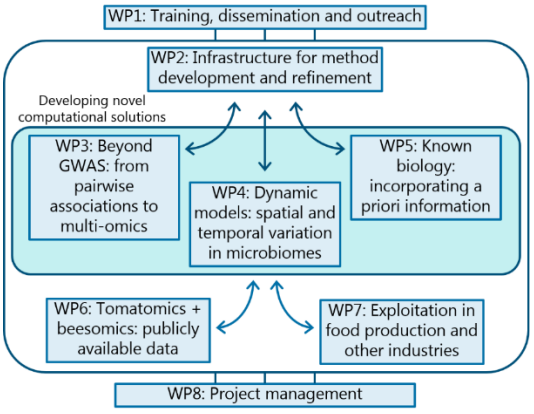
August 19, 2021

How is FindingPheno put together?

Updated: Jan 3, 2024

Activities in **FindingPheno** are allocated to various Work Packages (WPs) as seen in the diagram below. This 'tool' shows us what we should do and achieve during the project's lifetime but also offers an interplay between ideas and technology making up **FindingPheno**.

Developing new computational methods



What strikes me with this diagram is the parallel but interacting WPs in the middle where we are building new computational methods. We have brought together researchers, more specifically data scientists from diverse scientific backgrounds with data science. This allows each group to use different angles to attack the fiendishly complex data integration problems spoken about in previous posts. 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 associations (*a and b are somehow related*). This WP builds on a method developed at the Champalimaud Foundation used to understand fish behaviour where they track and record the movement patterns of many fish in a tank, then use unsupervised machine learning to identify, track, and characterize each fish within that data. Being unsupervised means that no biological information or assumptions are included and the machine learning algorithm focuses on finding structure and networks within the data as it is. This removes biases from what we think we should be seeing and opens the possibility of finding something new or unexpected. In addition, we are adding structural causal models currently developed at the University of Copenhagen to give directionality to these networks.

The **second approach**: WP4 focuses on the dynamic nature of living systems where the omic and meta-omic data changes over time or across different parts of the organism. By modelling

these changes we can focus on associations and pathways vital across different conditions, and improve the robustness of our predictions. To so do, we combine expertise from The Centre for Ecological Research in modelling microbial community development and evolution with the University of Turku's expertise in modelling the spatial patterns in the human microbiome. This new collaboration brings together ideas from ecology like game theory and mutualistic interactions with traditional data science methods like machine learning/AI and probabilistic statistics and applies them to food production species in **FindingPheno**.

The **third approach**: WP5 takes the opposite approach to WP3. It starts with existing biological knowledge to focus on the things most likely to be meaningful. This knowledge includes annotations and information from publications, functional databases like GO or KEGG, and evolutionary knowledge from UCPH. WP 5 led by the University of Copenhagen, develops a hierarchical Bayesian framework integrating this information with omics data to create an inference model to predict phenotype outcomes. We aim to balance biology-agnostic causal modelling (WP3) against dynamic modelling (WP4) and biology-driven prediction models (WP5) to give the best overlap between the novelty and certainty of our results.

Other project strengths

Two other strengths of FindingPheno include WP2 – the testbeds and data handling infrastructure underpinning the development work described above. The European Bioinformatics Institute leads this WP as a world-class leader in data management, processing and storage of multi-omics data. This WP provides a unified and well-documented collection of hologenomic data for partners to build on, keeps everyone on the same page, and offers a shared computational infrastructure based on Embassy Cloud and Amazon Web Services.

The other strength is the real-world validation in WPs 5 and 6, both scheduled for the later part of the project. The initial tool development will use fully integrated hologenomic data from salmon, chicken and maize, *i.e.* DNA, mRNA, protein, metabolites and microbiome collected at the same time from the same plant or animal in a controlled experiment. However, these integrated data are still relatively rare and generally come from different experiments and do not always match. 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. We will work with industry partners (Chr. Hansen, Qiagen and Njorth Bio) to test these prototypes in their companies to demonstrate the commercial potential of our results.

This 'simple' diagram works together to make a cool project where we get to try many diverse things and make some interesting new connections beween ideas, while having a chance to develop methods that work and can do something meaningful in the real world.

Written: Shelley Edmunds
Updated: Marie Sorivelle