Scaling up disease model discovery

In December of 2002, fresh on the heels of the successful Human Genome Project, the Mouse Genome Sequencing Consortium published the full sequence of the C57BL/6J mouse genome. During that time, at a meeting in the Banbury Center of Cold Spring Harbor Laboratories, leaders in genomic research were faced with a difficult question: now that we have the mouse genome, what do we do with it?

To help answer that question, a large-scale effort to knock out each of the mouse's ~20,000 protein-coding genes was begun, and eventually matured into the International Mouse Phenotyping Consortium (IMPC), a global collaboration of researchers and institutions with expertise in mouse biology, genetics, and bioinformatics, and a common goal: to understand the function of mammalian genes and their role in disease.

"The development of the IMPC was really driven by the community," says Terry Meehan, a biochemist turned bioinformatician and member of the IMPC, "because the mouse biologists saw that by working together on a large resource project like this, there would be benefits to everyone. So the IMPC really sprang right up from the mouse genome project."

In addition to systematically generating knockout mouse lines, the IMPC uses a high-throughput and standardized phenotyping platform across institutions to compare wild-type and mutant animals. Reporting in *Nature Genetics*, the group now presents results from 3,328 mutant mouse lines and the discovery of new models for 360 human diseases (*Nat. Genet.* doi: 10.1038/ng.3901, published online 26 June 2017).

In addition to the sheer scale of the paper's data and analysis (3,328 genes knocked out, 28,406 phenotype annotations, and ~20 million data points), the results have a unique quality that only a consortium-style operation could produce: a hypothesis-free

and unbiased resource for discovering novel disease models.

"A key benefit to a non-hypothesis-driven research project is that we are free to find phenotypes that a researcher wouldn't have suspected or that they weren't interested in finding, so it could be great for novel disease allele discovery," says Damian Smedley, IMPC member and corresponding author of the new paper. For disease model generation, "hypothesis free" translates to "disease agnostic," which could be a boon to those in the orphan disease community where research models are difficult to come by. "We're hoping that for a number of orphan diseases, a gene that's associated with that disease will have a mouse model in the IMPC with phenotype data that will be useful," says Smedley. Example orphan disease models from their latest results include those for Bernard-Soulier syndrome type C, Bardet-Biedl syndrome-5, and Gordon Holmes syndrome.

But this agnostic approach to disease modeling comes with significant technical hurdles, even beyond the logistics of knockout mouse generation and phenotyping on such a large scale (Meehan notes that it took nearly 18 months of phone calls and meetings to standardize the pipelines across the IMPC). Once data is in hand for thousands of mutant mouse lines, the IMPC has to figure out what to do with it. Unlike a typical knockout mouse project, where a lab already has a specific gene target, disease state, and tailored set of phenotypes in mind, the IMPC has to sort through a mountain of standardized data and decide which mouse model best aligns with human pathologies. This is where the efforts of bioinformaticians like Meehan and Smedley kick in.

Using a previously published analytical tool, PhenoDigm (*Database* **9**, bat025; 2013), developed by Smedley and colleagues—including the Monarch Initiative, which maps phenotypes between humans and

other animals—the IMPC takes phenotype data and ranks their mutant lines as models for human diseases.

"The goal is to use phenotype mapping along with whole genome variant prioritization and filtering to land on the most likely causative variant along with the best model," says Smedley. "But really, the secret sauce of the software is using these phenotype comparisons." Successful implementation of these analysis tools requires coordination with other teams and databases outside of the IMPC, including, among others, the Mouse Genome Informatics database at Jackson Laboratories, and the Human Phenotype Ontology (also developed by the Monarch Initiative).

With the next 5-year phase of funding now kicking in, the IMPC will continue to generate and phenotype mutant mice for the remaining protein-coding genes, as well as move into new territory, such as conducting longitudinal studies on specific models of interest, and adopting new technologies like automated behavior analysis from home cage monitoring. With good experience under their belts, Meehan is confident in the IMPC's ability to efficiently move forward into uncharted waters. "Because of the work done in the previous phase, we know we to have the phone calls and we have to have the meetings to standardize protocols and to break them down into data parameters that will work for our goals."

In addition to the mouse models being generated by the IMPC, all of which are available to the research community, Meehan is hopeful the phenotype data alone will continue to help drive research forward. "Increasingly we're seeing people who just use our data because it's all freely available, so in those specific use cases where that postdoc needs that last bit of evidence, they can find it on the IMPC portal." (IMPC portal: www.mousephenotype.org).

Dustin M. Graham