

STEM CELL (RE)PROGRAMMING: Computing New Recipes

By Sarah C.P. Williams

Most scientists seeking to turn back adult cells' developmental clocks rely on go-to recipes that—when followed just right—will yield stem cells. A dash of one reprogramming factor, a sprinkle of another, and let the mixture stew. Likewise, when researchers want stem cells to remain stem cells or, alternatively, when they want them coaxed down a particular developmental pathway, they have cocktails they turn to. Most of these recipes were concocted using trial and error over the past few years, and then they've been passed between labs. Whether they're the best ways to derive or control stem cells, or the most efficient, is unclear.

Now, by harnessing the power of big data, modeling, and computational biology, scientists are starting to write new—and potentially better—protocols for creating and maintaining stem cells, based on a better understanding of how large networks of genes and proteins interact to influence cellular development and differentiation.

"It takes time for stem cell researchers to embrace these kinds of systems level methods," says **Avi Ma'ayan, PhD**, of the Icahn School of Medicine at Mount Sinai. But as these approaches have started yielding results, he says, there's much more interest in giving them a shot.

Reasoning a Better Reprogramming Recipe

At the Hebrew University of Jerusalem, one team of researchers was getting frustrated by the low yield of the stem cell reprogramming methods they were using to coax adult cells into a pluripotent state—able to become any cell in the body.

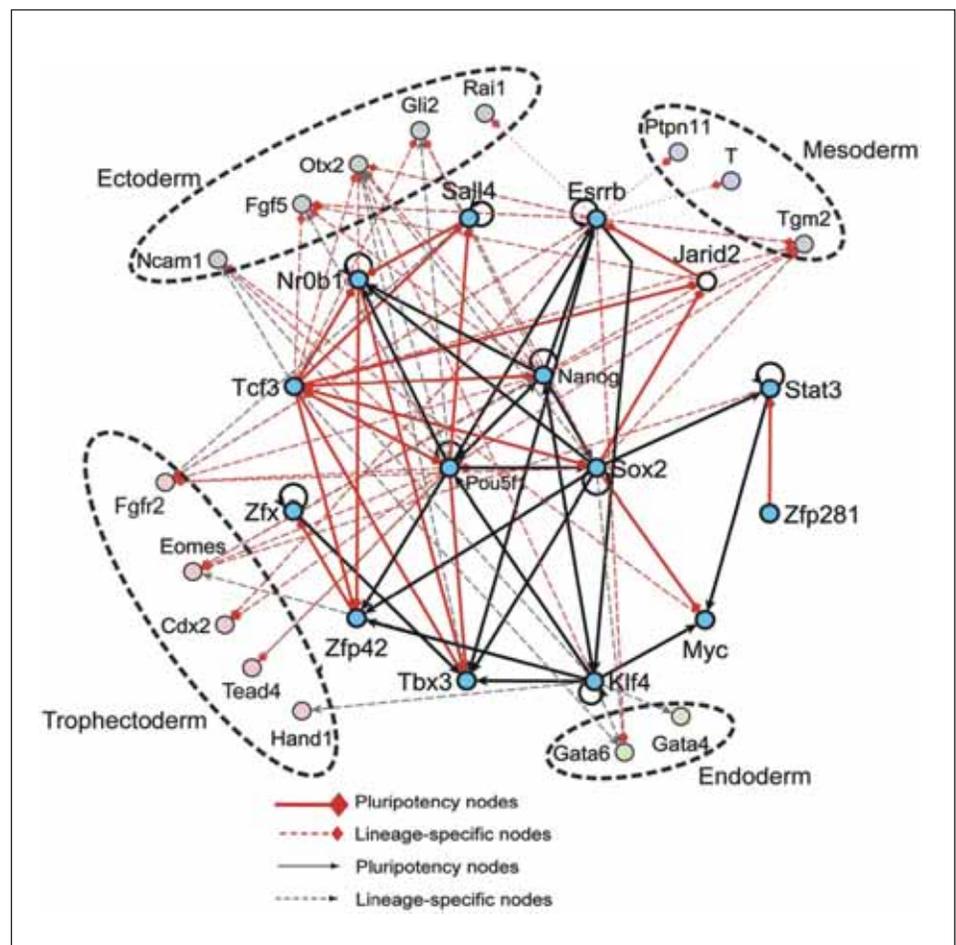
"The problem was that this program was very inefficient; only a small number of cells became stem cells, and then you'd have to use single cell technologies to capture these pluripotent cells," says **Yosef Buganim, PhD**, of the Hebrew University of Jerusalem. Moreover, once those induced pluripotent stem cells (iPSCs) were isolated, the quality of them varied. Only about 20 percent of the mouse iPSCs, Buganim says, had the capability to develop into a whole mouse—the true test for a stem cell.

Buganim's team was using a well-known mixture of transcription factors, dubbed OSKM for its four main ingredients: Oct4, Sox2, Klf4, and Myc. The scientists began to wonder whether the OSKM factors were turning on not only the genetic programs that led to pluripotency, but other programs that were contrary to this goal. So, using a combination of lab work and bioinformatics, they started figuring out how OSKM influence 48 other transcription factors that were turned on during the reprogramming process.

The final network of genes they uncovered revealed just what they wanted: four transcription factors turned on by OSKM

which could, themselves, induce pluripotency without turning on other, unwanted genetic programs. "It would take you a hundred years if you just tried culturing cells with all these different combinations of factors," Buganim says. But with bioinformatics, they could analyze the gene expression patterns much more quickly.

When Buganim's group used the new mixture—SNEL for Sall4, Nanog, Esrrb, and Lin28—on adult mouse cells, they were able to generate higher quality iPSCs than ever before. Eighty percent could generate a whole mouse that could live for more than year, Buganim says. The results were



Ma'ayan and his collaborators connected 15 known pluripotency regulators to 15 lineage markers in this network, which shows how various combinations push the cell toward four different fates (circled in dotted lines). Reprinted from Xu H, Ang Y-S, Sevilla A, Lemischka IR, Ma'ayan A (2014) Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells. PLoS Comput Biol 10(8): e1003777. doi:10.1371/journal.pcbi.1003777.

published in *Cell Stem Cell* in September.

Now, his team is focused on uncovering how even more genes interact with the pluripotency program. Rather than analyzing just 48 genes already suspected to play a role, they're using big data to look at all the genes in the cells as they reprogram from adult to iPSC.

"We have the technology to probe the transcriptome of the entire cell, and that makes the bioinformatics analysis that much more important," says Buganim. "When you're talking about 20,000 genes

programming. They then clustered genes into functional modules and studied their interactions, using a combination of tandem knockdown experiments and a novel software tool they developed: HiTSelect. The proteins they ended up validating were involved in a range of different cellular processes including transcription, chromatin regulation, vesicle-mediated transport and cell adhesion. The work was reported in a July 2014 *Cell* paper. Now, Diaz says, they are able to add shRNAs—or other molecules that selectively block

been integrated before.

"Finding the data is not very hard," Ma'ayan says, "but putting it together is a real challenge."

So his group, using a database program they developed called ESCAPE—for Embryonic Stem Cell Atlas of Pluripotency Evidence—took on this challenge. They manually collected and organized each piece of evidence to fit it into ESCAPE, then added it to their pile of evidence. The new network, a dense spider web of arrows between the 15 transcription fac-

"I think it's going to be more and more mandatory for biologists to have some background in computation," says Diaz.

instead of 48, it would be a nightmare to analyze by hand."

Knocking Down Barriers to Reprogramming

With the growing use of bioinformatics in biology labs, Buganim's group isn't the only one using computational modeling approaches to work out better reprogramming recipes. Aaron Diaz, PhD, an applied mathematician at the University of California, San Francisco, recently used a massive library of short hairpin RNA (shRNA) to selectively block genes in cells as they were being reprogrammed toward pluripotency. Diaz and colleagues then analyzed the results of this genome-wide screen, using systems biology and bioinformatics approaches, and discovered key pathways regulating the transition to pluripotency.

Each shRNA—from a library of shRNAs targeting more than 19,000 genes in human cells—was packaged inside a viral particle that had a unique barcode added. Then, each of the hundreds of thousands of unique viruses were added to human fibroblasts. Using the classic OSKM technique, Diaz and colleagues then coaxed the cells to become iPSCs. If a cell contained a shRNA that blocked a gene necessary for reprogramming, it would fail to turn into an iPSC. Using high-throughput next-generation sequencing, the researchers could then determine which shRNAs were present in cells that became iPSCs, and which shRNAs were enriched in cells that failed to reprogram.

Next, they turned to bioinformatics to analyze these results and filter out off-target effects—false positives, essentially. More than a thousand genes originally appeared in the screen as influencing reprogramming, but the analysis honed the list down to about 20 that, if blocked, enhanced repro-

gramming. They then clustered genes into functional modules and studied their interactions, using a combination of tandem knockdown experiments and a novel software tool they developed: HiTSelect. The proteins they ended up validating were involved in a range of different cellular processes including transcription, chromatin regulation, vesicle-mediated transport and cell adhesion. The work was reported in a July 2014 *Cell* paper. Now, Diaz says, they are able to add shRNAs—or other molecules that selectively block

these barrier genes—to the OSKM cocktail to help lift these blocks on reprogramming. "I think it's going to be more and more mandatory for biologists to have some background in computation," says Diaz. "With advances in single-cell sequencing, for example, it is becoming routine for us to generate hundreds of genome-wide profiles per experiment. There's just no way you can analyze that many datasets without modern data science approaches."

Computationally Informed Differentiation

Once iPSCs are generated, researchers want to know how to engineer the fate of these cells—either sending them down a pathway to become brain, blood, bone, or any other type of somatic cell, or keeping them dividing as stem cells. Looking deeply at broader cell networks by analyzing the expression levels of many genes can definitely help, says Ma'ayan. "There's a lot of excitement around using bioinformatics to improve differentiation protocols," he says.

In an August *PLoS Computational Biology* paper, Ma'ayan and collaborators described a new model of how 15 different transcription factors and 15 lineage markers interact with each other to influence the differentiation of stem cells. Other researchers, Ma'ayan says, have generated a plethora of data on these transcription factors—using techniques ranging from cDNA microarrays, RNA-seq, chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq), mass spectrometry proteomics and phospho-proteomics, and RNAi screens. But the data from all these approaches has never

tors and 15 lineage markers, shows how the increased expression of one factor can push a cell toward one of four different fates: ectoderm, mesoderm, trophoectoderm, and endoderm; this network was then validated experimentally in living cells by knocking down individual or combinations of factors and then measuring the changes in expression of the rest of the nodes in the network model.

"We didn't find anything earth shattering, but now we have a global framework to work from," Ma'ayan says. "In principle, if the model works and it becomes predictive and large enough, we can use it to improve differentiation protocols and reprogramming strategies."

Researchers agree that the future of stem cell research—and therapeutics based on stem cells—requires the ability to quickly and efficiently create personalized stem cells from a patient's own adult cells, and then coax these iPSCs into whatever

"There's a lot of excitement around using bioinformatics to improve differentiation protocols," Ma'ayan says.

healthy cell is needed by that patient. To meet that end, though, the field needs more predictable methods to direct stem cell fates. Computational models are helping to achieve this goal. □