nature genetics

Combined harvest of knowledge

Investment in national infrastructure should include a scalable open informatics solution for agricultural genomics, germplasm and crop traits. This is a priority measure for economic stimulus and food security. As building this knowledge harvester should be simpler than the infrastructure required for precision medicine, it will also pave the way to that goal.

Genomic data deposited by agricultural scientists could be used to generate research publications if they were stored in the right place, in the right format and with the associated trait information and measurements. Indeed, this journal has published more than fifty Analysis articles in the last three years, nearly all of which made new use of human genome variation and required no new data generation. Most of these reanalysis publications are above average in impact, as judged by their citations and article access. Genetic markers and gene–trait association information are also needed by agricultural breeders to generate improved crops, so provision of this information in a format useful to breeders would result in an economic stimulus.

Pilot studies are underway in many countries to store genomic and medical record information at the population level (for example, https://www.genomicsengland.co.uk/the-100000-genomes-project/).

Because human data are subject to consent, medical confidentiality and national privacy legislation, it is intrinsically more difficult to build an open infrastructure for their storage. Consequently, human research is relatively restricted, and even if genomic information is considered pre-competitive, because of its long path to translational medicine, it is relatively difficult to regard human data as an immediate economic incentive. In contrast, gene–trait information is literally seed corn for agriculture.

Unfortunately, the genomes of many crops are equal to or greater in size than the human genome (for example, wheat (*Nature* **491**, 705–710, 2012 and *Capsicum* (*Nat. Genet.* **46**, 270–278, 2014)). Maize, at about two-thirds the size of the human genome, vastly exceeds the human genome in complexity and diversity. One study on 103 maize lines found 55 million simple genetic variants (SNPs; *Nat. Genet.* **44**, 803–807, 2012). We do not yet have good graph-based storage formats for multiple genomes and their variants. For many genomes, we do not know how many reference sequences will be needed to capture the pan-genomic diversity. Thus, before the genomic information can be compressed, we will have many years in which multiple sequences will accumulate and need to be searched and computed upon. The trait information generated from field measurements by imaging, robots and drones will of course rapidly dwarf the storage required for crop genome diversity before we work out how to reduce its dimensions and extract the key measurements. Human genomic medicine has the same informatics problems, so a similar but open infrastructure for agricultural genomic research and breeding would allow common solutions to be found, through open discussion and interaction among data generators and data users, including both researchers and seed companies. Solutions could be developed in parallel to promote both human health and security via both the farming and biomedical informatics paths.

Incentives to stimulate agricultural prosperity can be local and traditional and can at the same time involve innovative solutions. For instance, Kentucky passed legislation in 2013 to enable reintroduction of the traditionally grown crop hemp for fiber, oil, nutritional supplements and pharmaceuticals (for example, those based on non-psychoactive cannabidiol; http://www.kyagr.com/marketing/history-of-hemp-in-kentucky.html). Perhaps ironically for those looking to find local solutions to resuscitate the local economy in the face of globalization, the only germplasm repository for original Kentucky hemp may be in St. Petersburg, Russia, thanks to the work of Vavilov (http://www.vir.nw.ru/hemp/hemp1.htm). Such local innovation as that of Kentucky hemp deserves national recognition at many levels but also, most importantly, the best quality informatics resources to support home-grown as well as internationally competitive science, such as the maize research discussed above. Of course, a national informatics network to track seed and gene variants for improved crop production does not preclude nationally advantageous improvements in international trade, particularly with nations that have made similar investments. Rather, these infrastructural stimuli will boost a virtuous spiral of innovation for global crop security, one country at a time. ■