GENETICS

For every protein its tag

New research explores *in vivo* protein function in the worm—at the genome scale.

It was one thing to sequence the entire genome of an organism using methods that don't require years and millions of dollars, but it's another, and not smaller, feat to understand how the instructions encoded in those sequences are actually being read by each of the cells in an organism.

Between the code and the protein stand multiple interconnected levels of gene expression regulation—transcription, splicing, post-translational modifications and so on—that make it hard to predict a protein's identity, expression dynamics or localization from its genomic signature, particularly in multicellular organisms.

A solution to this problem is to genetically label the protein of interest with a tag (fluorescent or affinity based) so that one can see exactly where it is *in vivo* at a given time or use affinity reagents to pull out other molecules it hangs out with. Mihail Sarov, Anthony Hyman and their colleagues at the Max Planck Institute in Dresden, Germany, have taken the time to perform this type of genetic protein tagging at a scale that approaches covering the entire genome of *Caenorhabditis elegans*.

The platform—called the '*C. elegans* TransgeneOme'—has now been released to the community through a dedicated web application (http://transgeneome.mpi-cbg.de/) (Sarov *et al.*, 2012).

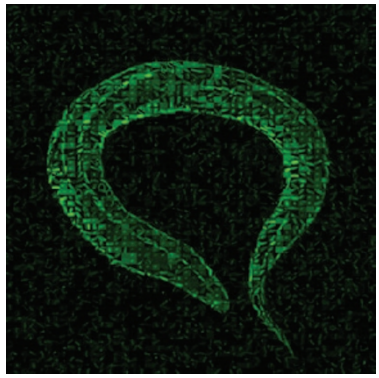
Of the approximately 20,000 protein-coding genes in the *C. elegans* genome, currently over 16,000 are covered by a fosmid clone—a genomic DNA construct that is sufficiently large to include all the important coding and regulatory sequences of a gene. The group set out to tag the genes in these fosmids and then generate stable transgenic worm lines with them.

"Initially the challenge was basically the [engineering] method," recalls Sarov, "how to be able to handle multiple steps of engineering efficiently … and to handle everything in liquid culture so that you don't have to plate everything in agar and pick single colonies." Over the years, Sarov, Hyman and their colleagues have optimized the process to enable efficient insertion, at any desired position, of a tag into a large genomic DNA fragment, and to do so at the throughput required for genome-scale coverage. The pipeline was first worked out using bacterial artificial chromosomes and mammalian cells (Poser *et al.*, 2008); now the group extends this methodology to tag proteins in whole worms.

After testing a lot of protein-tag combinations, Sarov and his colleagues found a good all-purpose label and tagged every single protein at the C terminus with the same fluorescent and affinity cassette, "accepting that for some proteins we would fail," he says. The group managed to tag 98% of the genes covered by a fosmid (which represent about 80% of the worm's genome in total).

Ultimately, one wants to know what every protein in an organism does and to understand each from a systems point of view, says Sarov, "but we still have the bottleneck of generating all the lines." Freely providing the collection of transgenes to the community while requesting that any generated transgenic worm lines be sent back to them for further analysis, the group is hoping to speed up this process with community support.

Although there will undoubtedly be some proteins that won't tolerate a C-terminal tag, the researchers are finding that the tags are benign to most proteins in the organism. However, if required, an N-terminal tag can be used instead. In the future, they also hope to use site-specific or targeted genome engineering methods to control where the transgenes are integrated and to extend the worm resource to achieve full genome coverage.



Mosaic of images from the *C. elegans* TransgeneOme project. Image courtesy of M. Sarov.

The resource got its field testing in the modENCODe project, where it was used for high-resolution mapping of transcription factor binding and localization. The data that Sarov and his colleagues have collected so far can be explored using the same web application. "We want to generate 4D positional maps for every protein in the lines and to go into protein interaction proteomics," says Sarov. In addition, his team is currently building similar types of resources for the fly, human cell lines and mouse embryonic stem cells.

After many years of technology optimization, Sarov is eager to move more toward the biology. "I'm interested in the bigger picture," he says. "The resource is a platform for discovery, and the unexpected is what we are most excited about."

Erika Pastrana

RESEARCH PAPERS

Poser, I. *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409–415 (2008).

Sarov, M. *et al.* A genome-scale resource for *in vivo* tag-based protein function exploration in *C. elegans*. *Cell* **150**, 855–866 (2012).