of individual genes or clusters of genes. The salient advantage of this combined approach, originally outlined by Jansen and Nap [16], is that it permits assessment, in a quantitative fashion, of the proportion of gene loci that display self-linkage (i.e., *cis*-proximal) relative to those that link to one or more distant or *trans* loci.

Brem et al. [20] were the first to employ this strategy experimentally to map genetic determinants of variation in gene expression in yeast. By testing 1528 genes differentially expressed between strains (see above) for linkage with 3312 markers spaced across the entire *S. cerevisiae* genome, Brem et al. found that 308 of these genes (20%) showed linkage to at least one locus at a high statistical threshold. Simulation experiments indicated that the approach should have 97% power to detect any single locus controlling expression variation for a particular gene. The simulation also suggested that the study should have roughly 40% power to detect up to five controlling loci of equal effect for any gene. Taken together, the results appear to imply that quantitative variation in the expression of most yeast genes is under the influence of multiple loci.

In a follow-on yeast study, Yvert et al. [25] attempted to differentiate *cis*-acting vs *trans*-acting genetic components directly. In addition to focusing on individual genes, they defined 798 clusters comprising two or more genes likely to be coregulated on the basis of similar expression patterns. Using 86 segregants and 3114 linkage markers, Yvert et al. identified 2294 individual genes and 304 gene clusters (collectively containing 1011 genes) that showed linkage to at least 1 position in the genome at a high significance threshold. They then identified a subset comprising 578 genes (25%) and 57 clusters (20%) that showed proximal *cis* linkage to within 10 kb of the expressed gene. Based on these findings, the authors concluded that most variation is due to *trans* effects. However, the sensitivity of this approach to *cis* effects is not clear, and it is not possible to account reliably for compound *"cis–trans"* effects, e.g., variation in *cis*-regulatory sequences of *trans*-regulatory genes.

Schadt et al. [22] employed a similar approach to analyze heritable patterns of gene expression in mice, maize, and humans. Analyzing an F2 cross constructed from two strains of inbred mice, they found 7861 of 23,574 genes to be differentially expressed between F2 animals or between the two parental strains. Linkage studies employing a genome-wide panel of microsatellite markers at 13 cM average density identified loci with lod scores of >4.3 for 2123 of

these genes, accounting for 25% of the observed transcriptional variation. Markers with lod scores of >7.0 were found for a subset of 965 genes, but could explain 50% of expression variation of linked genes. Overall, 40% of genes with LOD >3.0 had their expression trait map to more than one locus, and 4% mapped to more than three loci.

A study of maize by the same investigators revealed even more pronounced effects. Of 18,805 genes differentially expressed in ear leaf tissue (see above), 6481 (34%) linked to marker regions with LOD >3.0. Strikingly, 80% of the regions linking with LOD >7.0 colocalized with the linked gene—the signature of proximal *cis*-acting effects on transcription.

A significant feature common to both mouse and maize studies was that the most pronounced lod scores were observed for loci with *cis*-linked effects. A priori, expression-linkage study designs are generally expected to be more sensitive to *cis* effects than to *trans* effects, since the latter are presumably spread more diffusely over larger numbers of genes and therefore more difficult to detect.

Three general conclusions are forthcoming from the studies discussed above. First, natural variation in gene expression appears to be a common phenomenon across the evolutionary spectrum. Second, even in relatively simple organisms, the expression of a large number of genes appears to be affected by more than one locus. Third, *cis*-acting variation appears to play a considerable role in determining genetic variability in gene expression on a genomic level. Just how common such variation is, however, remains in question.

Limitations of global expression profiling-based approaches

One drawback of using microarrays to measure individualized gene expression relates to sensitivity. Part of this can be mitigated with innovative experimental designs and quantitative techniques such as those employed by Jin et al. [18] and Oleksiak et al. [21]. However, part of it is intrinsic to the experimental platform and may limit sensitivity for *cis*-acting effects in the context of nonhaploid genomes. In the studies discussed above, significance thresholds for interindividual variation typically correspond to reproducible 1.5- to ≥2-fold changes in gene expression. For diploid organisms, these measurements incorporate expression from both alleles, and may therefore underestimate allele-specific *cis*-acting effects. Comparing a given gene between two heterozygous individuals, the observation of 50% interindividual variability in overall expression of a given gene exceeding 50% (i.e., the 1.5-fold threshold employed in most studies) may imply an underlying inter-allelic variability exceeding 100%. A 2-fold change in aggregate expression of a gene between individuals may imply an allelic *cis*-regulatory imbalance of >300%.

Another drawback of current microarray-based survey approaches is the difficulty in distinguishing trans-regulatory effects on gene expression from those due to environmental factors. For simpler organisms such as yeast, even slight perturbations in growth conditions or availability of nutrients may have wide-ranging effects. This problem may be considerably more significant in the case of naturally occurring populations (such as the human) that cannot be raised in a controlled environment. One possible route for ameliorating this deficit would be to examine a broad range of controlled environmental conditions.

In cases in which different cis-regulatory variants of the same locus contribute to the expression phenotype, it may yet be possible to differentiate these components through further application of the genetic dissection strategy. For example, by constructing an F2 population from two inbred strains of mice and estimating the additive and dominance effects at a study locus, it may be possible to approximate the relative abundance of different allelic species.

Direct examination of allelic variation

An alternative to expression profiling is direct measurement of allelic expression, which may be effected through design of allele-specific primers that exploit known coding sequence polymorphisms. Quantitative methods of allele discrimination can then be applied to individual subjects who are heterozygous for the marker polymorphism to measure relative allelic expression [16]. A key advantage of allelic studies is that they circumvent many of the difficulties noted above for microarray-based transcript profiling. First, decomposition of the aggregate expression signal into allelic components significantly increases sensitivity. Second, comparison of alleles within an individual enables each allele to act as an internal control for confounding factors that impact the overall expression of a gene, including tissue quality, sample preparation, environmental influences, and the influences of *trans*-regulatory proteins. Third, since allelic variation is by definition reflective of *cis*-acting influences, it can be used to measure directly the proportion of genes subject to *cis* variation.

On a genomic level, what proportion of genes exhibit allelic variation in expression? A series of recent studies shed light on this important question. Because they employ a variety of designs, it is useful to preface closer scrutiny with a brief consideration of the factors that may mitigate the success of a particular approach for detecting allelic differences. Here, the overall sensitivity of the approach is determined by three factors: (i) the threshold of detection of expression variation, (ii) the number of individuals analyzed, and (iii) the number of tissues analyzed. The first is a basic property of the assay used. If executed properly, quantitative PCR-based assays should be capable of reliably detecting >20% differences in the relative abundance of two transcripts (assuming an appropriate level of replication) [26]. The second factor is an expected consequence of the fact that the vast majority of polymorphic alleles in human

populations have frequencies of <10%. The third factor recognizes that genetic variation may affect regulatory motifs that are instantiated by tissue-specific regulatory factors, and therefore the effect of such variation will be manifest only in a particular tissue environment. An additional limiting factor for studies employing PCR relates to the availability of suitable transcribed sequence variants permitting allele-specific discrimination, a situation that is not expected to obtain for all genes.

Allelic variation in mammalian gene expression

Apart from one study conducted in mice [27], systematic surveys of allelic variation in gene expression performed to date have analyzed human populations. Salient finds from these studies are summarized in Table 1. An important preliminary observation that may be made from these findings is that while *cis*-regulatory variability appears to be a common phenomenon among genes, the alleles that underlie such variability for a given gene appear to be uncommon or rare.

In an attempt to discern the proportion of genes subject to *cis*-regulatory variation in mice, Cowles et al. [27] examined allelic expression in 69 mouse genes for which coding sequence variants enabled design of allele-specific primers. They then screened spleen, liver, and brain tissues of two F1 hybrid mice from each of five genetic backgrounds. Using an allelic ratio threshold of 1.5 (i.e., at least 50% difference between alleles of the same gene), Cowles et al. identified 4 genes (6%) that clearly displayed *cis*-linked variation in expression. Significantly, for 2 of these genes, the manifestation of allelic variation followed a tissue-specific pattern and was evident only in the liver.

Using a different quantitative approach that enabled a lower but statistically rigorous threshold of detection (allelic ratio of >1.2), Yan et al. [28] tested 13 genes for allelic variation expression in lymphoid cells from a sample of 96 individuals from the CEPH families. For any given gene, between 17 and 37 individuals were found to be heterozygous and therefore informative for allelic ratio. Allelic variation was observed for 6/13 genes (46%), with allelic ratios ranging from 1.3:1.0 to 4.3:1.0. For two of these genes, Yan et al. further established that allelic variation was inherited in a Mendelian fashion. The higher proportion of varying genes compared with Cowles et al. may be explained in part by the lower significance threshold, but also by the larger number of individuals analyzed. Presumably, if Yan et al. had screened additional tissue types, even more variability might have been detected. Importantly, not all genes exhibiting variation did so in all individuals tested. For example, *catalase* gene expression exhibited allelic imbalance (1.4:1) in only 1 of 37 heterozygous individuals examined, while *p73* expression varied between alleles in 30% of heterozygotes, among whom the allelic imbalances ranged from 1.5:1 to 4.3:1. This suggests a significant role for genetic or epigenetic background.

Figures comparable to those of Yan were obtained by Bray et al. [29], who found 7/15 (47%) brain-expressed genes to evince allelic variation in expression in a sample of 60 unrelated subjects. Examining a larger set of genes ($n = 126$) in lymphoid cells of 60 individuals selected from five CEPH pedigrees, Pastinen et al. [30] found allelic variation in 23 (18%) genes though at a more rigorous allelic ratio cutoff of >1.5:1.0. Significantly, for genes displaying allelic variation, most exhibited both up- and down-regulation (in the context of different individuals), suggesting that activating and repressing polymorphisms occur with appreciable (and perhaps comparable) frequency. Another intriguing finding was discordant manifestation of allelic variation in two pairs of siblings with complete sharing of alleles identical-by-descent. This indicates that allelic expression itself may in certain cases exhibit incomplete penetrance, presumably due to variabil-

Table 1

Study	Organism	No. of genes surveyed	No. of individuals/gene	No. of tissues	Threshold allelic ratio	Inter-individual variation	Monoallelic expression	Tissue-specific variation	Comments
Cowles et al. [25]	MM	69	2 (10)[a]	3	1.5:1	4 (6%)	–	Y	
Yan et al. [26]	HS	13	17–37[b]	1	1.2:1	6 (46%)	–	–	3/6 genes showed allelic variation in only 1 individual
Bray et al. [27]	HS	15	8–26[c]	1	1.2:1	7 (47%)	–	–	5/7 genes showed allelic variation in only 1 individual
Pastinen et al. [28]	HS	126	5–53[d]	1	1.5:1	23 (18%)	3 (2%)	–	Allelic variation observed in 3 to 13 individuals for each gene
Lo et al. [29]	HS	602	1–5[e]	2	2.0:1[f]	326 (54%)	10 (1.6%)	Y	207/326 genes showed allelic variation in only 1 individual.

[a] Two adult F1 females from each of five hybrid backgrounds.
[b] Analysis of 96 individuals from CEPH families yielded between 17 and 37 heterozygotes per gene.
[c] Analysis of 60 unrelated individuals yielded between 8 and 26 heterozygotes per gene.
[d] Analysis of 63 unrelated individuals yielded between 5 and 53 informative heterozygotes per gene.
[e] Liver and kidney tissue from one and five fetuses examined for each gene.
[f] 170 genes showed allelic imbalance > 4.0:1.0.

ity in the cognate *trans*-regulatory factor(s) of a *cis*-regulatory variant or through an epigenetic mechanism.

All of the aforementioned studies relied on standard quantitative PCR techniques for transcript measurement. Taking a different approach with considerably more scalable potential, Lo et al. [31] adapted the Affymetrix HuSNP chip system to analyze allele-specific gene expression. Of 1494 SNPs assayed by the chip, 1063 fall within coding regions. This enabled simultaneous assay of 602 genes expressed in fetal kidney and/or liver tissue from seven heterozygous individuals. Use of a microarray platform diminished sensitivity relative to quantitative PCR, necessitating a higher allelic ratio cutoff of >2.0:1.0. Despite this, allelic variation in either liver or kidney tissue was observed for 326 genes (54%). Lo et al. further observed that the manifestation of this variation was tissue-specific, analogous to the findings of Cowles et al. in mice.

Taken together, the aforementioned studies establish allelic variation in gene expression as a common feature of human genes and suggest that between 25 and 50% of human gene regulation is mediated in cis. Although polymorphism in *cis*-regulatory elements is likely to be a major cause of this variability, allelic imbalance in gene expression does not necessarily imply the action of this class of variants. Alternative mechanisms that can produce allelic variation include imprinting [32,33], random monoallelic expression [34], allele-specific splice variants [35], unequal allele-specific mRNA decay rates [36], and lack of linkage disequilibrium between the allelic assay polymorphism and the causative *cis* defect. A small subset of the genes in the studies of Pastinen et al. ($n = 3$; 2%) and Lo et al. ($n = 10$; 1.6%) exhibited monoallelic expression (i.e., nearly complete silencing of one allele), compatible with imprinting or random monoallelic expression. Notably, these figures are considerably higher than current estimates of the proportion of imprinted human genes [37]. On a cautionary note, one factor that will require further clarification through acquisition of larger data sets with a greater degree of replication is the role of experimental vs biological noise in interpreting findings that are suggestive of allelic expression.

Implications for the genomic organization of gene regulation

The prevalence of allelic variation in human gene expression suggests a high frequency of polymorphism in *cis*-regulatory sequences. The existence of such variability, in combination with certain distinguishing architectural features of *cis* regulation in higher eukaryotes, may have played an important role in shaping complex genomes.

Vertebrate gene expression is regulated by several different classes of *cis*-regulatory DNA sequences, including enhancers, silencers, insulators, and promoters [38–40]. The core promoter is the site of formation of the transcription complex. Enhancers and silencers act over distances—typically several kilobases but up to many hundreds of kilobases—to potentiate or repress Pol II coalescence and function. Insulator sequences prevent enhancers and silencers targeted to one gene from inappropriately regulating a neighboring gene. Larger, more complex elements comprising multiple enhancers and/or silencers that coordinate the activity of linked genes over large chromosomal domains ("locus control regions" or "domain control regions") have come to light [41,42].

The disparate structure of mammalian—and particularly human—*cis*-regulatory systems is thus highly distinguished from that of lower organisms such as *Saccharomyces*, in which gene-proximal (promoter) sequences appear to predominate. Two additional features of this structure add further complexity. The first is the fact that *cis*-regulatory sequences and genic sequences are frequently commingled [43]. For example, regulatory sequences controlling one gene may be located within the intron of a nearby unrelated gene [44], or a gene may be separated from its regulatory sequences by one or more entire intervening genes [41,42,44]. The second complicating feature is that many *cis*-regulatory elements appear to be "multipotential," i.e., capable of activating functionally diverse linked genes [45] or even heterologous genes that have been approximated through naturally occurring translocations [46].

One predicted result of *cis*-regulatory commingling is preservation of syntenic relationships between unrelated genes, examples of which abound and have been linked directly with regulation in a few cases [44,47,48]. A predicted consequence of the ability of certain *cis*-regulatory elements to activate multiple linked genes is the emergence of clusters of coexpressed (if not coregulated) genes. Such clustering has been observed for functionally unrelated but coexpressed genes in a variety of species [49–53], though it has not yet been definitively connected with particular *cis*-regulatory sequences.

In the human genome, clustering of coexpressed genes appears to be most prominent in the case of widely expressed (so-called "housekeeping") genes [49]. Interestingly, such genes comprise a large share of those for which interindividual and allelic variability in expression has been observed [30,31,54]. An important question therefore arises concerning the mechanism of this variability: is it specific to the gene in question, or is it a feature of the domain in which the gene is ensconced? This can be tested directly by systematically examining genes neighboring those found to exhibit allelic or interindividual variability. The total number of genes examined to date in studies of allelic variation has not been large enough to address this issue comprehensively since few genes were juxtaposed. Pastinen et al. and Lo et al. both identified small sets of physically colocated genes exhibiting allelic variability. However, in both cases these genes were known to reside within imprinted domains.

The significance of *cis*-regulatory variability for genomic evolution has not been considered in detail, owing in part to

the paucity of data available until very recently. It is likely that commingling of *cis*-regulatory regions between juxtaposed genes combined with prevalent functional genetic variability in *cis* regulation provided the substrate for widespread localized selective effects, which should have left a detectable footprint on the organization of complex genomes.

Implications for human disease

Common diseases are characterized by polygenic inheritance and by quantitative variation in specific phenotypic traits. A major biological mechanism contributing to quantitative phenotypic variation is expected to be heritable variation in the regulation of gene expression, which has been predicted to reside principally within *cis*-regulatory sequences [55]. Since individual *trans*-regulatory transcriptional factors typically interact with a wide network of genes, variation affecting these proteins would be expected to have pleiotropic effects and comparatively dramatic phenotypes and are therefore anticipated to be quite rare. An example of this phenomenon may be found in inherited defects in transcriptional factors that give rise to marked early-onset Type 2 diabetes phenotypes [56,57].

Since transcriptional factors require interaction with *cis*-regulatory sites for their effects to be manifest, defects in the genomic target sites of these factors may produce similar (though quantitatively more subtle) physiological consequences. However, the effects of *cis*-regulatory variations should impact directly only their cognate gene(s). *Cis*-regulatory variation could manifest functionally in a variety of ways by impacting (a) the magnitude of gene expression, (b) the regulation of tissue specificity, (c) the control over timing of expression during development and differentiation, (d) the response to environmental stimuli (such as pharmacologic agents), or (e) some combination thereof. Given the overall prevalence of human genetic variation, lesions in one or more of the cognate *cis*-regulatory sites of a transcriptional factor should be comparatively common. When the multiple regulatory factors that interact with each regulatory sequence of each gene are considered, such *cis* variation would provide the ideal substrate for a complex, quantitatively varying phenotype.

On the level of specific diseases, examples have now emerged to suggest that in vivo, even small differences in allelic expression can have dramatic phenotypic consequences. For example, a modest (<25%) decrease in total APC expression can result in a dramatic increase in risk of development of adenomatous polyposis coli and malignant lesions [58]. Expression polymorphism is expected to be particularly significant in the case of enzymes, for which reaction rates may depend more on quantity that on parameters dictated by coding sequence changes (e.g., K_m, K_{cat}) [59]. For genes that exhibit a "threshold" effect in activity such as receptors, the effect may be pronounced: for example, even a very small difference in the total amount of CFTR transcript can dramatically attenuate the cystic fibrosis phenotype [60,61].

Conclusion

Our understanding of the process of transcriptional regulation on a genomic scale, while nascent, is poised for rapid expansion. Clear evidence is now emerging for heritable variation in gene expression, a significant component of which, at least in humans, appears to be *cis*-regulatory variability. Quantitative approximation of this component is likely to be conservative given that much *cis* variation may be subtle and therefore beyond the limits of detection using current methodologies. The situation is further complicated by the fact that regulatory variability may be manifest in different tissue environments or within discrete windows during development or differentiation.

Despite these limitations, it should now be feasible to apply the approaches discussed above to screen large numbers of genes for the signature of *cis* regulation. The next phase will be to identify the causative variants and to discern whether allelic variability is most commonly due to closely linked variation at the promoter level or at the level of distant (and perhaps multipotential) regulatory sequences.

An optimal approach would combine allelic gene expression data with a catalogue of candidate polymorphisms within cognate *cis*-regulatory regions, thus enabling identification of causative lesions. In the context of linkage studies, this combination would permit mapping of *cis*-linked determinants of gene expression to specific regulatory elements and, indirectly, to specific *trans*-acting factor pathways. Applied systematically, such an approach could underpin detailed dissection of the architecture of gene regulation on a genomic scale and would reveal major regulatory networks of direct relevance to human disease. However, the key enabling step for this approach has been elusive, namely, an effective approach for large-scale identification of *cis*-regulatory sites in vivo at a resolution sufficient for targeted discovery of candidate regulatory genetic variation. Such technologies are now becoming available [62] and would have obvious and important consequences for our understanding of the genomics of gene regulation.

References

[1] Y. Pilpel, P. Sudarsanam, G.M. Church, Identifying regulatory networks by combinatorial analysis of promoter elements, Nat. Genet. 29 (2001) 153–159.

[2] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, N. Friedman, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat. Genet. 34 (2003) 166–176.

[3] M. Markstein, P. Markstein, V. Markstein, M.S. Levine, Genome-

wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo, Proc. Natl. Acad. Sci. USA 99 (2002) 763–768.

[4] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Rubin, G.M. Rubin, M.B. Eisen, Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, Proc. Natl. Acad. Sci. USA 99 (2002) 757–762.

[5] A. Stathopoulos, M. Van Drenth, A. Erives, M. Markstein, M. Levine, Whole-genome analysis of dorsal–ventral patterning in the Drosophila embryo, Cell 111 (2002) 687–701.

[6] M. Rebeiz, N.L. Reeves, J.W. Posakony, SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data: site clustering over random expectation, Proc. Natl. Acad. Sci. USA 99 (2002) 9888–9893.

[7] M.S. Halfon, Y. Grad, G.M. Church, A.M. Michelson, Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model, Genome Res. 12 (2002) 1019–1028.

[8] A. Ureta-Vidal, L. Ettwiller, E. Birney, Comparative genomics: genome-wide analysis in metazoan eukaryotes, Nat. Rev. Genet. 4 (2003) 251–262.

[9] R.H. Waterston, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.

[10] L. Elnitski, R.C. Hardison, J. Li, S. Yang, D. Kolbe, et al., Distinguishing regulatory DNA from neutral sites, Genome Res. 13 (2003) 64–72.

[11] E.T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, et al., Numerous potentially functional but non-genic conserved sequences on human chromosome 21, Nature 420 (2002) 578–582.

[12] G.G. Loots, R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, K.A. Frazer, Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, Science 288 (2000) 136–140.

[13] G.G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak, E.M. Rubin, rVista for comparative sequence-based discovery of functional transcription factor binding sites, Genome Res. 12 (2002) 832–839.

[14] D.P. Mortlock, C. Guenther, D.M. Kingsley, A general approach for identifying distant regulatory elements applied to the Gdf6 gene, Genome Res. 13 (2003) 2069–2081.

[15] N.D. Trinklein, S.J. Aldred, A.J. Saldanha, R.M. Myers, Identification and functional analysis of human transcriptional promoters, Genome Res. 13 (2003) 308–312.

[16] R.C. Jansen, J.P. Nap, Genetical genomics: the added value from segregation, Trends Genet. 17 (2001) 388–391.

[17] M.C. King, A.C. Wilson, Evolution at two levels in humans and chimpanzees, Science 188 (1975) 107–116.

[18] W. Jin, R.M. Riley, R.D. Wolfinger, K.P. White, G. Passador-Gurgel, G. Gibson, The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster, Nat. Genet. 29 (2001) 389–395.

[19] R. Sandberg, R. Yasuda, D.G. Pankratz, T.A. Carter, J.A. Del Rio, L. Wodicka, M. Mayford, D.J. Lockhart, C. Barlow, Regional and strain-specific gene expression mapping in the adult mouse brain, Proc. Natl. Acad. Sci. USA 97 (2000) 11038–11043.

[20] R.B. Brem, G. Yvert, R. Clinton, L. Kruglyak, Genetic dissection of transcriptional regulation in budding yeast, Science 296 (2002) 752–755.

[21] M.F. Oleksiak, G.A. Churchill, D.L. Crawford, Variation in gene expression within and among natural populations, Nat. Genet. 32 (2002) 261–266.

[22] E.E. Schadt, S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, P.S. Linsley, M. Mao, R.B. Stoughton, S.H. Friend, Genetics of gene expression surveyed in maize, mouse and man, Nature 422 (2003) 297–302.

[23] A.R. Whitney, M. Diehn, S.J. Popper, A.A. Alizadeh, J.C. Boldrick, D.A. Relman, P.O. Brown, Individuality and variation in gene expression patterns in human blood, Proc. Natl. Acad. Sci. USA 100 (2003) 1896–1901.

[24] V.G. Cheung, L.K. Conlin, T.M. Weber, M. Arcaro, K.Y. Jen, M. Spielman, R.S. Spielman, Natural variation in human gene expression assessed in lymphoblastoid cells, Nat. Genet. 33 (2003) 422–425.

[25] G. Yvert, R.B. Brem, J. Whittle, J.M. Akey, E. Foss, E.N. Smith, R. Mackelprang, L. Kruglyak, Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors, Nat. Genet. 35 (2003) 57–64.

[26] G. Lahr, A. Starzinski-Powitz, A. Mayer, Analysis of specific gene expression, Methods Enzymol. 356 (2002) 271–281.

[27] C.R. Cowles, J.N. Hirschhorn, D. Altshuler, E.S. Lander, Detection of regulatory variation in mouse genes, Nat. Genet. 32 (2002) 432–437.

[28] H. Yan, W. Yuan, V.E. Velculescu, B. Vogelstein, K.W. Kinzler, Allelic variation in human gene expression, Science 297 (5584) (2002) 1143.

[29] N.J. Bray, P.R. Buckland, M.J. Owen, M.C. O'Donovan, Cis-acting variation in the expression of a high proportion of genes in human brain, Hum. Genet. 113 (2003) 149–153.

[30] T. Pastinen, R. Sladek, S. Gurd, A. Sammak, B. Ge, P. Lepage, K. Lavergne, A. Villeneuve, T. Gaudin, H. Brandstrom, A. Beck, A. Verner, J. Kingsley, E. Harmsen, D. Labuda, K. Morgan, M.C. Naumova, A.K. Naumova, D. Sinnett, T.J. Hudson, A survey of genetic and epigenetic variation affecting human gene expression, Physiol. Genom., (In press).

[31] H.S. Lo, Z. Wang, Y. Hu, H.H. Yang, S. Gere, K.H. Buetow, M.P. Lee, Allelic variation in gene expression is common in the human genome, Genome Res. 13 (8) (2003) 1855–1862.

[32] A.C. Ferguson-Smith, M.A. Surani, Imprinting and the epigenetic asymmetry between parental genomes, Science 293 (2001) 1086–1089.

[33] T. Sakatani, M. Wei, M. Katoh, C. Okita, D. Wada, K. Mitsuya, M. Meguro, M. Ikeguchi, H. Ito, B. Tycko, M. Oshimura, Epigenetic heterogeneity at imprinted loci in normal populations, Biochem. Biophys. Res. Commun. 283 (2001) 1124–1130.

[34] D. Watanabe, D.P. Barlow, Random and imprinted monoallelic expression, Genes Cells 1 (1996) 795–802.

[35] S. Jeong, Y.J. Lee, J.S. Jang, C.W. Park, J.H. Chung, J.K. Seong, K.K. Lee, D.Y. Yu, A novel epigenetic control operating on Vmel+ locus leads to variegated monoallelic expression, Biochem. Biophys. Res. Commun. 279 (2000) 884–890.

[36] A.B. Khodursky, J.A. Bernstein, Life after transcription—revisiting the fate of messenger RNA, Trends Genet. 19 (3) (2003) 113–115.

[37] F. Sleutels, D.P. Barlow, The uniqueness of the imprinting mechanism, Curr. Opin. Genet. Dev. 10 (2000) 229–233.

[38] G. Felsenfeld, M. Groudine, Controlling the double helix, Nature 421 (2003) 448–453.

[39] J.E. Butler, J.T. Kadonaga, The RNA polymerase II core promoter: a key component in the regulation of gene expression, Genes Dev. 16 (2002) 2583–2592.

[40] R. van Driel, P.F. Fransz, P.J. Verschure, The eukaryotic genome: a system regulated at different hierarchical levels, J. Cell Sci. 116 (2003) 4067–4075.

[41] Q. Li, K.R. Peterson, X. Fang, G. Stamatoyannopoulos, Locus control regions, Blood 100 (2002) 3077–3086.

[42] R.C. Hardison, New views of evolution and regulation of vertebrate beta-like globin gene clusters from an orphaned gene in marsupials, Proc. Natl. Acad. Sci. USA 98 (2001) 1327–1329.

[43] N. Dillon, Gene autonomy: positions, please..., Nature 425 (2003) 457.

[44] F. Spitz, F. Gonzalez, D. Duboule, A global control region defines a chromosomal regulatory landscape containing the HoxD cluster, Cell 113 (2003) 405–417.

[45] G. Blom van Assendelft, O. Hanscombe, F. Grosveld, D.R. Greaves, The beta-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner, Cell 56 (1989) 969–977.

[46] C.A. Heckman, T. Cao, L. Somsouk, H. Duan, J.W. Mehew, C.Y.

J.A. Stamatoyannopoulos, Critical elements of the immunoglobulin heavy chain gene enhancers for deregulated expression of bcl-2, Cancer Res. 63 (2003) 6666–6673.

[47] F. Santagati, K. Abe, V. Schmidt, T. Schmitt-John, M. Suzuki, K. Yamamura, K. Imai, Identification of cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny, Genetics 165 (2003) 235–242.

[48] M.A. Nobrega, I. Ovcharenko, V. Afzal, E.M. Rubin, Scanning human gene deserts for long-range enhancers, Science 302 (2003) 413.

[49] B.A. Cohen, R.D. Mitra, J.D. Hughes, G.M. Church, A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression, Nat. Genet. 26 (2000) 183–186.

[50] M.J. Lercher, A.O. Urrutia, L.D. Hurst, Clustering of housekeeping genes provides a unified model of gene order in the human genome, Nat. Genet. 31 (2002) 180–183.

[51] T. Blumenthal, et al., A global analysis of Caenorhabditis elegans operons, Nature 417 (2002) 851–854.

[52] P.T. Spellman, G.M. Rubin, Evidence for large domains of similarly expressed genes in the Drosophila genome, J. Biol. 1 (2002) 5.

[53] M.J. Lercher, T. Blumenthal, L.D. Hurst, Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes, Genome Res. 13 (2003) 238–243.

[54] M.V. Rockman, G.A. Wray, Abundant raw material for cis-regulatory evolution in humans, Mol. Biol. Evol. 19 (2002) 1991–2004.

[55] R.W. Doerge, Mapping and analysis of quantitative trait loci in experimental populations, Nat. Rev. Genet. 3 (1) (2002 (Jan)) 43–52.

[56] S.M. Mitchell, T.M. Frayling, The role of transcription factors in maturity-onset diabetes of the young, Mol. Genet. Metab. 77 (2002) 35–43.

[57] A. Stride, A.T. Hattersley, Different genes, different diabetes: lessons from maturity-onset diabetes of the young, Ann. Med. 34 (2002) 207–216.

[58] H. Yan, Z. Dobbie, S.B. Gruber, S. Markowitz, K. Romans, F.M. Giardiello, K.W. Kinzler, B. Vogelstein, Small changes in expression affect predisposition to tumorigenesis, Nat. Genet. 30 (2002) 25–26.

[59] D.L. Crawford, D.A. Powers, Molecular basis of evolutionary adaptation at the lactate dehydrogenase-B locus in the fish Fundulus heteroclitus, Proc. Natl. Acad. Sci. USA 86 (1989) 9365–9369.

[60] A.S. Ramalho, S. Beck, M. Meyer, D. Penque, G.R. Cutting, M.D. Amaral, Five percent of normal cystic fibrosis transmembrane conductance regulator mRNA ameliorates the severity of pulmonary disease in cystic fibrosis, Am. J. Respir. Cell. Mol. Biol. 27 (2002) 619–627.

[61] N. Rave-Harel, E. Kerem, M. Nissim-Rafinia, I. Madjar, R. Goshen, A. Augarten, A. Rahat, A. Hurwitz, A. Darvasi, B. Kerem, The molecular basis of partial penetrance of splicing mutations in cystic fibrosis, Am. J. Hum. Genet. 60 (1997) 87–94.

[62] P.J. Sabo, R. Humbert, M. Hawrylycz, J.C. Wallace, M.O. Dorschner, M. McArthur, J.A. Stamatoyannopoulos, Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries, Proc. Natl. Acad. Sci. USA 101 (2004) 4537–4542.