# New study uses computer learning to provide quality control for genetic databases

July 19 2016, by Kathryne Metcalf

DNA doesn't exist in a vacuum: even though every cell contains the entire genome of its host organism, they know how to differentiate, to become part of an eye, or a bone, or a leaf. These differences are related to each cell's transcriptome—the array of messenger RNA (mRNA) that describe which parts of the genome are expressed as they are translated into proteins.

A new study published in *The Plant Journal* helps to shed light on the transcriptomic differences between different tissues in *Arabidopsis*, an important model organism, by creating a standardized "atlas" that can automatically annotate samples to include lost metadata such as tissue type. By combining data from over 7000 samples and 200 labs, this work represents a way to leverage the increasing amounts of publically available 'omics data while improving quality control, to allow for large scale studies and data reuse.

"As more and more 'omics data are hosted in the public databases, it become increasingly difficult to leverage those data. One big obstacle is the lack of consistent metadata," says first author and Brookhaven National Laboratory research associate Fei He. "Our study shows that metadata might be detected based on the data itself, opening the door for automatic metadata re-annotation."

The study focuses on data from microarray analyses, an early high-throughput genetic analysis technique that remains in common use. Such data are often made publically available through tools such as the

National Center for Biotechnology Information's Gene Expression Omnibus (GEO), which over time accumulates vast amounts of information from thousands of studies.

Though this abundance of data opens the door for large and inexpensive studies, there are often issues integrating multiple data sets. For example, University of Illinois bioengineer and Carl R. Woese Institute for Genomic Biology affiliate Sergei Maslov describes, "tissue type is a major metadata point for a sample. However, different researchers use different vocabularies to describe the same tissue, [... and] errors exist during the data submission process."

Because the sheer amount of data precludes manual correction or quality control, Maslov, He and collaborators were inspired to create an automated solution that could deduce metadata from the expression profiles themselves by identifying similarities between tissue types. Their findings suggest that expression profiles remain remarkably similar between samples of the same tissue type, even when taken from plants grown under very different conditions.

By identifying the most similar samples with tissue types already annotated, researchers were able to teach their algorithm to identify other samples of the same type with an excellent degree of accuracy. The team generated over 10,000 entries of metadata, and was even able to correct some mistaken annotation in another lab's study by confirming with the original author. The end result is a massive "atlas" of well-annotated data that can be used for future studies.

"Our ultimate goal is to provide cloud-based computer infrastructure for the study of energy/agriculture related plants, such as poplar and maize," says Maslov. "If our strategies have been successfully applied on *Arabidopsis*, they can be applied on other species as well."

Meanwhile, adds He, their integrated *Arabidopsis* atlas is itself an important contribution to plant genetics. "It can be used for constructing coexpression networks, one of the popular methods to leverage transcriptome data for annotation of gene function. We hope it will become a gold standard dataset in many applications."

**More information:** Fei He et al, Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis, *The Plant Journal* (2016). DOI: 10.1111/tpj.13175

Provided by University of Illinois at Urbana-Champaign