Open Science Grid

Open Science Grid Contributes to Genetic Diversity and Food Security Research

In their influential 1990 book, *Shattering: Food, Politics, and the Loss of Genetic Diversity*, Cary Fowler and Pat Mooney issue a warning: “Without the genetic diversity from which farmers traditionally breed for resistance to diseases, crops are more susceptible to the spread of pestilence. Tragedies like the Irish Potato Famine may be thought of today as ancient history; yet the U.S. corn blight of 1970 shows that technologically based agribusiness is a breeding ground for disaster.”

The National Center for Genetic Resources Preservation (NCGRP), a U.S. Department of Agriculture facility in Ft. Collins, Colorado, is a primary repository for germplasm (living genetic resources) and is part of a larger system of such facilities around the country. It is a gene bank—or, seed bank—holding around 570,000 specimens and 15,000 species of plants. Most of the collection is stored as the seeds of crop varieties or their wild relatives. Part of the center’s job is to find the best ways to keep the seeds alive for the long term (storage physiology).

(http://www.opensciencegrid.org/wp-content/uploads/2015/05/Richards.jpeg)

Patrick Reeves and Christopher Richards are part of a NCGRP group working on best practices for obtaining and distributing germplasm. Since this mostly involves the wild relatives of crop species, they are determining how to efficiently create collections from nature by trying to predict the geographic distribution of agriculturally-important genes.

One aspect of their work uses the Open Science Grid (OSG) to help determine genetically distinct groups within a species and understand how they are distributed across the landscape. Another aspect that uses the OSG evaluates the performance of a variety of different methods for identifying such genetic groups.

Christoper Richards: courtesy photo

We are interested narrowly in understanding the performance of genetic clustering algorithms using simulated data," said Reeves. "In this study that uses the OSG, we are trying to understand the impact of error in our data sets. It may seem trivial, but it is rather important because some of the methods for rapidly producing genome-wide data are error-prone. We need to know how that affects the accuracy with which we infer genetic clusters."

A typical study using the OSG simulates data under a variety of plausible conditions and asks whether analytical methods can return the correct answer. "This is useful because it gives us some confidence that we can apply what we learn to natural systems where we don't know the answer ahead of time," said Reeves.

Patrick Reeves: courtesy photo

A software tool that they have found valuable is InStruct (http://cbsuapps.tc.cornell.edu/InStruct.aspx), developed at Cornell University. It employs a Markov Chain Monte Carlo (MCMC) method for identifying natural groupings of individuals from genetic data.

Reeves pointed out that data sets are becoming very large. "Because it is getting cheaper to produce data, we are now dealing with genome-wide samples for many individuals from many populations. A typical data matrix might be a million or so cells wide by a couple thousand deep," said Reeves.

As data sets increase in size, the OSG becomes all the more critical. "This is kind of unique for high performance computing," said Reeves. "Typically, analyses are broken up into smaller bits that run a couple of seconds to an hour or so each." In contrast, some NCGRP analyses could run for two weeks, and researchers may have several thousand to complete. "Most high performance computing systems kick out what's running after a period of time. Our runs are long, so we had to find uninterrupted nodes."

Reeves said they started out using iPlant (http://www.iplantcollaborative.org/), another National Science Foundation-funded project, but ran into a problem—they were using too many resources. So, iPlant's Nirav Merchant put them in touch with the OSG, where they got help from Suchandra Thapa at the University of Chicago.

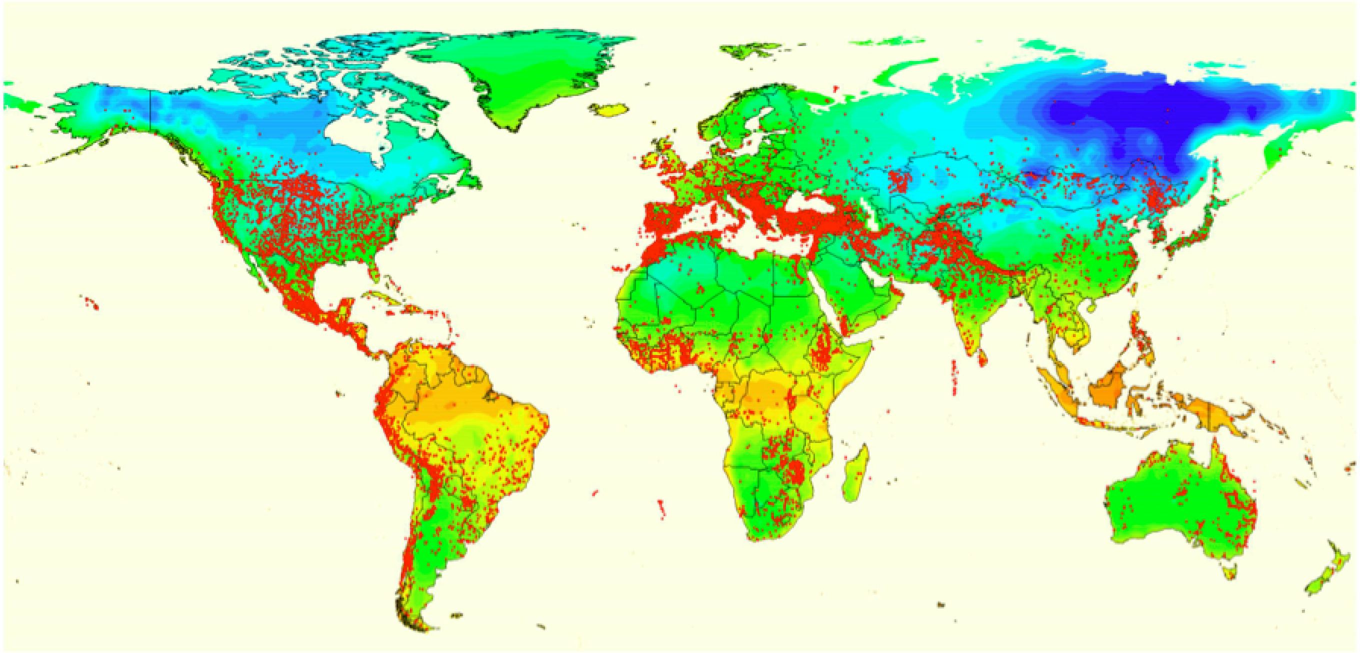
"A combination of brute force and long-run nodes make it possible," said Reeves. "Before iPlant, we used a small onsite Condor cluster with about 150 processors. We still use them, often simultaneously with the OSG, but OSG gave us many more processors that greatly increased our capacity." The project used around 12,000 OSG wall hours in the last six months alone.

"We found that OSG was easy to work with and transition to. Suchandra Thapa got us started with the right variables to get us to the right nodes and helped with limiting the number of jobs," said Reeves. "He also told us about DAGman, an HTCondor scheduler, so we could control the number of simultaneous runs automatically and not have to manually set them up."

"My recommendation for other scientists looking into using the OSG is don't be intimidated, find someone to work with, and don't hesitate to reach out," continued Reeves. "The competence and concern of the staff, and their ability to communicate with non-computer scientists are a huge help. It seems like OSG really has a service mentality."

Reeves referred to thought-leader and author Cary Fowler, who argues that crop diversity is mankind's most important creative act and maintaining food security is critical. "This is an ongoing process," noted Reeves, "and we can't rest or take it for granted. There are constantly new challenges like climate change, disease, and drought. Our society has put elaborate systems in place to maintain food security, from developing basic theory like our genetic diversity research, to breeding improved crop varieties, to farmers planting seeds."

In other words, there are many people involved in maintaining an edge over natural processes. The USDA is a part of that effort, and the work being done by Richards and Reeves is an early step in maintaining that edge.



(http://www.opensciencegrid.org/wp-content/uploads/2015/05/usda.png)*Location of source populations for the nearly 15,000 plant species (~570,000 unique accessions) stored as seeds at the National Center for Genetic Resources Preservation. Image courtesy USDA (http://www.ars-grin.gov/npgs/faq/):*

- Greg Moore

Funded by National Science Foundation (http://nsf.gov) and Department of Energy (http://science.energy.gov/)