nature genetics

Growing access to phenotype data

Plant genomes are the index that will allow plant breeders and researchers to access the information contained in the world's seed banks, with each allele linking germplasm, genotype and phenotype. The journal endorses the international DivSeek initiative and will work with authors to ensure access to phenotype data linked to published genetic data.

Researchers and plant breeders spend too much time searching for information, and this provides a powerful incentive to link physical and informatics resources together. If it is worth your time to collect phenotypic data and relate it to the genotype to publish, then these linked phenotypic data should be shared as part of your publication. If a seed is worth preserving among the approximately seven million crop accessions in a germplasm center (seed bank), its genomic constitution and phenotypic potential should be recorded, linked and accessible. Prediction of traits from genomic information is the shortcut to avoiding years of field trials in a very large search space, for example, among the hundreds of potato varieties or hundreds of thousands of rice varieties. Once the linked informational infrastructure is in place, it makes sense to then store and catalog the much larger range of wild relatives and local domestications (landraces) of the world's plants that have crop potential (*Nature* **499**, 23–24, 2013).

Because of these reasons, the DivSeek initiative has recently been launched (http://www.divseek.org/white-paper/) to ensure that genotype and phenotype information are stored with the seed in a form accessible to curator, breeder and researcher alike. We endorse the aims of this initiative and, to help to populate their databases with the highest quality information, we will work with authors of existing and future publications in plant genetics and genomics to curate and deposit related phenotypic information. We recommend that authors of publications rich in data resources publish a data descriptor with Nature Publishing Group's *Scientific Data* (http://www.nature.com/sdata/). We ask peer referees of current publications to help us identify submissions that will benefit from a parallel publication describing the deposition of phenotypic data.

Not all crop data can be stored at germplasm centers, nor is there yet funding for detailed phenotypic data curation and structured databasing in one place. Therefore, the data descriptor allows for data sets held in different places and for these data sets to be harmonized and linked together in a high-profile publication that explains where the data have been deposited, together with the experimental conditions. It also links related genotype data, explains how the phenotype data were defined and collected—together with the standards, definitions and ontologies used—and links to other measurements made on the same samples and populations.

We certainly do not want standardization to get in the way of data access. Indeed, measurements worth making have already been made in easily described, standard ways. Adding these metadata to the measurements will go a long way in establishing the common ways to carry out measurement and which measures to include in the data sets made accessible.

Our food security rests on too narrow a base, despite the planet's enormous plant diversity, with rice, wheat, maize and potato providing most of the energy for most of the human population. Plant improvement needs to keep pace with human population demands and climate change. The mechanistic basis for agricultural traits is also an excellent substrate for basic and translational research. The biology of trait stability is of basic interest to geneticists and has translational value in agriculture. Information on highly variable phenotypes and unstable genotype-environment combinations arising from plant breeding programs and experimental quantitative trait locus (QTL) discovery efforts based on differences in mean trait values is often unavailable to the basic research community who would be interested and capable of interpreting the underlying genetic and biological causes (*PLoS Biol.* **11**, e1001595, 2013).

A lot of effort in plant breeding is focused on yield. Yield data are contextual, as not only the environmental conditions, latitude and timing of planting and harvest influence yield, but also the relationship of other phenotypic traits. For example, in cereal grain yield, typical influences on plant morphology include tillering, branching angles, internode lengths, root architecture and depth, and leaf morphology. Each of these traits can be measured in different ways at different times and under different conditions. There are two main implications of this observation. First, the more trait information we link, the better we can predict yield. Second, an agreement on what to measure needs to be rapidly reached. We think that phenotypic data release via the data descriptor mechanism is a good way to publicize and evolve such community standards. ■