[Test] FindingPheno Newsletter Aug 2021

1 message

View this email in your browser

Newsletter 1 | August 2021

I've been trying to write this blog post for a while, something to introduce all the different bits that go into our project. But, while I understand the project very well, the framing of how to explain that in the blog was eluding me. Then I saw a tweet from another cool project (PROMICON) with a diagram from their grant proposal which triggered the idea of how to get started - and this is the result.
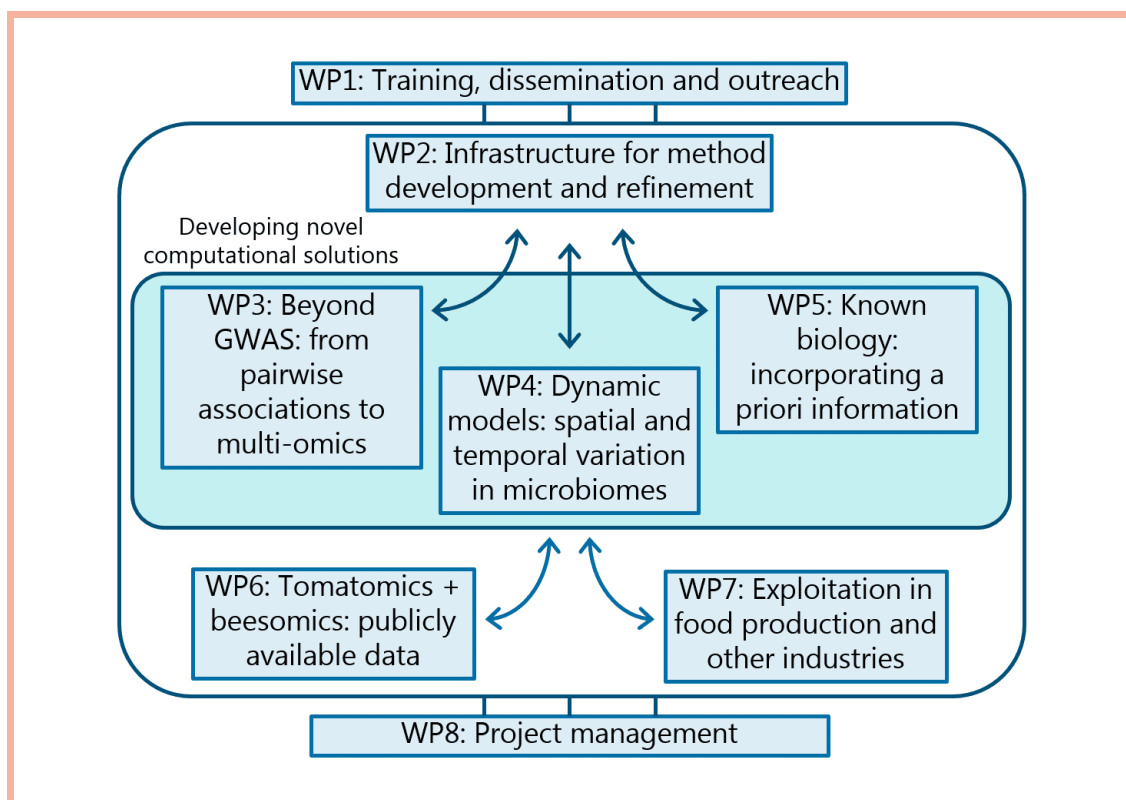
How is FindingPheno put together?

As with all publicly funded projects, the activities in FindingPheno are divided into different Work Packages (WPs) to give our overall project structure, as shown in the diagram above. While this is primarily a tool to tell us what we should do and achieve during the project, it is also a nice way to see the interplay between the ideas and technology that actually make up FindingPheno.

Developing new computational methods

What strikes me with this diagram is the parallel but interacting tracks in the middle of the diagram where we are building new computational methods. We have brought together researchers who, yes, are all data scientists in one way or another, but who also all have very different scientific backgrounds underlying the data science. This lets each group come from different angles to attack the fiendishly complex data integration problems I have talked about previously. Our project structure streamlines these angles into three main approaches for the best chance of success.

The first approach, WP3, focuses on developing multi-way association models within each set of different omic data types, with the end goal of finding directional pathways (a causes b) rather than just associations (a and b are somehow related). This WP builds on a method developed at the Champalimaud Foundation for understanding fish behaviour where they tracked and recorded the movement patterns of many fish within a tank, then used unsupervised machine learning to identify, track, and characterise each individual fish within that data. Being unsupervised means that no biological information or assumptions are included and the ML algorithm instead focuses on finding structure and networks within the data as it really is. This removes biases from what we think we should be seeing and opens the possibility of finding something new or unexpected. On top of this we are adding structural causal models currently being developed at Copenhagen University to give the directionality to these networks.

Meanwhile, WP4 focuses on the dynamic nature of the living system where the omic and meta-omic data changes over time or across different parts of the organism. By modelling these changes we can focus in on the associations and pathways that remain important even across different conditions, thus improving the robustness of our predictions. To do this we are combining the expertise at The Centre for Ecological Research for modelling microbial community development and evolution with the experience at University of Turku with modelling the spatial patterns in the human microbiome. This is a new collaboration, bringing together ideas from ecology such as game theory and mutualistic interactions with traditional data science methods such as machine learning/AI and probabilistic statistics then applying them to food production species here in FindingPheno.

Lastly, WP5 takes an opposite approach to WP3, instead starting from the existing biological knowledge to focus in on the things which are most likely to be meaningful. This knowledge includes annotations and information from publications, functional databases such as GEO or KEGG, and evolutionary knowledge from UCPH. Led by the University of Copenhagen, we are developing a hierarchical

Bayesian framework to integrate this information with the omics data to create an inference model for predicting phenotype outcomes. In the long run, we aim to balance the biology-agnostic causal modelling from WP3 against the dynamic modelling in WP4 and the biology-driven prediction models from WP5 to give the best overlap between novelty and certainty in our results.

Other project strengths

Along with these three key approaches, the project diagram also shows the two other strengths of FindingPheno. First is WP2 – the testbeds and data handling infrastructure that is being developed to underpin all the development work. We are lucky to have the European Bioinformatics Institute from EMBL leading this WP as they are world leaders in data management, processing and storage for multi-omics data sets. This WP not only provides a unified and well documented collection of hologenomic data sets for the other partners to build on, keeping everyone on the same page from the start, but also a shared computational infrastructure based on Embassy Cloud and Amazon Web Services.

The other strength is the real world validation in WP5&6, both scheduled for the later part of the project. The initial tool development will use fully integrated hologenomic data sets for salmon, chicken and maize, i.e. DNA, mRNA, protein, metabolites and microbiome collected at the same time from the same plant or animal within a controlled experiment. However, these integrated sets are still relatively rare and generally data comes from different experiments and does not always match. Therefore, we will adapt our new tools to work with data assembled from different sources, with the plan to focus on bees and tomatoes as they are relatively well studied and important species that do not yet have integrated data sets available. Meanwhile, we will also work with our industry partners (Chr. Hansen, Qiagen and Njorth Bio) to try out these prototype tools in their companies to demonstrate the commercial potential of what we are creating.

So this simple little diagram actually works all together to make a really cool project where we get to try a lot of different things and make some interesting new connections between ideas, while also having a good chance of developing methods that work and can do something meaningful in the real world.

Read and comment on the full post here.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952914.