EQUINOX GRAPHICS/SCIENCE PHOTO LIBRARY

A representation of the gene regulatory network of the bacterium *Escherichia coli*, showing the interactions that control gene expression.

GENE CIRCUITS MADE SIMPLE

Tools that untangle a cell's wiring let researchers find key regulators of behaviour. **By Jeffrey M. Perkel**

When it comes to dissecting how a cell's regulatory circuits are wired, some researchers turn to their pipettes. Emily Miraldi turns to her keyboard.

A computational and systems biologist at Cincinnati Children's Hospital in Ohio, Miraldi uses mathematics to understand what makes cell systems tick, and to predict how they respond to their environment. As a postdoc, she worked with computational biologist Richard Bonneau and immunologist Dan Littman at New York University in New York City. In 2006, Bonneau and his colleagues built a computational modelling tool called the Inferelator[1] that uses gene-expression data to deduce how DNA-binding proteins called transcription factors control the expression of particular genes. Researchers can use the resulting network maps to track the flow of information through the cell, identifying – and perhaps reverse-engineering – the regulators that control key processes.

But inferring the structure of these circuits is complicated. Even the simplest gene-expression data can be explained by multiple network architectures, and interactions that seem direct might not be. Transcription factors often work in concert, are modified by enzymes and can act tens or hundreds of thousands of DNA bases away from their target gene. Although some 1,600 transcription factors have been identified in the

human genome, information on the exact sequences (or 'motifs') where they bind to DNA is lacking for many. Furthermore, genomic DNA in the cell is packaged with proteins in a complex called chromatin, which can stop transcription factors binding.

To resolve some of these issues, Bonneau's team folded in another type of experimental data to improve the Inferelator. They used information from a technique that reveals which regions of chromatin in the genome are unpackaged and available for transcription-factor binding. The method is called ATAC-seq – assay for transposase-accessible chromatin with high-throughput sequencing. By reconfiguring the software to use these data, the team were able to work out which genes changed expression in tandem, and which transcription-factor DNA-binding motifs were available to influence that expression.

In what Bonneau, now at Genentech Research and Early Development in South San Francisco, California, calls a "tour de force" study[2], Miraldi and her colleagues used this updated Inferelator to trace networks comprising thousands of transcription factors in a class of white blood cells called type 17 T-helper cells. They found that the transcription factors STAT3 and FOXB1 in these cells are key regulators of genes that are implicated in inflammatory bowel disease.

"This paper was the first time where we were able to validate that if you start with just RNA-seq and ATAC-seq [data], you can get a more accurate gene-regulatory network relative to gene-expression data alone," Miraldi says.

Today, the Inferelator is just one of a fast-growing collection of software tools for gene-regulatory network (GRN) inference, whether at the level of populations or individual cells. These might rely on gene-expression data alone, but some exploit other data types or simulate systematic disruption of regulatory networks. Others are helping to tease out the sequences that direct transcription-factor activity. If you want to predict the behaviour of cells, Miraldi says, "you need to understand how they're wired".

A matter of inference

Researchers can tease out regulatory networks experimentally. Using methods such as chromatin immunoprecipitation (which uses antibodies to identify where and when transcription factors bind to DNA) and gene-expression analysis, for instance, researchers can correlate transcription-factor binding with gene expression, and identify the DNA regions where they act. From there, they can build networks to explain the data. But these methods are labour-intensive, and might require antibodies that either haven't been made or are of poor quality. They tend to focus on a single protein at a time. And the cell type of interest might be unavailable or impractical to obtain in the laboratory. GRN inference allows researchers to circumvent these issues by mining gene-expression data to deduce these networks computationally. The resulting networks can then inform experimental design, which in turn can refine computational models.

The simplest approaches to GRN inference rely on correlation – the tendency of the expression of pairs of genes to rise and fall in sync. "If I see that from cell to cell these two genes always go up and down together, they always correlate, then there is a high chance that there is a regulatory relationship between them," says Xiuwei Zhang, a computational scientist at Georgia Institute of Technology in Atlanta, who has built her own GRN-inference tools.

Another GRN-inference tool ,called SCENIC+, exploits machine learning, says Seppe De Winter, a PhD student at the Catholic University of Leuven (KU Leuven) in Belgium, who helped to develop it. Alternatively, researchers can reduce GRNs to mathematical equations. In January, Joanna Handzlik, then a computational-science graduate student at the University of North Dakota in Grand Forks, used a modelling approach called gene circuits – a system of coupled differential equations, each of which describes a single gene – to deduce the regulatory relationships between a dozen transcription factors and target genes involved in blood-cell maturation[3].

"We saw that the model actually agreed with what would be experimentally expected."

Because such models are computationally intensive, researchers tend to simplify them by incorporating fewer proteins or reducing them to Boolean systems, in which each interaction is either on or off. Instead, Handzlik threw computational power at the problem. She ran 100 computer-processing cores on the university's high-performance computing cluster in parallel for days, solving the equations tens of millions of times until she arrived at a set of parameters for her model that mirrored experimental data. Then, Handzlik simulated what would happen if she eliminated or reduced the expression of either of two transcription factors, called PU.1 and GATA1. "We saw, remarkably, that the model actually agreed with what would be experimentally expected," she says.

'A-ha' moment

Aviv Regev, a pioneer in single-cell biology who is now executive vice-president of Genentech Research and Early Development, has spent most of her career pursuing GRNs. One of the motivations that has driven her team to design ever-more-subtle methods for processing and profiling single cells, she says, "was derived from how important that topic was to me".

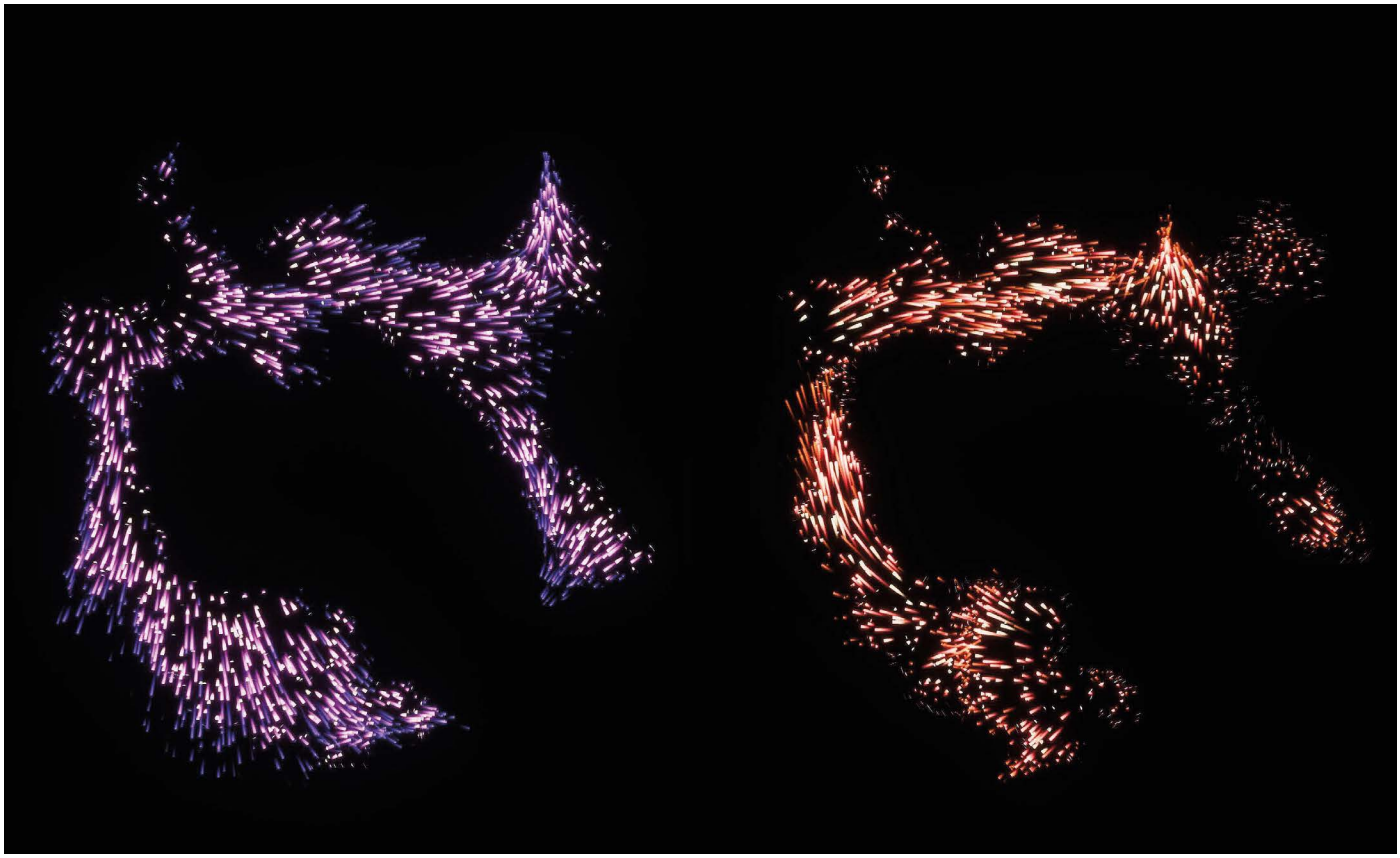
Suppose, she says, that you perturb a single gene in a population of cells. By observing which genes are affected, you can model a regulatory circuit. But to confirm your hypothesis, you might need to disrupt dozens or even hundreds of other genes. That quickly becomes impractical, she says – but not at the single-cell level, where each cell is its own data set. "We thought that in single-cell genomics we would be able to do something that we were simply not able to do in bulk."

Regev and her team applied single-cell methods and new computational approaches to study how a sample of 18 specialized immune cells from bone marrow, called dendritic cells, respond to a component of bacterial cell walls. Those 18 cells, they say, actually represented two populations. Focusing on the larger subpopulation, they discovered that although all were stimulated with the bacterial molecule at the same time, not all had responded to the same extent. Exploiting that subtle variation between the cells, the team deduced a simple related circuit that marked the transcription factors STAT2 and IRF7 as 'master regulators' of antiviral activity[4]. "You can do quite a lot just from this variation between single cells," she says.

For Anthony Gitter, a computational biologist at the University of Wisconsin–Madison, Regev's work represented an 'a-ha' moment. By examining each single-cell profile for clues to their relative position along a cell-differentiation pathway, he saw, it would be possible to organize them chronologically in 'pseudotime'.

"Pseudotime allows you to order cells so you can see which causes precede effects," Gitter says. It attempts to "estimate a time point for each cell by using the expression measurements of that one cell relative to the others". Researchers can then use those pseudotime estimates to build GRNs.

Gitter's team created a tool called SINGE based on this idea[5], and applied it to mouse embryonic stem cells as they developed into endodermal cells. It worked, but the results, he says, were underwhelming. "There still seems to be some fundamental limit on how much you can learn about gene regulation if the only data you're going to look at is gene expression." The problem, says Jason Buenrostro, co-director of the Gene Regulation Observatory at the Broad Institute of Harvard and MIT in Cambridge, Massachusetts, is that gene-expression data alone cannot sufficiently 'constrain' the number of possible networks that could explain the data. For instance, two correlated genes could be regulated by the same transcription factor, or by two different ones regulated by a third, distinct transcription factor.



KENJI KAMIMOTO/SAMANTHA MORRIS LAB

CellOracle software visualizes the gene regulatory networks that change a cell's identity. In their natural state (left panel), mouse blood cells differentiate into red (lower left) or white cells (upper left); the right panel shows the same process when a key transcription factor is deleted.

In a 2020 study, computer scientist T. M. Murali at Virginia Tech in Blacksburg and his team described a computational pipeline called BEELINE, which they used to test a dozen GRN-inference methods based on single-cell RNA sequencing against gold-standard and synthetic data sets[6]. "Most methods do a relatively poor job of inference," Murali says, at least when it comes to deducing interactions — performing about as well as a random predictor, he notes. The solution, he says, is to include extra data.

Buenrostro's team, for instance, has developed a computational framework called FigR. It uses data from single-cell RNA sequencing and ATAC-seq to integrate expression of transcription factors and their target genes with identification of protein-binding motifs and data on chromatin accessibility. "When we did that, we started to see really nicely that a lot of transcription factors that were co-expressed with our favourite gene don't actually have sequence enriched at our favourite gene." This means there's no place for the transcription factor to bind and regulate the gene, so "they get removed from the analysis", he says. "We also see lots of sequences that are enriched, but the transcription factor is not even expressed."

The latest version of the Inferelator also makes use of single-cell ATAC-seq data. But it further constrains that information by considering transcription-factor activity.

"A transcription factor's expression level doesn't indicate anything about what it's doing at the time that you observe it from sequencing data," explains Claudia Skok Gibbs, who led the development of the updated version[7]. That's because some of them act with partners, or must be chemically modified to become active. Alternatively, their binding sites might be unavailable for binding. Inferelator 3.0 looks at the expression level of target genes together with databases of transcription-factor motifs and the chromatin accessibility of potential binding sites in the genome. This means it can determine which transcription factors are available to stimulate or repress a target gene in a given cell type. Those activity scores are then plugged into one of three network-building algorithms.

"You could watch all the transcriptional responses at once to understand the real underlying function of the gene."

But for computational models, the more variables they incorporate the better they tend to be, Bonneau says. In many cases, that performance increase comes down to noise. To balance those competing forces, he says, the software gives a 'penalty' to each protein in the model — unless that protein seems to be active at the gene of interest. "If this transcription factor has a binding site near that target gene that is also shown to be open in the ATAC-seq data for that cell type, we say it doesn't have to pay as large a penalty."

Skok Gibbs has used Inferelator 3.0 to identify regulators in brain cells called transmedullary neurons in *Drosophila* fruit flies[8]. These neurons have several forms, and it's possible to convert one to another by altering the expression of a single gene. "I was able to show that I could find the specific transcription factor and what genes it was targeting that were responsible for this," she says.

Data on genetic variation can also inform GRN inference. Over the past decade, network biologist John Quackenbush at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, and his team have created a virtual 'zoo' of algorithms with names such as PANDA, LIONESS and CONDOR. These methods exploit a machine-learning strategy called message passing, as well as knowledge of where transcription factors could bind in the genome, to guess and then optimize a GRN. The team's most recent iteration, EGRET, uses information on genetic variants to tailor GRNs to specific individuals and cell types. It does so essentially by factoring in how sequence

variations called polymorphisms could affect transcription-factor binding[9].

The resulting networks can reveal how variants in the non-coding parts of the genome could lead to disease. In an analysis of 119 individuals descended from the Yoruba people of West Africa, Quackenbush and his colleagues showed that polymorphisms associated with coronary artery disease mainly affected GRNs in cardiac cells, and those associated with autoimmune disease affected immune cells[9]. "We see our predicted disruptions in gene regulation for disease-related transcription factors in the most relevant cell type that we looked at," says study co-author Deborah Weighill.

Knockout plans

In 2016, Regev and cell biologist Jonathan Weissman at the Massachusetts Institute of Technology in Cambridge, and their colleagues, authored a pair of studies[10,11] describing Perturb-seq, a pooled screening approach based on the gene-editing technique CRISPR. Perturb-seq allows researchers to reduce or knock out selected genes, using single-cell RNA-sequencing as a readout. Previous CRISPR-screening approaches tended either to use genetic reporters or to look at specific phenotypes, Weissman says. But a lot of biology will fly under the radar of such strategies. "Aviv and I independently hit on this idea that, with RNA sequencing, you could basically watch all the transcriptional responses at once," Weissman says. "That would give you much more information, and lead you to understand what the real underlying function of the gene was."

In one study[10], the researchers used Perturb-seq to analyse the effect of 24 transcription factors on genes involved in the stimulation of bone-marrow-derived dendritic cells. In the other[11], they targeted genes associated with a cell-stress pathway called the unfolded protein response. Since then, Regev has migrated the method into animals, and coupled it with protein quantitation in a method called Perturb-CITE-seq. Meanwhile, Weissman's team has taken Perturb-seq genome-wide, knocking down nearly 10,000 human genes in more than 2.5 million cells[12]. "So now you've sort of shaken the cell in every possible way, and you're asking, how does it respond?" Weissman says.

Alternatively, researchers can perturb genetic networks *in silico*. Kenji Kamimoto, a stem-cell and developmental biologist in Samantha Morris's lab at the Washington University School of Medicine in St. Louis, Missouri, created CellOracle, a software tool that blends single-cell RNA-sequencing and ATAC-seq data to first infer a GRN and then disrupt it. By examining changes in the resulting maps of cell fate, researchers can visualize how transcription-factor disruption can alter a cell population.

Kamimoto has applied CellOracle to systematically investigate the proteins that can reprogram connective-tissue cells so that they form other cell types, identifying factors that can substantially increase the efficiency of this transition[13]. At least 5 peer-reviewed studies and 13 preprints have used the tool as well, Morris says. In one[14], biomedical engineer Tim Herpelinck at KU Leuven and his colleagues used CellOracle to model the loss of the transcription factor Sox9 in bone development. "Knockout experiments take a huge amount of time, especially if you want to do them *in vivo*," Herpelinck says. And Sox9 is particularly difficult for such analysis, he adds, because loss of the gene is lethal in developing embryos.

Validate, validate, validate

To properly exploit ATAC-seq data, researchers must know where transcription-factor binding sites are. Usually, says Miraldi, researchers find them using what is essentially a text-matching algorithm. But in July, she and her team described another option: using deep neural networks to find these sites in ATAC-seq data. According to Miraldi, researchers can use the algorithm, called maxATAC, to simulate chromatin immunoprecipitation and DNA sequencing in rare cells for which it isn't practical to conduct such an experiment, including in samples from patients. Miraldi's team used maxATAC to implicate the transcription factors MYB and FOXP1 in a common autoimmune disorder called atopic dermatitis[15].

"You're that much more accurate in identifying transcription-factor binding sites."

The algorithm was about four times better than conventional transcription-factor-motif scanning at finding binding sites, Miraldi says. This should "directly translate to improvements in gene-regulatory network inference because you're that much more accurate in identifying transcription-factor binding sites". But it cannot find everything: maxATAC includes models for only 127 out of the nearly 1,600 identified human transcription factors.

To help close the gap, researchers can again turn to deep learning. In 2021, computational biologist Anshul Kundaje at Stanford University, California, and Julia Zeitlinger at the Stowers Institute for Medical Research in Kansas City, Missouri, described a convolutional neural network called BPNet. This uses a form of chromatin immunoprecipitation data called ChIP-nexus to learn, with single-nucleotide resolution, precisely which DNA sequences transcription factors bind to – at least in the cells for which the researchers have data[16]. The pair applied the approach to the four transcription factors used to make induced pluripotent stem cells – Oct4, Sox2, Klf4 and Nanog – and detected unexpected subtleties in how these proteins bind to DNA in stem cells. For instance, it turns out that Nanog typically partners with Sox2, but only if the protein's binding sites are spaced 10.5 bases apart, a distance that corresponds to the periodicity of the DNA helix. "Even for four very well studied pluripotency factors, we find new modes of cooperativity," Kundaje says.

Whichever GRN method you choose, at the end of the day it is only a hypothesis. Like all bioinformatics problems, GRN inference will always return an answer. But to determine whether that answer makes sense, says Morris, you need to "validate, validate, validate".

As the methods get more complicated, Regev says, the challenge becomes one of scale: at some point, it becomes impossible to test every variable and combination. "There aren't enough cells in the world," she says. But, she notes, it might be possible to design experiments efficiently enough for researchers to predict other experimental outcomes without actually testing them.

A different way of using Perturb-seq offers one solution, by looking at the effect of multiple perturbations in the same cell. In their 2016 paper[10], for instance, Regev and her team found some cells that had received as many as three CRISPR-targeting RNAs per cell. Comparing those to cells that had received just one or two targeting RNAs, they found cases in which the effects were synergistic, suggesting regulatory interactions. Such combinatorial studies, she says, are "the frontier – that's where the field is going."

And once researchers are able to work out the cellular wiring, they can tinker with it to engineer cells or repair them. "Arguably," says Buenrostro, "it's the most important problem in biology."

Jeffrey M. Perkel is technology editor at *Nature*.

1. Bonneau, R. *et al. Genome Biol.* **7**, R36 (2006).
2. Miraldi, E. R. *et al. Genome Res.* **29**, 449–463 (2019).
3. Handzlik, J. E. & Manu *PLoS Comput. Biol.* **18**, e1009779 (2022).
4. Shalek, A. K. *et al. Nature* **498**, 236–240 (2013).
5. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. *Cell Rep.* **38**, 110333 (2022).
6. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. *Nature Methods* **17**, 147–154 (2020).
7. Gibbs, C. S. *et al. Bioinformatics* **38**, 2519–2528 (2022).
8. Özel, M. N. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.05.01.490216 (2022).
9. Weighill, D., Ben Guebila, M., Glass, K., Quackenbush, J. & Platig, J. *Genome Res.* **32**, 524–533 (2022).
10. Dixit, A. *et al. Cell* **167**, 1853–1866 (2016).
11. Adamson, B. *et al. Cell* **167**, 1867–1882 (2016).
12. Replogle, J. M. *et al. Cell* **185**, 2559–2575 (2022).
13. Kamimoto, K. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.07.01.497374 (2022).
14. Herpelinck, T. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.03.14.484345 (2022).
15. Cazares, T. A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2022.01.28.478235 (2022).
16. Avsec, Ž. *et al. Nature Genet.* **53**, 354–366 (2021).