DivSeek International Network Inc (DIN)
110 Gymnasium Place Saskatoon, Saskatchewan, S7N 4J8 Canada
www.divseek.org

Submission of views and information on "Digital Sequence Information" in Plant Genetic Resources for Food and Agriculture
October 8, 2019

In preparation for the 8th Session of its Governing Body, the International Treaty for Plant Genetic Resources in Food and Agriculture (Plant Treaty) has asked stakeholders for information on digital sequence information (DSI), including: (1) terminology used in this area; and (2) types and extent of uses of DSI on PGRFA in breeding and genetic improvement, identification, characterization and conservation of plant genetic resources for food and agriculture (PGRFA). DivSeek International Network Inc (DIN) is pleased to offer its views on these issues for consideration by the Governing Body of the Plant Treaty.

The DivSeek International Network (DIN) is a global, community-driven initiative that aims to harness the genetic potential of crop diversity using science and technology. DIN facilitates networking, sharing of community standards and best practices in order to enable breeders and researchers to mobilize plant genetic variation to accelerate the rate of crop improvement and furnish food and agricultural products to the growing human population. DIN has 67 Member institutions from 28 different countries, as well as four Observers who participate in DIN meetings but do not have voting rights in the organization (http://www.divseek.org/partners).

1. Terminology Used in this Area

DIN suggests use of the term "genetic sequence data" rather than "digital sequence information."

The "data" generated by DNA sequencing refers to a collection of raw observations about the physical order of the four bases, thymine (T), adenine (A), cytosine (C), and guanine (G) in a molecule of DNA or RNA. When the data are processed, organized, and interpreted in a given context, they become meaningful and useful to users and are referred to as "information."

"Genetic sequence data" about PGR is abundant and can be accumulated, stored, processed, organized, shared, utilized and interpreted to generate many different kinds of "information."

2. Uses of Genetic Sequence Data

a. Breeding and genetic improvement.

Advances in whole genome sequencing have transformed modern breeding in the last decade by allowing researchers to more easily quantify diversity, identify genetic variants contributing to agronomically important traits, and develop genomic prediction models.

- Genomic sequence data plays an increasingly important role in genetics as it applies to all aspects of crop breeding, including understanding domestication origins and subsequent radiation of crops, varietal identification, pedigree verification, tracking of germplasm, gene discovery, marker assisted selection, and genomic selection.
- Genomic-breeding approaches have been thoroughly integrated into the breeding pipelines for most large commodity crops and are now considered essential for minor crops, as new sequencing platforms have dramatically reduced the costs.
- Genomic sequence data have created new opportunities to dissect the architecture of complex traits and infer the evolutionary history of genes and genomic regions, making it possible to trace the ancestry of specific genes and alleles back to landrace varieties and/or wild crop relatives. Identifying new traits and alleles in exotic germplasm, and introgressing those alleles into modern varieties can help adapt existing crops to new environments and changing climates. It also allows breeders to begin to harness the wealth of untapped variation that exists in wild and exotic gene pools.
- Lastly, genomic data for crops are being leveraged to build quantitative genetic models that can predict field performance based solely on genetic sequence data. Genomic selection modelling is a promising application for increasing breeding efficiency; it facilitates efforts to combine favorable traits into the same genotype more precisely and with fewer selection cycles.

As the prevalence of genomic sequence data has increased, the number, diversity and complexity of genomic and bioinformatic resources has also increased. These computational resources are essential for managing genomic sequence data, enabling researchers to systematically accumulate, store, process, organize, interpret, share, and utilize the data.

b. Identification, characterization and conservation of PGRFA.

Genomic sequence data are being widely generated for the world's gene bank holdings. DIN initiated a survey of gene bank curators from its membership. From this survey and other input from national gene banks, we documented that genomic characterization is moving from individual experiments on select accessions to more systematic screening of diversity across entire collections. There is a broad consensus that the availability of genomic sequence data aids gene bank managers in identifying, organizing and selecting accessions for users, and that these data increase the efficiency of gene bank management and will enhance utilization of collections in the future.

The utilization of genetic sequence data for gene bank curation activities can be organized into three main categories:

- *More accurate estimates of identity and integrity of an accession.* Genomic sequence data can be efficiently used to identify and discriminate individual accessions. Variation at individual nucleotides can be used to assess levels of fine scale diversity, providing curators with information about duplication, contamination and possible gaps in the collection.
- *More precise estimation of the phylogenetic relationships among and genetic history of accessions.* Understanding the genetic relationships among accessions is vital in gene bank management. Structuring the collection among species at the phylogenetic level and then subdividing the accessions into genetic population clusters has become a routine procedure in many global collections. With the advent of a high-density single nucleotide polymorphisms (SNPs) the clarity of these estimations is greatly increased.
- *More efficient development of core sets, custom reference sets and improved access to diversity.* Subsets of large collections can be identified that contain a representative fraction of the overall diversity. Subsets can either use whole genomic data or can be used to target specific user-defined regions of the genome to meet a user's needs.

c. Basic research in the biological, ecological and environmental sciences.

Genomic sequence data are essential to fundamental research in biology, ecology and environmental science, with important applications in agriculture, medicine, natural resources and energy. Fundamental research areas driven by the widespread availability of genomic sequence data include gene discovery across biological kingdoms, gene regulation (including epigenetics), gene and genome evolution, pan-genome conservation and diversity, cell, tissue and organismal development, the microbiome, ecophysiological modeling, plasticity of populations and ecosystems. These and related areas of research generate fundamental understanding that impacts the well-being of human populations and the value of PGRFA.

3. Additional Considerations Regarding Genetic Sequence Data

DIN acknowledges that different communities have different requirements for genetic sequence data, data analysis and data management tools, and for data utilization and breeding strategies. DIN supports efforts to meet community-specific needs and objectives. In addition, DIN strongly supports the need for ongoing capacity building, knowledge-exchange, and technology-sharing to augment the ability of all parties to interpret and utilize genetic sequence data in their efforts to accelerate the development of resilient and sustainable agricultural practices. The amount of genetic sequence data available for each crop or species varies considerably, and the process by which communities extract

benefit from these data also differs because they have different needs and objectives. For this reason, capacity building must be needs-based and targeted to the community and crop or species at issue.

A central aim of DIN is to facilitate the development of best practices, shared vocabularies, and community standards to enable collaboration, capacity building, and data integration, once the terms and conditions for access and benefit-sharing are agreed upon. In so doing, DIN aims to ensure that genetic sequence data are not only widely accessible, but that the benefits they provide are universally actionable.