Customer case study
Field of study: Computational biology

Rockefeller University Scientist Builds Mutation Analysis Tools with HGMD

In New York, Yuval Itan is developing new tools to help scientists sort through exome and genome data and find disease-causing mutations more rapidly. HGMD® provided the foundation for tools designed to improve gene- and mutation-level data analysis

At Rockefeller University in New York, research associate Yuval Itan has built two computational tools that may fundamentally change how scientists associate genetic mutations with human disease.

As a member of the Casanova lab, Itan puts his computational biology background to use in the effort to understand the genetic underpinnings of Mendelian disease. The lab focuses primarily on children, tracking genetic mutations that confer heightened susceptibility to diseases such as flu or tuberculosis.

Itan and his colleagues use whole exome and whole genome sequencing to analyze patients, mostly with severe cases of infectious disease that are suspected of having a genetic component, often comparing those patients to healthy parents or to other patients with similar phenotypes to find relevant mutations. "I'm interested in correlating the specific genotypes of the patients to the specific phenotypes they are displaying," he says. "We deal with huge data — it's really a needle-in-the-haystack problem."

In Itan's view, the scientific problem is just as much about finding the right needle as it is about not getting distracted by the rest of the haystack. In the quest for a causative mutation, chasing down unrelated mutations is costly, both in time and resources, he says. His goal in two recent projects was to develop tools that would significantly improve the accuracy of mutation prediction and effectively remove as many false leads as possible. These tools will enable other scientists to sift through the massive numbers of mutations revealed in any exome or genome and efficiently home in on the mutation of interest.

To accomplish this goal, Itan turned to a resource he trusts: Human Gene Mutation Database (HGMD), a carefully annotated collection of data associated with observed mutations. HGMD contains all known human inherited disease mutations, allowing scientists to determine quickly whether a variant is novel or has been reported previously and to discover the genetic basis of a disease.

HGMD proved to be just the repository Itan needed to build the new tools that will be freely available to the genomics community: the gene damage index and the mutation significance cutoff. Both tools provide scientists a streamlined workflow to find causal mutations as well as greater confidence in their results.

GDI

The gene damage index (GDI) was born from the observation that highly mutated genes rarely harbor the kinds of disease-causing changes that Itan and his fellow scientists are looking for when they dive into the variant lists generated from exome or genome sequencing. "Genes that are highly mutated in healthy individuals are very unlikely to cause the severe

diseases we're investigating," Itan says. "A big proportion of the mutations we find belong to genes that are naturally very highly mutated in the general population. These genes can be removed from the analysis, and that would remove a lot of the noise."

Of course, pulling so many mutations out of the investigation requires an objective, high-confidence, gene-level approach focused on detecting genes irrelevant to disease (rather than methods for detecting relevant genes such as RVIS and de novo excess) — an approach that didn't exist when Itan began looking. He reasoned that a careful analysis of how mutated each gene was would provide the foundation for this approach.

Itan used HGMD to help with this gene-level analysis. He began by detailing the accumulated mutational damage in each human gene, assigning a GDI score. The higher the score, the more heavily mutated a gene is in healthy individuals — and therefore less likely to be associated with disease causation. Once each gene was indexed, Itan had to determine an appropriate threshold between the potentially disease-causing and the overly mutated genes. "I took all the disease-causing genes from HGMD and looked at the profile of the GDI for genes that are known to be disease-causing," he says. "Then I set the cutoff above which GDI would not be disease causing."

Evaluations of the GDI approach show that "these genes can be very safely removed from the analysis of patients," Itan says, noting that the method allows for stripping out up to 60 percent of the thousands of mutations detected by exome or genome sequencing. Using this method "massively increases the chance that investigators will see the true mutation," Itan says. "In many cases mutations that seem very promising are not disease-causing and can be a waste of time."

MSC

For the other tool, Itan zoomed down from the gene level to the mutation level. Aware that many scientists use tools such as CADD, PolyPhen-2, or SIFT to determine whether a mutation will be deleterious or not, Itan looked into these widely available tools and found a disturbing trend: many mutations that were known to cause disease and that are not extremely rare in the general population were predicted to be benign. "It's very risky to remove mutations from the analysis based on these predictions," he says. "If you remove the true mutation, then your project is dead."

Itan believes the problem lies in a uniform benign/damaging cutoff for all human genes, based on a single instance per known disease-causing mutation. "When frequency of the variant in the population is taken into account, then about 60 percent of true disease-causing mutations would be removed from the analysis," he says. The tools he assessed seemed to have a strong correlation to the rarity of a variant.

That set Itan on the path to creating the mutation significance cutoff (MSC), a tool designed to provide similar information to scientists as existing tools but with a more

> "We improved the prediction rate from about 40 percent with current methods to about 95 percent when the mutations' population frequency is taken into consideration.

nuanced understanding of the DNA in question. "There is a big difference in evolutionary pressure acting upon different human genes, but currently available methods use a fixed cutoff to predict damaging or benign," he says. With his approach, that cutoff is far more tailored, allowing him to more accurately cull the damaging mutations from all others.

Again, Itan relied on HGMD to get started. He looked up each human gene associated with disease and then checked existing tools' prediction scores for the disease-causing mutations in every gene. The scores varied significantly, meaning that an appropriate cutoff for one gene would be completely useless for another gene. For every gene with at least two known disease-causing mutations, Itan calculated a confidence interval based on the expected range of predictions and established a safe cutoff value. "I would not have been able to do it without HGMD," he says. "It was crucial. There was no other way to do this with such high quality and large scale."

That work paid off. By setting gene-specific cutoff values and incorporating those into prediction scores for mutations, "we improved the prediction rate from about 40 percent with current methods to about 95 percent when the mutations' population frequency is taken into consideration," Itan says. "Also, you can now be relatively confident that if you remove a variant in a specific gene, there is only a 5 percent chance that you have removed the true disease-causing mutation."

For both tools, Itan focused on user-friendliness, designing them for use by biologists and bioinformaticians alike. "I really hope the GDI will help people to have very clean data and to increase the discovery rate of genes that are relevant to disease, and with MSC I hope it will help to remove variants that are truly benign and help people to infer mutations that are truly disease-causing," he says.

QIAGEN Silicon Valley
1700 Seaport Blvd. Third Floor
Redwood City, CA 94063
Tel. +1 650 381 5100
Fax. +1 650 381 5190
info@ingenuity.com
www.ingenuity.com