PHYS.ORG

Supercomputing the evolution of a model flower

January 27 2015, by Jorge Salazar

Arabidopsis thaliana, a model flowering plant studied by biologists, has climate-sensitive genes whose expression was found to evolve. Credit: Penn State.

Scientists using supercomputers found genes sensitive to cold and drought in a plant help it survive climate change. These findings increase basic understanding of plant adaptation and can be applied to improve crops.

The computational biology study on the flowering mustard weed *Arabidopsis* thaliana appeared in the journal Molecular Biology Evolution in September 2014. The iPlant Collaborative and the supercomputers Stampede, Lonestar and Ranger of the Texas Advanced Computing Center aided in the research. Study funding came from the National Science Foundation (NSF) and the U.S. Department of Agriculture.

"We found pretty good evidence, certainly the best evidence to date, that the evolution of gene expression is an important way that plant populations adapt to local environments," said study co-author Jesse Lasky, an Earth Institute fellow at Columbia University.

Thomas Juenger is another co-author and a faculty member in the Department of Integrative biology of The University of Texas at Austin. The Juenger Lab has studied *Arabidopsis* thaliana for over a decade. "It's one of the model plants that biologists study," Juenger said. *Arabidopsis* has one of the smallest genomes of any plant, and in 2000 it was the first plant genome to be completely sequenced.

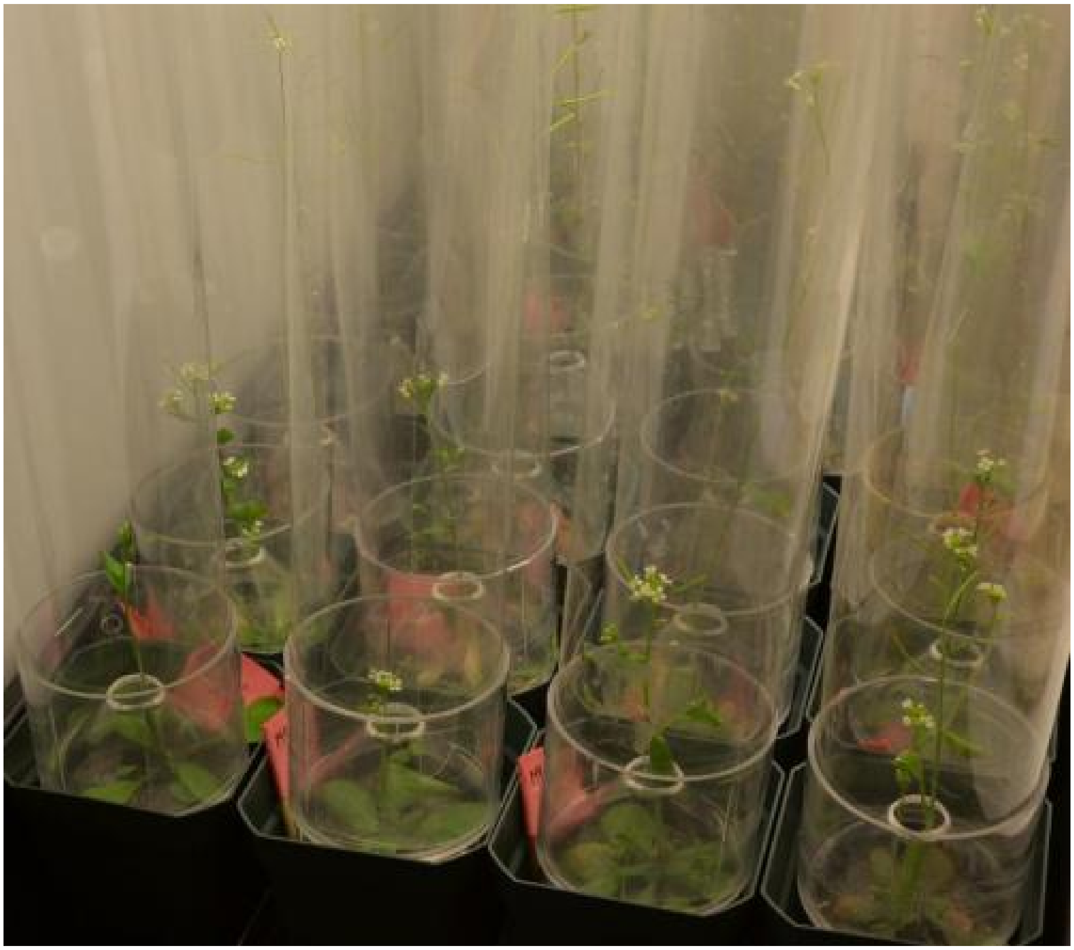
Plant biologists consider *Arabidopsis* to be like the fruit fly of their genetic research. But instead of knocking out or ramping up genes with genetic engineering, Juenger studies natural variation in genes. "We want to understand how they've evolved in response to the processes of natural selection and gene flow and mutation in the field," he said.

To date, plants have stumped scientists' understanding of how life adapts to climate, specifically the details of gene expression, which can vary wildly in a hardy plant species like *Arabidopsis* that thrives in environments as diverse as Scandinavia, North Africa, and Central Asia. Genes, or snippets of the four-letter DNA molecule, carry not only the code for which proteins make for its survival but also the instructions for how many to make, or express. Gene expression "... is the part of the organism that we show here is strongly involved in local adaptation to environment," Lasky said.

Because plants are rooted, they have to stand their ground against changes in temperature, soil moisture, and insect attacks to name a few. Juenger explained that one way they cope with environmental change is to change their gene expression.

"As a plant starts to sense dropping temperatures, a cascade of gene expression can allow the plant to acclimatize to cold temperatures, and in effect prepare itself for the coming freezing conditions," Juenger said. So his science team used prior lab work that exposed seedlings of *Arabidopsis* to artificial cold and drought stress to measure changes in gene expression across the entire genome.

Juenger described the problem of finding the right gene like finding a needle in a haystack. *Arabidopsis*' relatively tiny genome still contains over 25,000 genes. The needle Juenger's team sought was what's called a SNP polymorphism, a single letter difference in the over 100 million DNA base pairs that comprise the genes of *Arabidopsis*. "This is a fundamental challenge in biology," Juenger said. "We're looking through tens of thousands of genes to find the right ones, the few that might actually matter."



Scientists took a computational approach using the Stampede and Lonestar supercomputers to compare lab data with reference genomes of over a thousand strains of *Arabidopsis* sampled throughout Europe and Asia. Credit: Juenger Lab

The scientists took the genes they found and compared them with genomic data from previous studies that sampled *Arabidopsis* from populations throughout Europe and Asia. They narrowed that reference data to 1,003 strains of the flowering mustard weed. Of those genes that showed changes in their response to their environment, the scientists needed to know if they also showed changes in DNA along environmental gradients. Such a pattern "suggests that there are changes in the DNA sequence that are adapted to those local conditions and that are associated with changes in gene expression," Lasky said.

The research team statistically tested for associations between climate and SNP polymorphism by making the hypothesis null, or assuming no association. They did that by shuffling the data and doing permutation testing. "We can randomize climatic variation with respect to SNP polymorphism variation and do that thousands and thousands of times and ask, what sort of test statistic might we observe by chance alone," Juenger said. "We can compare that to our real, empirical data."

The computational challenges were daunting, involving thousands of individual strains of *Arabidopsis* with hundreds of thousands of markers across the genome and testing for a dozen environmental variables. "It's impossible to do this on a standard desktop computer, and it requires some of the throughput that we can have on a cluster like Stampede or Lonestar," Juenger said. "The computational time on the clusters at TACC allowed us to evaluate the hypothesis that generated from the SNP data."

Lasky added that "to run these models across the genome, you quickly run out of time. It's really just a problem where you do lots of little things many, many times. It's much easier to accomplish that when you can run that problem on many cores across a cluster. That was the challenge."

"I didn't have any experience with high performance computing before this," Lasky confided.

Lasky called on Weijia Xu, the group lead for the Data Mining and Statistics Group at TACC. "He helped me orient myself to what kind of problem I had and how to scale that up to run it on some of the clusters," Lasky said. Xu also helped by writing a parametric job launcher, which allowed Lasky to get his separate runs across the genome started more easily.

"It was a code I developed to launch multiple R jobs in parallel using an MPI interface," Xu said of the launcher. Scientists commonly use the R statistical programming language; and MPI is short for Message Passing Interface, which is a software library that breaks up large computing jobs into smaller ones to run in parallel on the nodes of a cluster.

The NSF-funded iPlant Collaborative helps life scientists use high performance computers. Juenger remarked that "iPlant, associated with TACC, has certainly been developing lots of new tools, simplifying computational tools for biologists, and giving us access to data storage as well as service units through high performance computing clusters like those at TACC. It's a helpful, timely program that's impacting plant biologists in individual labs around the country."

Lasky notes that while the results of the experiment with *Arabidopsis* are promising, more confirmation is needed. "We have experimental work here, but we haven't experimentally shown that the genes that we identified are causing localized adaptations."

More information: Study: mbe.oxfordjournals.org/content … /05/21/molbev.msu170

Provided by University of Texas at Austin

Citation: Supercomputing the evolution of a model flower (2015, January 27) retrieved 17 April 2024 from https://phys.org/news/2015-01-supercomputing-evolution.html