Project Number: 952914

Project Acronym: FindingPheno

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

D5.1 Knowledge resources for chicken, salmon and maize

WP5 DYNAMIC MODELS: KNOWN BIOLOGY: INCORPORATING *A PRIORI* INFORMATION

Due date of deliverable:	31 May 2022
Submission date:	31 May 2022
Author(s):	István SCHEURING, Shelley EDMUNDS, Robert FINN, Lorna RICHARDSON, Shyam GOPALAKRISHNAN
Dissemination level:	Confidential

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	Changes	Page
1.0	31/05/2022	Initial edition	All

Table of Content

1 Background

FindingPheno's major objective is to develop better computational solutions for the challenges posed by the vast amount of multi-omics data that is currently being produced. To solve these challenges, we rely on cutting edge technology and develop new statistical methods for working with this data. Work Package 5, and especially Tasks 5.1-3, develops a statistical inference framework that builds on what we already know about the genes, proteins and metabolic pathways active within both host and microbiome when analysing multi-omics data. The aim is to decrease multiple testing burden by removing data points unlikely to contribute to phenotype, improving our predictions of truly causal molecular interactions. In order to accomplish these tasks we rely on external public databases to provide evolutionary and biological knowledge that we can integrate into our models. It is important that the information contained within these datasets can be trusted and remains accessible both during and after our project duration, so we focus on those resources which are supported by high quality publications and which follow FAIR data sharing principles.

2 List of Resources

We have complied the following list (Table 1) of useful resources which fulfil the above criteria to use when developing our new analysis models in WP5 below. This list is also being developed into a page on our website as a public resource for other researchers in this area, see https://www.findingpheno.eu/knowledgeresources.

Table 1: List of knowledge resources for chicken, salmon and maize

Name	Acronym	Description	Location	Journal Reference
Kyoto Encyclopedia of Genes and Genomes	KEGG2	Database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.	A collection of different databases, table of contents here: https://www.genome.jp/kegg/kegg2.html	https://doi.org/10.1093/nar/gkaa970
KEGG Organisms	KEGG Organisms	Collections of genes and proteins in complete genomes of cellular organisms generated from publicly available resources, mostly from NCBI RefSeq and GenBank, and annotated by KEGG in the form of KO (KEGG Orthology) assignment.	Atlantic salmon: https://www.genome.jp/kegg-bin/show_organism?org=sasa Chickens: https://www.genome.jp/kegg-bin/show_organism?org=gga Maize: https://www.genome.jp/kegg-bin/show_organism?org=zma	
Gene Ontology Annotations	GO	The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.	Download the ontology: http://geneontology.org/docs/download-ontology/ Download latest GO annotation files: http://current.geneontology.org/products/pages/downloads.html Chickens: https://www.ebi.ac.uk/GOA/chicken_release	https://doi.org/10.1038/75556 https://doi.org/10.1093/nar/gkaa1113
Ensembl	Ensembl	Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data.	https://www.ensembl.org/index.html	https://doi.org/10.1093/nar/gkab1049
MGnify	MGnify	MGnify offers an automated pipeline for the analysis and archiving of microbiome data to help determine the taxonomic diversity and functional & metabolic potential of environmental samples. Users can submit their own data for analysis or freely browse all of the analysed public datasets held within the repository. In addition, users can request the assembly and analysis of any appropriate dataset within the European Nucleotide Archive (ENA).	https://www.ebi.ac.uk/metagenomics/	https://doi.org/10.1093/nar/gkz1035

The Genome Taxonomy Database	GTDB	The Genome Taxonomy Database provides a curated, phylogenetically consistent and rank-normalized database of microbial genomes. It is sourced on the NCBI Assembly database, and receives regular updates.	https://gtdb.ecogenomic.org/	https://doi.org/10.1093/nar/gkab776
Genomic Evolutionary Rate Profiling	GERP	GERP++ is a tool that uses maximum likelihood evolutionary rate estimation for position-specific scoring and, in contrast to previous bottom-up methods, a novel dynamic programming approach to subsequently define constrained elements	https://bio.tools/gerp	https://doi.org/10.1371/journal.pcbi.1001025
Online Mendelian Inheritance in Animals	OMIA	A catalogue/compendium of inherited disorders, other (single-locus) traits, and associated genes and variants in 346 animal species (other than human and mouse and rats and zebrafish, which have their own resources) co-authored by Professor Frank Nicholas and Associate Professor Imke Tammen of the University of Sydney, Australia, with help from many people over the years. OMIA information is stored in a database that contains textual information and references, as well as links to relevant PubMed and Gene records at the NCBI, and to OMIM and Ensembl.	https://omia.org/home/	https://doi.org/10.1093/nar/gkg074 https://doi.org/10.1111/age.13010
The NHGRI-EBI GWAS Catalog	GWAS Catalog	A high-quality curated collection of all published genome-wide association studies enabling investigations to identify causal variants, understand disease mechanisms, and establish targets for novel therapies	https://www.ebi.ac.uk/gwas/	https://doi.org/10.1093/nar/gky1120
Functional Annotation of ANimal Genomes project	FAANG	FAANG is the Functional Annotation of ANimal Genomes project. We are working to understand the genotype to phenotype link in domesticated animals.	https://www.faang.org/	https://doi.org/10.1186/s13059-015-0622-4 https://doi.org/10.1146/annurev-animal-020518-114913
InterPro	InterPro	InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.	https://www.ebi.ac.uk/interpro/	https://doi.org/10.1093/nar/gkaa977
Genome Properties	Genome Properties	Genome properties is an annotation system whereby functional attributes can be assigned to a genome, based on the presence of a defined set of protein signatures within that genome. Properties (which often describe pathways) are composed of steps, with each step defining a protein required for the function of the pathway/property. Genome properties use protein signatures as evidence to determine the presence of each step within a property.	https://www.ebi.ac.uk/interpro/genomeproperties/	https://doi.org/10.1093/nar/gky1013

Reactome	Reactome	Reactome is a free, open-source, curated and peer-reviewed pathway database.	https://reactome.org/	https://reactome.org/cite
Interactive Pathways Explorer	iPath	A web-based tool for the visualization, analysis and customization of various pathway maps	https://pathways.embl.de/	https://doi.org/10.1093/nar/gky299
MetaCyc metabolic pathway database	MetaCyc	MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life. MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes. The goal of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each experimentally elucidated pathway.	https://metacyc.org/	https://doi.org/10.1093/nar/gkz862
Molecular Signatures Database	MSigDB	A collection of annotated gene sets including genes grouped by their location in the human genome, canonical pathways and experimental signatures curated from publications, genes sharing cis-regulatory motifs up- or downstream of their coding sequences, clusters of genes co-expressed in microarray compendia, genes grouped according to gene ontology (GO) categories, signatures of oncogenic pathway activation, and a large collection of immunological conditions. All of the gene sets in MSigDB are manually reviewed, curated, and annotated.	https://www.gsea-msigdb.org/gsea/msigdb/index.jsp	https://doi.org/10.1093/bioinformatics/btr260 https://doi.org/10.1016/j.cels.2015.12.004
STRING	STRING	STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases.	https://string-db.org/	https://doi.org/10.1093/nar/gkaa1074
IntAct Molecular Interaction Database	IntAct	IntAct provides a free, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions.	https://www.ebi.ac.uk/intact/home	https://doi.org/10.1093/nar/gkt1115
AlphaFold Protein Structure Database	AlphaFold DB	AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. AlphaFold DB provides open access to 992,316 protein structure predictions for the human proteome and other key proteins of interest, to accelerate scientific research.	https://alphafold.ebi.ac.uk/	
Sorting Intolerant from Tolerant	SIFT	SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. SIFT can be applied to naturally occurring nonsynonymous polymorphisms and laboratory-induced missense mutations.	https://sift.bii.a-star.edu.sg/ (includes chicken)	https://doi.org/10.1101/gr.176601
Gallus GBrowse	Gallus GBrowse	Online access to genomic and other information about the chicken, Gallus gallus. Includes predicted genes and Gene Ontology (GO) terms, links to Gallus In Situ Hybridization Analysis (GEISHA), Unigene and Reactome, the genomic positions of chicken genetic markers, SNPs and microarray probes,	https://www.animalgenome.org/cgi-bin/gbrowse/gallus/ (the original seems to now be offline)	https://pubmed.ncbi.nlm.nih.gov/17933775/

		and mappings from turkey, condor and zebra finch DNA and EST sequences to the chicken genome. We also provide a BLAT server (http://birdbase.net/cgi-bin/webBlat) for matching user-provided sequences to the chicken genome.		
GalBase	GalBase	Galbase is a chicken multi-omics database that hosts reference genomes, annotations, high-quality genetic variants, transcriptomes, histone modifications, open chromatin regions, GWAS, and QTL. Galbase allows users to retrieve genomic variations in geographical maps, gene expression in heatmaps, and epigenomic signal in peak patterns, and also provides modules for batch annotation of genes, regions, and loci based on multi-layered omics data. Galbase integrated the UCSC Genome Browser, the WashU Epigenome Browser, BLAT, BLAST, and LiftOver, to facilitate search and visualize sequence features.	http://animal.nwsuaf.edu.cn/code/index.php/ChickenVar	https://doi.org/10.1186/s12864-022-08598-2
Chicken SNP database	ChickenSD	Chicken SNP Database (ChickenSD) is a data container for the variation information of chicken (Gallus gallus) genome. The aim of this database is to construct an SNPs detector and online visualization tool for the chicken research communities on population, evolution, phenotype and life habit studies. Currently, ChickenSD contains ~33 million whole genome non-redundant SNPs with well annotated information, which identified from 865 samples (167 wild, 697 domesticated and 1 hybrid).	https://ngdc.cncb.ac.cn/chickensd/	
Chicken Quantitative Trait Locus Database	Chicken QTLdb	Chicken QTL and association data curated from published data. The database is designed to facilitate the process for users to compare, confirm, and locate the most plausible location for genes responsible for quantitative traits important to chicken production. We have been striving our best to curate all available data, and adding tools to the QTLdb for users to accomplish many data meta-analysis and comparison tasks.	https://www.animalgenome.org/cgi-bin/QTLdb/GG/index	
SalmoBase	SalmoBase	Salmobase is a tool for making molecular genomic resources for salmonid species publicly available in a framework of visualizations and analytic tools.	https://salmobase.org/	https://doi.org/10.1186%2Fs12864-017-3877-1
Maize Genetics and Genomics Database	MaizeGDB	MaizeGDB is a community-oriented, long-term, federally funded informatics service to researchers focused on the crop plant and model organism Zea mays. It is a USDA/ARS funded project to integrate the data found in MaizeDB and ZmDB into a single schema, develop an effective interface to access this data, and develop additional tools to make data analysis easier.	https://www.maizegdb.org/	https://doi.org/10.1186/s12870-021-03173-5