Aiding the journey from data to publication in the plant sciences

Robert Davey[1], Vicky Schneider-Gricar[1], Susanna Sansone[2], Paul Kersey[3], Jim Beynon[4], Ruth Bastow[4], Mario Caccamo[1]

[1] Presenting author: robert.davey@tgac.ac.uk

[1]The Genome Analysis Centre, Norwich, UK, [2]The University of Oxford, Oxford, UK, [3]The European Bioinformatics Institute, Hinxton, UK, [4]The University of Warwick, Coventry, UK

The development of new experimental technologies has opened the way to new, data-generative approaches to plant research, particularly within the genomics field but also high-throughput transcriptomics, proteomics and metabolomics.

Furthermore, publication models are moving away from the traditional "data late" approach, and are shifting towards the "data early", with particular pressure being made by funding agencies to see data publicised quickly. Alongside the wealth of these large plant science datasets held in public and private laboratories around the globe, there are a large number of tools to help researchers disseminate, analyse and publish those datasets. However, the disparate nature of the tools, data formats and scientific problems in light of this expansive experimental development has resulted in a lack of interoperable, production-quality software available for data analysis and dissemination.

We present a newly funded project, **Collaboratively Open Plant Omics** (COPO), to address this disparity in interoperability and easy access to important services in the 'omics data realm. We will develop a framework to utilise existing services to facilitate the description, deposition and publication of datasets, but also to enable the identification and citation of datasets, thereby increasing transparency and reproducibilty. Promoting reward for making data available is a central aim of the project.

By developing high quality, stable Application Programming Interfaces (APIs) and virtualised resources we will be tying together existing services such as: the ISATools metadata suite; EBI data repositories; Galaxy and iPlant analytical platforms; figshare, Research Object, Scientific Data and Gigascience platforms; to:

(i) develop and use community-accepted standards for data representation
(ii) facilitate submission to persistent archival resources, for data publication and citation
(iii) enable seamless transition from data to analysis platforms
(iv) provide aggregated provenance and suitable publication markup to link citable resources

All project code will be open source under an appropriate licence, such as GPL, LGPL or BSD, and made available from project outset on TGAC's GitHub repository.