GENE EXPRESSION

The ATGC of gene expression

The advent of whole-genome approaches, microarray technologies and improved computation has given us new insights into the regulation of gene expression, although relating the expression of regulatory genes to that of the genes under their control remains difficult. Beer and Tavazoie have recently reported on a systematic approach to the problem, based on the identification of 5′-upstream DNA sequences of the genes of interest.

The relationship between mRNA levels of a transcription factor and the gene it regulates might not be direct owing to, for example, post-transcriptional regulation. But correlating the abundance of the active transcription-factor protein to the mRNA level of the gene it regulates is technically challenging. Fortunately, the new method of Beer and Tavazoie sidesteps these difficulties by building sequence-to-gene networks, using short 5′-upstream DNA sequence elements as a surrogate for active transcription-factor protein levels.

Working in yeast, the authors began by looking for groups of genes that were coexpressed under a range of experimental conditions (for example, heat shock or diamide treatment). In all, they assigned 2,587 genes to 49 'expression patterns'. Next, they identified overrepresented sequence motifs within 800 bp upstream of the genes in each pattern. The rationale was that such sequence elements were likely to be involved in the regulation of the corresponding genes. Indeed, many of the motifs that were pinpointed closely matched known regulatory elements. The authors used a Bayesian approach to apply further constraints to the motifs, such as orientation and distance to ATG, so that regulatory 'rules' could be inferred. They also took into account combinations of motifs. For example, the two elements PAC and RRPE were both found upstream of a high proportion of genes in a particular expression pattern, indicating that they coregulate genes in this group. It emerged that the order and distance between the two elements also strongly affected the degree of correlation between genes.

So, having identifed the upstream sequence elements that are involved in transcriptional regulation, along with the positional and combinatorial constraints that govern their role, the authors tested the predictive power of their approach. Based on promoter sequences alone, they attempted to predict the expression patterns of 'test' sets of genes, not used while learning the rules. Impressively, their predictions were accurate in 73% of the genes, and fine-tuning of the system is likely to improve on this.

As high-quality mRNA expression data becomes readily available in different organisms, the process reported here will be invaluable to our understanding of the regulation of genes, and of cellular behaviour more generally. The authors have already begun to apply their new approach to multicellular organisms, starting with *Caenorhabditis elegans*. In a preliminary study using expression data collected over a time course from oocyte to adult, they were able to predict the expression patterns of half the genes. As these studies are expanded to take into account further complexities, such as downstream or intronic regulatory elements, it should become possible to unravel the transcriptional regulatory mechanisms behind diverse spatiotemporal processes.

Ruth Kirby, Nature Publishing Group

🌐 **References and links**
ORIGINAL RESEARCH PAPER Beer, M. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004)
WEB SITE
Tavazoie's laboratory:
http://genomics.princeton.edu/tavazoie/index.htm

HUMAN GENETICS

Narrowing down the candidates for asthma

In 2001, a genome-wide scan in asthma identified chromosome 7p as 1 of 6 possible asthma-susceptibility loci. Kere and colleagues have now narrowed the candidate region from a 20-cM to a 133-kb region, which contains 2 genes — *GPRA* and *AAA1*.

The authors genotyped 874 subjects from the Finnish Kainuu subpopulation by interspersing successive rounds of genotyping that increased the density of markers (SNPs and microsatellites) with analysis using the haplotype pattern mining (HPM) algorithm, which searches large sets of unrelated haplotypes for allele patterns that are shared between several haplotypes. Having identified a strong association of a conserved 47-kb haplotype pattern in this way, Kere and colleagues sequenced 133 kb that encompassed the 47-kb region from a homozygous asthmatic patient. Comparison with the public sequence identified 80 new polymorphisms.

Asthmatics from North-eastern Quebec and individuals with high serum immunoglobin E levels from North Karelia, Finland, also had a 133-kb haplotype pattern with the same limits, for which most SNPs were conserved. For the 3 populations combined, 7 alternative haplotypes were formed by 13 SNPs across the most-conserved region of 77 kb. Sequencing of the 133-kb region from 6 individuals who were homozygous for the remaining 6 haplotypes confirmed that they had different SNP compositions. Phylogenetic analysis showed these risk haplotypes to be related and distinct from non-risk haplotypes. By SNP-tagging the risk haplotypes, the authors confirmed that they confer risk in all three populations, which is consistent with the common-disease/common-variant hypothesis.

With the 133-kb region confirmed as a susceptibility locus, the authors looked for open reading frames. They found 2: exons 3 to 5 of a gene that they named *GPRA* (for G-protein-coupled receptor for asthma susceptibility) and, on the opposite strand, exons 3 to 10 of a gene that they called *AAA1* (for asthma-associated alternatively spliced gene 1). Both *GPRA* and *AAA1* are probably alternatively spliced, but *AAA1* might not be a protein-coding gene (for example, *in vitro* translation of its longest open reading frame, which encodes only 74 amino acids, did not produce a stable polypeptide, and no recombinant protein was produced by transiently transfected cells).

GPRA's two main transcripts, A and B, encode proteins of 371 and 377 amino acids, respectively. Whereas the A isoform is mainly expressed by smooth-muscle cells, the B isoform is mainly found in epithelial cells; however, in asthma patients, the B isoform is strongly expressed by smooth-muscle cells.

TECHNOLOGY

A new approach to microarrays

Reporting in *Genome Research*, Mark Chee and colleagues introduce an approach that looks set to realize the potential of randomly assembled microarrays.

Unlike conventional microarrays, the location of probes on randomly assembled microarrays is unknown. However, random assembly allows more effective miniaturization and higher packing densities to be achieved, which improves cost-effectiveness and statistical power. Until now, though, identifying the probes after assembly required complex dye chemistries and labelling processes, which limited attempts to develop randomly assembled microarrays.

Chee and colleagues have developed a simple and efficient algorithm that can accurately decode many thousands of DNA-linked objects, in this case, DNA-coated beads, using only a few labels and several sequential rounds of hybridization.

To demonstrate this approach, the authors prepared 1,520 different oligonucleotide-linked bead types (each bead type represents a unique 22–24-base oligonucleotide sequence that acts as both a probe and as a decodable sequence). These bead types were pooled and assembled into the wells of 96 optical-fibre bundles — each bundle has ~50,000 wells and each well holds one bead — to create 96 arrays at a time. The ~30-fold redundancy per bead type per array increases both the precision and the robustness of this approach, which also has the advantage of using a standard 96-well microtitre plate layout.

For the decode process, three different sets of oligonucleotides that are complementary to the bead sequences were used — two sets labelled with different fluorescent dyes and the other set unlabelled. Oligonucleotides from these sets were used to create eight decoder pools, each with a different combination of oligonucleotides. Although seven successive hybridizations can decode the bead types, the authors included an extra hybridization step that allows for error checking and provides the first functional test for the quality of every probe on a microarray. The fluorescent signals imaged from all eight hybridization stages are decoded by an algorithm that uses the fluorescence intensity profiles of each bead to provide a combinatorial code that can identify them. Encouragingly, decoding randomly assembled microarrays using this approach has a very low error rate, with fewer than five misclassified beads per array.

Not only is this high-performance and cost-effective approach general, and could therefore be applied to any DNA-linked fixed objects or molecules, but it can also be scaled up and is proven in large-scale genomics applications, such as SNP genotyping for the International HapMap Project. In the future, this technology could also help to speed the use of microarray-based assays for large-scale clinical applications.

Natalie Wilson

🌐 **References and links**
ORIGINAL RESEARCH PAPER Gunderson, K. L. *et al*. Decoding randomly ordered DNA arrays. *Genome Res*. **14**, 870–877 (2004)

So, Kere and colleagues suggest that the balance between the A and B isoforms might be altered by polymorphisms in the risk haplotypes. In support of a role for GPRA in asthma pathogenesis, *Gpra* mRNA is significantly upregulated in a mouse model of ovabulmin-induced lung inflammation.

Although these are not the first candidate asthma-susceptibility genes, this study identifies *GPRA* as the strongest candidate so far. And, most encouragingly, the properties of GPRA as a G-protein-coupled receptor, which might be used by a novel ligand as part of a new signalling pathway in asthma, mean that it also has potential as a possible drug target.

Natalie Wilson

🌐 **References and links**
ORIGINAL RESEARCH PAPER Laitinen, T. *et al*. Characterization of a common susceptibility locus for asthma-related traits. *Science* **304**, 300–304 (2004)
WEB SITES
Juha Kere's laboratory:
http://www.cbt.ki.se/groups/jke;
http://www.ktl.fi/diseasegenetics/HUIPPU_MULTIFACTORIAL.HTM