Compiling big data in a human-centric way

May 11 2017

A depiction of the double helical structure of DNA. Its four coding units (A, T, C, G) are color-coded in pink, orange, purple and yellow. Credit: NHGRI

When a group of researchers in the Undiagnosed Disease Network at Baylor College of Medicine realized they were spending days combing

through databases searching for information regarding gene variants, they decided to do something about it. By creating MARRVEL (Model organism Aggregated Resources for Rare Variant ExpLoration) they are now able to help not only their own lab but also researchers everywhere search databases all at once and in a matter of minutes.

This collaborative effort among Baylor, the Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital and Harvard Medical School is described in the latest online edition of the *American Journal of Human Genetics*.

Big data search engine

"One big problem we have is that tens of thousands of human genome variants and phenotypes are spread throughout a number of databases, each one with their own organization and nomenclature that aren't easily accessible," said Julia Wang, an M.D./Ph.D. candidate in the Medical Scientist Training Program at Baylor and a McNair Student Scholar in the Bellen lab, as well as first author on the publication. "MARRVEL is a way to assess the large volume of data, providing a concise summary of the most relevant information in a rapid user-friendly format."

MARRVEL displays information from OMIM, ExAC, ClinVar, Geno2MP, DGV, and DECIPHER, all separate databases to which researchers across the globe have contributed, sharing tens of thousands of human genome variants and phenotypes. Since there is not a set standard for recording this type of information, each one has a different approach and searching each database can yield results organized in different ways. Similarly, decades of research in various model organisms, from mouse to yeast, are also stored in their own individual databases with different sets of standards.

Dr. Zhandong Liu, assistant professor in pediatrics - neurology at Baylor,

a member of the Jan and Dan Duncan Neurological Research Institute at Texas Children's and co-corresponding author on the publication, explains that MARRVEL acts similar to an internet search engine.

"This program helps to collate the information in a common language, drawing parallels and putting it together on one single page. Our program curates model organism specific databases to concurrently display a concise summary of the data," Liu said.

Supporting researchers

A user can first search for a gene or variant, Wang explains. Results may include what is known about this gene overall, whether or not that gene is associated with a disease, whether it is highly occurring in the general population and how it is affected by certain mutations.

"MARRVEL helps to facilitate analysis of human genes and variants by cross-disciplinary integration of 18 million records so we can speed up the discovery process through computation," Liu said. "All this information is basically inaccessible unless researchers can access it efficiently and apply it to their own work to find causes, treatments and hopefully identify new diseases."

Collaboration

This project started as a necessity for the Model Organism Screening Center for the Undiagnosed Disease Network at Baylor, but as it grew, the group began reaching out to researchers in different disciplines for feedback on how MARRVEL might benefit them.

"This program is just the start. I think our tool is going to be a model for us to help clinicians and basic scientists more efficiently use the

information already publicly available," Wang said. "It will help us understand and process all of the different mutations that researchers are discovering."

"The most exciting part is how this project is bringing so many different researchers together," Liu said. "We are working with labs we might not have normally collaborated with, trying to put together a puzzle of all this data."

Both Wang and Liu are thankful to the contributions from the genetics communities allowing them access to the databases as they developed MARRVEL.

Others who contributed to the findings include Drs. Rami Al-Ouran, Seon-Young Kim, Ying-Wooi Wan, Michael Wangler, Shinya Yamamoto, Hsiao-Tuan Chao, and Hugo Bellen (Howard Hughes Medical Institute at Baylor) all with Baylor College of Medicine; Yanhui Hu, Aram Comjean, Stephanie E. Mohr, and Norbert Perrimon (Howard Hughes Medical Institute at Harvard Medical School) all with Harvard Medical School.