Hackathon: Initiating Linked Public Datasets for Agriculture and Food

Date: Wednesday 6th – Thursday 7th March 2019
Venue: Garden Room. EMBL-EBI Main Building, Wellcome Genome Campus, Hinxton, Cambridge, UK

Focusing on wheat, can we gain field-level insight from molecular data?

Molecular based plant breeding is widely used to develop climate and disease resilient crops. Production of healthy, high yielding crops depends on crop plant genetics and detailed knowledge of how this influences physiological responses to the environment, including the pathogens and pests that the crops encounter.

Public research effort has already generated several crop and pathogen genomes as well as data about the molecular interactions between crops and major disease pathogens. This has contributed to marker assisted crop genetic improvement and provided targets for the development of disease and pest control agents. Data to inform about the field, weather and cultivation practice conditions in which crops are grown is also increasingly gathered and being made openly available. Concurrent effort has been devoted to developing public tools and resources to manage and link these big data in meaningful ways to assist plant breeding efforts.

Faced with major global challenges of food security, climate change and the drive to safeguard the environment, industry is asking new questions on how we can work with the available germplasm to develop crops of the future. In this Hackathon we aim to address two challenges. We want to find out the extent and scale at which we can link existing public data to address these questions, identify opportunities to improve data gaps and increase the utility of these data.

Crop Yield & Quality Challenge

The Challenge: Climate change is creating unpredictable weather patterns such as the warmer and drier summer observed throughout Europe in summer 2018. As a result, there may be new pests, disease, and abiotic stress (e.g. drought) problems in major crops such as wheat. Breeders would like to understand which wheat variety, genes, SNPs, environmental conditions and geographical locations have a positive or negative impact on the production of quality wheat.

Crop Protection Challenge

The Challenge: Several fungicides and pesticides are at risk of losing efficacy due to arising pathogen resistance. Coupled with the fact that several commonly used pesticides are restricted or are being banned to mitigate their wider environmental impact. This presents challenges for new product development pipelines in Industry. A major objective is the identification of novel targets for broad-spectrum development of pesticides that can disrupt host-pathogen interactions.

Industry organisers & collaborators

Derek Scuffell (Syngenta); Erick Antezana (Bayer CropScience); Fei-ling Lim (Unilever); Sally Gibbs (Unilever); Jan Taubert (KWS)

Anticipated workshop outcomes

This will be a proof of concept hackathon, to demonstrate the extent and scale at which we are able to connect agriculture and food datasets; using publicly available data from EnsemblGenomes (including plants, fungi, bacteria, metazoa and protists); AHBD recommended list variety trials data (incorporating agronomy, yield and quality trait data); PHI-base (incorporating some FRAC data); AgriMetrics Field Explorer data (UK field location; weather, soil and cropping); CerealsDB (varietal marker data); Expression Atlas for plants (for tissue and/or developmental specific gene expression data) and other data resources as needed.

Desirable overall outcomes can be summarised as:

- An understanding of the efforts required to link relevant datasets and gaps in interoperability.
- Delivery of data frameworks that can be used across industries
- Building a community of interest to continue data links
- Discussion on precompetitive activity to sustain activity

Participants...

will include a wide range of **data scientists;** domain experts from organizations and institutes engaged in creating public data resources aimed at delivering data for the AgriFood industries; **plant breeders**, key opinion **leaders in farming and agriculture** as well as **plant scientists** and others with biological expertise to identify and translate the most appropriate data parameters for linking through to relevant molecular biology/omics data.

Confirmed data resource experts

- Cyril Pommier (INRA) - Computer scientist and a bioinformatician specialized in plant scientific information systems; deputy lead of team in charge of INRA plant national database

- Monika Solanki (AgriMetrics) - Principal Data Scientist, with a research focus on developing ontology design patterns, models and applications in the area of knowledge representation and reasoning using Semantic Web technologies and Linked data principles.

- Bastiaan Brak (AHDB) - Research Data Analyst, applying data analysis and visualisation to create decision support tools (e.g. Sclerotinia infection alerts,BYDV management tool, AHDB WeatherHub) for farmers.

- Martin Urban (Rothamsted) - Wheat pathogenomics/Data Scientist with a research focus on the discovery of strategies for controlling plant pathogenic microbes that cause considerable losses in crop yield and contaminate food products with toxins.

- Paul Wilkinson (Bristol University) - Bioinformatician; Cereals Genomics Group.

- Guy Naamati (EMBL-EBI) - Senior Bioinformatician on Ensembl Plants, an integrative platform for accessing and comparing plant data for over 50 species. His main focus

is on wheat as part of the Designing Future Wheat research program, working on developing the next generation of wheat germplasm.

- Simon Jupp (EMBL-EBI) - Bioinformatician in EMBL-EBI's Samples, Phenotypes and Ontologies team with a special interest in the development of ontologies and linked open data for the life sciences.

Programme

The order and the timing of activities might change before the workshop. The start and end times will not be altered to allow you to make appropriate travel arrangements.

Day One: Wednesday 6th March 2019

Time	Activity	Who
09:00-10:00	**Arrival, registration and coffee**	
10:00-10:05	Welcome and introductions	Effie Mutasa-Gottgens
10:05-10:20	Competency Question Pitches • Crop Yield and Quality Challenge • Crop Protection Challenge	Industry Organisers
	Data Resources Signposts – flash talks	
10:20-10:25	EnsemblGenomes/plants/fungi	Guy Naamati (EMBL-EBI)
10:25-10:30	CerealsDB	Paul Wilkinson (UoBristol)
10:30-10:35	PHI-base	Martin Urban (Rothamsted)
10:35-10:40	AHDB RL - Wheat variety trial data	Bastiaan Brak (AHDB)
10:40-10:45	Field explorer	Monika Solanki (AgriMetrics)
10:45-10:50	Ontologies & Biosamples	Simon Jupp (EMBL-EBI)
10:50-10:55	BrAPI tools	Cyril Pommier (INRA)
	Hack Team Formation & Hack	
10:50-11:00	**Coffee break**	
11:00-11:15	Top Tips for Hackers	Matt Shearer & Francis Rowland
11:15-11:30	Cloud compute resource, useful links & documentation	Effie Mutasa-Gottgens
11:15-	Decide on CQ to hack; create teams including relevant agri-domain knowledge expert. Plan; prep (data match making) and begin hack… **Refreshments will be supplied throughout**	All, with help from data resource experts…

13:00-14:00	**LUNCH BREAK**	Hinxton Hall Restaurant
14:00-18:00	Continue with hack	
18:00-18:30	Reflect on the day's activities and forward plan for Day 2	Effie/organisers
18:30-19:00	**Transport to Tickell Arms Duxford**	All
19:00-21:30	**Drinks Reception & Dinner** (followed by taxi transfers to accommodation)	All

Day Two: Thursday 7thth March 2019

Time	Activity	Who
08:30-09:00	Registration & morning refreshments	
09:00-09:15	Plan for the day	Effie/Organisers
09:15-15:00	Hacking in teams …/continues	
10:00-	**Tea/coffee refreshments available**	
12:30-13:30	**WORKING LUNCH**	
	Hack team presentations (refreshments will be available)	
15:00-15:10	**Team 1** Solution(s) applied to answer Q(s) + list of identified opportunities for improvements	Spokesperson
15:10-15:20	**Team 2:** as above	Spokesperson
15:20-15:30	**Team 3:** as above	Spokesperson
15:30-15:40	**Team 4:** as above	Spokesperson
15:40-15:50	**Team 5:** as above	Spokesperson
15:50-16:00	**Tea break while organisers data resource custodians and end-users review team outputs**	TBD
16:00-16:10	**Feedback to teams**	Industry organisers
16:10–16:30	**Final discussion** • Key take-home messages • Next steps	All/Moderated
16:30	**End of workshop**	