

On the Emerging Complexity of Gene Expression

J. E. Purvis

It has been suggested that a cell's genetic program is comparable to an onion: as soon as one property is discovered and characterized, another layer of complexity is found lurking beneath¹. In light of recent studies of global gene expression, however, we may be forced to reject this metaphor on the grounds that, unlike a real onion, the structure of genetic variation seems to contain no core, an infinite number of layers, and ever-increasing size and complexity as each new property is revealed^{2,3}. Indeed, many would argue that the much-heralded advances in genetic data-acquisition (e.g., shotgun-sequencing, microarray technology, high throughput screening) have provided just as much frustration to our understanding of genetic variation as they have improved it^{2,4}. Studies now provide more information than can be readily digested, and are often criticized for exploiting a technology at the expense of lacking a hypothesis⁵. Yet the data-driven approach has afforded us a completely different perspective on biological systems—a “bird's eye view” of global processes⁶, and through the tangled mess of noisy and complex data, a faint outline of the genetic basis of variation is beginning to emerge.

This phenomenon of ever-increasing complexity is not new to biology^{7,8}, yet it has been recently illustrated by a wave of “genome-wide” linkage and association analyses. These studies are based on the notion that the variation in the abundance of a given mRNA transcript is itself a heritable trait. Most of the experimental approaches to establishing this heritability employ the same strategy: search for correlations between transcript levels and genetic markers. The result is a two dimensional landscape of intra-genomic connections

that can span millions of data points in each dimension (see ref. 9 for an example figure). By establishing a map between (possibly multiple) genetic loci and the amount of transcribed RNA, one can begin to enumerate the *quantitative trait loci* (QTL) that are responsible for generating the observed levels of RNA transcripts. A related but opposite phenomenon called *pleiotropy* occurs when a single gene influences multiple phenotypic traits. As will be shown, both phenomena play a large role in modulating a cell's expression profile.

From a practical perspective, the choice to study gene expression as a quantitative trait is a sensible one. Thousands of message levels can be simultaneously measured using complimentary DNA (cDNA) microarrays. Likewise, tissue or organism samples can be genotyped in a similar fashion using distinguishing sets of hybridization probes. Moreover, millions of genetic markers exist for the human genome, including a vast collection of single nucleotide polymorphisms (SNPs) that have been catalogued as part of the International HapMap project⁹. There also are demographical and medical reasons to study expression QTLs. Gene expression correlates with phenotypic diversity among humans^{10,11} and is believed to account for some of the differences in individuals' susceptibility to disease or reaction to therapeutic drugs¹². In fact, most medically important traits do not follow simple monogenic, or Mendelian, inheritance, but are instead *complex traits*—the cumulative effect of several genetic factors¹².

If gene expression is a trait like any other and amenable to genetic analysis¹³, one must concede that, unlike petal color¹⁴ or dumpy wings¹⁵, the “gene expression phenotype¹⁶” is extraordinarily complex. For instance, only 3% of highly heritable transcripts in a common yeast strain are consistent with single-locus inheritance¹⁷; in the BXD mouse strain, a single regulatory locus on chromosome 6 modulates the abundance of over 1500 transcripts¹⁸; and in humans, more than 10% of the genome is not just altered at the single-nucleotide level,

but is physically rearranged to either disrupt or alter transcription. This paper aims to stress the profound complexity of gene expression as a quantitative trait and its relationship with genetic variation, drawing on specific examples from recent studies in the field.

One of the first questions addressed by “genetical genomics¹⁹” was whether quantitative traits are determined by a few loci with large effects, or many loci with small contributions. It turns out that substantial evidence exists for both models. Brem and colleagues found a locus in the yeast genome that controls the abundance of more than 90 transcripts, including a mitochondrial “expression cluster” of 52 genes.²⁰ At the same time, levels for over a thousand messages that differed between two parent strains did not show linkage in the cross, suggesting that these messages (which comprised almost 80% of the ones tested) were affected by multiple weak-acting loci. Similar results for human gene expression were reported by Morley *et al*, who describe a complex network of “master transcriptional regulators” that affect the baseline levels of many genes with similar expression profiles. Notably, many of these closely linked genes were controlled by the same *cis*-acting regulators (determinants in close physical proximity to the gene), implying that transcriptional regulation is partly influenced by the spatial arrangement of genetic elements. Cheung and others investigated these questions in the context of complex human diseases. In one study, they found that individuals who were carriers for the recessive disease ataxia telangiectasia (AT) could be identified by a group of 71 genes whose expression was significantly altered in heterozygotes²¹. This finding underscored the sensitivity and utility of gene expression as a phenotype, since carriers of AT were previously undetectable by physical examination.

It quickly became clear that the workload for control of gene expression is unevenly divided, with overlapping layers of control and hierarchical organization of interacting

genetic determinants. To begin to understand complex control mathematically, the *additive model*²² was proposed to provide a simplified framework for understanding the contributions of each QTL. Under the model, each loci contributes a small positive amount to a given trait (known as an *effect size*), much like individual votes count toward the total vote tally in an election. For statisticians, there is an obvious problem with mapping complex traits under this model: the more genes sharing control of a certain quantitative trait, the more difficult it becomes to detect significant associations. This is especially an issue when, as was found in yeast, most loci account for less than a third of the total parental expression differences¹⁷. To illustrate, using a conservatively parameterized simulation of QTL detection, Brem could only detect loci for traits that were controlled by less than seven determinants or a single locus controlling more than 25% of a given trait. She found that half of the transcripts required models of more than five controlling determinants, and that a third of the genes required more than eight. These results suggest a theoretical upper bound on the number of simultaneously acting determinants that can be identified using current methods. The issue was also explored quantitatively by Cheung and colleagues, who performed simulations to determine the theoretical cohort size required to detect associations with weak but significant linkage. Assuming an idealized determinant that is virtually identical to the genetic marker, it would require samples of 500 individuals to detect a QTL explaining 10% of a trait, 80% of the time. Besides the computational barrier it creates, the concept of the “percent variance explained” by a certain QTL¹⁷ stresses the existence of the carefully modulated effects of many determinants on a single trait.

With the extensive use of statistical inference to draw qualitative conclusions from the data,⁴ it is easy to lose sight of the fact that the true effects of genetic variation on gene expression are molecular.²³ If global analyses provide a loose sketch of functional genetic

interactions, how can we use this data to identify functional gene regulating components? Bystrykh and others have proposed a way to do this by narrowing their search to a specific locus and searching for candidate genes that control a single trait¹⁹. After discovering a QTL on mouse chromosome 11 called stem cell proliferation-2 (*Scp2*) that is responsible for how long cells remain in S phase, they identified all *cis*-regulated transcripts that mapped to this interval. By looking at segregating expression patterns in a pedigree of genetically distinct mice, they were able to pinpoint differences in expression patterns between HSCs causing the S-phase phenotype and identify eight candidate genes on this interval. Of these, four transcripts are known to be involved in DNA replication and repair machinery. But there is a lesson to be learned here as well: by narrowing the search to a specific locus of candidate genes, one cannot be guaranteed that the region functions exclusively as single functioning unit (in this case, as a stem-cell regulating region). In fact, after correlating other transcripts elsewhere in the genome, this region was found to contain a variety of seemingly unrelated functional elements (e.g. two ontologically unrelated seven-transmembrane receptors). This example illustrates the key tradeoff between genome-wide and reductionist methods: global methods are comprehensive, but lack sufficient resolution; focused approaches provide sufficient detail, but can be harried by extraneous influences. The ultimate map of gene expression will probably require us to fuse the detail from focused studies with genome-wide analyses.

Taking functional analyses one step further, Morley *et al.* wrapped up their study of genome-wide association by examining the functional connection between a genetic marker (rs755467) and its regulated target gene chitinase 3-like 2 (*CHI3L2*). It was found that a single nucleotide change in rs755467 (G to T) was responsible for doubling the expression of *CHI3L2* by creating a more favorable binding site for RNA polymerase II at the

promoter. While it would take a lifetime to follow up each association detected by GWA with the appropriate biochemical studies, it is clear that global analyses have created the means to identify possible interactions for further study.

In 2005, Cheung and colleagues shifted the focus from *where* genetic control elements were located in the human genome to *which* allelic variants (at these sites or others) were responsible for causing the expression phenotype. Using a dense collection of SNPs near linkage peaks, they used microarray expression data to identify sets of allelic determinants linked to the gene expression phenotypes. This technique of *association mapping*²⁴ represents a more finely tuned picture of the causes of variation in gene expression since each genetic determinant can exist in one of several allelic states. As an example, a single SNP marker (rs6700) caused an eightfold difference in the expression of the human phosphoserine phosphatase-like (*PSPHL*) gene, depending on which allele was present at that site. They also characterized the physical distribution of these regulatory sites with respect to the target genes, showing that the same number of sites was located at the 5' and 3' ends, as well as within the gene region itself.

A crucial feature of association analysis is that it provides a handle on the actual functional variants that control gene expression. Although an associated marker itself might not be the cause for an observed expression pattern, it is likely to be in linkage disequilibrium with the true causal variant given the block-like haplotype structure of the human genome (stretches of SNPs are often in strong linkage disequilibrium)²⁵. Thus, mapping of one SNP to a given trait may be the second-to-last step in identifying the true functional variant.²⁶

In both linkage and association mapping, one of the highlighted findings was that the determinants of gene expression could be located anywhere in the genome, not just close to the target gene. For decades, transcriptional regulation was generally thought to involve elements located close to the affected gene (within a few kilobases), which influence transcription by recruiting or blocking soluble factors at the transcription start site.²⁷ While many genome-wide studies have confirmed this type of *cis*-acting linkage, the discovery of numerous *trans*-acting determinants has demonstrated that many genes are strongly influenced by determinants beyond this immediate range, including elements located on other chromosomes. In practice, the difference between *cis*- and *trans*-acting has been somewhat arbitrarily defined. Cheung *et al.* set the cut-off at 50 kb in a study of humans, while others have extended this range up to 20 Mb¹⁹. Regardless of the designation, the concept is the same: many genetic determinants act through soluble diffusing factors that travel through the cell to effect a change in expression. Consequently, this class of determinants can be tougher to detect (relative to *cis*-acting determinants), as observed by Hubner *et al.* in a genome-wide linkage analysis of gene expression in rats²⁸. The authors found that as the genome-wide significance of the QTL linkage was relaxed (higher *P* value), they could detect a higher proportion of *trans*-acting QTLs relative to *cis*-acting determinants. This result suggests a stronger effect from *cis*-acting QTLs and the more likely polygenic activity of *trans*-acting determinants.

Global expression studies have shed light on several interesting features of complex inheritance. For example, not only can a combination of several determinants with small effect size give rise to large differences in a quantitative phenotype (recall the additive model), but the blending of two genomes via sexual reproduction can actually uncover, and polarize, differences in gene expression. This pattern was observed for in a study of two

genetically distinct yeast strains and their progeny²⁰, in which it was found that almost half of the differentially expressed genes showing linkage in the cross were *not* originally identified as different in the parental strains. In other words, the progeny had manifested quantitative traits that extended beyond the parental phenotype. This phenomenon is fittingly known as ‘*transgressive*’ segregation and involves two controlling loci that act in opposite directions but are decoupled after meiosis²⁹.

Epistatic interactions comprise another class of gene expression patterns that result when one QTL influences another QTL.³⁰ For example, suppose the *A*-allele of gene *X* can cause disease only when the *B*-allele of gene *Y* is present. To date, most of the studies of model organisms have ignored epistasis since it is not generally detectable in QTL-mapping studies.³¹ Some groups have attempted to model epistatic effects on complex inheritance by testing for a difference between mean expression levels for parent and offspring mean phenotype values,³² reasoning that the means should be equal under an additive inheritance model. These results can be highly processed,¹⁷ however, and indicate that many transcripts experience complex epistatic effects that are undetectable by current methods.

If the diversity in gene expression due to allelic variation was not already sufficiently complicated, another layer of complexity was recently brought to the forefront by two seminal papers that showed that gross chromosomal alterations comprise a large proportion of genotypic variation^{33,34} and represent a significant fraction of observable phenotypic diversity³. One of these types of alterations, *copy number variation* (CNV), refers to insertions, deletions, and duplications of 1 kb or larger that are present at variable copy number compared to a reference genome.³⁵ CNVs can disrupt coding sequences, interfere with long-range genetic interactions, or alter *gene dosage* if multiple copies of the transcript are synthesized from repeated regions. As an example of how CNVs can affect the expression

phenotype, the *CYP2D6* gene encodes a member of the cytochrome P450 superfamily of enzymes and is important in drug-metabolism. Alleles of this gene have been shown to vary more than tenfold between populations, increasing metabolism of drugs in proportion to gene dosage.³⁶ A comprehensive look at CNV in the human genome was performed by Redon and colleagues,³³ who used SNP genotyping arrays and *comparative genomic hybridization* (which quantifies gains and losses of cellular DNA content) to measure the abundance of transcripts in the HapMap collection.⁹ CNVs were found to cover 12% of the genome, including regions known to be involved in genomic disorders and complex traits.

To assess the prevalence of intermediately-sized structural variations in the human genome, Tuzun and colleagues mapped half a million pairs of fosmid end sequences from an anonymous female donor to their best locations in the human reference genome, letting the fosmid ends align or ‘criss-cross’ with the reference sequence. Instances where the orientation of the end sequences was inconsistent between fosmid end pairs and the reference genome provided the locations of putative insertions, deletions, or inversions. They found hundreds of such variants, ranging in size from ~5 kb to almost 2 Mb for some deletions. In a separate study, Stranger *et al.* compared the relative contributions of SNPs and CNVs to gene expression levels in the HapMap population.³⁷ Although more genes were associated with SNPs than with CNV or CGH clones, both sources of genetic variation caused a substantial effect on gene expression. Of course, this type of comparison is limited by the density of each type of probing set (~700,000 SNPs and ~25,000 CNV clones were used), yet it made clear that both forms of polymorphisms (single-point and structural) will be important in future investigations of genetic contributions to gene expression patterns.

The overarching theme in each of these studies is that the most immediate manifestation of the information content in genes, namely, the expression of gene message, is itself a heritable trait and subject to genetic transmission rules. If there is any uncertainty remaining regarding the tremendous complexity of genetic effects on gene expression, we are now faced with evidence that both strands of the human genome are probably transcribed and involve interleaving mRNA segments which are themselves multifunctional in nature.² This presents a problem for the current method of data collection, since microarrays are normally designed such that each probe is a discrete sensor for measuring the abundance of a single transcript (usually cDNA from a single message). As such, they are not capable of measuring the abundance of overlapping transcripts. Furthermore, as mentioned previously, most mapping methods are designed to detect additive QTL effects and not epistatic interactions or transgressive segregation.³⁰ Thus, new tools and methods will be necessary to probe the genome at the level of detail for which it is already known to display variation. In the meantime, existing linkage and association mapping techniques provide a generalized method that can be improved by more markers and smaller clone sizes, to be followed up with careful functional and statistical analyses. Clearly, there is much work to be done in filling in the fine detail of what is already an astonishingly complicated outline.

References

1. Shanmugam, K.T. Personal Communication. (2001).
2. Kapranov, P., Willingham, A.T. & Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **advanced online publication**(2007).
3. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nature Reviews Genetics* **7**, 862-872 (2006).
4. Kell, D.B. & Oliver, S.G. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99-105 (2004).
5. Allen, J.F. Bioinformatics and discovery: induction beckons again. *Bioessays* **23**, 104-107 (2001).
6. Ko, M.S.H. Expression profiling of the mouse early embryo: Reflections and perspectives. *Developmental Dynamics* **235**, 2437-2448 (2006).

7. Limburg, K.E. Increasing Complexity and Energy-Flow in Models of Food Webs. *Ecological Modelling* **29**, 5-25 (1985).
8. Warth, J.A. & Rucknagel, D.L. The Increasing Complexity of Sickle-Cell-Anemia. *Progress in Hematology* **13**, 25-47 (1983).
9. Altshuler, D. et al. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
10. Dermitzakis, E.T. & Stranger, B.E. Genetic variation in human gene expression. *Mammalian Genome* **17**, 503-508 (2006).
11. Spielman, R.S. et al. Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics* **39**, 226-231 (2007).
12. Lander, E.S. & Schork, N.J. Genetic Dissection of Complex Traits. *Science* **265**, 2037-2048 (1994).
13. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747 (2004).
14. Jain, S.K. & Joshi, B.C. Estimation of Linkage + Penetrance Parameters in Study of Petal Color in Pigeon Pea. *Genetics* **49**, 611-& (1964).
15. Stroman, P. Pyrimidine-Sensitive Drosophila Wing Mutants - Withered (Whd), Tilt (Tt) and Dumpy (Dp). *Hereditas* **78**, 157-167 (1974).
16. Cheung, V.G. & Spielman, R.S. The genetics of variation in gene expression. *Nature Genetics* **32**, 522-525 (2002).
17. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572-1577 (2005).
18. Chesler, E.J. et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics* **37**, 233-242 (2005).
19. Bystrykh, L. et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genetics* **37**, 225-232 (2005).
20. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752-755 (2002).
21. Watts, J.A. et al. Gene expression phenotype in heterozygous carriers of ataxia telangiectasia. *American Journal of Human Genetics* **71**, 791-800 (2002).
22. Slate, J. Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology* **14**, 363-379 (2005).
23. Watson, J.D. & Crick, F.H.C. The Structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology* **18**, 123-131 (1953).
24. Risch, N. & Merikangas, K. The Future of Genetic Studies of Complex Human Diseases. *Science* **273**, 1516-1517 (1996).
25. Wall, J.D. & Pritchard, J.K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics* **4**, 587-597 (2003).
26. Cheung, V.G. et al. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365-1369 (2005).
27. Jacob, F. & Monod, J. Genetic Regulatory Mechanisms in Synthesis of Proteins. *Journal of Molecular Biology* **3**, 318-& (1961).
28. Hubner, N. et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* **37**, 243-253 (2005).
29. Rieseberg, L.H., Archer, M.A. & Wayne, R.K. Transgressive segregation, adaptation and speciation. *Heredity* **83**, 363-372 (1999).
30. Carlborg, O. & Haley, C.S. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* **5**, 618-U4 (2004).
31. Flint, J. & Mott, R. FINDING THE MOLECULAR BASIS OF QUANTITATIVE TRAITS: SUCCESSES AND PITFALLS. *Nature Reviews Genetics* **2**, 437-445 (2001).
32. Lynch, M.W., B. *Genetics and Analysis of Quantitative Traits*, (Sinauer, Sunderland, Massachusetts, 1998).
33. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).
34. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nature Genetics* **37**, 727-732 (2005).
35. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nature Reviews Genetics* **7**, 85-97 (2006).
36. Xu, C. et al. An in Vivo Pilot Study Characterizing the New CYP2A6*7, *8, and *10 Alleles. *Biochemical and Biophysical Research Communications* **290**, 318-324 (2002).

37. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853 (2007).