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

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

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

2 OmicsDriven, Toelloese, Denmark

3 Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany

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.

26 The information is modelled in a graph-based database, enabling flexibility for querying and

27 future extensions. Tools such as BLAST and Cytoscape.js are available together with the

28 option of performing GO enrichment analysis. The database also enables the user to browse

29 information such as Orthologs, Protein domains and publications citing a given gene.

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

31 gene.

32

33 **Availability and Implementation**

34 Currently, Omicsdb supports networks for 10 organisms and is freely available for public

35 use at http://pathogens.omicsdb.org

36

Introduction

38 Plant pathogens are a great threat to the ever-growing population and the following need for

39 food. To combat them, an understanding of their mode of action and of how the pathogenic

40 processes work is essential. To be able to obtain useful knowledge from genomes, it is

41 necessary to be able to correctly assign functions to gene products. Manual annotation is not

42 a feasible option to annotate the ever-growing number of available genomes. Scientists,

43 therefore, must rely more and more on predictions.

44 In-silico methods have gained popularity for automated annotation of gene functions, though

45 for it to be useful, it is important that the accuracy of the predictions, and number of genes

46 that can be annotated is high.

47 The guilt by association hypothesis is based on the observation that genes that participate in

48 similar biological processes or are involved in similar regulatory pathways tend to display

49 similar expression profiles (Wolfe et al., 2005). An increasingly popular method to utilize this

50 observation is using gene co-expression networks. Networks are a convenient way of

51 representing biological data and allow for the utilization of graph theory. By analysing the

52 structure of the networks, it is possible to identify communities within the network, where the

53 genes are more closely related to each other than to the rest of the network and are

54 enriched for specific biological processes.

55 The two main applications for co-expression network analysis are 1) To use a bait-gene from

56 a known pathway or with a known function, as a query, and identify possible genes within

57 the same pathway or with a similar function (Itkin et al., 2013) and 2) to suggest the

58 biological process or pathways given gene is involved in, based on the functions of its

59 neighborhood. With the increase in available biological networks, an increasingly important

60   usage is the identification of common network patterns across species, this allows for more

61   reliable transfer of knowledge from and across model species.

62   Inter-network comparisons serve multiple purposes, they can improve identification of

63   functionally related orthologs across species (Hansen et al., 2014), and aid in the

64   identification of conserved subgraphs within or across species (Stuart et al., 2003).

65

66   The amount of data a researcher working within molecular biology needs to handle has

67   grown exponentially during later years. Not just does that put a heavy load on the IT

68   infrastructure, but also on the researcher that might not be a bioinformatics specialist.

69   Handling and processing omics data are often a requirement these days, to be able to

70   perform analysis on this kind of data, researchers need to be familiar with scripting

71   languages such as Python or R.

72

73   Curated databases have shown to be powerful tools for integration and analysis of data.

74   This has led to an emergence of integrated web-based databases, changing the analysis of

75   networks from being a task for specialist bioinformaticians, into a simple routine task for

76   experimental molecular biologists, investigating specific genes or conditions. However, most

77   of these databases focus on the major model species, such as Maize (Andorf et al., 2016),

78   Budding yeast (Kim et al., 2014) or fission yeast (Vo et al., 2016). There exist very few

79   options for researchers working on non-model species.

80   Plant pathogens have until now had little representation in these databases. Electronic

81   resources for *Fusarium graminearum* eFG (Liu et al., 2013) a model species in plant

82   pathology (Zhang et al., 2019), contains functional annotation as well as annotations of

83   transcription factors and curated and predicted pathogenic genes for fusarium. PHI-base

84   (Urban et al., 2017) is a great resource for curated Pathogen-Host Interactions but is

85   manually curated and thus limited in size. PhytoPath (Pedro et al., 2016), is a resource for

86   genomic information related to plant pathogens. Other databases with microbial data that

87   cover mainly model species are also available (Kim et al., 2014, 2015; Oughtred et al., 2019)

88

89   To tackle this problem, we here present OmicsDB::Pathogens, an integrated database of

90   networks from 8 species: 6 model species (*Escherichia coli, Saccharomyces cerevisiae,*

91   *Saccharomyces pombe, Arabidopsis thaliana, Mus musculus, Homo sapiens*), and 4 plant

92   pathogens (*Fusarium graminearum, Ustilago maydis, Blumeria graminis, Neurospora crassa*

93   *and Magnaporthe oryzae*), as well as orthologs from 158 other fungal pathogens, to ease

94   the knowledge transfer to and from these.

95

3

96 OmicsDB:: Pathogens is a database developed to assist experimental molecular biologists in

97 accessing and analyzing omics data.

98 The database provides intuitive navigation, allowing the researchers to store, browse,

99 analyse and compare their data, as well as handling metadata of both samples and

100 workflows. Another important use of the database is that it allows comparisons between

101 species that can allow the experimental researcher not just to infer possible functions of

102 genes not previously studied experimentally in detail. It also makes it possible to rank genes

103 of interest to study enabling more focussed research in sorting out gene function. We

104 provide an extensive amount of already processed publicly available genomic and

105 transcriptomic data.

106 # Results and discussion

107 OmicsDB is a database that includes high throughput experimental data and information for

108 multiple species. Since we store data from biological networks, it was decided to go for the

109 graph-based database Neo4J for storing most of the data. Neo4J has previously been

110 evaluated by Have and Jensen for several graph processing problems related to

111 bioinformatics and compared with PostgreSQL, they found Neo4J to be faster in many

112 cases, though they also conclude that graph-based databases are not necessarily the best

113 choice for all problems (Have and Jensen, 2013). Neo4J uses the property graph model,

114 which means that nodes and edges can have key/value properties associated. Neo4J uses

115 its own query language, Cypher for querying the graph, cypher allows queries to be

116 formulated in terms of paths, which allows them to be concise and intuitive compared with

117 equivalent SQL queries, which is often complicated by joins, and difficult to read. Due to the

118 simplicity of traversing edges, and accessing data through the Cypher language, we decided

119 to use Neo4J for all data gene expression data, that is stored in an SQLite database, and flat

120 files that are stored in an S3 compatible object storage, with their paths and metadata stored

121 in Neo4J linked to the relevant organism. A simplified overview of our Neo4J data model can

122 be seen in Fig. 1 An overview of the infrastructure and data processing steps can be seen in

123 Fig. 2. The processed expression data is stored in a SQL database, with common identifiers

124 shared between the other systems.

125
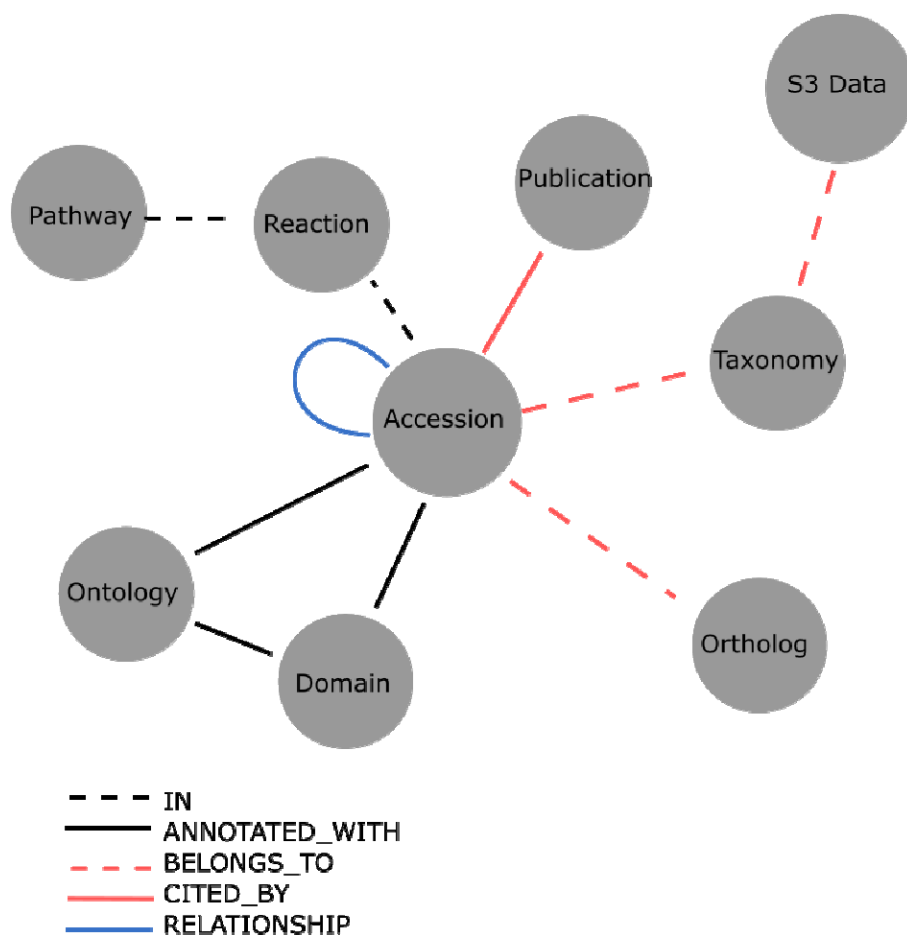
Fig 1. The part of our Neo4J data model. This shows how the different nodes are related.
Properties are left out for simplicity. Gathering data across any node is easy, for example, if
one wishes to know all publications related to a given Domain its possible to traverse the
graph from Domain over Accession 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.

141
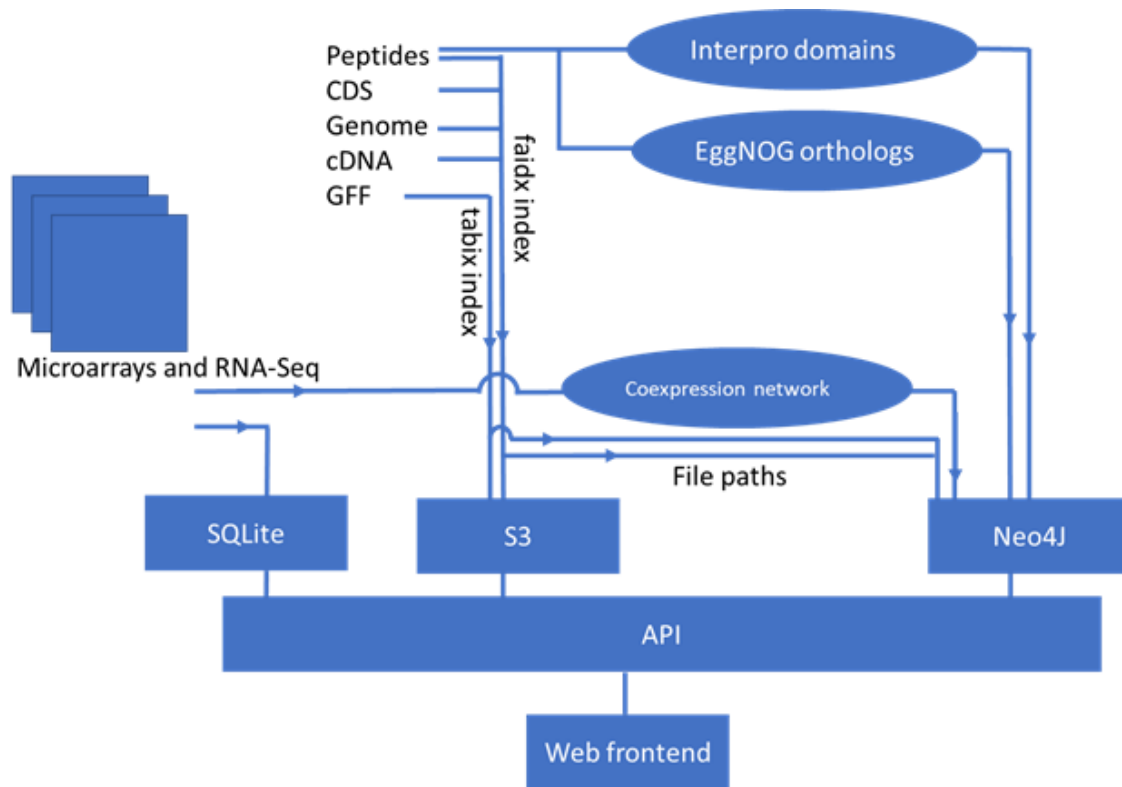
Figure 2. Storage and pipelines included in the data processing and storage. Microarray and RNA-seq data are stored processed and normalized (not shown). The TPMs are then stored in an SQLite database. The TPMs is also processed to generate a co-expression network, stored in Neo4J. Peptide, CDS, cDNA, 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 InterProScan to find conserved domains. Both orthologs and domains are stored in Neo4J, connected to their respective peptides.

150 Basic functions of OmicsDB::Pathogens

151 The main source of information is the gene page that offers important information regarding

152 each gene. It provides different information for each gene. An overview can be seen in Fig 3.

153 The top of the page (Fig 3A) shows the basic information: The gene model with intron and

154 exon information. Gene Ontology annotations (Ashburner et al., 2000). Ortholog groups from

155 EggNOG (Huerta-Cepas et al., 2016) and protein domains from InterProScan (Hunter et al.,

156 2009; Jones et al., 2014) and the coding DNA sequence and the protein sequence. Further

157 down the page (Fig 3B) is a list of co-expressed genes and a visualization of the co-

158 expression network. As well as A bar chart of gene expression displaying the average of

159 replicates for a given experiment (Fig 3C).
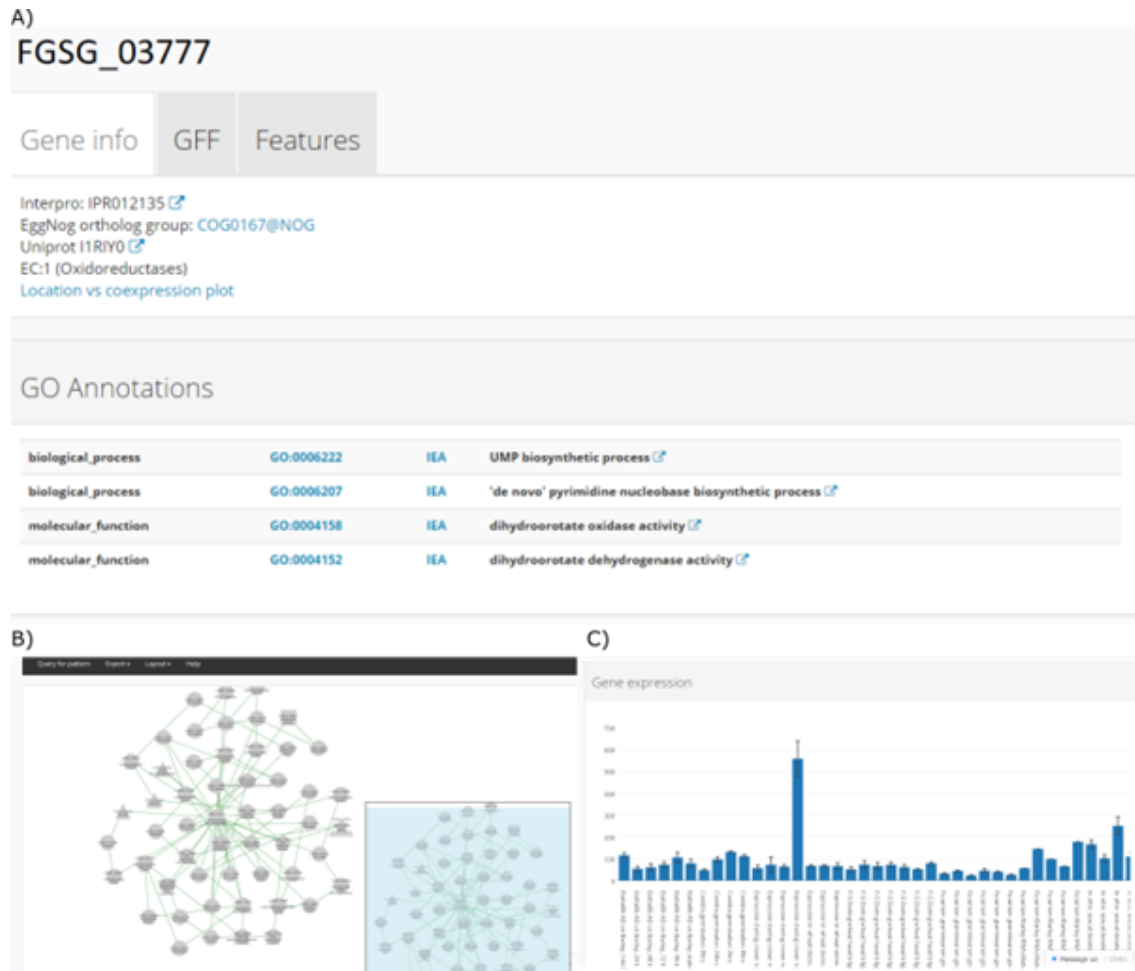
160

161 Fig 3. An example of the gene profile page. A) Shows the structural and functional

162 information on the gene as well as ortholog information, possible pathways, and InterPro

163 domains. B) Shows an example of a co-expression network. C) Shows the expression profile

164 of a given gene. Both A, B and C are found on the same page in the browser.

165

166 This page also serves as a gateway to explore the information within and across species.

167 For the species without networks, this data is limited to protein domains, ortholog families,

168 DNA and protein sequences.

169 The network representation of gene interactions allows the users to query for the pattern of

170 co-regulated orthologs across other species, to see if this regulated "module" is preserved.

171 This function can ease the annotation transfer of "biological process" terms, relying not only

172 on sequence similarity but also on the biological context.

173 To give a quick overview of what types of proteins the genes in the network code for, shapes

174 of the nodes are used. The default shape is round, but it will be given a different shape in the

175 case of transcription factors, or if a protein is classified with one of the 1-6 Enzyme

176   commission classifications (Cornish-Bowden, 2014). The node representing the bait gene is

177   displayed in red. An overview of the shapes can be found in Supplementary Table 1. To

178   represent different types of relationships, the edges are colored using the following scheme,

179   Green edges: co-expression, red edges: physical interaction, blue edges: genetic interaction.

180   The physical and genetic interactions are from The BioGrid database (Oughtred et al.,

181   2019), the co-expression data is calculated using the Mutual Rank (MR) method (Obayashi

182   et al., 2009) modified to 1/MR with a MR threshold of 50.

183

184   From the gene page is it possible to access other pages with information. Ortholog pages:

185   Each EggNOG ortholog family has a dedicated page, with aggregated information on this

186   family. Including genes and species in the family, functional annotations as well as papers

187   citing the family. This enables users to easily check at a glance what information is available

188   from other species. As well as how diverse the family is by visualizing the distribution of

189   members per species in a pie chart. An example of an ortholog page can be seen in Fig 4.

190   Domain pages are similar to ortholog pages, but centred around a domain, for example,

191   PF00331 or PS51760.
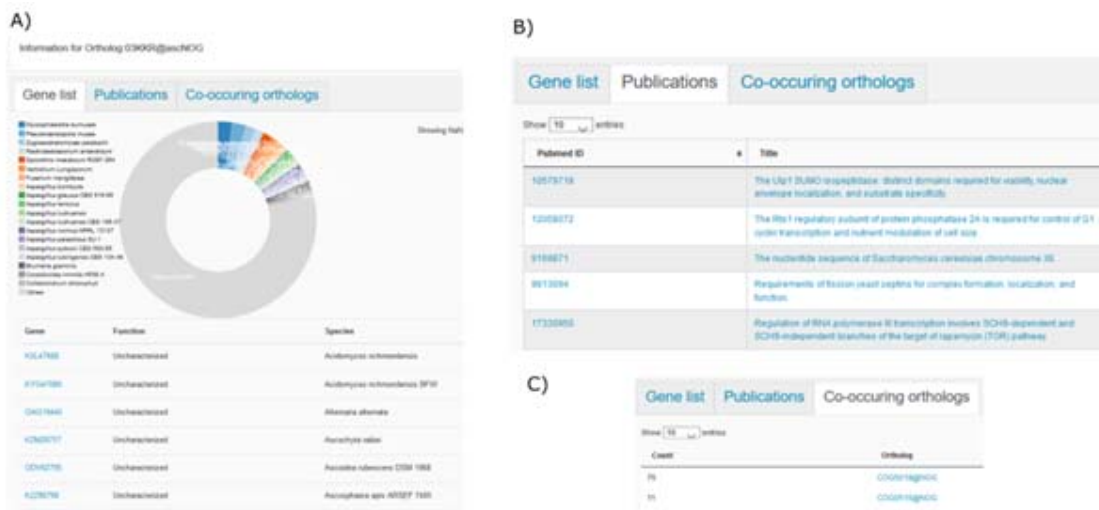
192

193   *Fig 4. Each Ortholog has a unique page with multiple tabs. A) it shows all that has been*

194   *assigned and the species they belong to. It is possible to sub select species using the pie*

195   *chart. The other tabs include B) publications citing genes within this ortholog group. As well*

196   *as C) orthologs that co-occur across the networks. Orthologs that often co-occur hints that*

197   *they might participate in a similar process.*

198   The Gene Ontology(Ashburner et al., 2000) has also been integrated, with pages for each

199   term, displaying all genes annotated with a given ontology term, as well as the position of the

200   term in the GO graph. It is thus possible to find networks in one species that resembles

201 those affiliated with a certain ontology in another species. Traditionally the user would have

202 had to download and annotate both datasets. However here it can be done with the click of a

203 button.

Advanced functions

205 OmicsDB also provides some tools enabling users to perform more advanced analyses. One

206 example is cross-species network alignment. Cross-species comparisons of biological

207 networks with interactions are still an emerging field, It allows for finding network patterns

208 that between orthologous genes across species, hinting towards similar function. This has

209 been used successfully in plants (Usadel et al., 2009; Movahedi et al., 2012; Hansen et al.,

210 2014). Most work on gene function prediction using gene co-expression so far has been

211 working on single or few species. Identifying conserved co-expression patterns across

212 orthologs in different species can identify highly relevant candidate genes sharing similar

213 functions or participate in the same pathway (Movahedi et al., 2012; Hansen et al., 2014).

214 An example of such a strategy with experimental confirmation is the study by Itkin et al.,

215 where comparative co-expression for tomato and potato was utilized, leading to the

216 discovery of a gene cluster that is related to the steroidal glycoalkaloids pathway (Itkin et al.,

217 2013).

218 It is also possible to use the surrounding network and look for enriched GO terms, using a

219 hypergeometric test. This is done at the bulk analysis page. Where a list of genes can be

220 analysed. A plot of the expression values for all genes in the list will be generated as well as

221 the GO enrichment. It is also possible to export both GO enrichments as well as peptide and

222 CDS sequences in fasta format.

Use case: Vacuolar iron uptake

224 Yeast Fet3/Ftr1 (YMR056W/YER145C) are genes for two proteins that are in yeast high-

225 affinity iron uptake at the plasma membrane (Askwith et al., 1994, 3; Stearman et al., 1996).

226 The proteins physically connect to each other. Ftr1 is a transporter and Fet3 is an

227 oxidoreductase. A similar system Fet5/Fth1 (YFL071W/YBR207W) exists at the vacuolar

228 membrane and is involved in regulating vacuolar iron storage (Urbanowski and Piper, 1999)

229 In *Fusarium graminearum* FGSG_02143 is known to be the plasma membrane Ftr1-like

230 protein and the homologous protein FGSG_05160 is probably the Fth1-like protein located at

231 the vacuole membrane. To test if this can be the case FGSG_05160 was used as bait and

9

232   the result is displayed together with the networks for yeast Ftr1 and Fth1. The aligned

233   network can be seen in Fig. 5. Red edges represent possible orthologs and black edges co-

234   expression within the species. The subcellular location for some genes products in Yeast

235   has been marked.

236   What can be seen is that the vacuolar and the plasma-membrane iron transport is more co-

237   regulated in the *F. graminearum* data than in the yeast data (direct strong co-regulation

238   between FGSG_05160 and FGSG_02143). The pattern of co-regulation of the other proteins

239   with orthologues in yeast also supports that FGSG_05160 is a vacuolar located Fth1-

240   orthologue. This has, however, to be confirmed by direct experiments.

241
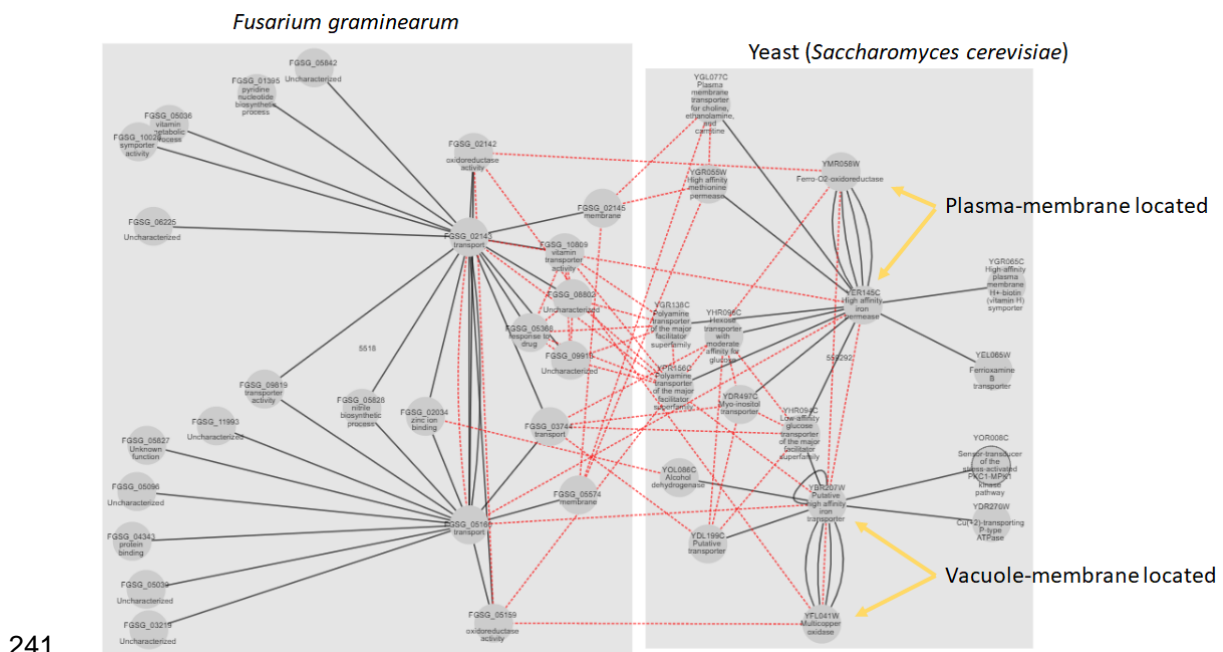
242   *Fig 5. The alignment of the co-expression neighbourhoods from Yeast and F. graminearum,*

243   *Using FGSG_05160 as a bait, and searching for counterparts in Yeast with similar co-*

244   *expression patterns. Black edges represent co-expression edges, red edges represent*

245   *possible orthologs relationship. It can be seen that large parts of the network are mirrored*

246   *across the organisms, hinting towards the genes carrying out the same functionality.*

Conclusions and future directions

247

248   In this study, we developed a system for handling biological omics data. We processed gene

249   expression data, as well as generated co-expression networks for 10 species.

250   We established an interactive web interface omicsdb.org to provide access to the data as

251   well as the analysis platforms for the public. We plan to continue to improve the quality and

252   functionality of this database, by regularly updating with new publicly available data.

253   Currently the database contains networks for 4 pathogens and 6 references, however, there

254   is a growing need for understanding how plant pathogens work, to alleviate this need we will

255   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 co-expression 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 co-expression, 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

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

288

289   With A and B represent genes. GFF files were stored in tabix format (Li, 2011) peptide and

290   DNA sequences were stored in faidx format implemented in Samtools (Li et al., 2009).

Functional Annotations

292   Protein domains including PFAM domains (Finn et al., 2014) and Panther (Mi et al., 2013)

293   were identified using InterProScan v. 5.29-68.0 (Hunter et al., 2009; Jones et al., 2014),

294   Ortholog groups were identified using EggNOG (Huerta-Cepas et al., 2016), which also

295   provided predicted GO terms. For *E. coli, S. pombe, S. cerevisiae A. thaliana, M. musculus*

296   *and Homo Sapiens*, experimentally validated GO terms could be downloaded from the

297   GeneOntology website (The Gene Ontology Consortium, 2019). The Gene Ontology website

298   also provides mappings from domains identified by InterProScan to their associated GO

299   terms. These maps were used to further improve the annotations.

300   Papers citing a given gene or gene family were retrieved from Uniprot (The Uniprot

301   Consortium, 2019)**.**

302   For reference, Arabidopsis thaliana, Mus musculus and Homo sapiens were included,

303   annotations were derived similarly to the pathogens.

Network alignment

305   The network alignment uses the "@NOG" ortholog families, if you provide a bait gene, it will

306   query all other genes from this family, and compare how many orthologs are shared in the

307   1st-degree neighbourhood. To calculate how much of the network around two genes is

308   similar, the Jaccard index for the gene families shared across the networks is then

309   calculated following

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

311   Where A and B represent the gene families in the network surrounding your bait gene and

312   the gene it is compared against respectively. The numerator is the intersection of gene

313   families, and denominator represents the union of all gene families in the 1st-degree

314   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.

344   doi:10.1093/nar/gkv1007.

345   Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available at:

346   http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed June 25, 2019].

347   Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000).

348   Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.

349   *Nat. Genet.* 25, 25–29. doi:10.1038/75556.

350   Askwith, C., Eide, D., Van Ho, A., Bernard, P. S., Li, L., Davis-Kaplan, S., et al. (1994). The

351   FET3 gene of S. cerevisiae encodes a multicopper oxidase required for ferrous iron

352   uptake. *Cell* 76, 403–410. doi:10.1016/0092-8674(94)90346-8.

353   Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al.

354   (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.

355   doi:10.1186/1471-2105-10-421.

356   Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., et al. (1998).

357   SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26, 73–79.

358   doi:10.1093/nar/26.1.73.

359   Cornish-Bowden, A. (2014). Current IUBMB recommendations on enzyme nomenclature

360   and kinetics. *Perspectives in Science* 1, 74–87. doi:10.1016/j.pisc.2014.02.006.

361   Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al.

362   (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222-230.

363   doi:10.1093/nar/gkt1223.

364   Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016).

365   Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32,

366   309–311. doi:10.1093/bioinformatics/btv557.

367   Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., et al.

368   (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature*

369   422, 859–868. doi:10.1038/nature01554.

370   Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., et al. (2002).

371   Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418, 387–391.

372   doi:10.1038/nature00935.

373   Hansen, B. O., Vaid, N., Musialak-Lange, M., Janowski, M., and Mutwil, M. (2014).

374   Elucidating gene function and function evolution through comparison of co-

375   expression networks of plants. *Front. Plant Sci.* 5. doi:10.3389/fpls.2014.00394.

376   Have, C. T., and Jensen, L. J. (2013). Are graph databases ready for bioinformatics?

377   *Bioinformatics* 29, 3107–3108. doi:10.1093/bioinformatics/btt549.

378   Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al.

379   (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional

380   annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 44,

15

381   D286–D293. doi:10.1093/nar/gkv1248.

382   Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009).

383   InterPro: the integrative protein signature database. *Nucleic Acids Res* 37, D211–

384   D215. doi:10.1093/nar/gkn785.

385   Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., et al. (2013).

386   Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by

387   clustered genes. *Science* 341, 175–179. doi:10.1126/science.1240230.

388   Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014).

389   InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30,

390   1236–1240. doi:10.1093/bioinformatics/btu031.

391   Kämper, J., Kahmann, R., Bölker, M., Ma, L.-J., Brefort, T., Saville, B. J., et al. (2006).

392   Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis.

393   *Nature* 444, 97–101. doi:10.1038/nature05248.

394   Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013).

395   TopHat2: accurate alignment of transcriptomes in the presence of insertions,

396   deletions and gene fusions. *Genome Biology* 14, R36. doi:10.1186/gb-2013-14-4-r36.

397   King, R., Urban, M., Hammond-Kosack, M. C. U., Hassani-Pak, K., and Hammond-Kosack,

398   K. E. (2015). The completed genome sequence of the pathogenic ascomycete

399   fungus Fusarium graminearum. *BMC Genomics* 16, 544. doi:10.1186/s12864-015-

400   1756-1.

401   Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., et al.

402   (2015). ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43,

403   D1113–D1116. doi:10.1093/nar/gku1057.

404   Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files.

405   *Bioinformatics* 27, 718–719. doi:10.1093/bioinformatics/btq671.

406   Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The

407   Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

408   doi:10.1093/bioinformatics/btp352.

409   Liu, X., Zhang, X., Tang, W.-H., Chen, L., and Zhao, X.-M. (2013). eFG: an electronic

410   resource for Fusarium graminearum. *Database (Oxford)* 2013.

411   doi:10.1093/database/bat042.

412   Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and

413   dispersion for RNA-seq data with DESeq2. *Genome Biol* 15. doi:10.1186/s13059-

414   014-0550-8.

415   Mi, H., Muruganujan, A., and Thomas, P. D. (2013). PANTHER in 2013: modeling the

416   evolution of gene function, and other gene attributes, in the context of phylogenetic

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

418   Movahedi, S., Bel, M. V., Heyndrickx, K. S., and Vandepoele, K. (2012). Comparative co-
419   expression analysis in plant biology. *Plant, Cell & Environment* 35, 1787–1798.
420   doi:10.1111/j.1365-3040.2012.02517.x.
421   Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. (2009). ATTED-II provides
422   coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.* 37, D987-991.
423   doi:10.1093/nar/gkn807.
424   Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., et al. (2019). The
425   BioGRID interaction database: 2019 update. *Nucleic Acids Res* 47, D529–D541.
426   doi:10.1093/nar/gky1079.
427   Palankar, M. R., Iamnitchi, A., Ripeanu, M., and Garfinkel, S. (2008). Amazon S3 for science
428   grids: a viable solution? in *Proceedings of the 2008 international workshop on Data-*
429   *aware distributed computing* DADC '08. (Boston, MA, USA: Association for
430   Computing Machinery), 55–64. doi:10.1145/1383519.1383526.
431   Pedro, H., Maheswari, U., Urban, M., Irvine, A. G., Cuzick, A., McDowall, M. D., et al. (2016).
432   PhytoPath: an integrative resource for plant pathogen genomics. *Nucleic Acids Res*
433   44, D688–D693. doi:10.1093/nar/gkv1052.
434   Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers
435   differential expression analyses for RNA-sequencing and microarray studies. *Nucleic*
436   *Acids Res* 43, e47–e47. doi:10.1093/nar/gkv007.
437   Stearman, R., Yuan, D. S., Yamaguchi-Iwai, Y., Klausner, R. D., and Dancis, A. (1996). A
438   Permease-Oxidase Complex Involved in High-Affinity Iron Uptake in Yeast. *Science*
439   271, 1552–1557. doi:10.1126/science.271.5255.1552.
440   Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A Gene-Coexpression Network for
441   Global Discovery of Conserved Genetic Modules. *Science* 302, 249–255.
442   doi:10.1126/science.1087447.
443   The Complete Genome Sequence of Escherichia coli K-12 | Science Available at:
444   https://science.sciencemag.org/content/277/5331/1453 [Accessed August 7, 2019].
445   The Gene Ontology Consortium (2019). Gene Ontology Resource: 20 years and still GOing
446   strong. *Nucleic Acids Res* 47, D330–D338. doi:10.1093/nar/gky1055.
447   The Uniprot Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic*
448   *Acids Res* 47, D506–D515. doi:10.1093/nar/gky1049.
449   Urbanowski, J. L., and Piper, R. C. (1999). The Iron Transporter Fth1p Forms a Complex
450   with the Fet5 Iron Oxidase and Resides on the Vacuolar Membrane. *J. Biol. Chem.*
451   274, 38061–38070. doi:10.1074/jbc.274.53.38061.
452   Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., et al.
453   (2009). Co-expression tools for plant biology: opportunities for hypothesis generation
454   and caveats. *Plant Cell Environ.* 32, 1633–1651. doi:10.1111/j.1365-

455   3040.2009.02040.x.

456   Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al.

457   (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat*

458   *Methods*, 1–12. doi:10.1038/s41592-019-0686-2.

459   Vo, T. V., Das, J., Meyer, M. J., Cordero, N. A., Akturk, N., Wei, X., et al. (2016). A

460   Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from

461   Yeasts to Human. *Cell* 164, 310–323. doi:10.1016/j.cell.2015.11.037.

462   Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J. P., Shatalina, M., Roffler, S., et al.

463   (2013). The wheat powdery mildew genome shows the unique evolution of an

464   obligate biotroph. *Nature Genetics* 45, 1092–1096. doi:10.1038/ng.2704.

465   Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general

466   applicability of "guilt-by-association" within gene coexpression networks. *BMC*

467   *Bioinformatics* 6, 227. doi:10.1186/1471-2105-6-227.

468   Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., et al. (2002).

469   The genome sequence of Schizosaccharomyces pombe. *Nature* 415, 871–880.

470   doi:10.1038/nature724.

471   Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., Staines, D. M., et

472   al. (2012). PomBase: a comprehensive online resource for fission yeast. *Nucleic*

473   *Acids Res* 40, D695–D699. doi:10.1093/nar/gkr853.

474   Zhang, L., Zhang, D., Chen, Y., Ye, W., Lin, Q., Lu, G., et al. (2019). Magnaporthe oryzae

475   CK2 Accumulates in Nuclei, Nucleoli, at Septal Pores and Forms a Large Ring

476   Structure in Appressoria, and Is Involved in Rice Blast Pathogenesis. *Front. Cell.*

477   *Infect. Microbiol.* 9. doi:10.3389/fcimb.2019.00113.