OmicsDB::Pathogens - A database for exploring functional networks of plant pathogens

Bjoern Oest Hansen 1,2,3 Stefan Olsson 3,4

[1] *OmicsDriven, Toelloese, Denmark*

[2] *Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany*

[3] *State Key Laboratory for Ecological Pest Control of Fujian and Taiwan Crops, College of Plant Protection, Fujian Agriculture and Forestry University, Fuzhou 350002, China.*

[4] *Plant Immunity Center, Haixia Institute of Science and Technology, College of Life Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China.*

Abstract

Plant pathogens are a great threat to food security. To combat them we need an understanding of how they work. Integrating large-scale omics datasets such as genomes and transcriptomes has been shown to provide deeper insights into many aspects of molecular biology. For a better understanding of plant pathogens, we aim to construct a platform for accessing genomic and gene co-expression networks for a range of pathogens and reference species. Currently we have integrated genomic and transcriptomics data from 10 species (*Fusarium graminearum, Ustilago maydis, Blumeria graminis, Neurospora crassa, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Escherichia coli, Arabidopsis thaliana, Mus musculus and Homo sapiens*).

Here we introduce OmicsDB::Pathogens (http://pathogens.omicsdb.org), a publicly available web portal with an underlying database containing genomic, and transcriptomic data and analysis tools. It allows non-bioinformaticians to browse genomic data, and inspect and compare biological networks across species.

The information is modelled in a graph-based database, enabling flexibility for querying and future extensions. Tools such as BLAST and Cytoscape.js are available together with the option of performing GO enrichment analysis. The database also enables the user to browse information such as Orthologs, Protein domains and publications citing a given gene.

Herein we describe how to use this platform for generating hypotheses for the function of a gene.

Availability and Implementation

Currently, Omicsdb supports networks for 7 organisms and is freely available for public use at http://pathogens.omicsdb.org

Introduction

Plant pathogens are a great threat to the ever-growing population and the following need for food. To combat them, an understanding of their mode of action and of how the pathogenic processes work is essential. To be able to obtain useful knowledge from genomes, it is necessary to be able to correctly assign functions to gene products. Manual annotation is not a feasible option to annotate the ever-growing number of available genomes. Scientists, therefore, must rely more and more on predictions.

In-silico methods have gained popularity for automated annotation of gene functions, though for it to be useful, it is important that the accuracy of the predictions, and number of genes that can be annotated is high.

The guilt by association hypothesis is based on the observation that genes that participate in similar biological processes or are involved in similar regulatory pathways tend to display similar expression profiles (Wolfe et al., 2005). An increasingly popular method to utilize this observation is using gene co-expression networks. Networks are a convenient way of representing biological data and allow for the utilization of graph theory. By analysing the structure of the networks, it is possible to identify communities within the network, where the genes are more closely related to each other than to the rest of the network and are enriched for specific biological processes.

The two main applications for co-expression network analysis are 1) To use a bait-gene from a known pathway or with a known function, as a query, and identify possible genes within the same pathway or with a similar function (Itkin et al., 2013) and 2) to suggest the biological process or pathways given gene is involved in, based on the functions of its

neighborhood. With the increase in available biological networks, an increasingly important usage is the identification of common network patterns across species, this allows for more reliable transfer of knowledge from and across model species.
Inter-network comparisons serve multiple purposes, they can improve identification of functionally related orthologs across species (Hansen et al., 2014), and aid in the identification of conserved subgraphs within or across species (Stuart et al., 2003).

The amount of data a researcher working within molecular biology needs to handle has grown exponentially during later years. Not just does that put a heavy load on the IT infrastructure, but also on the researcher that might not be a bioinformatics specialist. Handling and processing omics data are often a requirement these days, to be able to perform analysis on this kind of data, researchers need to be familiar with scripting languages such as Python or R.

Curated databases have shown to be powerful tools for integration and analysis of data. This has led to an emergence of integrated web-based databases, changing the analysis of networks from being a task for specialist bioinformaticians, into a simple routine task for experimental molecular biologists, investigating specific genes or conditions. However, most of these databases focus on the major model species, such as Maize (Andorf et al., 2016), Budding yeast (Kim et al., 2014) or fission yeast (Vo et al., 2016). There exist very few options for researchers working on non-model species.
Plant pathogens have until now had little representation in these databases. Electronic resources for *Fusarium graminearum* eFG (Liu et al., 2013) a model species in plant pathology (Zhang et al., 2019), contains functional annotation as well as annotations of transcription factors and curated and predicted pathogenic genes for fusarium. PHI-base (Urban et al., 2017) is a great resource for curated Pathogen-Host Interactions but is manually curated and thus limited in size. PhytoPath (Pedro et al., 2016), is a resource for genomic information related to plant pathogens. Other databases with microbial data covergr mainly models species are also available (Kim et al., 2014, 2015; Oughtred et al., 2019)

To tackle this problem, we here present OmicsDB::Pathogens, an integrated database of networks from 7 species: 3 model species (E. coli, S. cerevisiae, S. pombe), and 4 plant pathogens (Fusarium graminearum, Ustilago maydis, Blumeria graminis, Neurospora crassa and Magnaporthe oryzae), as well as orthologs from 158 other fungal pathogens, to ease the knowledge transfer to and from these.

OmicsDB:: Pathogens is a database developed to assist experimental molecular biologists in accessing and analyzing omics data.
The database provides intuitive navigation, allowing the researchers to store, browse, analyse and compare their data, as well as handling metadata of both samples and workflows. Another important use of the database is that it allows comparisons between species that can allow the experimental researcher not just to infer possible functions of genes not previously studied experimentally in detail. It also makes it possible to rank genes of interest to study enabling more focussed research in sorting out gene function. We provide an extensive amount of already processed publicly available genomic and transcriptomic data.

Results and discussion

OmicsDB is a database that includes high throughput experimental data and information for multiple species. Since we store data from biological networks, it was decided to go for the graph-based database Neo4J for storing most of the data. Neo4J has previously been evaluated by Have and Jensen for several graph processing problems related to bioinformatics and compared with PostgreSQL, they found Neo4J to be faster in many cases, though they also conclude that graph-based databases are not necessarily the best choice for all problems (Have and Jensen, 2013). Neo4J uses the property graph model, which means that nodes and edges can have key/value properties associated. Neo4J uses its own query language, Cypher for querying the graph, cypher allows queries to be formulated in terms of paths, which allows them to be concise and intuitive compared with equivalent SQL queries, which is often complicated by joins, and difficult to read. Due to the simplicity of traversing edges, and accessing data through the Cypher language, we decided to use Neo4J for all data gene expression data, that is stored in an SQLite database, and flat files that are stored in an S3 compatible object storage, with their paths and metadata stored in Neo4J linked to the relevant organism. A simplified overview of our Neo4J data model can be seen in Fig. 1 An overview of the infrastructure and data processing steps can be seen in Fig. 2. The processed expression data is stored in a SQL database, with common identifiers shared between the other systems.

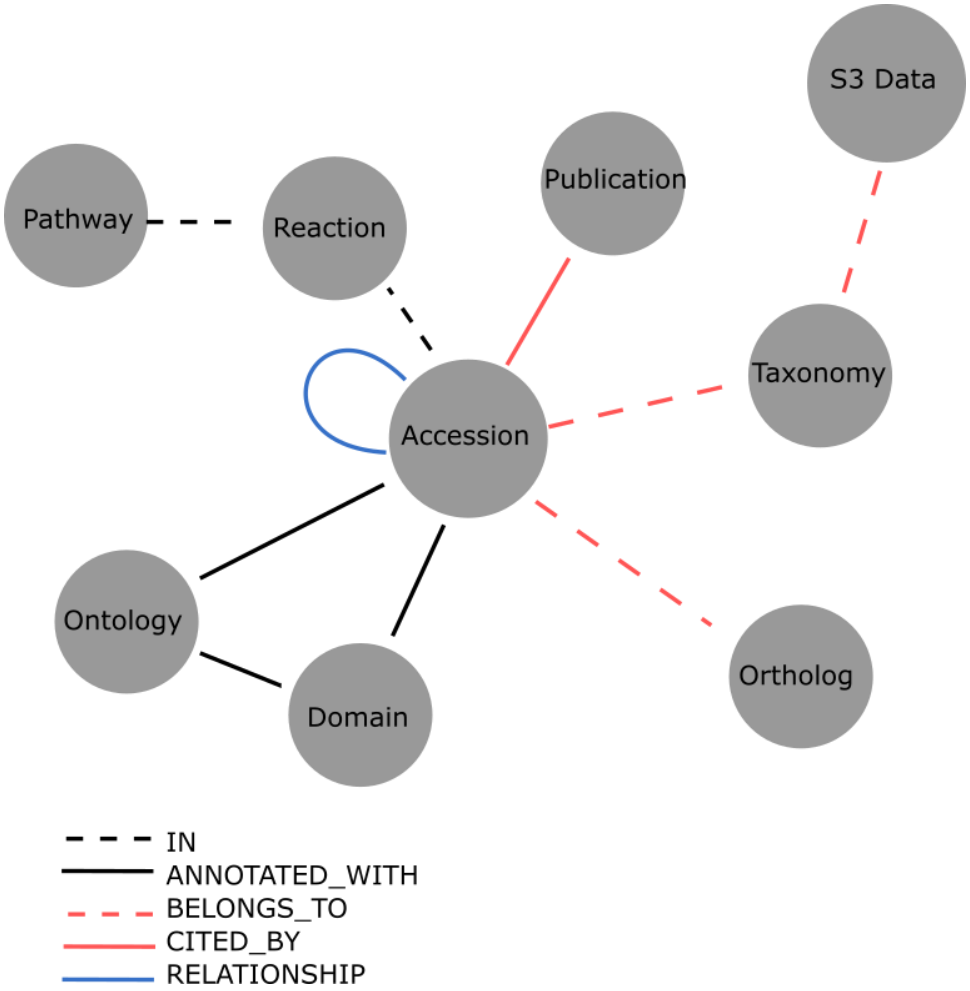

Fig 1. The part of our Neo4J data model. This shows how the different nodes are related. Properties are left out for simplicity. Gathering data across any node is easy, for example, if one wishes to know all publications related to a given Domain its possible to traverse the graph from Domain over Accession to Publication. We believe this gives a more intuitive way of accessing data than through a classical relational database.

FASTA and GFF files, are stored using an S3 compatible object storage (Palankar et al., 2008), together with their respective indexes. Accessing this information is easy since we can traverse the graph from a given Accession to the S3 data, and use the Accession together with the file index to drag out only what we need from the flat-file, while benefiting from the features of S3 buckets, especially scalability and accessibility across servers.

To ease access to our data, we developed a web-based interface for OmicsDB::Pathogens (http://pathogens.omicsdb.org). An overview of the website is shown in Fig 3. The site is centered around gene information pages, where each page tries to summarize the knowledge about a given gene, Users can search for the genes of interest using keywords, gene identifiers or by performing a BLAST search against their sequence of interest.

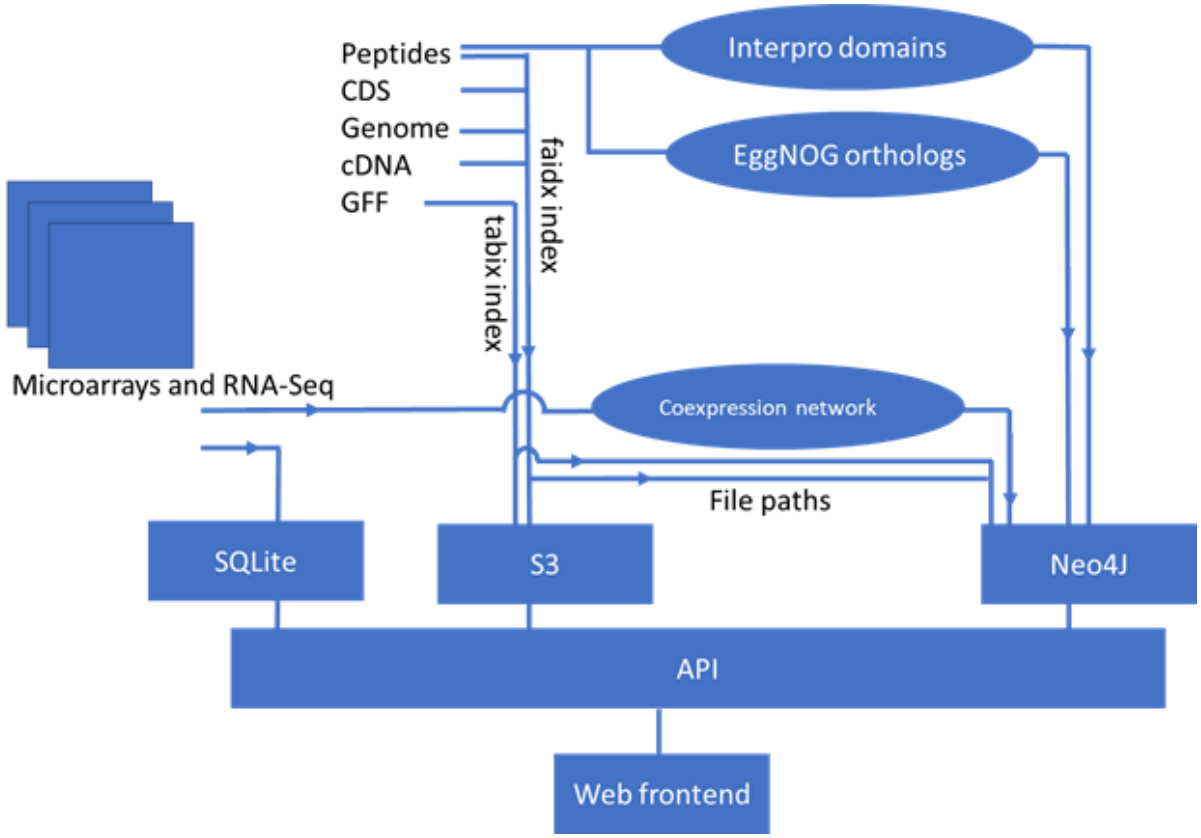

Figure 2. Storage and pipelines included in the data processing and storage. Microarray and RNA-seq data are stored processed and normalized (not shown). The TPMs are then stored in an SQLite database. The TPMs is also processed to generate a co-expression network, stored in Neo4J. Peptide, CDS, cDNA, Genomes and GFF files are stored in an S3 compatible object storage, together with their respective faidx and tabix indexes (Li et al., 2009; Li, 2011). Peptide sequences are processed by eggNOG (Huerta-Cepas et al., 2016) to identify ortholog families, and by Interpro scan to find conserved domains. Both orthologs and domains are stored in Neo4J, connected to their respective peptides.

Basic functions of OmicsDB::Pathogens

The main source of information is the gene page that offers important information regarding each gene. It provides different information for each gene. An overview can be seen in Fig 3.

The top of the page (Fig 3A) shows the basic information: The gene model with intron and exon information. Gene Ontology annotations (Ashburner et al., 2000). Ortholog groups from EggNOG (Huerta-Cepas et al., 2016) and protein domains from Interproscan (Hunter et al., 2009; Jones et al., 2014) and the coding DNA sequence and the protein sequence. Further down the page (Fig 3B) is a list of co-expressed genes and a visualization of the co-expression network. As well as A bar chart of gene expression displaying the average of replicates for a given experiment (Fig 3C).

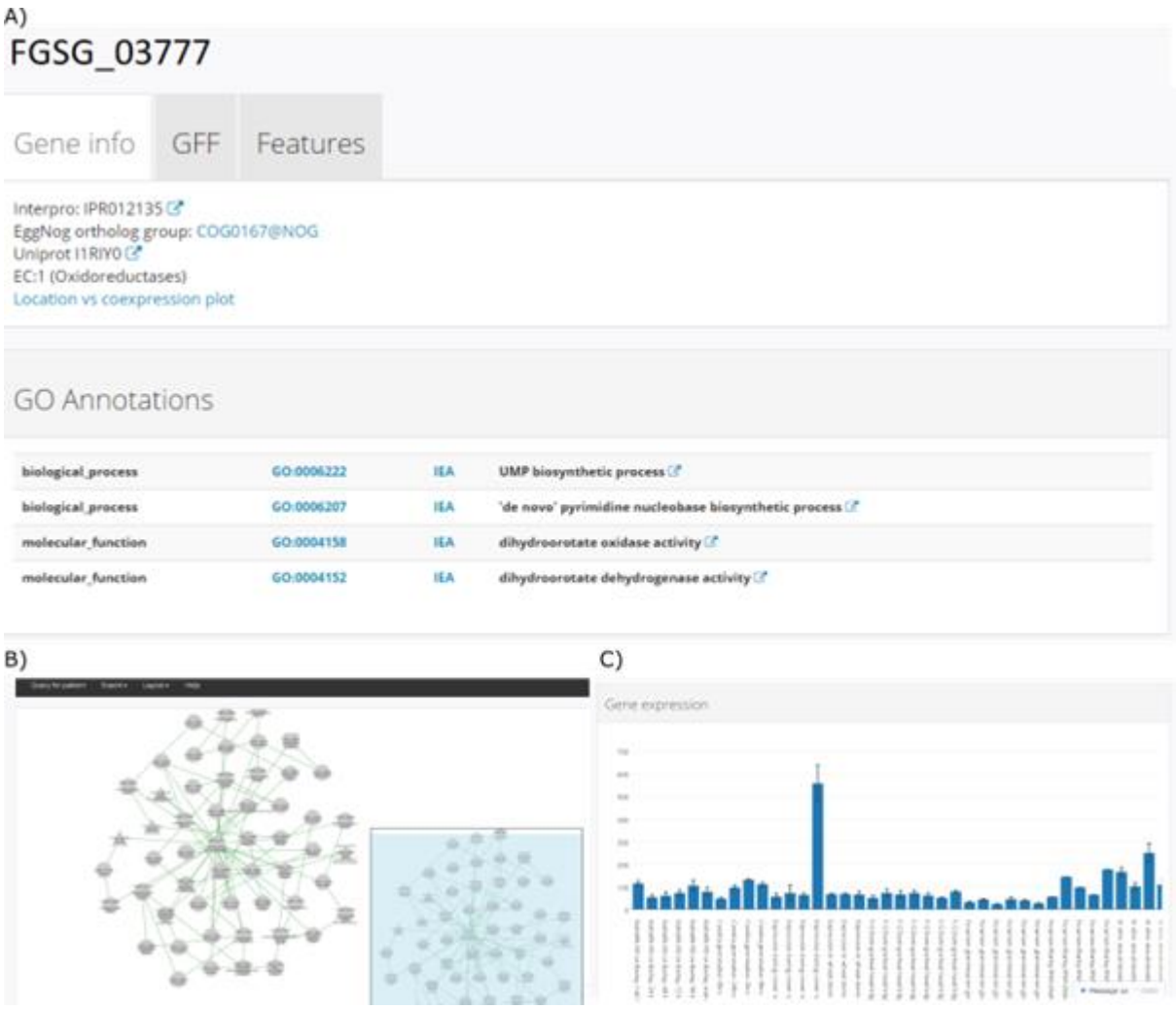

Fig 3. An example of the gene profile page. A) Shows the structural and functional information on the gene as well as ortholog information, possible pathways, and InterPro domains. B) Shows an example of a co-expression network. C) Shows the expression profile of a given gene. Both A, B and C are found on the same page in the browser.

This page also serves as a gateway to explore the information within and across species. For the species without networks, this data is limited to protein domains, ortholog families, DNA and protein sequences.

The network representation of gene interactions allows the users to query for the pattern of co-regulated orthologs across other species, to see if this regulated “module” is preserved. This function can ease the annotation transfer of “biological process” terms, relying not only on sequence similarity but also on the biological context.
To give a quick overview of what types of proteins the genes in the network code for, shapes of the nodes are used. The default shape is round, but it will be given a different shape in the case of transcription factors, or if a protein is classified with one of the 1-6 Enzyme

commission classifications (Cornish-Bowden, 2014). The node representing the bait gene is displayed in red. An overview of the shapes can be found in Supplementary Table 1. To represent different types of relationships, the edges are colored using the following scheme, Green edges: co-expression, red edges: physical interaction, blue edges: genetic interaction. The physical and genetic interactions are from The BioGrid database (Oughtred et al., 2019), the co-expression data is calculated using the Mutual Rank (MR) method (Obayashi et al., 2009) modified to 1/MR with a MR threshold of 50.

From the gene page is it possible to access other pages with information. Ortholog pages: Each EggNOG ortholog family has a dedicated page, with aggregated information on this family. Including genes and species in the family, functional annotations as well as papers citing the family. This enables users to easily check at a glance what information is available from other species. As well as how diverse the family is by visualizing the distribution of members per species in a pie chart. An example of an ortholog page can be seen in Fig 4.

Domain pages are similar to ortholog pages, but centered around a domain, for example, PF00331 or PS51760.

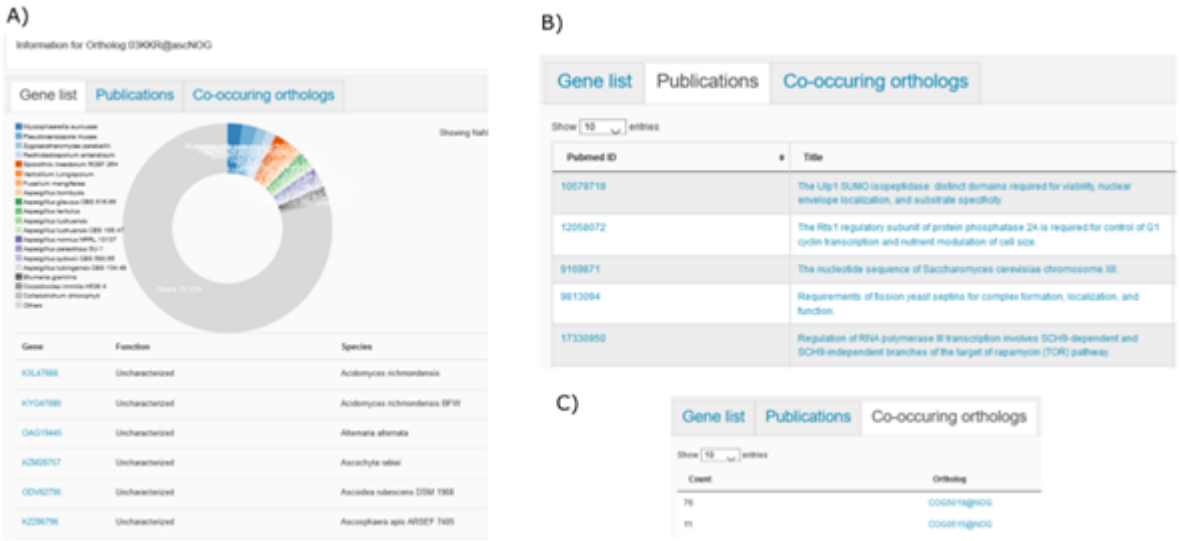

Fig 4. Each Ortholog has a unique page with multiple tabs. A) it shows all that has been assigned and the species they belong to. It is possible to subselect species using the pie chart. The other tabs include B) publications citing genes within this ortholog group. As well as C) orthologs that co-occur across the networks. Orthologs that often co-occur hints that they might participate in a similar process.

The Gene Ontology(Ashburner et al., 2000) has also been integrated, with pages for each term, displaying all genes annotated with a given ontology term, as well as the position of the

term in the GO graph. It is thus possible to find networks in one species that resembles those affiliated with a certain ontology in another species**.** Traditionally the user would have had to download and annotate both datasets. However here it can be done with the click of a button.

Advanced functions

OmicsDB also provides some tools enabling users to perform more advanced analyses. One example is cross-species network alignment. Cross-species comparisons of biological networks with interactions are still an emerging field, It allows for finding network patterns that between orthologous genes across species, hinting towards similar function. This has been used successfully in plants (Usadel et al., 2009; Movahedi et al., 2012; Hansen et al., 2014). Most work on gene function prediction using gene co-expression so far has been working on single or few species. Identifying conserved co-expression patterns across orthologs in different species can identify highly relevant candidate genes sharing similar functions or participate in the same pathway (Movahedi et al., 2012; Hansen et al., 2014).

An example of such a strategy with experimental confirmation is the study by Itkin et al., where comparative co-expression for tomato and potato was utilized, leading to the discovery of a gene cluster that is related to the steroidal glycoalkaloids pathway (Itkin et al., 2013).
It is also possible to use the surrounding network and look for enriched GO terms, using a hypergeometric test. This is done at the bulk analysis page. Where a list of genes can be analyzed. A plot of the expression values for all genes in the list will be generated as well as the GO enrichment. It is also possible to export both GO enrichments as well as peptide and cds fasta sequences.

Use case: Vacuolar iron uptake

Yeast Fet3/Ftr1 (YMR056W/YER145C) are genes for two proteins that are in yeast high-affinity iron uptake at the plasma membrane (Askwith et al., 1994, 3; Stearman et al., 1996). The proteins physically connect to each other. Ftr1 is a transporter and Fet3 is an oxidoreductase. A similar system Fet5/Fth1 (YFL071W/YBR207W) exists at the vacuolar membrane and is involved in regulating vacuolar iron storage (Urbanowski and Piper, 1999)

In *Fusarium graminearum* FGSG_02143 is known to be the plasma membrane Ftr1-like protein and the homologous protein FGSG_05160 is probably the Fth1-like protein located at

the vacuole membrane. To test if this can be the case FGSG_05160 was used as bait and the result is displayed together with the networks for yeast Ftr1 and Fth1. The aligned network can be seen in Fig. 5. Red edges represent possible orthologs and black edges co-expression within the species. The subcellular location for some genes products in Yeast has been marked.

What can be seen is that the vacuolar and the plasma-membrane iron transport is more co-regulated in the *F. graminearum* data than in the yeast data (direct strong co-regulation between FGSG_05160 and FGSG_02143). The pattern of co-regulation of the other proteins with orthologues in yeast also supports that FGSG_05160 is a vacuolar located Fth1-orthologue. This has, however, to be confirmed by direct experiments.

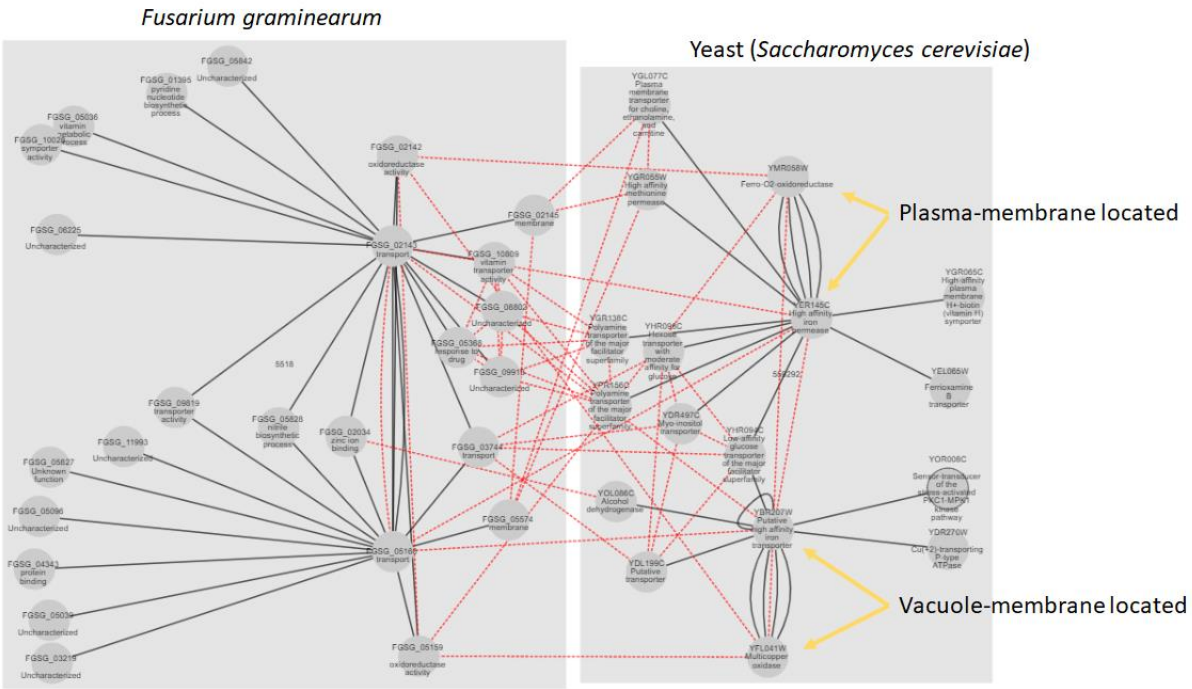

Fig 5. The alignment of the co-expression neighborhoods from Yeast and F. graminearum, Using FGSG_05160 as a bait, and searching for counterparts in Yeast with similar co-expression patterns. Black edges represent co-expression edges, red edges represent possible orthologs relationship. It can be seen that large parts of the network are mirrored across the organisms, hinting towards the genes carrying out the same functionality.

Conclusions and future directions

In this study, we developed a system for handling biological omics data. We processed gene expression data, as well as generated co-expression networks for 7 species.

We established an interactive web interface omicsdb.org to provide access to the data as well as the analysis platforms for the public. We plan to continue to improve the quality and functionality of this database, by regularly updating with new publicly available data.

At the moment the database only contains networks for **4** pathogens and 5 references, however, there is a growing need for understanding how plant pathogens work, to alleviate this need we will include new species as when they become available.

Materials and methods

External tools

BLAST+ (Camacho et al., 2009) has been integrated to enable the users to easily find their gene of interest, allowing the user to BLAST their bait sequence against our database. This is practical in case they have identifiers not present in our database, or the user working on species not yet included.

Construction of gene coexpression data

For the construction of co-expression database for plant pathogens, we selected 4 pathogenic species (Fusarium graminearum, Ustilago maydis, Blumeria graminis, Neurospora crassa), based on the availability of gene expression data, as well as 6 reference species (Escherichia coli, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Arabidopsis thaliana, Mus musculus, Homo sapiens).

Expression data were downloaded from Arrayexpress (Kolesnikov et al., 2015), and normalized using the following methods: For Agilent data, the R-package limma (Ritchie et al., 2015) were used . For Affymetrix, Affymetrix Powertools v. was used, running the RMA algorithm. For NimbleGen, DEVA V 1.02 (Roche) was used. RNA-Seq data were subjected to QC using FastQC (Andrews), reads were mapped using TopHat2 (Kim et al., 2013), counted using HTseq-count (Anders et al., 2015) and normalized using the VST algorithm implemented in DESeq2 (Love et al., 2014), run in R v. 3.4 (R Core Team, 2013).

Genome annotations were derived from Pombase for *S. pombe* (Wood et al., 2012), SGD for *S. cerevisiae* (Cherry et al., 1998), and Ensembl for the remaining species. Transcription factor annotations were downloaded from (Liu et al., 2013), Enzyme Commision numbers were obtained from Uniprot (The Uniprot Consortium, 2019). Biological, chemical and genetic interactions were derived from BioGrid (Oughtred et al., 2019)**.**

The Mutual Rank value of the weighted Pearson's correlation coefficient was used as the measure of coexpression, as described by (Obayashi et al., 2009). A threshold of 50 was applied. To make it comparable with protein interactions, the MR was normalized by taking 1/MR ensuring that 1 was equivalent to the best value. Mutual Rank is calculated as

$$MR(A,B) = \sqrt{(Rank(A \rightarrow B) \times Rank(B \rightarrow A))}$$

With A and B represent genes. GFF files were stored in tabix format (Li, 2011) peptide and DNA sequences were stored in faidx format implemented in Samtools (Li et al., 2009).

Functional Annotations

Protein domains including PFAM domains (Finn et al., 2014) and Panther (Mi et al., 2013) were identified using InterProScan v. 5.29-68.0 (Hunter et al., 2009; Jones et al., 2014),

Ortholog groups were identified using EggNOG (Huerta-Cepas et al., 2016), which also provided predicted GO terms. For the Escherichia coli, Schizosaccharomyces pombe and Saccharomyces cerevisiae, experimentally validated GO terms could be downloaded from the GeneOntology website (The Gene Ontology Resource: 20 years and still GOing strong, 2019). The Gene Ontology website also provides mappings from domains identified by InterproScan to their associated GO terms. These maps were used to further improve the annotations.

Papers citing a given gene or gene family were retrieved from Uniprot (The Uniprot Consortium, 2019)**.**

For reference, Arabidopsis thaliana, Mus musculus and Homo sapiens were included, annotations were derived similarly to the pathogens.

Network alignment

The network alignment uses the “@NOG” ortholog families, if you provide a bait gene, it will query all other genes from this family, and compare how many orthologs are shared in the 1st-degree neighborhood. To calculate how much of the network around two genes is similar, the Jaccard index for the gene families shared across the networks is then calculated following

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B represent the gene families in the network surrounding your bait gene and the gene it is compared against respectively. The numerator is the intersection of gene families, and denominator represents the union of all gene families in the 1st-degree neighbourhood.

Ontology Enrichment

Enrichment of Gene Ontology terms in sets of genes is calculated using the Hypergeometric distribution function, implemented in Scipy (Virtanen et al., 2020). For the calculation of background terms, following the true path rule, all ancestors of any given annotation are considered.

Database implementation

The system runs on Linux v. 18.04 LTS Bionic Beaver. The web-service was implemented in a Python-based web application framework, Flask v. 1.0.2, with SQLite and Neo4J 3.4 (www.neo4j.com) as the backend databases for the expression data and everything else respectively. The website is being served using gunicorn and NGINX. The co-expressed gene networks as well as the directed acyclic graphs for the GO terms were visualized using Cytoscape.js v. 3.14 (Franz et al., 2016), gene expression plots were generated using dc.js (https://dc-js.github.io/dc.js/). For the HTLM page layout, the bootstrap framework was used (www.getbootstrap.com), and for general purpose usability features the Jquery.js library (https://jquery.com) was used

BLAST+ (Camacho et al., 2009) was installed and a custom HTML interface was developed. Icons were obtained from Font Awesome v. 5 (www.fontawesome.com).

Conflict of Interest

No conflicts are declared.

Author Contributions

BOH and SO designed the study, BOH developed the database and processed the data. BOH and SO wrote the manuscript

References

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638.

Andorf, C. M., Cannon, E. K., Portwood, J. L., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., et al. (2016). MaizeGDB update: new tools, data and interface for the maize

model organism database. *Nucleic Acids Res.* 44, D1195-1201. doi:10.1093/nar/gkv1007.

Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed June 25, 2019].

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556.

Askwith, C., Eide, D., Van Ho, A., Bernard, P. S., Li, L., Davis-Kaplan, S., et al. (1994). The FET3 gene of S. cerevisiae encodes a multicopper oxidase required for ferrous iron uptake. *Cell* 76, 403–410. doi:10.1016/0092-8674(94)90346-8.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., et al. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26, 73–79. doi:10.1093/nar/26.1.73.

Cornish-Bowden, A. (2014). Current IUBMB recommendations on enzyme nomenclature and kinetics. *Perspectives in Science* 1, 74–87. doi:10.1016/j.pisc.2014.02.006.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222-230. doi:10.1093/nar/gkt1223.

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32, 309–311. doi:10.1093/bioinformatics/btv557.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., et al. (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature* 422, 859–868. doi:10.1038/nature01554.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418, 387–391. doi:10.1038/nature00935.

Hansen, B. O., Vaid, N., Musialak-Lange, M., Janowski, M., and Mutwil, M. (2014). Elucidating gene function and function evolution through comparison of co-expression networks of plants. *Front. Plant Sci.* 5. doi:10.3389/fpls.2014.00394.

Have, C. T., and Jensen, L. J. (2013). Are graph databases ready for bioinformatics? *Bioinformatics* 29, 3107–3108. doi:10.1093/bioinformatics/btt549.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional

annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44, D286–D293. doi:10.1093/nar/gkv1248.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res* 37, D211–D215. doi:10.1093/nar/gkn785.

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., et al. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* 341, 175–179. doi:10.1126/science.1240230.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031.

Kämper, J., Kahmann, R., Bölker, M., Ma, L.-J., Brefort, T., Saville, B. J., et al. (2006). Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. *Nature* 444, 97–101. doi:10.1038/nature05248.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14, R36. doi:10.1186/gb-2013-14-4-r36.

King, R., Urban, M., Hammond-Kosack, M. C. U., Hassani-Pak, K., and Hammond-Kosack, K. E. (2015). The completed genome sequence of the pathogenic ascomycete fungus Fusarium graminearum. *BMC Genomics* 16, 544. doi:10.1186/s12864-015-1756-1.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., et al. (2015). ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43, D1113–D1116. doi:10.1093/nar/gku1057.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* 27, 718–719. doi:10.1093/bioinformatics/btq671.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Liu, X., Zhang, X., Tang, W.-H., Chen, L., and Zhao, X.-M. (2013). eFG: an electronic resource for Fusarium graminearum. *Database (Oxford)* 2013. doi:10.1093/database/bat042.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15. doi:10.1186/s13059-014-0550-8.

Mi, H., Muruganujan, A., and Thomas, P. D. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic

trees. *Nucleic Acids Res.* 41, D377-386. doi:10.1093/nar/gks1118.

Movahedi, S., Bel, M. V., Heyndrickx, K. S., and Vandepoele, K. (2012). Comparative co-expression analysis in plant biology. *Plant, Cell & Environment* 35, 1787–1798. doi:10.1111/j.1365-3040.2012.02517.x.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. (2009). ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.* 37, D987-991. doi:10.1093/nar/gkn807.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 47, D529–D541. doi:10.1093/nar/gky1079.

Palankar, M. R., Iamnitchi, A., Ripeanu, M., and Garfinkel, S. (2008). Amazon S3 for science grids: a viable solution? in *Proceedings of the 2008 international workshop on Data-aware distributed computing* DADC '08. (Boston, MA, USA: Association for Computing Machinery), 55–64. doi:10.1145/1383519.1383526.

Pedro, H., Maheswari, U., Urban, M., Irvine, A. G., Cuzick, A., McDowall, M. D., et al. (2016). PhytoPath: an integrative resource for plant pathogen genomics. *Nucleic Acids Res* 44, D688–D693. doi:10.1093/nar/gkv1052.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43, e47–e47. doi:10.1093/nar/gkv007.

Stearman, R., Yuan, D. S., Yamaguchi-Iwai, Y., Klausner, R. D., and Dancis, A. (1996). A Permease-Oxidase Complex Involved in High-Affinity Iron Uptake in Yeast. *Science* 271, 1552–1557. doi:10.1126/science.271.5255.1552.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* 302, 249–255. doi:10.1126/science.1087447.

The Complete Genome Sequence of Escherichia coli K-12 | Science Available at: https://science.sciencemag.org/content/277/5331/1453 [Accessed August 7, 2019].

The Gene Ontology Resource: 20 years and still GOing strong (2019). *Nucleic Acids Res* 47, D330–D338. doi:10.1093/nar/gky1055.

The Uniprot Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47, D506–D515. doi:10.1093/nar/gky1049.

Urbanowski, J. L., and Piper, R. C. (1999). The Iron Transporter Fth1p Forms a Complex with the Fet5 Iron Oxidase and Resides on the Vacuolar Membrane. *J. Biol. Chem.* 274, 38061–38070. doi:10.1074/jbc.274.53.38061.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., et al. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation

and caveats. *Plant Cell Environ.* 32, 1633–1651. doi:10.1111/j.1365-3040.2009.02040.x.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, 1–12. doi:10.1038/s41592-019-0686-2.

Vo, T. V., Das, J., Meyer, M. J., Cordero, N. A., Akturk, N., Wei, X., et al. (2016). A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. *Cell* 164, 310–323. doi:10.1016/j.cell.2015.11.037.

Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J. P., Shatalina, M., Roffler, S., et al. (2013). The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nature Genetics* 45, 1092–1096. doi:10.1038/ng.2704.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics* 6, 227. doi:10.1186/1471-2105-6-227.

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., et al. (2002). The genome sequence of Schizosaccharomyces pombe. *Nature* 415, 871–880. doi:10.1038/nature724.

Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., et al. (2012). PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res* 40, D695–D699. doi:10.1093/nar/gkr853.

Zhang, L., Zhang, D., Chen, Y., Ye, W., Lin, Q., Lu, G., et al. (2019). Magnaporthe oryzae CK2 Accumulates in Nuclei, Nucleoli, at Septal Pores and Forms a Large Ring Structure in Appressoria, and Is Involved in Rice Blast Pathogenesis. *Front. Cell. Infect. Microbiol.* 9. doi:10.3389/fcimb.2019.00113.