© 2020. Published by The Company of Biologists Ltd | Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

RESEARCH ARTICLE

Informing epidemic (research) responses in a timely fashion by knowledge management – a Zika virus use case

Angela Bauch[1,‡], Johann Pellet[2], Tina Schleicher[1], Xiao Yu[3], Andrea Gelemanović[4], Cosimo Cristella[3], Pieter L. Fraaij[5], Ozren Polasek[4], Charles Auffray[2], Dieter Maier[1], Marion Koopmans[5,*] and Menno D. de Jong[3,*]

ABSTRACT

The response of pathophysiological research to emerging epidemics often occurs after the epidemic and, as a consequence, has little to no impact on improving patient outcomes or on developing high-quality evidence to inform clinical management strategies during the epidemic. Rapid and informed guidance of epidemic (research) responses to severe infectious disease outbreaks requires quick compilation and integration of existing pathophysiological knowledge. As a case study we chose the Zika virus (ZIKV) outbreak that started in 2015 to develop a proof-of-concept knowledge repository. To extract data from available sources and build a computationally tractable and comprehensive molecular interaction map we applied generic knowledge management software for literature mining, expert knowledge curation, data integration, reporting and visualization. A multi-disciplinary team of experts, including clinicians, virologists, bioinformaticians and knowledge management specialists, followed a pre-defined workflow for rapid integration and evaluation of available evidence. While conventional approaches usually require months to comb through the existing literature, the initial ZIKV KnowledgeBase (ZIKA KB) was completed within a few weeks. Recently we updated the ZIKA KB with additional curated data from the large amount of literature published since 2016 and made it publicly available through a web interface together with a step-by-step guide to ensure reproducibility of the described use case. In addition, a detailed online user manual is provided to enable the ZIKV research community to generate hypotheses, share knowledge, identify knowledge gaps, and interactively explore and interpret data. A workflow for rapid response during outbreaks was generated, validated and refined and is also made available. The process described here can be used for timely structuring of pathophysiological knowledge for future threats. The resulting structured biological knowledge is a helpful tool for computational data analysis and generation of predictive models and opens new avenues for infectious disease research. ZIKV Knowledgebase is available at www.zikaknowledgebase.eu.

[1]Biomax Informatics AG, 82152 Planegg, Germany. [2]European Institute of Systems Biology and Medicine, 69390 Lyon, France. [3]Department of Medical Microbiology, Amsterdam UMC, University of Amsterdam, Amsterdam 1105 AZ, the Netherlands. [4]Department of Public Health, University of Split School of Medicine, 21000 Split, Croatia. [5]Department of Viroscience and Department of Paediatrics, Erasmus Medical Centre, 3000 CA Rotterdam, the Netherlands.
*These authors contributed equally to this work

[‡]Author for correspondence (angela.bauch@biomax.com)

A.B., 0000-0002-8712-4414

Received 29 May 2020; Accepted 28 October 2020

KEY WORDS: Zika virus, Emerging epidemics, Knowledge base, Molecular interaction map

INTRODUCTION

The response to a (re-)emerging infectious disease (ID) epidemic requires a rapid compilation of existing pathophysiological knowledge to inform research priorities guiding basic and clinical research. Gaps in understanding of the underlying mechanisms make it difficult to design effective disease-modifying therapies. Hence, during an emerging ID outbreak, the available information at the time of its emergence and the subsequent rapid accumulation of scientific knowledge from various sources needs to be captured and analyzed in a timely and comprehensive fashion. Responding to an ID outbreak therefore would benefit from the use of a knowledge repository that organizes the disease-related knowledge into pathway, molecular interaction and disease maps. Such maps are a relatively new concept that have been used in neurodegenerative and heart diseases (Fujita et al., 2014; Nim et al., 2015), but which have had limited application in the field of ID thus far (Guo et al., 2010; Le Breton et al., 2011; Matsuoka et al., 2013).

Molecular interaction and disease maps are dynamic computer-based knowledge repositories developed to integrate data and information across information sources, in a manner that is customized to the research domain of interest. Data types include interactions between molecular components, such as genes, pathogens, compounds and diseases.

The Platform for European Preparedness Against (Re-)emerging Epidemics (PREPARE) is an EU-funded research consortium and clinical research network with the aim to rapidly respond to severe ID outbreaks, generating real-time evidence to inform optimized clinical management of patients and public health response. The 2015 Zika virus (ZIKV) outbreak was considered as a test case in the context of the PREPARE network, as the pathogenesis of neurologic or immune disease induced by ZIKV is not fully understood. ZIKV is a flavivirus belonging to the Flaviviridae family and had only marginally been researched prior to the 2015 epidemic was minimal (Anderson et al., 2016; Pierson and Diamond, 2018; Pardy and Richer, 2019). Outbreaks of ZIKV disease have been recorded in Africa, the Americas, Asia and the Pacific. Acute ZIKV infections are mostly asymptomatic or associated with mild and self-limiting symptoms of fever, rash, conjunctivitis, headache or joint pain (Murray, 2017; Sharma et al., 2019). However, the unexpected association of ZIKV infection with pregnancy and the subsequent severe neurodevelopmental problems in offspring and with the occurrence of neurological illnesses such as Guillain-Barre syndrome (GBS) or meningoencephalitis in acutely infected patients, led to widespread global concerns and a Public Health Emergency of International Concern (PHEIC) declaration by World Health Organization (WHO) in 2016 (Pierson and Diamond, 2018).

We used the ZIKV virus outbreak as a case study to develop and test the steps, tasks, protocols and tools necessary to rapidly gather and

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

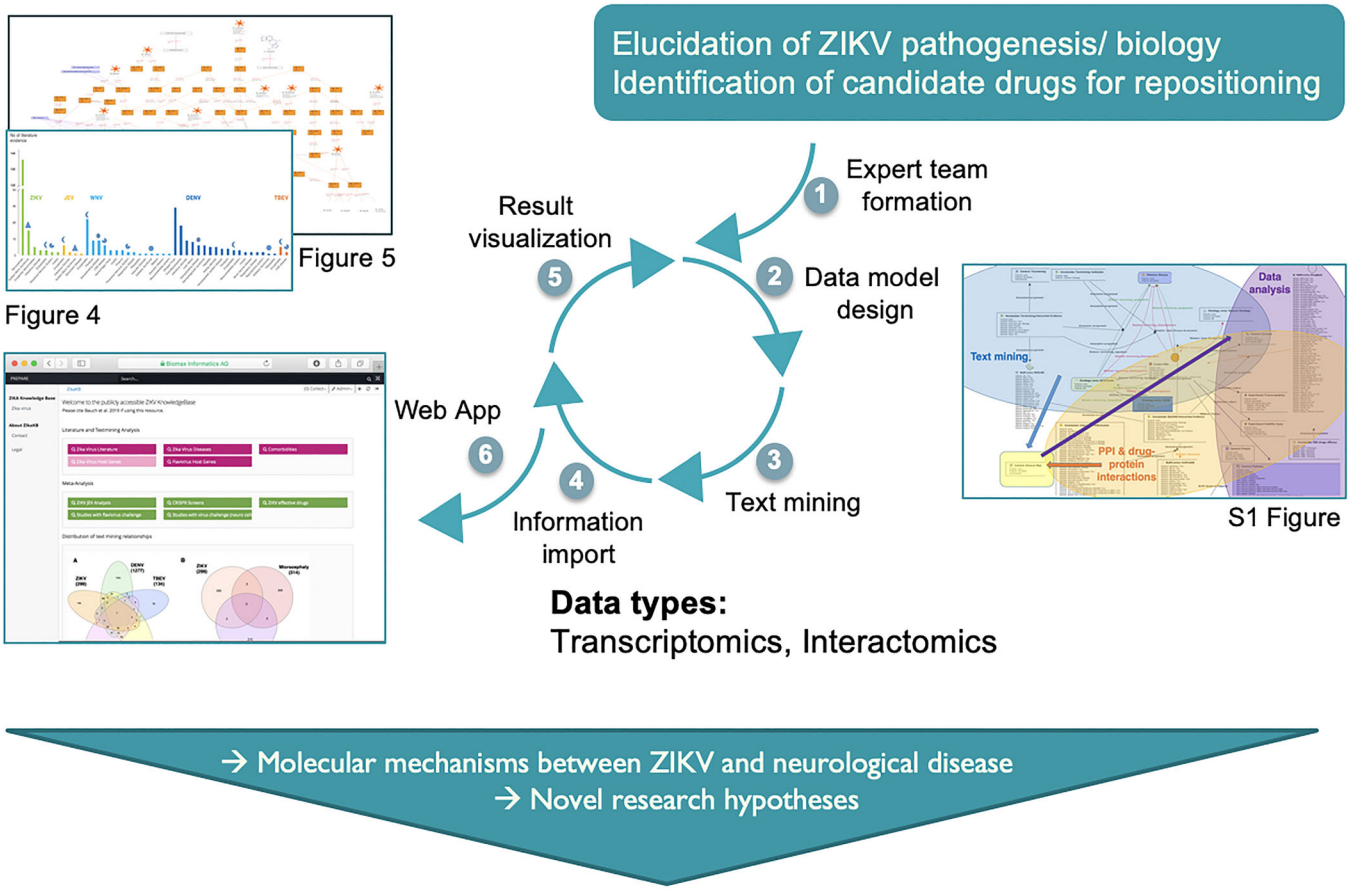

Fig. 1. ZIKV KnowledgeBase generation process – overview. Based on the research objectives and knowledge provided by clinical/virology domain experts a six-step process was applied. In the first step a multidisciplinary expert team is assembled, in step 2, a semantic representation ('data model') was designed by the knowledge management experts. This model includes details about the data sources for integration, how to transfer data into the system and how to report, visualize and export results, as well as the definition of the semantic context for objects, such as 'gene', 'cell type' and 'strain'. In a third step, a natural language processing algorithm was applied to the integrated PubMed literature source. In step 4 the relevant data, including literature mining results, were imported into the system and semantically mapped to the data model. In step 5, queries, views and reports were formulated. In the last step a web-browser based user interface was implemented to enable clinical/virology experts to review, validate and refine the integrated information.

integrate existing and emerging knowledge and to inform research priorities (Fig. 1). Based on the available data and information we aimed to obtain a general overview of pathophysiological knowledge on ZIKV infection and its associated clinical manifestations described in the public domain. Other neurotropic flaviviruses, such as Dengue virus (DENV), West Nile virus (WNV), Japanese encephalitis virus (JEV) and Tick-borne encephalitis virus (TBEV) also cause nervous system infections, in particular encephalitis, but no association with neurodevelopmental disorders or GBS have been reported (Carod-Artal, 2016). To see whether including these viruses would shed additional light on ZIKV pathogenesis we compared available ZIKV information to other neurotropic flaviviruses in terms of neurovirulence and disease severity.

RESULTS

Semantic representation of ZIKV infection

The data model implemented to provide a semantic representation of ZIKV infection is described in detail in Fig. S1. Briefly, the model focuses on genes, diseases, pathogens and drugs, and distinguishes between associations derived from literature mining and those provided by experimental data such as protein–protein interactions (PPIs).

Text mining results

We searched PubMed with the terms 'Zika virus', 'Dengue', 'West Nile virus', 'Japanese encephalitis virus', 'Tick-borne encephalitis virus', 'Microcephaly' and 'Guillain-Barre Syndrome' initially in December 2016 and most recently in September 2018. The recent search resulted in 4927 hits for 'Zika virus' and 19,974, 7700, 5918, 5213, 14,248 and 8615 hits for the other search terms, respectively. During the analyzed time frame, literature on ZIKV increased substantially from 1414 in 2016 to the current 4927 hits (250%), whereas for all other terms, the increase in publications was closer to 10%. Accordingly, the recent search identified additional disease phenotypes, including carditis and skin diseases, which were reported to be associated with ZIKV that were not present in the previous search. An additional set of 236 open access full text articles about ZIKV were included. A natural language processing algorithm was applied to these sets of documents to efficiently extract the fast growing information in the biomedical literature. The text mining extracted a total of 11,916 relationships, which were manually evaluated to 2982 verified relationships (Table 2). The distribution of the curated relationships is depicted in Fig. 2, indicating that the largest overlap was for ZIKV and DENV and for DENV and WNV. The curated set of relationships was used

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

Table 1. External data sources integrated into ZIKV KnowledgeBase

Source database	Information type	Current statistics	Level of curation	Update frequency and version
ATC	Anatomical therapeutic classification system	6064		2016
BioGRID	Protein–protein interaction	293,022	Manually curated from literature Different evidence codes	Updated monthly Version 3.4.137
BioGRID	Protein–drug interaction	10,722	Manually curated from literature Different evidence codes	Updated monthly Version 3.4.137
ChEBI	Compound information	161,090	Curated from different data sources	Updated weekly
DisGeNET	Gene–disease associations	429,036	Integrated from several public data sources and literature Score for ranking associations	Permanently updated Version 4.0.0.0
Disease ontology	Standardized ontology for human disease	15,043	Manually curated	Updated weekly
DrugBank	Drug and drug target database	8203	Manually curated from literature	Updated weekly
EntrezGene	Gene functional information	>24 million	Curated information integrated from different databases, based on RefSeq genomes	Updated weekly
Human Phenotype Ontology	Standardized vocabulary of phenotypic abnormalities in human disease	11,592		2016
KEGG	Pathways and reactions	273	Manually curated from literature	2008
NCI Thesaurus	Controlled vocabulary of the National Cancer Institute	118,502	Manually curated from literature	Updated weekly
OMIM	Gene–disease relations	21,395	Curated from literature	Updated weekly
PubMed	Literature	>24 million	Automatic collection with manual curation	Updated weekly
Reactome	Pathways and reactions	5334	Manually curated from literature	Updated quarterly Aug 2016
UniProtKB	Protein sequences	>65 million	Manually curated	Updated bi-weekly
VirHostNet	Virus/host molecular interactions	44,310	Manually curated from literature	2.0 (March 2016)

for further analyses, including generation of molecular interaction and disease maps and querying for virus-associated genes or diseases.

Integrated data

Overall, the ZIKV KnowledgeBase (ZIKA KB) contains a network of 337,332 human and host-pathogen PPI integrated from BioGRID and VirHostNet, as well as 18905 protein–drug interactions integrated from BioGRID and DrugBank, and 450,431 gene-disease associations from DisGeNET (Table 1). Recently, a variety of ZIKV- and other flavivirus-related large-scale data sets, including microarray gene expression (Kumari et al., 2016; Nowakowski et al., 2016), RNAseq (Tang et al., 2016) as well as CRISPR/Cas data (Zhang et al., 2016), have become publicly available and were integrated to identify host factors that are affected during viral infection.

Molecular interaction and disease maps

Curated text mining results were used to populate the initial ZIKV molecular interaction and disease map. In a second step the map was extended with interaction data (PPI and protein–drug interaction) by applying a network search to implement the breadth-first algorithm (Moore, 1959), which connected genes extracted from text mining relationships based on the overall network. This set of interaction data can be filtered and explored interactively. In a systems medicine approach, a multidisciplinary expert team systematically analyzed literature, public databases and experimental resources to create a formal, structured model of molecular and cellular ZIKV–host interactions (‘molecular interaction and disease map’).

Table 2. Text mining analyses

Text corpus	Key word	Number of documents	Number of relationships	Curated relationships
ZIKV virus FT	Zika virus	4927; 236 FT[a]	1192	298
Medline	Dengue	19,974	5868	1277
Medline	West Nile virus	7700	1586	387
Medline	Japanese encephalitis virus	5918	1594	354
Medline	Tick-borne encephalitis virus	5213	508	134
Medline	Microcephaly	6896	749	314
Medline	Guillain-Barre syndrome	4133	419	218
Total			11,916	2982

[a]FT: full text.

Publicly available ZIKA KB

After an assessment period of internal use, a web-browser based user interface was implemented to make the PREPARE ZIKA KB available to all ZIKV researchers. By openly sharing the collected data and information, the ZIKA KB allows researchers to generate hypotheses, identify knowledge gaps and interactively explore and interpret data. All data are currently in the public domain. Upon request, data submission can be modified to allow registered users to specify that submitted data should not be publicly available.

Use of the ZIKA KB

In the following we provide several example use cases. For instance, publicly available interaction data, such as the PPI and protein–drug interaction data can be used to visualize drug targets and host factors involved in ZIKV pathogenesis. Alternatively, users can filter for PPIs whose source or target is a drug or refine search results to include only proteins localized to a specific cellular compartment, such as the endoplasmic reticulum. The returned networks can be interrogated subsequently to identify host factors that are targeted by the virus and to search for drugs that interact with these host factors and thus might contribute to drug repositioning for future treatment options for ZIKV infection. The maps can also be explored further by using integrated expression and knockout data.

To explore the integrated literature knowledge for relevance or obtain an overview of drug targets or identify critical genes within the network consisting of gene-disease-pathogen relationships,

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

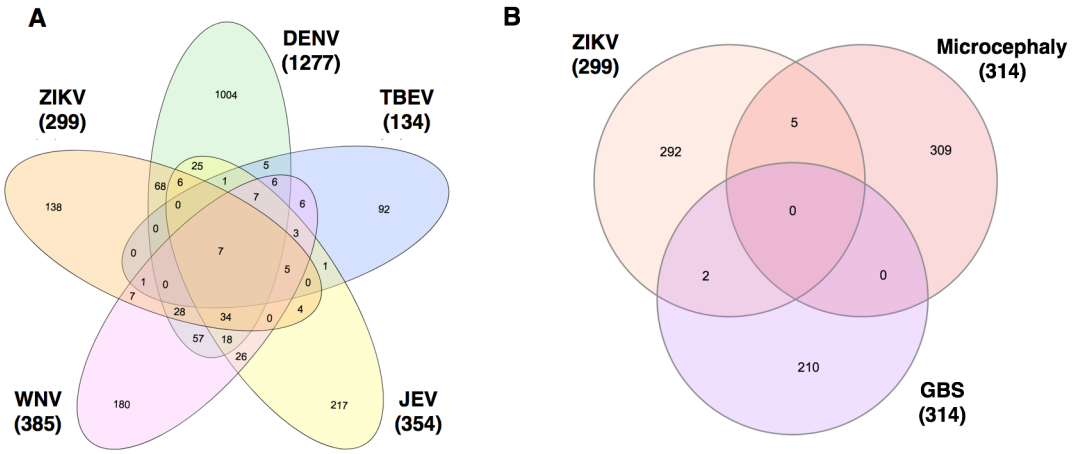

Fig. 2. Distribution of text mining relationships. Numbers represent the sum of different types of relationships, such as gene-pathogen, gene-disease, gene-gene, gene-compound and pathogen-disease relations, found for each virus in total as well as in overlap with other viruses (A), or ZIKV in overlap with text mining analyses of microcephaly and Guillain-Barre syndrome (B).

predefined perspectives were overlaid onto the default map. The association of ZIKV with microcephaly was reported most frequently across all ZIKV literature and this association is visualized by the thickness of the edges (Fig. 3A). Known drug targets interacting directly with ZIKV or microcephaly were highlighted in green for potential intervention evaluation (Fig. 3B). Genes playing a role in

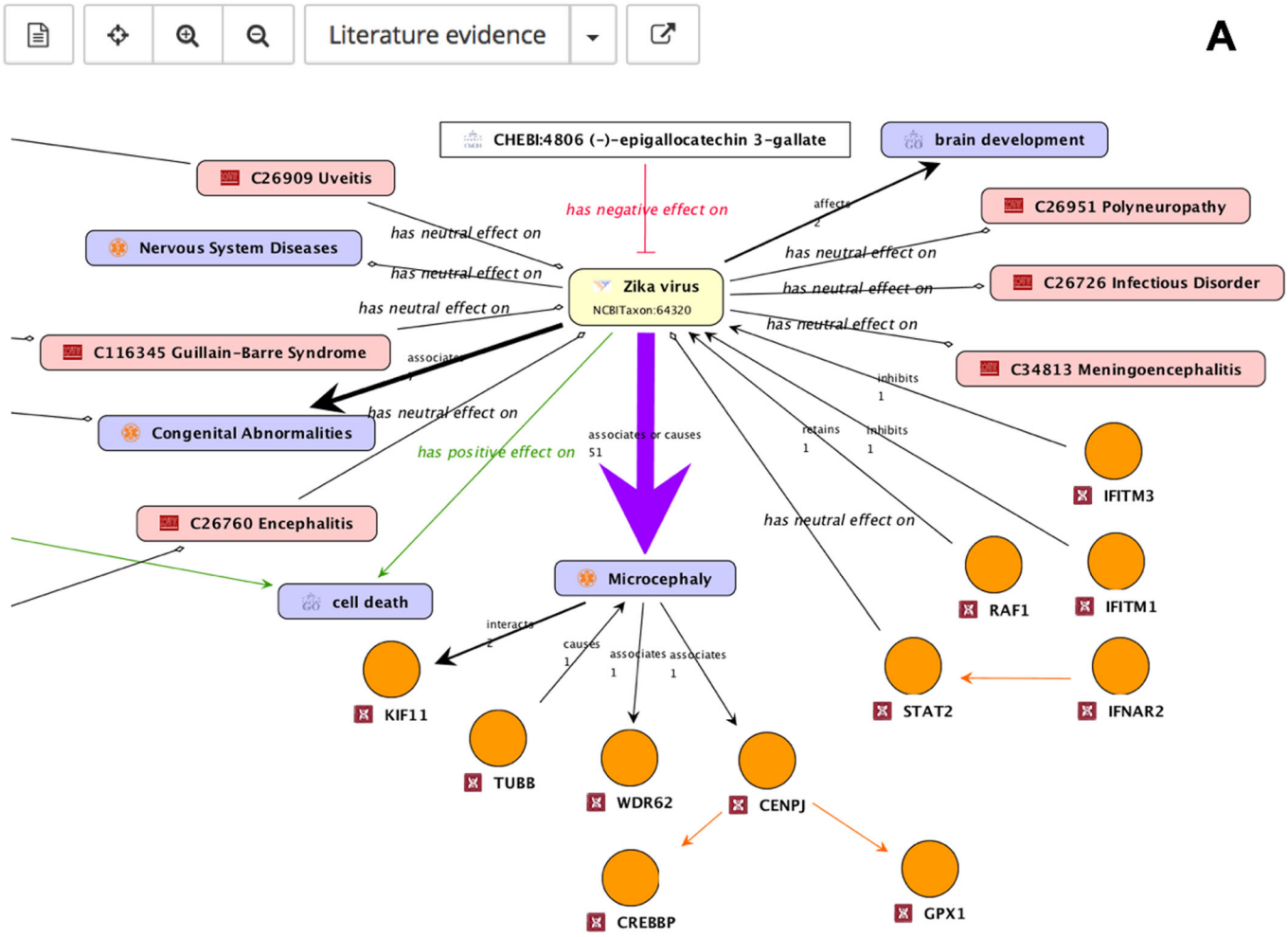

Fig. 3. Different ZIKV molecular interaction and disease map perspectives. (A) The amount of literature evidence is depicted by relation strength: the thicker the edge the bigger the amount of literature evidence, more than 20 sources of literature evidence are highlighted by a purple colored arrow. Genes are represented by circular nodes, diseases by pink rectangles, GO processes by violet rectangles, ChEBI compounds by white rectangles, and flaviviruses by yellow rectangles. (B) Genes known to be drug targets are color coded: the gene is a target for one (green) or five or more (blue) drugs or no drugs (orange). (C) Genes (hNPCs challenged with ZIKV, Tang et al., 2016) are color coded to indicate up- (red) and down regulation (blue).

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

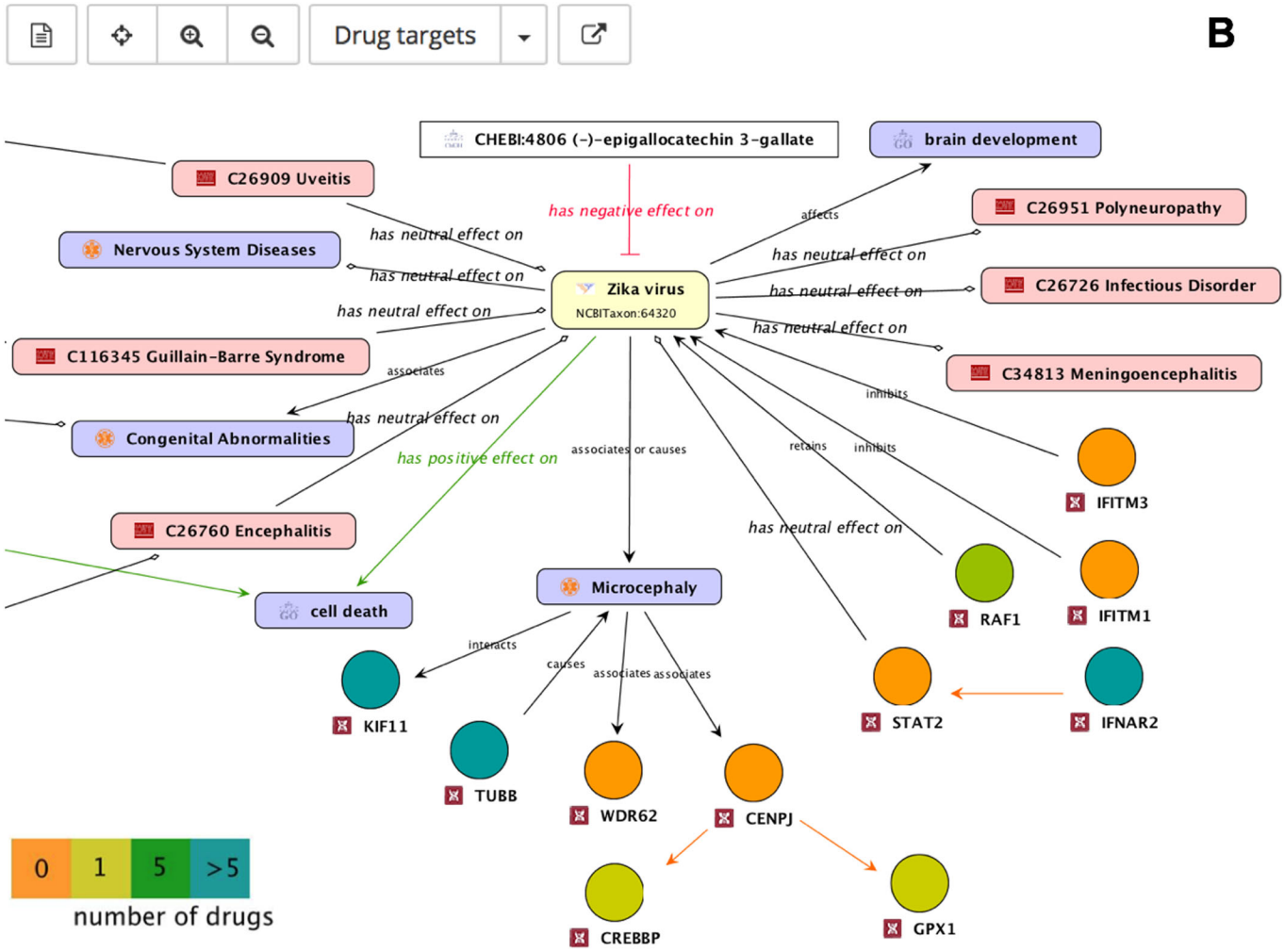

Fig. 3. continued.

ZIKV infected human neural progenitor cells (hNPCs) were also highlighted for comparative analyses of complementary experimental analyses (Fig. 3C).

Diseases associated with flaviviruses

To further explore and validate the knowledge derived from the curated text mining analyses, diseases associated with a virus of interest were queried. The results are displayed in Fig. 4. It was assumed that the disorders that were most frequently associated with a virus were the primary disorder for infection with the virus. Microcephaly and GBS, for instance, are the most frequently mentioned disorders associated with ZIKV infection. Dengue fever and hematopoietic system disorders (e.g. thrombocytopenia) are frequently listed for DENV, whereas encephalitis is the most frequently mentioned disease for WNV, JEV and TBEV. Each disease thus represents the corresponding primary manifestation of these viral infections. Encephalitis is equally frequently associated with ZIKV and DENV confirming that ZIKV is also an aetiological agent in encephalitis.

Potential inhibitors of ZIKV infection

Currently there is no approved therapy to treat ZIKV infection. Barrows et al. recently performed a screen of 774 FDA-approved drugs to identify agents that could potentially be repositioned as treatment options for ZIKV infection (Barrows et al., 2016). Of these, 24 potential inhibitors of ZIKV infection were identified and validated in human neural stem cells and primary amnion cells. In addition to their potential use for treatment, these compounds provide a resource to study ZIKV pathogenesis and can contribute to insights into the biology of ZIKV. To this end, the ZIKV molecular interaction and disease map described in Fig. 3 was extended and filtered to include these potential 'ZIKV effective drugs', which were connected to genes associated to ZIKV through PPIs (Fig. 5). After this extension ten of the identified ZIKV effective drugs were part of the new map that we then used to gain insight into potential drug mechanisms and ZIKV biology. One of the drugs, Bortezomib, is a known antiviral compound that inhibits replication of flaviviruses (Choy et al., 2015). Bortezomib is a proteasome inhibitor, suggesting that proteasome action is essential for ZIKV replication. This conclusion is in agreement with published CRISPR screen data (Zhang et al., 2016) identifying genes associated with protein degradation required for ZIKV infectivity. Interestingly, four of the predicted ZIKV effective drugs (Mefloquine, Mebendazole, Sorafenib and Dactinomycin) are associated with genes, which, through PPIs, are involved in ErbB signaling. *ErbB* is associated with the development of neurodegenerative diseases when inactivated (Bublil and Yarden, 2007). Four of these genes (*MYC*, *GSK3B*, *BRAF* and *MAP2K2*) are reported to be upregulated in a published RNAseq analysis (Tang

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

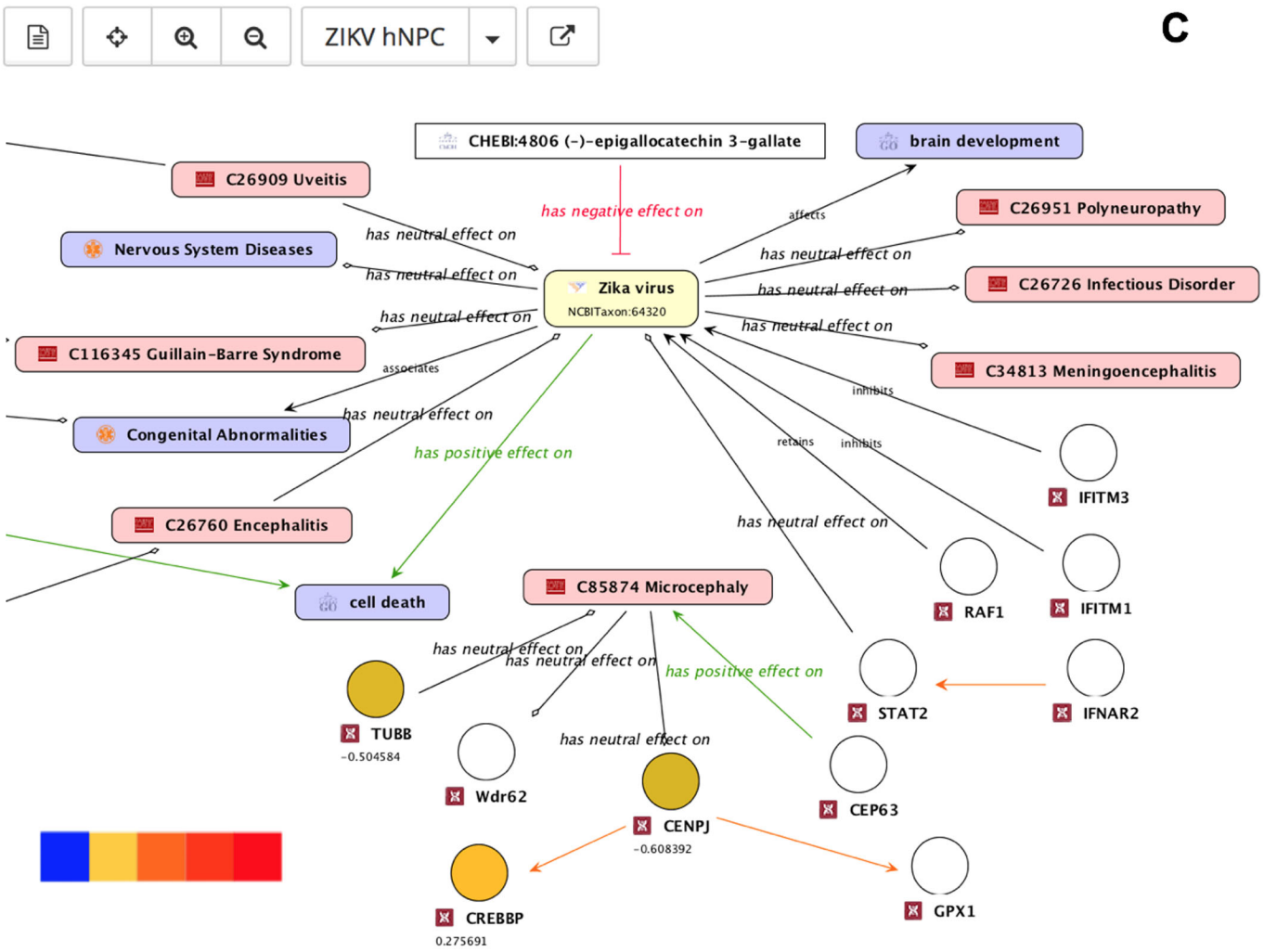

Fig. 3. continued.

et al., 2016) performed in human embryonic cortical neural progenitor cells (Fig. 6). These genes can serve as an entry point to be tested in specific assays designed to unravel molecular mechanisms between of ZIKV involvement in microcephaly.

Another predicted ZIKV effective drug is Sorafenib, a multi-target tyrosine kinase inhibitor. The ZIKV map was used to identify the effective target of Sorafenib:

1. Sorafenib interacts with four target genes, *FLT3*, *BRAF*, *VEGFR* (also known as *KDR*) and *PDGFR*. The latter two genes are known to interact with additional drugs, such as Sunitinib, Pazopanib, Dasatinib and Imatinib. These additional drugs were among those that had no effect in the ZIKV infection assay.
2. In ZIKV infection expression data, none of the genes producing protein products that interact with VEGFR and PDGFR through known PPI (orange edges) with ZIKV are differentially expressed (Fig. 7). In contrast, *BRAF* and *SOCS2*, a FLT3 interactor, were unregulated upon ZIKV infection.

Based on the observations above drawn from the ZIKV map, we hypothesize that FLT3 or BRAF are the effective targets of Sorafenib in ZIKV infection, rather than VEGFR and PDGFR. This exemplifies how molecular interaction and disease maps can be used to provide further insight into ZIKV biology.

The network analysis described above could also be used to rank drugs according to their distance to known ZIKV associated genes, such as STAT2, to suggest a metric for prioritization in screening assays (Grant et al., 2016). The discussed extended ZIKV map contains 429 target gene products that interact with FDA approved drugs. Evaluation of the distance between STAT2 and drug targets via experimentally proven PPIs revealed that targets of ZIKV effective drugs were on average more proximal to STAT2 compared to other targets.

Combining the proximity measure with additional knowledge, for example the 'FDA pregnancy' label, reduces the number of drugs to be screened from 774 to 64.

DISCUSSION

In case of an emerging epidemic, public health as well as clinical and preclinical research responses are typically hampered by a lack of structured, curated and actionable knowledge. The results of this study describe an approach to knowledge extraction and mapping that can quickly provide an overview of existing and missing information if done by a dedicated and trained group of experts. The developed workflow does not follow formal expert consensus seeking processes, such as Delphi (Dalkey and Helmer, 1963), systematic literature review processes such as Cochrane (Higgins and Green, 2011) and PRISMA (Liberati

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

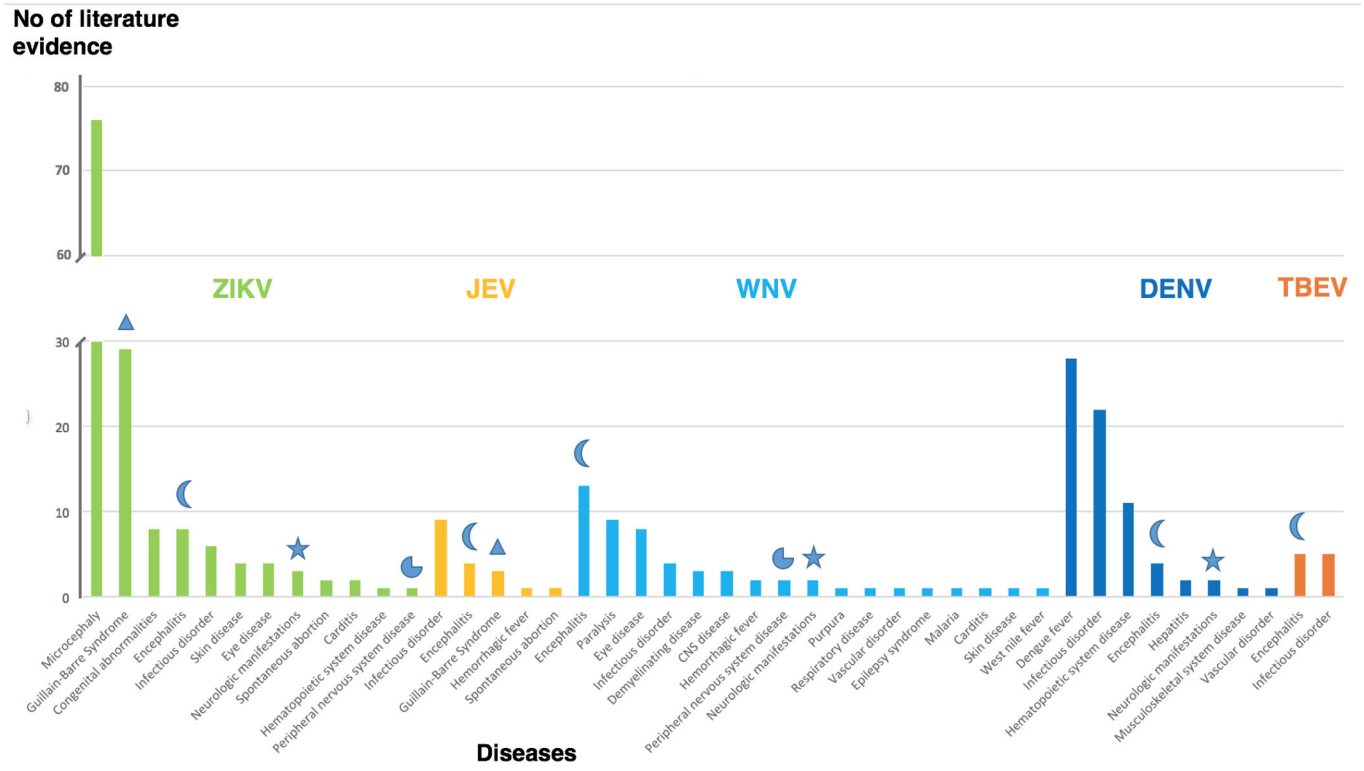

Fig. 4. Diseases associated with flaviviruses. The amount of literature evidence (y-axis) for each of the diseases (x-axis) is grouped for the five flaviviruses. Symbols highlight neurological diseases associated with more than one virus. Triangle, GBS; moon, encephalitis; pie-chart, peripheral nervous system disease; star, neurologic manifestations.

et al., 2009), or medical guideline related processes (Turner et al., 2008) as these processes are not compatible with the need for speed during emerging epidemics. Nevertheless, the workflow adopts several important aspects of good practice: it is systematic, independent and transparent, provides evidence for all integrated information and uses appropriate quality criteria. Combined with the software tools employed in the process, this pragmatic approach enabled much faster knowledge generation than more traditional methods.

The tools employed in the process need to be able to semantically integrate disparate structured resources of heterogeneous data with ease. However, much of the knowledge that represents scientific research advancements is locked within the unstructured text of classical publications, such as journal articles, newsfeeds or free-form web publications (e.g. Zika-related clinical information at http://www.ovid.com/site/zika/resources.html). The sheer volume of this published information grows constantly and exponentially and, for the most active areas of research, far exceeds the capacity of individual scientists and medical doctors to identify and read all relevant articles. Literature mining, a well-established technology to extract meaningful information from text, provides valuable assistance in structuring the massive amounts of text data and, therefore, is an indispensable tool in the process of guidance generation. Dynamic integration of objects and the relationships they participate in that are present in the literature through the use of structured resources and experimental data is a pre-requisite for analysis and distinguishes the ZIKA KB from text mining-only solutions, such as ContentMine (http://contentmine.org/) or databases dedicated to specific questions such as SncRNAs linking to disease symptoms (http://zikadb.cpqrr.fiocruz.br/zika/).

Beyond our initial analysis presented here, users can explore the ZIKA KB within a web-based user interface with the use of the online manual and the step-by-step guide (www.zikaknowledgebase.eu) to reproduce the presented results. We will collect and highly appreciate any user feedback to optimize user experience for broad usage. In contrast to alternative useful resources such as the Virus Pathogen Resource (viprbrc.org), which focus on gene and protein sequences, the ZIKA KB integrates genetic, phenotypic and drug knowledge about ZIKV to facilitate the generation of hypotheses, define research priorities and enable better understanding of viral pathogenesis. In addition to interactive exploration, a corresponding ranking of connections in a network based on integration of multiple pieces of biological evidence can also be performed systematically and on a large scale, for example, by applying the ChainRank method (Tényi et al., 2016), which we plan to integrate in the future.

Based on the text mining analyses performed here, disease profiles for the set of five neurotropic flaviviruses were confirmed. Common knowledge was retrieved along with underlying literature evidence and rare manifestations, such as encephalitis, associating with ZIKV and DENV.

Using a molecular interaction and disease map based on ZIKV, microcephaly, and GBS text mining analyses results, we showed that further exploration of the described map can provide insight into, for example, ZIKV biology, propose conclusions for research decisions, predict drug efficacy, as exemplified in the results section, as well as propose hypotheses on specific host factors and signaling pathways affected by ZIKV. The map can help to distinguish between multiple potential targets of a ZIKV effective drug. The integration of information about effectiveness of other

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

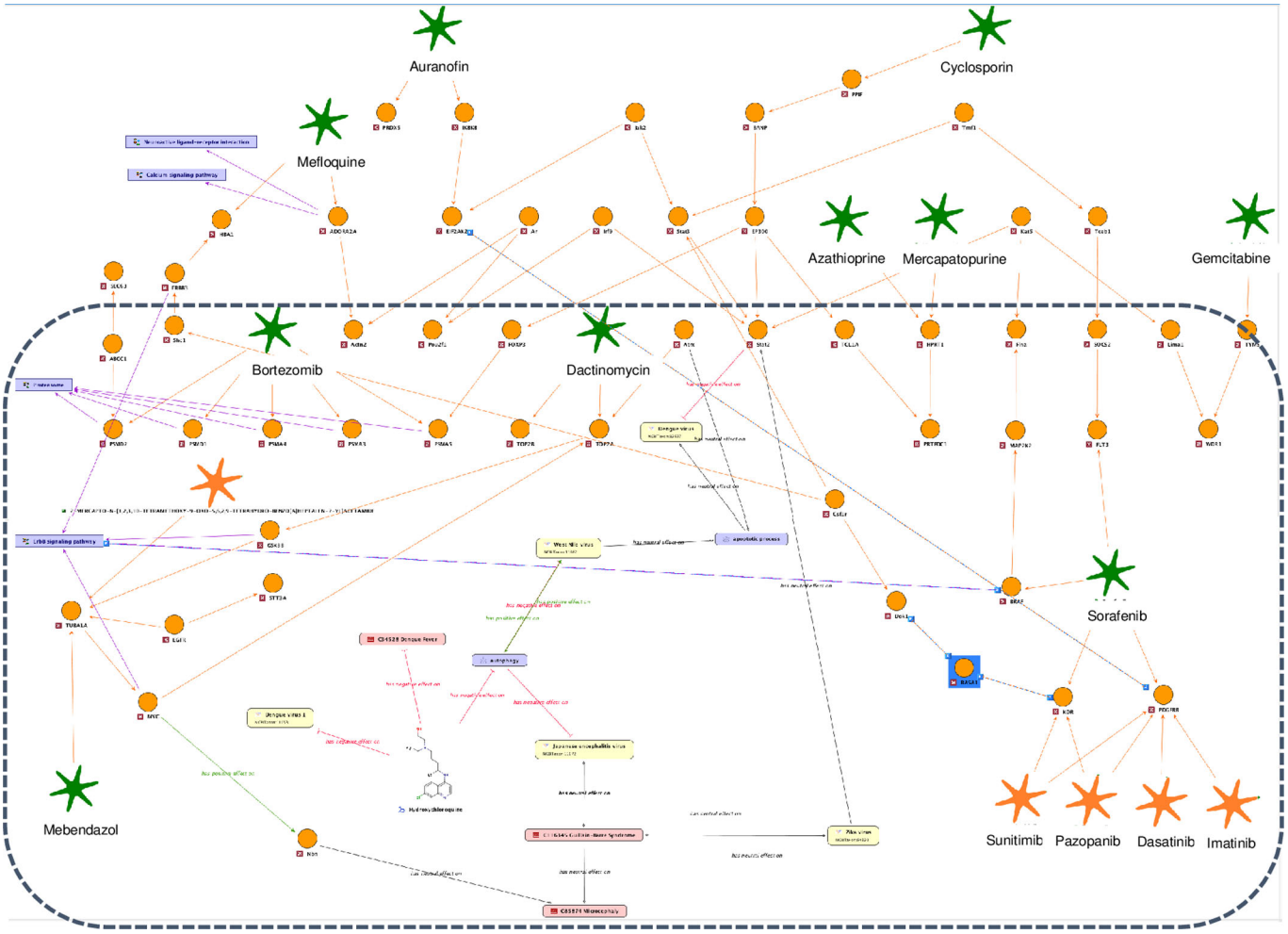

Fig. 5. ZIKV molecular interaction and disease map extended for ZIKV effective drugs. The following are the symbols used in the map. Orange circles, genes; green stars, ZIKV effective drugs; yellow rectangles, flaviviruses; pink rectangles, diseases; violet rectangles, GO processes or KEGG signaling pathways. The following edge colors are used in the map. Black edges, relationships derived from text mining; orange, protein-drug or PPIs; violet edges, participation in pathways. PPI or protein–drug relationships were obtained by applying a network search algorithm selecting the drug target of a ZIKV effective drug as start and STAT2 (contained in a direct relation with ZIKV) as the end point.

drugs as well as their target genes and the information about genes whose expression is affected during ZIKV infection indicated that Sorafenib likely acts via its target genes, *FLT3* and/or *BRAF*, but not via its alternative target genes *VEGFR* or *PDGFR*. In addition, the number of drugs to be screened was reduced from 774 to 64 by filtering potential drug candidates based on their network distance to ZIKV infection associated genes and additional phenotype relevant additional knowledge, such as contained in the 'FDA pregnancy' label.

The conclusions that can be drawn are limited by the initially low number of available publications and limited experimental data, a situation which is inherent to most emerging epidemics. Nevertheless, the work presented shows that the use of a knowledge integrating system can provide guidance for clinical and research responses, such as follow-up studies regarding the association between ZIKV, microcephaly and epilepsy, the validation of candidate drugs for ZIKV treatment, and the validation of candidate genes in specific functional assays to better understand molecular ZIKV infection mechanisms. Or to complement existing functional genomic approaches with proteomics studies, such as the integrated proteomics approach identifying cellular targets of ZIKV proteins (Scaturro et al., 2018; Scaturro et al., 2019). These studies allow additional comparative analyses between ZIKV and other flavivirus family members in terms of virulence and pathogenic traits.

Another limitation of the system is the restricted types of information which can be retrieved by text mining. While qualitative associations between genes/proteins, drugs, diseases and organisms are readily amendable to automatic approaches, it is currently almost impossible to extract, for example, clinical study designs, detailed quantitative information or complex treatment plans.

Finally, the ZIKA KB in its current stage enables exploration of the integrated information, as well as generation and curation of text-mining analysis but is not a public tool for molecular interaction and disease map generation. The functions required for these tasks will need further refinement before they can be made available in a general way.

In summary, this approach in our opinion provides a feasible way to collect and integrate existing knowledge to better understand the molecular mechanisms of an emerging pathogen. In addition our approach helps to identify gaps in knowledge and, together with the other features, guides rapid and effective responses to future epidemics. We have made the specific outcome of our approach, the ZIKA KB, publicly available as a hopefully valuable resource to the ZIKV research community.

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

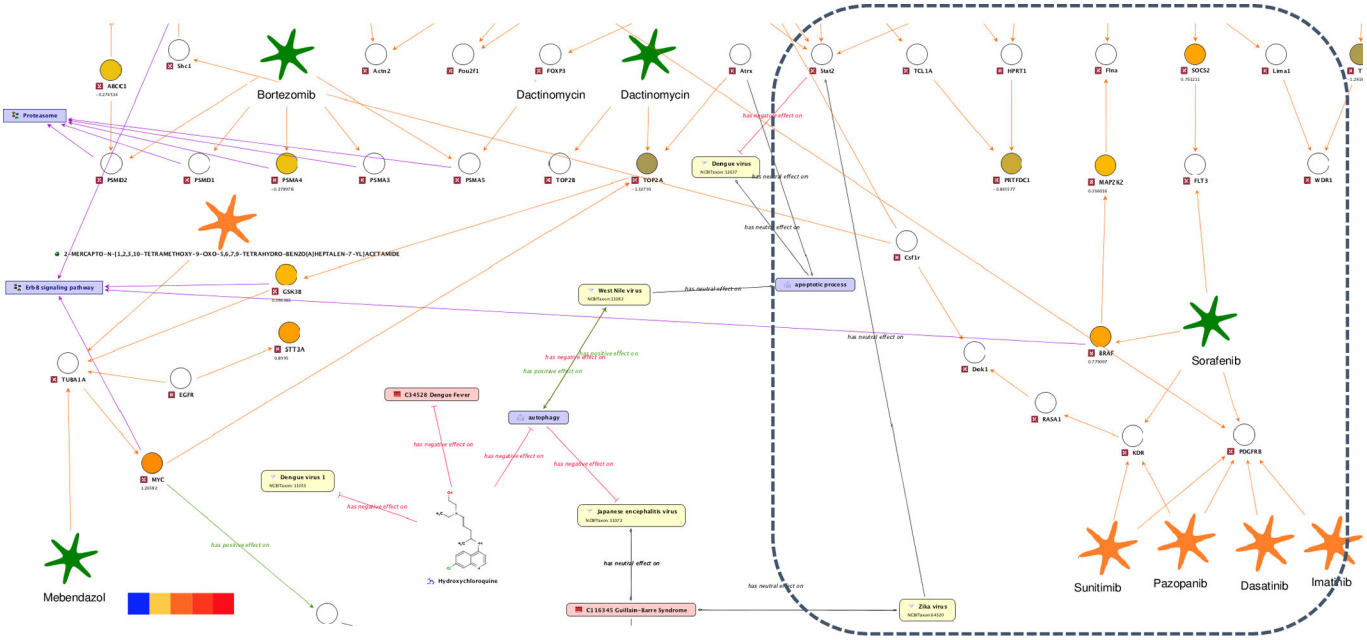

Fig. 6. ZIKV molecular interaction and disease map extended for ZIKV effective drugs (enlarged perspective). Up- and downregulated genes (hNPCs challenged with ZIKV, Tang et al., 2016) are highlighted by a color code ranging from red (upregulation) to blue (downregulation), from enlarged perspectives surrounded by dotted boxes in Fig. 5.

In the light of the current COVID-19 pandemic we are now applying the described workflow to SARS-CoV-2 and other coronaviruses and made the developed resource available (https://ailani.ai).

MATERIALS AND METHODS

Rapid response protocol

Many procedures have been published to collect knowledge from literature and experts, including systematic literature reviews (Liberati et al., 2009), clinical guideline consensus building (Dalkey and Helmer, 1963) and literature mining (Suh et al., 2010). Based on these approaches we developed a dedicated six step protocol with a focus on rapid assembly of existing knowledge (see Fig. 1 and online www.zikaknowledgebase.eu).

Team organization and process management

A multidisciplinary team of clinicians, virologists, bioinformaticians and knowledge management specialists was formed by a dedicated project enabler who contacted and invited team members, provided information on the overall aim, background and tools to use and moderated meetings. The aim was to collaboratively extract existing ZIKV related knowledge from the literature and from public databases with a focus on relationships between genes, diseases, pathogens and drugs. The extracted relationships were curated based on defined inclusion and exclusion criteria, integrated into a consistent summary and further connected with experimental data, molecular and pharmaceutical information. To enable an efficient and consolidated initial result, tasks were distributed between individuals and results were discussed and integrated in weekly online conferences. The detailed protocol for knowledge base generation was developed in this initial exercise and is presented in the Results section.

Knowledge management

For data organization, integration and development of molecular interaction and disease maps, a dedicated knowledge management tool is required. In this case, one of the PREPARE partners contributed the BioXM™ Knowledge Management Environment, a generic platform for dynamic modelling, visualization and analysis of biological and biomedical networks (Losko and Heumann, 2009). For knowledge representation we applied a semantic network approach as described previously (Maier, et al., 2011; Cano et al., 2014). Concepts required to represent domain knowledge were elucidated from domain experts and mapped to existing ontologies by knowledge engineers, essentially following the ‘concept maps’ approach (Castro et al., 2006). Based on input from clinical and virology experts, the concepts required to represent existing pathophysiological knowledge of infectious diseases were modelled with objects, such as ‘genes’, ‘strains’, ‘is expressed in’ or ‘interacts with’. For the ZIKV KB, we focused on concepts required to represent text-mining results and information from structured databases of PPI and drug–protein interactions, namely genes, diseases, pathogens and drugs. Relationships between pathogens, genes, drugs and compounds extracted by text-mining were represented by three types of relations: upregulation, downregulation and regulation (for further details see Fig. S1). Where possible, each concept was referenced to unique entries from reference databases or ontologies such as ChEBI (Degtyarenko et al., 2008) for chemicals and ICD10 (‘The international conference for the tenth revision of the International Classification of Diseases. Strengthening of Epidemiological and Statistical Services Unit. World Health Organization, Geneva’, 1990) for diseases. The defined semantic concepts become directly available in a natural-language-like query and reporting language. This language can be used to address specific questions and to summarize and visualize available knowledge. For example, the query ‘Retrieve all drugs which are interacting with human proteins which are expressed from genes affected by Zika virus infection’ retrieves the number of drugs that interact with a protein of interest and generates a visualization that applies a color coding to genes that indicates to the number of associated drugs.

Text mining

The integrated text mining tool uses syntactic text parsing and dictionary-based named-entity recognition to extract semantically typed associations (such as ‘inhibits’, ‘activates’) between the defined semantic concepts (such as ‘gene’, ‘strain’) (Losko et al., 2006). The initial task creates a defined text corpus, which includes uploaded relevant full text articles if applicable. In principle the textual materials for mining can be derived from PubMed abstracts, text from the WHO or other news feeds or any document in the portable document format (PDF), Microsoft Word or American Standard Code for Information Interchange (ASCII) formats. For the case study described here we used all ZIKV PubMed abstracts and publicly openly

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

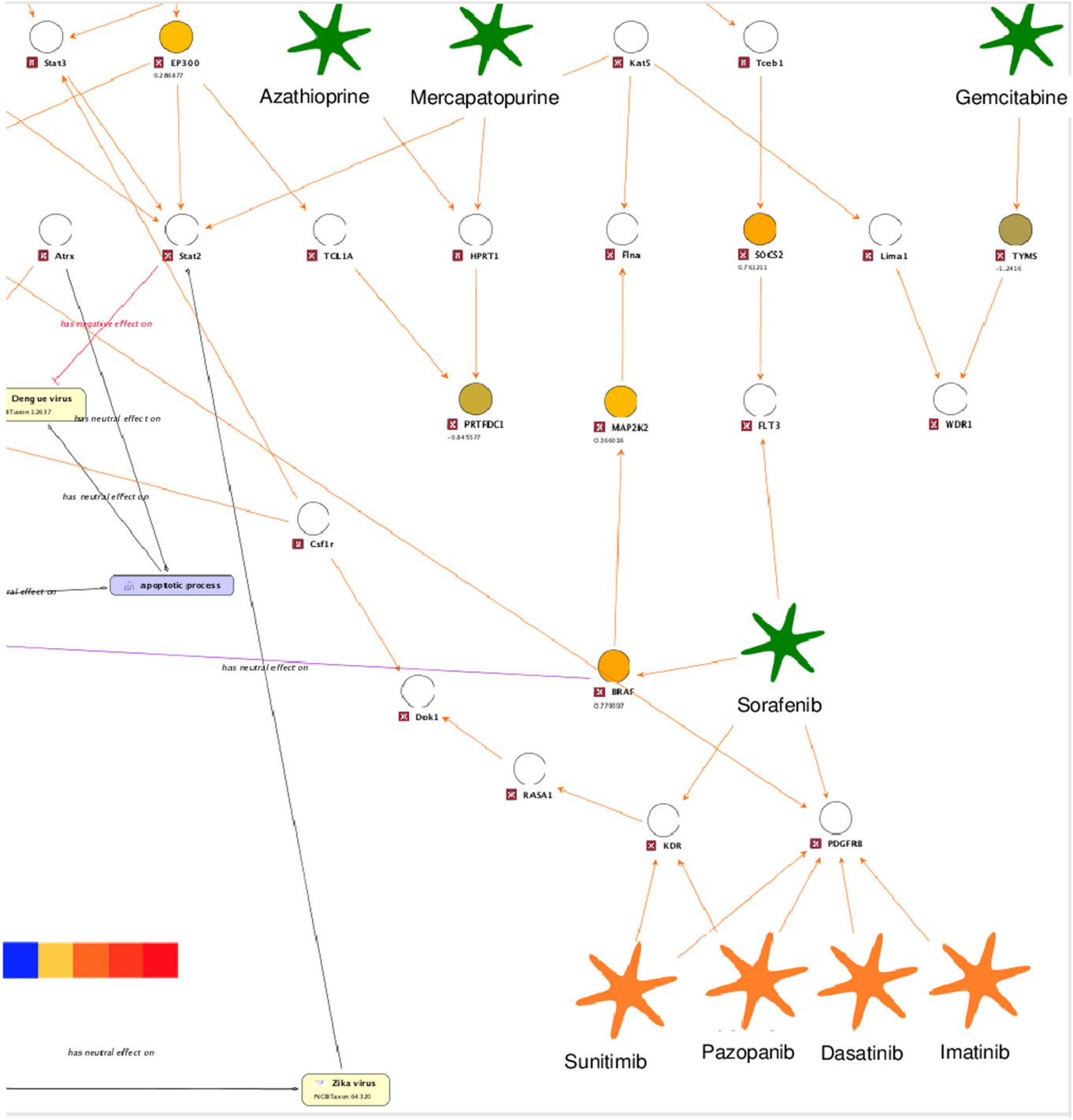

Fig. 7. ZIKV molecular interaction and disease map extended for ZIKV effective drugs (enlarged perspective). Up- and downregulated genes (hNPCs challenged with ZIKV, Tang et al., 2016) are highlighted by a color code ranging from red (upregulation) to blue (downregulation) from enlarged perspectives surrounded by dotted boxes in Fig. 6.

available full text articles. From these sources, relationships between genes, diseases, pathogens and drugs were extracted. The extracted associations consist of a subject, an object and the linking predicate and are enriched by their supportive evidence and additional metadata. For example, one such relationship is 'Zika virus (subject) causes (predicate) microcephaly (object)' (Fig. 8). Genes, diseases, pathogens and drugs, can be used as subjects and as objects and the sum of all extracted associations form an initial knowledge network. Genes, diseases, pathogens and drugs were defined by dictionaries that we curated from public sources as described below. Each dictionary consists of a well-defined set of ontologies (including synonyms) or reference databases tailored to the research question of interest. For instance, the disease dictionary consists of disease ontology entries (Schriml et al., 2012) and relevant branches of the NCI Thesaurus (de Coronado et al., 2004), the organism dictionary of NCBI taxonomy entries (Wheeler et al., 2000), the compound dictionary of ChEBI entries (Degtyarenko et al., 2008), as well as of KEGG (Goto et al., 1998) and NCI Thesaurus compounds. The gene dictionary is based on genes derived from human and flavivirus genomes. Predicates are derived from a

Biology Open (2020) 9, bio053934. doi:10.1242/bio.053934

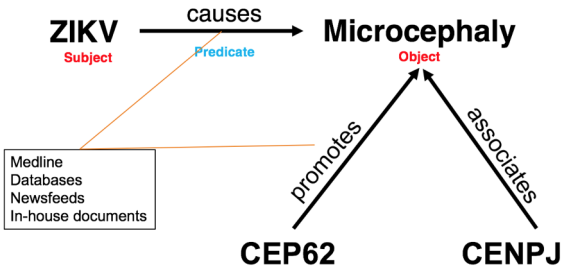

Fig. 8. Predicated text mining relationship. A text mining relationship consists of a subject (ZIKV), an object (microcephaly) and a linking predicate (causes). Subject and object are defined by dictionaries consisting of ontologies or reference databases, whereas predicates are derived from a fixed set of verbs with assertions from integrated sources, such as Medline. The term 'microcephaly' is a reusable scientific concept that participates not just in one 'Subject-Predicate-Object' construct detected, but in all such constructs detected that mention ' microcephaly'. Supplementary information is associated with the 'microcephaly' object, including, for example, information from the disease ontology, and other integrated resources, such as Gene-Disease-Association data (DisGeNET). This expandable set of relationships forms a large network of knowledge that enables new knowledge to be inferred by 'reasoning' based on the logic encoded in those relationships.

set of verbs, which can be modified. These predicates describe mainly molecular interactions but can also indicate causal associations between proteins or compounds and diseases (for instance 'activates', 'restricts', 'targets'). To optimize recall and specificity of the mining, we extended the dictionaries for viral names, acronyms and interaction predicates as well as defined a black-list of acronyms causing mostly false positives.

Finally, the extracted relationships can be curated to manually optimize quality and information content. A curation user interface was implemented to enable the expert team to support or refute the automatically generated relationships. At least two independent researchers (the '4-eye review mode') manually evaluated the evidence for every extracted relationship. In the case that the evaluations from the two researchers conflicted, the conflicts were either resolved during the weekly online conferences or were excluded, as our goal was to maximize specificity (correctness) rather than sensitivity (completeness) of the integrated information.

In addition, experts could expand the network with any relevant supporting evidence from other integrated sources, such as public or proprietary databases and experimental data.

Semantic mapping of experimental results, public data sources and ontologies

Semantic mapping describes the process of identifying and linking concepts that are shared between two information sources. We integrated the databases listed in Table 1 using existing concepts such as genes, pathogens or diseases which were identified by ontological descriptors. Semantically identical objects are mapped to descriptive data from literature and databases to allow informed and efficient querying of the overall collected information (e.g. 'Dengue disease' is mapped to the following synonyms: 'Breakbone fever', 'Dengue disorder', 'Dengue fever' and 'Dengue') (Maier, et al., 2011). To this end, mapping scripts are created to resolve a given input data format and match the provided entity identifiers or ontology terms. Experimental data from key publications are mapped by the same approach. While these data are henceforth available for search and reporting they are not yet displayed as part of any specific molecular interaction and disease map.

Querying and visualization of integrated information in tables, networks and disease maps

To help the expert team establish a specific molecular interaction and disease map we defined a number of queries to explore the collective knowledge. These queries were used, for example, to find diseases and genes associated with a virus of interest to find diseases associated with genes prioritized according to experimental evidence.

Based on these queries we developed a streamlined, wizard-based user interface to create disease maps by selecting the relevant relationships from the curated text mining from query results (Fig. S2). This basic network was further extended with interaction data (e.g. PPI and protein-drug interaction) by applying a network search algorithm based on genes extracted from text mining relationships. Finally, we defined queries to overlay additional information, such as literature evidence, experimental data, drug targets or host factors to obtain different perspectives of the same underlying molecular interaction or disease map.

Deployment of an open access, web-based user interface

To make the results of our internal test case generally available and to support ZIKV research, we provide and maintain a regularly updated ZIKA KB at the following URL, www.zikaknowledgebase.eu. As we continue to extend this resource user registration for access will be implemented to ensure the knowledge base is used for research only.

Acknowledgements

The authors wish to thank their colleagues for discussion and feedback, and Sheridon Sauer and Shannon Frances for critical review of the article.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: A.B., D.M., M.D.d.J.; Methodology: A.B.; Validation: J.P., C.C.; Investigation: A.B.; Data curation: T.S., X.Y., A.G., P.L.F.; Writing - original draft: A.B., O.P., D.M., M.K., M.D.d.J.; Writing - review & editing: A.B., J.P., P.L.F., O.P., C.A., D.M., M.K., M.D.d.J.; Visualization: A.B.; Supervision: D.M.; Project administration: A.B., M.D.d.J.; Funding acquisition: C.A., D.M., M.D.d.J.

Funding

This work was supported by the European Commission's Seventh Framework Research Programme project PREPARE [FP7-Health n°602525] and ZIKALLIANCE [MK, H2020; No 734548].

Supplementary information

Supplementary information available online at
https://bio.biologists.org/lookup/doi/10.1242/bio.053934.supplemental

References

Anderson, K. B., Thomas, S. J. and Endy, T. P. (2016). The Emergence of Zika virus: a narrative review. *Ann. Intern. Med.* **165**, 175-183. doi:10.7326/M16-0617

Barrows, N. J., Campos, R. K., Powell, S. T., Prasanth, K. R., Schott-Lerner, G., Soto-Acosta, R., Galarza-Muñoz, G., Mcgrath, E. L., Urrabaz-Garza, R., Gao, J. et al. (2016). A screen of FDA-approved drugs for inhibitors of Zika virus infection. *Cell Host Microbe* **20**, 259-270. doi:10.1016/j.chom.2016.07.004

Bublil, E. M. and Yarden, Y. (2007). The EGF receptor family: spearheading a merger of signaling and therapeutics. *Curr. Opin. Cell Biol.* **19**, 124-134. doi:10.1016/j.ceb.2007.02.008

Cano, I., Tényi, Á., Schueller, C., Wolff, M., Migueláñez, M. M. H., Gomez-Cabrero, D., Antczak, P., Roca, J., Cascante, M., Falciani, F. et al. (2014). The COPD Knowledge Base: enabling data analysis and computational simulation in translational COPD research. *J. Transl. Med.* **12**, S6. doi:10.1186/1479-5876-12-S2-S6

Carod-Artal, F. J. (2016). Epidemiología y complicaciones neurológicas de la infección por el virus del Zika: un nuevo virus neurotropo emergente. *Revista de Neurología* **62**, 317-328. doi:10.33588/rn.6207.2016152

Castro, A. G., Rocca-Serra, P., Stevens, R., Taylor, C., Nashar, K., Ragan, M. A. and Sansone, S.-A. (2006). The use of concept maps during knowledge elicitation in ontology development processes–the nutrigenomics use case. *BMC Bioinformatics* **7**, 267. doi:10.1186/1471-2105-7-267

Choy, M. M., Zhang, S. L., Costa, V. V., Tan, H. C., Horrevorts, S. and Ooi, E. E. (2015). Proteasome inhibition suppresses dengue virus egress in antibody dependent infection. *PLoS Negl. Trop. Dis.* **9**, e0004058. doi:10.1371/journal.pntd.0004058

Dalkey, N. and Helmer, O. (1963). An experimental application of the DELPHI method to the use of experts. *Manage. Sci.* **9**, 458-467. doi:10.1287/mnsc.9.3.458

de Coronado, S., Haber, M. W., Sioutos, N., Tuttle, M. S. and Wright, L. W. (2004). NCI Thesaurus: using science-based terminology to integrate cancer research results. *Stud. Health Technol. Inform.* **107**, 33-37.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008). ChEBI: a

database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344-D350. doi:10.1093/nar/gkm791

Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M. A. et al. (2014). Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol. Neurobiol.* **49**, 88-102. doi:10.1007/s12035-013-8489-4

Goto, S., Nishioka, T. and Kanehisa, M. (1998). LIGAND: chemical database for enzyme reactions. *Bioinformatics (Oxford, England)* **14**, 591-599. doi:10.1093/bioinformatics/14.7.591

Grant, A., Ponia, S. S., Tripathi, S., Balasubramaniam, V., Miorin, L., Sourisseau, M., Schwarz, M. C., Sánchez-Seco, M. P., Evans, M. J., Best, S. M. et al. (2016). Zika virus targets human STAT2 to inhibit Type I interferon signaling. *Cell Host Microbe* **19**, 882-890. doi:10.1016/j.chom.2016.05.009

Guo, X., Xu, Y., Bian, G., Pike, A. D., Xie, Y. and Xi, Z. (2010). Response of the mosquito protein interaction network to dengue infection. *BMC Genomics* **11**, 380. doi:10.1186/1471-2164-11-380

Higgins, J. and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. 5.1.0. (The Cochrane Collaboration). Available at: http://handbook.cochrane.org.

Kumari, B., Jain, P., Das, S., Ghosal, S., Hazra, B., Trivedi, A. C., Basu, A., Chakrabarti, J., Vrati, S. and Banerjee, A. (2016). Dynamic changes in global microRNAome and transcriptome reveal complex miRNA-mRNA regulated host response to Japanese Encephalitis Virus in microglial cells. *Sci. Rep.* **6**, 20263. doi:10.1038/srep20263

Le Breton, M., Meyniel-Schicklin, L., Deloire, A., Coutard, B., Canard, B., De Lamballerie, X., Andre, P., Rabourdin-Combe, C., Lotteau, V. and Davoust, N. (2011). Flavivirus NS3 and NS5 proteins interaction network: a high-throughput yeast two-hybrid screen. *BMC Microbiol.* **11**, 234. doi:10.1186/1471-2180-11-234

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J. and Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med.* **6**, e1000100. doi:10.1371/journal.pmed.1000100

Losko, S. and Heumann, K. (2009). Semantic data integration and knowledge management to represent biological network associations. *Method. Mol. Biol.* **563**, 241-258. doi:10.1007/978-1-60761-175-2_13

Losko, S., Wenger, K., Kalus, W., Ramge, A., Wiehler, J. and Heumann, K. (2006). Knowledge Networks of Biological and Medical Data: An Exhaustive and Flexible Solution to Model Life Science Domains. In *Data Integration in the Life Sciences*, pp. 232-239. Springer Berlin / Heidelberg (Lecture Notes in Computer Science).

Maier, D., Kalus, W., Wolff, M., Kalko, S. G., Roca, J., De Mas, I. M., Turan, N., Cascante, M., Falciani, F., Hernandez, M. et al. (2011). Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Syst. Biol.* **5**, 38. doi:10.1186/1752-0509-5-38

Matsuoka, Y., Matsumae, H., Katoh, M., Eisfeld, A. J., Neumann, G., Hase, T., Ghosh, S., Shoemaker, J. E., Lopes, T. J. S., Watanabe, T. et al. (2013). A comprehensive map of the influenza A virus replication cycle. *BMC Syst. Biol.* **7**, 97. doi:10.1186/1752-0509-7-97

Moore, E. F. (1959). The shortest path through a maze. In Proceedings of the International Symposium on the Theory of Switching. Harvard University Press, pp. 285-292.

Murray, J. S. (2017). Understanding Zika virus. *JSPN* **22**. doi:10.1111/jspn.12164

Nim, H. T., Furtado, M. B., Costa, M. W., Kitano, H., Rosenthal, N. A. and Boyd, S. E. (2015). CARFMAP: a curated pathway map of cardiac fibroblasts. *PLoS ONE* **10**, e0143274. doi:10.1371/journal.pone.0143274

Nowakowski, T. J., Pollen, A. A., Di Lullo, E., Sandoval-Espinosa, C., Bershteyn, M. and Kriegstein, A. R. (2016). Expression analysis highlights AXL as a candidate Zika virus entry receptor in neural stem cells. *Cell Stem Cell* **18**, 591-596. doi:10.1016/j.stem.2016.03.012

Pardy, R. D. and Richer, M. J. (2019). Zika Virus pathogenesis: from early case reports to epidemics. *Viruses* **11**, 886. doi:10.3390/v11100886

Pierson, T. C. and Diamond, M. S. (2018). The emergence of Zika virus and its new clinical syndromes. *Nature* **560**, 573-581. doi:10.1038/s41586-018-0446-y

Scaturro, P., Stukalov, A., Haas, D. A., Cortese, M., Draganova, K., Płaszczyca, A., Bartenschlager, R., Götz, M. and Pichlmair, A. (2018). An orthogonal proteomic survey uncovers novel Zika virus host factors. *Nature* **561**, 253-257. doi:10.1038/s41586-018-0484-5

Scaturro, P., Kastner, A. L. and Pichlmair, A. (2019). Chasing intracellular Zika virus using proteomics. *Viruses* **11**, 878. doi:10.3390/v11090878

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G. and Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* **40**, D940-D946. doi:10.1093/nar/gkr972

Sharma, V., Sharma, M., Dhull, D., Sharma, Y., Kaushik, S. and Kaushik, S. (2019). Zika virus: an emerging challenge to public health worldwide. *Can. J. Microbiol.* **66**, 87-98. doi:10.1139/cjm-2019-0331

Suh, K. S., Park, S. W., Castro, A., Patel, H., Blake, P., Liang, M. and Goy, A. (2010). Ovarian cancer biomarkers for molecular biosensors and translational medicine. *Expert Rev. Mol. Diagn.* **10**, 1069-1083. doi:10.1586/erm.10.87

Tang, H., Hammack, C., Ogden, S. C., Wen, Z., Qian, X., Li, Y., Yao, B., Shin, J., Zhang, F., Lee, E. M. et al. (2016a). Zika virus infects human cortical neural progenitors and attenuates their growth. *Cell Stem Cell* **18**, 587-590. doi:10.1016/j.stem.2016.02.016

Tényi, Á., De Atauri, P., Gomez-Cabrero, D., Cano, I., Clarke, K., Falciani, F., Cascante, M., Roca, J. and Maier, D. (2016). ChainRank, a chain prioritisation method for contextualisation of biological networks. *BMC Bioinformatics* **17**, 17. doi:10.1186/s12859-015-0864-x

'The international conference for the tenth revision of the International Classification of Diseases. Strengthening of Epidemiological and Statistical Services Unit. World Health Organization, Geneva' (1990). *World Health Statistics Quarterly. Rapport Trimestriel de Statistiques Sanitaires Mondiales* **43**, 204-245.

Turner, T., Misso, M., Harris, C. and Green, S. (2008). Development of evidence-based clinical practice guidelines (CPGs): comparing approaches. *Implementation Science* **3**, 45. doi:10.1186/1748-5908-3-45

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A. and Rapp, B. A. (2000). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **28**, 10-14. doi:10.1093/nar/28.1.10

Zhang, R., Miner, J. J., Gorman, M. J., Rausch, K., Ramage, H., White, J. P., Zuiani, A., Zhang, P., Fernandez, E., Zhang, Q. et al. (2016). A CRISPR screen defines a signal peptide processing pathway required by flaviviruses. *Nature* **535**, 164-168. doi:10.1038/nature18625