HUMAN GENOMICS

Cracking the regulatory code

A collection of papers catalogues the associations between genetic variation and gene expression in healthy tissues — the largest analysis of this kind so far. SEE ARTICLE P.204 & LETTERS P.239, P.244 & P.249

MICHELLE C. WARD & YOAV GILAD

How does the same DNA sequence, present in almost every cell in the body, give rise to diverse tissues that have distinct functions? The Genotype-Tissue Expression (GTEx) Consortium aims to answer this question by using a strategy called expression quantitative trait loci (eQTL) mapping. This technique allows the researchers to generate a comprehensive catalogue of associations between genetic variation and gene expression across many tissues in many individuals. In four papers[1–4] in this issue, the consortium presents the second phase of their project, and the largest survey of this type so far.

Over the past two decades, considerable progress has been made towards understanding the molecular mechanisms that underlie the dynamic gene-regulatory programs that direct development, differentiation and function in specific cell types. The outstanding challenge is to understand, and ultimately to predict, how genetic differences between individuals contribute to specific traits, including susceptibility to disease.

A large body of work[5] shows that genetic variants that drive inter-individual differences in complex traits, including disease, are often found in non-protein-coding regions of the genome that might determine how and when genes are expressed. As a result, biologists have set out to catalogue and understand how genetic variation in both coding and non-coding regions affects dynamic and tissue-specific gene-expression programs. The GTEx project, established in 2010, represents a coordinated attempt to achieve this goal.

In 2015, the GTEx Consortium described a pilot study[6] in which gene-expression data from multiple tissues were collected from 237 recently deceased donors. The current iteration of the project involves substantially more samples — a total of 7,051 from 449 individuals (Fig. 1). The consortium combined gene-expression measurements from 44 tissues with nucleotide information from each person taken from about 12.5 million DNA bases known to vary between individuals. This involved a concerted collaborative effort to overcome the ethical, legal and technical challenges associated with obtaining post-mortem samples on a large scale.

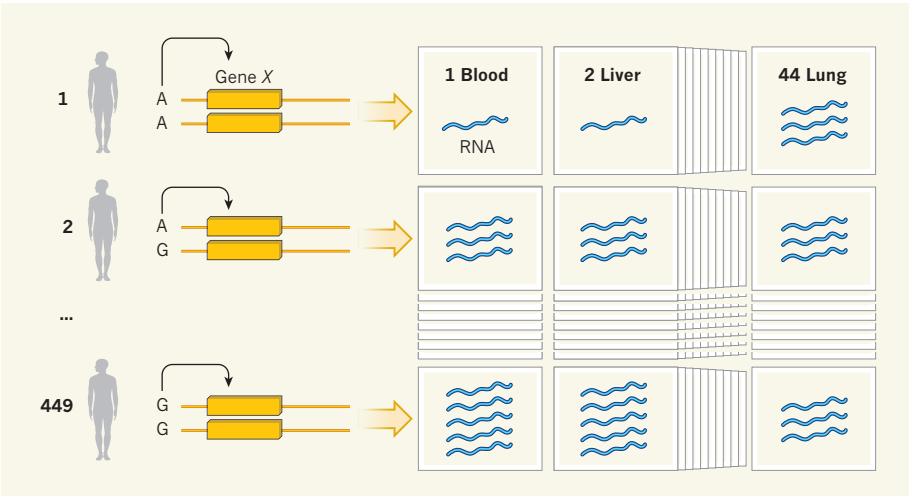

Figure 1 | Data collection by the Genotype-Tissue Expression (GTEx) Consortium. The consortium[1–4] collected tissue samples from 44 tissues in 449 human individuals. The researchers analysed these samples to look for genetic differences between individuals — in this example, one individual harbours two adenosine bases (As) at a particular point on two sister chromosomes, another harbours one A and one guanine (G), and a third harbours two Gs. The authors measured RNA levels to determine whether such genetic variation was associated with differences in gene expression (here, in the levels of RNA transcribed from gene *X*). Different genetic variants were associated with different expression in different tissues.

In the first paper[1] (page 204), the consortium took advantage of its large data set to show that the expression of almost all genes in the human genome is affected by genetic variation. Most of the variants that affect gene expression are located within a few kilobases of the affected gene, and are dubbed *cis*-eQTLs. These variants are typically located in regions of genetic sequence that modify the regulation of only one of a person's two copies of the affected gene — for example in regulatory elements called promoters, enhancers and repressors. The consortium also identified several hundred *trans*-eQTLs, which affect the expression of genes that are located farther away, or even on a different chromosome. These variants typically alter the regulation of both copies of a gene, for example because they encode transcription factors or small RNAs.

The authors showed that *cis*-eQTLs tend to alter gene expression in most tissues examined. By contrast, *trans*-eQTLs generally seem to affect expression in just one or very few tissues. Many of the variants tested had previously been found to be associated with complex diseases and, interestingly, the consortium found that about half of these were associated with altered gene expression in some of the tissues that they tested. This observation demonstrates the usefulness of large eQTL studies for identifying genes and pathways affected by disease-associated genetic variation.

In the second paper[2] (page 239), the authors extended their analyses to specifically examine the effects of rare variants on gene expression. Every individual has tens of thousands of rare non-coding variants, which are often ignored in a clinical context and in disease studies. These variants are also not typically considered in eQTL analyses, which focus on common genetic variation. The authors present a statistical method that integrates DNA-sequence and gene-expression data from the same individual. Their findings underscore the importance of rare variation in determining gene expression. Their statistical approach could ultimately be used to predict which DNA variants in individual genomes

cause cellular changes that lead to disease.

In the third and fourth papers (pages 244 and 249, respectively), the consortium combined its GTEx data with other data sets to investigate how variants associated with altered gene expression can regulate two phenomena — RNA-editing processes[3] and X-chromosome inactivation[4].

In addition to the results presented, the GTEx project has provided a valuable resource for the community, making its raw data available in the dbGaP database (www.ncbi.nlm.nih.gov/gap), and processed data available in an interactive website (www.gtexportal.org). The sample collection, quality control, data standardization and organization of the project are perhaps no longer cutting-edge, because the study was conducted over several years. Nonetheless, these aspects of the work are more thorough than is typical for large consortium projects, so the data can be readily interrogated by other researchers to address specific questions using more-nuanced analyses.

As the GTEx project moves forward and examines more people, it will be necessary to consider three main challenges. First, although the consortium identified almost 1 million genetic variants associated with differences in gene expression, it could be that most don't directly cause gene-expression differences. DNA variants are often correlated across the genome, passed down together from one generation to the next. This means that, in addition to the causal variant for any given trait, numerous related, non-causal associations can be found. Therefore, some causal variants might not yet have been identified by the consortium. A complete genome sequence from each individual will be needed to identify all these associated variants, and should be used alongside new methods to predict the causal variant. The ability to manipulate genetic variants using CRISPR–Cas9 genome editing and to analyse any subsequent changes in gene expression, as the authors do in a handful of cases, should also allow researchers to determine causal genetic variation.

Second, although the GTEx analyses represent the most comprehensive tissue set catalogued so far, all tissues consist of many cell types, which probably contributes to the observed variation in gene expression. Testing for genetic effects on gene expression at a higher resolution in individual cells using single-cell processing technologies will help to distil the signal.

Third, to move beyond descriptive work to an understanding of the actual mechanisms that underlie gene-regulatory programs, multiple functional genomic assays that profile factors affecting gene expression (for example, chromosome accessibility, transcription-factor binding and the modification of DNA by methyl groups) should be performed in the same cells. Genetic variants can affect aspects of the gene-regulatory cascade other than levels of RNA, and these should also be examined. The rate of gene transcription, the mechanism of RNA processing and the rate of translation are three such examples. Some of these aspects of gene regulation will be examined by the ongoing 'Enhancing GTEx' project, as outlined in a Commentary published in *Nature Genetics*[7].

But for many of these dynamic experiments, frozen tissue samples, such as those used in the current study, might not be optimal. Future efforts could use stem-cell models, or study differentiated cells *in vitro* as a complement to generating data from frozen tissue.

Nonetheless, the extensive catalogue generated by the GTEx Consortium takes us a step closer to decoding the regulatory code of the genome. The consequences of genetic variation on gene expression are gradually becoming clearer. ■

Michelle C. Ward *and* **Yoav Gilad** *are in the Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA.*
e-mails: mcward@uchicago.edu; gilad@uchicago.edu

1. The GTEx Consortium. *Nature* **550,** 204–213 (2017).
2. Li, X. *et al. Nature* **550,** 239–243 (2017).
3. Tan, M. H. *et al. Nature* **550,** 249–254 (2017).
4. Tukiainen, T. *et al. Nature* **550,** 244–248 (2017).
5. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. *Am. J. Hum. Genet.* **93,** 779–797 (2013).
6. The GTEx Consortium. *Science* **8,** 648–660 (2015).
7. Stranger, B. E. *et al. Nature Genet.* doi:10.1038/ng.3969 (2017).

MATHEMATICS

A pariah finds a home

Pariahs are fundamental building blocks in a branch of mathematics called group theory, but seem to be unconnected to both physics and other areas of mathematics. Such a connection has now been identified.

TERRY GANNON

The idea of a group is intrinsic to mathematics — it is simply a collection of actions called elements. For example, the symmetries of an equilateral triangle form a group consisting of six elements (three reflections and three rotations), and the shuffling of a deck of 52 playing cards forms a group that has about 8×10^{67} elements (the different ways in which the cards can be arranged). If something is fundamental to mathematics, then it is usually fundamental to physics. Indeed, the Lorentz group is central to Einstein's special theory of relativity, and the gauge group is central to the standard model of particle physics[1]. However, certain groups called pariahs were thought to have no connection to the physical world. Writing in *Nature Communications*, Duncan *et al.*[2] report the discovery of such a connection, which could have implications for both mathematics and physics.

Points on a plane are identified using their x, y coordinates. Because these coordinates are a pair of numbers, a plane can be referred to as 2-space. Similarly, we can speak of 3-space (if we include a third dimension), 4-space (if we also include time), and so on. Groups can act on n-space (where n is any number between 1 and infinity) by, for example, rescaling, rotating or reflecting points. These actions, known as representations, are well understood and computer-friendly, and feature in many areas of mathematics and physics. For example, every particle in high-energy physics corresponds to a representation of the Lorentz group[1].

Humans think reductively: we understand something complicated in terms of its basic components. Like the clicking together of Lego blocks, a large group can be obtained by clicking together smaller (usually simpler) groups. We do this by putting the smaller groups side-by-side, and then allowing one-way communication between them — analogous to fitting the prongs of one Lego block into the underside of another. The archetypal example is the addition of multi-digit numbers: when adding together 27 and 45, we first add 7 and 5 in one column to get 12, 'carry' the 1 and then add 1, 2 and 4 in a second column. In doing so, we click together two copies of what is known as the addition modulo 10 group (one copy for each column), with one-way communication taking place through the 'carry' process.

Just as we can write any number as a product of prime numbers (for example, $60 = 2^2 \times 3 \times 5$), we can write any group as a clicking together of so-called simple groups. To some extent, group theory can be reduced to understanding the simple groups (the Lego blocks) and the different ways that they can be clicked together. One of the great accomplishments of twentieth-century mathematics was the determination of the complete list of simple groups that contain a finite number of elements[3]. Almost all of these groups belong to one of 18 'infinite families' — for example, the nth simple group in one of the families consists of half of the ways in which n playing cards can be arranged. But there are also 26 isolated groups called the sporadics.