Project Number: 952914

Project Acronym: FindingPheno

Project Title: Unified computational solutions to disentangle biological interactions in multi-omics data

D6.1 Summary on challenges in curating public data to apply FindingPheno's solutions

WP6 TOMATOMICS + BEESOMICS: APPLICATION TO PUBLICLY AVAILABLE DATA

Due date of deliverable:	30/11/2022
Submission date:	23/02/2023
Authors:	Alejandra ESCOBAR, Lorna RICHARDSON, Robert D. FINN
Dissemination level:	Public

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

DOCUMENT HISTORY

Version	Date	Description	Changes
0.1	25/01/2023	First draft	
0.2	27/01/2023	Second draft	Report formatted and internally reviewed
0.3	23/02/2023	Final draft	Report internally reviewed

Table of Contents

1 Introduction

As part of the FindingPheno mission, we aimed to identify suitable public datasets for use in the construction of models and their subsequent evaluation for elucidating genotype-phenotype associations, through the integration of multi-omics datasets. To address this, we used the European Nucleotide Archive (ENA) as a starting point for the identification of such datasets. The premise for starting with nucleotide datasets, especially the (meta-) genomic DNA, is that this data-type provides the foundational omics data upon which other datasets can be layered. Moreover, studies that contain metagenomic (metaG) and metatranscriptomic (metaT) data from the **same sample** are particularly valuable as they facilitate the best multi-omic integration, through the prediction of both coding potential (via the assembly and analysis of metaG) as well as a measure of actual coding levels (via the analysis of metaT, which may involve an assembly step, but always involves mapping the raw reads to the assembled metaG and/or metaT data).

2 Challenges

One of the important challenges that FindingPheno aims to address is the use of publicly available ‘omics datasets to discern the genotype-phenotype associations. The project focuses on genomics, metagenomics, and transcriptomics data from two important species, namely, tomato and honey bees. In light of the challenges in data access from the original data sources, we have included barley as another species of interest. In collecting and collating publicly available multi-omics data for tomatoes, bees and barley to which we would apply our FindingPheno solutions, we encountered the following challenges:

2.1 Missing, incorrect and inconsistent sample and experimental metadata

An initial list of studies was generated from a broad query against the sample metadata in the ENA to identify datasets of interest. The primary targets were metaG and metaT datasets associated with honeybee (13 and 11 studies respectively), tomato (12 and 5)-, barley (1 and 1)-, and soybean-rhizosphere (7, 0) (i.e. the host organisms of interest). We typically query the ENA API for all studies containing raw sequencing reads, not only those specifically labelled as derived from a "metagenomic" source, as prior experience demonstrated that inconsistencies in the data definition can result in the omission of important datasets. This broad approach also facilitates the capture of datasets where focus was on the host, rather than the microbiome, but where the microbiome is also captured. However, this approach requires additional manual curation and filtering.

For the FindingPheno search, to automatically identify the subset of results tagged as metaG and metatT sequencing data, we used the Library_strategy and Library_source fields. Biomes of interest were selected using relevant keywords (like tomato, *Solanum lycopersicum*, honeybee, *Apis mellifera*, etc) in the 'Study_description' and 'Host' fields. This resulted in a list of datasets potentially tractable for use in FindingPheno. Subsequently, the raw-reads were fetched and processed for assembly, functional annotation, and in the case of metaG, the generation of metagenome-assembled genomes (MAGs) where there was sufficient sequence depth to achieve a good level of assembly and the sample preparation was not selected for viruses.

Despite this approach, there remained cases where there was a mismatch between the metadata-based description and the submitted file content. A subset of specific examples are shown in **Table 1**. The most frequent mislabeling observed was 'RNA-Seq' in the Library_strategy field of a dataset that was in fact amplicon/barcoding data. While the amplification may correspond to the gene encoding the ribosomal small subunit RNA (SSU, RNA) marker genes (i.e. 16S or 18S), it is the gene that is amplified, not the rRNA. In addition, an array of other mismatches were also identified. In the case of metaG sequence data, the expected combination of metadata fields would be Library_strategy=WGS, Library_source=metagenomics, whereas for metaT the expected combination would be Library_strategy=RNA-Seq, Library_source=metatranscriptomics. As can be seen from the examples shown in **Table 1**, there were a variety of combinations identified. In such cases, it is necessary to use the associated publication (where one existed and was linked in the study object) to determine the correct data type. In cases where a dataset is not associated with a publication, it is only when analysis is started that it becomes apparent that the dataset has been incorrectly labelled (which can result in a waste of computational resources and person time). While comprehensive guidelines do exist for the submission of data (and its descriptive metadata to the ENA), it is not possible to rely on data being correctly described in submission. Reading the free-text description associated with the study is another way to discern the likely sample source and experimental methods, which is time consuming and unstructured making it difficult to access programmatically.

Table 1: Examples of inaccurate metadata related to sequencing data records in the ENA

Biome	Study acc examples	Lib strategy	Lib source	Expected sample type	Curated sample type
Honeybee **Tomato** **Barley**	ERP104960 SRP252362 SRP252362	RNA-Seq	metagenomics	metaG or metaT	amplicon
Honeybee	ERP131595	WMS	metagenomics	metaG	amplicon
Honeybee	SRP043685	RNA-Seq	metagenomics	metaG or metaT	metaT
Honeybee	SRP003774	WGS	genomic	metaG	metaT

For purposes of compiling the list of appropriate datasets for FindingPheno, datasets were excluded from the list where there appeared to be missing or mis-assigned metadata and no publication existed, thus diminishing potential statistical power that comes from large inputs. It is also worth noting that while the ENA provides the facility to link metaG and metaT datasets pertaining to the same sample, by assigning both sequencing datasets to the same BioSamples record, this is not routinely applied by users when submitting. Thus, it can be difficult to identify multi-omic datasets in this way without the manual overhead of consulting the original publication. Mismatched or mis-specified BioSamples records for matched metaG and metaT samples can result in the removal of these samples from FindingPheno datasets, again resulting in a reduction in statistical power.

2.2 Connecting processed nucleotide datasets to raw-reads

With a view to minimizing unnecessary compute, and fully representing existing publicly available data, prior to assembly and MAG-generation we sought to determine if the raw sequence data submitter pre-generated these derived sequence products and associated them with the raw sequence data. As metagenomic assembly and MAG generation become increasingly common, this has not been paralleled by the appropriate submission of these derived data products, as journals and funders only stipulate the need for submission of the raw sequence data. For resources such as MGnify, this represents a missed opportunity and limits downstream multi-omics integration until the assembly is regenerated. For the FindingPheno project we mitigated against this through a combination of querying the ENA portal API and/or identifying the data via the corresponding publication. In cases where MAGs were publicly available but not submitted to INSDC (such as available within an FTP location of figshare resource), we contacted the authors to encourage them to upload their MAGs to the ENA. This not only allowed us to include them in the MAG catalogues, but also demonstrated best practice of data stewardship and thus greatly increased capacity for broader re-use. FindingPheno is well placed to provide direction to the research community on the merit of well-described and correctly archived data, made available in freely accessible public repositories.

2.3 Submitting derived sequence data to the ENA

Having identified and filtered the datasets for FindingPheno, these were then systematically analyzed by the MGnify pipelines. The standard process undertaken for public studies was: fetch raw-reads; assemble raw-reads; submit assemblies to the ENA; fetch the assemblies for analysis; (in the case of metaG derived data) bin the assemblies; generate MAGs from the bins; submit the MAGs to the ENA. When the derived data products are submitted to the ENA, these are submitted as third party analyses associated with the original raw sequences and sample data. One unusual challenge encountered was an incompatibility between NCBI and the ENA regarding the rules for labelling ‘organism name’ in metagenomic data. This invalidated the submission of assemblies generated from the study *SRP199631*, blocking the usual workflow of the sample processing. The study (*SRP199631*) is a tomato virome, with the most abundant expected viral family in the samples "Geminiviridae" used as the taxonomic label in the original submission, preventing us from submitting the derived assembly as metatranscriptomic.

2.4 Utility of the data for FindingPheno approaches

Another challenge to utilizing public data for FindingPheno is the variability in sequencing depth. Examples of studies that could not be used for MAG generation for this reason included the following honeybee studies - SRP298034, SRP116602, SRP271039; and tomato studies - SRP017205, SRP167946, SRP338795. In cases where there were multiple related samples with insufficient data from a single sequence dataset to perform assembly and MAG generation, we employed a co-assembly approach to overcome this challenge. However, the limited number of datasets available for soybean and barley biomes made it impossible for co-assembly to overcome low-sequencing depth issues in those cases.

2.5 Linking other data types to nucleotide sequences

We addressed issues faced with obtaining and connecting nucleotide datasets. However, the methods established as part of FindingPheno are not limited to nucleotide data, with both metabolomics and (meta-) proteomics data of relevance. The latter types are submitted to public data archives that operate independently to the ENA. One solution that is increasingly being recognized as a way of connecting different datasets across different archives is the accessioning of samples, using the BioSample resource. Once a sample has an accession, regardless of the subsequent experiment, be it DNA sequencing, metabolomics or even imaging, the resulting data can be connected via the BioSample accession if it is included in the metadata of the submission. However, not all data archives require the use of BioSamples accessions. The MetaboLights data resource for depositing metabolomic data has this as a structured field, but it is not mandatory, while the PRIDE resource (for archiving proteomics data) does not have a specific BioSample field. This depends on the submitter's understanding and utilizing the process, but, as these different 'omics techniques are often performed by different groups and different facilities, using different internal sample accessioning systems, this tends to mean that the sample accessions rapidly diverge. For all public datasets identified in the ENA as part of FindingPheno, no other 'omics data type has been associated (see below), even after referring to associated publications. Notably, another dataset highlighted in the proposal was a study of tomatoes that encompasses genomes, transcriptomes, and metabolomics, however only the sequence data has been submitted[1]. Even when multi-omic datasets are linked through BioSamples records there can be limitations to fully integrating the data. By way of example, metabolomics data can suffer from poor spectral interpretation resulting in a small fraction of metabolites being identified and so available to be overlaid on metagenomics datasets. This sparsity of data can limit the interpretation of data. In addition, there is a lack of user-friendly tools to permit this overlay of metabolomic data onto metagenomic data.

2.6 Access to phenotypes

Currently no public resource exists that routinely accepts phenotypic data from a broad range of samples/species. Major resources like dbGaP and gene2phenotype gather genetic variation and phenotypic data, which is associated with human data. There are more specific datasets focusing on host, microbiomes and phenotypes (e.g. Phenotype Database contains information on human, diet and linked phenotypes), but these are relatively narrow in scope and numbers of dataset. A challenge in capturing phenotypic information for use in FindingPheno's computational tools, is structuring of the phenotypic information. While ontologies are being developed to enable structuring of such data, standards need to be developed to allow the integration of controlled vocabularies with metrics, including how those metrics were obtained (experimentally or derived mathematically).

2.7 Ability to share public datasets

As partners in the HoloFood project, researchers at MGnify have a good understanding of the HoloFood data and how it can contribute to FindingPheno. The HoloFood data have been generated and submitted to public archives with comprehensive metadata that addresses most of the aforementioned challenges. The samples are crosslinked across resources through the creation of BioSamples records, which phenotypic and other data that are not typically submitted to archives, added to the BioSample record in a structured, machine-readable fashion. While the HoloFood data will represent the gold-standard for data integration and data description, much of it has not yet been made publicly available, and so has hindered plans for its use thus far within FindingPheno. Much of the delay in data release stem from effects of COVID-19 pandemic, resulting in delays in the HoloFood wet-laboratory work.

[1] https://doi.org/10.1016/j.cell.2017.12.019

3 Potential solutions to mitigate challenges

While it is impossible to address all the challenges, we have listed some potential solutions that would reduce the impact of these issues while allowing future multi-omics datasets to be more readily utilized. Some solutions represent developments that are beyond the scope and jurisdiction of the FindingPheno consortium. However, where we have provided or could offer a solution, it is recorded as an interim measure.

1. While the ENA submission interface may seem daunting to new users, fundamentally it facilitates the appropriate submission of all data types, its associated metadata and connectivity of different sequence types. However, the lack of uptake suggests that **improved user interface design and training** would aid the organization and submission of multi-omic data.

    - ***Interim solution (provided by FindingPheno):*** *we will update the MGnify user-manual pages to provide an example of best practice for the submission of multi-omic datasets. We will gather feedback from the consortium on the user-manual to ensure clarity, updating where necessary.*

2. **Improving validation** within ENA submission tools would help identify when library strategies are mislabeled, as metabarcodes have bias in sequence composition (due to the sequencing of the barcode tag) compared to shotgun approaches. There are quality control tool in MGnify capable of detecting this phenomenon. Previous ENA submission processes did not include this element due to the computational overhead, but as more validation is pushed to the client this may become feasible.

    - ***Interim solution:*** *share MGnify quality control strategies with the ENA team to consider for future inclusion in ENA validation tools. At the same time, MGnify quality control measures will mitigate against errors in FindingPheno.*

3. ENA (and all INSDC members) allow only the original data provider/account owner to modify records. MGnify researchers tried mitigating this by informing the ENA of errors observed with the FindingPheno data. ENA is in the process of obtaining permission to change theses record from data provider, which is time consuming, often times with no response from the data provider (as they may have moved to a new position and the email address is no longer valid). Allowing authoritative sources, like MGnify, to **submit evidence based corrections to metadata** may remedy the situation.

    - ***Interim solution:*** *for datasets with identified errors in the sample and/or experimental metadata, and been unable to have the ENA record updated, we will submit proposed changes to the Contextual Data ClearingHouse (CDCH) to enable a discoverable record.*

4. Increasing awareness of the need to archive derived sequences, often requires extensive computational resources to produce among journals and funding agencies. While some datasets of derived sequences are available via *Figshare* or similar resources, it lacks structural organization and discoverability required by the scientific community. The solution proposed here can be addressed using the same mechanism as solution 1 - **improved user interface design and training**.

5. MGnify team's inability to submit data back to the ENA due to inconsistent taxonomic labels represents a specific case of inconsistencies between two INSDC member databases. The team is **working directly with ENA helpdesk** to address the issue and facilitate submission of the derived sequence datasets.

6. Discovering linked datasets across multiple archival resources can be arduous and time consuming, especially for less familiar resources. While better crosslinking between the different archives is desirable, this is beyond the control of the MGnify team. **Increasing the uptake of BioSamples** will aid in identifying

data pertaining to the same sample across different resources.

- ***Interim solution:*** *During FindingPheno we are working on a solution to aid the gathering of sample-related datasets via the development of a publicly available Jupyter notebook simplifying methods required to gather and simply overlay results in R. This can be built upon by project partners and others to connect to more elaborate data analysis methods.*

7. New modalities of sharing data between scientists are required. Currently, INSDC operates a single user per study policy, making it difficult to control data access in a secure fashion. However, with increasing complexity of experiments, often conducted as part of large consortia, this policy does not work for many areas of sciences. The HoloFood data could have been made available to FindingPheno in a pre-publication fashion with a more elaborate data access control model. However, this functionality is costly to both implement and maintain by the INSDC archives.

4 Conclusion

The capture of phenotypic traits remains a key, unaddressed issue. This is well beyond the scope of the FindingPheno project, and may represent one of the critical blockers in the widespread adoption and application of the FindingPheno approaches. In any relevant publication, we will highlight the critical need of this data type to be captured and the fact that the HoloFood project managed to overcome this limitation by extension of the BioSample record. We will also raise this issue through our contact with the Genome Standards Consortium, to understand whether standards are being developed for the capture of phenotypic data.