High volume data and their integration in a clonal plant breeding program

Reinhard Simon[1], Raul Cordova[2], Vilma Hualla[2], Sofia Tejada[2], Mirella Flores[2], Elisa Salas[2], Merideth Bonierbale[2]

Centro Internacional de la Papa[1], International Potato Center[2]

Plant breeding consists in the creation and selection of new genotypes. This involves not only keeping records across generations and environments but also accommodating data of increasing resolution on genotypes, phenotypes, and growth environments. Some such high-resolution characterization methods are Near-Infrared spectroscopy, metabolomics, next-generation sequencing and high resolution spatial-temporal-spectral photos. A first need is the integration and retrieval of this information. Such an integrated and complete set can be described in breeder's terms in six dimensions: a plant phenotype (P) is the result of a genotypes (G) interaction with its environment (E) given certain field management (M) practices. In addition, data on the administrative (A) context should be kept including staff involved, objectives and, if applicable, projects and donors; as well as on data documentation standards (S) like ontologies. The latter play an important part in exchanging and aggregating information. Here we describe the adoption of the biomart database for this purpose. While biomart was developed originally to accommodate gene and sequencing data at a genomic scale we describe here how it can be used for breeding program data. This is being illustrated by current data warehousing in the potato breeding program at the International Potato Center (CIP). Particularly, genotype and phenotype can be transparently combined for further analysis in the decision process for the selection of new genotypes.