DISCOVERING LESSER KNOWN MOLECULAR PLAYERS AND MECHANISTIC PATTERNS IN ALZHEIMER'S DISEASE USING AN INTEGRATIVE DISEASE MODELLING APPROACH

Von der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

angenommene

Kumulative Dissertation

zur Erlangung des Doktorgrades

Doctor Rerum Naturalium (Dr. rer. nat.)

in Fachgebiet

Computational Life Sciences

universitätbonn

vorgelegt von

Shweta Bagewadi Kawalia

aus Belagavi, Indien

Bonn, 2018

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter:
1. Prof. Dr. rer. nat. Martin Hofmann-Apitius
2. Prof. Dr. rer. nat. Andreas Weber

Fachnahes Mitglied : Prof. Dr. rer. nat. Stefan Wrobel
Fachfremdes Mitglied : Prof. Dr. rer. nat. Diana Imhof

Tag der Promotions : 4. September 2018
Erscheinungsjahr : 2018

"Science is all about: building and communicating knowledge. You may have a beautiful experiment in your lab notebook, or in your head, but it isn't science until you make it available to others so that they can build on it."

— Nobel Laureate Oliver Smithie

Abstract

Convergence of exponentially advancing technologies is driving medical research with life changing discoveries. On the contrary, repeated failures of high-profile drugs to battle Alzheimer's disease (AD) has made it one of the least successful therapeutic area. This failure pattern has provoked researchers to grapple with their beliefs about Alzheimer's aetiology. Thus, growing realisation that Amyloid-β and tau are not 'the' but rather 'one of the' factors necessitates the reassessment of pre-existing data to add new perspectives. To enable a holistic view of the disease, integrative modelling approaches are emerging as a powerful technique. Combining data at different scales and modes could considerably increase the predictive power of the integrative model by filling biological knowledge gaps. However, the reliability of the derived hypotheses largely depends on the completeness, quality, consistency, and context-specificity of the data. Thus, there is a need for agile methods and approaches that efficiently interrogate and utilise existing public data.

This thesis presents the development of novel approaches and methods that address intrinsic issues of data integration and analysis in AD research. It aims to prioritise lesser-known AD candidates using highly curated and precise knowledge derived from integrated data. Here much of the emphasis is put on quality, reliability, and context-specificity. This thesis work showcases the benefit of integrating well-curated and disease-specific heterogeneous data in a semantic web-based framework for mining actionable knowledge. Furthermore, it introduces to the challenges encountered while harvesting information from literature and transcriptomic resources. State-of-the-art text-mining methodology is developed to extract miRNAs and its regulatory role in diseases and genes from the biomedical literature. To enable meta-analysis of biologically related transcriptomic data, a highly-curated metadata database has been developed, which explicates annotations specific to human and animal models. Finally, to corroborate common mechanistic patterns — embedded with novel candidates — across large-scale AD transcriptomic data, a new approach to generate gene regulatory networks has been developed.

The work presented here has demonstrated its capability in identifying testable mechanistic hypotheses containing previously unknown or emerging knowledge from public data in two major publicly funded projects for Alzheimer's, Parkinson's and Epilepsy diseases.

Zusammenfassung

Die Konvergenz exponentiell fortschreitender Technologien treibt die medizinische Forschung mit lebensverändernden Entdeckungen voran. Andererseits das wiederholte Versagen von hochkarätigen Medikamenten gegen die Alzheimer-Krankheit hat sie zu einem der am wenigsten erfolgreichen Therapiegebiet gemacht. Dieses Versagensmuster hat Forscher dazu veranlasst, sich mit ihren Überzeugungen über die Alzheimer-Ätiologie auseinanderzusetzen. Die wachsende Erkenntnis, dass Aβ und tau nicht die Faktoren, sondern einer der Faktoren sind, macht eine Neubewertung bereits vorhandener Daten erforderlich, um neue Perspektiven zu eröffnen. Um eine ganzheitliche Betrachtung der Krankheit zu ermöglichen, entwickeln sich integrative Modellierungsansätze zu einer wirkungsvollen Methode. Die Kombination von Daten aus verschiedenen Ebenen und Modi wird die Vorhersagekraft des integrativen Modells erheblich erhöhen, indem biologische Wissenslücken geschlossen werden. Die Zuverlässigkeit der abgeleiteten Hypothesen hängt jedoch in hohem Maße von der Vollständigkeit, Qualität, Konsistenz und Kontextspezifität der Daten ab. Daher bedarf es agiler Methoden und Ansätze, die öffentlich verfügbare Datensätze effektiv und effizient abfragen und nutzen.

Diese Arbeit stellt die Entwicklung neuer Ansätze und Methoden vor, die sich mit wesentlichen Fragen der Datenintegration und -analyse in der Alzheimer-Forschung befassen. Sie zielt auf die Priorisierung von weniger bekannten Alzheimer-Kandidaten mit Hilfe von hochgradig kuratiertem und präzisem Wissen, das aus integrierten Daten gewonnen wird. Dabei wird der Schwerpunkt auf Qualität, Zuverlässigkeit und Kontextspezifität gelegt. Diese Arbeit zeigt den Nutzen der Integration gut kuratierter und krankheitsspezifischer heterogener Daten in ein semantisches web-basiertes Framework für die Gewinnung von handlungsfähigem Wissen. Darüber hinaus werden die Herausforderungen bei der Extraktion von Informationen aus Literatur und transkriptomischen Ressourcen vorgestellt. Modernste Text-Mining-Methodik werden entwickelt, um miRNAs und ihre regulatorische Rolle bei Krankheiten und Genen aus der biomedizinischen Literatur zu extrahieren. Um die Metaanalyse von biologisch verwandten transkriptomischen Daten zu ermöglichen, wird eine hochgradig kuratierte Metadaten-Datenbank entwickelt, die Annotationen spezifisch für menschliche und tierische Modelle bereitstellt. Schließlich wird ein neuer Ansatz zur Generierung von Genregulationsnetzwerken entwickelt, um gemeinsame mechanistische Zusammenhänge

nachzuweisen, die mit neuartigen Kandidaten in umfangreichen transkriptomischen Alzheimer-Daten eingebettet sind.

Die hier vorgestellte Arbeit hat gezeigt, dass sie in der Lage ist, testbare mechanistische Hypothesen zu identifizieren, die bisher unbekannte oder neu entstehende Erkenntnisse aus bestehenden öffentlich verfügbare Daten enthalten. Diese Daten stammen aus zwei öffentlich finanziert Projekten, die sich mit Alzheimer-, Parkinson- und Epilepsie-Erkrankung beschäftigen

Acknowledgement

What a PhD thesis needs?

A very strong motivation to pursue your research dream when your close friends are moving to industry jobs; of course, lured by better financial conditions. "*Two Gurus*", Dr. Martin Hofmann-Apitius and Dr. Philipp Senger, to inspire, support, encourage, guide, and forge new opportunities to pave the right professional path. I thank them for making me what I am today. A co-referent, Prof. Dr. Andreas Weber, who reviews and believes in the potential of this thesis. Two thesis reviewers, Prof. Dr. Stefan Wrobel and Prof. Dr. Diana Imhof, who also believe in the value of this work and complete the thesis jury.

It needs a professional environment and a work place: Fraunhofer SCAI. It needs administrative angels like Meike Knieps, Alina Enns, and Heike Gross without whom I would be lost in all the paper work. A very patient project manager, Stephan Springstubbe, who taught me how to handle tense and stressful situations. It needs Michael Krapp and Sabrina Diaz to communicate my research work to the world through their intellectual marketing strategies. It needs Horst Schwichtenberg, Jan Peterson, and Stefan Bach to fix computer and hardware related problems so that I can focus on my research. Highly esteemed colleagues: Dr. Juliane Fluck. Dr. Roman Klinger, Theo Mevissen, Sumit Madan, Christian Ebeling, Erfan Younesi, and co-PhD students, who influenced my scientific knowledge. A bilingual colleague, Sumit Madan to translate the English abstract to German language. A couple of master and bachelor students, who helped me achieve the project deadlines; especially a talented student like Tamara Raschka.

It needs funding from Neuroallianz's D10 and IMI's AETIONOMY projects to support novel research work and get me through my financial expenses. Many anonymous journal reviewers and editors, whose critical comments enriched the worthiness of my submitted manuscripts. It needs project partners, for constructive and insightful discussions strengthening the scientific work. It needs public data contributed by wet-lab researchers worldwide.

It needs a best friend and a husband, Dr. Amit Kawalia, as the pillar of support to succeed through personal and professional challenges. His love, patience, and critics have helped me overcome many obstacles in life. My son, Avyaan, who constantly keeps me busy and reminds me with his actions that *"mistakes can open doors to new opportunities"*. Of course, he also likes to contribute to this work by randomly typing *"rlopökä-,'<dgthzvbjuikmloöåcdfvgbhjnm"*, when my laptop is left unattended. A mother, Surekha Bagewadi, who is always there to support you no matter what. Her help and cooperation to look after Avyaan enabled me to write my thesis during my parental leave. A wifi physician, my brother-in-law, Dr. Gaurav Kawalia, whose constant advise with medications backed my physical health after the delivery, giving me the strength for writing this thesis. A father, Late Vishwanath Bagewadi, who always believed in my potential and capability. A best friend for life, Thileepan Sekaran, his wife Shanthini, and his parents for supporting me during the final phase of writing. And a loving family and friends who are the great force of support.

And at last, to all the people who taught me the real meaning of the below quote:

"There's always something good that comes out of every experience. Good times become good memories. Bad times become good lessons. You can never lose, you only grow from life".

— *Ryan Ferreras*

And most definitely, not to forget it needs lot of strong coffee and good Bollywood music.

Statutory Declaration

I hereby declare that this dissertation has been solely composed by me and has not been submitted, in whole or in part, to any other faculty or university. I confirm that the work submitted is my own, except where the work is jointly published. Appropriate credit has been given within this dissertation where reference has been made to the work of others.

Declaration of contribution as co-author

I am the first author or shared first author in all the studies included in this dissertation. I have contributed majorly or equally (if shared first author) during the design, conceptualization, implementation, analysis and manuscript writing in the studies listed in Chapter 3 and 6. I have obtained the consent of the joint authors and my supervisor, Prof. Dr. Hofmann-Apitius, to use these shared publications in my thesis.

Declaration of name change

I hereby declare that my name has been changed from *Shweta Bagewadi* to *Shweta Bagewadi Kawalia* after marriage. I confirm that I am the author (first or shared first author) of publications used in this dissertation using either of these names.

Signature: .....

Author: Shweta Bagewadi Kawalia

Table of Contents

List of Figures

List of Abbreviations

Abbreviation	Term
ACH	Amyloid Cascade Hypothesis
AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
Alzforum	Alzheimer Research Forum
APOE4	e4 allele of APOE gene
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
Aβ	Amyloid-Beta
BAMS	Brain Architecture Knowledge Management System
BC3Net	Bagging the C3NET
BD2K	Big Data to Knowledge
BEL	Biological Expression Language
BioPAX	Biological Pathway Exchange Database
BIRN	Biomedical Informatics Research Network
BMBF	Bundesministerium für Bildung und Forschung
C3NET	Conservative Causal Core Networks
Ca2+	Calcium ions
CENs	Co-Expression Networks
CLR	Context Likelihood of Relatedness
CPN	Cross-Platform Normalisation
DARPA	Defense Advanced Research Projects Agency
DB	Database
DBCLS	Database Center for Life Sciences
DE	Differentially Expressed

DL	Description Logic
DNA	Deoxyribonucleic Acid
DREAM	Dialogue for Reverse Engineering Assessment and Methods
EBI	European Bioinformatics Institute
EE	Event Extraction
EFPIA	European Federation of Pharmaceutical Industries and Associations
EMBL	European Molecular Biology Laboratory
EMIF	European Medical Information Framework
EOAD	Early-Onset Alzheimer's Disease
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
GAAIN	Global Alzheimer's Association Interactive Network
GENIE3	Gene Network Inference with Ensemble of Trees
GEO	Gene Expression Omnibus
GRDDL	Gleaning Resource Descriptions from Dialects of Languages
GRNs	Gene Regulatory Networks
GWAS	Genome-Wide Association Study
HBP	Human Brain Project
HCLSIG	Health Care and Life Sciences Interest Group
IE	Information Extraction
IMI	Innovative Medicines Initiative
INCF	International Neuroinformatics Coordination Facility
IR	Information Retrieval

IRIs	Internationalized Resource Identifiers
JPND	EU Joint Programme – Neurodegenerative Disease Research
JSON-LD	Javascript Object Notation for Linked Data
KAON2	Karlsruhe Ontologie 2
KD	Knowledge Discovery
KEGG	Kyoto Encyclopedia of Genes and Genomes
LND	Linked Neuron Data
LOAD	Late-Onset Alzheimer's Disease
LODD	Linking Open Drug Data
MCI	Mild Cognitive Impairment
MI	Mutual Information
MIAME	Minimum Information about a Microarray Experiment
MINISEQE	Minimum Information about a High-throughput Nucleotide Sequencing Experiment
MINT	Molecular Interaction Database
MiRNA	MicroRNAs
mRNA	Messenger RNA
MRNET	Maximum Relevance/Minimum Redundancy Network
MSigDB	Molecular Signatures Database
MTIs	MiRNA-Target Interactions
MySQL	My Structured Query Language
NCBI	National Center for Biotechnology Information
NDD	Neurodegenerative Diseases
NDG	Neuroscience Database Gateway
NER	Named Entity Recognition
NFTs	Neurofibrillary Tangles

NGS	Next generation Sequencing
NI	Network Inference
NIF	Neuroscience Information Framework
NIFSTD	NIF Standard (NIFSTD) ontology
NIH	National Institutes of Health
NLP	Natural Language Processing
NMDA	*N*-methyl-D-aspartate
OBO	Open Biological and Biomedical Ontology
OMIM	Online Mendelian Inheritance in Man
Open PHACTS	Open Pharmacological Concept Triple Store
OPHID	Online Predicted Human Interaction Database
OWL	Web Ontology Language
PHFs	Paired Helical Filaments
PPIs	Protein-protein Interactions
PySB	Systems Biology Modeling in Python
R2RML	RDB to RDF Mapping Language
RDB	Relational Database
RDF	Resource Description Framework
RDFa	RDF in Attributes
RDFS	RDF Schema
RE	Relation Extraction
RIF	Rule Interchange Format
RNA	Ribonucleic Acid
RO	Relational Ontology
SKOS	Simple Knowledge Organization System
snoRNA	Small nucleolar RNAs

SPARQL	SPARQL Protocol and RDF Query Language
SW	Semantic Web
SWAN	Semantic Web Applications in Neuromedicine
SWRL	Semantic Web Rule Language
TMKB	Translational Medicine Knowledge Base
Turtle	Terse RDF Triple Language
UCB	Union Chimique Belge
URLs	Uniform Resource Locators
URIs	Universal Resource Identifiers
USA	United States of America
W3C	World Wide Web Consortium
WGCNA	Weighted Correlation Network Analysis
WHO	World Health Organization
XML	Extensible Markup Language

Glossary

This glossary provides information on the tools applied in this thesis. Further information on the listed tools can be found under the URLs provided.

Cytoscape	A software platform for visualization and analysis of complex biological networks along with integration of experimental of data http://www.cytoscape.org/
Knowtator	A text annotation tool integrated in the Protégé for manual information extractions tasks http://knowtator.sourceforge.net/index.shtml
ProMiner	A named entity and concepts recognition tool used in the field of life sciences https://www.scai.fraunhofer.de/de/geschaeftsfelder/bioinformatik/produkte/prominer.html
Protégé	A knowledge representation framework for ontology development and management https://protege.stanford.edu/products.php
SCAIView	A semantic search engine for biomedical concepts and entities from scientific literature using comprehensive biomedical terminologies and disease ontologies https://www.scaiview.com/en/introduction.html
MySQL	A SQL-based relational database management system https://www.mysql.com/
tranSMART	A leading knowledge management platform that integrates data storage and data mining applications needed for translational research and genomic research http://transmartfoundation.org/

Chapter 1 Introduction

The irreversible and debilitating nature of neurodegenerative diseases (NDD) — with no cure — has made it a daunting medical and socio-economic issues of our time. A focused interdisciplinary effort to transform our biological understanding of the brain, driven by technological advancements and large-scale data, aims to treat and eradicate NDDs. Today, it is possible to sequence a human genome in a day with cost of approx. $1000 compared to the cost of $3 billion and several years of effort needed for the first human genome sequencing. Yet, the multifactorial nature of these diseases has made it difficult to unravel its molecular underpinnings; leading to repeated drug failures. Thus, innovative paradigms are needed to discover meaningful players and gain biological insights from high dimensional feature space.

1.1 Alzheimer's disease: A looming global crisis

NDDs share a common property of progressive dysfunction and loss of neurons, which is the major cause of motor (ataxia) and mental dysfunction (dementia). In 2016, 47 million people were demented with an estimated global cost of $818 billion [1]. Owing to 100% drug attrition rate in the last two decades[1] [2], WHO has recognised dementia as the "public health priority" [3]. Alzheimer's disease (AD) is the most prevalent form of NDD, representing approximately 60–70% of the dementia cases. This global epidemic is currently the 6th leading cause of death and costs $160 billion in the USA alone, which will spike to $1 trillion by 2050 [4]. Moreover, AD prevalence has increased from less than 1% to 2.5% as the first baby boomers turned 65 [4]. If unaddressed, AD's economic burden will simply become unsustainable, driving millions below the poverty line.

AD is characterised clinically by progressive cognitive decline and neuropathologically by the presence of intraneuronal neurofibrillary tangles (NFTs) and extracellular amyloid-beta (Aβ) deposits — hallmark pathological features [5,6]. It begins with slowly progressing memory loss and advances to deteriorate higher intellectual and cognitive abilities, namely language, recognition, and personality [7]. The actual AD neuropathology(-ies) is thought

[1] https://www.ohe.org/publications/dementia-rd-landscape (this and subsequent URLs have been last accessed on 15th March 2018)

to begin 20–25 years before any apparent clinical symptoms, making it difficult for early diagnosis and treatment [8]. Moreover, a very thin line delineates the memory loss in the initial phase of normal ageing and AD [9]. This awareness has recently led to the refinement (first revision since 1984) of current AD guidelines and diagnostic criteria [10]. Thus, based on the disease continuum, AD is now classified as: (i) preclinical AD (newly defined stage) represents asymptomatic individuals with evidence of amyloidosis, synaptic dysfunction, and not overtly evident cognitive changes [11] (ii) in AD-MCI stage noticeable changes in memory and thinking are observed, disrupting day-to-day activities [12] (iii) AD dementia causes severe impairments of memory, thinking, and behaviour, needing support in everyday life [13].

Furthermore, two major categorisations of AD cases are: (i) Based on the inheritance pattern — familial and sporadic (ii) Based on the age of onset — early-onset AD (EOAD) and late-onset AD (LOAD). Familial AD exhibits the mendelian autosomal dominant pattern of inheritance attributed to several and varied highly penetrant mutations in more than 20 genes [14]. Accounting for 95% of the AD cases, sporadic AD is the commoner form whose precise aetiology is not yet known. However, it is attributed to multiple inheritances that include low penetrant genetic variants and non-genetic factors such as environmental risks [15]. Since sporadic AD mostly occurs after the age of 65 years, it is synonymously used with LOAD. EOAD accounts for 1-2% of all the AD cases with age of onset earlier than 65 years and accounts for 10% of familial AD cases [16,17].

1.2 AD etiopathogenesis

Although the AD cause-consequence debate still continues, many researchers have tried to elucidate its insidious features since its first description by Alois Alzheimer in 1907 [18]. Indeed, with the advent of molecular revolution in the mid-1980s, identification of AD genetic risks offered a promise of more rapid development in unravelling the AD aetiology [15]. However, some of the elementary questions asked decades ago about Aβ and NFTs, although highly topical, remains unanswered.

1.2.1 Amyloid cascade hypothesis

At this point of time, the central role of “neurotoxic Aβ plaques” is very strong in AD pathogenesis and believed to be “too big to fail”; Joseph *et al.* described it as the “Church of the Holy Amyloid” [19,20]. It has long been hypothesised that the core of Aβ plaques

formation is due to disordered proteolytic actions of α-, γ- and β-secretases on APP processing leading to abnormal folding of Aβ peptides, aggregating as insoluble plaques. Hardy and Allsop [21] postulated this Aβ dyshomeostasis as the primary event in AD pathological chain, known as the amyloid cascade hypothesis (ACH). Nonetheless, it was later transpired that (a) the mutations in the familial AD genes caused overproduction of Aβ 42 peptide (b) e4 allele of APOE (APOE4) gene is a potent risk factor [22] (c) decreased Aβ clearance was observed in LOAD cases [7] (d) soluble Aβ oligomers were primary neurotoxic agents [23] and (e) trisomy 21 (Down syndrome) led to overexpression of APP gene [24]. Most researchers accept that the downstream effect of Aβ plaques initiate tau hyperphosphorylation, leading to NFT formation, further synapse destruction, and subsequently causing neuronal death. However, studies also report that Aβ accumulation is observed in elderly individuals who show no signs of cognitive decline [25].

1.2.2 Tau hypothesis

The tau hypothesis identifies hyperphosphorylated tau protein as the possible culprit of AD pathogenesis and that tau tangles (also known as NFTs) occur prior to Aβ plaques formation [26,27]. Hyperphosphorylated tau loses its ability to bind to microtubules causing it to aggregate into insoluble tangles (known as paired helical filaments (PHFs)) to eventually form NFTs[2]. There is good evidence that hyperphosphorylated tau and its aggregates lead to the disruption of axonal transport, resulting in synaptic dysfunction [28]. Recent imaging studies, involving a large autopsy cohort (3618 brains), have linked tau deposits more closely to age at onset of cognitive impairment, disease duration and dementia than Aβ deposits [29]. Thus, tau is a speculated to be a better and more robust predictor of different stages when patients transition from healthy to severe AD [30].

1.2.3 Alternative hypotheses

With the passage of time, growing evidence reject the linear structure of either Aβ or tau being the singular cause in the cascade of AD pathogenesis. Conversely, we should not ignore the entirety of these hypotheses, rather revisit them with an assumption that they are

[2] NFTs are bundles of PHFs found in the cytosol of neurons

the consequence of paradoxical associations. To better unify and reconcile the existing hypotheses, alternative perspectives are proposed [31].

A steadily growing body of evidence suggests that the brain may compensate for the effect caused by Aβ, but the combined work of Aβ and tau drive the dramatic decline of healthy neurons [32]. The duo effect takes place either when tau renders neuronal dendrites to Aβ toxicity, or Aβ and tau synergistically amplify each other's toxic effect [28]. Some researchers have repositioned the causal molecular events of AD within the ageing spectrum as the histological boundaries between them are not absolute. Thus, AD's onset could represent the failure of the ageing brain to revert back the altered cell functioning due to events such as injury, infection, stress, negative life event, to name a few [33,34].

A recent study by AddNeuroMed Consortium [35], posits mitochondrial dysfunction as the primary pathology; reported altered mitochondrial genes in blood before any clinical diagnosis of AD. Furthermore, strong indications of oxidative stress and DNA damage in early AD pathology due to redox imbalance is reported [36,37]. Several neuroscientists argue that continuum of abnormalities in the cholinergic system [38], autophagy and/or lysosomal pathways [39,40], hormonal imbalance [41] and Ca2+ homeostasis [42] lies in the core of AD pathology. The vital role of neuroinflammation [43,44], highly active innate immune system [45,46], and disrupted insulin signalling [47,48] are strongly argued. Figure 1.1 depicts this conceptualization using three of the major AD contributing factors, refer [9,50,52,53].

Evidence suggests that some or all the above-mentioned events may augment to Aβ plaque formation and tau hyperphosphorylation, forming a vicious cycle that promotes AD pathogenesis [49,50]. Given this, Prof. Garrett proposes to approach the AD pathology as 'A Complexity Theory' where the effects of several causal variances are seemingly independent but rely on each other in ways, yet unknown, to bring about systematic malfunctions linking to the disease symptoms [51].

1.3 Status quo of AD therapeutics

Since 1960, ACH hypothesis has maintained supremacy in driving AD drug development strategies. Additional burden has been casted by gnawing controversies and major gaps in the basic biology and clinical pathophysiology. Despite huge investments, AD is one of the

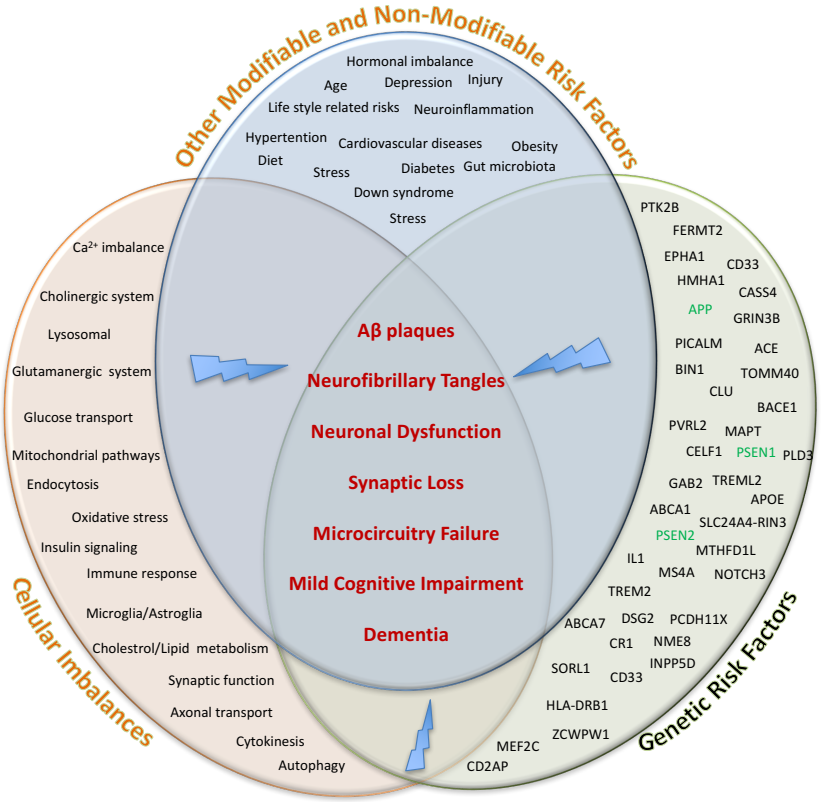


Figure 1.1: Overview of several factors that contribute to the clinical symptoms of AD.

Three major AD contributing factors are represented in the ovals: cellular imbalances (orange), genetic risk factors (green), and modifiable and non-modifiable factors (blue). Genes emphasized in green indicate their role in EOAD. The intersection between any two ovals remain empty to depict interaction between factors in ways currently unknown, contributing (trigger symbol) to common AD symptoms (in red).

least successful therapeutic areas with 0.5% success rate and with no blockbuster drug yet [52]. Many pharma companies are wary about investing significantly (both time and money) in AD after a series of high profile late stage failures, questioning investment in AD research [53]. To increase the probability of being successful and to speed up the hunt for AD's holy grail, several public-private initiatives such as the Innovative Medicines Initiative (IMI)[3] are providing platforms for collaborative projects to boost pharmaceutical innovation. Given the burning need for AD (prevention) therapy, FDA is now granting *fast-track* designations to potential interventions to reach the global deadline set in G8 summit — to prevent or effectively treat AD by 2025 [54,55]. Several countries have joined this global fight by encouraging a number of key strategic initiatives [55,56].

[3] http://www.imi.europa.eu/content/home

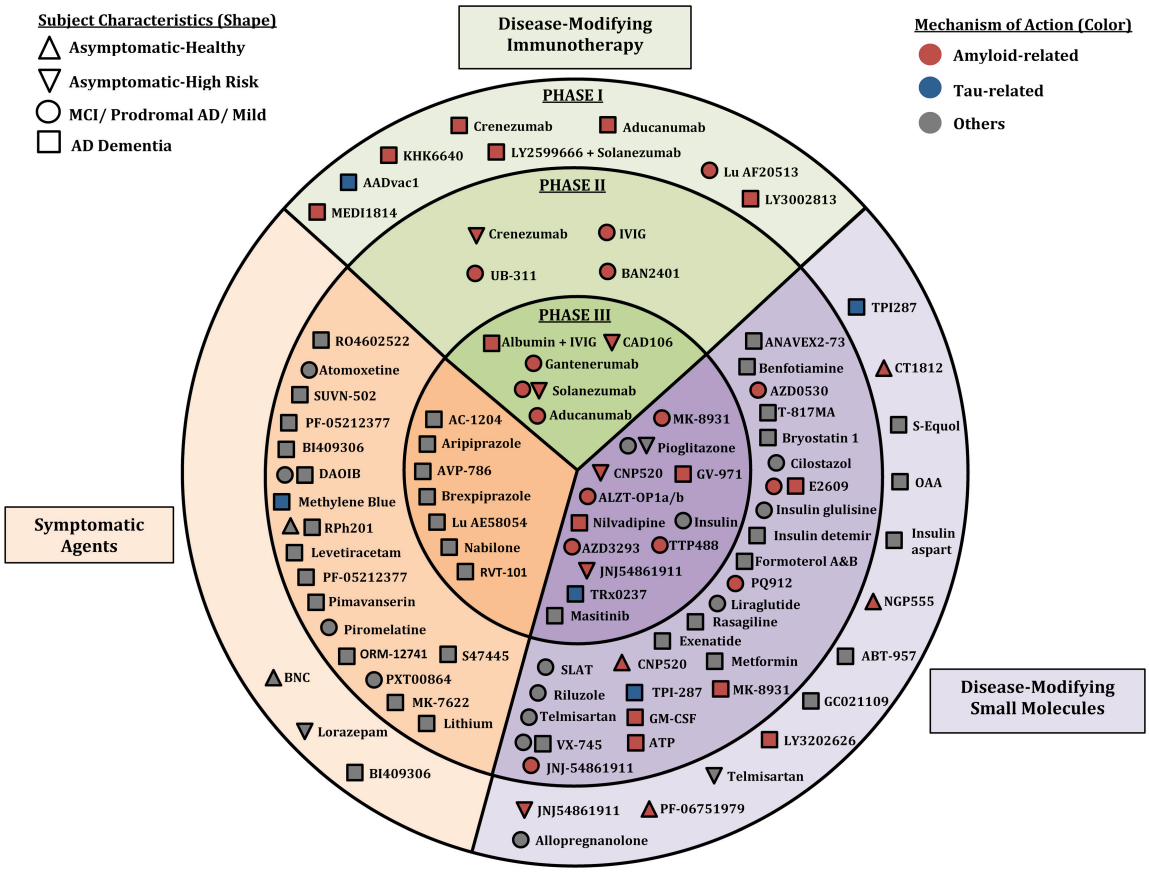


Figure 1.2: Overview of the ongoing clinical trials for AD therapeutics, reported according to their mechanism of action, phase of study, type of agents and targeted subjects.

Reproduced from Cumming *et al.* [57] under Creative Commons Attribution-NonCommercial-No Derivatives License.

Knowledge of neurotransmitter disturbances led to the development of currently approved AD palliative treatments [6]: (i) inhibitors of cholinesterase (tacrine, donepezil, galantamine, and rivastigmine) and (ii) NMDA (*N*-methyl-D-aspartate) antagonist (memantine). These drugs show no evidence of modifying the disease pathology but rather aim to slow the decline in quality of life — symptomatic treatment. However, increasing knowledge of AD's multifactorial nature has amplified the drug discovery ecosystem and rationales for modification in therapeutic strategies [58,59]. Although lagging behind Aβ-directed agents, some of these disease-modifying agents have advanced to initial human trials. Figure 1.2 provides an overview of the currently investigated AD drugs along with

the details of their mechanism of action, broadly categorised under (visit the PhRMA Foundation[4] for more details) [9,58,60,61]:

- *Aβ related* —targeting α-, γ- and β-secretases, clearance of Aβ aggregates, proteases, chaperones, immunotherapy (both active and passive)
- *tau related* — inhibition/clearance of tau aggregates, immunotherapy (both active and passive), targeting kinases and phosphatases, stabilizing microtubules
- *others* — modulating abnormalities in multiple neurotransmitters pathways such as cholinergic, glutamatergic, and GABAergic system, microglia-mediated inflammatory response, modifying epigenetics and/or epidemiological factors such as mitochondrial dysfunction, and metabolic disorders

Adding to the decade of bitter disappointments, two of the potential ground-breaking drugs have recently failed, verubecestat (MK-8931) and solanezumab, despite promising results in phase 2 trials [62]. Nevertheless, repeated drug failures have not yet unequivocally disproved the ACH belief [63]. Examining closer, failed trials provide no evidence of target being the problem but rather acknowledges methodological weaknesses: lack of drug/placebo difference, unacceptable toxicity, misdiagnosis of the enrolled patients, low dosage, and so on [58,59,64,65]. Many experts argue that the fundamental problem is the lack of awareness between cause and visible effects [66] and the lack of translation from mouse to human [67].

Increasing interest in combination drug therapies involving a "cocktail" of medications[5], aimed at several targets — with common associated biology — could address the profound complexity in AD, similar to current treatments in cancer [68]. This represents an important future direction in AD therapeutics; genuinely considering the systems biology approach and ending the vigorous debate between TAUists and BAPtists [69,70]. Moreover, neuroscientists suggest intervening early in the disease process before irreversible neuronal dysfunction prevails; similar to treating hypertension years before the incidence of cerebral infarction [66,71]. Unravelling new pathways amenable to neuronal changes (genetics or epigenetics) could improve the disease understanding and provide new therapeutic approaches [72].

[4] http://phrma-docs.phrma.org/sites/default/files/pdf/medicines-in-development-drug-list-alzheimers.pdf

[5] http://www.alz.org/research/science/alzheimers_treatment_horizon.asp

1.4 Elucidating AD mechanisms through computational approaches

Modern biomedical research is driven by technological advancements with growing prodigious amount of disparate data; drowning with information. Yet, it is the one that bedevils the progress as we still starve for knowledge. With millions of data points and myriad clinical information[6], the life science industry faces an increasing challenge of converting the harvested (complex-)data into actionable knowledge. In addition, there remain incredible barriers that have significantly stigmatized AD diagnosis and therapeutics: the high complexity of brain, the inaccessibility to good quality brain tissue, the lack of direct access to brain tissue in living patients, the lack of well characterised animal models, inadequate molecular diagnosis for cohort selection, huge cost and the time for extensive drug development processes, and current graveyard of AD clinical trials [73]. Thus, to stay in AD treatment race, pharma companies need to remain agile by skilfully drawing meaningful insights in a relatively short time, from limited observations and sparse data [74]. To increase the prediction accuracy, maximising the yield and biological relevance in downstream processes and critical evaluation of planned research, they need to leverage on huge volume of accumulated prior knowledge [75,76].

To derive a realistic model on modular nature of cellular architecture and functioning, taking stock of available knowledge on physical and functional associations between biomolecules have become a standard approach. In contrast, not all domain expert's knowledge is explicitly stated and manual interpretation is a daunting task; often leading to the question *"How can I realise the potential of these resources to construct a systems-level understanding?"*

Data integration approaches capable of describing complex systems and supporting broader interoperability are key to efficient integrative data analysis. Through these approaches, we may bring together previously overlooked factors (may or may not involve indicative biomarkers) that can uncover essential mechanistic relationships between molecular changes and diseases [77]. Biological information about diseases, genetic variants, experimental datasets, protein-protein interaction (PPIs), among others are well-

[6] https://www.nia.nih.gov/research/blog/2016/12/increasing-usability-big-data-alzheimers-research

documented and well-annotated in databases. However, these databases represent only a small percentage of information when the bulk of scientific publications are taken into account. Furthermore, experimental data, not being fully exploited contain compelling evidence for biological understanding (or validating) a new hypothesis. Moreover, to harness the full potential of the integrated data, the inferred biological hypothesis must assume the form of biological networks such as gene regulatory networks or projected onto previously compiled pathways. On the other hand, one must take into account that these public resources are fragmented, lack harmonisation and reproducibility is by and large inconsistent[7,8] [78–82].

Technological and data resources required to determine links to diseases are pieces of the puzzle that when put together, promise to reveal novel regulators in pathomechanisms. The following paragraphs introduce fundamentals and applications of various bioinformatic approaches and resources used in this thesis: first, introduction to integrative approaches applied to the AD domain, focusing on semantic web (SW) technology; second, a detailed description of the technologies and methods applied to distil knowledge from existing resources before integration.

1.5 Connecting the dots: semantic data integration to boost identification of AD driving mechanisms

To generate new insights into AD, several researchers have developed methods/tools to combine and show the extraordinary value of a wide variety of existing data through innovative re-analyses. Fowler *et al.* prioritized two genes involved in neuronal oxidative damage, to stratify patient subsets based on gender and APOE status: NEUROD6 for APOE4+ female and SNAP25 APOE4+ male patients; through the integration of publicly available gene expression datasets, a disease associated SNP datasets, and multiple databases [83]. Chen *et al.* developed a heuristic algorithm and scoring method to rank-order proteins based on their functional relevance in an AD-PPI network [84]. To derive a highly-connected AD-PPI network, an initial seed of AD-related genes was extracted from the OMIM database, which was further enriched with PPIs from OPHID database using a

[7] http://www.alzforum.org/news/research-news/replication-challenge-quest-alzheimers-blood-test

[8] http://protomag.com/articles/replication-in-research-problem

nearest-neighbour expansion method. Similar to Chen *et al.*, Soler-López *et al.* applied an interaction discovery strategy using initial seed gene list, and interaction network from public databases to prioritize novel genes, suggesting a link between plaque formation and inflammatory processes [85]. Krauthammer *et al.* presented a molecular triangulation to predict unknown genetic variants by computing the graph-theoretic distance between expert-curated seed genes and other biomolecules in the literature-derived molecular network [86]. Considerable studies integrated information about genetics, functional, dysregulated expression, and interaction from public data to unveil novel AD candidates and provide new hypotheses for mechanisms underlying AD [87]; for drug discovery [88–90]. Among others, most widely used data integration strategies include data warehousing, data centralization, and federated databases [91].

Problems for data integration are rooted in the data itself; resources are broadly distributed, across the web and encompass heterogeneity and diversity in data formats, concepts, semantics, syntax, to name a few. Of course, these approaches have strengths and limitations and there is no "one-size fits all" solution [75,92]. Furthermore, these integration solutions are local and fail to operate at the global level and cannot cope with the updates of resources such as newly added information and changes in the data structure, formats, and naming convention. In general, most of these approaches overlook the importance of data quality, and context-specificity. Semantic web technology is the first truly global integrative solution revolutionising the lossless exchange and formalisation of data, calling it "smart data" [93].

Since the inception of World Wide Web, its inventor Tim Berners-Lee *et al.* envisioned SW as "intelligent agents" capable of universal integration and exchange of data through the incorporation of machine-readable meaning (or semantics) and logical relations between data elements [9] [94]. Thus, resulting in a network of linked data, whose formalization allows identification of new implicit connections by reasoning over the data. To realize the vision of SW, World Wide Web Consortium (W3C) focused on empowering SW technologies, among which RDF, OWL, SKOS, and SPARQL have become *de facto* standards[10]. The 7-layered Semantic Web Stack, depicted in Figure 1.3, shows, how the

[9] the definitions and descriptions of SW are taken from W3C's web page https://www.w3.org/TR/

[10] https://www.w3.org/standards/semanticweb/

proposed technologies (although still evolving) realise each other's capabilities towards building SW. Interested readers are referred to Glimm *et al.* for the overview of the SW developments since its inception 15 years ago [95].

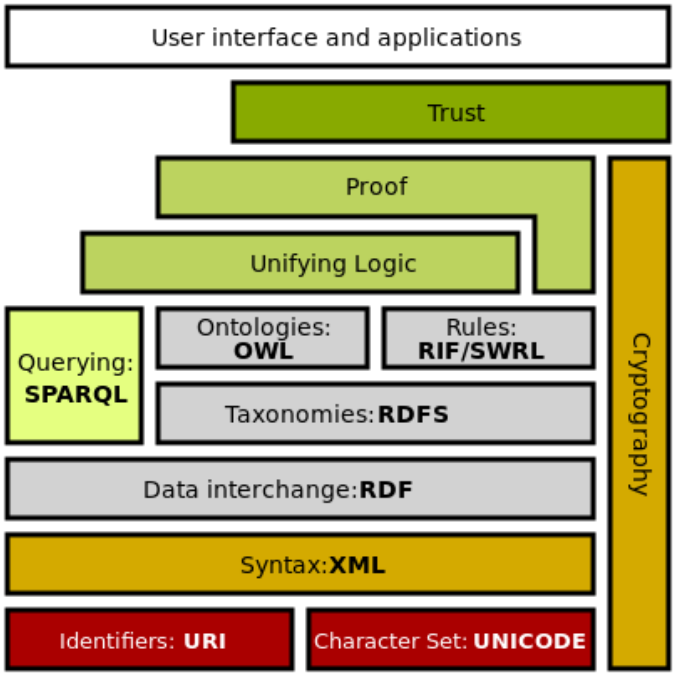


Figure 1.3: Semantic Web Architecture, also informally known as "layer cake"

This figure is taken from Wikipedia page[11] under Creative Commons (CC0) License. The original illustration proposed by Tim Berners-Lee is available here[12] Copyright © 2015 W3C® (MIT, ERCIM, Keio, Beihang)[13].

1.5.1 Semantic web technology standards

Resource Description Framework (RDF) is a W3C's proposed standard for publishing and exchanging data on the web [96]. The core concept of RDF lies in the usage of a unique global identification system called as "Universal Resource Identifiers (URIs)" and more recently IRIs (Internationalized Resource Identifier) [97]. RDF data model uses the syntax of Extensible Markup Language (XML) to impose structural constraints for representing the data as graph structures. Due to high flexibility and cost effectiveness of graph databases (introduced in sub-section Biological databases), RDF-centric databases have

[11] https://en.wikipedia.org/wiki/Semantic_Web_Stack

[12] https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)

[13] http://www.w3.org/Consortium/Legal/2015/doc-license

become a choice for managing highly connected data. RDF triplestore is a semantic graph database which stores semantic facts as a network of links containing directed edges; termed as RDF statements or triples, hence the name triplestore [98]. Figure 1.4 shows the anatomy of a basic triple statement; subjects and objects are concepts (called resources), connected using verbs which represents the relationship types, called predicates and literals (also a resource) are the constant values mapped to the resources.

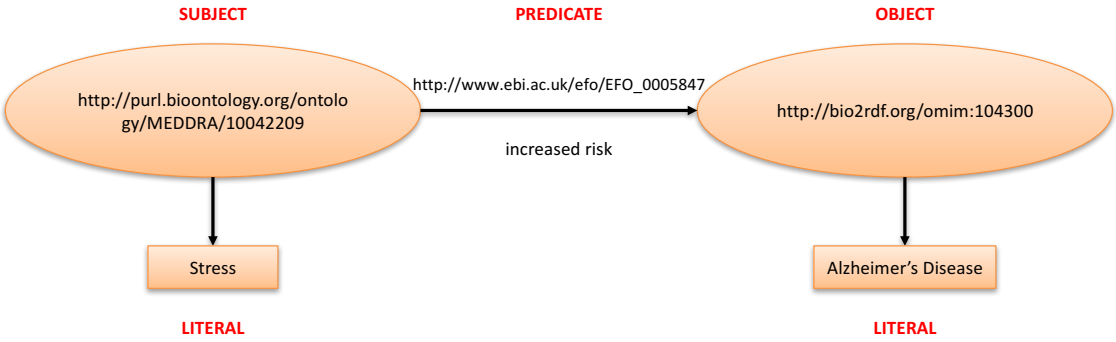


Figure 1.4: Anatomy of a triple statement.
Ovals represent subjects and objects; rectangle literals; arc predicates.

The RDF's simple triple format — although simpler to implement — does not allow higher levels of expressiveness such as the union of existing concepts, hierarchical relations between concepts, reasoning, among others [99]. Thus, W3C introduced two data modeling languages: RDF schema (RDFS) and Web Ontology Language (OWL). RDFS is objected oriented in nature and formally describes RDF data properties as taxonomies of object classes, and relationship properties using ontologies. Simply put, it defines a metamodel of concepts such as Resource, Literal, Class, and Datatype and relationships such as subClassOf, domain, and subPropertyOf. Ontologies allow explicit formal description of the terms — rich vocabulary with highest level of expressivity — in the data to map distinct terms to the same concept. Semantic Web Rule Language (SWRL) is an extension of OWL to provide more powerful "deductive reasoning" capabilities [99]. Although it is a daunting task to provide data linked to ontologies, the merit lies in higher interoperability. SPARQL, a self-referencing acronym for SPARQL Protocol and RDF Query Language, is a SQL-like query language for accessing RDF data. SPARQL queries can be federated, meaning one can access diverse and evolving data from various RDF resources in one query. The advantage of SPARQL over other query languages is the availability of a query interface

called SPARQL endpoint. Several programming environments have been developed to parse RDF data: PerlRDF for Perl, and Apache Jena for Java framework. Data exchange standards[14] for RDF include: RDFa, RDF/XML, N-Triples, Turtle, Javascript Object Notation-LD (JSON-LD), among others. Several optimized databases[15] are available for storing RDF data, namely Virtuoso, GraphDB™, Stardog, and AllegroGraph. Among the variety of extraction tools, strategies, and interfaces to transform non-RDF into RDF resources: GRDDL transforms XML into RDF, and R2RML maps relational databases to RDF.

RDF is capable of handling intuitive and powerful semantic queries (set of inference rules) to infer new triples (logical consequences) out of asserted facts; turning information into knowledge. For example, *"If two diseases have common genes, then they affect each other's incidence"*. This gives a competitive edge for most pharma industries by creating more value and easily scaling up the derived knowledge into smart solutions. RIF (Rule Interchange Format) is a standard for exchanging rules between disparate semantic data models by combining ontologies. Using semantic reasoners, one can infer implicit facts out of explicit statements, thus uncovering hidden relationships. Since OWL provides Description Logic (DL)-based reasoning capabilities, a number of reasoners including Racer [100], Fact++ [101], Pellet [102], and KAON2[16] have been developed. Mishra *et al.* [103] have published an overview of semantic reasoners recently.

1.5.2 Bridging the knowledge gap through semantic web: focus on neuroscience

Despite the youth of SW technologies, active researchers and developers have been developing tools and infrastructures, both open source and commercial, that foster FAIR data (findable, accessible, interoperable and reusable) principles [104]. The OBO Foundry and BioPortal serve as an umbrella of ongoing collaborative efforts for standardisation, storage, and linking of public biomedical ontologies [105]. The Identifiers.org registry provides persistent official identifiers to scientific terms. Observing the immense advantage

[14] https://www.w3.org/TR/rdf11-new/

[15] https://www.w3.org/wiki/LargeTripleStores

[16] http://kaon2.semanticweb.org/

of linked open data, a number of existing bioinformatics resources have adopted SW as data exchange standards: Entrez Gene, EBI data resources, KEGG, and many more [106,107]. The importance of semantic data integration and mining in the life science domain was brought to limelight by the Bio2RDF and subsequently by Linking Open Drug Data (LODD) project [82,107]. These projects demonstrated the possibility of querying heterogeneous life science resources (public) as linked data. Steady efforts by W3C's Semantic Web interest group to focus on Health Care and Life Sciences (HCLSIG) has led to the launch of projects such as for modeling ontologies [108], RDF-based graph system (LinkHub [109]), biological pathways (BioPAX [110]), and drug discovery (AlzPharm [111], TMKB [112][113]). A series of annually hosted DBCLS BioHackathons[17] serve as the driving force to integrate life science databases using SW technologies, through improved interaction between providers of data and bioinformatics tools. Recently, "The Monarch Initiative[18]" has taken the semantic route to enable reasoning over genotype-phenotype equivalence (similarity analysis of biochemical models) for generating new hypothesis and prioritising candidates/variants within and across species.

A number of initiatives and projects are striving to advance neuroscience research by allowing sustained interlinking between data using SW technologies. Government-sponsored endeavours such as the USA's BRAIN[19] initiative and Europe's Human Brain Project (HBP)[20] are among a few leveraging on SW technologies for data management. Under the premise of the US government, the Neuroscience Information framework (NIF)[21] project, an initiative of NIH Blueprint for Neuroscience, aims to advance neuroscience research by providing "one-stop-shop" to public neuroscience data and tools in a semantically enhanced networked environment. Some of the NIF – backend research outcome includes BIRN, NIFSTD, NeuroLex, and many more[22]. Other NIH-funded

[17] http://www.biohackathon.org/

[18] https://monarchinitiative.org/

[19] http://www.darpa.mil/program/our-research/darpa-and-the-brain-initiative

[20] https://www.humanbrainproject.eu/en/

[21] http://neuinfo.org

[22] https://neuinfo.org/Resources/search?q=%2A&l=

projects that use SW technologies include BD2K [114], and Commons[23]. The Open Science Framework[24], the International Neuroinformatics Coordination Facility's (INCF)[25], the Neuroscience Database Gateway (NDG) [115], NeuroML[26], DARPA's Big Mechanism program [116], and others are fostering community efforts to use SW technologies as the core to catalyse neuroscience research. Originally as a part of HBP, the SenseLab project provides a suite of interrelated databases to gain insights into the neuronal basis of behaviour [117]. As part of SenseLab, the BrainPharm database[27] stores information of NDD drugs/agents targeting neuronal receptors and signal transduction pathways (differentiates between diseased and healthy). The Linked Neuron Data (LND) [118] provides a platform for integration of multi-scale brain and neuroscience data and knowledge sources with the aim to understand the association between cognitive functions and brain diseases. Currently, LND integrates structured neuroscience knowledge from Allen Brain Atlas, NeuroLex, NeuroMorpho, Mesh terms, etc. The IMI's Open PHACTS project [119] aimed to integrate diverse chemical and biological data resources for pharmacological research. Apart from these, BioGateway, BAMS, NeuroMorpho, NeuroSynth database, Entrez Neuron, DisGeNet, Cognitive Atlas, to name a few also apply SW technologies [120]. For more details refer to Nielsen for neuroinformatics databases [121] and Okano *et al.* for brain mapping projects [120].

Among the several collaborative AD projects that receive European Union funding and that are under Framework Programme, European Medical Information Framework (EMIF)[28], AETIONOMY[30], EU Joint Programme – Neurodegenerative Disease Research (JPND)[31], and ELIXIR[32] utilise(-d) (partially-)SW technologies for data management. One notable community effort in AD is The Alzheimer Research Forum (Alzforum) [122], which

[23] https://datascience.nih.gov/commons

[24] https://osf.io/

[25] https://www.incf.org/

[26] https://www.neuroml.org/

[27] https://senselab.med.yale.edu/brainpharm/

[28] www.emif.eu

[30] http://www.aetionomy.eu/en/vision.html

[31] http://www.neurodegenerationresearch.eu/

[32] https://www.elixir-europe.org/about-us/how-funded/eu-projects

benefits from both social and technical solutions to nurture productive discussion and informal discourse to advance AD therapeutics. Alzforum's SW initiative, AlzSWAN (SWAN — Semantic Web Applications in Neuromedicine) is a hypotheses management system that captures a significant amount of AD scientific discourse from hypotheses, claims, dialogues, publications, and digital repositories. Active participation of Alzforum in the HCLSIG has led AlzPharm development – integrates BrainPharm and SWAN [111]. The Global Alzheimer's Association Interactive Network (GAAIN) project [80] attempts to build a global collaborative platform for sharing AD data such as ADNI, by overcoming data-sensitive sharing impediments.

1.6 Knowledge discovery: Needles in stacks of needles

An important requirement for any data integration approach is to first capture the relevant data from diverse resources in an efficient and effective way. However, the most common problem is to condense useful information from these data mountains and transform them into actionable knowledge. Several data mining methods have shown great promise in closing the gap between large disparate data and discovering hidden/new knowledge. In this thesis, we have focused on three major data sources: databases, literature, and transcriptomic data.

A compendium of transcriptomic studies provides quantitative information on the state of the gene in a cell. Most certainly, modeling of interactome and regulatory relations represent confirmed knowledge when derived from omics data. However, they do not account for the domain knowledge, which is important in any scientific discovery. Databases provide a systematised collection of biologically important information; increasing in number every year according to *Nucleic Acids Research* journal's annual compendium of peer reviewed databases [123]. However, they do not fully represent the current state of rapidly growing knowledge. Conversely, vast collections of literature data are a massive body of existing current knowledge that can fill knowledge gaps and assist in informed decision making. However, due the data deluge, it becomes unmanageable. Below we discuss how data mining approaches applied to these data models contribute to knowledge discovery.

1.6.1 Omics data analysis: complex biological data streams

The advent of high-throughput technologies has fuelled the search for unique molecular markers that govern the information flow in "the central dogma" framework of molecular biology. Omics-based approaches are now broadly used for the identification of disease markers and understanding underlying pathomechanisms; supporting hypothesis-free elucidation. Particularly, it provides a holistic view of genes (genomics), gene expression (transcriptomics), proteins (proteomics), and metabolites (metabolomics) through a variety of techniques including mRNA and miRNA arrays, NGS, and mass spectrometry [124]. Recent research has led to the revelation that RNA is not just a simple genetic messenger but rather plays a central role in translating genetic code into protein, gene silencing, post-transcriptional regulation, and as a modulator of epigenetic elements [125]. Relative to the fixed nature of DNA sequence variation, gene expression varies tremendously between tissues, cells, and response to external stimuli [126]. Thus, gene expression profiling, encompassing many species of RNA such as miRNA, mRNA, and snoRNA, represents a rich source for early diagnosis by revealing altered transcriptome signatures of the cells, and tissues under a given biological state [127].

Quantification of RNA abundance using microarray technologies [128] or, more recently, developed RNA-Seq [127] methods have led to the accumulation of large amounts of data in public repositories. As of 15th March 2018, GEO contains 2,429,236 samples from 4,348 datasets[129] and ArrayExpress hosts 70,834 experiments [130]. Although RNA-Seq is a substantially advanced technology with several advantages over microarray [127,131], microarrays are still widely used as they are less expensive, more consistent with the already existing wealth of data, and there exist substantial number of robust statistical methods [132,133]. Extracting biological information from these data is done by identifying individual genes (differentially expressed (DE) genes) associated with a particular biological effect (such as fold change) or finding global signatures composed of multiple gene expression changes. A more consistent and robust approach is looking for genes that share a particular biological characteristic [134]. However, low reproducibility and low overlap with similar studies performed by other study groups, render gene expression levels incomparable, mainly arising due to several technical and biological variabilities like applied analytical methods, different platforms, and dependency on library preparation [131,135,136]. Due to difficulties in acquiring human brain tissue and the

associated cost, NDD research experiments are composed of small sample sizes, resulting in less robust gene signatures and missing out on less apparent signals [137]. Thus, there is a need to fully exploit the existing data for more compelling evidence that could pave way for next ground-breaking discoveries.

Combining multiple transcriptomic studies (termed as "cross-platform normalisation (CPN)") or their results (termed as "horizontal meta-analysis") has been advocated to increase the power of derived conclusion by overcoming the biases of individual studies [137–140]. These approaches have been used to uncover disease subtypes, predict survival, discover new biomarker and therapeutic targets [141–144]. Although it is argued that CPN is more powerful, it is less frequently used as: (i) it fails to eliminate batch effects across experiments (ii) very few well-established algorithms are available and (iii) increased complexity of data integration [145]. Refer to Rudy and Valafar for detailed comparative analysis of different CPN methods [146]. Most current meta-analysis methods are gene-centric, combining DE genes based on majority voting, gene rank aggregation, and combining univariate summary statistics, such as p-values and effect sizes [147–149]. A more consistent and robust approach is through functional enrichment of the identified genes using established pathway knowledge such as KEGG [150], MSigDB [151], or, more recently, NeuroMMSig [152]. For the majority of the meta-analysis approaches, functional enrichment has become a standard follow-up. However, these approaches often have a tendency to converge towards genes that express in large magnitudes and generated hypotheses are restricted by current understanding of pathways. Moreover, these approaches do not shed light on the coordinated genes that collectively orchestrates the underlying (patho-)mechanism, unravelling dysregulated events heralding known and unknown patterns. Network-based approaches that rely on the coherence of functionally dependent genes could ameliorate DE gene's dependency and increase confidence in biological validation by collapsing the number of testable hypotheses with regulatory clues.

1.6.2 Biological network inference

Cellular and molecular components work in concert with a large number of dynamic partners — directly or indirectly — to execute or govern cell/tissue phenotypes [153]. The power of biological networks resides precisely in simplifying the complex systems merely as nodes (biomolecules) and edges (intramolecular interactions) in the form of pathways, protein-protein interactions (PPIs), miRNA-target interactions (MTIs), among others. In an

attempt to dissect higher level organisation of molecular and cellular communications, information on modules of genes, proteins, or miRNAs that are physically associated, or functionally co-ordinated are translated into physical and functional networks. Physical interaction networks represent how biomolecules of interest interact with each other. Functional networks aim to connect not only interacting but also non-interacting biomolecules that depict functional or regulatory dependencies; examples include pathways, co-expression networks (CENs), and gene regulatory networks (GRNs). Analysis of these networks relies on characteristic topological properties, which serve as scaffold information for global and local graph theory [154–156]. Through these networks, useful discoveries for identification of putative biomarkers, understanding the disease-driving mechanisms, and insights into the research findings can be made.

Network inference (NI) methods have recently emerged as highly effective "reverse engineering framework" to reconstruct biological networks based on educated inference from data profiles, reducing the cost and time associated with the experimental investigation by prioritising putative candidates [157–161]. In the last few years, we have seen a swarm of NI approaches that majorly fall under: (i) deconvolution methods applied to the literature [162,163], databases [150], and multi-omics data and (ii) prediction algorithms based on thermodynamic stability, and sequence similarity [164,165].

Networks inferred from literature and databases represent "what is already known" and are usually used as reference or gold standard to put the inferred results from other NI approaches in a specific biological context. Although, these networks as quite large, most of the interactions cannot be easily filtered for a specific biological context and data formats are not easily interchangeable. Recent endeavours have led to the development of standardised languages that use rich semantics for modelling networks: OpenBEL, and PySB [166]. On the other hand, use of genomic profiling technologies is more reliable for uncovering previously unknown and underappreciated mechanistic links along with involved putative candidates. Genomic data-based NI approaches have transformed biological research by enabling comprehensive monitoring of co-expressed and co-regulated components. I refer the reader to Lee *et al.* [167] for conceptually different GRN methods, Markowetz *et al.* [168] for other NI approaches, and a book chapter by Vert [169] for machine-learning based NI approaches. Series of *The Dialogue for Reverse Engineering*

Assessment and Methods (DREAM) challenges allow comparison of strengths and weaknesses of different network inference methods[33].

Gene regulatory networks and co-expression networks

Most researchers mistakenly use GRNs and CENs synonymously, however, the latter may contain co-regulated genes that represent the former [170–173]. CENs comprise of gene clusters where the edge (non-directed) represent similarity or dependency in expression patterns between two genes across tissues/cells. Similarities are usually quantified by Pearson correlation, Spearman correlation, mutual information, or linear modelling [174–176]. On the other hand, GRNs (directed graphs) capture regulatory relationships (such as causal influence, and transcription regulation) assuming that changes in expression level of regulatory elements should be mirrored in expression levels of its regulated elements [170]. Allen *et al.* proposes four co-expression measures to define gene similarity metric for inferring GRNs [177]: (i) probabilistic-based, e.g. Bayesian networks [178] (ii) correlation-based, e.g. Weighted Correlation Network Analysis (WGCNA) [179] (iii) partial correlation [180], and (iv) mutual information-based (MI), e.g. ARACNE [181], MRNET [182], and CLR [183]. For details of these measures, I refer the reader to Song *et al.* [184] and Kiani *et al.* [185].

Failure to identify more complex dependencies between the genes by correlation-based methods is overcome by MI methods. Moreover, MI-based methods apply refinement approaches to eliminate indirect interactions for a given threshold using empirical distribution (CLR), Data Processing Inequality (ARACNE), maximum relevance/minimum redundancy criterion (MRNET), and predictions based on estimates of MI values with a maximisation step (C3NET [186]). Although Bayesian inferred networks are capable of modelling higher order dependencies, they lack feedback loops and some are limited to time series data [187,188]. A recent trend, ensemble-based methods are reported to improve stability and accuracy by formulating feature selection using random forests [189], gradient boosting [190], least angle regression [191], and partial least squares [192]. Briefly explained, these methods apply NI approach(-es) on bootstrapped data, aggregating the results in a final network; examples include BC3Net [193], and GENIE3.

[33] http://dreamchallenges.org/publications/

Clearly, these methods have an advantage of being straightforward and efficient on large computer clusters. A recently patented approach by Leiserson *et al.*, called *the heat diffusion based genetic network analysis* [194], identifies known and unknown pathways by determining local neighbourhood influence of each mutated gene via physics of heat diffusion in the network [195,196] has brought a lot of attention to GRNs.

Choice of the method to infer GRNs depends on the studied conditions such as type of data (real or simulated), network size, number of samples, noise level, experimental design (intervention, observational), underlying interaction structure (scale-free, random), error measure (local, global), among others [197]. However, recently, de Matos Simoes *et al.* demonstrated for C3Net, BC3Net, and ARACNE that the differences are not large if one focuses on the biological consistency rather than technical [198].

Application of GRNs and CENs in AD

Using public transcriptomic data, Rhinn *et al.* identified key regulatory molecules (APBA2, FYN, RNF219 and SV2A) and pathways (endocytosis, intracellular trafficking) involved in APOE-based risk for LOAD [199,200]. Their work focuses on differences between diseased and healthy co-expression patterns. Zhang *et al.* [201] identified eight immunity- and microglia-specific genes, including TYROBP, strongly dominating the LOAD pathology; inferred from GRNs generated using 1,647 post-mortem brain tissue of LOAD and non-demented subjects. From these results, the authors concluded that the causal network framework was a useful predictor of response to gene perturbations and could be used to test models of disease mechanisms underlying LOAD. Forabosco *et al.* [202] reported TREM2 to be a hub gene in 5 out of 10 brain regions in neuropathologically normal individuals using the co-expression network analysis. Additionally, they found highly enriched genes in TREM2-containg module that are genetically implicated in AD, sharing common pathways centred on microglia functioning. Miller *et al.* [203] identified convergent and divergent co-expression modules between 18 human and 20 mouse public microarray datasets. Significantly, they determined three hub genes (for human only) with zinc-finger motifs, whose exact functioning in dementia was previously unknown. In a similar approach, Ray *et al.* [204] revealed transcriptional commonalities that might explain the co-occurrence of cardiovascular diseases and AD. Additionally, several efforts have been made to unfold the links and common mechanisms between AD and ageing [134].

The current multi-omics-based GRN approaches have an intrinsic limitation of dependency on (i) known interactions catalogued in databases and literature for follow-up analysis (ii) well-known gene candidates to refine networks within its proximity and (iii) restricting inference on genes that exhibit a clear shift in expression behaviour. This means that the derived results have a tendency to converge to "what is already known", missing out on lesser-known candidates. Additionally, none of the above approaches elaborates on context-specificity and completeness of the generated networks, undermining the modules that approximate the biological truth.

1.6.3 Biological databases

Databases provide convenient, searchable, visualisable, and computable access to organised prior knowledge. They are indispensable research tools for translating "big data to big discovery", hosting enriched and pertinent information, [205]. Biological databases are developed for diverse purposes and encompass heterogeneous data types, and formats, refer Kumari *et al.* [206]. These databases can be classified based on:

- *type*: primary, secondary (curated or/and value-added)
- *nature of stored information*: RNA, drugs, pathway, miRNAs, among others
- *curation*: expert-curated, community curated (crowd-sourced)
- *storage type*: MySQL, NoSQL, flat files

Primary databases host experimentally derived raw sequence read data (for proteins, nucleotide, so on) or macromolecular structure, which is directly submitted by the researchers; examples include ArrayExpress [207], miRBase [208], and GEO [129]. Secondary databases, mostly curated, build upon the information derived from primary databases. For example, ExpressionAtlas derives knowledge about gene expression patterns from ArrayExpress archive [209]. Zou *et al.* provides a comprehensive overview of human databases, categorised based on the data type and nature of information stored [205]. Highly knowledgeable and experienced biocurators critically assimilate and review the information before being stored in expert-curated databases, namely UniProt [210,211]. Crowd-sourced databases have proven to be an efficient, economical, and faster way to harness knowledge from the scientific community with broad coverage; RiceWiki is a good example [212]. Relational databases (e.g. MySQL and PostgreSQL) are an efficient way to access structured information using declarative query language (e.g. SQL) for a pre-specified set of operations and schema. NoSQL databases are whiteboard friendly that

reflect our natural thinking and support agile development for dynamic schemas and are quickly scalable for additional data integration [213]. NoSQL databases include key-value stores (e.g. Berkeley DB), document stores (e.g. MongoDB), wide column stores (e.g. Cassandra), and graph databases (e.g. Neo4j) [214].

Community efforts keep researchers abreast with the growth of bioinformatics tools and databases in the form of trusted directories/databanks, such as Biological Links Directory [215], and OmicTools [216]. Several core bioinformatics organisations provide an amalgamation of multiple primary and secondary databases covering different data types and species; examples include NCBI [217], EMBL-EBI [34], and Swiss Institute of Bioinformatics [218]. In addition, several commercial ventures such as Ingenuity Pathway[35], NextBio[36], and MetaCore™[37] provide a wealth of curated information in a structured form. However, most of the databases vary in data quality and become obsolete over time [219].

Primary transcriptomic repositories, namely GEO and ArrayExpress, provide a wealth of molecular data to conduct integrative meta-analysis and inferring GRNs. This allows researchers to reproduce and/or reanalyse existing data for new discoveries, especially when the data availability is limited (see section Omics data analysis: complex biological data streams). To consistently integrate heterogeneous data, accurate details of the associated metadata information including patient's age, gender, pathological diagnosis, and comorbidity are crucial in clinical practice. In addition, mapping of this information to standard ontologies could increase the compatibility and usability across studies. The Ioannidis study [220] highlighted the importance of metadata information in reproducible science. Pioneering attempts to adopt guidelines for submission of minimum metadata information required for data reproducibility such as MIAME[38] and MINISEQE[39], still lack compliance. Often, the data submitter and data generator are different persons, increasing

[34] https://www.ebi.ac.uk/services/all

[35] https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/

[36] http://www.nextbio.com/b/nextbio.nb

[37] https://lsresearch.thomsonreuters.com/pages/solutions/1/metacore

[38] https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html

[39] http://fged.org/projects/minseqe/

the risk of errors and missing information [221]; sometimes leading to the metadata information being scattered in the associated publication(s). There is no easy way to tackle this problem and extensive manual effort is needed. Several (non-)commercial value-added databases and tools such as NextBio, and ArrayExpress have invested considerably to manually extract and correct the missing and erroneous metadata information. Indeed, to our knowledge none of the value-added databases capture NDD specific metadata information. In addition, they fail to explicate annotations that distinguish diseased human from NDD-induced immortalised cell lines, and mouse/rat strains, which are critical for translating preclinical studies to drug trials. Thus, there is a need for a dedicated approach to extract and refine metadata annotations catered to NDD research.

1.6.4 Text mining: discovering hidden connections

Biomedical literature is the key communication channel for scientific findings and hypotheses in the form of research articles, conference proceedings, reviews, books, and monographs [222]. Advancing with impressive speed, automated technologies — text mining — have complemented the manual reading for extracting and reconstructing mosaic of non-trivial and implicit knowledge from unstructured (or semi-structured) text, along with provenance [223]. Although not trivial, text mined information has the capability to shed perspective on modeling complex biological systems by summarising the entirety of prior research [224]. Text mining techniques can be simply abstracted to four phases: information retrieval (IR), information extraction (IE), knowledge discovery (KD), and hypothesis generation [225,226].

IR deals in identifying and triaging relevant textual sources to seek background information, addressing a research question at hand. For the overview of current IR tools/services refer to Lu *et al.* [227]. In the biomedical domain, IE involves identification of predefined classes of biomedical entities (genes/proteins, miRNAs, drugs, etc.) and relations between these entities (drug-gene, gene-miRNAs) from the text. Tagging key biological entity mentions in the text is the first step in IE, called as named entity recognition (NER), performed using predefined vocabulary (dictionary-based), applying rules (rule-based), or classifying (machine-learning-based) on the basis of training data [228,229]. Relation extraction (RE) adds context to the identified entities by extracting relationship(s) between them through association-based (co-occurrence and tri-occurrence) or natural language processing (NLP)-based methods [222,230,231]. More narrowed

application of RE is event extraction (EE) [232], which focuses on identifying specific events such as phosphorylation, and inhibition. To benchmark the developed methodology, several expert-annotated reference corpora are available, but they do not cover all the entities and relations [230,231,233]. In contrast to IE, which extracts nuggets of information, KD aims to extract new knowledge for answering biomedical queries [234]. Furthermore, hypothesis generation infers novel and testable insights from hidden clues in the text that are not easily derivable through expert reading [235]. However, text-mined information is error prone and must be crucially assessed by human experts as they are inherently limited by variable quality, lack of systematisation, and absence of reporting negative data [236–238]. Several annotation tools are in place to speed up manual curation process [239–241]. Specialised databases are established to provide standardised and structured accessibility of the harvested literature knowledge [242–245].

Many cellular functions — biological and pathological — are a result of cross-talk between different bio-entities, namely genes, proteins, transcription factors, and miRNAs. Therefore, to fully uncover the modular organization of the cellular networks it is crucial to elucidate these players. Although extensively researched, protein-coding genes represent only 2% of the human genome suggesting that PPIs are just half of the story in AD biology. With the beginning of miRNA era in 2001, non-coding RNAs have become attractive targets and research topic for novel diagnostic and therapeutic approaches [246]. Information about miRNA's regulatory roles has been widely discussed in the literature. Thus, utilising biomedical text-mining approaches to extract new evidence from existing literature has become very crucial to drive AD research.

MicroRNAs (miRNAs) are highly conserved small non-coding RNAs (21-25nt), post-transcriptionally regulating 30% of protein-coding genes through mRNA degradation and translational inhibition. Previous studies have reported on the essential roles of miRNAs in neuronal functioning, and survival and its potential implications in modulating AD genes [247–249]. The cross-talks between AD-related miRNAs and genes/proteins are far more intricate and dynamic than anticipated but poorly understood. Significant research efforts in dissecting miRNA-related associations (e.g. miRNA-target, miRNA-disease) have resulted in high-quality databases, curated networks, and prediction algorithms. In reality, relative to PPIs, automated text-mining methods dedicated to the identification of miRNA-related relationships are limited and not widely adopted [250]. Indeed, resulting in a lack

of expert-annotated corpus needed for benchmarking the developed tools [251]. Thus, there is a clear need for automated text-mining approaches/tools/resources to support and drive the miRNA research for a new perspective on diseases at post-transcriptional level.

Chapter 2 Goals and Objectives

Enabling scientists to reuse and extend the work of other researchers is frequently perceived to have "an upper hand" in generating time and cost efficient testable hypotheses, assisting biomarker and mechanism discovery. Repurposing public data using data collation and integrative approaches could act as an evidence store for deriving new knowledge where traditional approaches fail to deliver [252]. Several studies and projects have employed data mining approaches on public data to identify novel mutations, genes, pathways, among others that were previously undetected in AD [253–255]. These reports are a proof that retrospective analysis of public data can provide important insights for clinical utility. However, they have also inherently raised several questions illuminating fundamental aspects of data reusability and retrospective analysis: (i) lack of high-quality and informative data (ii) lack of sufficient metadata information for more focused reanalysis (iii) overfitting the developed models to prior knowledge (iv) data bias, and (iv) a need of a high-quality pre-competitive infrastructure for data integration.

2.1 Issues addressed and goal of this thesis

Motivated by the need of novel approaches for integrative analysis and considering the scepticism around public data, the primary goal of my PhD thesis is to provide improved and novel solutions for accessibility, reusability, and unbiased retrospective analysis of high-quality public data with a potential impact on uncovering previously unattended AD insights towards real world drug development. In this thesis, I offer a perspective on the looming central issues of data reusability and limitation of current approaches that hinder the progress in research on AD therapeutics.

Taking into account my previously made statement: *"Technological and data resources required to determine links to diseases are pieces of the puzzle that when put together, promise to reveal novel regulators in pathomechanisms. — refer page 9"*, the objectives of my thesis work are summarised in Figure 2.1. The specific objectives of this thesis addressing the issues stated above and in Chapter 1 are:

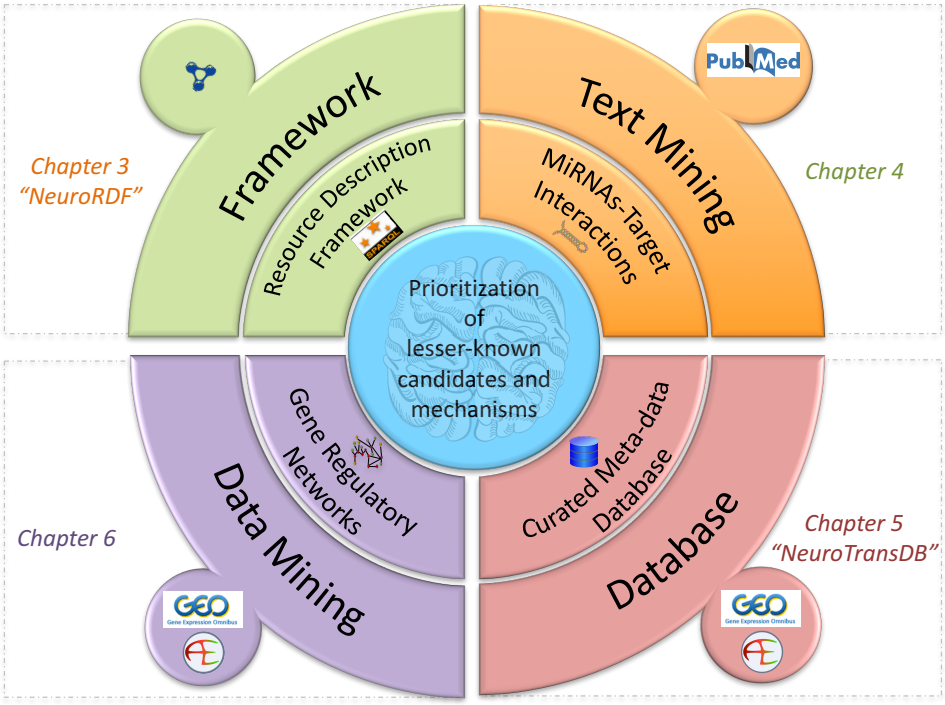


Figure 2.1: Objectives of my thesis work

1. **Problem Statement:** *Problems for data integration are rooted in the data itself; resources are broadly distributed, across the web and encompass heterogeneity and diversity in data formats, concepts, semantics, syntax, to name a few. Of course, these approaches have strengths and limitations and there is no "one-size fits all" solution [75,92]. Furthermore, these integration solutions are local and fail to operate at the global level and cannot cope with the updates of resources such as newly added information and changes in the data structure, formats, and naming convention. In general, most of these approaches overlook the importance of data quality, and context-specificity. — page 10*

   Objective: To build a high-quality and domain-specific pre-competitive infrastructure for intellectual integration of existing resources to facilitate interrogation of the distributed data legacy; enabling a systematic and objective prioritisation of molecular protagonists and mechanisms in AD

2. **Problem Statement:** *Many cellular functions — biological and pathological — are a result of cross-talk between different bio-entities, namely genes, proteins, transcription factors, and miRNAs. Therefore, to fully uncover the modular organization of the cellular networks it is crucial to elucidate these players. Although extensively researched, protein-coding genes represent only 2% of the*

human genome suggesting that PPIs are just half of the story in AD biology. With the beginning of miRNA era in 2001, non-coding RNAs have become attractive targets and research topic for novel diagnostic and therapeutic approaches [247]. Information about miRNA's regulatory roles has been widely discussed in the literature. Thus, utilising biomedical text-mining approaches to extract new evidence from existing literature has become very crucial to drive AD research. — page 25

Objective: To develop an automated text-mining method for extracting new interaction evidence from existing scientific literature using miRNA research domain as an example.

3. **Problem Statement:** *Indeed, to our knowledge none of the value-added databases capture NDD specific metadata information. In addition, they fail to explicate annotations that distinguish diseased human from NDD-induced immortalised cell lines, and mouse/rat strains, which are critical for translating preclinical studies to drug trials. Thus, there is a need for a dedicated approach to extract and refine metadata annotations catered to NDD research. — page 24*

   Objective: To develop a comprehensive and highly curated metadata database for public NDD gene-expression studies that allow precise selection of data subsets for meta-analysis and translational research

4. **Problem Statement:** *The current multi-omics-based GRN approaches have an intrinsic limitation of dependency on (i) known interactions catalogued in databases and literature for follow-up analysis (ii) well-known gene candidates to refine networks within its proximity and (iii) restricting inference on genes that exhibit a clear shift in expression behaviour. This means that the derived results have a tendency to converge to "what is already known", missing out on lesser-known candidates. Additionally, none of the above approaches elaborates on context-specificity and completeness of the generated networks, undermining the modules that approximate the biological truth. — page 22*

   Objective: To establish an approach for constructing a more reliable and complete large-scale AD GRNs that is not biased towards prior knowledge, but rather extends the scope to not so obvious players

2.2 Thesis organisation

The dissertation consists of four publications, fulfilling the above listed objectives, organised as individual chapters[40]:

RDF Framework for data integration In Chapter 3, an integrative approach based on RDF technology for modelling curated knowledge to prioritise potential AD candidates and mechanisms is introduced. This study utilises the data resources reported in Chapter 4, and Chapter 5.

Extracting new interaction evidence from literature — applied to miRNA domain In Chapter 4, the state-of-the-art text-mining methodology for extracting miRNA mentions and its relations from biomedical literature is reported. In addition, the generated benchmark corpus to evaluate this and previously reported similar studies is described.

Capturing metadata information from public transcriptomic studies In Chapter 5, systematic harvesting and curation of NDD-specific metadata from publically available transcriptomics studies is reported, which lead to the development of *NeuroTransDB* database.

Knowledge-instructed gene regulatory network construction The high-quality and NDD-specific data harvested in Chapter 5 can contribute to more comprehensive meta-analysis to derive new-novel insights in AD. Therefore, a new computational approach to construct more reliable GRNs across large-scale omics studies has been developed in Chapter 6. This approach is capable of identifying lesser-known players and mechanisms in AD.

Conclusion and outlook The key aspects and utilisation of the work presented in this dissertation are concluded in Chapter 7.

[40] Supplementary files, other than supplementary figures and tables, from the publications are not included in this thesis. Please refer to the publication's webpage for downloading these files.

2.3 List of all the publications

2.3.1 Thesis publications

1. **Shweta Bagewadi Kawalia¶**, Tamara Raschka[¶], Mufassra Naz, Ricardo de Matos Simoes, Martin Hofmann-Apitius, and Philipp Senger. *"Analytical strategy to prioritize Alzheimer's disease candidate genes in gene regulatory networks using public expression data."* **Journal of Alzheimer's Disease** 2017; 59(4) [IF 3.9]
2. Anandhi Iyappan¶, **Shweta Bagewadi Kawalia¶**, Tamara Raschka, Martin Hofmann-Apitius, and Philipp Senger. *"NeuroRDF: semantic integration of highly curated data to prioritize biomarker candidates in Alzheimer's disease."* **Journal of Biomed Semantics** 2016, 7:45 [IF 2.4]
3. **Shweta Bagewadi**, Subash Adhikari, Anjani Dhrangadhariya, Afroza Khanam Irin, Christian Ebeling, Aishwarya Alex Namasivayam, Matthew Page, Martin Hofmann-Apitius, and Philipp Senger. *"NeuroTransDB: Highly Curated and Structured Transcriptomic Meta-Data for Neurodegenerative Diseases."* **Database** 2015: bav099. [IF 3.9]
4. **Shweta Bagewadi**, Tamara Bobić, Martin Hofmann-Apitius, Juliane Fluck, and Roman Klinger. *"Detecting miRNA Mentions and Relations in Biomedical Literature."* **F1000Research** 2014, 3:205

2.3.2 Other publications

1. Jiali Wang, **Shweta Bagewadi Kawalia**, Mehdi Ali, Philipp Senger, Reinhard Schneider, Serge Haan, and Venkata P. Satagopam. *"miRSystec: an integrated web portal of human miRNA-target interaction."* **Frontiers in Genetics** 2017 (submitted) [IF 3.78]
2. Martin Hofmann-Apitius, Gordon Ball, Stephan Gebel, **Shweta Bagewadi**, Bernard de Bono, Reinhard Schneider, Matt Page, Alpha Tom Kodamullil, Erfan Younesi, Christian Ebeling, Jesper Tegnér and Luc Canard. *"Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in*

[¶] The authors contributed equally to this work

Neurodegenerative Disorders." **International Journal of Molecular Sciences** 2015, 16, 29179-29206. [IF 3.48]

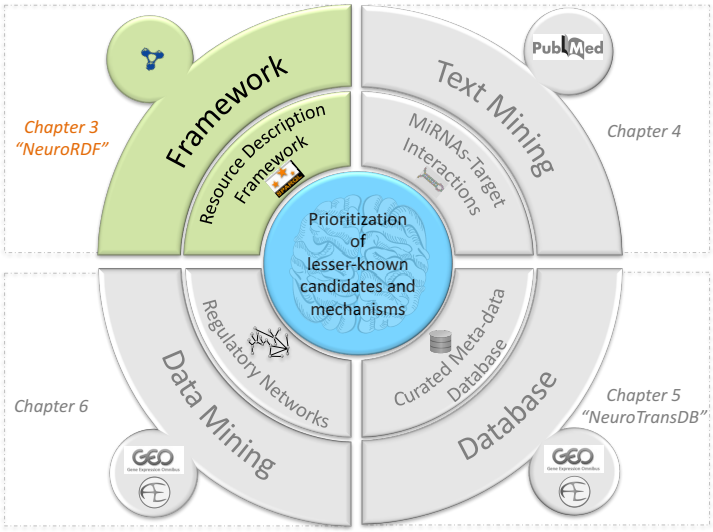
3. Avisek Deyati, **Shweta Bagewadi**, Philipp Senger, Martin Hofmann-Apitius, and Natalia Novac. *"Systems approach for the selection of micro-RNAs as therapeutic biomarkers of anti-EGFR monoclonal antibody treatment in colorectal cancer."* **Scientific Reports** 2015, 5:8013 [IF 4.8]
4. Alpha Tom Kodamullil, Erfan Younesi, Mufassra Naz, **Shweta Bagewadi**, and Martin Hofmann-Apitius. *"Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis."* **Alzheimer's & Dementia** 2015, ISSN 1552-5260 [IF 13.29]
5. Ashutosh Malhotra, Erfan Younesi, **Shweta Bagewadi**, and Martin Hofmann-Apitius. *"Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease."* **Genome Medicine** 2014, 6:97 [IF 7.07]
6. Anandhi Iyappan¶, **Shweta Bagewadi¶**, Matthew Page, Martin Hofmann-Apitius, and Philipp Senger. *"NeuroRDF: Semantic Data Integration Strategies for Modeling Neurodegenerative Diseases."* In Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (**SMBM2014**). Aveiro, Portugal.

2.3.3 Other published posters/abstracts

1. Mohammad Asif Emon, **Shweta Bagewadi Kawalia**, Aliaksandr Masny, Philipp Senger, Henri A. Vrooman and Martin Hofmann-Apitius. *"NeuroMap: Modeling of amyloid beta and tau spreading across brain circuitry in Alzheimer's disease using BEL language."* **The International Conference on Systems Biology**, Barcelona, Spain, 2016
2. Tamara Raschka, **Shweta Bagewadi, Mufassra Naz**, Martin Hofmann-Apitius, and Philipp Senger. *"Analytical strategy to unravel novel candidates from Alzheimer's disease gene regulatory networks using public transcriptomic studies."* 6th **International Conference on Genomics & Pharmacogenomics**, Berlin, Germany, 2016
3. Philipp Senger and **Shweta Bagewadi**. *"Automatic Quality Assessment Of Microarray Datasets Using Ensemble Methods."* The 13th European Conference on Computational Biology (**ECCB'14**), Strasbourg, France, 2014

4. **Shweta Bagewadi**, Erfan Younesi, Alpha Tom Kodamullil, and Martin Hofmann-Apitius. *"Identifying Unconventional Role of MiRNAs in Alzheimer's Disease Through Cause-and-Effect Model."* The 8th World Congress on Controversies in Neurology (**CONy**), Berlin, Germany, 2014

Chapter 3 Semantic-based Integrative Strategy for Candidate Prioritization and their Mechanistic Analysis



3.1 Introduction

Combining data and knowledge in a robust, scalable, shareable, and extensible framework is the fundamental need to understand the dynamics of neurodegenerative mechanisms. It increases the confidence of the derived hypotheses if the consensus is shown by different data resources. Indeed, such a framework should provide the capability to refine the search paths for dynamic exploration of the data without losing the underlying biological meaning. Disparate data types coming from literature, databases, imaging, omics experiments, GWAS studies, among others are an integral part of such a framework. However, assembling ever-growing information from several disciplines that represent complex and heterogeneous data is far from trivial. This is due to high variability in distribution, quality, representation, and applied statistical methods. If not addressed, these issues have a huge impact on the derived hypotheses and decisions made.

This publication presents an approach using RDF framework to facilitate representation and integration of heterogeneous AD data. It shows that well-curated and precise data enables novel biomarker and mechanism discovery. It additionally points out the need and effort of manual curation for precise modelling.

3.2 Publication

Iyappan *et al. Journal of Biomedical Semantics* (2016) 7:45
DOI 10.1186/s13326-016-0079-8

Journal of Biomedical Semantics

RESEARCH Open Access

NeuroRDF: semantic integration of highly curated data to prioritize biomarker candidates in Alzheimer's disease

Anandhi Iyappan[1,2†], Shweta Bagewadi Kawalia[1,2*†], Tamara Raschka[1,3], Martin Hofmann-Apitius[1,2] and Philipp Senger[1]

Abstract

Background: Neurodegenerative diseases are incurable and debilitating indications with huge social and economic impact, where much is still to be learnt about the underlying molecular events. Mechanistic disease models could offer a knowledge framework to help decipher the complex interactions that occur at molecular and cellular levels. This motivates the need for the development of an approach integrating highly curated and heterogeneous data into a disease model of different regulatory data layers. Although several disease models exist, they often do not consider the quality of underlying data. Moreover, even with the current advancements in semantic web technology, we still do not have cure for complex diseases like Alzheimer's disease. One of the key reasons accountable for this could be the increasing gap between generated data and the derived knowledge.

Results: In this paper, we describe an approach, called as *NeuroRDF*, to develop an integrative framework for modeling curated knowledge in the area of complex neurodegenerative diseases. The core of this strategy lies in the usage of well curated and context specific data for integration into one single semantic web-based framework, RDF. This increases the probability of the derived knowledge to be novel and reliable in a specific disease context. This infrastructure integrates highly curated data from databases (Bind, IntAct, etc.), literature (PubMed), and gene expression resources (such as GEO and ArrayExpress). We illustrate the effectiveness of our approach by asking real-world biomedical questions that link these resources to prioritize the plausible biomarker candidates. Among the 13 prioritized candidate genes, we identified MIF to be a potential emerging candidate due to its role as a pro-inflammatory cytokine. We additionally report on the effort and challenges faced during generation of such an indication-specific knowledge base comprising of curated and quality-controlled data.

Conclusion: Although many alternative approaches have been proposed and practiced for modeling diseases, the semantic web technology is a flexible and well established solution for harmonized aggregation. The benefit of this work, to use high quality and context specific data, becomes apparent in speculating previously unattended biomarker candidates around a well-known mechanism, further leveraged for experimental investigations.

Keywords: RDF, Semantic web, Data integration, Data curation, Data harmonization, Disease modeling, Neurodegenerative diseases, Alzheimer's disease

* Correspondence: shweta.bagewadi@scai.fraunhofer.de
†Equal contributors
[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2]Bonn-Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, 53113 Bonn, Germany
Full list of author information is available at the end of the article

Background

Alzheimer's disease (AD), the most prominent neurodegenerative disease (NDD), has become a global threat to the aging society, affecting nearly 115 million people by 2050 [1]. The imperfect understanding of the AD etiology has created a large gap in translating the preclinical findings into clinical trials dominantly observed in high drug attrition rates [2]. Early diagnosis and preventive interventions could facilitate substantial reduction in the number of affected cases to 9 million by 2050 [3, 4]. Particularly, reliable biological markers of disease and disease progression could assist in early diagnosis and treatments catered to the patient [5]. In this direction, considerable global research efforts have been dedicated to investigate molecular players underlying AD pathogenic events, contributing to an ever-growing wealth of disparate data. Refinement of this information into actionable knowledge representations requires a good interoperable and formalized framework, capable of inferring potential biomarkers across different facets of the molecular physiology. Additionally, in silico disease models that integrate complementary data from various resources are capable of recapitulating key mechanisms for a given condition [6–8].

Among others, most widely used data integration strategies include data warehousing (e. g., Pathway Commons [9]), data centralization (e. g., UniProt [10], IntAct [11]), and federated databases (e. g., BioMart [12]). An example of a data integration framework is tranSMART [13], which consists of a data warehouse covering various types of data and related data mining applications required for translational research and biomarker discovery workflows. Such a harmonized aggregation of heterogeneous data sources facilitates interpretation over a large knowledge space [14].

However, one fundamental challenge with most of these integration approaches is to cope with the variability and heterogeneity in content, language, and formats of incoming data from different source repositories. Moreover, regular updates of these data resources are necessary to keep up with newly added information and to avoid incompleteness. The inaccessibility to the integrated data resources, due to altered database structure or change in the naming conventions is unavoidable [15]. Semantic web technologies have overcome the above described challenges up to an extent by revolutionizing the lossless exchange of data and formalizing the data format into a computable knowledge [16], calling it “smart data" [17]. The capability of using rich formal descriptions for data and its standardized mapping allows complex querying in a more efficient way without information loss.

Resource Description Framework (RDF) is the World Wide Web Consortium (W3C) proposed standard for semantic integration and modeling of data. RDF uses the syntax of Extensible Markup Language (XML) and imposes structural constraints to represent the meta-data as a set of triples containing directed edges. One big advantage lies in the usage of common namespaces across the different data domains encoded as Unified Resource Identifiers (URIs). Initiatives such as Identifiers.org [18] provide persistent official identifiers in the biomedical domain, allowing sustained interlinking between distinct data resources. This allows high levels of seamless interoperability between data sources and the capability to access and map against additional related data unambiguously, called data federation. On the contrary, large efforts are still needed during an initial definition of the ontologies to build the schema for data representation.

Semantics in life sciences

The idea of semantic web prevails in various domains, including life sciences. Recently, "The Monarch Initiative" [19] has taken the semantic route to enable reasoning over genotype-phenotype equivalence within and across species. They leverage on ontologies to link external curated data resources for generating new hypothesis and prioritizing candidates/variants based on the phenotypic similarity. Stevens et al. [20] launched TAMBIS, multi-data application tool, which allows biologists to formulate complex molecular biology questions to databases such as Swiss-Prot [21], Enzyme [22], CATH [23], BLAST [24], and Prosite [25] through well-defined semantics.

Among the early users of RDF, Lindemann et al. [26] applied it to centralize and flexibly access the heterogeneous and varying quality of medical data obtained from several clinical partners. The importance of semantic mining in the life science domain was brought to limelight by the Bio2RDF project [27], which demonstrated the possibility of querying life science knowledgebases by linking public bioinformatics databases and providing public SPARQL endpoints. Subsequently, Linking Open Drug Data (LODD) [16] demonstrated linking drug data information from DrugBank [28] and clinical trials resources. Chem2Bio2RDF [29] demonstrates the potential usage of the above two mentioned RDF repositories in the field of chemoinformatics.

Observing the immense advantage of linked open data, several major publicly available life science databases such as UniProt, DisGeNet [30], Protein Data Bank Japan (PDBj) [17], and EBI resources such as Gene Expression Atlas [31], ChEMBL [32], BioModels [33], Reactome [34], and BioSamples [35], have made their data available in the form of RDF. Thus, the RDF platform has been increasingly adopted as a standard for data exchange. Amidst prime users of RDF in elucidating disease pathophysiology, Shin et al. [36] demonstrated systematic querying of linked experimental data to

explore the effect of genes that are regulated by volatile organic compounds in human blood. Qu et al. [6] showed semantic framework capability in drug repurposing by proposing Tamoxifen, an FDA approved drug for Breast Cancer, as a candidate drug for Systemic Lupus Erythematosus. The above reported association has already been tested in mice by Sthoeger et al. [37], showing a leverage of semantic web in a real world scenario. Furthermore, Willighagen et al. [38] presented the linkage of several RDF technologies in molecular chemoinformatics and proteochemometrics.

To our knowledge, there has been very limited application of semantic web approaches to the research of neurodegenerative diseases. Linked Brain Data (LBD) [39] is an upcoming initiative which focusses on understanding the brain functionality by integrating resources such as genomic, proteomic, anatomical and biochemical resources with respect to neuroscience. Using such a multi-level knowledgebase, they aim to understand the association between cognitive functions and brain diseases. Lam et al. [40] made the first attempt to develop an e-Neuroscience data integration framework, AlzPharm [41]. They extracted AD-related drug information from BrainPharm [42] to be further integrated with manually inferred hypotheses from the scientific literature and published articles (SWAN [43]). They demonstrated the usage of such a model by clustering AD drugs based on their molecular targets and to filter publications (claims and hypotheses) specific to Donepezil effect on treatment of AD. Although AlzPharm made use of manually inferred hypothesis, they lack the validation of their findings with experimental data such as gene expression and pathways.

Motivation

Despite the current advancements in semantic web technology, we still do not have cure for complex diseases like AD. One of the key reasons accountable for this could be the increasing gap between generated data and the derived knowledge. In order to increase the probability of the derived knowledge to be novel, data quality and data reliability is highly desirable. Moreover, the contextual specificity of the data is of paramount importance.

Compared to relational database management system (RDBMS) technologies, in RDF the relations have explicit meaning (expressiveness) in a given context and are directly accessible; allowing the user to extract meaningful knowledge from the data as opposed to an unstated structured data. In addition, RDF structures are more adaptive and flexible, allowing fluidity in the data relationships. This overcomes the fragility of RDBMS; where if the underlying representation of the keys and flat table changes, the tentacled connections are lost. Moreover, triples from RDF can be transformed into RDBMS structure and vice-versa. One another advantage of RDF is its graph representation that enables us to better explore relations through network topological characteristics such as relatedness, network perturbation, centrality, influence, etc. The usage of automated reasoners have largely been beneficial to understand the semantics and to expand the associated relations [44].

In this paper we propose *NeuroRDF*, an approach harnessing the potential of RDF as a framework for modeling neurodegenerative diseases to enable a close, biologically sensitive integration of well curated, complementary, and multi-faceted data. The fundamental principle of this strategy is to take advantage of semantics to develop a context specific, multi-layered in silico disease model, represented as a formalized, and computationally processable domain knowledge. A fine-grained analysis of the metadata from various data resources empowers the user to ask more focused questions around a hypothesized pathomechanism involving previously neglected or hidden candidates, further leveraged for experimental investigations. Considerable efforts have been invested to process and manually curate huge amounts of data that is required to build such a knowledge base around a specific indication. This includes for example the in-depth assessment of the respective phenotype, the type of tissue used in an experiment, and information around the donor of the tissue like gender, age, and possible comorbidities. Querying such a highly curated and focused knowledgebase increases the chances of unraveling novel hypothesis, which could have been lost over time or pave way to newly emerging knowledge.

We used SPARQL to traverse each of these knowledge graphs (derived from distinct resources) in an integrative manner, allowing highly disease specific analysis of the underlying data. Using this approach, we demonstrate an example on how to prioritize novel candidates in AD mechanism.

Methods

The developed generic semantic web-based workflow integrating heterogeneous data resources is outlined in Fig. 1. This multi-layered model integrates data from various public resources such as databases, literature, and gene expression information. The harmonization of heterogeneous data to build RDF models was achieved by using several data/file parsers. The workflow also includes a pre-processing step to monitor the quality of each incoming data type for specificity.

Data collection and resources

This subsection depicts briefly the different data resources integrated into the *NeuroRDF*.

Database-derived interactions for healthy brain

A closer look into the healthy human brain interactions could improve identification of the dysregulated

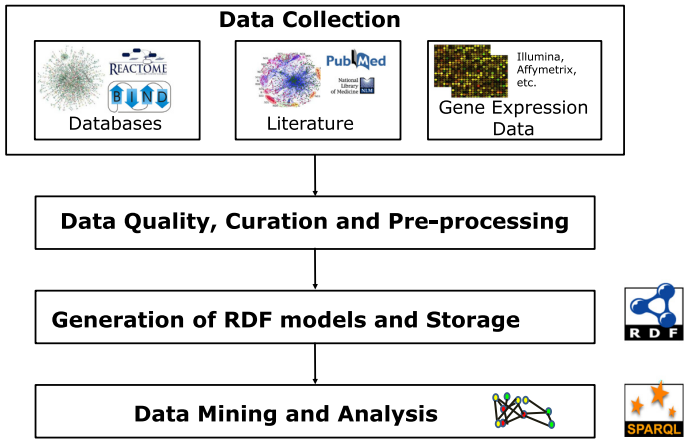


Fig. 1 Overall workflow of NeuroRDF. The workflow illustrates the collection of data from various resources such as databases, and literature, followed by steps taken to pre-process and prune the collected data. These high-quality data are represented semantically as RDF models and stored in a triplestore. The stored knowledge can later be queried for biologically interesting questions

mechanisms which further surges the plausibility of identifying AD drugs in clinical trails [45, 46]. However, the mainstream AD research is biased towards the well known disrupted events such as APP, and tau rather than recognizing their role in normal brain functions [47].

Several publicly available databases provide protein-protein interactions (PPIs) and microRNA-target interactions (MTIs), which can be derived using multiple sources and methodologies. For instance, Human Protein Reference Database [48], Molecular INTeraction database [49], and miRTarBase [50] focus on experimentally verified interactions that are manually mined from literature by expert biologists. In addition to literature-derived information, Biomolecular Interaction Network Database [51] centralizes interactions from high-throughput technologies. Few other databases such as STRING [52], and miRWalk [53] also provide predicted interactions. However, none of these databases mine interactions specific to a given context (for example AD pathology or normal physiology).

A lot of published healthy state PPIs are not directly measured in human cells but in artificial conditions such as human cell lines, human genes transfected into yeast cells, etc., missing out on the biological plausibility in humans and context specificity [54]. Hence, considerable effort by Bossi and Lehner [55] was invested to verify the tissue specificity of PPI interactions from 21 databases (including a few above mentioned) using human gene expression data. Furthermore, this additional action to ensure validity of the interactions in normal state aids improved prediction of genes in disease state [56]. In that direction, our group has extracted a subset of these experimentally confirmed PPIs belonging to healthy brain physiology [57]. Currently, the healthy brain PPI network contains 7,192 genes and 45,001 PPIs.

Extracting AD-specific interactions from literature

The bridging factor between researchers and scientific accomplishments are published as texts, warehoused in large repositories like PubMed [58]. These biomedical articles are the major information source of functional factors such as proteins, genes, microRNAs (miRNAs), etc. However, their functional descriptions are scattered as unstructured text in literature [59]. Text-mining methods could help us mine these articles and retrieve the associated relations/evidence for a given context. Since proteins are the chief players in almost all biological processes and miRNAs have been established in the last decade as important regulators of gene expression, we focus our current research on MTIs and PPIs.

In order to harvest AD-specific knowledge from the literature, we used our in-house state-of-the-art named entity recognition (NER) system ProMiner [60] and the semantic search engine SCAIView [61]. Identification of genes/proteins and disease mentions was accomplished using dictionaries. The disease dictionary was built using MeSH [62], MedDRA [63], and Allie [64] databases. Currently, it contains 4,729 concepts and 64,776 synonyms [65], which are normalized to MeSH names. Human Genes/Proteins dictionary [60] was compiled from three different resources: SwissProt, EntrezGene [66], and HGNC [67]. Currently, this dictionary consists of 36,312 entries and 515,191 synonyms. All the identified gene/protein names were normalized to HUGO gene symbols for maintaining homogeneity across all

data resources and also for future comparisons and visualizations.

To identify MTIs from MEDLINE abstracts, we applied our previously developed approach [65]. Here we extracted novel miRNA mentions using a regular expression. These mentions were normalized to miRBase database identifiers [68]. In addition, relation dictionary containing the major classes of relationship terms between miRNAs and their target genes/proteins was also developed. A tri-occurrence based approach was used to extract the MTIs (co-occurring with a relation term) at the sentence level.

Using the above-mentioned dictionaries, our group previously harvested AD specific PPIs from MEDLINE abstracts and full text articles [69]. Here we used the interaction terms compiled by Thomas et al. [70]. A state-of-the-art machine learning based approach [71] was applied to retain true pairs of PPIs in a given sentence. Both approaches have been optimized for recall. Hence, the obtained relations have been manually filtered for false positives. After manual inspection, 339 PPIs for 301 proteins and 99 MTIs for 36 microRNAs that are specific to AD were obtained. Articles published in languages other than English could lead to increased information content, however a dedicated approach to harvest them is needed. Moreover, separate parsers are needed. Thus, for this work we extracted interactions from the biomedical literature in English.

AD gene expression data

A standard approach to test any generated hypothesis is to assess the gene expression of the involved candidates between affected and healthy patients or in the absence of human data we fall back to animal models or derived cell cultures [72–75]. High-throughput technologies such as microarray, RNA-seq provide potential to measure gene expression simultaneously for different experimental/biological conditions. These studies are assembled in widely adopted public archives: The NCBI Gene Expression Omnibus (GEO) [76] and ArrayExpress [77].

For querying AD-specific gene expression data, we used previously developed database, NeuroTransDB [78], which contains highly curated meta-data information for eligible AD studies. It assembles studies from public resources namely, GEO and ArrayExpress, using a keyword based search approach. Among the 45 prioritized AD human studies, we filtered for microarray studies that measure gene expression in brain tissue extracted from both AD and healthy patients. In addition, availability of raw data was a mandate for applying uniform pre-processing. In total, we obtained eight microarray studies to be integrated in *NeuroRDF*: GSE12685, GSE1297, GSE28146, GSE5281, E-MEXP-2280, GSE44768, GSE44770, and GSE44771.

To assess the quality of the arrays we applied ArrayQualityMetrics [79] package. The selected studies (independent of the platform type) were pre-processed using Bioconductor (Version 3.0) packages in R [80], by applying similar methods for maintaining consistency by reducing variance. All studies conducted on Affymetrix chips were normalized by robust multi-array average method (*rma*) [81]. Similarly, package *limma* [82] was applied on Rosetta/Merck Human 44 k 1.1 microarray chip. All the chips were normalized for background correction and quantile normalization. The normalized intensity values were log2-transformed and duplicate probes were averaged. To identify the differentially expressed genes between healthy and Alzheimer's patients we used *limma* package by applying Benjamini and Hochberg's method to control for false discovery rate (adjusted p-value ≤ 0.05).

Data curation

Although the current text-mining methods have started to leverage expert curators to extract PPIs, MTIs, etc. from text, the extracted information are still prone to false positives [83]. Moreover, it is not straightforward to use these systems for retrieval of context-specific triples due to technological limitations [84]. Hence, the meticulousness of the identified triples to occur in a certain cell type, disease state, or events captured in AD-specific documents is not guaranteed. Thus, the need for manual verification is unavoidable, especially when considering the full text articles. The previously published test corpus used for evaluating the constructed AD PPI network contained AD-specific PPIs extracted by the machine learning approach from 200 full text articles [69]. Manual inspection by the authors resulted in retaining PPIs from 38 articles that are truly specific to AD, thus discarding 81 % of the originally retrieved articles. Similarly, we retained only 68 abstracts from 250 articles (27 %) that were retrieved using our tri-occurrence based approach for AD MTIs [65]. Thus, we can conclude that only about 20–30 % of the (relation extraction based) extracted PPIs and MTIs are truly relevant to AD, pointing out the need of manual curation.

Similarly, in our recent publication [78], we have highlighted the key issues related to retrieval and reusability of the datasets from public transcriptomics archives, such as GEO and ArrayExpress. We showed that a simple keyword based search not necessarily asserts the specificity of the retrieved datasets to the queried disease or organism. When manually inspected, we reported nearly 20 % of these retrieved studies to be irrelevant for AD query. In addition, basic metadata annotations such as age, gender, etc., which strongly contributes to the differential estimates, were observed to be incomplete. Brazma et al. [85] had earlier reported that not all the data submitted to GEO or ArrayExpress are MIAME compliant [86]. We additionally noticed these missing annotations being scattered as unstructured prose in database webpages,

publications, supplementary material, figures, etc., leading to a steep increase in the needed curation effort. Although the published research articles are rich in annotations, a large number of experiments have missing citations [87], which have to be added manually. Moreover, inconsistencies between the information stored in the archives and in the associated publications were also noted. On an average, about 30 min to 2 h of curation effort was needed to retrieve pertinent information for a single dataset. The outcome of this work resulted in a highly curated metadata database, NeuroTransDB, which is used in this work for extracting relevant AD gene expression studies.

Generation of RDF models

RDF data model

RDF allows the generation of models for processed data that exchanges information on the Web [82]. The RDF data model stores all the relationships between different entities as triples (subject-predicate-object). In RDF terminology, the subject, the predicate and the object are known as resources and are represented by a unique "Uniform Resource Identifiers (URIs)" in order to support global data exchange. Literals are constant values such as numbers and strings mapped to the resources. Literals can only be used as objects but never as subjects or predicates.

RDF schemas

We constructed the RDF schemata by abiding the standard RDF graph notation where an ellipse represents Resource, an arrow for Property, and rectangle for Literal. In all the RDF schemas, we have maintained a common resource representation for the "Gene" namespace adapted from Bio2RDF that maps to the NCBI gene database. For the namespaces with no available ontologies, we created an internal namespace, called "SCAI". Some of the properties were described using URIs from Dubin Core Metadata Element [88].

Four separate schemas (for each data resource) have been generated that are centered on genes for interoperability, associating each gene product to its official gene symbol. In the AD PPI schema (see Fig. 2), proteins and their interactions were represented using the Uniprot Core Ontology [89]. Supporting literature evidence were adapted to URIs from Bio2RDF namespaces. The article resource was linked to its PubMed ID, sentence in which the interaction has been mentioned, and the associated journal. Experimental evidence that validates the given interaction (if any) were mapped to BioPax [90], MGED [91], ONTOAD [92], and SCAI namespaces. In the MTI models (see Fig. 3), literature, genes, and proteins namespaces were adapted similarly to the PPIs. To represent the miRNAs, we applied the Bio2RDF namespace that links it to miRBase database [93].

For the PPI schema encoding the healthy state, as seen in Fig. 4, we used the same ontologies as in case of AD PPI. Additional interaction evidence such as brain region, reference database, experimental evidence, and literature information were described using Core, BioPax, and Bio2RDF namespaces.

The microarray schema has two branches that are linked to the experiment: sample details and differential expression analysis. The majority of the resources and properties are linked to URIs from EBI's Atlas (atlas) [94] and MGED [91] namespaces, cf. Fig. 5. Gene expression experiments could contain several samples that are measured in different conditions. A detailed description of

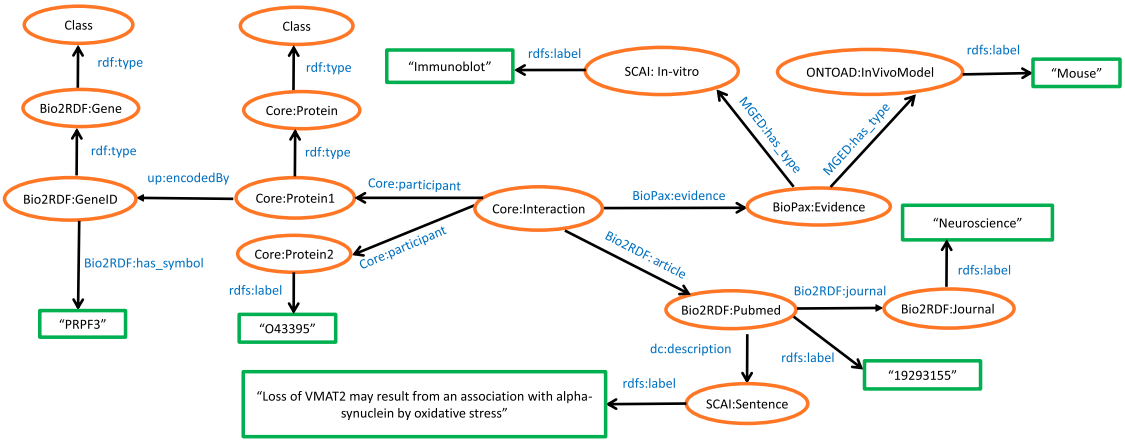


Fig. 2 Schematic representation of the Diseased PPIs in RDF. The figure describes AD specific PPI interactions along with supporting evidence mined from literature

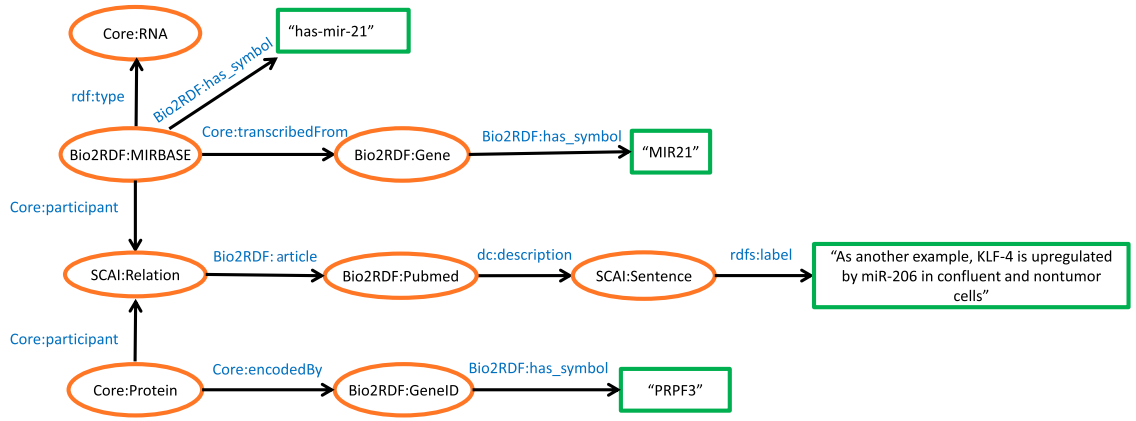


Fig. 3 Schematic representation of MiRNA-target interactions in RDF. The figure encapsulates miRNA mentions along with their corresponding gene identifier from literature

each sample is needed for accurate analysis. Thus, we associated each sample to its meta-data annotations, namely age, gender, organism, organism part, platform, and phenotype. Organism under investigation is mapped to NCBI Taxonomy URIs [95]. The factor value of each sample, i.e., the phenotype information, is described using the EFO ontology [96]. Each platform array is made up of multiple probes that may represent a gene. To be able to retain the expression values for individual probes, we linked the probe ID resource to platform. However, for better reasoning, quantitative values retrieved from statistical analysis are linked to genes and not to probes. The meta-analysis results, derived from *limma* [82], such as differential expression value of a gene and its associated p-value are all linked to the gene symbols.

Construction, validation and storage of RDF models

We modeled all the triples (represented in the schemas) using the Apache Jena API [97]. Resources, and Properties as Java classes were created from the ontologies using the corresponding in-built methods in the API and with the help of Schemagen [98].

In order to check for the correctness of our generated RDF models, we made use of the online service RDF validator [99]. By using such a service, we verified the models using their graph and triples representation.

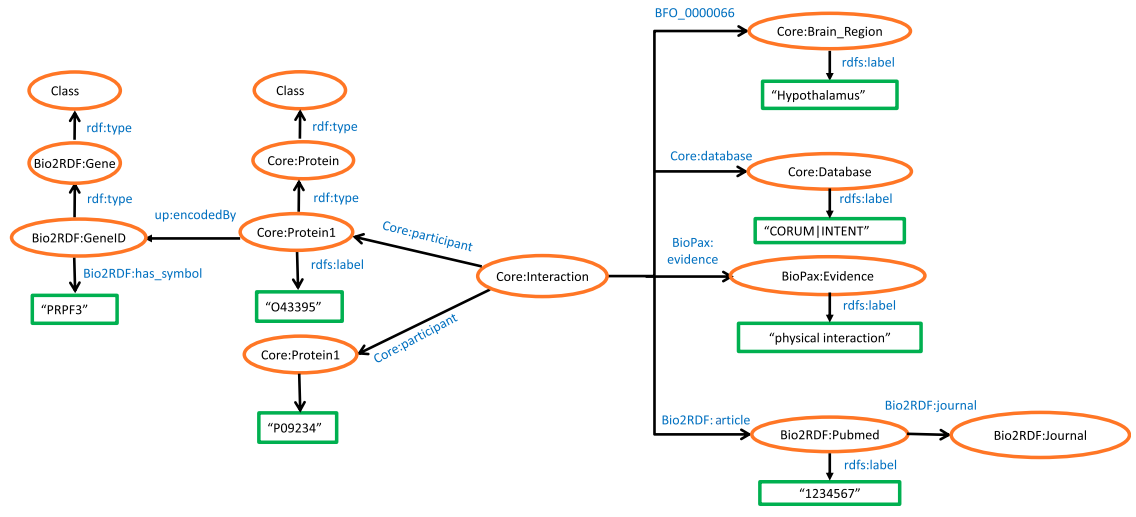


Fig. 4 Schematic representation of Healthy PPIs in RDF. The figure represents PPIs of healthy subjects extracted from literature and PPI specific databases. The schema also contains meta-information about these PPIs

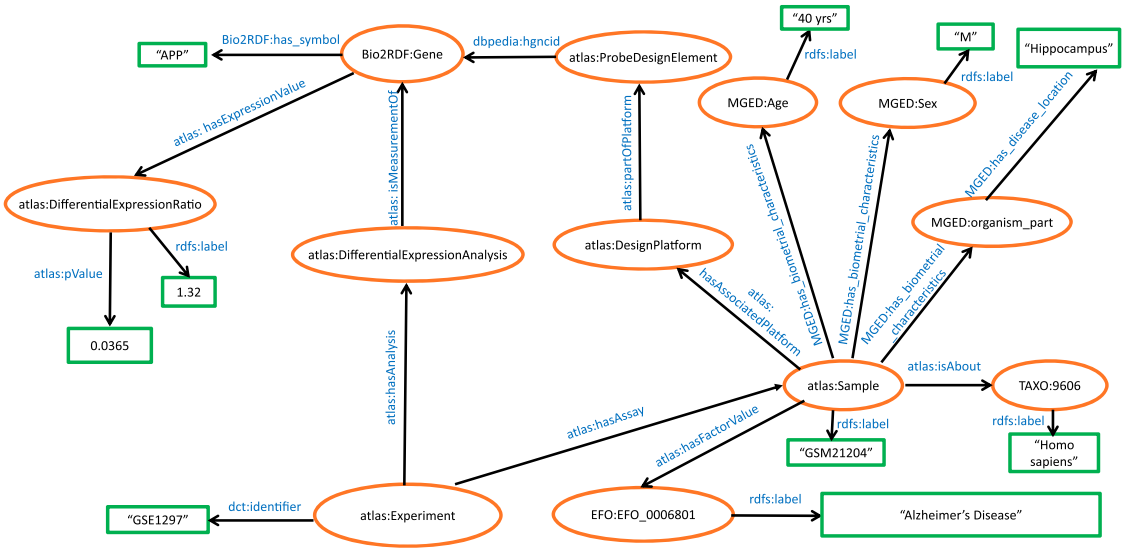


Fig. 5 Schematic representation of Gene Expression Data in RDF. This figure represents gene expression data obtained from public resources such as GEO and ArrayExpress

Triple stores, such as Virtuoso [100], provides an opportunity to store individual or integrated RDF models in one endpoint. Taking advantage of this, we stored all the generated RDF models as individual graphs in a single Virtuoso instance. Using common URIs (e.g., "Gene" identifier) as the connecting link between these models, it is possible to traverse through them integratively.

Data mining and analysis

In RDF, all the stored triples are accessible using a common query language, SPARQL Protocol and RDF Query Language (SPARQL) [101]. We generated a Java library with embedded SPARQL queries to ask our endpoint and the underlying networks biologically relevant questions. Queries were generated from individual models, which were further integrated as nested queries to traverse different graphs. Each query uses the common Gene URI namespace (which is common across all models) to pass on the results used to the next nested query. One possibility to visualize the query results is the SemScape Cytoscape [102], to represent the return values as (sub-) graphs again.

Results and discussions

NeuroRDF covers a wide range of curated AD related data resources, stored as four separate RDF models in a single Virtuoso endpoint. It tries to address the main concepts (complementary) that contributes significantly to unraveling AD pathology.

Differentially expressed genes

For the eight selected microarray datasets, gene expression analysis was performed between healthy and diseased patients. Among these, GSE1297, GSE28146, and E-MEXP-2280 resulted in no differential genes for adjusted p-value cutoff 0.05. From the remaining studies, only genes that exhibited a log2 fold change of > 1.5 were selected for analysis. In total, GSE5281 resulted in 4,278 genes under p-value cutoff and 2 up-, and 48 down-regulated genes for the defined fold change cutoff. Similarly, GSE44770 provided 254 differentially expressed genes, among which 16 up- and 11 down-regulated were selected further. In case of GSE44771, we obtained 335 differential genes that contain 11 up and 11 down-regulated genes that show > 1.5 log2 fold change. For both, GSE12685 and GSE44768, we obtained 1 and 51 genes under the p-value cut-off. However, there were no genes that had log2 fold change of >1.5. The list of all the differentially expressed genes that were selected for further analysis is provided in Additional file 1.

RDF models

Table 1 summarizes the content of the generated triple store by providing some statistics of all integrated networks. In total, there are 8353 unique triples in AD PPI, 1,204,194, 667 unique triples in Healthy PPI, and 20,454 unique triples in gene expression RDF models (Additional file 2). The number of unique predicates (relations) for AD and healthy PPIs are 11, whereas for MTI there are 5 and the gene expression model

Table 1 Statistics of generated RDF models stored in Virtuoso endpoint

Models	No. of triples	No. of entries	No. of properties	Size (mb)
Alzheimer's disease PPI	8353	19900	11	0.894
Healthy State PPI	1204194	78852	11	99.102
MTI	667	300	5	0.095
Microarray	20454	9477	16	303.5

consists of 16. The number of entities present in these models range from 300 to 78,852 (cf. Table 1). In case of the gene expression data, to avoid large triples we excluded the gene expression values of individual probes and included information only related to differential expression. Uploading and querying these models was not computationally expensive due to lower set of predicates and relatively small file size.

Prioritization of AD candidates

To illustrate the potential of NeuroRDF approach and to determine novel AD candidates from the high quality integrated data, we exploit the underlying biological association between the different data resources and identify the previously unknown information.

Our prioritization criteria was based on the notion that every data resource brings with it a piece of missing biological information which is needed to understand the mechanism of a certain candidate. We tried to associate this distributed information by systematically addressing the following questions:

(1) Whether candidates in the diseased network tend to be associated with normal physiology. If yes, what are the common players that could help us in the differential estimates (called as causal candidates);
(2) Which microRNAs regulate the selected causal candidates that could give insights into their post-transcriptional dysregulation;
(3) Have any of the selected causal candidates assessed for their level of differential expression in an unbiased data source (e. g., gene expression data);
(4) How strong is the influence of the neighboring genes on the casual candidates. This is based on the assumption that strong candidates tend be surrounded by dysregulated genes and have an influence on the candidate itself;
(5) Is there any functional relatedness between the causal candidates and their neighbors;

To answer these questions, we generated a set of SPARQL queries. Figure 6 is an example SPARQL query syntax used to obtain miRNAs that regulate the genes in the AD networks. Similar querying has been applied to build a system of faceted searches for the above described questions. Firstly, we identified common genes between the healthy and AD PPI networks. This query resulted in 230 intersecting genes. Looking into the MTIs, we found 13 of these genes to be regulated by at least one microRNA (cf. Table 2). Among these 13 genes, 9 were observed to be differentially expressed: APP, BACE1, ADAM10, IL1B, MAPK3, DLG4, LRP1 PTGS2, and TGFB1. Except for APLP2, and IL6, all the other genes contained differentially expressed neighbors either in AD or in healthy PPIs. There were no miRNAs that were common to these 13 genes.

Sub-networks from the AD and healthy PPIs were extracted to investigate the prioritized candidates (see Figs. 7 and 8). As observed from Fig. 8, for healthy PPIs there was one larger sub-network (containing APP, ADAM10, BACE1, MIF, MAPT, and LRP1) and a smaller one containing two genes (PTGS2, and IL1B). On the other hand, for diseased PPIs in Fig. 7, there were two large sub-networks containing four (STAT4, JUN, MAPK3, and STMN2) and five genes (APP, LRP1, BACE1, DLG4, and TGFB1). The third sub-network was made up of two genes (MAPT, and TUBA4A). Among the prioritized candidates, APLP2 and IL6 had no common links to other prioritized candidates. Thus, they were discarded for further analysis.

Relevance of prioritized AD candidates

The remarkability of complementing wet lab research using the predictability and reproducibility of measured outcomes is one of the core reasons why researchers are

```
SELECT ?Gene ?Rel ?Mirna ?Gene2
    from <http://localhost:8890/MiRNA>
    where {
        ?Gene <http://purl.uniprot.org/core/encodedBy> ?Protein .
        ?Protein <http://purl.uniprot.org/core/participant> ?Rel .
        ?Mirna <http://purl.uniprot.org/core/participant> ?Rel .
}
```

Fig. 6 Example SPARQL query for information retrieval from NeuroRDF. SPARQL query as seen in the figure retrieves the miRNAs for a given gene

Table 2 Prioritized AD candidate genes

Intersected genes between healthy and AD PPI	MiRNAs	Differentially expressed neighbors		Number of literature articles for intersected genes
		Healthy PPI	AD PPI	
APP	MIR101-1,	ADAM10,	TGFB1,	
	MIR106A,	MAPT,	BACE1,	
	MIR106B,	MIF,	LRP1	
	MIR124-1,	BACE1,		24550
	MIR137,	LRP1		
	MIR153-1,			
	MIR181-C,			
	MIR29A,			
	MIR520C,			
	MIR19-1			
BACE1	MIR107,			
	MIR124-1,		APP,	
	MIR145,	APP	LRP1	1883
	MIR298,			
	MIR29A,			
	MIR29B1,			
	MIR328,			
	MIR9-1			
ADAM10	MIR451,			
	MIR144,			
	MIR1306,	APP	-	231
	MIR107,			
	MIR103			
IL1B	MIR146A,			
	MIR155	PTGS2	-	1099
MAPK3	MIR15A,	-	STMN2,	276
	MIR155		JUN	
MAPT	MIR16-1,	APP	TUBA4A	3367
	MIR132			
APLP2	MIR153-1	-	-	134
DLG4	MIR485	-	LRP1	151
IL6	MIR27B	-	-	748
JUN	MIR144	-	STAT4,	142
			MAPK3	
LRP1	MIR205	APP	DLG4,	305
			APP,	
			BACE1	
PTGS2	MIR146A	IL1B	-	474
TGFB1	MIR155	-	APP	276

This table summarizes the literature based evidences of intersected genes between healthy and AD PPI and their corresponding miRNAs and differentially expressed genes

more inclined to the field of bioinformatics. Therefore, in silico validation of predicted candidates for its relevancy is of utmost importance. In this direction, we pinpoint the relevance of our prioritized candidates through a literature survey.

AD established candidates

Although there are no FDA approved biomarkers for AD, researchers focus on some of the key candidates that are hypothesized to be involved in AD. In the current NDD research practice, APP has been established as the widely used biomarker candidate. The classical pathological hallmark of AD is formation of amyloid-beta aggregates (leading to plaques) in brain. This is reported to be caused by faulty proteolytic processing of APP that releases amyloid-beta [103]. Another hallmark of AD is tau pathology (MAPT gene), regulated by amyloid-beta. Hyperphosphorylation of tau causes accumulation of neurofibrillary tangles due to the disrupted functioning of axonal transport [5]. However, it is also interesting to note the paradigm shift in AD research due to recently failed drug trails that focused mostly around these hypotheses [2]. Nevertheless, several neuroscientists still believe in the potential of APP and the tau hypothesis for elucidation of the underlying pathomechanism. As observed from our generated sub-networks, our largest sub-network was established around the APP gene.

When compared to APP, BACE1 has not been so frequently studied. However these genes often fall into the "most interesting gene zone" as far as AD is concerned since it is involved in the formation of amyloid-beta. BACE1 is the major enzyme (beta secretase) involved in the cleaving of APP at beta site and generating soluble amyloid-beta [104]. However, increased BACE1 activity has been reported to be associated with amyloid-beta aggregation in AD patients [105]. Bu et al. have detailed out the evidence that LRP1 is a receptor for APOE, a contributing factor to AD [106]. Furthermore, in 1993, Strittmatter, Roses and colleagues [107] have identified APOE4 as the major risk for late-onset AD. TGFB1 polymorphism has been widely associated with an increased risk of late-onset AD. Deficiency in TGFB1 signaling leads to neurofibrillary tangle formation increasing the advancement of mild cognitive impairment patients to AD, by increasing the depressive symptoms [108]. DLG4 is a post-synaptic scaffolding protein that interacts with postsynaptic receptors such as NMDA receptors for efficient postsynaptic response [109]. However, its impairment has largely contributed to the synaptic degeneration in AD. Mutations in ADAM10 gene have been associated to late-onset AD. ADAM10 enzyme has alpha-secretase activity to cleave amyloid-beta, however BACE1 competes with ADAM10 for cleavage. Thus, its decreased expression has been implicated in AD pathogenesis [110].

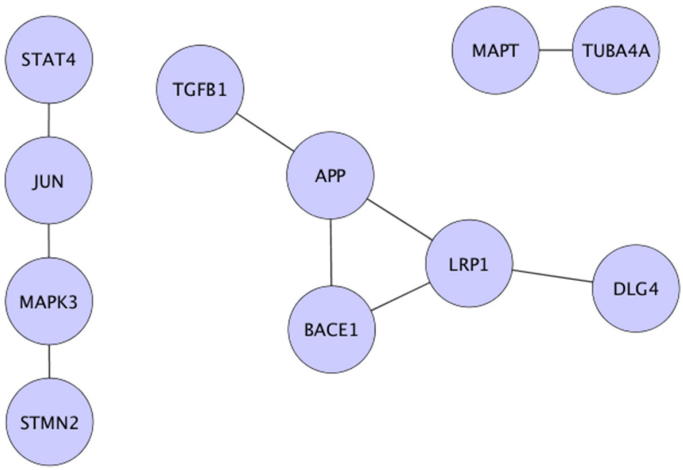


Fig. 7 Extracted sub-networks from AD PPIs network. This figure symbolizes the diseased sub-graphs that were generated using prioritized candidates and their differentially expressed neighbors

AD emerging candidates

To identify emerging knowledge in the context of AD, we performed an individual gene analysis using SCAIView for publications in PubMed. Here, we measured the co-occurrence of the causal genes (including its differential neighbors) and AD over a period of last 10 years, see Fig. 9. Since the number of articles for the APP gene was relatively too high each year, we normalized the number of literature evidence of other candidates using the APP gene's article count for that year. Hence, the normalized range for the literature distribution is between 0 and 1, where 1 is the highest number of articles for that year (the APP gene). Please refer to Additional file 3 for details of the literature counts. Inspecting literature evidence, we found that all the prioritized causal candidates have been studied in the context of AD. Moreover, among their differentially expressed neighbors, STMN2 (8 articles), MAPK4 (1 article), TUBA4A (2 articles), and MIF (15 articles) contained fewer articles related to AD. Among these genes, STMN2 and MIF have been recently studied in the context of AD, whereas, MAPK4, STMN2, and TUBA4A were implicated in AD nearly two decades before but failed to establish as robust biomarker candidates.

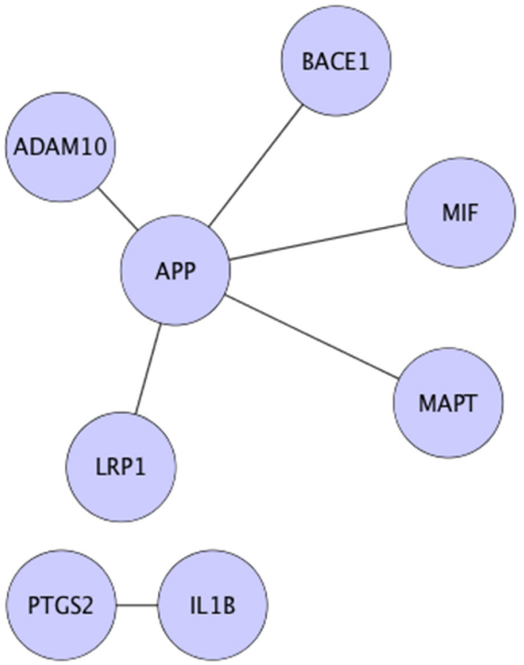


Fig. 8 Extracted sub-networks from healthy PPIs network. This figure symbolizes the healthy sub-graphs that were generated using prioritized candidates and their differentially expressed neighbors

MIF's role in AD

Macrophage Migration Inhibitory Factor (MIF) has for long been known to participate in tumor proliferation due to its pro-inflammatory cytokine functionality [111]. In general, MIF acts as a key regulator of inflammatory activities such as innate and adaptive immunity [112]. Apart from that, it is also known to play a significant role as an anti-apoptotic factor of neutrophils as well as macrophages [113].

The MIF gene has been well studied in cancer and inflammation. However, recent studies are emerging around a plausible role of MIF in neurodegenerative diseases, in particular AD. Moreover, Flex et al. [114] have earlier reported that MIF polymorphisms are not linked

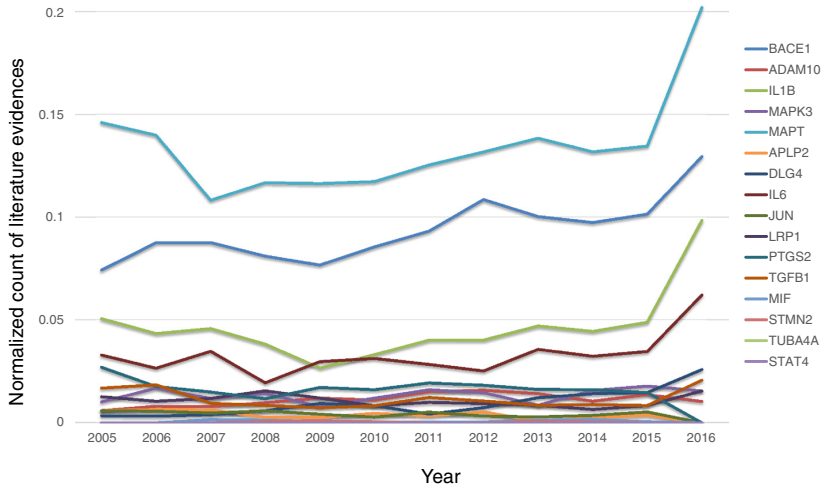


Fig. 9 Statistics of the literature evidence for emerging candidate genes in the speculated sub-networks

to AD, but confirmed its complex immune and inflammatory activities. Although, APP and tau have been associated to play a key role in the pathophysiology of AD, many researchers strongly believe in the role of inflammatory processes subsidizing to the pathology of AD. This stems from the fact that activated microglial cells discharge immunoregulatory cytokines which result in various side-effects such as neuronal dysfunction and inhibition of hippocampal neurogenesis [115]. MIF is one such pro-inflammatory cytokine which is known to bind with amyloid-beta protein and enhance the plaque removal and neuronal debris from the brain during normal conditions [116]. Also, MIF has been identified to play a role in neuronal survival by inhibiting the activation of ERK-1/MAP kinases [117] (regulatory role in cell proliferation and glucocorticoid action) as well as its ability to surpass the p53 mediated apoptosis [118]. Although, the precise molecular function of MIF in the context of AD is unknown, it is known to play a role in inflammatory processes around the plaque formation. MIF is also highly expressed in the neurons of rat hippocampus, one of the primary regions to be affected by AD [117]. Bryan et al. [119] also report on the abnormal expression of MIF in both microglia and in the hippocampal neurons in human. This all makes MIF a plausible biomarker for inflammatory responses in AD.

Conclusion

NeuroRDF approach has been designed to identify new knowledge through semantic mining. The proposed integrative approach takes advantage of the RDF technology to integrate well-curated data from various sources within a specific indication area. From our perspective, it is necessary to focus on one indication or at least a group of indications to build such a knowledge base for precise modeling and analysis due to the high curation effort one has to spend in order to reach the necessary details. We showed how to harmonize three major heterogeneous resources (databases, gene expression data, and literature) used in the research area to generate hypotheses for underlying disease mechanisms. This approach supports identification of novel insights without compromising over quality. Furthermore, new data resources can be included without altering the overall framework. The usage of well-accepted ontologies provides the advantage for further integration of external resources and databases (e.g., federated queries). Using such an approach, we were able to prioritize MIF gene as an emerging candidate due to its role in inflammatory processes implicated in AD pathogenesis.

The advantage of using an RDF schema is that it is highly supportive for data interoperability. Although this work represents the usage of the RDF schema specific for AD, we have also extended the same to other disease models such as Parkinson's and Epilepsy. However, the curated data and the generated hypothesis for these two diseases will be released in future under the terms of a Neuroallianz agreement [120]. Also, these resources are constantly kept up-to-date as they are transferred to various upcoming projects such as AETIONOMY [121].

Additional files

Additional file 1: List of differentially expressed genes. This file contains the list of differentially expressed genes (for each dataset used) that fall under the adjusted p-value cutoff of 0.05. The differential expression analysis was performed using *limma* package in R statistical environment. The file is provided in an Excel format. (XLSX 68 kb)

Additional file 2: The developed RDF models and the SPARQL queries used are made available at: http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/downloads/neurordf.html. (ZIP 178 kb)

Additional file 3: Detailed count of literature evidences for prioritized candidates. This file contains the detailed count of number of evidences available for each prioritized candidate for each year since 2005 in context of Alzheimer's disease. These statistics were retrieved using SCAIView knowledge discovery tool (as of 18 May, 2016). (XLSX 35 kb)

Acknowledgement
We are grateful to Matthew Page, Translational Bioinformatics, UCB Pharma for providing his valuable inputs during the design of the project and reviewing the manuscript. We are thankful to Erfan Younesi and Ashutosh Malhotra for providing the healthy state PPI and AD-PPI network respectively for this work. We also want to thank Christian Ebeling for his support in building the resources for gene expression data. We would like to acknowledge the Semantic Mining in Biomedicine (SMBM2014) conference organizers, participants, and reviewers for inspiring discussions during the conference. The authors express gratitude to the SMBM2014 conference organizers for providing an opportunity to submit to the Journal of Biomedical Semantics an extended version of the initially published conference proceeding paper.

Funding
This study was funded by a grant from the German Federal Ministry for Education and Research (BMBF) within the BioPharma initiative "Neuroallianz".

Authors' contributions
AI, SBK, PS, and MHA conceived and designed the overall research strategy required for data integration. PS is the scientific supervisor to this work. SBK and AI are the main contributors to manuscript writing. TR contributed to the analysis of gene expression data. PS, and MHA reviewed the content. All authors read and approved the final version of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Declarations
The underlying principles of this article have been previously published in Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM2014), Aveiro, Portugal, 2014.

Author details
[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany. [2]Bonn-Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, 53113 Bonn, Germany. [3]University of Applied Sciences Koblenz, RheinAhrCampus, Joseph-Rovan-Allee 2, 53424 Remagen, Germany.

Received: 1 March 2015 Accepted: 23 May 2016
Published online: 08 July 2016

References

1. International AD. Policy brief for heads of government: the global impact of dementia 2013–2050. 2013. http://www.alz.co.uk/research/G8-policy-brief.
2. Golde TE, Schneider LS, Koo EH. Anti-Aβ therapeutics in Alzheimer's disease: the need for a paradigm shift. Neuron. 2011;69:203–13. doi:10.1016/j.neuron.2011.01.002.
3. Brookmeyer R, Johnson E, Ziegler-Graham K, et al. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 2007;3:186–91. doi:10.1016/j.jalz.2007.04.381.
4. Norton S, Matthews FE, Barnes DE, et al. Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. Lancet Neurol. 2014;13:788–94. doi:10.1016/S1474-4422(14)70136-X.
5. Rachakonda V, Pan TH, Le WD. Biomarkers of neurodegenerative disorders: how good are they? Cell Res. 2004;14:347–58. doi:10.1038/sj.cr.7290235.
6. Qu XA, Gudivada RC, Jegga AG, et al. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. BMC Bioinf. 2009;10 Suppl 5:S4. doi:10.1186/1471-2105-10-S5-S4.
7. Le Masson G, Przedborski S, Abbott LF. A computational model of motor neuron degeneration. Neuron. 2014;83:1–14. doi:10.1016/j.neuron.2014.07.001.
8. Talwar P, Silla Y, Grover S, et al. Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. BMC Genomics. 2014;15:199. doi:10.1186/1471-2164-15-199.
9. Pathway Commons Database. http://www.pathwaycommons.org/about/. This and all the subsequent URLs have been accessed on 31 May 2016.
10. UniProt Database. http://www.uniprot.org/
11. IntAct Database. http://www.ebi.ac.uk/intact/
12. BioMart. http://www.biomart.org/
13. Szalma S, Koka V, Khasanova T, et al. Effective knowledge management in translational medicine. J Transl Med. 2010;8:68. doi:10.1186/1479-5876-8-68.
14. Rodriguez-Esteban R, Loging WT. Quantifying the complexity of medical research. Bioinformatics. 2013;29:2918–24. doi:10.1093/bioinformatics/btt505.
15. Aoki-Kinoshita KF, Kinjo AR, Morita M, et al. Implementation of linked data in the life sciences at BioHackathon 2011. J Biomed Semantics. 2015;6:3. doi:10.1186/2041-1480-6-3.
16. Samwald M, Jentzsch A, Bouton C, et al. Linked Open drug data for pharmaceutical research and development. J Cheminform. 2011;3:19. doi:10.1186/1758-2946-3-19.
17. Kinjo AR, Suzuki H, Yamashita R, et al. Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format. Nucleic Acids Res. 2012;40:453–60. doi:10.1093/nar/gkr811.
18. Identifiers.org. http://identifiers.org
19. The Monarch Initiative. http://monarchinitiative.org/page/about
20. Stevens R, Baker P, Bechhofer S, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. 2000;16:184–5. doi:10.1147/sj.402.0532.
21. Swiss-Prot Database. http://web.expasy.org/docs/
22. Enzyme Database. http://enzyme.expasy.org
23. CATH Database. http://www.cathdb.info
24. BLAST. http://blast.ncbi.nlm.nih.gov/Blast.cgi
25. Prosite Database. http://prosite.expasy.org
26. Lindemann G, Schmidt D, Schrader T, et al. The resource description framework (RDF) as a modern structure for medical data. Int J Biol Life Sci. 2008;4:89–92. http://waset.org/publications/3109/the-resource-description-framework-rdf-as-a-modern-structure-for-medical-data.
27. Belleau F, Nolin MA, Tourigny N, et al. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008;41:706–16. doi:10.1016/j.jbi.2008.03.004.
28. DrugBank Database. http://www.drugbank.ca
29. Chen B, Dong X, Jiao D, et al. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinf. 2010;11:255. doi:10.1186/1471-2105-11-255.
30. Furlong LI. DisGeNET : from MySQL to nanopublication, modelling gene-disease associations for the semantic Web. Paris: Proc 5th Int Work Semant Web Appl Tools Life Sci; 2012. *Fr Novemb 28–30, 2012 2012* Published Online First: 2012. http://ceur-ws.org/Vol-952.
31. Kapushesky M, Adamusiak T, Burdett T, et al. Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res. 2012;40:D1077–81. doi:10.1093/nar/gkr913.
32. CHEMBL Database. https://www.ebi.ac.uk/chembl/
33. BioModels Database. http://www.ebi.ac.uk/biomodels-main/
34. Reactome Ontology. http://www.reactome.org
35. BioSamples Database. http://www.ebi.ac.uk/biosamples/
36. Shin GH, Kang YK, Lee SH, et al. MRNA-centric semantic modeling for finding molecular signature of trace chemical in human blood. Mol Cell Toxicol. 2012;8:35–41. doi:10.1007/s13273-012-0005-9.
37. Sthoeger ZM, Zinger H, Mozes E. Beneficial effects of the anti-oestrogen tamoxifen on systemic lupus erythematosus of (NZBxNZW)F1 female mice are associated with specific reduction of IgG3 autoantibodies. Ann Rheum Dis. 2003;62:341–6. doi:10.1136/ard.62.4.341.
38. Willighagen EL, Alvarsson J, Andersson A, et al. Linking the resource description framework to cheminformatics and proteochemometrics. J Biomed Semantics. 2011;2 Suppl 1:S6. doi:10.1186/2041-1480-2-S1-S6.
39. Linked Brain Data. http://www.linked-neuron-data.org/about.jsp?link=link6
40. Lam HYK, Marenco L, Clark T, et al. Semantic Web Meets e-Neuroscience : An RDF Use Case, Semant Web - ASWC 2006 first Asian semant web conference. 2006. p. 158–70.
41. Lam HYK, Marenco L, Clark T, et al. AlzPharm: integration of neurodegeneration data using RDF. BMC Bioinf. 2007;8 Suppl 3:S4. doi:10.1186/1471-2105-8-S3-S4.

42. BrainPharm Database. http://senselab.med.yale.edu/BrainPharm
43. SWAN Ontology. http://www.w3.org/TR/hcls-swan
44. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform. 2015;16: 1069–80. doi:10.1093/bib/bbv011.
45. Douaud G, Refsum H, de Jager CA, et al. Preventing Alzheimer's disease-related gray matter atrophy by B-vitamin treatment. Proc Natl Acad Sci. 2013;110:9523–8. doi:10.1073/pnas.1301816110.
46. Tagawa K, Homma H, Saito A, et al. Comprehensive phosphoproteome analysis unravels the core signaling network that initiates the earliest synapse pathology in preclinical Alzheimer's disease brain. Hum Mol Genet. 2015;24:540–58. doi:10.1093/hmg/ddu475.
47. Kodamullil AT, Younesi E, Naz M, et al. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. Alzheimer's Dement. 2015;11:1329–39. doi:10.1016/j.jalz.2015.02.006.
48. Human Protein Reference Database (HPRD). http://www.hprd.org/
49. The Molecular INTeraction Database (MINT). http://mint.bio.uniroma2.it/mint/Welcome.do
50. Chou CH, Chang NW, Shrestha S, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 2015;5712121:gkv1258. doi:10.1093/nar/gkv1258.
51. Biomolecular Interaction Network Database (BIND). http://bioinformatics.ca/links_directory/database/9267/bind-biomolecular-interaction-network-database
52. STRING Database. http://string-db.org/
53. miRWalk Database. http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/
54. Schaefer MH, Lopes TJS, Mah N, et al. Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. PLoS Comput Biol. 2013;9, e1002860. doi:10.1371/journal.pcbi.1002860.
55. Bossi A, Lehner B. Tissue specificity and the human protein interaction network. Mol Syst Biol. 2009;5:260. doi:10.1038/msb.2009.17.
56. Magger O, Waldman YY, Ruppin E, et al. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. PLoS Comput Biol. 2012;8, e1002690. doi:10.1371/journal.pcbi.1002690.
57. Younesi E, Hofmann-Apitius M. Biomarker-guided translation of brain imaging into disease pathway models. Sci Rep. 2013;3:3375. doi:10.1038/srep03375.
58. PubMed Database. http://www.ncbi.nlm.nih.gov/pubmed
59. Krallinger M, Erhardt RA, et al. Text mining approaches in molecular biology and biomedicine. Drug Discov Today. 2005;10:439–45.
60. Fluck J, Mevissen HT, Dach H, et al. ProMiner: recognition of human gene and protein names using regularly updated dictionaries, Proceedings second BioCreative challenge evaluation work. Madrid: CNIO; 2007. p. 149–51.
61. SCAIView. http://www.scaiview.com/en/scaiview-distributions/scaiview-academia.html
62. National Library of Medicine's MeSH Controlled Vocabulary. http://www.ncbi.nlm.nih.gov/mesh
63. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf. 1999;20:109–17. doi:10.2165/00002018-199920020-00002.
64. Allie Database. http://allie.dbcls.jp/
65. Bagewadi S, Bobić T, Hofmann-Apitius M, et al. Detecting miRNA mentions and relations in biomedical literature. F1000 Res. 2014; doi: 10.12688/f1000research.4591.2
66. NCBI's Entrez Gene Database. http://www.ncbi.nlm.nih.gov/gene
67. HUGO Gene Nomenclature Committee (HGNC). http://www.genenames.org/
68. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73. doi:10.1093/nar/gkt1181.
69. Malhotra A, Younesi E, Sahadevan S, et al. Exploring novel mechanistic insights in Alzheimer's disease by assessing reliability of protein interactions. Sci Rep. 2015;5:13634. doi:10.1038/srep13634.
70. Thomas P, Solt I, Klinger R, et al. Learning to extract protein – protein interactions using distant supervision. In: Proceedings of robust unsupervised and semi-supervised methods in natural language processing, Workshop at international conference recent advances in natural language processing. 2012.
71. Bobić T, Klinger R, Thomas P, et al. Improving distantly supervised extraction of drug-drug and protein-protein interactions, Proc 13th Conf Eur Chapter Assoc Comput Linguist. 2012. p. 35–43.
72. Kogelman LJA, Cirera S, Zhernakova DV, et al. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. BMC Med Genomics. 2014;7:57. doi:10.1186/1755-8794-7-57.
73. Krämer A, Green J, Pollard J, et al. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics. 2014;30:523–30. doi:10.1093/bioinformatics/btt703.
74. Van Dam D, De Deyn PP. Animal models in the drug discovery pipeline for Alzheimer's disease. Br J Pharmacol. 2011;164:1285–300. doi:10.1111/j.1476-5381.2011.01299.x.
75. McDermott JE, Wang J, Mitchell H, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn. 2012;7:1–15. doi:10.1517/17530059.2012.718329.
76. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30:207–10. doi:10.1093/nar/30.1.207.
77. Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 2003;31: 68–71. doi:10.1093/nar/gkg091.
78. Bagewadi S, Adhikari S, Dhrangadhariya A, et al. NeuroTransDB : highly curated and structured transcriptomic metadata for neurodegenerative diseases. Database. 2015;2015:bav099. doi:10.1093/database/bav099.
79. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics - A bioconductor package for quality assessment of microarray data. Bioinformatics. 2009;25: 415–6. doi:10.1093/bioinformatics/btn647.
80. Bioconductor. http://www.bioconductor.org/
81. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4:249–64. doi:10.1093/biostatistics/4.2.249.
82. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47. doi:10.1093/nar/gkv007.
83. Czarnecki J, Shepherd AJ. Mining biological networks from full-text articles. Methods Mol Biol. 2014;1159:135–45. doi:10.1007/978-1-4939-0709-0_8.
84. Krallinger M, Vazquez M, Leitner F, et al. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinf. 2011;12:S3. doi:10.1186/1471-2105-12-S8-S3.
85. Brazma A. Minimum Information About a Microarray Experiment (MIAME)–successes, failures, challenges. Sci World J. 2009;9:420–3. doi:10.1100/tsw.2009.57.
86. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001;29:365–71. doi:10.1038/ng1201-365.
87. Piwowar H, Chapman W. Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers. J Biomed Discov Collab. 2010;5:7–20. doi:10.5210%2Fdisco.v5i0.2785.
88. Dublin Core Metadata Element Set. http://dublincore.org/documents/dces/
89. Uniprot Core Ontology. http://lov.okfn.org/dataset/lov/vocabs/uniprot
90. Biological Pathway Exchange (BioPax). http://www.biopax.org/
91. Whetzel PL, Parkinson H, Causton HC, et al. The MGED ontology: a resource for semantics-based description of microarray experiments. Bioinformatics. 2006;22:866–73. doi:10.1093/bioinformatics/btl005.
92. Ontology of Alzheimer's Diseases and Related Diseases (ONTOAD). http://bioportal.bioontology.org/ontologies/ONTOAD
93. The miRBase Database. http://www.mirbase.org/.
94. Atlas RDF Ontology. https://www.ebi.ac.uk/fgpt/ontologies/gxaterms.html.
95. NCBI Taxonomy Namespace. http://www.ncbi.nlm.nih.gov/taxonomy.
96. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an experimental factor ontology. Bioinformatics. 2010;26:1112–8. doi:10.1093/bioinformatics/btq099.
97. Jena Tutorial. https://jena.apache.org
98. Schemagen Documentation. http://jena.apache.org/documentation/tools/schemagen.html
99. RDF Validator. http://www.w3.org/RDF/Validator
100. Virtuoso. http://virtuoso.openlinksw.com
101. Sparql. http://www.w3.org/TR/rdf-sparql-query
102. Cytoscape Tool. http://apps.cytoscape.org/apps/semscape
103. Golde TE, Petrucelli L, Lewis J. Targeting Aβ and tau in Alzheimer's disease, an early interim report. Exp Neurol. 2010;223:252–66. doi:10.1016/j.expneurol.2009.07.035.
104. Cole SL, Vassar R. The Alzheimer's disease beta-secretase enzyme, BACE1. Mol Neurodegener. 2007;2:22. doi:10.1186/1750-1326-2-22.
105. Washington PM, Morffy N, Parsadanian M, et al. Experimental traumatic brain injury induces rapid aggregation and oligomerization of amyloid-beta in an Alzheimer's disease mouse model. J Neurotrauma. 2014;31:125–34. doi:10.1089/neu.2013.3017.

106. Bu G. Apolipoprotein E, and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. Nat Rev Neurosci. 2009;10:333–44. doi:10.1038/nrn2620.
107. Strittmatter WJ, Saunders AM, Schmechel D, et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci. 1993;90:1977–81. doi:10.1073/pnas.90.5.1977.
108. Bosco P, Ferri R, Grazia Salluzzo M, et al. Role of the transforming-growth-factor-β1 gene in late-onset Alzheimer's disease: implications for the treatment. Curr Genomics. 2013;14:147–56. doi:10.2174/1389202911314020007.
109. Leuba G, Vernay A, Kraftsik R, et al. Pathological reorganization of NMDA receptors subunits and postsynaptic protein PSD-95 distribution in Alzheimer's disease. Curr Alzheimer Res. 2014;11:86–96. doi:10.2174/1567205013106660170.
110. Vassar R. ADAM10 prodomain mutations cause late-onset Alzheimer's disease: not just the latest FAD. Neuron. 2013;80:250–3. doi:10.1016/j.neuron.2013.09.031.
111. Choi S, Kim H-R, Leng L, et al. Role of macrophage migration inhibitory factor in the regulatory T cell response of tumor-bearing mice. J Immunol. 2012;189:3905–13. doi:10.4049/jimmunol.1102152.
112. Calandra T, Roger T. Macrophage migration inhibitory factor: a regulator of innate immunity. Nat Rev Immunol. 2003;3:791–800. doi:10.1038/nri1200.
113. Baumann R. Macrophage migration inhibitory factor delays apoptosis in neutrophils by inhibiting the mitochondria-dependent death pathway. FASEB J. 2003;17:2221–30. doi:10.1096/fj.03-0110com.
114. Flex A, Pola R, Serricchio M, et al. Polymorphisms of the macrophage inhibitory factor and C-reactive protein genes in subjects with alzheimer’s dementia. Dement Geriatr Cogn Disord. 2004;18:261–4. doi:10.1159/000080026.
115. Dong CJ, Guo Y, Ye Y, et al. Presynaptic inhibition by 2 receptor/adenylate cyclase/PDE4 complex at retinal Rod bipolar synapse. J Neurosci. 2014;34:9432–40. doi:10.1523/JNEUROSCI.0766-14.2014.
116. Oyama R, Yamamoto H, Titani K. Glutamine synthetase, hemoglobin α-chain, and macrophage migration inhibitory factor binding to amyloid β-protein: their identification in rat brain by a novel affinity chromatography and in Alzheimer's disease brain by immunoprecipitation. Biochim Biophys Acta Protein Struct Mol Enzymol. 2000;1479:91–102. doi:10.1016/S0167-4838(00)00057-1.
117. Mitchell RA, Metz CN, Peng T, et al. Sustained Mitogen-activated Protein Kinase (MAPK) and Cytoplasmic Phospholipase A2 Activation by Macrophage Migration Inhibitory Factor (MIF): regulatory role in cell proliferation and glucocorticoid action. J Biol Chem. 1999;274:18100–6. doi:10.1074/jbc.274.25.18100.
118. Mitchell RA, Liao H, Chesney J, et al. Macrophage migration inhibitory factor (MIF) sustains macrophage proinflammatory function by inhibiting p53: Regulatory role in the innate immune response. Proc Natl Acad Sci. 2002;99:345–50. doi:10.1073/pnas.012511599.
119. Bryan KJ, Zhu X, Harris PL, et al. Expression of CD74 is increased in neurofibrillary tangles in Alzheimer's disease. Mol Neurodegener. 2008;3:13. doi:10.1186/1750-1326-3-13.
120. The Neuroallianz Consortium. http://www.neuroallianz.de/en/mission.html
121. Aetionomy. www.aetionomy.eu

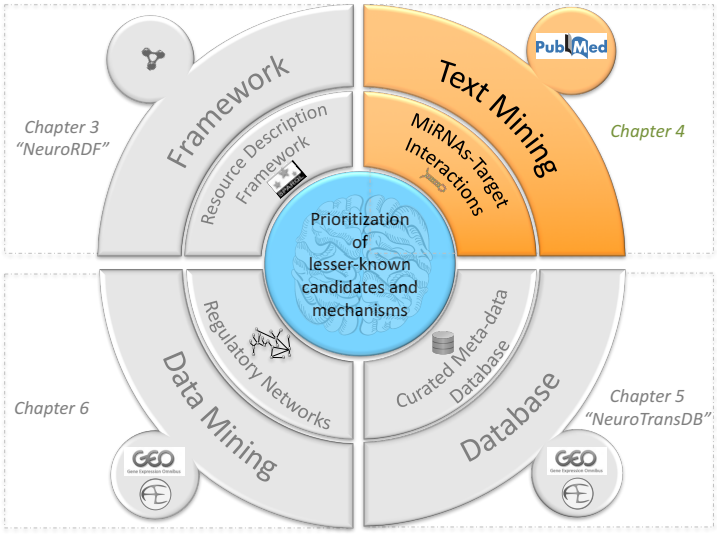
3.3 Summary

This work leverages on semantic-web technologies to develop a context-specific integrative and interoperable disease model. The proposed approach, *NeuroRDF*, harmonises and integrates data from three public resources: PPI databases, text-mined MTIs and PPIs from literature, and gene dysregulation information derived from transcriptomic studies. It outlines the pre-processing steps applied to each data resource to monitor quality and context-specificity. In addition, it reports on the need of huge manual effort for inspecting missing and incorrect metadata information. The main benefits of this work lie in enabling semantic interoperability across heterogeneous data to foster innovation in NDD research.

Querying such a framework empowers the user to ask more fine-grained questions across the knowledge graphs. Exemplary questions, as SPARQL queries, that account for patho-mechanisms and network topology illustrated the power of *NeuroRDF* in identifying previously unattended candidates in AD. The prioritised candidate, MIF, has a pro-inflammatory cytokine functionality that is crucial in inflammatory responses implicated in AD plaque formation. Thus, such an approach has the capability to assist in identification of reliable biomarkers for early diagnosis and treatment. The overall work presented here can be easily extended to other domains. The manual effort needed may vary depending on the formalization and interoperability of the selected data.

Integration of other data resources such as GWAS, NGS, imaging, so on can increase the biological confidence of the derived hypothesis in *NeuroRDF*. Additionally, drug-target information can aid in drug repurposing. The methodologies and approaches to harvest and integrate multiple resources are not well established in AD research. Particularly, knowledge representing miRNA's regulatory roles in AD and regulatory relationships between expression levels of genes. Work addressing these two topics are presented in Chapter 4, Chapter 5, and Chapter 6.

Chapter 4 Hypothesis-driven Knowledge Discovery



4.1 Introduction

A wealth of untapped information is available in the ever-growing biomedical literature, Delivering crucial background knowledge, it can maximise insights into AD research. Automated approaches are required to provide rapid access to this information by gleaning over large textual documents. Most of these methods unearth relationships between biological entities and additionally enrich it with a multitude of contextual information. When integrated with other data, the harvested knowledge allows generation and exploration of novel hypotheses. Text-mining approaches are well-established in some fields to identify biological entities and their relationships — like PPIs, drug-drug interaction. However, its development in other domains — like miRNAs and its interactions — is limited.

MiRNAs are small non-coding RNAs that post-transcriptionally regulate gene expression. Due to its involvement in several normal and pathophysiological processes, miRNAs are considered as potential biomarkers for diagnosis, prognosis, and therapeutics. This chapter describes the automated text-mining methods developed to extract miRNA's regulatory roles in diseases and gene expression from Pubmed abstracts. It provides step-by-step guidelines for annotating entities and their relationships for generating the benchmark corpus. The work here evaluates different relations extraction approaches for identification

of miRNA-disease and miRNA-target associations. This work demonstrates how domain-specific highly relevant information — for miRNAs — can be extracted from existing scientific literature.

4.2 Publication

F1000Research F1000Research 2015, 3:205 Last updated: 01 OCT 2015

METHOD ARTICLE

REVISED Detecting miRNA Mentions and Relations in Biomedical Literature [version 3; referees: 2 approved, 1 approved with reservations]

Shweta Bagewadi[1,2], Tamara Bobić[3], Martin Hofmann-Apitius[1,2], Juliane Fluck[1], Roman Klinger[4]

[1]Fraunhofer SCAI, Bioinformatics, Schloss Birlinghoven, 53754, Sankt Augustin, Germany
[2]University of Bonn, B-IT, Dahlmannstr. 2, 53113 Bonn, Germany
[3]Hasso Plattner Institute Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Potsdam, Germany
[4]Semantic Computing Group, CIT-EC, Bielefeld University, 33615 Bielefeld, Germany

v3 **First published:** 28 Aug 2014, **3**:205 (doi: 10.12688/f1000research.4591.1)
Second version: 23 Dec 2014, **3**:205 (doi: 10.12688/f1000research.4591.2)
Latest published: 01 Oct 2015, **3**:205 (doi: 10.12688/f1000research.4591.3)

Abstract

Introduction: MicroRNAs (miRNAs) have demonstrated their potential as post-transcriptional gene expression regulators, participating in a wide spectrum of regulatory events such as apoptosis, differentiation, and stress response. Apart from the role of miRNAs in normal physiology, their dysregulation is implicated in a vast array of diseases. Dissection of miRNA-related associations are valuable for contemplating their mechanism in diseases, leading to the discovery of novel miRNAs for disease prognosis, diagnosis, and therapy.

Motivation: Apart from databases and prediction tools, miRNA-related information is largely available as unstructured text. Manual retrieval of these associations can be labor-intensive due to steadily growing number of publications. Additionally, most of the published miRNA entity recognition methods are keyword based, further subjected to manual inspection for retrieval of relations. Despite the fact that several databases host miRNA-associations derived from text, lower sensitivity and lack of published details for miRNA entity recognition and associated relations identification has motivated the need for developing comprehensive methods that are freely available for the scientific community. Additionally, the lack of a standard corpus for miRNA-relations has caused difficulty in evaluating the available systems.

We propose methods to automatically extract mentions of miRNAs, species, genes/proteins, disease, and relations from scientific literature. Our generated corpora, along with dictionaries, and miRNA regular expression are freely available for academic purposes. To our knowledge, these resources are the most comprehensive developed so far.

Results: The identification of specific miRNA mentions reaches a recall of 0.94 and precision of 0.93. Extraction of miRNA-disease and miRNA-gene relations lead to an F_1 score of up to 0.76. A comparison of the information extracted by

Open Peer Review

Referee Status: ✓ ? ✓

	Invited Referees 1	2	3
REVISED version 3 published 01 Oct 2015			
REVISED version 2 published 23 Dec 2014	✓ report		✓ report
version 1 published 28 Aug 2014	? report	? report	

1 **Sofie Van Landeghem**, Ghent University Belgium

2 **Filip Ginter**, University of Turku Finland

3 **Robert Leaman**, National Institutes of Health USA

Discuss this article

Comments (0)

our approach to the databases *miR2Disease* and *miRSel* for the extraction of Alzheimer's disease related relations shows the capability of our proposed methods in identifying correct relations with improved sensitivity. The published resources and described methods can help the researchers for maximal retrieval of miRNA-relations and generation of miRNA-regulatory networks.

Availability: The training and test corpora, annotation guidelines, developed dictionaries, and supplementary files are available at http://www.scai.fraunhofer.de/mirna-corpora.html

Corresponding author: Shweta Bagewadi (shweta.bagewadi@scai.fraunhofer.de)

How to cite this article: Bagewadi S, Bobić T, Hofmann-Apitius M *et al.* **Detecting miRNA Mentions and Relations in Biomedical Literature [version 3; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2015, **3**:205 (doi: 10.12688/f1000research.4591.3)

Grant information: Shweta Bagewadi was supported by University of Bonn. Tamara Bobić was partially funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School during her contribution to this work at Fraunhofer SCAI.

Competing interests: No competing interests were disclosed.

First published: 28 Aug 2014, **3**:205 (doi: 10.12688/f1000research.4591.1)

REVISED **Amendments from Version 2**

The final revised version of the manuscript includes changes as per the reviewers' recommendation. We have mainly modified text in the "Corpus selection, annotation and properties" section to simplify the ambiguous texts, as pointed out by Robert Leaman. Additionally, grammatical errors pointed out by the reviewers have also been corrected. Please read the response provided to reviewers' comments for detailed information of the changes.

See referee reports

Introduction

Functionally important non-coding RNAs (ncRNAs) are now better understood with the progress of high-throughput technologies. Discovery of the major class of ncRNAs, microRNAs (miRNAs[1]) has further facilitated the molecular aspects of biomedical research.

MicroRNAs are a large group of small endogenous single-stranded non-coding RNAs (17–22nt long) found in eukaryotic cells. They post-transcriptionally regulate gene expression of specific mRNAs by degradation, translational inhibition, or destabilization of the targets (transcripts of protein-coding genes)[2]. Esquela-Kerscher *et al.* have reported on miRNAs involvement in almost every regulation aspect of biological processes such as apoptosis, and stress response[3]. Wubin *et al.* demonstrated that miR-29a regulatory circuitry plays an important role in epididymal development and its functions[4]. Additionally, tissue-specificity of miRNAs has been shown to provide a better clue of their fundamental roles in normal physiology[5].

Dysregulation of miRNAs and their ability to regulate repertoires of genes (as well as co-ordinate multiple biological pathways) has been linked to several diseases[6,7]. One example is chronic lymphocytic leukemia where (in about 68% of the cases) miRNA genes (*miR15* and *miR16*) are missing or down-regulated[8]. Thus, uncovering the relations between miRNAs and diseases as well as genes/proteins is crucial for our understanding of miRNA regulatory mechanisms for diagnosis and therapy[9,10].

Several databases, prediction algorithms and tools are available, providing insight into miRNA-disease and miRNA-mRNA associations. Although the detailed target recognition mechanism is still elusive, several algorithms attempt to predict miRNA targets. However, a limited precision of 0.50 and recall of 0.12 has been reported when evaluated against proteomics supported miRNA targets[11]. Despite the fact that these resources provide insight into miRNA-associated relationships, the majority of relations are scattered as unstructured text in scientific publications[12]. Figure 1 shows the growth of publications in MEDLINE and in addition depicts the normalized growth of publications that reference the keyword "microRNA".

Some databases such as *miR2Disease* and *PhenomiR* store manually extracted relations from literature. The *miR2Disease* database[13] contains information about miRNA-disease relationships with 3273 entries (as of the last update on March 14, 2011). *PhenomiR*[14] is a database on miRNA-related phenotypes extracted from published experiments. It consists of 675 unique miRNAs, 145 diseases, and 98 bioprocesses from 365 articles (Version 2.0, last updated on February 2011). *TarBase*[11] hosts more than 6500 experimentally validated miRNA targets extracted from literature.

However, manual retrieval of relevant articles and extraction of relation mentions from them is labor-intensive. A solution is to use text-mining techniques. Moreover, the vast majority of the research in this direction is mainly focused around extraction of

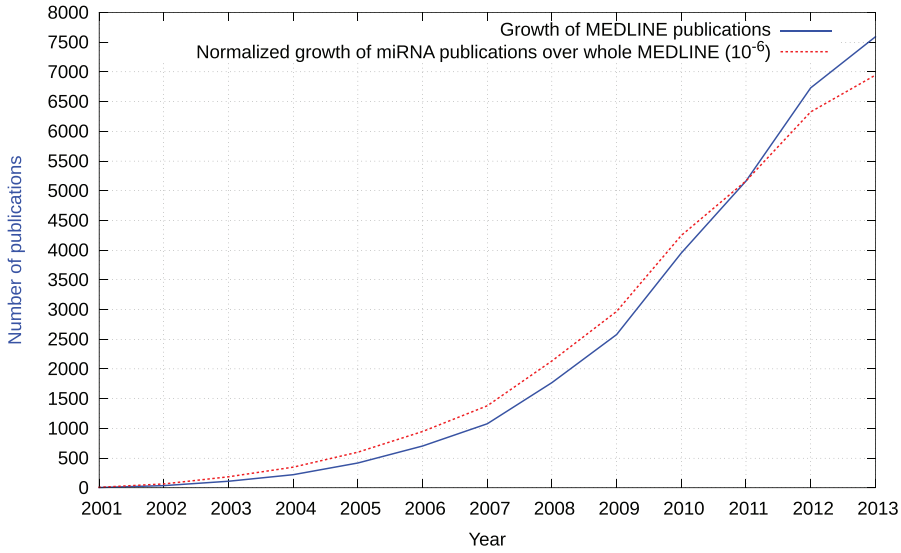


Figure 1. Growth of miRNA-related publications in comparison with the growth of MEDLINE. The dotted line points out the relative increase of miRNA-related publications per year in comparison to the growth of MEDLINE (as of 31 December, 2013).

protein-protein interactions[15]. On the contrary, miRNA relation extraction is still naive. The shift of focus towards identification of miRNA-relations is slowly establishing with the rise in systems approaches to investigate complex diseases. The manually curated database *miRTarbase*[16] incorporates such text-mining techniques to retrieve miRNA-related articles. Recently, the *miRCancer* database has been constructed using a rule-based approach to extract miRNA-cancer associations from text[17]. As of June 14, 2014, this database contains 2271 associations between 38562 miRNAs and 161 human cancers from 1478 articles.

Related work

Text-mining technologies are established for a variety of applications. For instance, the BioCreative competition[18,19] and BioNLP Shared Task[20–22] series have been conducted to benchmark text mining techniques for gene mention identification, protein-protein relation extraction and event extraction, among others.

To our knowledge, only limited work has been carried out in the area of miRNA-related text-mining. Murray *et al.* considered miRNA-gene associations from PubMed database using semantic search techniques[23]. For their analysis, experimentally derived datasets were examined, combined with network analysis and ontological enrichment. Regular expressions were used to detect miRNA mentions. The authors claim to have optimized the approach to reach 100% accuracy and recall for detecting miRNAs mentions as in *miRBase*. Relations were identified based on a manually curated rule set. The authors extracted 1165 associations between 270 miRNAs and 581 genes from the whole MEDLINE.

The freely available *miRSel*[12] database integrates automatically extracted miRNA-target relationships from **PubMed Abstract**s. A set of regular expressions is used for miRNA recognition that matches all *miRBase* synonyms and generic occurrences. The authors reach a recall of 0.96 and precision of 1.0 on 50 manually annotated abstracts for miRNA mention identification. Further, the relations between miRNA and genes were extracted at sentence level employing a rule-based approach. They evaluated on 89 sentences from 50 abstracts resulting in a recall of 0.90 and precision of 0.65. Currently, it hosts 3690 miRNA-gene interactions[11].

Since the miRNA naming convention has been formalized very early in comparison to other biological entities such as genes and proteins, applying text-mining approaches is relatively simple[17]. Thus, most of the previously applied text mining approaches for miRNA detection has been based on keywords. *miRCancer* uses keywords to obtain abstracts from PubMed, further miRNA entities have been identified using regular expressions based on prefix and suffix variations. Similarly, *miRWalk* database uses keyword search approach to download abstracts and applies a curated dictionary (compiled from six databases) for miRNA identification of human, rat, and mouse species[24]. *TarBase*, *miR2Disease*, *miRTarBase*, and several others have followed related search strategies. However, several authors still tend to use naming variations for acronyms, abbreviations, nested representations, *etc.* for listing miRNAs. Additionally, in contrast to the previous text-mining approaches focusing purely on miRNA gene relations, we extend the information extraction approach additionally to retrieve miRNA-disease relations. Furthermore, we evaluate our approach using a larger corpus to achieve robustness. We differentiate between actual miRNA mentions (refered to as Specific miRNAs) and co-referencing miRNAs (Non-Specific miRNAs), which could in addition enhance keyword search. We evaluated three different relation extraction approaches, namely co-occurrence, tri-occurrence and machine learning based methods.

To support further research, our corpora are made publicly available in an established XML format as proposed by Pyysalo *et al.*[25], as well as the regular expressions used for miRNAs named entity recognition. In addition, our dictionary for trigger term detection and general miRNA mention identification are made available. To our knowledge, the annotated corpora as well as the information extraction resources are the most comprehensive developed so far.

Methods

Data curation and corpus selection

Named entities annotation. Mentions of miRNAs consisting of keywords (case-insensitive and not containing any suffixed numerical identifier) such as *"Micro-RNAs"* or *"miRs"* are annotated as Non-Specific miRNA. Names of particular miRNAs such as *miRNA-101*, suffixed with numerical identifiers are labeled as Specific miRNA. Numerical identifiers (separated by delimiters such as ",", "/", and "and") occurring as part of specific miRNA mentions are annotated as a single entity. Box 1 depicts the annotation of specific miRNA mentions (including an example for part mentions). In addition, Disease, Gene/Protein, Species, and Relation Trigger are annotated. The detailed annotation guideline for annotating specific miRNA mentions is available as a supplementary file.

Box 1. Example of miRNAs annotations. Here "*-181b*", and "*-181c*" are the part mentions annotated as a single entity along with "*miR-181a*" in box. A non-specific miRNA mention is shown in italics.

Interesting results were obtained from miR-181a, -181b, and -181c. These set of brain-enriched *miRNAs* are down-regulated in glioblastoma. However, miR-222, and miR-128 are strongly up-regulated.

Mentions of disease names, disease abbreviations, signs, deficiencies, physiological dysfunction, disease symptoms, disorders, abnormalities, or organ damages are annotated as Disease. Only disease nouns were considered, adjective terms such as "*Diabetic patients*" are not marked; however, specific adjectives that can be treated as nouns were marked, e.g. "Parkinson's disease patients". Mentions referring to proteins/genes which are either single word (*e.g. "trypsin"*), multi-word, gene symbols (*e.g. "SMN"*), or complex names (including of hyphens, slashes, Greek letters, Roman or Arabic numerals) are annotated as Gene/Protein. Only those organisms that are having published miRNA sequences and annotations represented in *miRBase* database are labeled as Species. Any verb, noun, verb phrase, or noun phrase associating miRNA mention to either labeled disease or gene/protein term is annotated as Relation Trigger.

Relations annotation. We restrict the relationship extraction to sentence level and four different interacting entity pairs: Specific miRNA-Disease (SpMiR-D), Specific miRNA-Gene/Protein (SpMiR-GP),

Non-Specific miRNA-Disease (NonSpMiR-D), and Non-Specific miRNA-Gene/Protein (NonSpMiR-GP). Relevant triples, an interacting pair (from one of the above-mentioned) co-occurring with a Relation Trigger in a sentence are defined to form a relation and can belong to one of the four above-mentioned Relation classes. On the contrary, if an interacting pair does not co-occur with any Relation Trigger then we do not tag such pair as a relation.

The annotation has been performed using Knowtator[26] integrated within the Protégé framework[27].

Corpus selection, annotation and properties. We develop a new corpus based on MEDLINE, annotated with miRNA mentions and relations. Shah *et al.*[28] showed that abstracts provide a comprehensive description of key results obtained from a study, whereas full text is a better source for biological relevant data. Thus, we choose to build the corpus for abstracts only. Out of 27001 abstracts retrieved using the keyword "miRNA", 201 were randomly selected as training and 100 as test corpus. Two annotators performed the annotation. The first annotator annotated the training corpus iteratively to develop guidelines and built the consensus annotation. The second annotator followed these guidelines and annotated the same corpus. Disagreeing instances were harmonized by both the annotators through manual inspection for correctness and its adherence to the guidelines. Any changes to the guidelines were made if needed. During the harmonization process only the non-overlapping instances between the two annotators were investigated. Decisions were based on the rule that only noun forms were to be marked (specific adjectives that can be treated as nouns were also considered). In case of partial matches, where conflicting parts could be interpreted as an adjective were not resolved. For example, in "chronic inflammation", marking either "chronic inflammation" or just "inflammation" were considered correct. Table 1 provides the inter-annotator agreement (measured as F_1, for both exact and boundary match, and Cohen's κ) for the test corpus. Exact string match occurs only when both the annotators annotate identical strings, whereas in partial match fraction of the string has been annotated by either of the annotators. It is evident (*cf.* Table 1) that in almost all cases partial match performs better than exact string match, indicating variations in span of mentioned entities. An example annotation is shown in Box 1.

Table 2 shows the number of annotated concepts in the training and test corpora for each entity class and the count for manually extracted relations (triplets), categorized for different interacting entity pairs. Table 3 provides the overall statistics of the published corpora (additional information about the corpus is given in the README supplementary file).

Table 1. Inter-annotator agreement scores for the test corpus.

Annotation Class	F_1 (Exact Match)	F_1 (Partial Match)	κ
Non-specific MiRNAs	0.9985	0.9985	0.996
Specific MiRNAs	0.9545	0.9779	0.916
Genes/Proteins	0.8343	0.8705	0.752
Diseases	0.8270	0.9575	0.853
Species	0.9329	0.9437	0.875
Relation Triggers	0.8441	0.9543	0.798

Automated named entity recognition

For identification of specific miRNA mentions in text (*cf.* Table 4), we developed regular expression patterns using manual annotations of miRNA mentions as the basis. Similarly, a dictionary has been generated for general miRNA recognition. The regular expression patterns are represented in the format as defined by Oualline *et al.*[29]. For simplicity and reusability, several aliases are defined (*cf.* Table 5) to be used in the final regular expression patterns for specific miRNA identification, given in Table 4. Detected entities are resolved to a unique miRNA name and disambiguated to adhere to standard naming conventions as authors use several morphological variants to report the same miRNA term. For example, miR-107 can be represented as miRNA-107, Micro RNA-107, MicroRNA 107, has-mir-107, mir-107/108, micro RNA 107 and 108, micro RNA (miR) 107 and so on. Thus, the identified miRNA entity has been resolved to its base form (*e. g. hsa-microRNA-21* to *hsa-mir-21* and *microRNA 101* to *mir-101*) following the miRBase naming convention. Manual inspection of the test corpus for species distribution revealed that 71% of the documents belonged to human, followed by mouse (15%), rat (8%). Pig has 2 abstracts, zebrafish, HIV-1, HSV-1, and *Caenorhabditis elegans* 1 each (*cf.* Supplementary Figure A for the distribution). Thus, we assumed that most of the abstracts belonged to human and resolved the identified miRNA entities to human identifier in miRBase. Unique miRNA terms are mapped to human *miRBase* database identifiers through the mirMaid Restful web service. For those names where we do not retrieve any database identifiers, we fall back to another organism mention found in the abstract (if any), using the NCBI taxonomy dictionary (see below)

Table 2. Manually annotated entities statistics. Counts of manually annotated entities in the training and the test corpora as well as annotated sentences describing relations.

Annotation Class	Corpus	
	Training	Test
Non-specific MiRNAs	1170	336
Specific MiRNAs	529	376
Genes/Proteins	734	324
Diseases	1522	640
Species	546	182
Relation Triggers	1335	625
SpMiR-D	171	127
SpMiR-GP	195	123
NonSpMiR-D	124	54
NonSpMiR-GP	77	16

Table 3. Statistics of the published miRNA corpora.

Occurrences in the corpus	Training	Test
Sentences	1864	780
Entities	5836	2483
Entity pairs	2001	868
Positive entity pairs	567	320
Negative entity pairs	1434	548

Table 4. Regular expression patterns used for miRNAs identification. Aliases used to form the final regular expression, see Table 5, are highlighted in bold.

Regular expression patterns	Description	Example of identified text
(Pref+(Lin,Let))	Detection of *Lin* and *Let* variations of miRNAs	lin-4; hsa-let-7a-1
(Pref+(miRNA, Onco)(S*Tail)(Sep Tail)*)	MiRNAs mentions for different separators	hsa-mir-21/22; Oncomir-17~92
(Pref+(miRNA, Onco) S*(D(Z([/]Z)*)+) ([\,] S*? (Pref+(miRNA, Onco) S*(D(Z([/]Z)*)+)*)))	Multiple miRNA mentions occurring progressively	miR-17b, -1a; hsa-miR-21,22, and hsa-miR-17

Table 5. MiRNAs regex aliases. Aliases used in regular expression patterns for miRNAs identification (highlighted in bold).

Description	Alias	Regular Expression Pattern
Digit sequences	**D**	(\d?\d*)
Admissible hypens with a trailing space	**Z**	([\-]?[\-]*)
Admissible hypens with a leading space	**S**	([\-]?[\-]*)
3-letter prefix for human followed by a hyphen	**Pref**	([hH][sS][aA][\-])
Non-specific miRNA mentions	**miRNA**	([mM][iI]([cC][rR][oO])+[rR]([nN][aA]s+)+)
Let-7 miRNA mention	**Let**	([lL][eE][tT]**S***[7]?\l+)
Lin-4 miRNA mention	**Lin**	([lL][iI][nN]**S***[4]?\l+)
Oncomir miRNA mention	**Onco**	([oO][nN][cC][oO][mM][iI][rR])
Admissible tilde and word boundaries	**Cluster**	(~[\b]-[\b]-*)
Admissible hyphen and separator *and* and *comma*	**Sep**	(**S***((and?,**S**,\/,)? **S***)+)
Admissible combination of upper and lower case alphabets	**UL**	(?\l?\l+,?\u?\u+)
Admissible alpha-numerical identifiers in specific miRNA mentions	**AN**	(**UL**((/, *and*,**D**+)? **UL**)+)
Admissible alpha-numerical identifiers in oncomir mentions	**Tail**	(**D**(**AN Cluster**+,\-**D AN**+)+)

(*cf.* Supplementary Figure B), otherwise we retain the unique normalized name (*cf.* Box 2).

Box 2. Un-normalized and normalized entities that are mapped to miRBase identifiers. Here MIR0000007, MIR0000008, and MIR0000005 are internal identifiers used by ProMiner.

MIR0000007:MIMAT0015092@MIRBASE|MI0000002@MIRBASE|cel-lin-4|lin-4
MIR0000008: miR-171|microRNA 171
MIR0000005:MIMAT0000416@MIRBASE|has-miR-1|miRNA-1

We detect SPECIES with a dictionary-based approach. The built dictionary consists of all the concepts from the NCBI taxonomy corresponding to only those organisms mentioned in *miRBase*.

Similarly, for identification of DISEASE and GENE/PROTEIN mentions in text we adapted a dictionary-based approach. To detect DISEASE, we apply three dictionaries: MeSH, MedDRA[30] and *Allie*. For GENE/PROTEIN, a dictionary[31] based on SwissProt, EntrezGene, and HGNC is included. Gene synonyms which could be potentially tagged as miRNAs are removed to overcome redundancy. For example, genes encoding microRNA, *hsa-mir-21* are named as *miR-21*, *miRNA21* and *hsa-mir-21*, the gene symbol of *MIR16 membrane interacting protein of RGS16* is *MIR16*, which can represent a miRNA mention.

The RELATION TRIGGER dictionary comprises of all interaction terms from the training corpus. After reviewing the training corpus for relation trigger terms, we retrieved not one but many variants of the same RELATION TRIGGER occurring in alternative verb-phrase groups. For example, "change in expression" can be represented as one of the following verb-phrases: Change MicroRNA-21 Expression, Expression of caveolin-1 was changed, Change in high levels of high-mobility group A2 expression, change of the let-7e and miR-23a/b expression, expression of miR-199b-5p in the non-metastatic cases was significantly changed, etc. To allow flexibility for capturing RELATION TRIGGER along with its variants spanning over different phrase length, we first manually represented all the relations in its root form, such as "regulate expression" to "regulate" (*cf.* Relation_Dictionary.txt file in Dataset 1). The base form has been extended manually to different spelling variants, *e.g.* regulate to regulatory, regulation, etc., the detailed listing of variants is provided in Word_variations.txt in Dataset 1. Not all combinations of

the root forms are logical; target and up-regulation terms cannot be combined to form a relation trigger. Thus, we additionally defined a set of relation combinations that are allowed (see Permutation_terms.txt in Dataset 1 for all combinations).

For all named entity recognition performed, the dictionary-based system ProMiner[31] is used. Supplementary Table A (Dataset 1) provides a quantitative estimate of the entities available in the dictionaries used in this work.

Relation extraction

We consider three approaches for addressing automatic extraction of interacting entity pairs from free text, described in the following.

The co-occurrence approach serves as a baseline. Assuming all interactions to be present in isolated sentences, this approach is complete but may be limited in precision. Reducing the number of false positives can be achieved by filtering with the dictionary of relation triggers occurring in the same sentence. The rationale behind this filter is that the interaction is more likely to be described if such a term is present (we refer to this as tri-occurrence).

To increase the precision, we use a machine learning-based approach formulating the relation detection as a binary classification problem: each instance (consisting of a pair of entities) is classified either as not-containing a relation or belonging to one of the four-relation classes. Our system uses lexical and dependency parsing features. We evaluate linear support vector machines (SVM)[32] as implemented in the LibSVM library, as well as LibLINEAR, a specialized implementation for processing large data sets[33], and naive Bayes classifiers[34]. For more details, we refer to Bobić *et al.*[35].

Lexical features capture characteristics of tokens around the inspected pair of entities. The sentence text can roughly be divided into three parts: text between the entities, text before the entities, and text after the entities. Stemming[36] and entity blinding is performed to improve generalization. Features are bag-of-words and bi, tri, and quadri-gram based. This feature setting follows Yu *et al.* and Yang *et al.*[37,38]. The presence of relation triggers is also taken into account, using the previously described manually generated list. Next to lexical features, dependency parsing (created using Stanford parser) provides an insight into the entire grammatical structure of the sentence[39] and was performed using the Stanford CoreNLP library (http://nlp.stanford.edu/software/corenlp.shtml). Deep parsing follows the shortest dependency path hypothesis[40]. We analyzed the vertices v (tokens from the sentence) in the dependency tree from a lexical (text of the token) and syntactical (POS tag) perspective. Edges e in the tree correspond to the information about the grammatical relations between the vertices. Extracting relevant information from the dependency parse tree is usually done following the shortest dependency path hypothesis[40]. Lexical and syntactical *e-walks* and *v-walks* on the shortest path are created by alternating sequence of vertices and edges, with the length of 3. We capture the information about the common ancestor vertex, in addition to checking whether the ancestor node represents a verb form (*e.g.* POS tag could be VB, VBZ, VBD, etc.). Finally, the length of the shortest path (number of edges) between the entities is considered as a numerical feature.

Results and discussion

Dataset 1. Version 2. Manually annotated miRNA-disease and miRNA-gene interaction corpora

http://dx.doi.org/10.5256/f1000research.4591.d40643

Please see README.txt in the zip file for precise details about the corpus and supplementary files. The updated zip file contains new files (Permutation_terms.txt, Non-Specific_miRNAs_Dictionary.txt and Word_variations.txt) and Table A has been updated.

In the following, we present results for named entity recognition and relation extraction. This section concludes with two use-case analyses.

Performance evaluation of named entity recognition

Among the 201 abstracts present in the training corpus, 82% contained general miRNA mentions, in comparison to specific miRNAs with 45%. In Table 6, results for miRNA entity recognition are reported. Non-specific miRNA recognition is close to perfect. Specific miRNA mention recognition has an F_1 measure of 0.94.

For disease mention recognition, combined dictionaries, based on three established resources, resulted in 0.79 and 0.69 F_1 score for the training and test corpus respectively. The low score for disease identification could be due to the variation in disease mentions, such as multi-word, synonym combination, nested names, etc. However, the partial matches result for diseases reported 0.88 of F_1, providing the possibility for detection of similar text strings for better recall (*cf.* Supplementary Table B in Dataset 1). Genes/proteins dictionary showed a performance of 0.84 and 0.85 of F_1 in training and test corpus respectively.

The evaluation of the relation trigger dictionary (*cf.* Table 6) suggests that it covers a substantial part of the vocabulary with recall of 0.86 for the training and 0.79 for the test corpus.

Relation extraction

We queried MEDLINE for "miRNA and Epilepsy" documents, among which 16 documents containing miRNA-related relations were manually selected (*cf.* Supplementary Figure C for the detailed distribution statistics). To avoid any biased approach we choose Epilepsy disease domain. Manual inspection of these articles revealed

Table 6. Evaluation results for miRNA entity classes. Here only complete match results are presented. The performance of named entity recognition is evaluated using recall (R), precision (P) and F_1 score.

Entity Class	R	P	F_1	R	P	F_1
	Training Corpus			Test Corpus		
Non-specific MiRNAs	1.000	0.995	0.997	1.000	0.997	0.999
Specific MiRNAs	0.921	0.928	0.924	0.936	0.934	0.935
Relation Triggers	0.864	0.885	0.874	0.790	0.842	0.815

11.5% of miRNA-related associations occur outside the sentence level. Thus, our work focused on relations at sentence level. Sentences in which co-occurring entity pairs do not participate in any relation are tagged as *false*. A comparison of the different relation extraction approaches is shown in Figure 2. Supplementary Table D in Dataset 1 provides statistical details of the applied approaches given in Figure 2. If all the entities are correctly identified then co-occurrence based approach leads to 100% recall for relation extraction. The recall is not diminished using the tri-occurrence approach, as the true entity pairs remain constant, while the precision increases between 4pp (percentage points) and 17pp when compared to the co-occurrence based approach, reducing false positives (*cf.* Figure 2). However, overall the precision reaches less than 60%. In our work, we assume that all the entities have been identified giving a recall of 100% for both co-occurrence and tri-occurrence based approaches. Using the machine-learning based classification, precision is increased up to 76% for specific miRNA-gene relations for both LibLINEAR and LibSVM methods, although Naïve Bayes is not far behind. Similarly, these two methods performed nearly the same for specific miRNAs-disease relations, the F_1 measure is not substantially different but a trade-off between precision and recall can be observed. An increase in F_1 measure is observed for non-specific miRNA relations when Naïve Bayes method is applied, out performing other strategies. Nevertheless, preference of the method highly depends on the compromise one chooses, whether better recall or precision. Overall, better recall and acceptable precision can be achieved with tri-occurrence method.

Most relation extraction approaches are dependent on the performance of named entity recognition. The impact of error propagation coming from automated entity recognizers is evaluated by applying the tri-occurrence method on the automatically annotated training and test corpus, here termed as "NERTri". Compared to the results on the gold standard entity annotation a drop of 13 pp for NonSpMiR-D, 7pp for NonSpMiR-GP, 22pp for SpMiR-D, and 30pp for SpMiR-GP in F_1 is observed for the test corpus. Overall performance of the NERTri approach on training and test corpus is detailed in Supplementary Table C in Dataset 1.

Use case analysis

For the impact analysis of the proposed approach, we compare the extracted information with two databases, namely *miR2Disease* and *miRSel*. We focus on relations and articles concerning Alzheimer's disease.

Alzheimer's disease (AD) is ranked sixth for causing deaths in major developed countries[41]. It affects not only individuals but also incurs a high cost to the society. Recently, miRNAs have shown close associations with AD pathophysiology[42,43]. Increasing the need to identify new therapeutic targets for AD, after major set backs due to failed drugs, motivates the need to look in this direction. *In silico* methods, such as the one proposed in this work, can aid in building miRNA-regulatory networks specific to AD, for further analysis such as identifying the mechanisms, sub-networks, and key targets.

Extracting miRNA-Alzheimer's disease relations from full MEDLINE

The database *miR2Disease* is queried to return all miRNA-disease relations occurring in Alzheimer's disease. For comparison, we retrieved miRNA-disease relations from MEDLINE using NERTri approach, resulting in 41 abstracts containing 159 relations. Obtained triplets have been manually curated to remove 51 false positives. False negatives have not been accounted, which may result in loss of information (*cf.* Relation extraction section). Comparison between the relations obtained from *miR2Disease* and NERTri are summarized in Table 7. The *miR2Disease* database returns 28 evidential statements from 9 articles. Among these, only 14 evidences are present in abstracts. Moreover, 16 evidences are extracted from one full text document[44]. Only two evidences are identified at abstract level among these 16 evidences. Overall, 26 miRNAs identified by *miR2Disease* refer to Alzheimer's disease.

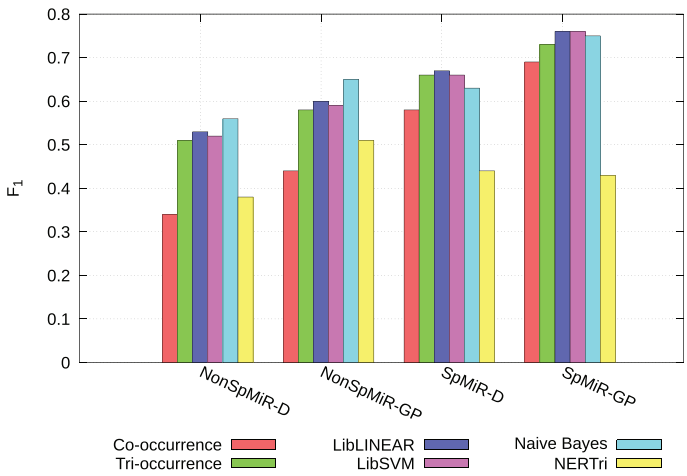


Figure 2. Comparison of different relation extraction approaches. On the x-axis, different entity pair relations are represented as SpMiR-D for Specific miRNA-Disease, SpMiR-GP for Specific miRNA-Gene/Protein, NonSpMiR-D for Non-Specific miRNA-Disease, and NonSpMiR-GP for Non-Specific miRNA-Gene/Protein.

Table 7. miR2Disease database comparison. MiRNA-Alzheimer's disease relation retrieved from MEDLINE and in *miR2Disease* database.

	miR2Disease	NERTri	True Positives in NERTri	NERTri and miR2Disease Overlap
Publications	9	41	36	8
Relations	28	159	108	11
Evidences (abstracts)	14	159	108	10
Unique miRNAs	26	46	40	16

Therefore, our text-based extraction proposes approximately three times more relations than the database provides.

The analysis of 17 false negative relations which are in the database but not found by our approach shows that most of the relations could be found only in full text and that the automatic system misses four miRNA-Alzheimer's disease relations from abstracts. Manual inspection reveals that in three out of these missing four evidences the disease name is not mentioned in the sentence (relation occurred at co-reference level).

Extraction of miRNA-gene relations for Alzheimer's diseases from full MEDLINE

Here we compare the performance of our relation detection NERTri with another text-mining database, *miRSel*. For comparison, 100 abstracts from PubMed were retrieved using the query *"alzheimer disease"[MeSHTerms] OR ("alzheimer disease"[All Fields] OR "alzheimer"[All Fields]) AND ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR "microrna"[All Fields]) AND ("2001/01/01"[PDAT]:"2013/7/4"[PDAT])*. Manual inspection of these articles leads to 184 miRNA-gene relations, at sentence level, (Table 8) in 37 abstracts.

NERTri approach was able to identify 140 of these found relations in 28 abstracts. Among the 37 abstracts from the PubMed query, *miRSel* contained only 12 abstracts with 56 miRNA-gene relations (cf. Table 8). False negatives in our approach when compared with *miRSel* could not be directly identified as the database is not downloadable and searchable for disease specific relations. However, low intersection between *miRSel* and NERTri can be observed.

Table 8. miRSel database comparison. Comparison of miRNA-gene relations retrieval for Alzheimer's disease in MEDLINE.

Approach	Articles	Relations
PubMed Query ("Alzheimer AND miRNA")	100	NA
PubMed Query with relations at sentence level	37	184
PubMed Query ∩ NERTri	28	140
PubMed Query ∩ miRSel	12	56
NERTri ∩ miRSel	14	22

In summary, our approach provides AD related gene-microRNA relations from PubMed which have not been available in the database before.

Overall, the results are promising when compared with the *miR2Disease* and *miRSel* databases and indicate that we can extend the databases to a large extent with new relations. Such an approach makes it much easier to keep databases up to date. Nevertheless full text processing would most certainly increase the recall of automatic processing.

Conclusion and future work

In this work, we proposed approaches for identification of relations between miRNAs and other named entities such as diseases, and genes/proteins from biomedical literature. In addition, details of named entity recognition for all the above entity classes have been described. We distinguished two types of miRNA mentions, namely Specific (with numerical identifiers) and Non-Specific (without numerical identifiers). Non-specific miRNAs entity recognition has enabled us to achieve better recall and precision in document retrieval. Three different relation extraction approaches are compared, showing that the tri-occurrence based approach should be the first reliable choice among all others. The tri-occurrence based approach is comparable to a machine learning-based method but considerably faster. In comparison to two well-established databases, we have shown that additional useful information can be extracted from MEDLINE using our proposed methods.

To best of our knowledge, this is the first work where manually annotated corpora containing information about miRNAs and miRNA-relations are published. Moreover, the corpora and methods provided represent useful basis and tools for extracting the information about miRNAs-associations from literature. This work serves as an important benchmark for current and future approaches in automatic identification of miRNA relations. It provides the basis for building a knowledge-based approach to model regulatory networks for identification of deregulated miRNAs and genes/proteins.

The proposed methods encourage the extension of this work to full-text articles, to elucidate many more relations from Biomedical literature. Non-specific miRNA mention identification could prove highly beneficial for co-reference resolution in full-text articles, in addition to abstracts. Proposed machine-learning approaches could be applied to only tri-occurrence based instances for reducing the false positive rates. Extending the current approach to other model organisms such as mouse, and rat could be helpful in revealing

important relations for translational research. Inclusion of additional named entities such as drugs, pathways, etc. could lead to an interesting approach for detection of putative therapeutic or diagnostic drug targets through a gene-regulatory network generated from identified relations.

Data availability

Corpora availability: http://www.scai.fraunhofer.de/mirnacorpora.html

Archived corpora at time of publication: F1000Research: Dataset 1. Version 2. Manually annotated miRNA-disease and miRNA-gene interaction corpora, 10.5256/f1000research.4591.d40643[45]

Author contributions

SB, RK, JF, and MHA conceived and designed the overall research strategy. SB carried out all the development work and performed the analysis. She is the major contributor of manuscript preparation and principal annotator. TB developed the machine learning-based workflow for relation extraction, transformed corpora into the standard format, and contributed to manuscript writing. JF supported in use-case analysis and paper writing. MHA is the scientific supervisor for this work. RK contributed to critical discussions, analysed the results and a major contributor in correcting and writing manuscript. All authors read and approved the final version of the manuscript.

Competing interests

No competing interests were disclosed.

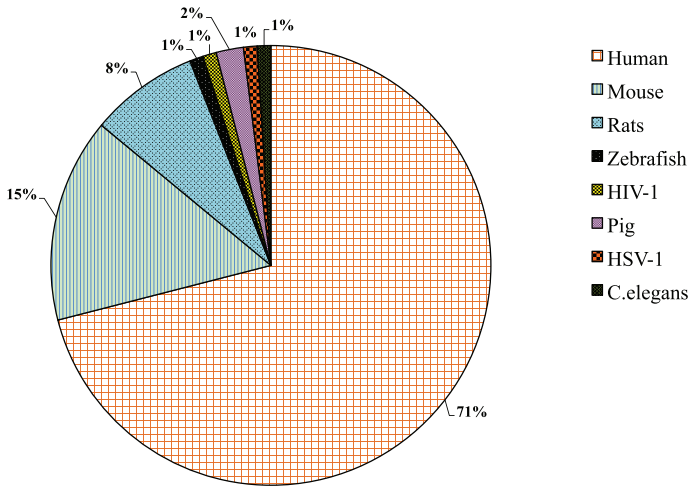
Grant information

Shweta Bagewadi was supported by University of Bonn. Tamara Bobić was partially funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School during her contribution to this work at Fraunhofer SCAI.

Acknowledgements

We would like to thank Heinz-Theo Mevissen for all the support during implementation of the dictionaries and regular expressions in ProMiner. We acknowledge Anandhi Iyappan for her contribution as the second annotator. We are also grateful to Harsha Gurulingappa for all his support and fruitful discussions during this work. We would like to thank Ashutosh Malhotra for proof reading the manuscript.

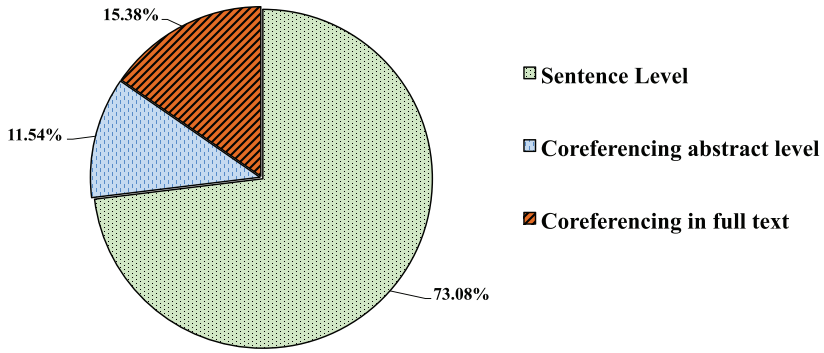
Supplementary figures



Supplementary Figure A. Distribution of organism mentions in training corpus.

Title: Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity.
Sentence : Interestingly, mir-125 is a putative homologue of lin-4.
Normalized miRNA name: dme-mir-125

Supplementary Figure B. A screenshot example of how we handle other organism miRNA normalization. There is no miR-125 entry related to human in miRBASE. Since the abstract mentions *Drosophila melanogaster* in the title, the miRNA is normalized to dme-mir-125.



Supplementary Figure C. Coverage of relations occurring in Epilepsy Documents.

References

1. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell.* 1993; **75**(5): 843–54.
PubMed Abstract | Publisher Full Text
2. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell.* 2004; **116**(2): 281–297.
PubMed Abstract | Publisher Full Text
3. Esquela-Kerscher A, Slack FJ: **Oncomirs microRNAs with a role in cancer.** *Nat Rev Cancer.* 2006; **6**(4): 259–69.
PubMed Abstract | Publisher Full Text
4. Ma W, Hu S, Yao G, *et al.*: **An androgen receptor-microrna-29a regulatory circuitry in mouse epididymis.** *J Biol Chem.* 2013; **288**(41): 29369–81.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Babak T, Zhang W, Morris Q, *et al.*: **Probing microRNAs with microarrays: tissue specificity and functional inference.** *RNA.* 2004; **10**(11): 1813–1819.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Bottoni A, Zatelli MC, Ferracin M, *et al.*: **Identification of differentially expressed microRNAs by microarray: a possible role for microRNA genes in pituitary adenomas.** *J Cell Physiol.* 2007; **210**(2): 370–377.
PubMed Abstract | Publisher Full Text
7. Wu X, Song Y: **Preferential regulation of miRNA targets by environmental chemicals in the human genome.** *BMC Genomics.* 2011; **12**(1): 244.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Calin GA, Dumitru CD, Shimizu M, *et al.*: **Frequent deletions and downregulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.** *Proc Natl Acad Sci U S A.* 2002; **99**(24): 15524–9.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Banno K, Yanokura M, Iida M, *et al.*: **Application of microRNA in diagnosis and treatment of ovarian cancer.** *BioMed Res Int.* 2014; **2014**: 232817.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell.* 2009; **136**(2): 215–33.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Vergoulis T, Vlachos IS, Alexiou P, *et al.*: **TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.** *Nucleic Acids Res.* 2011; **40**(Database issue): D222–229.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Naeem H, Küffner R, Csaba G, *et al.*: **miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature.** *BMC Bioinformatics.* 2010; **11**: 135.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Jiang Q, Wang Y, Hao Y, *et al.*: **miR2Disease: a manually curated database for microRNA deregulation in human disease.** *Nucleic acids Res.* 2009; **37**(Database issue): D98–104.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Ruepp A, Kowarsch A, Schmidl D, *et al.*: **PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes.** *Genome Biol.* 2010; **11**(1): R6.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Czarnecki J, Nobeli I, Smith A, *et al.*: **A text-mining system for extracting metabolic reactions from full-text articles.** *BMC Bioinformatics.* 2012; **13**(1): 172.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Hsu SD, Lin FM, Wu WY, *et al.*: **miRTarBase: a database curates experimentally validated microRNA-target interactions.** *Nucleic acids Res.* 2011; **39**(Database issue): D163–9.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Xie B, Ding Q, Han H, *et al.*: **miRCancer: a microRNA-cancer association database constructed by text mining on literature.** *Bioinformatics.* 2013; **29**(5): 639–44.
PubMed Abstract | Publisher Full Text
18. Smith L, Tanabe LK, nee Ando RJ, *et al.*: **Overview of BioCreative II gene mention recognition.** *Genome Biol.* 2008; **9**(Suppl 2): S2.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Arighi CN, Lu Z, Krallinger M, *et al.*: **Overview of the BioCreative III Workshop.** *BMC Bioinformatics.* 2011; **12**(Suppl 8): S1.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Nedellec C, Bossy R, Kim JD, *et al.*: **Proceedings of the BioNLP Shared Task 2013 Workshop**. Association for Computational Linguistics, Sofia, Bulgaria, 2013.
Reference Source
21. Tsujii J, Kim JD, Pyysalo S: **Proceedings of BioNLP Shared Task 2011 Workshop**. Association for Computational Linguistics, Portland, Oregon, USA, 2011.
Reference Source
22. Tsujii J: **Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task**. Association for Computational Linguistics, Boulder, Colorado, 2009.
Reference Source
23. Murray BS, Choe SE, Woods M, *et al.*: **An *in silico* analysis of microRNAs: mining the miRNAome.** *Mol Biosyst.* 2010; **6**(10): 1853–62.
PubMed Abstract | Publisher Full Text
24. Dweep H, Sticht C, Pandey P, *et al.*: **miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes.** *J Biomed Inform.* 2011; **44**(5): 839–47.
PubMed Abstract | Publisher Full Text
25. Pyysalo S, Airola A, Heimonen J, *et al.*: **Comparative analysis of five protein-protein interaction corpora.** *BMC Bioinformatics.* 2008; **9**(Suppl 3): S6.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Ogren PV: **Knowtator: A Protégé plug-in for annotated corpus construction**. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations.* New York, Association for Computational Linguistics. 2006; 273–275.
Publisher Full Text
27. Gennari JH, Musen MA, Fergerson RW, *et al.*: **The evolution of Protégé: an environment for knowledge-based systems development.** *Int J Hum Comput Stud.* 2003; **58**(1): 89–123.
Publisher Full Text
28. Shah PK, Perez-Iratxeta C, Bork P, *et al.*: **Information extraction from full text scientific articles: where are the keywords?** *BMC Bioinformatics.* **4**: 20.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Oualline S: **Vi iMproved**. New Riders Publishing, Thousand Oaks, CA, USA, 2001.
Reference Source
30. Brown EG, Wood L, Wood S: **The medical dictionary for regulatory activities**

(MedDRA). *Drug Saf.* 1999; **20**(2): 109–17.
PubMed Abstract | Publisher Full Text

31. Fluck J, Mevissen HT, Oster M, *et al.*: **ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries**. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, Spain. 2007; 149–151.
Reference Source

32. Cortes C, Vapnik V: **Support-vector networks**. In *Machine Learning*, 1995; **20**(3): 273–297.
Publisher Full Text

33. Fan E, Chang K, Hsieh C, *et al.*: **LIBLINEAR: A Library for Large Linear Classification.** *Machine Learning Research.* 2008; **9**: 1871–1874.
Reference Source

34. John GH, Langley P: **Estimating continuous distributions in Bayesian classifiers**. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, UAI'95, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. 1995; 338–345.
Reference Source

35. Bobić T, Klinger R, Thomas P, *et al.*: **Improving distantly supervised extraction of drug-drug and protein-protein interactions**. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, Avignon, France, Association for Computational Linguistics. 2012; 35–43.
Reference Source

36. Porter M: **An algorithm for suffix stripping.** *Program.* 1980; **14**(3): 130–137.
Publisher Full Text

37. Yu H, Qian L, Zhou G, *et al.*: **Extracting protein-protein interaction from biomedical text using additional shallow parsing information**. In *Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference on*, 2009; 1–5.
Publisher Full Text

38. Yang Z, Lin H, Li Y: **BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using svm and rich feature sets.** *J Biomed Inform.* 2010; **43**(1): 88–96.
PubMed Abstract | Publisher Full Text

39. De Marneffe MC, Manning CD: **Stanford typed dependencies manual**. 2010.
Reference Source

40. Bunescu RC, Mooney RJ: **A shortest path dependency kernel for relation extraction**. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics. HLT '05, Stroudsburg, PA, USA. 2005; 724–731.
Publisher Full Text

41. Thies W, Bleiler L, Alzheimer's Association: **2011 Alzheimer's disease facts and figures.** *Alzheimers Dement.* 2011; **7**(2): 208–244.
PubMed Abstract | Publisher Full Text

42. Cheng L, Quek C, Sun X, *et al.*: **Deep-sequencing of microRNA associated with Alzheimer's disease in biological fluids: From biomarker discovery to diagnostic practice.** *Frontiers in Genetics.* 2013; **4**(150).

43. Wang WX, Rajeev BW, Stromberg AJ, *et al.*: **The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1.** *J Neurosci.* 2008; **28**(5): 1213–23.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Hébert SS, Horré K, Nicolaï L, *et al.*: **Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression.** *Proc Nat Acad Sci U S A.* 2008; **105**(17): 6415–6420.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Bagewadi S, Bobi T, Hofmann-Apitius M, *et al.*: **Dataset, 1 version 2 in: Detecting miRNA Mentions and Relations in Biomedical Literature.** *F1000Research.*
Data Source

4.2.1 Supplementary Tables

Table A. Count of entries available in the dictionaries.

			Genes/Proteins			Relation Trigger	
Dictionaries	MeSHAbbr	MedDRA	Original	Processed	Species	Dictionary	Spelling variants
Entries	4,683	15,436	39,386	34,392	158	207	386
Synonyms	60,554	54,885	721,455	677,943	1,330	-	-

Table B. Performance evaluation of the disease dictionaries on training corpus (CM=Complete match and PM=Partial match).

	R		P		F_1	
	CM	PM	CM	PM	CM	PM
MeSH	0.62	0.74	0.70	0.70	0.66	0.72
MedDRA	0.50	0.59	0.73	0.86	0.60	0.70
MeSHAbbr	0.72	0.85	0.77	0.91	0.74	0.88

Table C. Results of tri-occurrence based approach for relation extraction using entities identified by ProMiner.

Interacting Entity Classes	Training Corpus			Test Corpus		
	R	P	F_1	R	P	F_1
NonSpMiR-D	0.60	0.21	0.31	0.49	0.3	0.38
NonSpMiR-GP	0.53	0.35	0.42	0.87	0.36	0.51
SpMiR-D	0.62	0.31	0.41	0.52	0.38	0.44
SpMiR-GP	0.43	0.41	0.42	0.45	0.41	0.43

Table D. Comparison of different relation extraction approaches on test corpus.

	NonSpMiR-D			NonSpMiR-GP			SpMiR-D			SpMiR-GP		
RE approaches	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
Co-occurrence	1.00	0.20	0.34	1.00	0.28	0.44	1.00	0.41	0.58	1.00	0.53	0.69
Tri-occurrence	1.00	0.35	0.51	1.00	0.41	0.58	1.00	0.50	0.66	1.00	0.58	0.73
ProMiner NER	0.49	0.3	0.38	0.87	0.36	0.51	0.52	0.38	0.44	0.45	0.41	0.43
Machine Learning Approaches												
LIBLINEAR	0.51	0.56	0.53	0.64	0.57	0.60	0.73	0.62	0.67	0.87	0.68	0.76
SVM	0.49	0.55	0.52	0.63	0.56	0.59	0.73	0.61	0.66	0.86	0.68	0.76
Naive Bayes	0.60	0.52	0.56	0.75	0.57	0.65	0.65	0.62	0.63	0.90	0.64	0.75

4.3 Summary

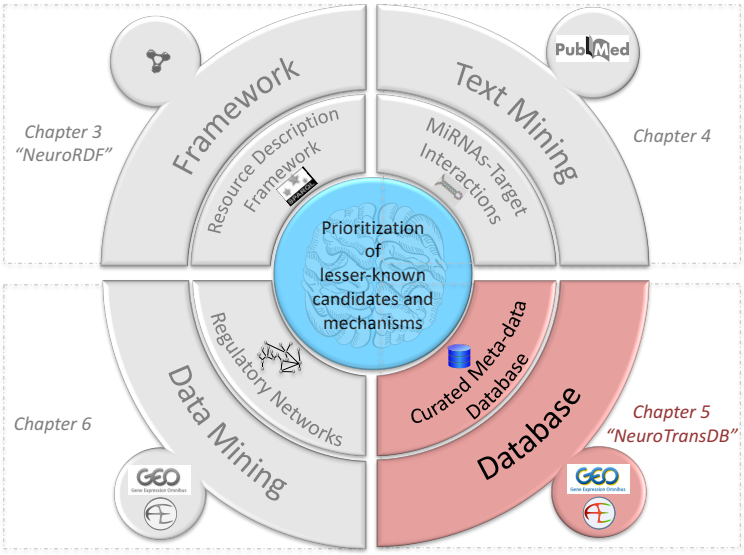
This publication details the development of the text-mining methods to automatically extract miRNAs and its associations from the text. To build these automated methods, firstly a training corpus was developed. A detailed description of the literature curation workflow and the annotated entities are provided. Here two types of miRNAs were distinguished for annotation: specific and non-specific. Specific miRNAs represent the

mentions that can directly be mapped to miRBase database identifiers, whereas non-specific mentions aid in co-reference resolution. Annotated relations include miRNA-disease and miRNA-gene associations. To our knowledge, this is the first work which has published a comprehensive corpus for miRNA research. Currently, this corpus serves as the gold standard for developing text-mining-based methods for miRNA research.

The automated text-mining workflow was implemented in the ProMiner tool. Specific miRNA mentions were identified using a set of regular expressions. Dictionaries were developed to extract non-specific miRNAs and relation trigger terms. Existing dictionaries were adapted for extracting genes/proteins and disease mentions. Three relation extraction approaches were evaluated: co-occurrence, tri-occurrence and machine learning-based. Among these, tri-occurrence-based and machine-learning approaches were comparable, the latter with a slightly better precision than the former.

To showcase the potential of the developed approach, an AD use-case was presented. The developed approach outperformed well-known databases: miR2Disease and miRSel. To conclude, this work serves as an important benchmark for current and upcoming attempts at aggregating highly specific and relevant miRNAs information.

Chapter 5 Discovery-based Data Harvesting



5.1 Introduction

With rapid advancements in high-throughput technologies, a number of omics studies have grown over the years. Indeed, the cumulative effort of several NDD researchers to decipher the underlying aetiology has added to these mountains of public data. More and more of these datasets are made available to the community to accelerate the much needed discoveries and innovations to aid NDD patients. Hence, there is an urgent need to organize the existing digital data objects for sharing and reusability. If reused and reanalyzed from a different perspective, such a large collection of data has huge potential to uncover hidden knowledge. However, only a small subset of these data holdings support efficient reusability. A potential reason is that most of the repositories allow data search based on the metadata provided by the researchers, which is often far from complete and are not adapted to standard ontologies. Moreover, it still requires significant effort to retrieve a relevant study or obtain experiment-related metadata information. In addition, these databases lack context-specific metadata information that is vital for comprehending disease-specific studies. For example, in NDD it is important to understand symptoms or comorbidities of the patients, treatment methods applied to induce neurodegeneration in animal models, etc. To retrieve pertinent studies and thus obtain the most context-specific

datasets for analysis, there is a need to harvest more granular, formalised, searchable, and context-specific metadata information.

In the following publication, a NDD-specific metadata database has been developed, named *NeuroTransDB*. Our study outlines the challenges faced during precise retrieval of omics data from GEO and ArrayExpress. Several examples have been provided that show the lack of compliance by data submitters. In addition, it describes the need for additional metadata fields needed for NDD research with good data coverage in humans and animal models. This publication emphasises on the approach and effort required to achieve FAIRness (Findable, Accessible, Interoperable, Reusable) of exisitng data in NDD.

5.2 Publication

DATABASE
The Journal of Biological Databases and Curation

Database, 2015, 1–17
doi: 10.1093/database/bav099
Original Article

Original Article

NeuroTransDB: highly curated and structured transcriptomic metadata for neurodegenerative diseases

Shweta Bagewadi[1,2,*], Subash Adhikari[3], Anjani Dhrangadhariya[1,2,†], Afroza Khanam Irin[1,2,†], Christian Ebeling[1], Aishwarya Alex Namasivayam[4], Matthew Page[5], Martin Hofmann-Apitius[1,2] and Philipp Senger[1]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany, [2]Rheinische Friedrich-Wilhelms-Universitaet Bonn, Bonn-Aachen International Center for Information Technology, 53113, Bonn, Germany, [3]Department of Chemistry, South University of Science and Technology of China, No 1088, Xueyuan Road, Xili, Shenzhen, China, [4]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg and [5]Translational Bioinformatics, UCB Pharma, 216 Bath Rd, Slough SL1 3WE, United Kingdom

*Corresponding author: Tel: +49-2241-14-2360, Fax: +49-2241-14-2656, Email: shweta.bagewadi@scai.fraunhofer.de

Correspondence may also be addressed to Philipp Senger. Tel: +49-2241-14-2280, Fax: +49-2241-14-2656, Email: philipp.senger@scai.fraunhofer.de

†These authors contributed equally to this work.

Citation details: Bagewadi,S., Adhikari,S., Dhrangadhariya,A. *et al. NeuroTransDB*: highly curated and structured transcriptomic metadata for neurodegenerative diseases. *Database* (2015) Vol. 2015: article ID bav099; doi:10.1093/database/bav099

Received 2 April 2015; Revised 7 September 2015; Accepted 10 September 2015

Abstract

Neurodegenerative diseases are chronic debilitating conditions, characterized by progressive loss of neurons that represent a significant health care burden as the global elderly population continues to grow. Over the past decade, high-throughput technologies such as the Affymetrix GeneChip microarrays have provided new perspectives into the pathomechanisms underlying neurodegeneration. Public transcriptomic data repositories, namely Gene Expression Omnibus and curated ArrayExpress, enable researchers to conduct integrative meta-analysis; increasing the power to detect differentially regulated genes in disease and explore patterns of gene dysregulation across biologically related studies. The reliability of retrospective, large-scale integrative analyses depends on an appropriate combination of related datasets, in turn requiring detailed meta-annotations capturing the experimental setup. In most cases, we observe huge variation in compliance to defined standards for submitted metadata in public databases. Much of the information to complete, or refine meta-annotations are distributed in the associated publications. For example, tissue preparation or comorbidity information is frequently described in an article's supplementary

tables. Several value-added databases have employed additional manual efforts to overcome this limitation. However, none of these databases explicate annotations that distinguish human and animal models in neurodegeneration context. Therefore, adopting a more specific disease focus, in combination with dedicated disease ontologies, will better empower the selection of comparable studies with refined annotations to address the research question at hand. In this article, we describe the detailed development of *NeuroTransDB,* a manually curated database containing metadata annotations for neurodegenerative studies. The database contains more than 20 dimensions of metadata annotations within 31 mouse, 5 rat and 45 human studies, defined in collaboration with domain disease experts. We elucidate the step-by-step guidelines used to critically prioritize studies from public archives and their metadata curation and discuss the key challenges encountered. Curated metadata for Alzheimer's disease gene expression studies are available for download.

Database URL: www.scai.fraunhofer.de/NeuroTransDB.html

Background

Considerable effort by the global research community has been dedicated to addressing a limited understanding of the pathogenic events underlying neurodegenerative disease (NDD) (1, 2). The cumulative output of these efforts has established an increased amount of deposited molecular data and published knowledge. As life expectancy continues to rise and treatment options for NDD remain limited, there is an increasing urgency to translate this amassed molecular data into biomarker tools for early diagnosis; to open the possibility of disease altering and preventative therapy (3, 4). Furthermore, biomarkers aiding the decision-making process for therapies targeting specific pathophysiological mechanisms will help to address the high drug attrition rate in the NDD pharmaceutical industry. Informatic efforts to facilitate the integration and interrogation of the distributed molecular data legacy for NDD can enable a systematic and objective prioritization of molecular protagonists and therefore potential biomarkers in NDD (5–8).

In this direction, we have previously developed a semantic framework, called *NeuroRDF* (9), for integration of heterogeneous molecular data types, extracted from biomedical literature, transcriptomic repositories and bespoke databases. *NeuroRDF* enables researchers to formulate biological questions that relate to the interplay of different facets of molecular biology as a formalized query. Even today, the most abundant source of quantitative molecular data remains transcriptomic data, which can support hypothesis-free, elucidation of biological function (10). When the same biological function is replicated in additional expression data sets, it increases the plausibility of the derived hypothesis (11).

The inaccessibility of the brain is a significant barrier to molecular analysis of NDD and this frequently limits the availability of samples from post-mortem tissue (12, 13). This is evident when simply comparing the availability of NDD studies to other disease domains, like cancer (14), in public archives such as Gene Expression Omnibus (GEO) (15) and ArrayExpress (16) (see Supplementary Figure S1). For instance, GEO contains 157 NDD studies in contrast to 16,910 cancer studies. Therefore, animal models are an important complement to human-derived samples but are at best an incomplete reflection of the human conditions. Assessing the biological complementarity of studies is important when considering a meta-analysis. Such an assessment can be a cumbersome process as searching in these public repositories is principally based on free text. Additionally, limited adoption of controlled vocabularies, such as the Experimental Factor Ontology (EFO) (17), to describe the metadata fields and lack of compliance to defined standards (18) has contributed to the dilemma. This has resulted in metadata being scattered as unstructured prose in public databases and as additional annotations, widely distributed in originating publications. Moreover, applying automated methods to retrieve information from these databases could compromise on the accuracy. On the other hand, capturing missing annotations through the manual curation can incur huge costs of trained labour.

Capturing the associated metadata in a standardized and precise fashion will empower integrative analysis by helping to control sources of variability that do not relate to the hypothesis under investigation (11, 19–21). Ober *et al.* (22) have reported on differing gene-expression patterns related to gender and suggest gender-specific gene architectures that underlay pathological phenotypes. Li *et al.* (23) observed distinct expression patterns, strongly correlated with tissue pH of the studied subjects; these patterns are not random but dependent on the cause of death: brief or prolonged agonal states. Thus, studies enriched

with metadata annotations provide the power to obtain more precise differential estimates.

Related work

Numerous approaches have been proposed to tackle the problem of identifying relevant gene-expression studies and annotating metadata information resulting in several databases, web servers and data exploration tools. These (value added) databases differ from one another based on their objectives, information content and mode of query.

AnnotCompute (24) is an information discovery platform that allows effective querying and grouping of similar experiments from ArrayExpress, based on conceptual dissimilarity. The dissimilarity measure used, Jaccard distance, which is derived from the MAGE-TAB fields submitted by the data owners. Another tool, Microarray Retriever (MaRe) (25) enables simultaneous querying and batch retrieval from both GEO and ArrayExpress for a range of common attributes (e.g. authors, species) (MAGE-TAB is a submission template, tab-delimited, for loading functional genomics data into ArrayExpress. https://www.ebi.ac.uk/fgpt/magetab/help/). GEOmetadb (26) is a downloadable database of structured GEO metadata with programmatic querying libraries in both R and MATLAB. However, all the above-mentioned resources suffer from a common limitation: they rely completely on the submitted data and do not provide solutions for missing metadata information.

Several value-added databases invest manual curation effort to enrich metadata information for gene-expression studies. Many Microbe Microarrays Database (M^3D) (27) contains manually curated metadata, retrieved from the originating publications, for three microbial species, conducted on Affymetrix platforms. Similarly, the Oncomine database (28) contains extensive, standardized and curated human cancer microarray data. A-MADMAN (19); an open source web application, mediates batch retrieval and reannotation of Affymetrix experiments contained in GEO for integrative analyses. Microarray meta-analysis database (M^2DB) (11) contains curated single-channel human Affymetrix experiments (from GEO, ArrayExpress and literature); categorized into five clinical characteristics, representing disease state and sample origin. However, experiments with missing link to the published paper in GEO and ArrayExpress were excluded. A substantial paucity of sample associated gender information in GEO and ArrayExpress motivated Buckberry *et al.* (29) to develop a *R* package, *massiR* (MicroArray Sample Sex Identifier) to label the missing and mislabelled samples retrospectively with gender information, based on data from Y chromosome probes. Apart from publicly available resources, there are various commercial products that contain manually curated transcriptomic metadata: NextBio, Genevestigator and InSilicoDB (30) (http://www.nextbio.com/b/nextbioCorp.nb and https://genevestigator.com/gv/). However, none of the above databases are optimized to capture detailed metadata specific to neurodegenerative disease. In addition, these databases fail to handle species-specific annotations; especially treatments applied on animal models to partially explicate or treat human-related NDD mechanisms, which may strongly contribute to increase the predictive power of translating preclinical results in NDD drug trials.

Here, we describe the detailed development of *NeuroTransDB,* a manually curated database containing metadata annotations for neurodegenerative studies and an enabling resource for supporting integrative studies across human, mouse and rat species. The participation of our group, at Fraunhofer Institute SCAI, in projects funded by the Neuroallianz Consortium (a part of the BioPharma initiative of the German Ministry of Education and Research) and the evident lack of a comprehensive NDD specific metadata archive has motivated us to develop **Neuro**degenerative **Trans**criptomic **D**ata**B**ase (*NeuroTransDB*) (http://www.neuroallianz.de/en/mission.html). This database now contains more than 20 dimensions of metadata annotations for human studies, as well as mouse and rat models, defined in agreement with disease experts. To demonstrate our approach, we chose to highlight Alzheimer's disease for this publication because it depicts a wide spectrum of the possible annotations across different types of metadata in neurodegeneration. Additionally, we have applied the same approach to all publicly available Parkinson's and Epilepsy studies, which shows that the overall approach is unspecific to the disease. However, the curated data for these two diseases will be released in the future under the terms of a Neuroallianz agreement. The database is updated every six months using highly trained curators. An interactive graphical user interface to access this data is currently being developed as part of the AETIONOMY IMI project (http://www.aetionomy.eu).

Curation of gene-expression studies: prerequisites, key issues and solutions

This section discusses the workflow we followed to retrieve relevant gene-expression datasets and to generate detailed metadata annotations for each study (Figure 1). First, we retrieved all functional genomics studies from GEO and ArrayExpress that reference Alzheimer's disease (AD) or a set of AD synonyms, along with the provided metadata (*cf.* Data Retrieval section). Each study

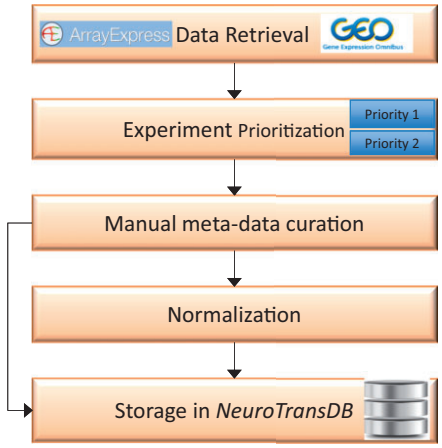


Figure 1. Overall workflow for curation of gene expression studies related to neurodegeneration from public archives. The first step involves automated retrieval of gene expression studies (along with metadata) from public archives such as GEO, and ArrayExpress. The related studies were further assigned to one of the two prioritization classes (priority 1 or priority 2), based on the specific experimental variables. Next, manual curation was applied to capture missing metadata information on priority 1 studies. All the harvested metadata was normalized using standard vocabularies. Both raw and normalized data are stored in *NeuroTransDB*.

was then prioritised (*cf.* Experiment Prioritization section) based on the disease relevancy, experimental type and sample source. Only studies in the top prioritization category were subjected to rigorous, semiautomated metadata curation (*cf.* Metadata Curation section). Annotations are standardized by reference to controlled vocabularies for each extracted metadata dimension (*cf.* Normalization of Metadata Annotations section). The curated Alzheimer's data is stored in *NeuroTransDB*, but in principle the proposed workflow can be applied with little adaptation to any disease indication, especially NDD.

Primary data resources

Together the GEO and ArrayExpress databases constitute a wealth of gene expression studies and are commonly reused for validating new hypotheses and identifying novel signatures through meta-analysis by multi-data integration (11). GEO is the largest public repository of functional genomic data; maintained by the National Center for Biotechnology Information (NCBI) in the USA. ArrayExpress is the European counterpart of GEO and consists of manually curated experimental information imported from GEO, in addition to the data that are directly submitted by the researchers. To support reuse of the deposited studies, each repository adheres to annotation standards for submission of transcriptomic data: 'Minimum Information about a Microarray Experiment' (MIAME) and 'Minimum Information about a high-throughput nucleotide SEQuencing Experiment' (MINSEQE) (http://fged.org/projects/miame/ and http://www.fged.org/projects/minseqe/). GEO allows data submission in Excel, SOFT or MINiML format and ArrayExpress as MAGE-TAB through Annotare webform tool (http://www.ncbi.nlm.nih.gov/geo/info/submission.html and http://www.ebi.ac.uk/arrayexpress/submit/overview.html).

Curation team

An obvious prerequisite for any curation process is to have access to specially trained personnel, who understand the key attributes required to adequately describe an expression experiment and are able to complete these attributes by reference to appropriate resources (31). Such individuals are known as biocurators. We assembled a team of candidate biocurators who have adequate biological experience. Each biocurator underwent extensive training in the fundamentals of curation, including the basics of gene expression study design, outlined by experts, scientists and disease experts. Clear curation guidelines (see Experiment Prioritization and Metadata Curation section) and a weekly meeting of the biocurators with one of the experts ensured good quality, consistency, and uniformity in curation procedure. In addition, this provided an opportunity to get feedback from the biocurators for improving and updating the defined guidelines. To keep abreast and eliminate any bias, the curated data was regularly exchanged between them for good interannotator agreement. The experts resolve any disagreement that may arise between the curators.

Data retrieval

Putative AD studies were programmatically retrieved from GEO and ArrayExpress by applying a recall-optimized keyword search approach, *cf.* Figure 2. The keywords include a set of AD synonyms such as 'Alzheimer', 'Alzheimer's' or 'AD' in combination with a species filter. Since ArrayExpress imports and curates the majority of GEO experiments, we firstly queried the former through its REST service (http://www.ebi.ac.uk/arrayexpress/help/programmatic_access.html). Conjointly, we further queried GEO using the *eSearch* Entrez Programming Utilities (E-utils) service to fetch additional identifiers (IDs), which were not picked up by the previous query (http://www.ncbi.nlm.nih.gov/geo/info/geo_paccess.html). The final list of unified experiment IDs was downloaded

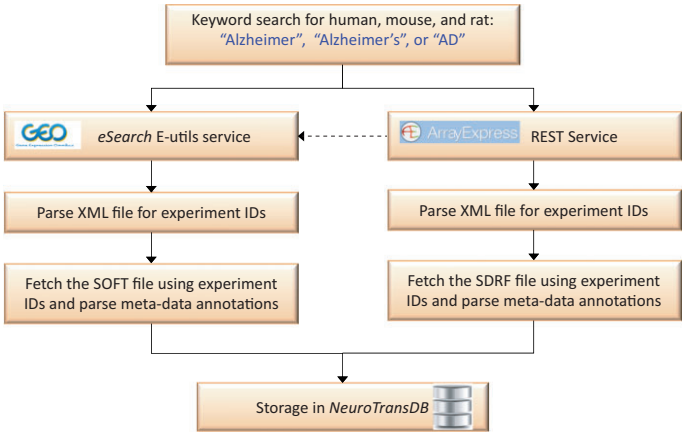


Figure 2. Automated data retrieval of Alzheimer's Disease specific gene expression studies from ArrayExpress and GEO. Here, the dotted line represents the sequence of query performed. Alzheimer's disease specific experiment IDs were automatically retrieved from GEO and ArrayExpress, using keywords, through *eSearch* and REST service respectively. Metadata information was extracted by automatically parsing sample information files (SDRF and SOFT) of these experiment IDs.

(along with their metadata) and stored in *NeuroTransDB*. Metadata information was captured from Sample and Data Relationship Format (SDRF) file of ArrayExpress and SOFT file of GEO (https://www.ebi.ac.uk/fgpt/magetab/help/creating_a_sdrf.html and http://www.ncbi.nlm.nih.gov/geo/info/soft.html). The above-described steps are fully automated; enabling an automatic update procedure we run every 6 months to obtain new published studies.

Experiment prioritization

For integrative meta-analysis, combining studies that address the same objectives could minimize biases from cohort selection (inclusion and exclusion criteria) and other design effects. Anatomical and functional heterogeneity arising from experimental sample source, imposes yet another challenge for integrative analysis. Moreover, keyword-based, recall optimized retrieval of experiments does not guarantee its clinical relevancy to the queried indication or organism. Thus, we propose a straightforward binning approach to select potentially eligible studies for AD as illustrated in Figure 3.

Firstly, we identified experiments relevant to AD indication, if not relevant we mark them as unrelated (referred as AD3 in the database). Relevancy is defined on the basis of the experiment's characteristics: investigation on AD mechanism, AD associated mechanism, AD genes or contains samples that belong to direct or implicated effects of or on AD. For example, GSE4757 is relevant to AD since it investigates the role of neurofibrillary tangle formation in Alzheimer patients between normal and affected neurons.

The retained AD-related experiment IDs were manually classified by biocurators into one of the two-prioritization categories (*cf*. Figure 3). To support this process, a set of classification rules were devised that capture two important considerations: organism specificity and source of the samples used in the study. Although curation with regards to these considerations is of obvious importance, no previously published guidelines were available for reference. To our knowledge, this is the first work where such a guideline has been explicitly detailed. A simplified description of the classification rules adopted for AD disease prioritization is provided below:

Priority 1

- Experiments that study AD pathophysiology in *in vivo* systems
- Studies containing samples from:
  - Human AD patients such as blood, brain tissue, serum, etc.
  - Animal model samples such as mouse brain tissue or rat brain, e.g. C57BL/6 mice, Sprague–Dawley rat, etc.
  - Animal models modified to study the role of an AD gene (knock-out models), or AD mechanism (transfected models), or diet/drug treatments (treated models), such as TgAPP23, APLP2-KO mice, etc.
- Experiments containing only healthy/normal samples from human/mouse/rat that are a part of a bigger study investigating AD

Priority 2

- Experiments that study AD pathophysiology in *in vitro* systems

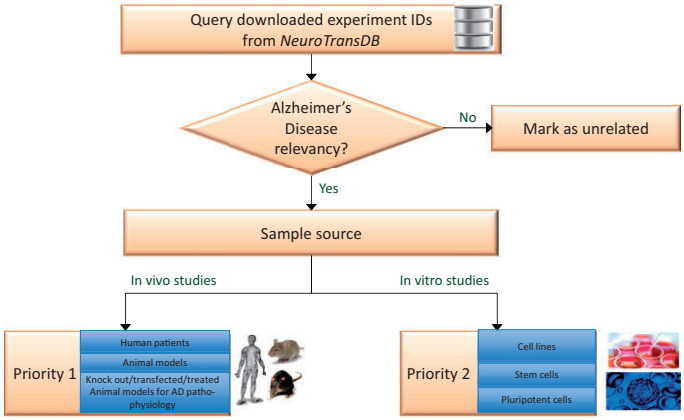


Figure 3. Experiment prioritization for metadata curation in *NeuroTransDB*. All the downloaded Alzheimer's Disease experiments were first checked for their disease relevancy. Those experiments which were falsely retrieved, are marked as unrelated. The remaining experiments were classified into one of two priority classes based on the experiment type: In vivo or In vitro studies. For priority 1, we considered direct/primary samples from human or animal models such as brain tissue, blood, etc. Experiments that were conducted on derived sample sources such as cell lines, were put into priority 2 class.

- Studies containing samples from derived or cultures sources:
  - Cell lines
  - Pluripotent cells
  - Stem cells

Incorrect organism or disease specificity

Although the experiment retrieval step was restricted to a specific organism and disease conditions, we observed differing levels of specificity. For example, some mouse studies were retrieved when querying for human studies. Similarly, we obtained experiments for related diseases such as Parkinson's disease, or diabetes, when querying for AD. Therefore, during study prioritization it was important to confirm the species of origin and relevancy of the study to AD. It's also possible that keyword-based retrieval may miss AD studies due to incorrect disease or organism tagging. However, we did not perform an exhaustive search for such falsely ignored studies, since it would require immense human effort.

Ambiguous species designation

In some studies, human cells such as embryonic stem cells are injected into animal models and post-mortem samples from these animal models are extracted for transcriptomic analysis (e.g. GSE32658 experiment in GEO). Such a study could arguably be classified as either human priority 2 or mouse priority 1. After several discussions, we concluded to prioritize such experiments based on the organism from which the final sample was extracted. In this case, although the mouse was grafted with human tissue, we prioritized it to mouse priority 1.

Superseries redundancy

During prioritization, we retrieved several superseries experiments from GEO. Manual inspection revealed that not all the subseries IDs of these superseries experiments were retrieved (see Data Retrieval section) (A SuperSeries is simply a wrapper to group of related Series (typically described in a single publication). It facilitates access to the entire dataset, and establishes a convenient reference entry that can be quoted in the publication (definition provided by the GEO team, as of 27 October 2014) and a subseries is an experiment that is a part of superseries.). With careful manual inspection, we included missing subseries, further subjected to priorization. Conversely, if the inclusion of superseries resulted in the duplication of experiments, we removed the duplicates. Having assigned priority categories to all retrieved AD studies, further metadata curation was focused on the priority 1 studies. Metadata curation steps are described below.

Metadata curation

Precisely and comprehensively capturing the accessory information for a transcriptomic study as meta-annotations, is an important precursor to identification of comparable experiments that address the biological question at hand. Unfortunately, the current, general, submission standards do not cater to the needs of metadata annotations, specific to a disease domain, during submission. In subsequent

sections, we discuss the metadata curation for NDD and key issues faced during the process.

Metadata annotation fields

We assembled a list of metadata annotations determined to be important for evaluating NDD studies in a process involving consultation by NDD domain experts. All the metadata fields were categorized as organism attributes and sample annotations, based on their relevancy to organism or sample source. Table 1 provides detailed descriptions of curated metadata fields including examples for human, mouse and rat.

Several animal models and *in vitro* systems have been defined that partially mimic the human diseased conditions. Animal models provide experimentally tractable systems for interrogating NDD, however, not all animal models faithfully mimic human pathophysiology. A dedicated set of meta-annotation was defined for NDD animal models to support assessments of inter-study comparability and translatability to human disease, *cf.* Table 2 These fields were defined with assistance from biologists and disease experts from industry.

Metadata curation workflow

To capture all the relevant meta-annotations, we designed a semiautomated curation workflow, illustrated in Figure 4. Firstly, we automatically retrieved all the available meta-annotations from GEO and ArrayExpress (*cf.* Figure 2). Annotations were captured in an Excel template as shown in Supplementary Figure S2 (A) and confirmed by our trained curators to rectify any inaccuracies.

To capture incomplete and newly defined meta-annotations, we followed a two-step approach. First, we check if the required meta-annotation entries are directly available in GEO, GEO2R or ArrayExpress (http://www.ncbi.nlm.nih.gov/geo/geo2r/). Where the required information is complete, we directly update *NeuroTransDB,* otherwise we move to a second step to manually harvest information for missing annotations. Missing information is retrieved from the originating publications and associated Supplementary files. When necessary, corresponding authors were contacted to request missing entries. The list of experiment IDs where we contacted the authors for further information, along with reason of contact (priority 1 experiments only) are provided in Supplementary Table S1. In most cases, the corresponding author or one of the coauthors responded to our queries; whereas, in few other cases the email addresses no longer remained valid. In the event that the authors do not respond or we were unable to contact them, information in primarily deposited database is used as the final authorative source. Once all the relevant data was captured, we updated the annotations in *NeuroTransDB*. If needed, we updated our automated retrieval iteratively.

To demonstrate the metadata curation process, here we relate our experience with study GSE36980 that includes a total of 79 samples. Common MIAME annotations such as gender, age and sample tissue were automatically captured from ArrayExpress and GEO. The associated publication contained further useful information on the enrolled patients, namely: disease stage, post mortem interval before sample extraction and preservation, pathological diagnosis and whether the patient suffered from comorbidities such as diabetes. This information was located in Supplementary File S2 of the associated publication (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128707/bin/supp_bht101_bht101supp_table2.xls). However, lack of a common ID to enable mapping between the sample entries in GEO and the associated Supplementary File S2 impeded curation. For example, sample GSM907797 in GEO is annotated as being derived from a 95-year-old female patient. However, in their Supplementary file, there are two entries that contain information for patients with same age and gender. The 'No.' column, assumed to be patient ID, in the Supplementary file was not helpful for mapping, since it was not mentioned in GEO. Thus, we contacted the authors for the missing link. They provided us an additional Excel sheet where the GEO sample ID was mapped to the 'No.' column in the Supplementary file (*cf.* Supplementary Figure S2 (B) and (C)). As a consequence, we achieved a 28.5% increase in the missing metadata information (*cf.* Table 1 for total number of fields) after contacting the authors.

Automated meta-annotation retrieval challenges

During automated retrieval of metadata fields, we observed several alternate representations of information for certain annotation types in the archives. For example, age information can be provided in the Characteristics section of GEO or ArrayExpress as 'age: 57 years' or 'Stage IV, male, 57 years' and so on. We attempted to prenormalize these diverse representations and automatically extract the correct information, however, due to the heterogeneity in data representation, manual curation was still required.

Although ArrayExpress and GEO provide programmatic access to their meta-annotations, much essential information appears in fields meant for general categories. For example, information about the sample source and clinical disease presentation appear in the sample title 'PBMC mRNA from Alzheimer's disease patient 2'. Adhering to the standard submission protocol for data

Table 1. Detailed description of Neurodegenerative disease metadata fields outlined for human, mouse and rat

Annotation type	Metadata fields	Description of the annotation	Relevancy for NDD	Examples	References
Organism attributes	Age	Age of the organism	Main factor for predisposition to disease	84 years, 9 months	(32–35)
	Gender	Gender of the organism	Possible disproportionate effect arising from difference in anatomy and hormonal composition	Male, female	(36, 37)
	Phenotype	Clinical phenotypes of the organism from which the sample was extracted	Supports comparative analysis for underlying pathomechanisms based on the observable/measurable characteristics	Healthy control, early incipient	(38)
	Behavioural Effect	Description of behavioural changes occurring in organism due to treatment or other effects	Impact of developed drug or other environmental factors to treat or reduce the disease/disease symptoms	Reduced agitation/aggression	(39, 40)
	Disease type	The disease occurrence is due to hereditary or effect of environmental factors	To distinguish the genetic variability and complexity between the two types during analysis	Sporadic, familial	(41)
	Stage	Disease stage of the organism from which the sample was extracted	Capability to distinguish severity of the affected disease	Incipient, severe, BRAAK II	(42)
	Cause of death	Reason for the organism's death	To determine if Alzheimer's disease or its associated comorbidities are major contributors to death rate	Respiratory disorder	(43)
	Comorbidity	Existence of another disease other than Alzheimer's	To determine the impact of another disease on Alzheimer's disease aetiology and progression	Type 2 diabetes	(44, 45)
Sample annotations	Post mortem duration (PMD)	Duration from death till the sample extraction from the dead organism	To assess quality and reliability of the sample obtained by measuring RNA integrity that is influenced by natural degradation of the sample after death	2.5 hours	(46, 47)
	pH	pH value of the extracted post-mortem sample	Indicator of agonal status and RNA integrity	6	(48–50)
	Functional effect	Description of functional effects observed	Observed changes such as gene expression, post-translation, or pathway due to external effects	Decreased expression of BDNF gene, reduced Aβ toxicity	(51, 52)
	Brain region	Brain region of the extracted sample	Provides information of pathogenesis and disease progression, as AD does not affect all the brain regions simultaneously	Hippocampus	(53, 54)
	Cell and cell parts	Type of cells or cell parts extracted from the sample for analysis (if any)	To determine cell type specific expression influencing pathogenesis and regional vulnerability	Synaptoneurosome, neurons and astrocyte	(55, 56)
	Body Fluid	Type of body fluid used for analysis	Could serve as biomarkers for early diagnosis and therapy monitoring	CSF, blood	(57–59)

The table provides a list of metadata fields, confirmed by disease experts, critical for NDD meta-analysis. The selected fields are classified as organism attributes and sample annotations based on their relevancy to organism or sample source.

Table 2. Detailed description of additional metadata fields, defined specifically for mouse and rat models

Annotation type	Metadata fields	Description of the annotation	Relevancy for NDD	Examples	References
Organism attributes	Physical injury	Method used to cause brain injury in animal models	Consideration for analysing plaque formation in animal models to mimic disease symptoms in human	Traumatic brain injury, ischemia reperfusion injury	(60, 61)
	Type of treatment	Description of chemical, drug, genetic or diet treatment	Consideration for determining the effect of treatment on animal models either to mimic or treat the disease/ symptoms	Long-term pioglitazone, BDNF treated	(62, 63)
	Dosage	Detailed description of the dosage associated with "type of treatment" description	Consideration of the right quantity of the substance for determining the effect on animal models either to mimic or treat the disease/ symptoms	Total polyphenol 6mg/ kg/day, received drinking water without ACE inhibitor	(64, 65)
	Mouse/rat strain name	Mouse model official or author given name	To determine the effect of different manipulated animal models in recapitulating key AD features capable of extrapolating to human studies	C57BL/6-129 hybrid, Sprague–Dawley rat	(66, 67)
	Mouse/rat weight	Weight of the animal model during analysis	Establishing a causative link to metabolic disruption	100–150 g	(68)

These additional metadata fields are defined by disease experts as critical for translating mouse/rat model outcomes to human, in the field of neurodegenerative diseases.

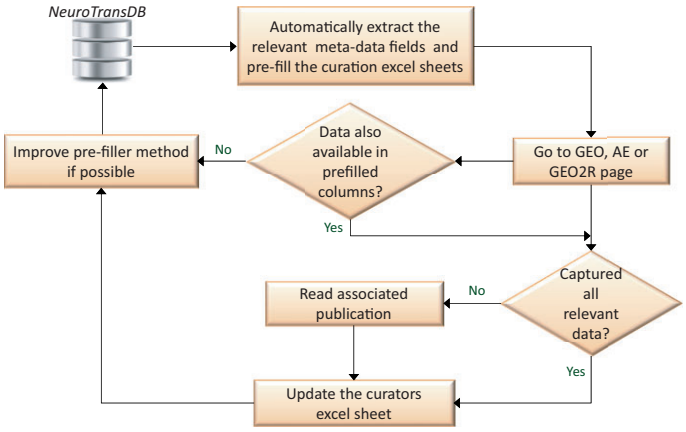


Figure 4. Semi-automated workflow for metadata curation. Automatically extracted metadata fields are rechecked by the curators. To capture the missing fields, curators browse through GEO, ArrayExpress (AE) or GEO2R experiment's description pages. For cases where the information is still incomplete, associated fulltext publications and their associated supplementary material are read. All the extracted metadata annotations are stored in *NeuroTransDB*. Intermediately, if feasible, automated extraction leverages on curator's experience for improvement. This process is carried out half yearly.

entry, this information should appear in the 'Characteristics column' of ArrayExpress and GEO. Again inconsistent adherence to annotation standards means that manual inspection is needed to capture correct and complete information from these archives.

Accessing linked publications

For annotation information that is not directly available from the source repositories, we refer to the associated full text publications. However, not all deposited studies link to an associated publication in PubMed, contributing to a

significant loss of information while curating. We attempted to overcome this by searching for an associated article using the study title with search engines such as SCAIView and/or Google (http://www.scaiview.com and https://www.google.com). Supplementary Figure S3 shows the percentage of articles that were retrieved with different search strategies. We are aware that not all the experiments in these databases are associated with published article (14%), but for 9% of the experiments (prioritized as 1) we were able to link them to publications through a title search. We strongly encourage study depositors to provide PubMed annotation whenever available to allow enhanced meta-annotation. Additionally, database owners should find a more robust way to update their resources.

Duplication and inconsistent sample counts

We observed differences in sample counts for some experiments between ArrayExpress and GEO, when downloaded automatically. For example, GSE49160 contained 36 samples in GEO and 72 samples in ArrayExpress. Following closer inspection at several similar experiments, we found that ArrayExpress duplicates sample IDs to provide separate links to different raw file formats or large raw files split into smaller ones (57%), processed raw files (17%), separate entry for each channel in double channel arrays (14%) and replicates (12%) (*cf.* Supplementary Figure S4); moreover, the duplicated samples mostly represented the same annotation information. Since, we used sample IDs as a unique entry in our database, the duplicated IDs were replaced with the last entry from the archive, in *NeuroTransDB*, as read by our algorithm; thus a risk of loosing the raw file or other non-duplicated annotation information.

Apart from duplication, occasionally some samples were missing in one archive relative to the other. For example, GSE47038 had some additional samples in ArrayExpress, which were not present in GEO. When we contacted the ArrayExpress team, they suggested that the experiment entry could be out of sync, since each entry from GEO is uploaded into ArrayExpress only once and is not updated if GEO deletes some samples later. However, they have now corrected the entry. This demonstrates a need for periodic review of study records in each database.

Missing RAW filenames

Public transcriptomic archives provide a gateway for the search and retrieval of studies for subsequent analysis outside of the platform. Therefore, one has to obtain the link between a sample's raw file name and corresponding phenotype. However, this is not the case when applying automated downloads. The majority of the raw file names present in public archives contain syntactical errors such as surrounded by brackets or separated by comma; moreover, such entries could be normalized through a simple script. In cases where no information about sample's raw file name is provided, manual intervention is required to link sample's raw file to its respective sample. This clearly indicates the need for standardization of the database entries for automation and to prevent loss of information.

Incorrect and incomplete metadata information

We also observed inconsistent meta-annotations between a study deposited in an archive and the information in the linked publication. In GEO for experiment GSE2880, the sample description page states that male Wistar rats have been used for the study. However, when we looked into the associated full text article, in the Methods section, the authors clearly mention using female Wistar rats (69). We are still waiting for the author's reply to correct the gender information for this entry. Another example is GSE18838, we observe that the ratio of male to female patients provided in GEO (male/female: 19/9) is different from that reported in the Supplementary file (male/female: 18/10); additionally, Supplementary Table S2 provides detailed challenges faced during mapping of age and gender information to samples. When searched in ArrayExpress, this experiment has been removed from the database, for unknown reasons. In yet another example, GSE36980, the age information for sample GSM907823 and GSM907823 vary between GEO (84 and 81 years, respectively) and ArrayExpress (74 and 86 years, respectively). From these anecdotal experiences, it is evident that one has to spend immense effort to obtain correct metadata information. Database owners and the submitters have to take utmost care to provide the correct data for reproducibility.

Information extraction from chained references

One further time consuming task included looking following chains of references to previous publications for human and animal model information such as mouse name, cross breeding steps applied and human subject information. In some cases, we had to tediously trace back 5–6 cross-referred publications to obtain the original source of information.

Normalization of metadata annotations

Meta-annotation involved the curation team extracting the original text as provided in GEO/ArrayExpress or in the published literature. We observed many different ways to express information for each annotation field, with obviously ramifications for accurate and efficient querying of *NeuroTransDB*. In an effort to standardize entries for different annotation fields specific controlled vocabularies were adopted during curation.

Age and gender normalization

We observed several different ways of representing age such as '24 yrs', '25 yo' and '23 ± 2 years old'. All age values were standardised by converting to simple decimal numbers, e.g. 24.00 for 24 years. Similarly for gender, we used a consistent representation of 'M' and 'F'. As an example, gender information for GSE33528 samples were reported in the associated article (40) as '70% of the participants were women'. Here, we annotated the information as '70% female'. Although the annotations such as ranges (e.g. '23 ± 2 years old'), ratios (male/female: 19/9), or percentages (70% female) (40) are study-level annotations, they were provided as sample level annotations; as they do not contribute to reasonable cohort selection we did not normalize them.

Phenotype, brain region and stage normalization

Disease phenotype and stage information contributes to specific details of clinical manifestations whereas the tissue source (hereafter brain region) caters to the sample origin. For all the curated phenotype mentions (human), we generated a binning scheme: diseased, control or treated. These binned terms were further mapped to controlled vocabularies provided by Alzheimer's Disease Ontology (ADO) (32). Other annotated terms that are not specific to AD were mapped to the Human Disease Ontology (33), Medical Subject Headings (MESH), Medicinal Dictionary for Regulatory Activities (MEDDRA) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) (34) ontologies (http://bioportal.bioontology.org/ontologies/MESH and http://bioportal.bioontology.org/ontologies/MEDDRA). This caters the need to query samples at a more abstract level, for downstream analysis. In total, for AD, we obtained 481 phenotype mentions assigned to at least one entry in the bins generated. Similarly, all the stage mentions (117 terms) were mapped to ADO, and ONTOAD (35). Mentions of brain region (41 unique terms) were tagged to Brain Region and Cell Type Terminology (BRCT) (http://bioportal.bioontology.org/ontologies/BRCT?p=summary). Please refer to Supplementary File S2 for detailed mapping of human annotation terms to controlled vocabularies.

Normalization of animal models

Similar to human phenotype normalization, we have normalized mouse and rat phenotype terms to EFO and SNOMED-CT. Different treatment procedures have been used to generate animal models that capture specific aspects of human diseases. At times, the incomplete nature of the models could lead to inadequate or misinterpretation of results. Thus, it is necessary to know the experimental procedures used on these animal models. To enhance this interpretation, we have binned all the captured animal model information, during the metadata curation, to a higher level of abstraction, further mapped to EFO, the National Cancer Institute Thesaurus (36), and the BioAssay Ontology (37). In addition, we mapped mouse and rat names to EFO, Jackson Laboratory database identifiers, and Sage Bionetworks Synapse Ontology (http://jax-mice.jax.org/query/f?p=205:1:0 and http://bioportal.bioontology.org/ontologies/SYN). This provides more flexibility during querying of samples from specifically treated animal models. Please refer to Supplementary Files S3 and S4 for mapping of mouse and rat-related terms to controlled vocabularies.

For some of the metadata terms, there were no controlled vocabularies available, e.g. 'Vehicle #1:non-transgenic' or 'BDNF-treated', describes that the mouse is non-transgenic and a vehicle in the former case, while in the second case it is specific gene treatment. Such terms were mapped to either of the phenotype's controlled vocabulary. In case of human stage mentions, specific stages such as Braak II or cognitive scores, such as CERAD, MMSE, etc. could not be mapped to any staging controlled vocabulary as most of the ontologies used higher level of staging, namely Braak. Moreover, in most of the ontologies cognitive tests are not classified under staging, but rather as cognitive tests. This has prompted us to generate a more detailed hierarchical representation of the above-mentioned binning schemes, which will be published separately as ontology, specifically for neurodegenerative gene expression studies. However, for current version, we stick to the already available controlled vocabularies, in addition to our internal classification.

Curation results and discussion

Compliance to standards

Authors tend to provide minimum information as required by the guidelines in archives; publishing major part of the experimental metadata annotations in associated publication. To test, whether the authors adhere to the minimum compliant standards, we performed an assessment of the complaint scores provided by ArrayExpress, the highest score being 5, for Alzheimer's studies. Figure 5 shows the trend in distribution of retrieved AD experiments (see Data Retrieval section) in ArrayExpress, based on the published MIAME and MINSEQE scores (for human, mouse and rat experiments). We observe the trend of submission is concentrated around the score of 4, showing that the submitted data are not fully MIAME or MINSEQE compliant; leading to variable levels of information stored in these archives.

To conclude that not all the submitters' abide 100% by the compliant standards, we investigated if this trend is same for all other disease domains; we chose one among the most studied cancer disease, Lung Cancer, and generated similar results to AD. Supplementary Figure S5 shows the distribution of compliant standards across Lung Cancer studies. From this observation, we show that the loss of information follow the same pattern across all submissions (varying mostly around score of 4). As a result, automated retrieval and meta-analysis is impeded, due to lack of information availability. Details of the experiment IDs investigated for AD and Lung Cancer, along with compliant scores is provided in Supplementary File S1.

Retrieval and prioritization of indication specific studies from GEO and ArrayExpress

Retrieval of experiment IDs using a keyword search (*cf.* Data Retrieval section) also acquires false positive experiments. Any non-disease specific experiment performed by an author named 'Alzheimer' is also retrieved when searching for AD specific experiments. For example, E-MTAB-2584 aims to investigate neuronal gp130 regulation in mechanonociception but was retrieved for AD since one of its author's name is Alzheimer. Moreover, we also obtained experiments for related diseases such as Epilepsy, or Breast Cancer, when querying for AD. For example, GSE6771, and GSE6773 are Epilepsy studies; GSE33500 belongs to Nasu Hakola Disease; all these studies were retrieved when queried for Alzheimer. Incorrect organism specificity was also noticed during prioritization. For example, GSE5281 was retrieved as rat study although it belonged to human. Similarly, GSE2866 was retrieved as mouse study but it belonged to zebra fish. Although incorrectly identified studies are not too high, this still indicates the need to include organism and disease specificity filter during prioritization. Additionally, we manually identified a few experiments that were not retrieved using these keywords, which were also included in the database.

Further on, just by applying these two filter criteria does not assure that all retained experiments were specific to AD. For example, there could be some experiments that aim at a certain pathway that are also relevant in the area of neurodegeneration, but the experiment submitted to the repository does not deal with AD pathology. As a consequence, additional disease relevancy conditions were included before prioritization (*cf.* Experiment Prioritization section). An overview of all the retrieved AD experiments, categorized to one of the priority classes is shown in Figure 6. In addition, a list of priority 1 experiments (for human, mouse and rat) is provided in Supplementary file S5. This figure indicates that nearly 20% of the retrieved studies are in any case not related to AD. On the other hand, to identify the remaining 80% of the experiments (prioritized as 1 and 2) we need massive manual filtering by trained personnel. Only if the archives take an initiative to apply such a structured classification for all uploaded experiments, individual time-cost can be reduced to a greater extent.

Some experiments contain cell lines or other disease samples in addition to Alzheimer's patient samples. Experiment GSE26927 additionally contain samples from patients suffering from Parkinson's disease, multiple sclerosis, etc. To be able to query only AD related samples for integrative

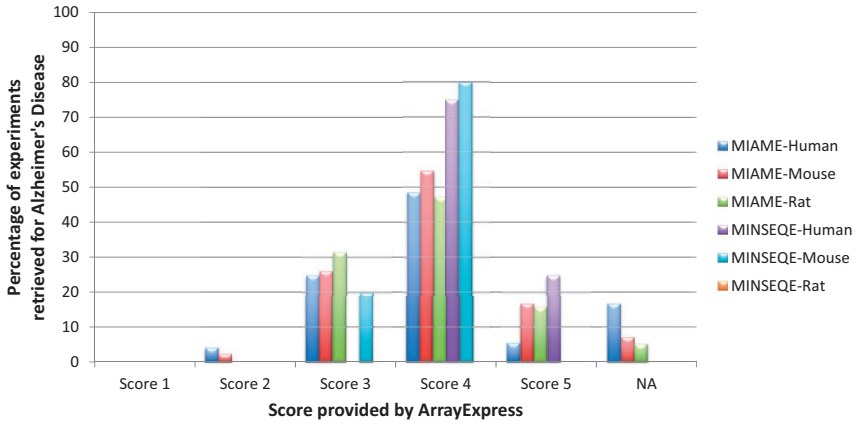


Figure 5. Distribution of MIAME and MINSEQE scores for all automatically retrieved Alzheimer's Disease gene expression experiments in ArrayExpress Database (for human, mouse and rat), as of December 2014. Percentage is calculated as (total number of AD experiments with a certain score)/(total number of AD experiments). 'NA' are the experiments which were not present in ArrayExpress. These scores reflect adherence to compliance standards by the data submitters, needed for re-investigation and reproducibility. It is observed that large percentage of experiments fall under score 4, shows that the required minimum information is still incomplete. The list of experiment IDs along with their associated scores, used for generating this statistics are provided in Supplementary File S1.

analysis, we additionally included priority information at sample level. For example, we tagged Alzheimer's disease samples to AD1 whereas multiple sclerosis samples to MS1. Please refer to the README.txt file for various priority notations used.

Metadata curation

The underlying metadata information for any gene expression study has been underrepresented and thus is largely under-utilized. To perform large-scale analysis, associated annotations are of utmost importance. With the availability of detailed annotation information, one is capable of selecting studies that focus on a particular attribute, such as stage or gender. Each priority class has a specific set of fields for curation; some fields are organism dependent. After prioritization of experiments (*cf.* Experiment Prioritization section), we expect to have ∼100% coverage of essential clinical and relational parameters during manual metadata curation for priority 1 studies. For example, age, gender, phenotype and stage are basic experimental variables for human studies. Additionally, in case of animal models, mouse and rat strain names are important for translational pipelines, as some strains are highly specific models for human NDD while others not (38). Irrespective of the organisms, samples mapped to their corresponding raw file identifiers are vital for running large-scale analysis. However, as shown in Figure 7, this does not hold true for human studies. From Figure 7, it is evident that even after performing thorough curation, we cannot achieve 100% in capturing information for these five basic metadata fields, a fact that is largely due to patient data privacy regulations. Similar is the case with mouse and rat information, see Supplementary Figure S6. Moreover, information related to animal models are much more scare, obstructing automated retrieval. Hence, manual curation accuracy is highly dependent on information availability, as curators cannot harvest information for annotation fields that are not available. On the contrary, the level of detail also depends on the type or aim of the experiment carried out. The authors and database owners obviously need to focus on the qualitative aspect of the experimental information, especially the phenotype of the sample, to allow normalized access for beginners, with standard prose, in order to support a robust computational analysis across all studies in ArrayExpress and GEO.

We selected five of the most common metadata fields (common to any disease domain such as age, gender, phenotype, stage and raw filename) and carried out a trend analysis of information availability versus time. Figure 8 (A) shows the trend over time for the metadata information provided in the archives versus the number of annotation fields that can be harvested after manual curation for human AD priority 1 experiments. Although a bit obscure, we can observe that the level of information submitted to the databases remains almost stable in the last decade (between 2 and 4 metadata fields). Moreover, with manual curation support, we were able to capture the majority of the remaining metadata from associated publications, Figure 8 (B) shows the shift in the mean value of the metadata availability. However, the trend is recently declining since the authors submit relatively lesser level of detailed information than in former times in the associated publications.

The incompleteness of metadata annotations contributed to a substantial increase in curation workload through an increased need for publication reading. This leads to a steep increase of the cost of the trained personal for curation. Overall, for the prioritization and metadata curation of AD gene expression studies, we spent about 1 year of four biocurators effort (working 10 h/week). This does not

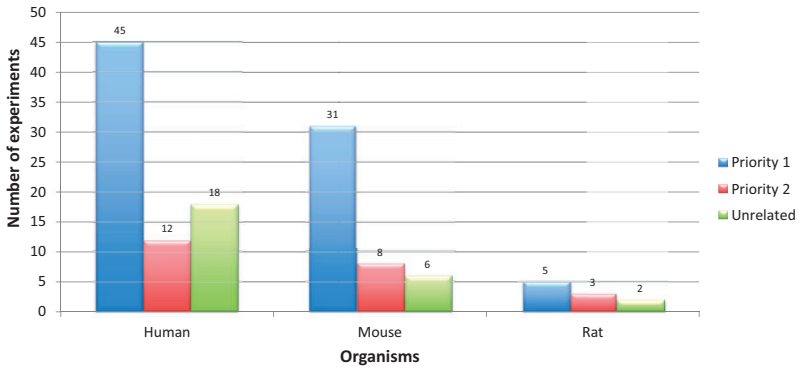


Figure 6. Priority classification statistics for Alzheimer's disease gene expression experiments retrieved from ArrayExpress and GEO (for human, mouse and rat). Alzheimer's disease experiments were retrieved using keywords. Applying the Experiment Prioritization guidelines, they were manually classified to one of the priority classes. Among them, 20% of the experiments were not related to Alzheimer's disease. The digits on the bars represent number of experiments.

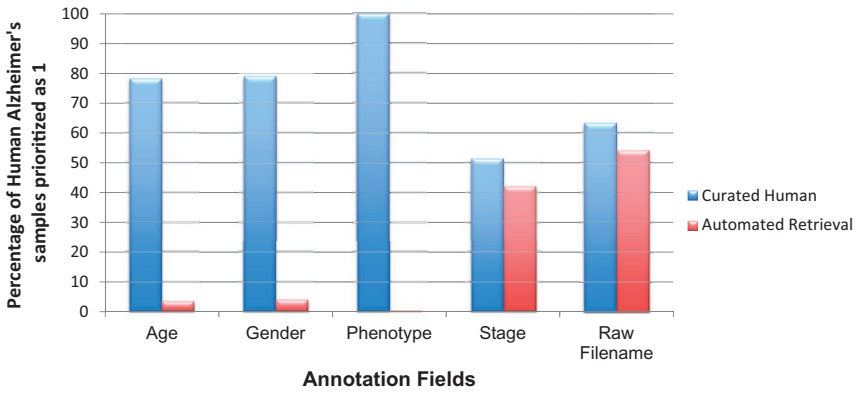


Figure 7. Coverage of basic metadata annotation fields for human AD priority 1 samples with automated retrieval and manual curation. Automated retrieval involved downloading the metadata information from ArrayExpress and GEO, programmatically. For missing meta-annotations, we applied manual curation step to harvest information from the published articles and their associated Supplementary materials. It is clear from the above statistics that manual curation accuracy for basic annotations, such as patient's clinical manifestations, and raw file information, is highly dependent on data availability.

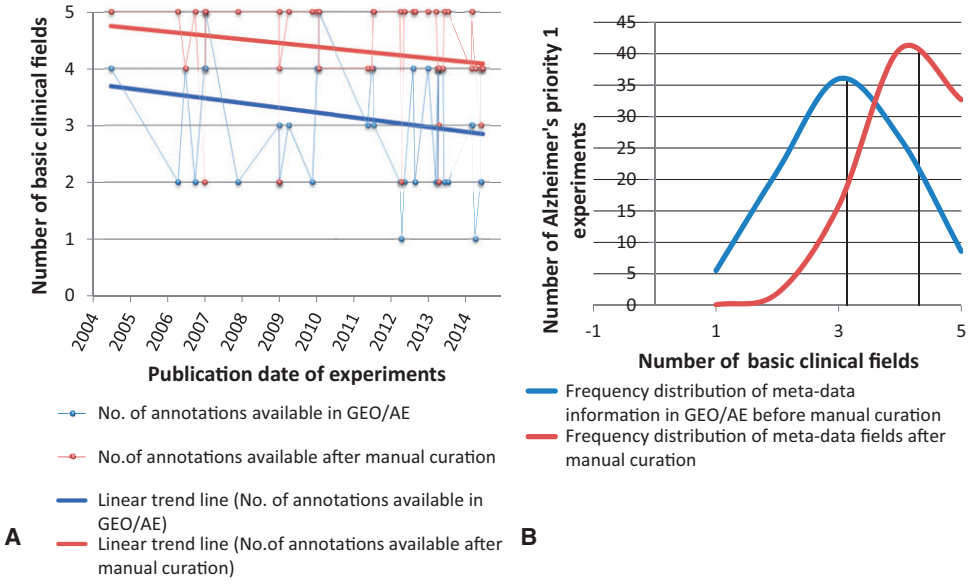


Figure 8. Frequency distribution and Trend Analysis of human priority 1 Alzheimer's disease gene-expression experiments for availability of five basic annotation fields in GEO/ArrayExpress sample page versus manual curation. The five basic annotations considered here are age, gender, stage, phenotype and raw filename. (**A**) Red and blue line represents the linear trend analysis of the availability of meta-annotations for experiments (represented as dots) over years, which has declined. (**B**) The black line represents mean value of the number of annotation fields filled. It is evident from the shift in mean of the distribution analysis that manual curation plays a very important role in capturing the missing metadata information.

include the expert's effort, who constantly provided guidance and monitored the curation work during the same duration.

Accessing *NeuroTransDB*

Metadata annotations for priority 1 AD gene expression studies for human, mouse and rat organisms, from GEO and ArrayExpress, are stored as MySQL tables separately; downloadable as dump files at Fraunhofer SCAI File Transfer Protocol (FTP) website: http://www.scai.fraunhofer.de/NeuroTransDB.html. Please refer to the README.txt for details of how to install and use MySQL dumps. Additionally, these tables are provided as Excel files to allow users to use the curated information in their preferred tools/interface. Currently, the data is in its non-

normalized form. Normalized data, tagged with standard ontologies (*cf.* Normalization of Metadata Annotations section), will be made available through the AETIONOMY Knowledge Base. Currently, we have provided human priority 1 studies normalized using our internal binning scheme. Half yearly updates are planned. Our ultimate goal is to make *NeuroTransDB* a comprehensive resource for researchers working on large-scale meta-analysis in the field of neurodegenerative diseases.

Conclusion and future directions

NeuroTransDB fills the gap for large-scale meta-analysis on publicly available gene-expression studies in the field of neurodegeneration. It joins bits of missing metadata information, scattered in public archives and associated publications, into a consistent, easily accessible and regularly updated data resource. Additionally, in this paper, we have systematically specified key issues encountered during selection of relevant gene expression studies from public archives, along with their associated metadata information. We observed a huge lack of structured metadata in these archives, hampering automated large-scale reusability on a usable level of abstraction. We present here recommendations, as guidelines, for prioritizing relevant studies and a step-by-step protocol for metadata curation. The challenges faced in the course of the development of these guidelines have been pointed out, and the huge manual effort has been made explicit.

The work presented here has listed metadata fields, which have been generated based on disease expert consultation. They are highly important for choosing the right subsets of expression studies to answer complex biological questions underlying a diseased pathology. Some additional fields are included for animal models studies to allow maximal use for translational research. For all the manually curated fields, we describe normalization strategies in an attempt to provide standards for more robust automated querying and interoperability. Our results show the amount of information that is scattered in various resources, requiring extensive manual effort to capture the same. Additionally, we report that even with comprehensive manual harvesting, we were not able to capture 100% of information to fill for the basic annotation fields. We demonstrate convincingly that data availability depends largely on the meticulousness of the submitters. Additionally, it also depends on the aim of the experiment carried out. On an average, considering all the retrieved AD experiments, the submitters provide about 60% of the most basic metadata information. The outlined guidelines could be of significant value to other researchers working on gene-expression studies. The described key issues we faced during such a curation work could influence the data submission and data storage architecture of public repositories.

Subsequently, we plan to extend the curation pipeline to other NDD diseases namely, Huntington's disease. A more gene-expression specific ontology will be built based on the curated annotations for selecting a subset of studies for meta-analyses. Although, microarray studies are the major contributors to the public repositories, RNA-Seq data are rapidly growing. We comprehend that it will be necessary for us to identify all the relevant RNA-Seq studies, since their large storage space has contributed to disperse nature of the available raw data.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank Dieter Scheller, whilst at UCB Pharma, for contributions as a disease expert. We are also grateful to Jonathan van Eyll for his inputs. We acknowledge Nidhi Singh for her contribution as a biocurator during the early stages of the project. The authors thank Dr. Erfan Younesi for suggesting various ontologies used during the normalization process. We are indepted to Jasmin Zohren, Anandhi Iyappan and Jiali Wang for their initial work on curation in gene-expression studies, the outcome of which guided us to develop the more robust workflow described in this manuscript. We additionally thank the researchers, who deposit experimental data for public use. The authors are grateful for the comments and valuable recommendations from two anonymous reviewers.

Funding

German Federal Ministry for Education and Research (BMBF) within the BioPharma initiative 'Neuroallianz', project D10 'In Silico Discovery for putative Biomarkers' (grant number: 1616060B); UCB Pharma GmbH (Monheim, Germany).

Conflict of interest. None declared.

References

1. Johnson,R., Noble,W., Tartaglia,G.G. *et al.* (2012) Neurodegeneration as an RNA disorder. *Prog. Neurobiol.*, **99**, 293–315.
2. Alzheimer's Association. (2014) Alzheimer's disease facts and figures. *Alzheimer's Dement*, **10**, e47–e92.
3. Herrup,K., Carrillo,M.C., Schenk,D. *et al.* (2013) Beyond amyloid: Getting real about nonamyloid targets in Alzheimer's disease. *Alzheimers Dement*, **9**, 452–458.e1.
4. Leidinger,P., Backes,C., Deutscher,S. *et al.* (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, **14**, R78.
5. Greco,I., Day,N., Riddoch-Contreras,J. *et al.* (2012) Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. *J. Transl. Med.*, **10**, 217.
6. Jedynak,B.M., Lang,A., Liu,B. *et al.* (2012) A computational neurodegenerative disease progression score: method and results

with the Alzheimer's disease Neuroimaging Initiative cohort. *Neuroimage*, **63**, 1478–1486.
7. Mayburd,A. and Baranova,A. (2013) Knowledge-based compact disease models identify new molecular players contributing to early-stage Alzheimer's disease. *BMC Syst. Biol.*, **7**, 121.
8. Stokes,M.E., Barmada,M.M., Kamboh,M.I. *et al.* (2014) The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data. *BMC Genomics*, **15**, 282.
9. Iyappan,A., Bagewadi,S., Page,M. *et al.* (2014) NeuroRDF: semantic data integration strategies for modeling neurodegenerative diseases. In: *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM2014)*. Aveiro, Portugal. pp. 11–8.
10. Lappalainen,T. and Sammeth,M. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*.
11. Cheng,W.-C., Tsai,M.-L., Chang,C.-W. *et al.* (2010) Microarray meta-analysis database (M(2)DB): a uniformly preprocessed, quality controlled, and manually curated human clinical microarray database. *BMC Bioinformatics*, **11**, 421.
12. Atz,M., Walsh,D., Cartagena,P. *et al.* (2007) Methodological considerations for gene expression profiling of human brain. *J. Neurosci. Methods*, **163**, 295–309.
13. Monoranu,C.M., Apfelbacher,M., Grünblatt,E. *et al.* (2009) measurement as quality control on human post mortem brain tissue: a study of the BrainNet Europe consortium. *Neuropathol. Appl. Neurobiol.*, **35**, 329–337.
14. American Cancer Society (2014). Cancer Facts and Figures.
15. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
16. Rocca-Serra,P., Brazma,A., Parkinson,H. *et al.* (2003) ArrayExpress: a public database of gene expression data at EBI. *C. R. Biol.*, **326**:1075–1078.
17. Malone,J., Holloway,E., Adamusiak,T. *et al.* (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**:1112–1118.
18. Brazma,A., Hingamp,P., Quackenbush,J. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
19. Bisognin,A., Coppe,A., Ferrari,F. *et al.* (2009) A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bioinformatics*, **10**, 201.
20. Piwowar,H. and Chapman,W. (2010) Recall and bias of retrieving gene expression microarray datasets through PubMed identifiers. *J. Biomed. Discov. Collab.*, **5**, 7–20.
21. Ramasamy,A., Mondry,A., Holmes,C.C. *et al.* (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.*, **5**, 1320–1332.
22. Ober,C., Loisel,D.A. and Gilad,Y. (2008) Sex-specific genetic architecture of human disease. *Nat. Rev. Genet.*, **9**, 911–922.
23. Li,J.Z., Vawter,M.P., Walsh,D.M. *et al.* (2004) Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum. Mol. Genet.*, **13**, 609–616.
24. Zheng,J., Stoyanovich,J., Manduchi,E. *et al.* (2011) AnnotCompute: annotation-based exploration and meta-analysis of genomics experiments. *Database*, **2011**, 1–14.
25. Ivliev,A.E., 't Hoen,P.C., Villerius,M.P. *et al.* (2008) Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. *Nucleic Acids Res.*, **36**, 327–331.
26. Zhu,Y., Davis,S., Stephens,R. *et al.* (2008) GEOmetadb: Powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**, 2798–800.
27. Faith,J.J., Driscoll,M.E., Fusaro, V. *et al.* (2008) Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Res., **36**, 866–870.
28. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.
29. Buckberry,S., Bent,S.J., Bianco-miotto,T. *et al.* (2014) massiR: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics*., **30**, 2084–5. doi:10.1093/bioinformatics/btu161.
30. Coletta,A., Molter,C., Duqué,R. *et al.* (2012) InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. *Genome Biol.*, **13**, R104.
31. Burge,S., Attwood,T.K., Bateman,A. *et al.* (2012) Biocurators and biocuration: surveying the 21st century challenges. *Database (Oxford)*, **2012**, bar059.
32. Glass D, Viñuela A, Davies MN, *et al.* Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology* 2013;**14**:R75. doi:10.1186/gb-2013-14-7-r75
33. Holland D, Desikan RS, Dale AM, *et al.* Rates of decline in Alzheimer disease decrease with age. *PloS one* 2012;**7**:e42325. doi:10.1371/journal.pone.0042325
34. Bernick C, Cummings J, Raman R, *et al.* Age and rate of cognitive decline in Alzheimer disease: implications for clinical trials. *Archives of neurology* 2012;**69**:901–5. doi:10.1001/archneurol.2011.3758
35. Mattsson N, Rosén E, Hansson O, *et al.* Age and diagnostic performance of Alzheimer disease CSF biomarkers. *Neurology* 2012;**78**:468–76. doi:10.1212/WNL.0b013e3182477eed
36. Carter CL, Resnick EM, Mallampalli M, *et al.* Sex and gender differences in Alzheimer's disease: recommendations for future research. *Journal of women's health (2002)* 2012;**21**:1018–23. doi:10.1089/jwh.2012.3789
37. Vest RS, Pike CJ. Gender, sex steroid hormones, and Alzheimer's disease. *Hormones and behavior* 2013;**63**:301–7. doi:10.1016/j.yhbeh.2012.04.006
38. Mirnics K, Pevsner J. Progress in the use of microarray technology to study the neurobiology of disease. *Nature neuroscience* 2004;**7**:434–9. doi:10.1038/nn1230
39. Knöchel C, Oertel-Knöchel V, O'Dwyer L, *et al.* Cognitive and behavioural effects of physical exercise in psychiatric patients. *Progress in neurobiology* 2012;**96**:46–68. doi:10.1016/j.pneurobio.2011.11.007
40. Cummings JL, Schneider E, Tariot PN, *et al.* Behavioral effects of memantine in Alzheimer disease patients receiving donepezil treatment. *Neurology* 2006;**67**:57–63. doi:10.1212/01.wnl.0000223333.42368.f1
41. Nacmias B. Genetics of familial and sporadic Alzheimer's disease. *Frontiers in Bioscience* 2013;**E5**:167. doi:10.2741/E605

42. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica* 1991;**82**:239–59. doi:10.1007/BF00308809
43. Helmer C. Mortality with Dementia: Results from a French Prospective Community-based Cohort. *American Journal of Epidemiology* 2001;**154**:642–8. doi:10.1093/aje/154.7.642
44. Solomon A, Dobranici L, Kåreholt I, *et al*. Comorbidity and the rate of cognitive decline in patients with Alzheimer dementia. *International journal of geriatric psychiatry* 2011;**26**:1244–51. doi:10.1002/gps.2670
45. Hiltunen M, Bertram L, Saunders AJ. Genetic risk factors: their function and comorbidities in Alzheimer's disease. *International journal of Alzheimer's disease* 2011;**2011**:925362. doi:10.4061/2011/925362
46. Sherwood KR, Head MW, Walker R, *et al*. RNA integrity in post mortem human variant Creutzfeldt-Jakob disease (vCJD) and control brain tissue. *Neuropathology and applied neurobiology* 2011;**37**:633–42. doi:10.1111/j.1365-2990.2011.01162.x
47. Durrenberger PF, Fernando S, Kashefi SN, *et al*. Effects of antemortem and postmortem variables on human brain mRNA quality: a BrainNet Europe study. *Journal of neuropathology and experimental neurology* 2010;**69**:70–81. doi:10.1097/NEN.0b013e3181c7e32f
48. Koppelkamm A, Vennemann B, Lutz-Bonengel S, *et al*. RNA integrity in post-mortem samples: influencing parameters and implications on RT-qPCR assays. *International journal of legal medicine* 2011;**125**:573–80. doi:10.1007/s00414-011-0578-1
49. Stan AD, Ghose S, Gao X-M, *et al*. Human postmortem tissue: what quality markers matter? *Brain research* 2006;**1123**:1–11. doi:10.1016/j.brainres.2006.09.025
50. Li JZ, Vawter MP, Walsh DM, *et al*. Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Human molecular genetics* 2004;**13**:609–16. doi:10.1093/hmg/ddh065
51. Long JM, Lahiri DK. MicroRNA-101 downregulates Alzheimer's amyloid-β precursor protein levels in human cell cultures and is differentially expressed. *Biochemical and biophysical research communications* 2011;**404**:889–95. doi:10.1016/j.bbrc.2010.12.053
52. Ly PTT, Wu Y, Zou H, *et al*. Inhibition of GSK3β-mediated BACE1 expression reduces Alzheimer-associated phenotypes. *The Journal of clinical investigation* 2013;**123**:224–35. doi:10.1172/JCI64516
53. Blalock EM, Buechel HM, Popovic J, *et al*. Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. *Journal of chemical neuroanatomy* 2011;**42**:118–26. doi:10.1016/j.jchemneu.2011.06.007
54. Ray M, Zhang W. Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. *BMC systems biology* 2010;**4**:136. doi:10.1186/1752-0509-4-136
55. Grolla AA, Sim JA, Lim D, *et al*. Amyloid-β and Alzheimer's disease type pathology differentially affects the calcium signalling toolkit in astrocytes from different brain regions. *Cell death & disease* 2013;**4**:e623. doi:10.1038/cddis.2013.145
56. Miller JA, Woltjer RL, Goodenbour JM, *et al*. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome medicine* 2013;**5**:48. doi:10.1186/gm452
57. Roed L, Grave G, Lindahl T, *et al*. Prediction of mild cognitive impairment that evolves into Alzheimer's disease dementia within two years using a gene expression signature in blood: a pilot study. *Journal of Alzheimer's disease: JAD* 2013;**35**:611–21. doi:10.3233/JAD-122404
58. Kiko T, Nakagawa K, Tsuduki T, *et al*. MicroRNAs in plasma and cerebrospinal fluid as potential markers for Alzheimer's disease. *Journal of Alzheimer's disease: JAD* 2014;**39**:253–9. doi:10.3233/JAD-130932
59. Koehler NKU, Stransky E, Shing M, *et al*. Altered serum IgG levels to α-synuclein in dementia with Lewy bodies and Alzheimer's disease. *PloS one* 2013;**8**:e64649. doi:10.1371/journal.pone.0064649
60. Bachstetter A, Webster S, Van Eldik L. Traumatic brain injury in a mouse model of Alzheimer's disease leads to persistent glial activation, chronic proinflammatory phenotype, and increased cognitive deficits. *Alzheimer's & Dementia* 2014;**10**:P337. doi:10.1016/j.jalz.2014.05.333
61. Washington PM, Morffy N, Parsadanian M, *et al*. Experimental traumatic brain injury induces rapid aggregation and oligomerization of amyloid-beta in an Alzheimer's disease mouse model. *Journal of neurotrauma* 2014;**31**:125–34. doi:10.1089/neu.2013.3017
62. Szczodry O, Van der Staay FJ, Arndt SS. Modelling Alzheimer-like cognitive deficits in rats using biperiden as putative cognition impairer. *Behavioural brain research* 2014;**274**:307–11. doi:10.1016/j.bbr.2014.08.036
63. Marlatt MW, Potter MC, Bayer TA, *et al*. Prolonged running, not fluoxetine treatment, increases neurogenesis, but does not alter neuropathology, in the 3xTg mouse model of Alzheimer's disease. *Current topics in behavioral neurosciences* 2013;**15**:313–40. doi:10.1007/7854_2012_237
64. Claxton A, Baker LD, Wilkinson CW, *et al*. Sex and ApoE genotype differences in treatment response to two doses of intranasal insulin in adults with mild cognitive impairment or Alzheimer's disease. *Journal of Alzheimer's disease: JAD* 2013;**35**:789–97. doi:10.3233/JAD-122308
65. Ho SW, Tsui YTC, Wong TT, *et al*. Effects of 17-allylamino-17-demethoxygeldanamycin (17-AAG) in transgenic mouse models of frontotemporal lobar degeneration and Alzheimer's disease. *Translational neurodegeneration* 2013;**2**:24. doi:10.1186/2047-9158-2-24
66. Chin J. Selecting a mouse model of Alzheimer's disease. *Methods in molecular biology (Clifton, NJ)* 2011;**670**:169–89. doi:10.1007/978-1-60761-744-0_13
67. James D, Kang S, Park S. Injection of β-amyloid into the hippocampus induces metabolic disturbances and involuntary weight loss which may be early indicators of Alzheimer's disease. *Aging clinical and experimental research* 2014;**26**:93–8. doi:10.1007/s40520-013-0181-z
68. Hassel B, Taubøll E, Shaw R, *et al*. Region-specific changes in gene expression in rat brain after chronic treatment with levetiracetam or phenytoin. *Epilepsia* 2010;**51**:1714–20. doi:10.1111/j.1528-1167.2010.02545.x

5.2.1 Supplementary Figures

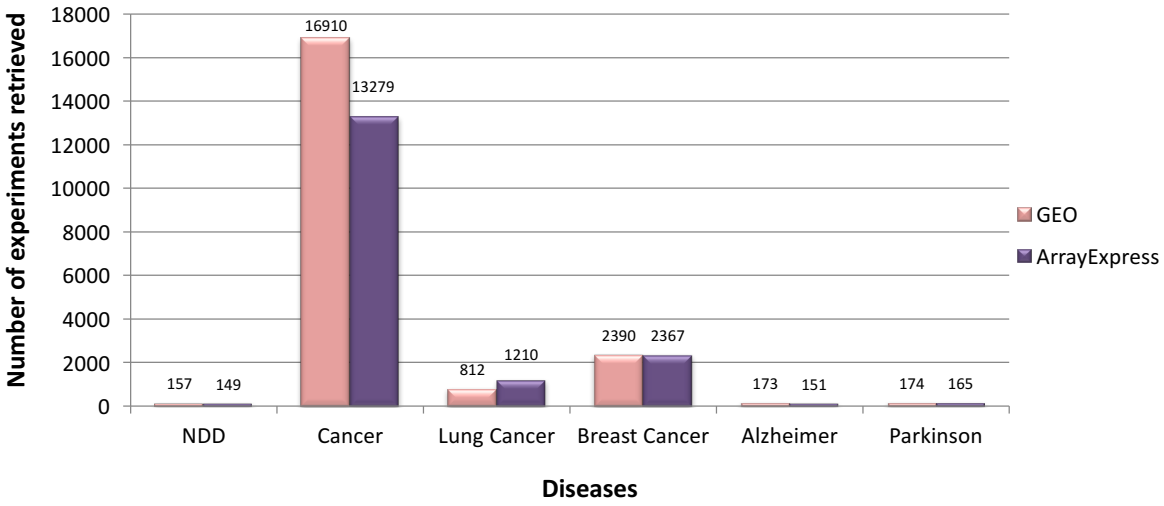


Figure S1: Landscape of transcriptomic data availability for neurodegenerative diseases and cancer in GEO and ArrayExpress. Here, we compare the data availability obtained by a simple keyword search for the two clinically problematic disease domains, neurodegenerative diseases (NDD) and cancer [2,14]; also includes the statistics for two most common diseases in each domain. The numbers on the bars represent quantification of the available experiments, for the selected disease in respective databases (as of 26th July, 2015). The above statistics clearly indicates that NDD field suffers from limited data availability.

title	Expression data from post mortem Alzheimer's disease brains										
summary	To identify molecular pathological alterations in AD brains, we performed interspecies comparative microarray analyses using RNAs prepared from postmortem human brain tissues donated for the Hisayama stu										
overall_design	We prepared RNA samples from the gray matter of frontal and temporal cortices and hippocampi derived from 88 postmortem brains, among which 26 cases were pathologically diagnosed as having AD or an A										
pubmed	http://www.ncbi.nlm.nih.gov/pubmed/23595620										
series_iid	sample_iid	title	type	keyword	domain_specificity	raw_filename	age_patient_or_cell	gender	phenotype	stage	comorbidity
E-GEOD-36980	GSM907792	AD_FC, biological rep1	total RNA	alzheimer	yes	GSM907792.CEL	84	male	AD	Braak 5 ; CERAD 3	Vascular Dementia
	GSM907793	AD_FC, biological rep2	total RNA	alzheimer	yes	GSM907793.CEL	105	female	AD	Braak 5 ; CERAD 3	
	GSM907794	AD_FC, biological rep3	total RNA	alzheimer	yes	GSM907794.CEL	88	female	AD	Braak 5 ; CERAD 3	Diabetes Mellitus
	GSM907795	AD_FC, biological rep4	total RNA	alzheimer	yes	GSM907795.CEL	88	male	AD like change	Braak 5 ; CERAD 3	prediabetes
	GSM907796	AD_FC, biological rep5	total RNA	alzheimer	yes	GSM907796.CEL	91	female	AD	Braak 6 ; CERAD 3	
	GSM907797	AD_FC, biological rep6	total RNA	alzheimer	yes	GSM907797.CEL	95	female	AD	Braak 6 ; CERAD 3	Dementia with Lewy bodies
	GSM907798	AD_FC, biological rep7	total RNA	alzheimer	yes	GSM907798.CEL	92	female	AD	Braak 5 ; CERAD 3	prediabetes
	GSM907799	AD_FC, biological rep8	total RNA	alzheimer	yes	GSM907799.CEL	95	female	AD	Braak 5 ; CERAD 3	
	GSM907800	AD_FC, biological rep9	total RNA	alzheimer	yes	GSM907800.CEL	101	female	AD	Braak 6 ; CERAD 3	
	GSM907801	AD_FC, biological rep10	total RNA	alzheimer	yes	GSM907801.CEL	94	male	AD	Braak 6 ; CERAD 3	
	GSM907802	AD_FC, biological rep11	total RNA	alzheimer	yes	GSM907802.CEL	89	male	AD	Braak 6 ; CERAD 3	
	GSM907803	AD_FC, biological rep12	total RNA	alzheimer	yes	GSM907803.CEL	100	female	AD	Braak 6 ; CERAD 3	Vascular Dementia
	GSM907804	AD_FC, biological rep13	total RNA	alzheimer	yes	GSM907804.CEL	99	male	AD like change	Braak 5 ; CERAD 3	
	GSM907805	AD_FC, biological rep14	total RNA	alzheimer	yes	GSM907805.CEL	83	male	AD like change	Braak 6 ; CERAD 3	
	GSM907806	AD_FC, biological rep15	total RNA	alzheimer	yes	GSM907806.CEL	90	male	AD	Braak 5 ; CERAD 3	

(A)

(B)

Scope: Self Format: HTML Amount:

Sample GSM907797

Status	Public on Apr 17, 2013
Title	AD_FC, biological rep6
Sample type	RNA
Source name	Frontal cortex of AD brain
Organism	Homo sapiens
Characteristics	tissue: Frontal cortex age: 95 Sex: F

(C)

Supplementary Table S2. Clinical and pathological data and RNA quality for subjects examined in the

No	Group	Sex	Age	PMI (h)	Brain weight (g)	CERAD score	Braak stage	Pathological diagnosis	DM or prediabetes
1	AD	M	84	17	1300	3	5	AD+VD	
2	AD	F	105	17	1150	3	5	AD	
3	AD	F	88		1200	3	5	AD	DM
4	AD	F	95	20	1150	3	6	AD like change	
5	AD	M	88	24	1120	3	5	AD like change	prediabetes
6	AD	F	91	3	960	3	6	AD	
7	AD	F	95	7	1050	3	6	AD+DLB	
8	AD	F	92	5	1020	3	5	AD	prediabetes
9	AD	F	95	12	1080	3	5	AD	
10	AD	F	101	6	1020	3	6	AD	
11	AD	M	94	14	1230	3	6	AD	
12	AD	M	89	14	1300	3	6	AD	
13	AD	F	100	16	1050	3	6	AD+VD	
14	AD	M	99	12	1150	3	5	AD like change	
15	AD	M	83	6	1330	3	6	AD like change	
16	AD	M	90	4	1150	3	5	AD	

Figure S2: Screenshot of the sample information provided by authors in GEO and associated publication for GSE36980. (A) Is an example of the excel sheet used by our curators for meta-data curation
(B) Shows the sample information (GSM907797) provided by the authors in GEO database. (C) Is a screenshot of the Supplementary Table 2 where additional information about the patients is provided. From (B) we know that the AD patient is a female and 95 years old, but in (C) there are two rows (marked in red boxes) that match this information. Thus, we could not map additional information such as stage, diagnosis, etc. to the respective samples.

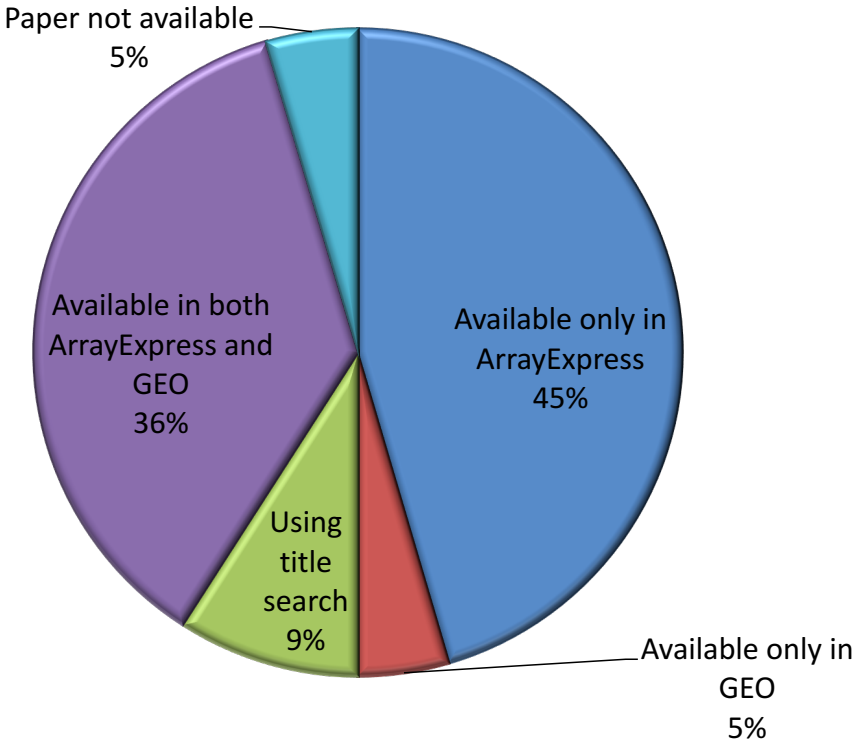


Figure S3: Search strategies applied to retrieve published articles linked to priority 1 Alzheimer's Disease gene expression studies (human, mouse and rat) for meta-data curation.

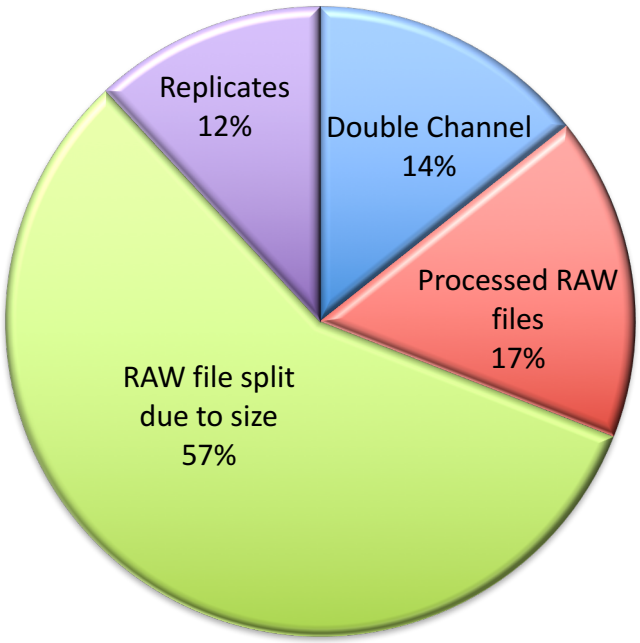


Figure S4: Probable reasons behind duplicated sample information in ArrayExpress for priority 1 Alzheimer's Disease gene expression studies (human, mouse, and rat).

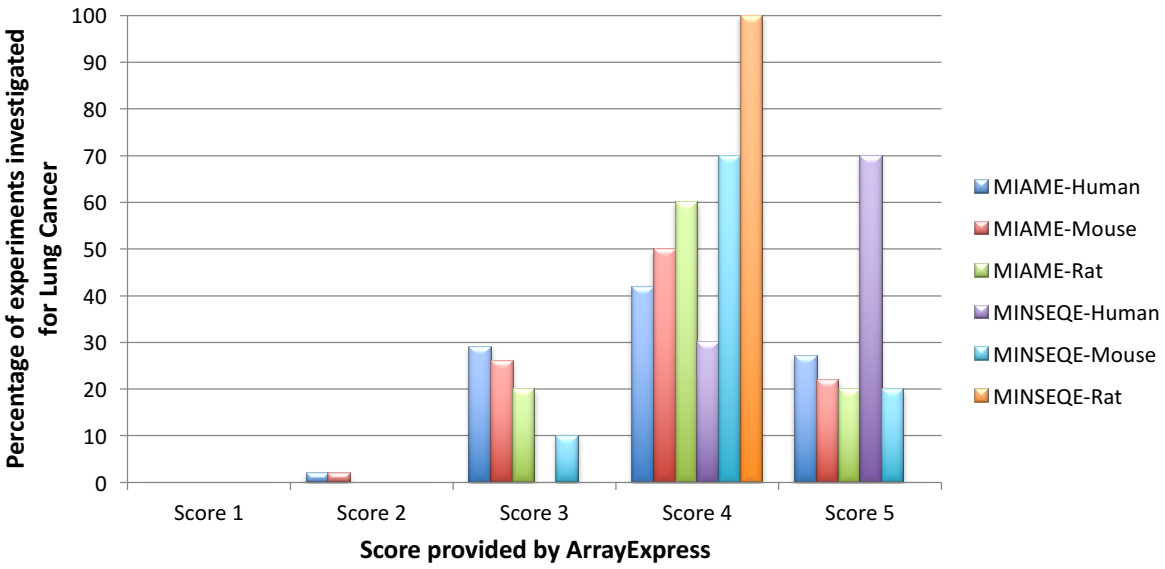


Figure S5: Distribution of MIAME and MINSEQE scores for randomly queried Lung Cancer gene expression experiments in ArrayExpress Database (for human, mouse, and rat), as of May 8, 2015. Percentage is calculated as (total number of Lung Cancer experiments with a certain score)/(Total selected Lung Cancer experiments). To align with the quantity of AD experiments, we investigated (randomly) 100 microarray and 10 sequencing experiments in ArrayExpress. The Lung Cancer domain resembles the trend pattern of Alzheimer experiments in adhering to compliant standards, concentrated mostly around score of 4. For rat, there was just one sequencing experiment. The list of experiment IDs along with their associated scores, used to generate this statistics, are provided in supplementary file S1.

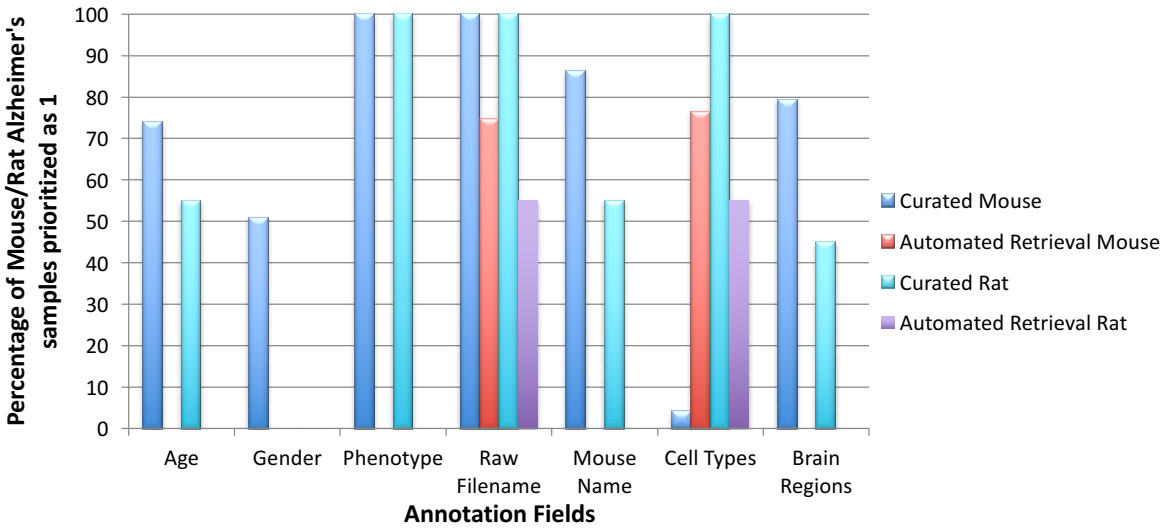


Figure S6: Coverage of basic meta-data annotation fields for mouse and rat AD priority 1 samples with automated retrieval and manual curation. It is clear from the above statistics that manual curation accuracy for basic annotations, such as age, gender, and raw file information, is highly dependent on data availability. The cell type information for mouse, which was automatically downloaded, contained both brain regions, and cell type information. Thus, after manual curation we observe reduced information in mouse cell type.

5.2.2 Supplementary Tables

Experiment ID	Contact Reason	Response of the authors
GSE47036	Mismatch in sample labelling between GEO database and the associated publication	Provided us the correct mapping file, in addition authors corrected the error in GEO database
GSE47038	Difference in number of samples between GEO (58 samples) and ArrayExpress (64 samples)	ArrayExpress synced the sample count to GEO database
GSE29652	Difficulty in mapping the sample information from GEO to the detailed information provided in publication (Table 1)	The authors provided us additional file mapped to Source name (*cf.* GEO sample page) along with detailed information of how to map these different IDs to the case information in Table 1
GSE36980	Could not map the patient information (Subject No.) provided in Supplementary Table 2 to the GEO sample IDs	Authors provided additional supplementary file which contained GEO sample IDs mapped to subject numbers in Supplementary Table S2
GSE26927	Unable to map the Cohort characteristics provide in Table 1 (or Supplementary Online Resource 1) of the paper with sample information in GEO. In addition, the number of samples used in the analysis (113), as described in the paper, differs from the number of samples in GEO (118).	Author provided the mapping information of sample title with cohort characteristics. Additionally, the authors reported that only 113 samples out of 118 were of good quality, thus the difference in sample numbers.

Table S1: **Detailed listing of GEO experiment IDs where we contacted the authors for missing or mismatched information.** Along with the experiment IDs we provide the reason for contacting the authors and their response.

Sample ID from GEO	Sample title from GEO (resemble patient ID)	Associated sample age provided in GEO website	Patient ID in supplementary file	Associated gender information in supplementary file
GSM466881	PD rep1	Male	PD8	M
GSM466882	PD rep2	Female	PD9	M
GSM466883	PD rep3	Male	PD11	M
GSM466884	PD rep4	Male	PD13	M
GSM466885	PD rep5	Male	PD14	M
GSM466886	PD rep6	Male	PD15	M
GSM466887	PD rep7	Male	PD15x	M
GSM466888	PD rep8	Female	PD17	F
GSM466889	PD rep9	Male	PD18	M
GSM466890	PD rep11	Female	PD20	F
GSM466891	PD rep12	Male	PD21	M
GSM466892	PD rep13	Male	PD22	M
GSM466893	PD rep14	Male	PD22x	M
GSM466894	PD rep15	Male	PD23	M
GSM466895	PD rep16	Male	PD23x	M
GSM466896	PD rep17	Female	PD27	M
GSM466897	PD rep18	Male	PD30	M
GSM466898	Ctrl rep_2	Female	CNT4	F
GSM466899	Ctrl rep_3	Female	CNT5	F
GSM466900	Ctrl rep_4	Male	CNT6	F
GSM466901	Ctrl rep_5	Male	CNT15	F
GSM466902	Ctrl rep_6	Male	CNT16	F
GSM466903	Ctrl rep_7	Female	CNT18	F
GSM466904	Ctrl rep_8	Female	CNT19	M
GSM466905	Ctrl rep_9	Male	CNT20	M
GSM466906	Ctrl rep_10	Male	CNT21	F
GSM466907	Ctrl rep_11	Female	CNT24	M
GSM466908	Ctrl rep_12	Male	CNT 25	F

Table S2: **Detailed mapping of GEO's Sample ID and Patient ID (from supplementary Table S6 of the associated publication) along with their associated gender information of experiment GSE18838.** Here we mapped the Sample and Patient IDs with an assumption that they are in natural sorted order. Samples for which we could not map the meta-data information between GEO archive and supplementary file (Table S6) of the associated published article are highlighted in green. The difference in male/female ratio between GEO samples (male/female: 19/9) and Supplementary Table S6 (male/female: 18/8) provides enough evidence that the information is misleading.

5.3 Summary

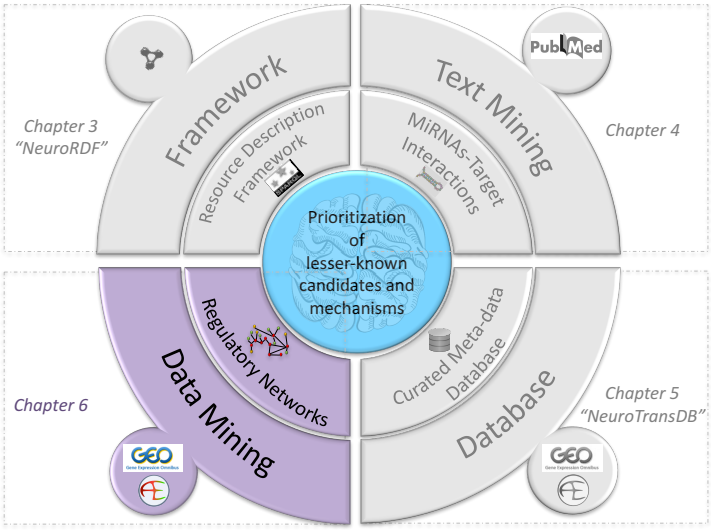
The database introduced here, *NeuroTransDB*, aims to assist researchers to re-use public omics data efficiently in NDD research. The presented work here is a good example to demonstrate that FAIRification of public omics data requires more than just reformatting and mapping to standard ontologies. Firstly, it reports on the key issues faced during retrieval of the subset of studies from public repositories. The keyword-based search possibilities provided by these repositories guarantees higher recall but misses out on the clinical relevancy to the queried indication and organism. Next, it draws attention to the need for comprehensive and precise harvesting of metadata annotations. Most of the data submitters do not comply to standards such as MIAME or MINSEQE during submission. This has resulted in missing or erroneous metadata annotations. Even when provided, the inconsistent and incomplete annotations are scattered as unstructured prose in the originating publications or associated supplementary files. Moreover, due to the varied nature of these files, automated extraction of annotations from these sources is not straightforward. Hence, large effort from dedicated individuals is required to harvest this valuable information. Finally, adding to the obstacles is the lack of uniform representation and mapping to standard vocabularies or ontologies.

In this work, we carry out FAIR transformation of the public omics studies to increase the added value for translational research in NDD; by navigating around the above-mentioned obstacles. To increase *Findability* and *Accessibilit*y of the data, relevant studies were prioritised based on a simple binning approach: *in vivo and in vitro* systems. A set of classification rules that account for sample source, organism and disease specificity were stated for prioritisation. Consulting the disease experts, a list of metadata fields that are needed for translational research (both human and animal models) was laid out. These metadata fields are very specific to NDD research and are scalable to a usable level of abstraction based on the research question at hand. Using a semi-automated curation workflow, missing and incorrect metadata fields were manually curated by gathering information from several sources: (i) directly from databases; (ii) subsections, tables, figures or supplementary files of originating publications; (iii) chained references for original source of information; (iv) contacting the data submitters or corresponding authors. Moreover, the trend analysis showed a drop in furnishing basic metadata information such age, and gender over the last decade. Harvested metadata fields were normalised using

public vocabularies and ontologies. To enable richer integration and *Interoperability*, these metadata fields were annotated with the URIs provided by the ontologies. Thus, qualifying for globally accepted standard for data exchange and knowledge representation on web, RDF. Finally, *Reusability* is greatly enhanced by providing data in a machine-readable format with provenance on original data source, context, and information on curated metadata. In addition, since *NeuroTransDB* is based on publicly-funded GEO and ArrayExpress databases, its data is free for public use without any restrictions.

To the best of our knowledge, this is the first database that systematically harvests metadata annotations specific to NDD research along the FAIR principles. The work presented here can act as a general guideline for prioritising and harvesting metadata information for any biomedical data. The transformed FAIR data increases the value of the public data in integrative data analysis, reported in Chapter 3; enabling researchers to ask questions that was previously not easily viable. The prioritised studies, with correct metadata annotations, enable researchers to precisely select samples for integrative meta-analysis; thereby increasing the accuracy of derived statistics. Chapter 6 proposes a new approach using gene regulatory networks to perform meta-analysis and identify robust expression patterns across heterogeneous datasets.

Chapter 6 Knowledge Instructed Gene Regulatory Networks



6.1 Introduction

Gene regulatory networks (GRNs) have attracted a lot of attention to model dependencies between the molecular entities from gene expression data. They serve as a blueprint of regulatory relations governing biological events, which can be used to derive biological hypothesis; to guide new experimental designs; as network based biomarkers; differential analysis. A variety of algorithms have been proposed to model GRNs that enhance our understanding of diseases: similarity measures (mutual information), regression-based, Bayesian models, or ensemble approaches. When considering biological consistency, the differences between these methods are not large. However, due to improved stability and accuracy, ensemble-based methods have gained popularity. Moreover, global gene expression analysis outperforms results derived from single experiment to classify subgroups or identify common patterns.

Meta-analysis approaches result in more robust and reliable gene signatures across heterogeneous datasets by enhancing the statistical power; overcoming the variability of individual experiments. Modelling GRNs using meta-analysis approach identifies molecular mechanisms and key drivers in an unbiased way. A more conventional approach is to merge results that are gene-centric (DE genes) or through functional enrichment of dysregulated genes. Despite the promising potential, these approaches tend to converge

towards ‘what is already known’. To improve the reliability without increasing the computational costs, one can exploit existing sources of knowledge to discard or enforce edges that are already known. But not all prior knowledge sources are reliable and guarantee a high level of confidence, which may lead to noisy results. Thus, there is a need for a more robust approach utilising prior knowledge such that the networks scale well, reduce computational effort, and do not converge to known players.

In this publication, we have developed an approach that iteratively self-instructs the generation of GRNs to unravel general principles of AD patho-mechanisms, firstly using literature knowledge and subsequently enriching with data-driven functional analysis. By integrating heterogeneous AD datasets, this approach has the capability to identify non-obvious subtle changes in expression level playing a central role in dysregulated events.

6.2 Publication

POST-PRINT VERSION

Analytical strategy to prioritize Alzheimer's disease candidate genes in gene regulatory networks using public expression data

Shweta Bagewadi Kawalia[1, 2, ¶], Tamara Raschka[1, 3, ¶], Mufassra Naz[1, 2], Ricardo de Matos Simoes[4], Philipp Senger[1] and Martin Hofmann-Apitius[1,2]

[1] Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany

[2] Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for Information Technology, Dahlmannstrasse 2, 53113 Bonn, Germany.

[3] University of Applied Sciences Koblenz, RheinAhrCampus, Joseph-Rovan-Allee 2, 53424 Remagen, Germany. [4] Dana-Farber Cancer Institute, 9450 Brookline Ave, 02215 Boston, United States.

[¶] Equal contributors

Correspondence:

Prof. Dr. Martin Hofmann-Apitius

Head of the Department of Bioinformatics

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

53754 Sankt Augustin, Germany

Email: martin.hofmann-apitius@scai.fraunhofer.de

Tel: +49-2241-14-2802

Fax: +49-2241-14-2656

Abstract

Background: Alzheimer's disease (AD) progressively destroys cognitive abilities in the aging population with tremendous effects on memory. Despite recent progress in understanding the underlying mechanisms, high drug attrition rates have put a question mark behind our knowledge about its etiology. Re-evaluation of past studies could help us to elucidate molecular-level details of this disease. Several methods to infer such networks exist, but most of them do not elaborate on context specificity and completeness of the generated networks, missing out on lesser-known candidates.

Method: In this study, we present a novel strategy that corroborates common mechanistic patterns across large scale AD gene expression studies and further prioritizes potential biomarker candidates. To infer gene regulatory networks (GRNs), we applied an optimized version of the BC3Net algorithm, named BC3Net10, capable of deriving robust and coherent patterns. In principle, this approach initially leverages the power of literature knowledge to extract AD specific genes for generating viable networks.

Results: Our findings suggest that AD GRNs show significant enrichment for key signaling mechanisms involved in neurotransmission. Among the prioritized genes, well-known AD genes were prominent in synaptic transmission, implicated in cognitive deficits. Moreover, less intensive studied AD candidates (STX2, HLA-F, HLA-C, RAB11FIP4, ARAP3, AP2A2, ATP2B4, ITPR2, and ATP2A3) are also involved in neurotransmission, providing new insights into the underlying mechanism.

Conclusion: To our knowledge, this is the first study to generate knowledge-instructed GRNs that demonstrates an effective way of combining literature-based knowledge and

data-driven analysis to identify lesser known candidates embedded in stable and robust functional patterns across disparate datasets.

Keywords: Alzheimer Disease; Gene Regulatory Networks; Microarray Analysis; Synaptic Transmission

Background

Alzheimer's disease (AD) is a very complex idiopathic disease contributing to immense personal and societal burden, with ~13.8 million people being affected by 2050 [1] in the US alone. High failure rate of AD drugs (98%) in Phase 3 trials have resulted in no new FDA approved drugs since 2003 [2]. Moreover, the five previously approved AD drugs just provide symptomatic relief [3]. Not all, but a substantial proportion of these studies focused on amyloid-beta (Aβ) and tau accumulations as being synonymous to the AD pathology [4], leading to an unprecedented wealth of molecular and clinical data. Despite the disappointing outcome of the clinical trials, neurology researchers still believe in the definiteness of these two hypotheses [5]. This reaffirms that pathological mechanisms underlying AD are much more complex than the current consideration, thus, opening up possibilities for new therapeutic targets. Working towards unraveling dysregulated events heralding known and unknown patterns could fill the gaps between AD hallmarks [6].

Existing experimental data, not being fully exploited, contain compelling evidence that have the potential to contribute next groundbreaking discoveries. The great challenge, however, lies in harmoniously integrating these data and interpreting them differently to derive new-novel insights while maintaining the biological connections. The term, "Horizontal Meta-analysis" implies the integration of results from several independent

studies [7], thereby increasing the statistical power of the derived conclusion. A more conventional gene-centric approach is to intercross differentially expressed (DE) genes across studies based on majority voting [8], merging gene ranks [9], and combining p-values [8]. However, differing factors can lead to a low overlap and discrepancies between studies such as the applied statistical methods, different platforms of the quantitative measurements, and heterogeneity of the patient cohorts [10]. Moreover, these approaches do not shed light on the co-ordinated genes that collectively orchestrates the underlying (patho-)mechanism. A more consistent and robust approach is through functional enrichment of the dysregulated genes using KEGG [11], MSigDB [12], and other sources of pathway knowledge. However, these approaches have a tendency to converge towards genes that express in large magnitudes and generated hypotheses are restricted by current understanding of pathways.

Network-based approaches that rely on the coherence of expression changes between functionally dependent genes, could provide an effective means to overcome the above-mentioned challenges. Such inferred networks have the capability to determine subtle expression shifts between correlated gene pairs that are linked to the dysregulation events. Particularly, these signatures are largely consistent across different studies; thus, emphasizing on its benefits for large scale meta-analysis. In the last few years, we have seen a swarm of methods that infer such networks based on co-expression, regulation, and causal information namely WGCNA [13], BC3Net [14], MRNET [15], ARACNE [16], GENIE3 [17], and CLR [18]. Among these, WGCNA and the bagging version of C3Net (BC3Net) are popular and computationally efficient methods. BC3Net is an ensemble method that statistically infers GRNs based on the strongest mutual dependencies between genes. Whereas, WGCNA clusters genes on the basis of calculated pairwise correlation

coefficients. In the meanwhile, BC3Net has been reported in providing meaningful biological insights for large-scale studies [19,20].

Traditional GRNs identify patterns through differential co-expression analysis [21–23], displaying grouping of patterns based on dysregulated and co-ordinated biomolecular changes. Integration of such priors drastically improves the context specificity of the inferred networks relative to using data as the sole source [24–26]. However, these spurious discriminative structures, in a given disease context, may vary since DE genes are highly inconsistent across studies [27]. Biologically speaking, one may argue that the differences in functionally enriched components, derived from DE genes are more consistent than gene-centric activities [28]. But this approach misses out on less informative and less studied non-DE genes, which act in groups, contributing to the observed phenotype or a part of cascade effect. Furthermore, overlaying the inferred networks with known interactions, cataloged in databases or harvested from published literature expand the knowledge space [29,30]. However, an intriguing question on completeness, veracity and context specificity of these interactions has proven to be a major setback [31].

Here we propose a new approach to identify common signature patterns across public AD studies and prioritize lesser known AD candidates that unravel the general principles of the intrinsic patho-mechanisms. To identify AD mechanistic footprints, we established an optimized workflow around BC3Net to extract more robust and coherent co-expressed gene patterns (named BC3Net10). The approach allows us to converge lesser known candidates into the final generated GRNs. Moreover, to generate context-specific GRNs, the main rationale applied was to leverage the power of prior knowledge and functionally enriched candidates in the data. Firstly, we identified most frequently discussed genes in the scientific literature using our literature mining environment SCAIView; this is called the

"seed". We are aware that the generated GRNs may be biased due to the incomplete nature of the prior knowledge. To overcome this limitation, we extended the seed by adding all the genes from the enriched pathways, determined for the high scoring inferred interactions from BC3Net10. Several iterations are performed until there are no more genes to be added to the seed. Finally, an aggregated GRN from all the iterations, for each dataset, is generated to prioritize functional context and determine lesser known candidates from the genetic variant analysis. Figure 1 presents an overview of the strategy in this study, and descriptions of the methodology are available in the Material and Methods section. This work suggests a context-specific strategy for future interpretation of the GRNs. Taken together, our work demonstrates that optimizing the GRN generation can provide a powerful resource to prioritize novel candidate genes (could serve as biomarkers) and common functional components that axles the disease progression.

Material and Methods

Selection of datasets

We collated eight Alzheimer's disease datasets (cf. Table 1) that are composed of 50 or more samples (for diseased and control phenotype) from the previously developed value-added database, *NeuroTransDB* [32]. Briefly, this database contains manually curated metadata annotations for eligible neurodegenerative studies. The datasets have been harvested from publicly available resources namely, Gene Expression Omnibus (GEO) [33] and ArrayExpress [34], using a keyword-based search approach.

Furthermore, datasets that fulfilled the following criteria were retained for generating gene regulatory networks: (i) oligonucleotide arrays for analysis consistency, (ii) availability of raw data to facilitate uniform pre-processing and (iii) expression profiling carried out on

brain tissue. A list of four potential datasets that comply with the above conditions is eligible for further analysis: GSE5281, GSE44771, GSE44770, and, GSE44768. An overview of the platform, stage, and brain region information for the same is given in Table 1. Among these, GSE44771, GSE44770, GSE44768 were from a single study reported by Zhang et al. [35] for late-onset AD.

Pre-processing and gene annotation

The four selected datasets were processed identically to reduce variance and to maintain consistent quality. All analysis was carried out with R (Version 3.1.3) [36], an open-source statistical language, using the packages from Bioconductor (Version 3.0) [37]. The overall step-by-step workflow is shown in Figure 1. To eliminate the variance effect of non-specific hybridization, all the downloaded raw data were uniformly normalized by performing background correction, quantile normalization, and averaging the expression values of duplicate probes on log2-transformed intensity values. For Affymetrix platform, robust multi-array average method (*rma*) [38] available in Bioconductor package *affy* was applied. Similar methods available in Bioconductor package *limma* [39] were applied on Rosetta/Merck Human 44k 1.1 microarray chip.

Affymetrix probes to gene symbols annotation mapping were obtained from the "hgu133plus2.db" Bioconductor package. In the case of Rosetta/Merck chip, the gene symbol annotations were provided directly along with the intensity values. For multiple probes mapping to the same gene within an array, average expression values were used. Unmapped probes were excluded from further analyses. As a result of this preprocessing step, we retained 20155 in GSE5281, 11254 in GSE44771, 10437 in GSE44770, and 12000 in GSE44768 genes for further analysis.

Quality control and outlier detection

Using the Bioconductor package *arrayQualityMetrics* [40], we assessed the array quality and removed the outlier samples. Describing shortly, *arrayQualityMetrics* determine outliers using three different metrics: (i) distance between samples using principal component analysis (ii) array intensity distributions of all samples on the array (iii) individual array quality through MA-plots. If a sample is detected as an outlier in either of the three metrics, we discard it from further analysis. In the four selected datasets, 9 in GSE5281, 19 in GSE44771, 27 in GSE44770, and 12 in GSE44768 arrays were outliers. The list of identified outlier arrays is provided in Supplementary File S3. The remaining arrays that passed the quality control were processed as described earlier.

Leveraging stable gene regulatory networks

In order to derive AD relevant GRNs, we divided the AD gene expression profile based on their phenotypes, disease and normal. Subsequently, BC3Net10 algorithm was applied only on diseased samples for AD seed genes, cf. Figure 1. GRNs were generated independently for each dataset, visualized as igraph objects in Cytoscape tool [41]. Network topological properties such as node degree, hub genes, etc. were determined using the Bioconductor package igraph.

Filter pre-processed data for seed gene list

Prior to GRN generation, each pre-processed dataset was restricted to the genes in the seed. Initially, it consists of a set of literature-derived genes that have high probability of direct or indirect involvement in AD pathogenesis (see Section Gathering initial seed genes). The rationale behind applying this filtration is to maintain the disease specificity and reduce

high run time due to bootstrapping in BC3NET. Furtheron, after every functional enrichment iteration we again restrict the expression data to the new seed.

Gathering initial seed genes

The backbone of the seed comprises of the results harnessed from our text-mining knowledge framework, SCAIView [42]. SCAIView is a knowledge discovery framework that supports named entity recognition, information retrieval, and information extraction on large textual sources. Its capability to rank documents and biomedical entities based on the relevancy score allows retrieval of significant players in a disease context [43,44]. Querying SCAIView for AD related genes resulted in 4808 genes, as of 2nd January 2016. Only the top 500 retrieved genes were used as the initial seed, depicted as i=1 in Figure 1.

Optimized GRN construction

For the construction of GRNs, R package *bc3net* was applied to the processed data with 100 bootstraps (B=100). Briefly explained, one aggregated network was generated by applying the C3Net algorithm on 100 bootstrapped data, which were inferred from given processed dataset. Statistically, non-significant edges inferred by C3Net and BC3Net were discarded using Bonferroni's multiple testing correction, $\alpha = 0.05$. In the resulting aggregated network, edge weights represent the frequency of a correlated gene pair in 100 random sampling, ranging from 0 to 1.

During random sampling, true and most prominent correlations are stochastically more likely to be selected than the non-correlated ones. This is reflected in BC3Net networks, where three independently generated GRNs, inferred from the same gene expression dataset (GSE5281), have an edge overlap of ~74% (for no edge weight cutoff) and ~89% (for edge weight>=0.5); the node overlap always remained 100%. The BC3Net parameters

used for performing this analysis are: boot=100, estimator= "pearson", disc= "equalwidth", mtc1=TRUE, alpha1=0.05, adj1= "bonferroni", mtc2=TRUE, alpha2=0.05, adj2= "bonferroni", weighted=TRUE, igraph=TRUE, verbose=FALSE and number of seed genes=4808 (see Section Gathering initial seed genes). However, less frequently appearing, yet plausible, edge interactions could offer the potential for promising candidates that are buried in expression data.

We observed that the intersection between independently generated GRNs saturated after 5-10 repetitions of the BC3Net algorithm on the same dataset. Thus, in order to expand the knowledge space around AD candidates and for completeness, we propose an optimization of the randomness to devise a more recall optimized GRNs. More specifically, we applied the BC3Net algorithm to the same dataset 10 times, named BC3Net10. Finally, we aggregated the 10 independently generated GRNs into one. The final edge weight is now the mean of the computed edge score from 10 GRNs. This increases the prospect of deducing more reasonable functional speculations in complex diseases with the high probability of novelty for further investigations.

Subnetwork selection and functional enrichment analysis

The choice of a threshold can significantly affect the integrity of the network and the co-expression modules derived from it. In this regard, computed edge weight (mean weight>0.5) from BC3Net10 was used as the filter criteria for selecting significant gene pairs in the generated GRNs. This increases the significance level by 50% for the inferred interactions in each dataset.

The overlap between the inferred interactions/edges was very low (zero genes common to all 4 subnetworks, see Figure 2) when BC3Net was applied on the initial seed. Several

reasons can be presumed for lack of common and stable genes such as different platforms, distinct brain tissues, diverse patient cohort, and treatment heterogeneity. However, numerous studies have already shown that the functional signatures are more stable relative to individual gene level information [45–48]. In this context, to extract the most representative biological pathways for genes in the subnetworks (separately for each dataset), we performed functional enrichment analysis (based on one-sided Fisher's exact test) for KEGG pathway information using ConsensusPathDB (CPDB) [49] (Release 30). Using the Bioconductor package, *org.Hs.eg.db* [50] we mapped the gene symbols to Entrez gene identifiers obtained from CPDB.

Identification of enriched candidates and seed gene list enrichment

We devised a strategy to expand the seed through functional enrichment analysis of the individual network modules inferred by the GRNs, enabling us to quantify the saturation of the inferred network. We extracted significant pathways (for the p-value<0.05) common between the determined subnetworks of the four datasets, generated using the initial seed. We added a new gene (called enriched candidate) to the seed when the gene belongs to the respective CPDB and KEGG pathway gene set that is significantly enriched across all 4 inferred GRNs and is not present in our initial seed. Further, we repeated the functional enrichment analysis to determine overlapping pathways for the enriched seed. We leveraged the identified enriched candidates in these pathways by subsequent inclusion in the seed iteratively until saturation. Once the seed has reached its saturation, we merge the networks of all iterations, separately for each dataset, to generate an aggregated network. This approach goes beyond just candidate enrichment, corresponding to a maximal AD specificity with minimal noise and harvesting lesser known genes in GRNs.

Gene list prioritization by genetic variant analysis

For the consensus network, we identified genes (involved in significant pathways and hub genes) to prioritize them using genetic variant analysis. Multiple genetic variants are attributed in the etiology of complex diseases. To investigate the impact of genetic variation, we extracted AD evidences for single-nucleotide polymorphisms (SNPs) from GWAS catalog [51], GWAS Central [52] and gwasDB [53], resulting in 11,314 SNPs. Further, linkage disequilibrium (LD) analysis was carried out to enrich the list of AD associated genetic variants, which were sorted based on their chromosome location. Linkage disequilibrium is SNP's property on a contiguous stretch of a chromosome that describes the degree to which an allele of one genetic variant is inherited or correlated with an allele of another genetic variant within a population. The LD analysis was performed using HaploReg v2 (developed by Broad Institute of MIT) [54] based on dbSNP-137 [55], motif instances (based on PWMs provided by the ENCODE project database) [56], enhancer annotations (adding 90 cell types from the Roadmap Epigenome Mapping Consortium) [57], and eQTLs (from the GTex eQTL browser) [58]. With LD threshold cutoff of r2 = 0.8, we obtained 115,782 SNPs. Furtheron, these SNPs were filtered based on the ENSEMBL SNP Effect predictor that estimates the influence of SNP variants on the respective transcripts of a gene and their gene products [59], shortlisting 4,831 SNPs. Genes obtained from the aggregated networks were boiled down to those associated with shortlisted SNPs. Finally, these refined genes were ranked using a cumulative score of their SNPs from RegulomeDB [60], dbSNP's functional annotation [55], ENSEMBL's Variant Effect Predictor [61] and regulatory feature annotation by ENSEMBL variant database [62]. RegulomeDB's ranking is based on the functional annotations from ENCODE database [63], chromatin states from the Roadmap Epigenome Consortium [64], DNase-

footprinting [65], position weighted matrix for transcription factor binding [66], and DNA methylation [67].

Results and Discussion

Algorithm convergence and network properties

Under the premise that lesser known genes are not prominently represented in literature, we extended the set of seed genes through functional enrichment (see Section Subnetwork selection and functional enrichment analysis). As depicted in Figure 1, BC3Net10 was applied on the identified four datasets for different seed lists to generate AD GRNs. Iteration 1, where we generated GRNs for SCAIView genes resulted in 10 overlapping and significant pathways between the four datasets. From these pathways, we obtained 820 genes that were earlier not present in the seed. Hence, there is a clear need for further enrichment of the seed, which is done by including these newly identified candidates to the seed and repeating the functional enrichment step. In the second iteration, we identified 38 overlapping significant pathways between the datasets. This iteration continues seven times until there are no newer candidates to be added. Table 2 provides the statistics of the number of pathways identified in each iteration, along with the number of enriched candidate genes that were added to the seed. A detailed list of the enriched candidates (as HGNC symbols) identified in each iteration is provided in Supplementary File S1. A sharp increase in the number of enriched candidates is observed in the first two iterations, which drops to zero in the seventh iteration. We assume that this indicates the completeness of the gene set that belongs to AD, specific to the selected four datasets.

Extension of the GRNs using enriched seed is a knowledge guided approach, which relies on the functional information derived from the gene expression data. We note that the

GRNs grow progressively, both inferred interactions and the participating candidates, but in the process, eliminates few of the previously inferred interactions. A potential reason is that the extension of the expression matrix with new seed contributes to a shift in the significance of the inferred interactions by BC3Net. It implies that although we obtain a final GRN for saturated seed (iteration 7), aggregating networks from earlier iterations could capture interactions that were previously inferred as potential. The fraction of nodes and edges from each iteration that makes up the aggregated network, for each dataset, is presented in Figure 3. We observe that the addition of nodes, in each iteration, across datasets, remained stable whereas the same cannot be said for the edges. The variance in edges could be presumed that the newly added set of genes bring in higher functional relevance through newly inferred interactions in one or the other iteration.

An assessment of the completeness of a GRN for AD specific genes can be precisely estimated by plotting the mean and the variance of the number of nodes and edges present in each dataset for each iteration. From Figure 4 (a), it is evident that the enrichment of the most relevant genes reach saturation. This increases the statistical significance of the GRNs suggesting an increment in the biological confidence. It is apparent that not all the genes present in the seed agree across platforms due to various differing experimental factors. However, we expect functional signatures across the datasets to be more agreeable. Analyzing edges, see Figure 4 (b), we observe that they orient three times, at saturation, to the number of nodes. The relative higher number of edges demonstrate that the gene sets are highly related, showing immense inter-connectivity between several functional modules. The high variance observed, in both nodes and edges, is contributed by the large network size of GSE5281 relative to the other three datasets. Details of the number of nodes and edges present in each dataset at each iteration is provided are Supplementary File S2.

Hub genes

Hub genes have a higher grade of lethality when dysregulated in a pathological condition, referred to as centrality lethality rule [68]. For each aggregated GRN, a gene was defined as a hub gene when it had a higher degree of distribution (>95% quantile). By this criterion, we identified 29 in GSE5281, 8 in GSE44768, 14 in GSE44770, and 1 in GSE44771 as hub genes. Table 3 displays the list of identified hub genes along with their node degree and pathway annotation (only for significant pathways, see Functional homogeneity across datasets section). Interestingly, there were no common hub genes between the four datasets. It was evident that six of the hub genes were perturbed in multiple pathways. Many of the hub genes were functionally enriched in neurotrophin signaling, endocytosis, and estrogen signaling pathways. Additional associated pathways with hub genes include calcium signaling, adipocytokine signaling, NOD-like receptor signaling, insulin signaling, apoptosis, thyroid signaling, and pancreatic secretion. The majority of these hub genes formed a connected subnetwork within each dataset, indicative of a possible cooperative effect in AD pathology (see Supplementary Figure S1). In the case of GSE44771, due to the presence of a single hub gene, we extracted the largest subnetwork associated with HSPA2.

Functional homogeneity across datasets

Are the core functional modules (set of interconnected-genes) unique to a human brain region or do they depict patterns reflecting the tight linkage between different regions of the brain? To address these questions, we compared the final determined significant pathways across the four aggregated GRNs (outlined in Methods). The functional enrichment analysis revealed 187 in GSE5281, 120 in GSE44768, 170 in GSE44770, and 43 in GSE44771 inferred modules within significant KEGG pathways. We computed a

simple overlap between the four GRNs to assess the conserved pathways, resulting in 34 pathways. Because this list contained pathways that were not directly relevant to the core pathophysiology of AD, we categorized them into subsets based on their pertinence to AD, see Table 4. Please refer to Supplementary File S4 for details of summary statistics. From these, we chose to focus on pathways that exacerbate the AD phenotype, classified as "Potential". Table 4 also provides the statistics of the number of genes enriched for these pathways in each dataset. Interestingly, there are no common genes between the four datasets when compared at the pathway level. However, many of the genes are shown to be involved in more than one potential pathway, providing the basis for functional connectivity in AD.

Regulatory underpinning across Consensus network

As described in Section Functional homogeneity across datasets, the genes in different GRNs are complementary for the top significant pathways. Thus, to provide a broader coverage than a single GRN and to infer stronger relationships through consensus, we merged the four aggregated GRNs into one, called consensus network. What we expect is to uplift the most promising pathways due to the assembly of more participating genes. To assess the concept of functional enrichment, we plot the p-values of all the significant pathways, listed in Table 4, for each of the aggregated and consensus GRNs, see Figure 5. From the figure, it is evident that these pathways have attained higher significance level (better p-values) in consensus GRN due to the gene complementarity from the aggregated GRNs.

Prioritizing through genetic variant analysis

We compiled 608 genes from listed significant pathways across datasets (see Table 4) and hub genes. We mapped these genes to the 4,831 shortlisted ENSEMBL SNPs (see Methods). For the obtained 167 mapped genes, we ranked them based on the calculated cumulative score for their potential functional consequences in a disease context. Restricting the ranked genes to the RegulomeDB score of 3, we generated a final list of 44 high ranked genes. In addition, we looked into the AD GWAS meta-analysis study carried out by Lambert et. al [69]. Among all their listed genes carrying genetic AD risks, we found three (AP2A2, DPYSL2, and EPHA1) of them to be present in our 608 gene list, including one (EPHA1) newly reported in their study; these three were added to our final gene list. Please refer to Table 5 for detailed ranking and RegulomeDB score. Additional investigation revealed 14 out of 47 genes from our final gene list are either validated by eQTLs studies or experimentally evident that the SNPs are linked to the active promoter region of the gene. These genes include IL1B, NSF, HLA-F, NOTCH4, VCL, PSAP, STX2, GGA2, STK11, CSF3R, LMNA, CTNNA2, HLA-C and RAB11FIP4. When we performed a comprehensive analysis of the biomedical literature, we found that many of these genes had no evidence of being linked to AD, but were rather known to be involved in AD co-morbidity diseases (see Supplementary Table S1).

Well known prioritized AD candidates

Apart from the new novel candidates, our method also determined well-known candidates (nearly 50 articles in AD) such as IL1B, NTRK2, GRIN2A, FYN, and DPYSL2. The IL1B gene is a pro-inflammatory cytokine that has been long studied for its modulatory effect in AD. It is reported that the expression of IL1B significantly increases with the increase of AD-related neurofibrillary pathology [70]. Synaptic plasticity, such as long-term

potentiation, is crucial for learning and memory. A neurotransmitter modulator, BDNF, mediates neuronal survival and plasticity by regulating neurotrophins through NTRK2. AD patients with cognitive deficits have been accounted with reduced levels of BDNF [71–73]. Similarly, GRIN2A is a subunit of NMDA receptors, whose reduced expression increases the vulnerability of neurons to excitotoxicity in AD, correlated with cognitive impairment due to reduced plasticity [74,75]. A strong correlation between lower levels of BDNF and cognitive deficits in AD patients was recently reported by Buchman et al. [76]. Recent research work has suggested BDNF as an upstream regulator of FYN gene, a Src family kinase, leading to enhanced cascade effect of NMDA mediated excitotoxicity and regulates the activity of hyperphosphorylated tau [77,78]. In addition, it mediates the synaptic deficits that are induced by Aβ [79]. DPYSL2 mediates synaptic signaling to facilitate neuronal guidance through regulation of calcium channels. Furthermore, FYN phosphorylates DPYSL2 within the brain and its hyperphosphorylation is causally related to Aβ neurotoxicity [80]. Taken together, these findings suggest that synaptic transmission is critical for regulating Aβ production in AD. Further studies, along these lines, may provide insights into the precise molecular mechanism underlying this part of AD etiology.

Mechanistic interpretation of newly prioritized candidates in neurotransmission

Neurotransmission is a pivotal brain function that declines with progressing age. However, in the case of AD there is a drastic and non-uniform deterioration of synaptic neurotransmission [81]. It is known that soluble oligomeric amyloid-β, rather than insoluble deposits that form plaques (extracellular), are detrimental to synaptic currents through calcium channel modulation, leading to excitotoxic cascades that mediate AD progression [82] and are related to the formation of neurofibrillary tangles (intracellular) [83]. Emerging research strongly supports the hypothesis of dysregulated calcium

homeostasis influencing the presence of neurotoxic Aβ in AD patients [84]. Increased endocytosis activity, enlarged endosomes, has been reported by Cataldo et al. [85] as the earliest intraneuronal neuropathologic feature of AD, subsequently impairing the modulation of NMDA receptor. NMDA excitotoxicity leads to the pathological overload of calcium resulting in synaptic impairment and ultimately neuronal death [86].

We observed that three of the "Potential" pathways are significantly involved in neurotransmission: calcium signaling, endocytosis, and synaptic vesicle cycle (see Figure 6). To assess the modularity of the prioritized candidates in these identified pathways, we extracted the functional relevance of their combination. This confirms our previous findings associated with well-known candidates (see Section Well known prioritized AD candidates). To gain new insights in this context we focused on lesser known prioritized candidates in Alzheimer's that are involved in these three pathways (less than 5 publications): STX2, HLA-F, HLA-C, RAB11FIP4, ARAP3, AP2A2, ATP2B4, ATP2A3, and ITPR2. Below, we briefly discuss the possibility of these candidates to presumably bear potential as new targets in AD (detailed description is provided in Supplementary File S5).

The presence of Aβ oligomers impairs the process of STX2 binding to SNARE proteins hindering the effective release of neurotransmitter during synaptic vesicle fusion in the presence of increased calcium influx [87,88]. From several previous studies, one can postulate that HLA-F and HLA-C mediated dysregulated trafficking of amyloid plaques in endocytosis could be correlated to the memory deficits in early AD [89–91]. Several recent evidence point to the fact that faulty amyloid-β processing can be detected in the membrane trafficking events (linked to RAB11 proteins) of early endosomes, promoting an effective early diagnosis [92,93]. ARAP3 modulates actin cytoskeleton's remodeling by regulating

ARF and RHO family members [94] and a growing body of evidence suggest that axonal transport defects due to its abnormality could be responsible for neurite degeneration and tau toxicity [95–98]. Impairment of APP shuttling by AP2A2 (part of AP-2 complex [99]) from the endocytotic pathway to autophagy degradation leads to intracellular aggregation of Aβ [100]. The next three candidates (ATP2B4, ATP2A3, and ITPR2) participate in neuronal calcium shuttling. A substantial body of evidence indicates ATP2B4, a plasma membrane Ca(2+) ATPases (PMCAs) is inhibited by Aβ peptides [101], causing cell death [102]. Similarly, ATP2A3's function in handling calcium load and release is perturbed by the mutation in PSEN1 (regulates the intramembrane Aβ processing) [103]. Increased expression of ITPR2 could lead to calcium toxicity in neurons and finally cell death [104,105].

Conclusion

The identification of biological mechanisms underlying normal physiology and – when dysregulated – contributing to or even directly causing disease phenotypes is a key objective of current integrative biology. Strategies, both data- and knowledge-driven, for mechanism-identification have shown to deliver valuable insights into disease mechanisms, however, both approaches have their specific drawbacks. Here, we demonstrate a new approach that combines literature-based knowledge and data-driven analysis through gene regulatory networks in a flexible and adaptive way. Thus, allowing us to identify stable and robust patterns of co-expressed genes across several large disparate datasets, in parallel, which enhances the interpretability around "interesting patterns" of co-regulated genes.

We developed an adapted version of BC3Net, called as BC3Net10, that supports a more fine-granular specification of functional context by "injecting" sets of seed genes (derived

from literature) into the algorithm. The seed genes were iteratively extended through functional enrichment applied on generated GRNs until convergence. Through several iterations of "selecting and injecting seed genes" and subsequent co-expression analysis, we come up with stable, knowledge-instructed GRNs across several experiments. We show the ability of our approach to identify functional context around subtle signals that would typically be expected for highly individual "modifier" functions not in the core of a dysregulation event, but have the potential to modulate the clinical path of a disease. Hence, making this approach ideally suited for biomarker identification. We show that by the enhanced functional interpretation of the GRNs shed more light on the role of neurotransmission physiology in early dysregulation events presumed to be part of Alzheimer´s Disease etiology. This warrant further investigation of their potential as therapeutic targets.

We would like to point out that there is more potential to the method presented here: in the course of IMI-project AETIONOMY we found limited coverage of signals in knowledge based models coming from the analysis of either gene expression or genetic variation information (GWAS studies). The methodology presented here bears the potential to establish biologically meaningful context around "isolated signals" in knowledge-based models to "embed" previously "non-interpretable" (at functional level) genes into a wider (knowledge based) context. Insights drawn from this approach could provide a novel foundation for the formation of new hypotheses. Although microarray data is the obvious starting point, the next logical step would be to extend this work to incorporate orthogonal datatypes such as NGS and single cell data. This could provide a broader view of disease etiology and enable comprehensive in silico investigations. It remains to be shown that the method we introduce here scales up to a really large number of experiments of different sample size.

Figures

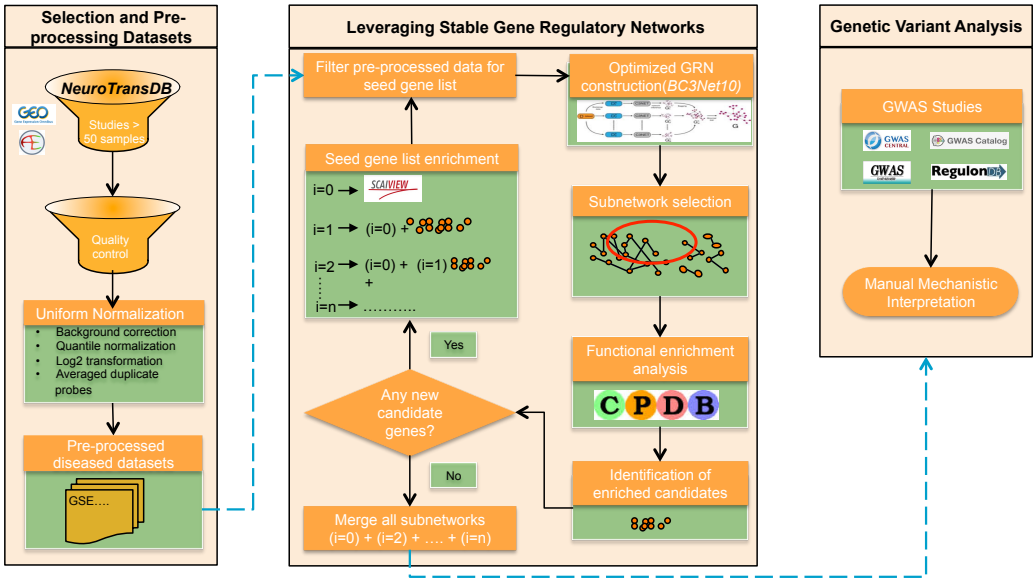


Figure 1 The overall strategy applied to obtain robust gene expression patterns across public Alzheimer's disease studies. Firstly, four gene expression datasets were shortlisted from NeuroTransDB database. The selected studies underwent preprocessing and quality control. In each dataset, the intensity values were limited to the seed gene list. To enrich the seed, functional enrichment was applied where genes from the identified significant pathways from each dataset's subnetwork (edge weight>0.5), generated using BC3Net10 approach, were included. When no additional genes were identified, subnetworks from each iteration, separately for each dataset, were merged into an aggregated network for further prioritization of the genes using genetic variant analysis

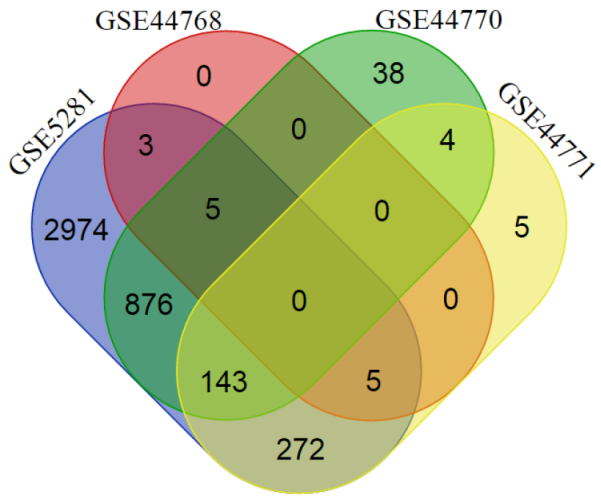
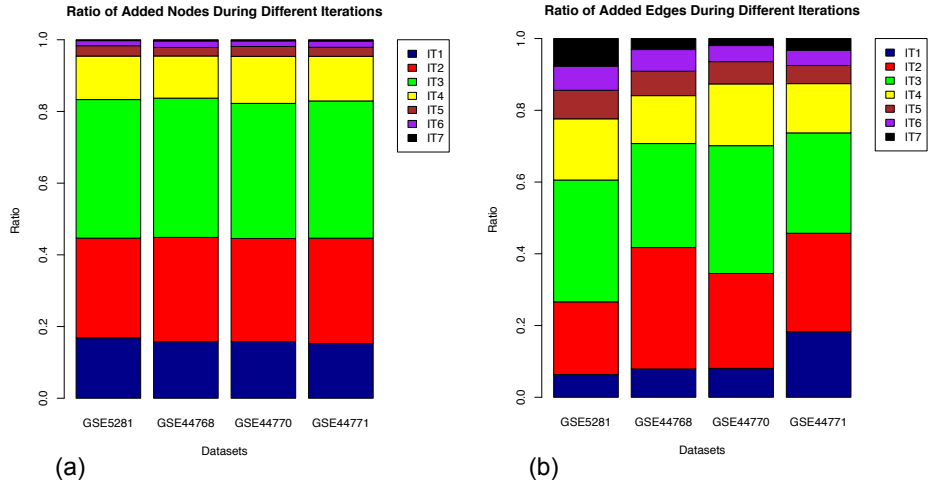


Figure 2 Venn diagram depicting the gene overlap between the subnetworks (edge weight>0.5) of the four datasets, generated using the initial seed. The initial seed was compiled from top 500 genes retrieved by querying SCAIView for Alzheimer's disease related genes. It is evident that there are no common genes among the four dataset's subnetworks. Differing factors between platforms, analytical methods, tissue source, etc. could contribute to such a behavior.



(a) Fraction of added nodes in different iterations (b) Fraction of added edges in different iterations

Figure 3 Stratification of the nodes and edges in four aggregated networks. Each stack in the bar plot represents the fraction of nodes added in that iteration (IT) relative to the aggregated network (considered as 1). The addition of nodes remained stable across the datasets in each iteration. However, the inclusion of edges varies, which could be presumed due to newly inferred interactions from the newly included nodes in each iteration

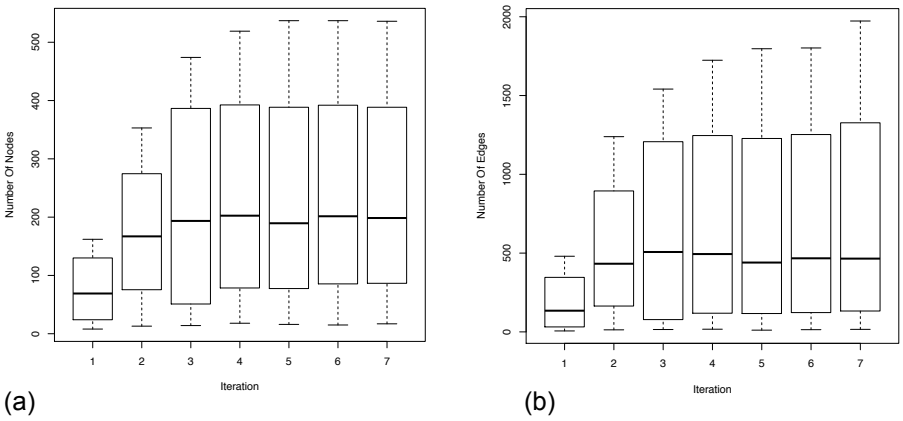


Figure 4 Mean and variance distribution across four datasets for the added nodes and edges in each iteration. Enrichment of nodes and edges reach saturation after 7th iteration, suggesting the completeness of the generated GRNs. Relatively high number of edges (see y-axis range) show immense inter-connectivity between the genes in the GRNs. (a) Boxplot for mean and variance distribution of nodes (b) Boxplot for mean and variance distribution of edges

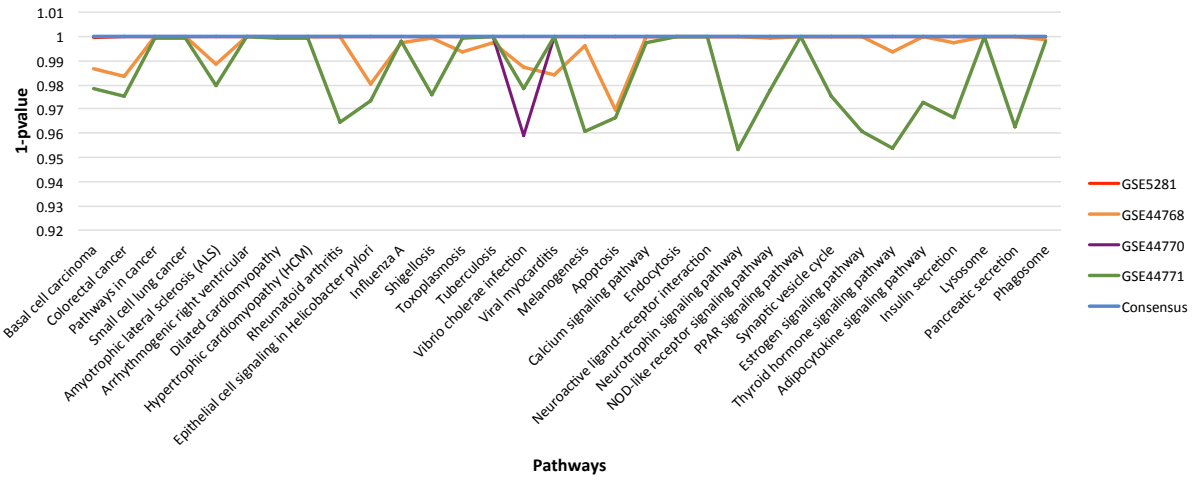


Figure 5 The landscape of p-value for the final list of significant pathways. For easy visualization, we have used 1-pvalue instead of p-value on Y-axis. Each line in the graph represents aggregated GRN for specified dataset (see chart legend). The listed pathways show higher significance level in consensus GRN in comparison to the individual dataset aggregated GRNs

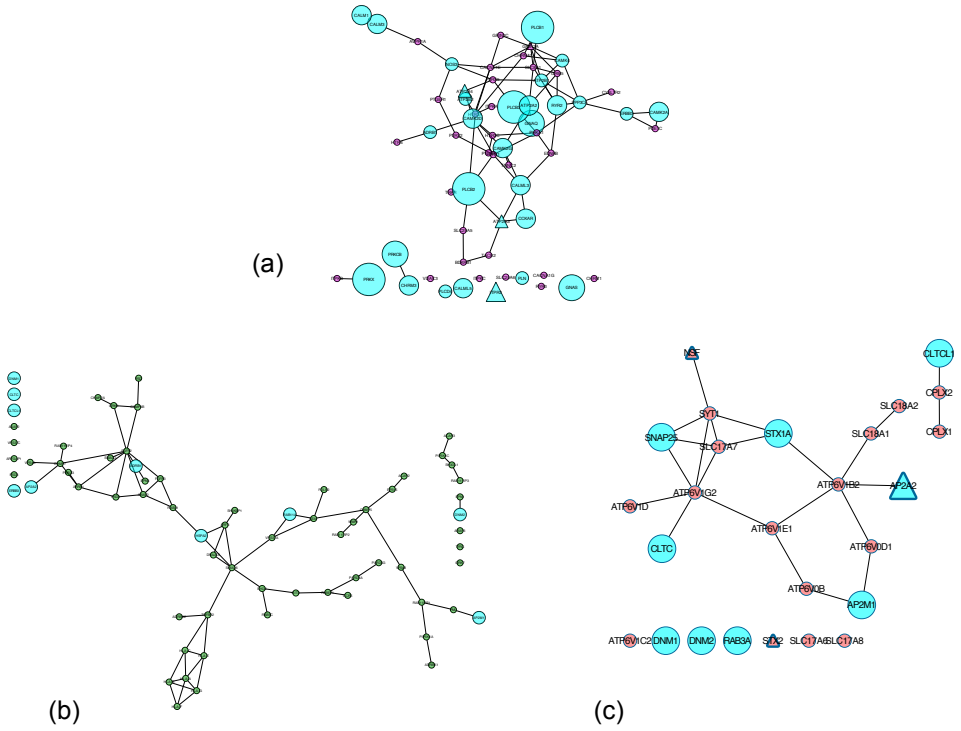


Figure 6 Subnetworks of the three shortlisted potential pathways (extracted from consensus network) involved in neurotransmission. Nodes in Cyan are involved in more than one pathways and the size of the nodes depends on the number of pathways involved. Triangle nodes represent the presence of a SNP. (a) Calcium signaling pathway (b) Endocytosis pathway (c) Synaptic vesicle cycle

Tables

GEO ID	**Number of Samples**		**Sample Source**	**Stage**	**Platform**
	Diseased	Control			
GSE5281	87	74	Entorhinal cortex, Hippocampus, Primary visual cortex, Prefrontal cortex, Medial	-	Affymetrix HG U133 Plus 2

			temporal gyrus, Superior frontal gyrus		
GSE44768	129	101	Cerebellum	LOAD	Rosetta/Merck Human 44k 1.1 microarray
GSE44771	129	101	Visual cortex	LOAD	Rosetta/Merck Human 44k 1.1 microarray
GSE44770	129	101	Dorsolateral prefrontal cortex	LOAD	Rosetta/Merck Human 44k 1.1 microarray
GSE13214	52	40	Hippocampal, Cortex frontal	Braak 4-6	Homo sapiens 4.8K 02-01 amplified cDNA
GSE15222	176	187	Cortical tissue	LOAD	Sentrix HumanRef-8 Expression BeadChip
GSE29676	350	200	Blood	-	Invitrogen ProtoArray v5.0
GSE33528	615	600	Blood	LOAD	Illumina Human-Hap650Yv2 Genotyping BeadChip

Table 1 List of datasets shortlisted from NeuroTransDB for generating gene regulatory networks. Final selected studies are highlighted in bold

Iteration (i)	**Seed**	**No. of overlapping pathways between the four datasets**	**No. of enriched candidate genes obtained from the overlapping pathways**
1	SCAIView (500)	10	820
2	i1+820	38	1148
3	i2 + 1148	30	361
4	i3+361	30	84
5	i4+84	32	41
6	i5+41	33	7
7	i6+7	37	-

Table 2 Statistics of the iterative functional enrichment approach

GEO ID	**Gene Symbols**	**Hub Degree**	**Pathway Annotation (CPDB)**	**Similar results in other datasets?**
GSE5281	HFE	244	-	-
	ATP2A3	162	Calcium signaling, pancreatic secretion	-
	GLP1R	150	Insulin Secretion	-
	ADRBK1	145	Endocytosis	GSE44770
	CACNG4, CACNG6	141	-	-

	KCNJ5	132	Estrogen signaling	-
	P2RX2	130	Calcium signaling	GSE44770
	KPNA2	122	-	-
	NOX1	118	-	-
	CACNG5	113	-	-
	EPN1	113	Endocytosis	-
	WAS	112	-	-
	CASP10	111	Apoptosis	-
	HSPB6, EPHA4	109	-	-
	ADNP	108	-	-
	DNAH3	106	-	-
	GRIN2A	105	Calcium signaling	-
	UBQLN1	101	-	-
	IL34, ATP5A1, UBE2L3	100	-	-
	DPYSL2	99	-	-
	FOLR2	98	Endocytosis	-
	NPR1	96	-	-
	DNM1L, KLC1,	92	-	-

	ATP5G3			
GSE44768	RASGRF1	80	-	-
	DNAL4	63	-	-
	EPHA1	60	-	-
	CHRND	59	-	-
	TRPC1	54	Pancreatic secretion	GSE5281, GSE44770
	PAK7	50	-	-
	NDUFA4	44	-	-
	CHMP4B	44	Endocytosis	-
GSE44770	IVNS1ABP	103	-	-
	FGF18	92	-	-
	ATF2	90	Estrogen signaling, Insulin secretion	-
	CTSG	88	-	-
	GABRE	86	-	-
	FBXL2	81	-	-
	GAPDH	75	-	-
	DIO1	72	Thyroid hormone signaling	-

	CACNB3, CDK2	66	-	-
	NFKBIB	66	Adipocytokine signaling, neurotrophin signaling, NOD-like receptor signaling	GSE44768
	PRDM4	64	Neurotrophin signaling	-
	MAPK9	63	Adipocytokine signaling, neurotrophin signaling, NOD-like receptor signaling	-
	PIK3CB	63	Apoptosis, estrogen signaling, neurotrophin signaling, thyroid hormone signaling	GSE5281
GSE44771	HSPA2	18	Endocytosis, estrogen signaling	-

Table 3 Hub genes identified in the aggregated network for the four datasets. The genes are sorted by their hub degree within each dataset. Only significant pathways are listed here (see Table 4 for the list)

Common Pathways	**Pathway Category**	**Total no. of genes in the pathway**	**Number of genes enriched for the pathway**				
			GSE5281	**GSE44768**	**GSE44770**	**GSE44771**	**Consensus**
Cancer	Basal cell carcinoma	55	5	2	7	1	15
Cancer	Colorectal cancer	62	6	2	8	1	14

Cancer	Pathways in cancer	398	64	27	40	3	119
Cancer	Small cell lung cancer	86	15	4	9	2	27
Comorbidity	Amyotrophic lateral sclerosis (ALS)	51	14	2	5	1	18
Comorbidity	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	13	5	8	2	21
Comorbidity	Dilated cardiomyopathy	90	17	6	7	2	26
Comorbidity	Hypertrophic cardiomyopathy (HCM)	83	14	7	7	2	23
Comorbidity	Rheumatoid arthritis	91	10	4	7	1	20
Infection	Epithelial cell signaling in Helicobacter pylori infection	68	9	2	6	1	16
Infection	Influenza A	177	25	4	22	2	46
Infection	Shigellosis	61	12	3	7	1	19
Infection	Toxoplasmosis	120	14	3	12	2	26
Infection	Tuberculosis	179	21	4	23	3	46
Infection	Vibrio cholera infection	54	9	2	2	1	13

Infection	Viral myocarditis	60	12	2	5	2	19
Others	Melanogenesis	101	18	3	10	1	30
Others	Neuroactive ligand-receptor interaction	275	57	24	30	3	98
Potential	Apoptosis	86	14	2	6	1	20
Potential	Calcium signaling pathway	180	43	12	16	2	62
Potential	Endocytosis	213	47	10	21	4	70
Potential	Neurotrophin signaling pathway	120	24	6	17	1	44
Potential	NOD-like receptor signaling pathway	57	9	3	6	1	16
Potential	PPAR signaling pathway	69	11	4	9	2	22
Potential	Synaptic vesicle cycle	63	15	4	8	1	26
Potential	Adipocytokine signaling pathway	70	17	6	8	1	27
Potential	Insulin secretion	86	18	3	10	1	28
Potential	Pancreatic secretion	96	21	5	9	1	30

Potential (hormones)	Estrogen signaling pathway	100	23	4	7	1	32
Potential (hormones)	Thyroid hormone signaling pathway	119	26	3	10	1	37
Potential (others)	Lysosome	122	13	7	11	4	33
Potential (others)	Phagosome	155	31	4	16	2	48

Table 4 Landscape of significant pathways (p-value<0.05) determined across datasets

Rank	**Gene Symbol**	**RegulomeDB score**	**No. evidences for Alzheimer's disease**	**Pathways involved**
1	IL1B	1b	1073	Apoptosis, NOD-like receptor signaling
2	NSF	1d	8	Synaptic vesicle cycle
3	HLA-F	1f	0	Endocytosis
4	NOTCH4	1f	3	Thyroid hormone signaling
5	VCL	1f	10	Shigellosis
6	PSAP	1f	3	Lysosome
7	STX2	1f	2	Synaptic vesicle cycle
8	GGA2	1f	4	Lysosome
9	STK11	1f	7	Adipocytokine signaling

10	CSF3R	1f	5	Pathways in cancer
11	LMNA	1f	11	Arrhythmogenic right ventricular cardiomyopathy(ARVC), Dilated cardiomyopathy, Hypertrophic cardiomyopathy(HCM)
12	CTNNA2	1f	3	Arrhythmogenic right ventricular cardiomyopathy(ARVC)
13	HLA-C	1f	1	Endocytosis
14	RAB11FIP4	1f	0	Endocytosis
15	GRIN2A	2a	52	Calcium signaling
16	RBX1	2a	0	Viral Myocarditis
17	KCNJ5	2a	0	Estrogen signaling
18	EPHA4	2b	18	Hub Genes
19	CACNG4	2b	0	Arrhythmogenic right ventricular cardiomyopathy(ARVC), Dilated cardiomyopathy, Hypertrophic cardiomyopathy(HCM)
20	PLA2G5	2b	7	Pancreatic secretion
21	ATP2B4	2b	1	Calcium signaling, pancreatic secretion
22	P2RY14	2b	0	Neuroactive ligand receptor interaction
23	P2RY13	2b	0	Neuroactive ligand receptor interaction

24	PTGER4	2b	11	Neuroactive ligand receptor interaction
25	ARAP3	2b	0	Endocytosis
26	FGF1	2b	22	Pathways in cancer
27	RPS6KA2	2b	0	Neurotrophin signaling
28	RAPGEF1	2b	0	Neurotrophin signaling
29	GABBR2	2b	1	Estrogen signaling
30	PRF1	2b	1	Viral myocarditis
31	ITGA8	2b	0	Arrhythmogenic right ventricular cardiomyopathy(ARVC), Dilated cardiomyopathy, Hypertrophic cardiomyopathy(HCM)
32	AP2A2	2b	0	Endocytosis, Synaptic vesicle cycle
33	ITPR2	2b	2	Calcium signaling, Estrogen signaling, pancreatic secretion
34	MED13L	2b	0	Thyroid hormone signaling
35	COL4A1	2b	0	Pathways in cancer
36	KCNJ6	2b	3	Estrogen signaling
37	ATP2A3	2b	0	Calcium signaling, Pancreatic secretion
38	ASAP2	3a	1	Endocytosis

39	FYN	3a	70	Viral myocarditis
40	NTRK2	3a	124	Neurotrophin signaling
41	PAK1	3a	7	Epithelial cell signaling in Helicobacter pylori infection
42	COL4A2	3a	0	Small cell lung cancer, Pathways in cancer
43	BMP4	3a	5	Thyroid hormone signaling
44	GABRB3	3a	0	Neuroactive ligand receptor interaction
45	CEBPB	3a	12	Tuberculosis
46	EPHA1	5	31	Hub Genes
47	DPYSL2	5	47	Hub Genes

Table 5 List of genes prioritized using genetic variant analysis

List of Abbreviations

AD Alzheimer's Disease

GRNs Gene Regulatory Networks

Aβ Amyloid Beta

DE Differentially Expressed

LD Linkage Disequilibrium

Funding

This work received financial support from the Innovative Medicines Initiative Joint Undertaking under grant agreements n◦ 115568 and n◦ 115736 resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2017-2013) and EFPIA companies' in kind contribution.

Competing interests

The authors declare that they have no competing interests.

Declaration

The current affiliation of Dr. Philipp Senger is CLS Head of Translational R&D, Bayer CropScience, Alfred-Nobel-Straße 50, 40789 Monheim, Germany. He contributed to this work during his employment at Fraunhofer SCAI.

Author's contributions

SBK, PS, and MHA conceived the idea and designed the strategy. SBK and TR are the main contributors to the work and manuscript writing under the guidance of PS. MHA critically reviewed the manuscript and contributed to the writing. MN supported with the genetic variant analysis and manuscript writing. RMS supported in understanding the technical aspects of BC3Net algorithm, data pre-processing, application of BC3Net and pathway analysis. All the co-authors read and approved the final manuscript.

Acknowledgements

We are grateful for the financial support received for this work from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY grant agreement n◦115568, resources of which are composed of financial contribution from the European

Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. We thank the anonymous reviewers for their valuable comments.

Supplementary material

GeneSeedList S1.xls—Enriched candidate genes in each iteration

This is a .xls file. It contains the list of enriched candidate genes that were added in each iteration to the seed gene list. The genes are represented as HGNC symbols.

NumberOfNodesAndEdges S2.xls — Network statistics for each dataset in each iteration

This is a .xls file. It provides detailed number of nodes and edges present in each dataset for each iteration.

Outlier S3.xlsx — Outlier arrays list

This is a .xlsx file. Each tab in the file provides the details of outlier arrays in each dataset that were discarded from our work due to low quality. In addition, the phenotype of the array is also provided.

Pathway Statistics S4.xls — Summary statistics of the enriched pathways

This is a .xls file. Each tab provides the detailed summary statistics for the individual dataset (including consensus) such as p-value, adjusted p-value, FDR, etc. of the enriched pathways that are obtained from ConsensusPathDB.

Mechanistic Details S5.doc — Detailed mechanistic information of the prioritized candidates

This is a .doc file. It provides detailed mechanistic information for each of the newly prioritized candidates.

Supplementary Figure S1.jpg—Connected subnetwork formed by the hub genes in each dataset

Supplementary Table S1.xls—comprehensive literature analysis of prioritized candidates This is a .xls file. It contains the list of diseases that the prioritized candidate

genes are involved in. The information was retrieved by querying the SCAIView knowledge discovery tool.

References

[1] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and costs now and in the future," 2016.

[2] S. Gauthier, M. Albert, N. Fox, M. Goedert, M. Kivipelto, J. Mestre-Ferrandiz, and L. T. Middleton, "Why has therapy development for dementia failed in the last two decades?," *Alzheimers. Dement.*, vol. 12, no. 1, pp. 60–4, Jan. 2016.

[3] World Health Organization, "Dementia: a public health priority," 2012.

[4] S. Report and P. F. Impact, "2016 Alzheimer's disease facts and figures," *Alzheimer's Dement.*, vol. 12, no. 4, pp. 459–509, Apr. 2016.

[5] H. Braak and E. Braak, "Frequency of stages of Alzheimer-related lesions in different age categories," *Neurobiol. Aging*, vol. 18, no. 4, pp. 351–357, 1997.

[6] K. Blennow, M. J. de Leon, and H. Zetterberg, "Alzheimer's disease.," *Lancet (London, England)*, vol. 368, no. 9533, pp. 387–403, Jul. 2006.

[7] A. C. Naj and G. D. Schellenberg, "Genomic variants, genes, and pathways of Alzheimer's disease: An overview," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 174, no. 1, pp. 5–26, Jan. 2017.

[8] K. Blennow, B. Dubois, A. M. Fagan, P. Lewczuk, M. J. De Leon, and H. Hampel, "Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease," *Alzheimer's Dement.*, vol. 11, no. 1, pp. 58–69, 2015.

[9] L. Mucke, "Neuroscience: Alzheimer's disease," *Nature*, vol. 461, no. 7266, pp. 895–897, Oct. 2009.

[10] T. J. Montine, C. H. Phelps, T. G. Beach, E. H. Bigio, N. J. Cairns, D. W. Dickson, C. Duyckaerts, M. P. Frosch, E. Masliah, S. S. Mirra, P. T. Nelson, J. A. Schneider, D. R. Thal, J. Q. Trojanowski, H. V. Vinters, and B. T. Hyman, "National institute on aging-Alzheimer's association guidelines for

the neuropathologic assessment of Alzheimer's disease: A practical approach," *Acta Neuropathol.*, vol. 123, no. 1, pp. 1–11, 2012.

[11] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, D. C. Park, E. M. Reiman, C. C. Rowe, E. Siemers, Y. Stern, K. Yaffe, M. C. Carrillo, B. Thies, M. Morrison-Bogorad, M. V. Wagster, and C. H. Phelps, "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 280–292, 2011.

[12] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, and C. H. Phelps, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 270–279, 2011.

[13] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 263–269, 2011.

[14] C. M. Karch, C. Cruchaga, and A. M. Goate, "Alzheimer's Disease Genetics: From the Bench to the Clinic," *Neuron*, vol. 83, no. 1, pp. 11–26, Jul. 2014.

[15] B. T. Hyman, H. L. West, G. W. Rebeck, S. V. Buldyrev, R. N. Mantegna, M. Ukleja, S. Havlin, and H. E. Stanley, "Quantitative analysis of senile plaques in Alzheimer disease: Observation of log-normal size distribution and molecular epidemiology of differences associated with apolipoprotein E genotype and trisomy 21 (Down syndrome)," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 8, pp. 3586–3590, 1995.

[16] R. J. Guerreiro, M. Baquero, R. Blesa, M. Boada, J. M. Brás, M. J. Bullido, A. Calado, R. Crook, C.

Ferreira, A. Frank, T. Gómez-Isla, I. Hernández, A. Lleó, A. Machado, P. Martínez-Lage, J. Masdeu, L. Molina-Porcel, J. L. Molinuevo, P. Pastor, J. Pérez-Tur, R. Relvas, C. R. Oliveira, M. H. Ribeiro, E. Rogaeva, A. Sa, L. Samaranch, R. Sánchez-Valle, I. Santana, L. Tàrraga, F. Valdivieso, A. Singleton, J. Hardy, and J. Clarimón, "Genetic screening of Alzheimer's disease genes in Iberian and African samples yields novel mutations in presenilins and APP," *Neurobiol. Aging*, vol. 31, no. 5, pp. 725–731, 2010.

[17] V. Dobricic, E. Stefanova, M. Jankovic, N. Gurunlian, I. Novakovic, J. Hardy, V. Kostic, and R. Guerreiro, "Genetic testing in familial and young-onset Alzheimer's disease: Mutation spectrum in a Serbian cohort," *Neurobiol. Aging*, vol. 33, no. 7, p. 1481.e7-1481.e12, 2012.

[18] A. Alzheimer, "Uber eine eigenartige Erkrankung der Hirnrinde.," *Allg Z Psychiat Psych-Gerichtl Med*, vol. 64, pp. 146–8, 1907.

[19] J. Joseph, B. Shukitt-Hale, N. A. Denisova, A. Martin, G. Perry, and M. A. Smith, "Copernicus revisited: amyloid beta in Alzheimer's disease.," *Neurobiol. Aging*, vol. 22, no. 1, pp. 131–46.

[20] R. J. Castellani and M. A. Smith, "Compounding artefacts with uncertainty, and an amyloid cascade hypothesis that is 'too big to fail'.," *J. Pathol.*, vol. 224, no. 2, pp. 147–52, Jun. 2011.

[21] J. Hardy and D. Allsop, "Amyloid deposition as the central event in the aetiology of Alzheimer's disease.," *Trends Pharmacol. Sci.*, vol. 12, no. 10, pp. 383–8, Oct. 1991.

[22] C. M. van Duijn, P. de Knijff, M. Cruts, A. Wehnert, L. M. Havekes, A. Hofman, and C. Van Broeckhoven, "Apolipoprotein E4 allele in a population-based study of early-onset Alzheimer's disease.," *Nat. Genet.*, vol. 7, no. 1, pp. 74–8, May 1994.

[23] L. Janssen, C. Keppens, P. P. De Deyn, and D. Van Dam, "Late age increase in soluble amyloid-beta levels in the APP23 mouse model despite steady-state levels of amyloid-beta-producing proteins," *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1862, no. 1, pp. 105–112, Jan. 2016.

[24] E. Head, D. Powell, B. T. Gold, and F. A. Schmitt, "Alzheimer's Disease in Down Syndrome.," *Eur. J. Neurodegener. Dis.*, vol. 1, no. 3, pp. 353–364, Dec. 2012.

[25] M. E. Murray and D. W. Dickson, "Is pathological aging a successful resistance against amyloid-beta

or preclinical Alzheimer's disease?," *Alzheimers. Res. Ther.*, vol. 6, no. 3, p. 24, 2014.

[26] M. G. Spillantini and M. Goedert, "Tau pathology and neurodegeneration," *Lancet Neurol.*, vol. 12, no. 6, pp. 609–622, Jun. 2013.

[27] H. Shi, G. Zhang, M. Zhou, L. Cheng, H. Yang, J. Wang, J. Sun, and Z. Wang, "Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations," *PLoS One*, vol. 11, no. 2, pp. 1–15, 2016.

[28] L. M. Ittner and J. Götz, "Amyloid-β and tau--a toxic pas de deux in Alzheimer's disease.," *Nat. Rev. Neurosci.*, vol. 12, no. 2, pp. 65–72, Feb. 2011.

[29] M. E. Murray, V. J. Lowe, N. R. Graff-Radford, A. M. Liesinger, A. Cannon, S. A. Przybelski, B. Rawal, J. E. Parisi, R. C. Petersen, K. Kantarci, O. A. Ross, R. Duara, D. S. Knopman, C. R. Jack, and D. W. Dickson, "Clinicopathologic and 11C-Pittsburgh compound B implications of Thal amyloid phase across the Alzheimer's disease spectrum.," *Brain*, vol. 138, no. Pt 5, pp. 1370–81, May 2015.

[30] M. R. Brier, B. Gordon, K. Friedrichsen, J. McCarthy, A. Stern, J. Christensen, C. Owen, P. Aldea, Y. Su, J. Hassenstab, N. J. Cairns, D. M. Holtzman, A. M. Fagan, J. C. Morris, T. L. S. Benzinger, and B. M. Ances, "Tau and A imaging, CSF measures, and cognition in Alzheimers disease," *Sci. Transl. Med.*, vol. 8, no. 338, p. 338ra66-338ra66, May 2016.

[31] E. Karran and B. De Strooper, "The amyloid cascade hypothesis: are we poised for success or failure?," *J. Neurochem.*, vol. 139, pp. 237–252, 2016.

[32] G. S. Bloom, "Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis.," *JAMA Neurol.*, vol. 71, no. 4, pp. 505–8, Apr. 2014.

[33] K. Herrup, "Reimagining Alzheimer's disease--an age-based hypothesis.," *J. Neurosci.*, vol. 30, no. 50, pp. 16755–62, Dec. 2010.

[34] M. Baumgart, H. M. Snyder, M. C. Carrillo, S. Fazio, H. Kim, and H. Johns, "Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective.," *Alzheimers. Dement.*, vol. 11, no. 6, pp. 718–26, Jun. 2015.

[35] K. Lunnon, A. Keohane, R. Pidsley, S. Newhouse, J. Riddoch-Contreras, E. B. Thubron, M. Devall, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, L. Schalkwyk, R. Dobson, A. N. Malik, J. Powell, S. Lovestone, and A. Hodges, "Mitochondrial genes are altered in blood early in Alzheimer's disease," *Neurobiol. Aging*, vol. 53, pp. 36–47, May 2017.

[36] A. Nunomura, G. Perry, G. Aliev, K. Hirai, A. Takeda, E. K. Balraj, P. K. Jones, H. Ghanbari, T. Wataya, S. Shimohama, S. Chiba, C. S. Atwood, R. B. Petersen, and M. A. Smith, "Oxidative damage is the earliest event in Alzheimer disease.," *J. Neuropathol. Exp. Neurol.*, vol. 60, no. 8, pp. 759–67, Aug. 2001.

[37] Y. Zhao and B. Zhao, "Oxidative Stress and the Pathogenesis of Alzheimer's Disease," *Oxid. Med. Cell. Longev.*, vol. 2013, pp. 1–10, 2013.

[38] L. A. Craig, N. S. Hong, and R. J. McDonald, "Revisiting the cholinergic hypothesis in the development of Alzheimer's disease," *Neurosci. Biobehav. Rev.*, vol. 35, no. 6, pp. 1397–1409, 2011.

[39] E. J. Goetzl, A. Boxer, J. B. Schwartz, E. L. Abner, R. C. Petersen, B. L. Miller, and D. Kapogiannis, "Altered lysosomal proteins in neural-derived plasma exosomes in preclinical Alzheimer disease," *Neurology*, vol. 85, no. 1, pp. 40–47, Jul. 2015.

[40] D. M. Wolfe, J.-H. Lee, A. Kumar, S. Lee, S. J. Orenstein, and R. A. Nixon, "Autophagy failure in Alzheimer's disease and the role of defective lysosomal acidification.," *Eur. J. Neurosci.*, vol. 37, no. 12, pp. 1949–61, Jun. 2013.

[41] R. S. Vest and C. J. Pike, "Gender, sex steroid hormones, and Alzheimer's disease.," *Horm. Behav.*, vol. 63, no. 2, pp. 301–7, Mar. 2013.

[42] M. J. Berridge, "Calcium regulation of neural rhythms, memory and Alzheimer's disease," *J. Physiol.*, vol. 592, no. 2, pp. 281–293, Jan. 2014.

[43] V. Calsolaro and P. Edison, "Neuroinflammation in Alzheimer's disease: Current evidence and future directions.," *Alzheimers. Dement.*, vol. 12, no. 6, pp. 719–32, Jun. 2016.

[44] M. T. Heneka, M. J. Carson, J. El Khoury, G. E. Landreth, F. Brosseron, D. L. Feinstein, A. H. Jacobs, T. Wyss-Coray, J. Vitorica, R. M. Ransohoff, K. Herrup, S. A. Frautschy, B. Finsen, G. C. Brown,

A. Verkhratsky, K. Yamanaka, J. Koistinaho, E. Latz, A. Halle, G. C. Petzold, T. Town, D. Morgan, M. L. Shinohara, V. H. Perry, C. Holmes, N. G. Bazan, D. J. Brooks, S. Hunot, B. Joseph, N. Deigendesch, O. Garaschuk, E. Boddeke, C. A. Dinarello, J. C. Breitner, G. M. Cole, D. T. Golenbock, and M. P. Kummer, "Neuroinflammation in Alzheimer's disease.," *Lancet. Neurol.*, vol. 14, no. 4, pp. 388–405, Apr. 2015.

[45] F. G. De Felice, M. V. Lourenco, and S. T. Ferreira, "How does brain insulin resistance develop in Alzheimer's disease?," *Alzheimers. Dement.*, vol. 10, no. 1 Suppl, pp. S26-32, Feb. 2014.

[46] G. Bedse, F. Di Domenico, G. Serviddio, and T. Cassano, "Aberrant insulin signaling in Alzheimer's disease: current knowledge.," *Front. Neurosci.*, vol. 9, no. MAY, p. 204, 2015.

[47] M. G. Moreno-Treviño, J. Castillo-López, and I. Meester, "Moving away from amyloid Beta to move on in Alzheimer research.," *Front. Aging Neurosci.*, vol. 7, no. 2, p. 2, Jan. 2015.

[48] K. Herrup, "The case for rejecting the amyloid cascade hypothesis.," *Nat. Neurosci.*, vol. 18, no. 6, pp. 794–9, Jun. 2015.

[49] M. D. Garrett, "Complexity Theory and Alzheimer's disease: A Call to Action," *Psychology Today*, 2016.

[50] D. Calcoen, L. Elias, and X. Yu, "What does it take to produce a breakthrough drug?," *Nat. Rev. Drug Discov.*, vol. 14, no. 3, pp. 161–162, 2015.

[51] J. Cummings, "What Can Be Inferred from the Interruption of the Semagacestat Trial for Treatment of Alzheimer's Disease?," *Biol. Psychiatry*, vol. 68, no. 10, pp. 876–878, Nov. 2010.

[52] J. Cummings, P. S. Aisen, B. DuBois, L. Frölich, C. R. Jack, R. W. Jones, J. C. Morris, J. Raskin, S. A. Dowsett, and P. Scheltens, "Drug development in Alzheimer's disease: the path to 2025," *Alzheimers. Res. Ther.*, vol. 8, no. 1, p. 39, Dec. 2016.

[53] Y. Joanette, E. C. Hirsch, and M. Goldman, "The Global Fight Against Dementia," *Sci. Transl. Med.*, vol. 6, no. 267, p. 267ed22-267ed22, Dec. 2014.

[54] A. D. International, "The Global Impact of Dementia 2013 – 2050 Policy Brief for Heads of

Government," 2013.

[55] K. Mullane and M. Williams, "Alzheimer's therapeutics: Continued clinical failures question the validity of the amyloid hypothesis - But what lies beyond?," *Biochem. Pharmacol.*, vol. 85, no. 3, pp. 289–305, 2013.

[56] G. R. Langley, "Considering a new paradigm for Alzheimer's disease research," *Drug Discov. Today*, vol. 19, no. 8, pp. 1114–1124, 2014.

[57] M. Citron, "Alzheimer's disease: strategies for disease modification.," *Nat. Rev. Drug Discov.*, vol. 9, no. 5, pp. 387–98, May 2010.

[58] W. V. Graham, A. Bonito-Oliva, and T. P. Sakmar, "Update on Alzheimer's Disease Therapy and Prevention Strategies.," *Annu. Rev. Med.*, vol. 68, no. 1, pp. 413–430, Jan. 2017.

[59] J. Carroll, "Another Alzheimer's drug flops in pivotal clinical trial," *Science (80-.).*, Feb. 2017.

[60] D. J. Selkoe and J. Hardy, "The amyloid hypothesis of Alzheimer's disease at 25 years.," *EMBO Mol. Med.*, vol. 8, no. e201606210, pp. 1–14, 2016.

[61] E. Karran and J. Hardy, "A critique of the drug discovery and phase 3 clinical programs targeting the amyloid hypothesis for Alzheimer disease.," *Ann. Neurol.*, vol. 76, no. 2, pp. 185–205, Aug. 2014.

[62] B. De Strooper, "Lessons from a failed γ-secretase Alzheimer trial.," *Cell*, vol. 159, no. 4, pp. 721–6, Nov. 2014.

[63] R. J. Castellani and G. Perry, "Pathogenesis and disease-modifying therapy in Alzheimer's disease: the flat line of progress.," *Arch. Med. Res.*, vol. 43, no. 8, pp. 694–8, Nov. 2012.

[64] D. J. Selkoe, "Resolving controversies on the path to Alzheimer's therapeutics," *Nat. Med.*, vol. 17, no. 9, pp. 1060–1065, Sep. 2011.

[65] D. Perry, R. Sperling, R. Katz, D. Berry, D. Dilts, D. Hanna, S. Salloway, J. Q. Trojanowski, C. Bountra, M. Krams, J. Luthman, S. Potkin, V. Gribkoff, R. Temple, Y. Wang, M. C. Carrillo, D. Stephenson, H. Snyder, E. Liu, T. Ware, J. McKew, F. O. Fields, L. J. Bain, and C. Bens, "Building

a roadmap for developing combination therapies for Alzheimer's disease," *Expert Rev. Neurother.*, vol. 15, no. 3, pp. 327–333, Mar. 2015.

[66] A. Mudher and S. Lovestone, "Alzheimer's disease-do tauists and baptists finally shake hands?," *Trends Neurosci.*, vol. 25, no. 1, pp. 22–6, Jan. 2002.

[67] J. R. Harrison and M. J. Owen, "Alzheimer's disease: the amyloid hypothesis on trial," *Br. J. Psychiatry*, vol. 208, no. 1, pp. 1–3, Jan. 2016.

[68] R. E. Becker and N. H. Greig, "Increasing the success rate for Alzheimer's disease drug discovery and development.," *Expert Opin. Drug Discov.*, vol. 7, no. 4, pp. 367–70, Apr. 2012.

[69] J. Cummings, T. Morstorf, and G. Lee, "Alzheimer's drug-development pipeline: 2016," *Alzheimer's Dement. Transl. Res. Clin. Interv.*, vol. 2, no. 4, pp. 222–232, 2016.

[70] H. Hampel, S. E. O'Bryant, S. Durrleman, E. Younesi, K. Rojkova, V. Escott-Price, J.-C. Corvol, K. Broich, B. Dubois, and S. Lista, "A Precision Medicine Initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling," *Climacteric*, vol. 0, no. 0, pp. 1–12, 2017.

[71] T. Schultz, "Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle," *Bull. Am. Soc. Inf. Sci. Technol.*, vol. 39, no. 5, pp. 34–40, Jun. 2013.

[72] C. M. Machado, D. Rebholz-Schuhmann, A. T. Freitas, and F. M. Couto, "The semantic web in translational medicine: Current applications and future directions," *Brief. Bioinform.*, vol. 16, no. 1, pp. 89–103, 2013.

[73] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery.," *Nat. Rev. Genet.*, vol. 13, no. 8, pp. 523–36, Jul. 2012.

[74] J. M. J. Derry, L. M. Mangravite, C. Suver, M. D. Furia, D. Henderson, X. Schildwachter, B. Bot, J. Izant, S. K. Sieberts, M. R. Kellen, and S. H. Friend, "Developing predictive molecular maps of human disease through community-based modeling," *Nat. Genet.*, vol. 44, no. 2, pp. 127–130, Jan. 2012.

[75] D. Galasko, T. E. Golde, C. Altar, D. Amakye, D. Bounos, J. Bloom, G. Clack, R. Dean, V. Devanarayan, D. Fu, S. Furlong, L. Hinman, C. Girman, C. Lathia, L. Lesko, S. Madani, J. Mayne, J. Meyer, D. Raunig, P. Sager, S. Williams, P. Wong, K. Zerba, A. Koyama, O. Okerere, T. Yang, D. Blacker, D. Selkoe, F. Grodstein, H. Soares, W. Potter, E. Pickering, M. Kuhn, F. Immermann, D. Shera, M. Ferm, R. Dean, A. Simon, F. Swenson, J. Siuciak, J. Kaplow, M. Thambisetty, P. Zagouras, W. Koroshetz, H. Wan, J. Trojanowski, L. Shaw, J. Doecke, S. Laws, N. Faux, W. Wilson, S. Burnham, C. Lam, A. Mondal, J. Bedo, A. Bush, B. Brown, K. De Ruyck, K. Ellis, C. Fowler, V. Gupta, R. Head, S. Macaulay, K. Pertile, C. Rowe, A. Rembach, M. Rodrigues, R. Rumble, C. Szoeke, K. Taddei, T. Taddei, B. Trounson, D. Ames, C. Masters, R. Martins, W. Hu, D. Holtzman, A. Fagan, L. Shaw, R. Perrin, S. Arnold, M. Grossman, C. Xiong, R. Craig-Schapiro, C. Clark, E. Pickering, M. Kuhn, Y. Chen, V. Van Deerlin, L. McCluskey, L. Elman, J. Karlawish, A. Chen-Plotkin, H. Hurtig, A. Siderowf, F. Swenson, V. Lee, J. Morris, J. Trojanowski, H. Soares, P. Mehta, T. Pirttila, B. Patrick, M. Barshatzky, S. Mehta, A. Fagan, D. Head, A. Shah, D. Marcus, M. Mintun, J. Morris, D. Holtzman, M. van Oijen, A. Hofman, H. Soares, P. Koudstaal, M. Breteler, N. Graff-Radford, J. Crook, J. Lucas, B. Boeve, D. Knopman, R. Ivnik, G. Smith, L. Younkin, R. Petersen, S. Younkin, J. Lambert, S. Schraen-Maschke, F. Richard, N. Fievet, O. Rouaud, C. Berr, J. Dartigues, C. Tzourio, A. Alpérovitch, L. Buée, P. Amouyel, E. Schrijvers, P. Koudstaal, A. Hofman, M. Breteler, L. Butterfield, D. Potter, J. Kirkwood, J. Yi, D. Craft, C. Gelfand, S. Lista, F. Faltraco, H. Hampel, M. Rifai, M. Gillette, S. Carr, S. van der Burg, M. Kalos, C. Gouttefangeas, S. Janetzki, C. Ottensmeier, M. Welters, P. Romero, C. Britten, and A. Hoos, "Biomarkers for Alzheimer's disease in plasma, serum and blood - conceptual and practical problems," *Alzheimers. Res. Ther.*, vol. 5, no. 2, p. 10, 2013.

[76] F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?," *Nat. Rev. Drug Discov.*, vol. 10, no. 9, p. 712, 2011.

[77] N. Ashish, P. Bhatt, and A. W. Toga, "Global Data Sharing in Alzheimer Disease Research," *Alzheimer Dis. Assoc. Disord.*, vol. 30, no. 2, pp. 160–168, 2016.

[78] P. S. Pillai, T.-Y. Leong, and Alzheimer's Disease Neuroimaging Initiative, "Fusing Heterogeneous Data for Alzheimer's Disease Classification.," *Stud. Health Technol. Inform.*, vol. 216, pp. 731–5, 2015.

[79] M. Samwald, A. Jentzsch, C. Bouton, C. S. Kallesøe, E. Willighagen, J. Hajagos, M. Scott Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, and S. Stephens, "Linked Open drug data for pharmaceutical research and development," *J. Cheminform.*, vol. 3, no. 5, p. 19, 2011.

[80] K. D. Fowler, J. M. Funt, M. N. Artyomov, B. Zeskind, S. E. Kolitz, and F. Towfic, "Leveraging existing data sets to generate new insights into Alzheimer's disease biology in specific patient subsets.," *Sci. Rep.*, vol. 5, no. October 2014, p. 14324, Sep. 2015.

[81] J. Y. Chen, C. Shen, and A. Y. Sivachenko, "Mining Alzheimer disease relevant proteins from integrated protein interactome data.," in *Pacific Symposium on Biocomputing*, 2006, pp. 367–78.

[82] M. Soler-López, A. Zanzoni, R. Lluís, U. Stelzl, and P. Aloy, "Interactome mapping suggests new mechanistic details underlying Alzheimer's disease.," *Genome Res.*, vol. 21, no. 3, pp. 364–76, Mar. 2011.

[83] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 42, pp. 15148–53, Oct. 2004.

[84] B. Liu, T. Jiang, S. Ma, H. Zhao, J. Li, X. Jiang, and J. Zhang, "Exploring candidate genes for human brain diseases from a brain-specific gene network," *Biochem. Biophys. Res. Commun.*, vol. 349, no. 4, pp. 1308–1314, Nov. 2006.

[85] L. Caberlotto and T.-P. Nguyen, "A systems biology investigation of neurodegenerative dementia reveals a pivotal role of autophagy.," *BMC Syst. Biol.*, vol. 8, no. 1, p. 65, 2014.

[86] L. Caberlotto, M. Lauria, T. P. Nguyen, and M. Scotti, "The central role of AMP-kinase and energy homeostasis impairment in Alzheimer's disease: A multifactor network analysis," *PLoS One*, vol. 8, no. 11, 2013.

[87] M. Schenone, V. Dančík, B. K. Wagner, and P. a Clemons, "Target identification and mechanism of action in chemical biology and drug discovery.," *Nat. Chem. Biol.*, vol. 9, no. 4, pp. 232–40, Apr. 2013.

[88] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and

genomic medicine," *J. Biomed. Inform.*, vol. 40, no. 1, pp. 5–16, Feb. 2007.

[89] E. L. Willighagen, J. Alvarsson, A. Andersson, M. Eklund, S. Lampa, M. Lapins, O. Spjuth, and J. E. Wikberg, "Linking the Resource Description Framework to cheminformatics and proteochemometrics.," *J. Biomed. Semantics*, vol. 2 Suppl 1, no. Suppl 1, p. S6, 2011.

[90] A. R. Kinjo, H. Suzuki, R. Yamashita, Y. Ikegawa, T. Kudou, R. Igarashi, Y. Kengaku, H. Cho, D. M. Standley, A. Nakagawa, and H. Nakamura, "Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 453–460, 2012.

[91] E. Neumann, "Finding the critical path: Applying the semantic web to drug discovery and development," *Drug Discov. World*, vol. 6, no. 4, pp. 25–33, 2005.

[92] B. Glimm and H. Stuckenschmidt, "15 Years of Semantic Web: An Incomplete Survey," *KI - Künstliche Intelligenz*, vol. 30, no. 2, pp. 117–130, Jun. 2016.

[93] R. Klapsing, G. Neumann, and W. Conen, "Semantics in Web engineering: Applying the resource description framework," *IEEE Multimed.*, vol. 8, no. 2, pp. 62–68, 2001.

[94] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Sci. Am.*, vol. 284, no. 5, pp. 34–43, May 2001.

[95] E. Miller, "An Introduction to the Resource Description Framework," *Bull. Am. Soc. Inf. Sci. Technol.*, vol. 25, no. 1, pp. 15–19, Jan. 2005.

[96] M. E. Holford, E. Khurana, K.-H. Cheung, and M. Gerstein, "Using semantic web rules to reason on an ontology of pseudogenes.," *Bioinformatics*, vol. 26, no. 12, pp. i71-8, Jun. 2010.

[97] V. Haarslev and R. Möller, "Consistency Testing: The RACE Experience," in *Automated Reasoning with Analytic Tableaux and Related Methods: International Conference, TABLEAUX 2000, St Andrews, Scotland, UK, July 3-7, 2000 Proceedings*, R. Dyckhoff, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 57–61.

[98] D. Tsarkov and I. Horrocks, "FaCT++ Description Logic Reasoner: System Description," in *Lecture*

Notes in Computer Science, Springer, 2006, pp. 292–297.

[99] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical OWL-DL reasoner," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 5, no. 2, pp. 51–53, Jun. 2007.

[100] R. B. Mishra and S. Kumar, "Semantic web reasoners and languages," *Artif. Intell. Rev.*, vol. 35, no. 4, pp. 339–368, Apr. 2011.

[101] M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. . 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, Mar. 2016.

[102] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Brief. Bioinform.*, vol. 16, no. 6, pp. 1069–1080, Nov. 2015.

[103] "Identifiers.org." [Online]. Available: http://identifiers.org.

[104] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, and A. M. Jenkinson, "The EBI RDF platform: Linked open data for the life sciences," *Bioinformatics*, vol. 30, no. 9, pp. 1338–1339, 2014.

[105] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," *J. Biomed. Inform.*, vol. 41, no. 5, pp. 706–716, 2008.

[106] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, K. Forsberg, Y. Gao, V. Kashyap, J. Kinoshita, J. Luciano, M. S. Marshall, C. Ogbuji, J. Rees, S. Stephens, G. T.

Wong, E. Wu, D. Zaccagnini, T. Hongsermeier, E. Neumann, I. Herman, and K.-H. Cheung, "Advancing translational research with the Semantic Web.," *BMC Bioinformatics*, vol. 8 Suppl 3, p. S2, 2007.

[107] P. E. Bourne, V. Bonazzi, M. Dunn, E. D. Green, M. Guyer, G. Komatsoulis, J. Larkin, and B. Russell, "The NIH Big Data to Knowledge (BD2K) initiative," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 6, pp. 1114–1114, Nov. 2015.

[108] D. Gardner and G. M. Shepherd, "A Gateway to the Future of Neuroinformatics," *Neuroinformatics*, vol. 2, no. 3, pp. 271–274, 2004.

[109] P. R. Cohen, "DARPA's Big Mechanism program.," *Phys. Biol.*, vol. 12, no. 4, p. 45008, 2015.

[110] S. F. Muldoon, *Encyclopedia of Computational Neuroscience*, no. Massaro 1975. 2013.

[111] Z. Yi, W. Dongsheng, Z. Tielin, and X. Bo, "Linked Neuron Data (LND): A Platform for Integrating and Semantically Linking Neuroscience Data and Knowledge," *Front. Neuroinform.*, vol. 8, 2014.

[112] A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons, "Open PHACTS: Semantic interoperability for drug discovery," *Drug Discov. Today*, vol. 17, no. 21–22, pp. 1188–1198, 2012.

[113] H. Okano and T. Yamamori, "How can brain mapping initiatives cooperate to achieve the same goal?," *Nat. Rev. Neurosci.*, vol. 17, no. 12, pp. 733–734, 2016.

[114] F. Å. Nielsen, "Brede tools and federating online neuroinformatics databases," *Neuroinformatics*, vol. 12, no. 1, pp. 27–37, 2014.

[115] T. Clark and J. Kinoshita, "Alzforum and SWAN: The present and future of scientific web communities," *Brief. Bioinform.*, vol. 8, no. 3, pp. 163–171, 2007.

[116] H. Y. K. Lam, L. Marenco, T. Clark, Y. Gao, J. Kinoshita, G. Shepherd, P. Miller, E. Wu, G. T. Wong, N. Liu, C. Crasto, T. Morse, S. Stephens, and K.-H. Cheung, "AlzPharm: integration of neurodegeneration data using RDF.," *BMC Bioinformatics*, vol. 8 Suppl 3, p. S4, 2007.

[117] D. J. Rigden, X. M. Fernández-Suárez, and M. Y. Galperin, "The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection.," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1-6, Jan. 2016.

[118] R. P. Horgan and L. C. Kenny, "'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics," *Obstet. Gynaecol.*, vol. 13, no. 3, pp. 189–195, Jul. 2011.

[119] J. S. Buguliskis, "The Epigenetic Insights of RNA-Seq," *Clin. Omi.*, vol. 3, no. 5, pp. 10–13, May 2016.

[120] D. M. Pedrotty, M. P. Morley, and T. P. Cappola, "Transcriptomic Biomarkers of Cardiovascular Disease," *Prog. Cardiovasc. Dis.*, vol. 55, no. 1, pp. 64–69, 2012.

[121] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics.," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[122] W. a Valdivia-granda and C. Dwan, *Chapter 6 MICROARRAY DATA MANAGEMENT*. 2006.

[123] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, 2002.

[124] "ArrayExpress Database." [Online]. Available: http://www.ebi.ac.uk/arrayexpress.

[125] R. Hitzemann, D. Bottomly, P. Darakjian, N. Walter, O. Iancu, R. Searles, B. Wilmot, and S. McWeeney, "Genes, behavior and next-generation RNA sequencing," *Genes, Brain Behav.*, vol. 12, no. 1, pp. 1–12, Feb. 2013.

[126] J. A. Thompson, J. Tan, and C. S. Greene, "Cross-platform normalization of microarray and RNA-seq data for machine learning applications.," *PeerJ*, vol. 4, p. e1621, 2016.

[127] K. A. Janes and M. B. Yaffe, "Data-driven modelling of signal-transduction networks," *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 11, pp. 820–828, Nov. 2006.

[128] P. P. Panigrahi and T. R. Singh, "Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: Revealed association

with aging and other diseases," *J. Theor. Biol.*, vol. 334, pp. 109–121, 2013.

[129] D. Liang, G. Han, X. Feng, J. Sun, Y. Duan, and H. Lei, "Concerted perturbation observed in a hub network in Alzheimer's disease," *PLoS One*, vol. 7, no. 7, 2012.

[130] P. H. Sudmant, M. S. Alexis, and C. B. Burge, "Meta-analysis of RNA-seq expression data across species, tissues and studies," *Genome Biol.*, vol. 16, no. 1, p. 287, Dec. 2015.

[131] C. Walsh, P. Hu, J. Batt, and C. Santos, "Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery," *Microarrays*, vol. 4, no. 3, pp. 389–406, 2015.

[132] G. C. Tseng, D. Ghosh, and E. Feingold, "Comprehensive literature review and statistical considerations for microarray meta-analysis," *Nucleic Acids Res.*, vol. 40, no. 9, pp. 3785–3799, 2012.

[133] J. S. Hamid, P. Hu, N. M. Roslin, V. Ling, C. M. T. Greenwood, and J. Beyene, "Data Integration in Genetics and Genomics: Methods and Challenges," *Hum. Genomics Proteomics*, vol. 2009, pp. 1–13, 2009.

[134] J. Taminau, C. Lazar, S. Meganck, and A. Nowé, "Comparison of Merging and Meta-Analysis as Alternative Approaches for Integrative Gene Expression Analysis," *ISRN Bioinforma.*, vol. 2014, pp. 1–7, 2014.

[135] M. S. Pepe and Z. Feng, "Improving Biomarker Identification with Better Designs and Reporting," *Clin. Chem.*, vol. 57, no. 8, pp. 1093–1095, Aug. 2011.

[136] P. A. Konstantinopoulos, S. A. Cannistra, H. Fountzilas, A. Culhane, K. Pillay, B. Rueda, D. Cramer, M. Seiden, M. Birrer, G. Coukos, L. Zhang, J. Quackenbush, and D. Spentzos, "Integrated Analysis of Multiple Microarray Datasets Identifies a Reproducible Survival Predictor in Ovarian Cancer," *PLoS One*, vol. 6, no. 3, p. e18202, Mar. 2011.

[137] L. Xu, A. Tan, R. L. Winslow, and D. Geman, "Merging microarray data from separate breast cancer studies provides a robust prognostic test," *BMC Bioinformatics*, vol. 9, no. 1, p. 125, 2008.

[138] C.-C. Liu, J. Hu, M. Kalakrishnan, H. Huang, and X. Zhou, "Integrative disease classification based on cross-platform microarray data," *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S25, 2009.

[139] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, "Key issues in conducting a meta-analysis of gene expression microarray datasets," *PLoS Med.*, vol. 5, no. 9, pp. 1320–1332, 2008.

[140] J. Rudy and F. Valafar, "Empirical comparison of cross-platform normalization methods for gene expression data," *BMC Bioinformatics*, vol. 12, no. 1, p. 467, 2011.

[141] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan, "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 25, pp. 9309–9314, 2004.

[142] E. Zintzaras and J. P. A. Ioannidis, "Meta-analysis for ranked discovery datasets: Theoretical framework and empirical demonstration for microarrays," *Comput. Biol. Chem.*, vol. 32, no. 1, pp. 39–47, Feb. 2008.

[143] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation.," *Bioinformatics*, vol. 19 Suppl 1, pp. i84-90, 2003.

[144] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 1–6, 2012.

[145] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[146] J. L. Wilson, M. T. Hemann, E. Fraenkel, and D. A. Lauffenburger, "Integrated network analyses for functional genomic studies in cancer," *Semin. Cancer Biol.*, vol. 23, no. 4, pp. 213–218, 2013.

[147] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks.," *Nature*, vol. 411, no. 6833, pp. 41–2, May 2001.

[148] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human

interactome," *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, Sep. 2006.

[149] E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka, "Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality.," *PLoS Comput. Biol.*, vol. 4, no. 8, p. e1000140, Aug. 2008.

[150] A. Aytes, A. Mitrofanova, C. Lefebvre, M. J. Alvarez, M. Castillo-Martin, T. Zheng, J. A. Eastham, A. Gopalan, K. J. Pienta, M. M. Shen, A. Califano, and C. Abate-Shen, "Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy," *Cancer Cell*, vol. 25, no. 5, pp. 638–651, May 2014.

[151] C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. Dalla-Favera, and A. Califano, "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers.," *Mol. Syst. Biol.*, vol. 6, p. 377, Jun. 2010.

[152] Y. Li and S. A. Jackson, "Gene Network Reconstruction by Integration of Prior Biological Knowledge.," *G3 (Bethesda).*, vol. 5, no. 6, pp. 1075–9, Mar. 2015.

[153] F. M. Giorgi, C. Del Fabbro, and F. Licausi, "Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana.," *Bioinformatics*, vol. 29, no. 6, pp. 717–24, Mar. 2013.

[154] S. Wang, Y. Yin, Q. Ma, X. Tang, D. Hao, and Y. Xu, "Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis," *BMC Plant Biol.*, vol. 12, no. 1, p. 138, 2012.

[155] J.-H. Chiang and H.-C. Yu, "MeKE: discovering the functions of gene products from biomedical literature via sentence alignment.," *Bioinformatics*, vol. 19, no. 11, pp. 1417–22, Jul. 2003.

[156] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio--the analysis and navigation of molecular networks.," *Bioinformatics*, vol. 19, no. 16, pp. 2155–7, Nov. 2003.

[157] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner, "MatInspector and beyond: promoter analysis based on transcription factor binding

sites.," *Bioinformatics*, vol. 21, no. 13, pp. 2933–42, Jul. 2005.

[158] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V Kel-Margoulis, and E. Wingender, "MATCH: A tool for searching transcription factor binding sites in DNA sequences.," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3576–9, Jul. 2003.

[159] T. Slater, "Recent advances in modeling languages for pathway maps and computable biological networks.," *Drug Discov. Today*, vol. 19, no. 2, pp. 193–8, Feb. 2014.

[160] W. P. Lee and W. S. Tzou, "Computational methods for discovering gene networks from expression data," *Brief. Bioinform.*, vol. 10, no. 4, pp. 408–423, 2009.

[161] F. Markowetz and R. Spang, "Inferring cellular networks – a review," *BMC Bioinformatics*, vol. 8, no. Suppl 6, p. S5, 2007.

[162] J.-P. Vert, "Reconstruction of Biological Networks by Supervised Machine Learning Approaches," in *Elements of Computational Systems Biology*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010, pp. 163–188.

[163] A. de la Fuente, "What are Gene Regulatory Networks?," in *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*, IGI Global, pp. 1–27.

[164] P. Khosravi, V. H. Gazestani, L. Pirhaji, B. Law, M. Sadeghi, B. Goliaei, and G. D. Bader, "Inferring interaction type in gene regulatory networks using co-expression data," *Algorithms Mol. Biol.*, vol. 10, no. 1, p. 23, Dec. 2015.

[165] K. J. Woolcock, R. Stunnenberg, D. Gaidatzis, H.-R. Hotz, S. Emmerth, P. Barraud, and M. Bühler, "RNAi keeps Atf1-bound stress response genes in check at nuclear pores.," *Genes Dev.*, vol. 26, no. 7, pp. 683–92, Apr. 2012.

[166] A. Tallam, T. M. Perumal, P. M. Antony, C. Jäger, J. V. Fritz, L. Vallar, R. Balling, A. del Sol, and A. Michelucci, "Gene Regulatory Network Inference of Immunoresponsive Gene 1 (IRG1) Identifies Interferon Regulatory Factor 1 (IRF1) as Its Transcriptional Regulator in Mammalian Macrophages," *PLoS One*, vol. 11, no. 2, p. e0149050, Feb. 2016.

[167] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 22, pp. 12182–6, Oct. 2000.

[168] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, no. 8, pp. 707–726, Aug. 2000.

[169] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.," *Pac. Symp. Biocomput.*, pp. 418–29, 2000.

[170] J. D. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao, "Comparing statistical methods for constructing large scale gene networks.," *PLoS One*, vol. 7, no. 1, p. e29348, 2012.

[171] A. V Werhli, M. Grzegorczyk, and D. Husmeier, "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks.," *Bioinformatics*, vol. 22, no. 20, pp. 2523–31, Oct. 2006.

[172] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis.," *BMC Bioinformatics*, vol. 9, p. 559, 2008.

[173] J. Peng, P. Wang, N. Zhou, and J. Zhu, "Partial Correlation Estimation by Joint Sparse Regression Models.," *J. Am. Stat. Assoc.*, vol. 104, no. 486, pp. 735–746, Jun. 2009.

[174] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.," *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, 2006.

[175] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks.," *EURASIP J. Bioinform. Syst. Biol.*, vol. 2007, no. 1, p. 79879, 2007.

[176] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.," *PLoS Biol.*, vol. 5, no. 1, p. e8, 2007.

[177] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC Bioinformatics*, vol. 13, no. 1, p. 328, 2012.

[178] N. A. Kiani, H. Zenil, J. Olczak, and J. Tegnér, "Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks," *Semin. Cell Dev. Biol.*, vol. 51, pp. 44–52, Mar. 2016.

[179] G. Altay and F. Emmert-Streib, "Structural influence of gene networks on their inference: analysis of C3NET.," *Biol. Direct*, vol. 6, no. 1, p. 31, 2011.

[180] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, Dec. 2004.

[181] B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. D'Alché-Buc, "Gene networks inference using dynamic Bayesian networks.," *Bioinformatics*, vol. 19 Suppl 2, p. ii138-48, Oct. 2003.

[182] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods," *PLoS One*, vol. 5, no. 9, p. e12776, 2010.

[183] J. Sławek and T. Arodź, "ENNET: inferring large gene regulatory networks from expression data using gradient boosting.," *BMC Syst. Biol.*, vol. 7, p. 106, Oct. 2013.

[184] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.," *BMC Syst. Biol.*, vol. 6, p. 145, Nov. 2012.

[185] S. Guo, Q. Jiang, L. Chen, and D. Guo, "Gene regulatory network inference using PLS-based methods," *BMC Bioinformatics*, vol. 17, no. 1, p. 545, Dec. 2016.

[186] R. de Matos Simoes and F. Emmert-Streib, "Bagging statistical network inference from large-scale gene expression data," *PLoS One*, vol. 7, no. 3, 2012.

[187] M. D. M. LEISERSON, F. VANDIN, H. T. Wu, and B. J. Raphael, "Heat diffusion based genetic network analysis," 2016.

[188] F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for Detecting Significantly Mutated Pathways in Cancer," *J. Comput. Biol.*, vol. 18, no. 3, pp. 507–522, Mar. 2011.

[189] M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nat. Genet.*, vol. 47, no. 2, pp. 106–114, Dec. 2014.

[190] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks.," *Front. cell Dev. Biol.*, vol. 2, no. August, p. 38, 2014.

[191] R. D. M. Simoes, M. Dehmer, and F. Emmert-Streib, "B-cell lymphoma gene regulatory networks: Biological consistency among inference methods," *Front. Genet.*, vol. 4, no. DEC, pp. 1–14, 2013.

[192] H. Rhinn, R. Fujita, L. Qiang, R. Cheng, J. H. Lee, and A. Abeliovich, "Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease.," *Nature*, vol. 500, no. 7460, pp. 45–50, Aug. 2013.

[193] V. Swarup and D. H. Geschwind, "Alzheimer's disease: From big data to mechanism," *Nature*, vol. 500, no. 7460, pp. 34–35, Jul. 2013.

[194] B. Zhang, C. Gaiteri, L. G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M. E. MacDonald, J. R. Lamb, D. A. Bennett, C. Molony, D. J. Stone, V. Gudnason, A. J. Myers, E. E. Schadt, H. Neumann, J. Zhu, and V. Emilsson, "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[195] P. Forabosco, A. Ramasamy, D. Trabzuni, R. Walker, C. Smith, J. Bras, A. P. Levine, J. Hardy, J. M. Pocock, R. Guerreiro, M. E. Weale, and M. Ryten, "Insights into TREM2 biology by network analysis of human brain gene expression data.," *Neurobiol. Aging*, vol. 34, no. 12, pp. 2699–714, Dec. 2013.

[196] J. a Miller, S. Horvath, and D. H. Geschwind, "Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 28, pp. 12698–703, Jul. 2010.

[197] M. Ray, J. Ruan, and W. Zhang, "Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases," *Genome Biol.*, vol. 9, no. 10, p. R148, 2008.

[198] D. Zou, L. Ma, J. Yu, and Z. Zhang, "Biological databases for human research.," *Genomics. Proteomics Bioinformatics*, vol. 13, no. 1, pp. 55–63, Feb. 2015.

[199] A. Kumari, S. Kanchan, R. P. Sinha, and M. Kesheri, "Applications of Bio-molecular Databases in Bioinformatics," in *Medical Imaging in Clinical Practice*, no. June, InTech, 2016, pp. 329–351.

[200] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. P. Pereira, E. Pilicheva, J. Rung, A. Sharma, Y. A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson, and U. Sarkans, "ArrayExpress update-trends in database growth and links to data analysis tools," *Nucleic Acids Res.*, vol. 41, no. D1, pp. 987–990, 2013.

[201] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D68–D73, Jan. 2014.

[202] R. Petryszak, T. Burdett, B. Fiorelli, N. a. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, N. Kryvych, J. McMurry, J. C. Marioni, J. Malone, K. Megy, G. Rustici, A. Y. Tang, J. Taubert, E. Williams, O. Mannion, H. E. Parkinson, and A. Brazma, "Expression Atlas update - A database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 926–932, 2014.

[203] C. Schaefer, A. Meier, B. Rost, and Y. Bromberg, "Snpdbe: Constructing an nsSnp functional impacts database," *Bioinformatics*, vol. 28, no. 4, pp. 601–602, 2012.

[204] Z. Zhang, J. Sang, L. Ma, G. Wu, H. Wu, D. Huang, D. Zou, S. Liu, A. Li, L. Hao, M. Tian, C. Xu, X. Wang, J. Wu, J. Xiao, L. Dai, L.-L. Chen, S. Hu, and J. Yu, "RiceWiki: a wiki-based database for community curation of rice genes," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1222–D1228, Jan. 2014.

[205] N. K. Gundla and Z. Chen, "Creating NoSQL Biological Databases with Ontologies for Query Relaxation," *Procedia Comput. Sci.*, vol. 91, pp. 460–469, 2016.

[206] C. T. Have, L. J. Jensen, and J. Wren, "Are graph databases ready for bioinformatics?," *Bioinformatics*, vol. 29, no. 24, pp. 3107–3108, 2013.

[207] M. D. Brazas, D. S. Yim, J. T. Yamada, and B. F. F. Ouellette, "The 2011 bioinformatics links directory update: more resources, tools and databases and features to empower the bioinformatics community," *Nucleic Acids Res.*, vol. 39, no. suppl, pp. W3–W7, Jul. 2011.

[208] V. J. Henry, A. E. Bandrowski, A.-S. Pepin, B. J. Gonzalez, and A. Desfeux, "OMICtools: an informative directory for multi-omic data analysis," *Database*, vol. 2014, p. bau069-bau069, Jul. 2014.

[209] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information.," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D7-19, Jan. 2016.

[210] SIB Swiss Institute of Bioinformatics Members, "The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases.," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D27-37, Jan. 2016.

[211] M. Helmy, A. Crits-Christoph, and G. D. Bader, "Ten Simple Rules for Developing Public Biological Databases," *PLoS Comput. Biol.*, vol. 12, no. 11, pp. 1–8, 2016.

[212] J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort, "Repeatability of published microarray gene expression analyses," *Nat. Genet.*, vol. 41, no. 2, pp. 149–155, Feb. 2009.

[213] J. Rung and A. Brazma, "Reuse of public genome-wide gene expression data.," *Nat. Rev. Genet.*, vol. 14, no. 2, pp. 89–99, 2013.

[214] J. Fluck and M. Hofmann-Apitius, "Text mining for systems biology," *Drug Discov. Today*, vol. 19, no. 2, pp. 140–144, 2014.

[215] M. A. Hearst, "Untangling text data mining," in *Proceedings of the 37th annual meeting of the*

Association for Computational Linguistics on Computational Linguistics -, 1999, pp. 3–10.

[216] S. M. Leach, H. Tipney, W. Feng, W. A. Baumgartner, P. Kasliwal, R. P. Schuyler, T. Williams, R. A. Spritz, and L. Hunter, "Biomedical Discovery Acceleration, with Applications to Craniofacial Development," *PLoS Comput. Biol.*, vol. 5, no. 3, p. e1000215, Mar. 2009.

[217] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen, "Biomedical text mining and its applications in cancer research," *J. Biomed. Inform.*, vol. 46, no. 2, pp. 200–211, 2013.

[218] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf, "Text-mining solutions for biomedical research: enabling integrative biology.," *Nat. Rev. Genet.*, vol. 13, no. 12, pp. 829–39, 2012.

[219] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature.," *Database (Oxford).*, vol. 2011, p. baq036, 2011.

[220] U. Leser and J. Hakenberg, "What makes a gene name? Named entity recognition in the biomedical literature.," *Brief. Bioinform.*, vol. 6, no. 4, pp. 357–69, Dec. 2005.

[221] S. Ananiadou, D. B. Kell, and J. Tsujii, "Text mining and its potential applications in systems biology.," *Trends Biotechnol.*, vol. 24, no. 12, pp. 571–579, 2006.

[222] S. Bagewadi, T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger, "Detecting miRNA Mentions and Relations in Biomedical Literature.," *F1000Research*, vol. 3, no. 0, p. 205, 2014.

[223] T. Bobić, R. Klinger, P. Thomas, and M. Hofmann-apitius, "Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions," *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguist.*, pp. 35–43, 2012.

[224] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop on BioNLP Shared Task - BioNLP '09*, 2009, no. June, p. 1.

[225] Y. WANG, J.-D. KIM, R. SÆTRE, S. PYYSALO, T. OHTA, and J. TSUJII, "IMPROVING THE INTER-CORPORA COMPATIBILITY FOR PROTEIN ANNOTATIONS," *J. Bioinform. Comput.*

Biol., vol. 8, no. 5, pp. 901–916, Oct. 2010.

[226] Y. Zhang, C. Tao, G. Jiang, A. a Nair, J. Su, C. G. Chute, and H. Liu, "Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network," *J. Biomed. Semantics*, vol. 5, no. 1, p. 33, 2014.

[227] A. Malhotra, E. Younesi, S. Bagewadi, and M. Hofmann-Apitius, "Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease," *Genome Med.*, vol. 6, no. 12, 2014.

[228] E. Seymour, R. Damle, A. Sette, and B. Peters, "Cost sensitive hierarchical document classification to triage PubMed abstracts for manual curation," *BMC Bioinformatics*, vol. 12, no. 1, p. 482, 2011.

[229] A. P. Davis, T. C. Wiegers, C. G. Murphy, and C. J. Mattingly, "The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database," *Database*, vol. 2011, pp. 1–12, 2011.

[230] M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, J. Vandenhaute, M. Galli, J. Yazaki, D. E. Hill, J. R. Ecker, F. P. Roth, and M. Vidal, "Literature-curated protein interaction datasets," *Nat. Methods*, vol. 6, no. 1, pp. 39–46, Jan. 2009.

[231] D. Campos, J. Lourenço, S. Matos, and J. L. Oliveira, "Egas: a collaborative and interactive document curation platform.," *Database (Oxford).*, vol. 2014, pp. 1–12, 2014.

[232] R. Rak, A. Rowley, W. Black, and S. Ananiadou, "Argo: An integrative, interactive, text mining-based workbench supporting curation," *Database*, vol. 2012, pp. 1–7, 2012.

[233] L. Hirschman, G. A. P. C. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenço, R. Nash, A. L. Veuthey, T. Wiegers, and A. G. Winter, "Text mining for the biocuration workflow," *Database*, vol. 2012, pp. 1–10, 2012.

[234] B. Xie, Q. Ding, H. Han, and D. Wu, "MiRCancer: A microRNA-cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, 2013.

[235] J. L. Rukov, R. Wilentzik, I. Jaffe, J. Vinther, and N. Shomron, "Pharmaco-miR: linking microRNAs and drug effects.," *Brief. Bioinform.*, Jan. 2013.

[236] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "HMDD v2.0: A database for experimentally supported human microRNA and disease associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 1–5, 2014.

[237] A. Ruepp, A. Kowarsch, D. Schmidl, F. Buggenthin, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, and F. J. Theis, "PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes," *Genome Biol.*, vol. 11, no. 1, p. R6, 2010.

[238] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, and I. Iliopoulos, "Protein–protein interaction predictions using text mining methods," *Methods*, vol. 74, pp. 47–53, Mar. 2015.

[239] D. P. Bartel, R. Lee, and R. Feinbaum, "MicroRNAs : Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[240] X. Li, W. Jiang, Y. Zhang, F. Meng, B. Lian, X. Chen, X. Yu, E. Dai, S. Wang, X. Liu, X. Li, and L. Wang, "Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease," *Bioinformatics*, vol. 29, no. 20, pp. 2596–2602, 2013.

[241] C. Delay and S. S. Hébert, "MicroRNAs and Alzheimer's Disease Mouse Models: Current Insights and Future Research Avenues.," *Int. J. Alzheimers. Dis.*, vol. 2011, p. 894938, 2011.

[242] P. Leidinger, C. Backes, S. Deutscher, K. Schmitt, S. C. Mueller, K. Frese, J. Haas, K. Ruprecht, F. Paul, C. Stähler, C. J. Lang, B. Meder, T. Bartfai, E. Meese, and A. Keller, "A blood based 12-miRNA signature of Alzheimer disease patients.," *Genome Biol.*, vol. 14, no. 7, p. R78, Jul. 2013.

[243] S. Gupta, K. E. Ross, C. O. Tudor, C. H. Wu, C. J. Schmidt, and K. Vijay-Shanker, "miRiaD: A Text Mining Tool for Detecting Associations of microRNAs with Diseases," *J. Biomed. Semantics*, vol. 7, no. 1, p. 9, 2016.

[244] G. Li, K. E. Ross, C. N. Arighi, Y. Peng, and C. H. Wu, "miRTex : A Text Mining System for miRNA- Gene Relation Extraction," pp. 1–24, 2015.

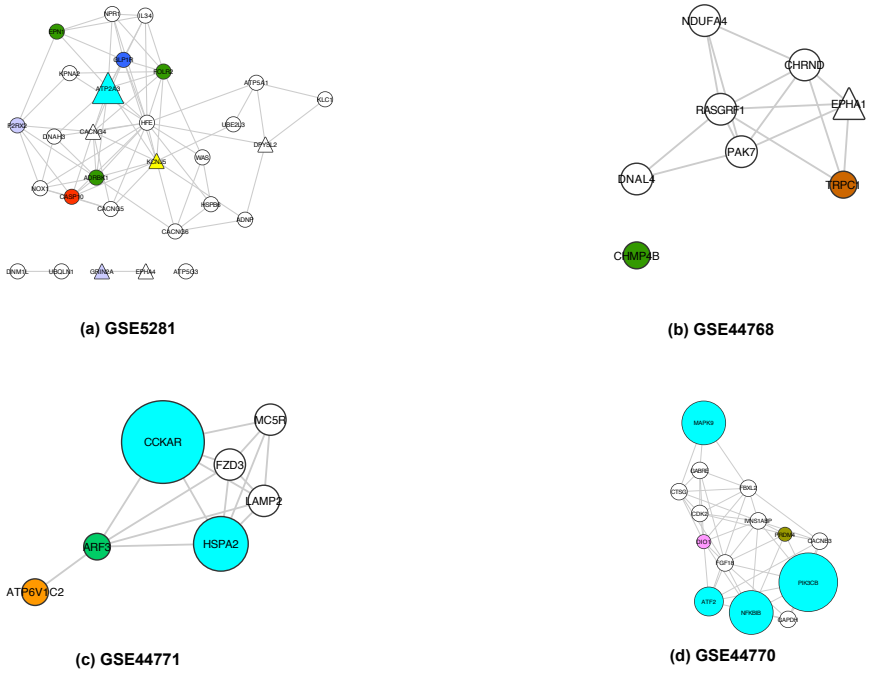
[245] F. N. Doubal, M. Ali, G. D. Batty, A. Charidimou, M. Eriksdotter, M. Hofmann-Apitius, Y. Kim, D. A. Levine, G. Mead, H. A. M. Mucke, C. W. Ritchie, C. J. Roberts, T. C. Russ, R. Stewart, W. Whiteley, and T. J. Quinn, "Big data and data repurposing - using existing data to answer new questions in vascular dementia research.," *BMC Neurol.*, vol. 17, no. 1, p. 72, Apr. 2017.

[246] Y. Berlyand, D. Weintraub, S. X. Xie, I. A. Mellis, J. Doshi, J. Rick, J. McBride, C. Davatzikos, L. M. Shaw, H. Hurtig, J. Q. Trojanowski, and A. S. Chen-Plotkin, "An Alzheimer's Disease-Derived Biomarker Signature Identifies Parkinson's Disease Patients with Dementia," *PLoS One*, vol. 11, no. 1, p. e0147319, Jan. 2016.

[247] J. Satoh, Y. Kino, and S. Niida, "MicroRNA-Seq Data Analysis Pipeline to Identify Blood Biomarkers for Alzheimer's Disease from Public Data," *Biomark. Insights*, p. 21, Apr. 2015.

[248] A. L. Young, N. P. Oxtoby, P. Daga, D. M. Cash, N. C. Fox, S. Ourselin, J. M. Schott, and D. C. Alexander, "A data-driven model of biomarker changes in sporadic Alzheimer's disease," *Brain*, vol. 137, no. 9, pp. 2564–2577, Sep. 2014.

6.2.1 Supplementary Figure



Supplementary Figure S1: Subnetwork derived for the hub genes in each dataset. Here the colors of the nodes represent the pathways involved. Lavender: Calcium signaling, Yellow: Estrogen signaling, Red: Apoptosis, Pink: Thyroid signaling, Orange: Synaptic signaling, and Olive green: Neurotrophin signaling. Nodes in Cyan are involved in more than one pathways and the size of the nodes depends on the number of pathways involved. Triangle nodes represent presence of a SNP

6.3 Summary

The presented study emphasises on the potential of literature knowledge and GRNs as a powerful framework for large-scale integrative meta-analysis to corroborate common AD mechanistic patterns. GRNs were inferred by leveraging the power of an ensemble-based method, BC3Net. To expand the knowledge space around previously unattended players, an optimised version of BC3Net was developed, BC3Net10. For generating AD specific GRNs, 500 most frequently discussed AD genes from the literature were injected into BC3Net10. Using this as the 'seed', first set of GRNs were generated independently for each dataset. Further, to overcome the incomplete nature of prior knowledge and identify new candidates from the grey-zone of knowledge, the seed was expanded with the genes from significant pathways that were common to these datasets. To obtain a stable and complete GRN, the enrichment and injection of seed was iteratively carried out until saturation. Although there were significant pathways common across datasets, the genes involved were not the same. Thus, to derive a stronger consensus and uplift the most

promising pathway the GRNs from different datasets were merged as one consensus GRN. The identified genes in these uplifted pathways, genetic variant analysis was applied. This resulted in 47 potential gene candidates, among which five well-known AD candidates were present. The value of this work comes from the identification of nine lesser known candidates that are mainly involved in pathways contributing to neurotransmission.

Although many sophisticated approaches and methods to generate GRNs using prior knowledge have been published in recent years, none have applied prior knowledge to self-instruct GRNs for identification of subtle signals that bear the potential to modulate the clinical path of the disease. To our knowledge, this is the first study to apply such an approach to retain inferred interactions that were lost due to shift in the significance considering the change in seed. Thus, our method is well-suited to provide a novel foundation for the generation of new hypotheses.

Chapter 7 Conclusion and Outlook

Integrative approaches are evolving into a promising way for translating big data into evidence-based decision support. Influencing healthcare and drug discovery research, they widen the knowledge space by contributing a specialised set of insights unique to individual resources. These approaches play a vital role in determining the constellation of interrelated yet diverse AD factors from disparate public data. However, each data resource and its associated computational approaches bring with it a set of unique challenges.

In this dissertation, I highlighted some of the challenges faced during the integration of biomedical data and approaches to address them. Mainly, I focused on solving the issues that revolve around two V's of big data that have an implication on data integration in NDD research: *Variety* and *Veracity*. The core of this work lies in collating highly curated, and context-specific heterogeneous public data in an integrative semantic framework for modelling diseases. To collate the disjoint data and to extract unbiased knowledge, new methods were developed and improvements to existing methods were achieved. Particularly, significant contribution to knowledge discovery, data mining, and network inference have been made. Using such a semantic framework, this dissertation reports deciphering previously unknown findings and set forth novel hypotheses that can accelerate innovation in AD research, deviating from the common Aβ- and tau-centric approaches.

7.1 Knowledge discovery and data mining contribution

In this dissertation, we developed novel approaches and methods to address the fundamental issues in integrative disease modelling for repurposing public data through several representative studies. In the first study, called *NeuroRDF*, we developed a semantic-based integrative approach using RDF framework to integrate well-curated and indication-specific data from databases, literature, and gene expression studies. This approach illustrated the potential of high-quality integrated data to exploit implicit associations and prioritise previously unattended AD candidates around well-known mechanisms. The remainder of the dissertation was dedicated to developing strategies and approaches to harvest, curate, extract, and analyse public data that were integrated into the *NeuroRDF* framework.

Although no consensus has yet been reached about the role of miRNAs in AD, several recent studies have suggested their value in early diagnosis. Hence, advancements in this direction are essential. In the second study, we developed text mining methods to automatically extract regulatory relationships of miRNAs with diseases and genes from the biomedical literature. Among the evaluated relation extraction approaches, tri-occurrence achieved state-of-the-art performance and outperformed existing approaches with comparable precision. This work provides the basis for building regulatory networks for identification of dysregulated miRNAs.

High-throughput data have a huge impact on drug discovery and their metadata annotations serve as the backbone for standardised retrieval, querying and consistent analysis. However, high variability of metadata information stored in public repositories hinders integrative meta-analysis. In the third study, we developed a database, named *NeuroTransDB*, primarily aimed to fill the gaps in the meta-analysis by joining bits of missing and scattered metadata information for public transcriptomic studies. This is the first highly curated metadata database that caters to the needs of NDD research by differentiating metadata fields for human and animal models. Additionally, we have systematically described the curation guidelines for building such a domain-specific database. To unlock the hidden potential of these harvested disparate data, a novel strategy that corroborates common mechanistic patterns across biologically related transcriptomic data is presented in the fourth study. This is the first study to generate knowledge-instructed gene regulatory networks for identifying lesser known candidates embedded in stable and robust functional patterns across heterogeneous AD datasets.

7.2 NDD research domain contribution

Comprehensive characterisation of the pathological events requires a systems level understanding by incorporating existing and new data. Semantic web technologies have proven to be particularly useful in providing a formalised framework for heterogeneous data integration and analysis. AlzPharm is a good example of a community effort for sharing AD data to advance hypothesis-driven therapeutic innovation. In general, a series of BioHackathons have aimed to increase interoperability between biological data (both structured and unstructured) and bioinformatics tools through semantic web. However, to pursue this ambitious aim in dementia and derive novel hypothesis, one needs to first address barriers associated with the data itself.

Public data in the field of dementia research are very heterogeneous with varying quality due to inconsistent diagnosis criteria and lack of adoption to standards; contributing to inconsistent reproducibility and reusability. Many commercial tools like MetaCore™, Ingenuity Pathway Studio, and NextBio have tried to address these issue by enriching with manual curation efforts, but rather fail to add domain context, extract missing values and verify incorrect information. In addition, they are not freely available for public use. In this dissertation, Chapter 3, *NeuroRDF* tries to address these issues and demonstrate that well-curated, precise, and formalised data have a huge impact on the deriving novel hypothesis from integrated public data. Even with limited yet well-curated data resources, we were able to prioritise MIF as a potential candidate to elucidate AD aetiology. Recently, a study conducted by Kassaar *et al.* [256] on human AD brain samples, has confirmed MIF hypothesis. These authors provide new evidence that implicates glucose modified and oxidised MIF as a potential link between dementia and diabetes. In addition, their results confirm an increase in the concentration of modified and oxidised MIF from early to late AD stages.

To start any research, the scientific literature is the primary source of knowledge; where scientists look for relevant and previous findings. It provides a comprehensive view across different disciplines and domains. Text mining technologies are a crucial part to extract knowledge from the vast growing literature, which otherwise is not achievable through manual reading. A recent monograph by Kostoff *et al.* [257], reported the usage of advanced text mining/information retrieval methodology to identify 600 actionable foundational causes of AD. Mining information from text is not an isolated problem, rather be considered as the evidence to be integrated with the experimental data; increasing the confidence of informed-decision making. Conventional preconceptions of cellular and molecular regulation, especially in neurology, — depended on genes and proteins — have changed with substantial progress in understanding of complex transcriptional landscape and non-coding RNA biology [258], especially miRNAs. This dissertation reports one of the early efforts for mining miRNA relations from text. The work presented in Chapter 4 is one of the first ones that contributed to text mining methods and manually curated corpus to the miRNA research community [259]. Using this work as a benchmark, several other text mining approaches have been built such as miRTex [251], MiNCor [259], and IBRel [260]. Among the miRNAs that were reported in *NeuroRDF*, recently, miR-132 has been confirmed to be associated with NFTs accumulation in subjects from two longitudinal

cohorts of ageing [261]. A most recent study, by Díez-Planelles *et al.* [262], linking the Huntington's disease diagnosis with dysregulated expression of circulating miRNAs in 15 patients, serve as a promising approach for early non-invasive diagnosis and prognosis in neurodegenerative diseases; paving way to precision medicine.

Although open science and global data sharing have gained momentum with recent initiatives like ADNI, JPND, and GAAIN. These future research projects/collaborations could leverage on existing experimental data to identify critical knowledge gaps in dementia. However, existing data resources are far from reusable as they are highly dependent on the quality and completeness of metadata. A recent call by IMI "FAIRification of IMI and EFPIA data – IMI2 – 2017-12-02"[41] recognizes this hurdle. In our third study, we have tried to highlight the need for FAIRification of the largest and most widely used public genomics databases, GEO and ArrayExpress, to cater to NDD research needs. This work clearly highlights that increasing interoperability of the metadata between resources is not sufficient to drive innovation. Moreover, complete and in-depth annotations that represent the heterogeneity and diversity in a specific domain like NDD are to be addressed first. Most of the existing research articles have tried to address this on a case-by-case basis [83,263]. Moreover, animal models have not shown high predictive validity in translating to AD clinical trials [264]. Onos *et al.* [265] have discussed the importance of detailed information on mouse models can avoid over interpretation of the derived results and help design more predictive mouse models the recapitulate clinical pathology for future experiments. Our work, reported in Chapter 5, is the only study which is dedicated to addressing these needs in the required breadth and depth for NDD, including detailed metadata for animal models, in a more structured way.

Representing biology as networks allow scientists to understand the cross-functioning of the constituent elements. Gene regulatory networks (GRNs) aim to identify organisational similarities between molecular and cellular players using expression data. GRNs offer insights into causal relationships, biomarkers, perturbation, predict expression changes, etc. to advance mechanistic understanding and prioritisation of potential candidates. Zhang *et al.* [201] used GRNs to identify influential modules to further prioritise TYROBP as a key

[41] https://www.up2europe.eu/calls/fairification-of-imi-and-efpia-data-imi2-2017-12-02_1941.html

regulator in LOAD patients, which was confirmed by in vitro experiments. However, most of the retrospective approaches undermine the lesser-known evidence that approximate the biological truth and clearly tend to converge to our current knowledge of NDD. In addition, none of these studies have addressed the context-specificity and completeness of the generated GRNs. Chapter 6 reports a novel approach that emphasises on using literature knowledge as the seed to iteratively determine the completeness of generated GRN, identification of robust mechanistic patterns across studies, and prioritisation of lesser-known candidates. This work has prioritised 5 well-known and 9 lesser-known AD candidates using public gene expression studies. The lesser-known candidates are observed to involved in three key pathways of neurotransmission in the generated GRNs: calcium signalling, endocytosis and synaptic vesicle cycle. A recent review by the Alzheimer's Association Calcium Hypothesis Workgroup [266] has placed Ca2+ in the centre of NDD; increasing the value and confidence the hypothesis derived from GRNs.

7.3 NDD projects contribution

In principle, the methods and approaches presented in this dissertation can be applied to any disease domain with little or no further adaptations. Moreover, the work presented here has made novel contributions to two projects:

1. **D10 "In Silico Discovery for putative Biomarkers"** – German Federal Ministry for Education and Research (BMBF) within BioPharma initiative 'Neuroallianz' (grant number: 1616060B)

2. **AETIONOMY "Development of Mechanism-based Taxonomy for Neurodegenerative Diseases"**– EU/EFPIA Innovative Medicines Initiative Joint Undertaking grant agreement n◦115568

In the D10 project, we applied the *NeuroRDF* framework to Parkinson's and Epilepsy diseases. We formulated complex queries on the integrated data that relates to molecule's biological role in the disease at the systems level. These queries were combined in a rationally informed weighting scheme as a set of features to rank putative biomarker candidates. The resulting prioritised novel candidates that provide novel mechanistic insights are currently being validated by our pharma partner, UCB Pharma. Depicted below is the overall workflow of the D10 project, which was previously published in a review:

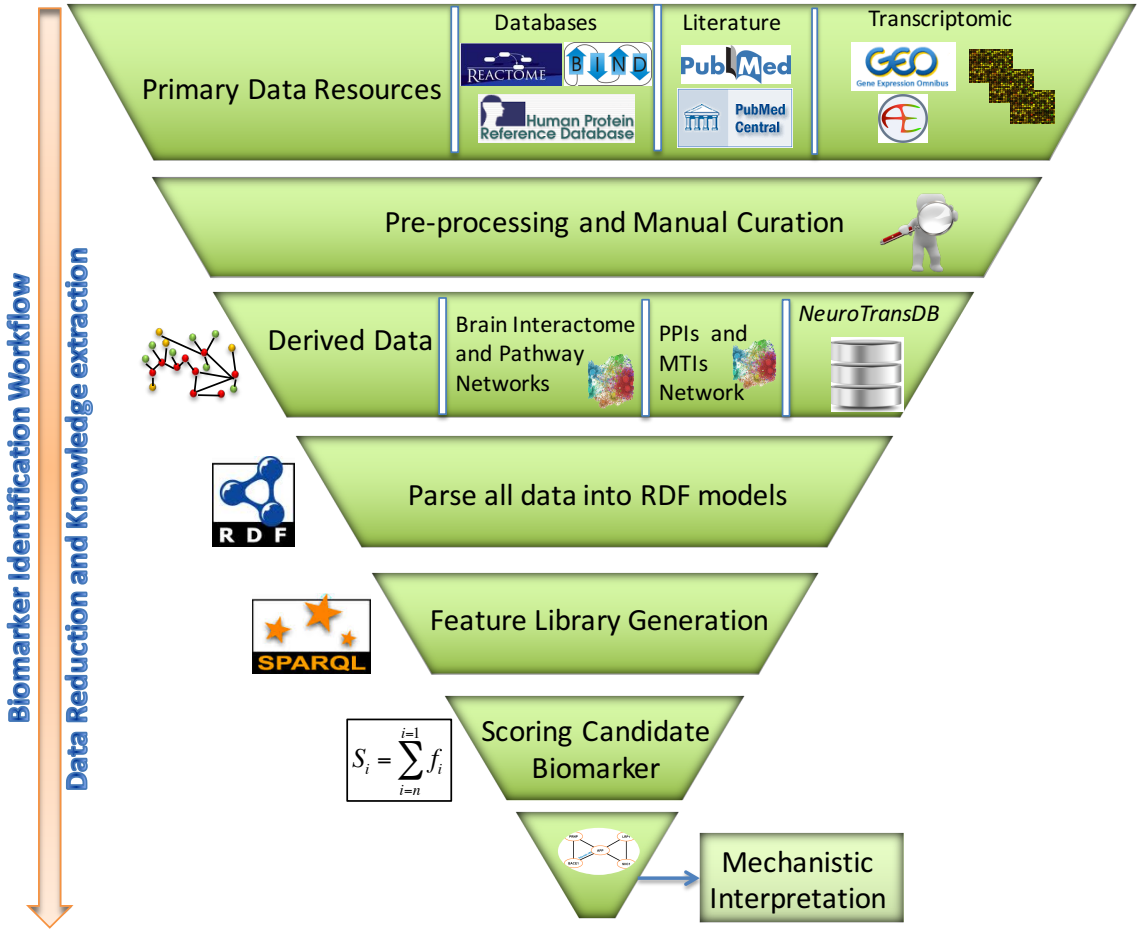


Figure 7.1: The D10 project workflow.

Here we present a top-down approach where highly-curated and context-specific data are integrated into the RDF framework. The candidates are scored based on the feature values determined for complex biological queries (features). Finally, the scored candidates are embedded in a mechanistic context for interpretation. Reproduced from Hofmann-Apitius *et al.* [267] under the Creative Commons Attribution License

The data harvested in the *NeuroTransDB* database is being utilised for mechanism-based patient subgroup identification for Alzheimer's and Parkinson's diseases in the AETIONOMY project. The data is currently migrated to the tranSMART framework to enable selection of relevant datasets for analysis. In addition, the prioritised candidates from the fourth study have been integrated into the web server for mechanism enrichment, NeuroMMSig[42].

[42] http://neurommsig.scai.fraunhofer.de/

7.4 Outlook

The work presented here shows that semantically integrating precise and context-specific data, derived from data mining and knowledge-discovery methods, support in identification of new molecular players. Certainly, the approach can be extended to other data types such as next generation sequencing, imaging, proteomics, so on. However, 'how to eliminate the influence of bias' in these data sources is still an open question worth exploring. To keep up with fast-evolving data, there is a need for more (semi-)automated approaches to make the submitted data more interoperable and identifiable, so as to reduce human efforts and cost. Recent crowdsourcing work such as CREEDS [268], and OMiCC [269] have proven to be efficient ways for annotating metadata and extracting gene signatures from GEO studies. Classifier can be trained on this high confidence and curated data to serve as an useful tool to automatically label and extract associations. However, the developed corpora are context specific. To extend to another domain similar human efforts are required. Active learning algorithms have shown to perform better, starting with only a few labelled examples, and dynamically improving its performance with user feedback. They have been applied to label experiments, text classification, entity recognition, interactions, and so on [270–272]. Similar active learning approaches can be developed for network reconstruction, which dynamically chooses the algorithm that optimizes to identify previously unknown candidates. One way forward to promote reusability and integration of public data at a global scale is to ensure that the data and resources developed by individual groups or in projects obtain long-term funding to maintain it.

References

[1] Prince M, Comas-Herrera A, Knapp M, Guerchet M, Karagiannidou M (2016) *World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and costs now and in the future*.

[2] Gauthier S, Albert M, Fox N, Goedert M, Kivipelto M, Mestre-Ferrandiz J, Middleton LT (2016) Why has therapy development for dementia failed in the last two decades? *Alzheimers. Dement.* **12**, 60–4.

[3] World Health Organization (2012) *Dementia: a public health priority*.

[4] Association A (2016) 2016 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **12**, 459–509.

[5] Braak H, Braak E (1997) Frequency of stages of Alzheimer-related lesions in different age categories. *Neurobiol. Aging* **18**, 351–357.

[6] Blennow K, de Leon MJ, Zetterberg H (2006) Alzheimer's disease. *Lancet (London, England)* **368**, 387–403.

[7] Naj AC, Schellenberg GD (2017) Genomic variants, genes, and pathways of Alzheimer's disease: An overview. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **174**, 5–26.

[8] Blennow K, Dubois B, Fagan AM, Lewczuk P, De Leon MJ, Hampel H (2015) Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease. *Alzheimer's Dement.* **11**, 58–69.

[9] Mucke L (2009) Neuroscience: Alzheimer's disease. *Nature* **461**, 895–897.

[10] Montine TJ, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Dickson DW, Duyckaerts C, Frosch MP, Masliah E, Mirra SS, Nelson PT, Schneider JA, Thal DR, Trojanowski JQ, Vinters H V., Hyman BT (2012) National institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: A practical approach. *Acta Neuropathol.* **123**, 1–11.

[11] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster M V., Phelps CH (2011) Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7**, 280–292.

[12] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7**, 270–279.

[13] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7**, 263–269.

[14] Karch CM, Cruchaga C, Goate AM (2014) Alzheimer's Disease Genetics: From the Bench to the Clinic. *Neuron* **83**, 11–26.

[15] Hyman BT, West HL, Rebeck GW, Buldyrev S V., Mantegna RN, Ukleja M, Havlin S, Stanley HE (1995) Quantitative analysis of senile plaques in Alzheimer disease: Observation of log-normal size distribution and molecular epidemiology of differences associated with apolipoprotein E genotype and trisomy 21 (Down syndrome). *Proc. Natl. Acad. Sci. U. S. A.* **92**, 3586–3590.

[16] Guerreiro RJ, Baquero M, Blesa R, Boada M, Brás JM, Bullido MJ, Calado A, Crook R, Ferreira C, Frank A, Gómez-Isla T, Hernández I, Lleó A, Machado A, Martínez-Lage P, Masdeu J, Molina-Porcel L, Molinuevo JL, Pastor P, Pérez-Tur J, Relvas R, Oliveira CR, Ribeiro MH, Rogaeva E, Sa A, Samaranch L, Sánchez-Valle R, Santana I, Tàrraga L, Valdivieso F, Singleton A, Hardy J, Clarimón J (2010) Genetic

screening of Alzheimer's disease genes in Iberian and African samples yields novel mutations in presenilins and APP. *Neurobiol. Aging* **31**, 725–731.

[17] Dobricic V, Stefanova E, Jankovic M, Gurunlian N, Novakovic I, Hardy J, Kostic V, Guerreiro R (2012) Genetic testing in familial and young-onset Alzheimer's disease: Mutation spectrum in a Serbian cohort. *Neurobiol. Aging* **33**, 1481.e7-1481.e12.

[18] Alzheimer A (1907) Uber eine eigenartige Erkrankung der Hirnrinde. *Allg Z Psychiat Psych-Gerichtl Med* **64**, 146–8.

[19] Joseph J, Shukitt-Hale B, Denisova NA, Martin A, Perry G, Smith MA Copernicus revisited: amyloid beta in Alzheimer's disease. *Neurobiol. Aging* **22**, 131–46.

[20] Castellani RJ, Smith MA (2011) Compounding artefacts with uncertainty, and an amyloid cascade hypothesis that is "too big to fail". *J. Pathol.* **224**, 147–52.

[21] Hardy J, Allsop D (1991) Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol. Sci.* **12**, 383–8.

[22] van Duijn CM, de Knijff P, Cruts M, Wehnert A, Havekes LM, Hofman A, Van Broeckhoven C (1994) Apolipoprotein E4 allele in a population-based study of early-onset Alzheimer's disease. *Nat. Genet.* **7**, 74–8.

[23] Janssen L, Keppens C, De Deyn PP, Van Dam D (2016) Late age increase in soluble amyloid-beta levels in the APP23 mouse model despite steady-state levels of amyloid-beta-producing proteins. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1862**, 105–112.

[24] Head E, Powell D, Gold BT, Schmitt FA (2012) Alzheimer's Disease in Down Syndrome. *Eur. J. Neurodegener. Dis.* **1**, 353–364.

[25] Murray ME, Dickson DW (2014) Is pathological aging a successful resistance against amyloid-beta or preclinical Alzheimer's disease? *Alzheimers. Res. Ther.* **6**, 24.

[26] Spillantini MG, Goedert M (2013) Tau pathology and neurodegeneration. *Lancet Neurol.* **12**, 609–622.

[27] Shi H, Zhang G, Zhou M, Cheng L, Yang H, Wang J, Sun J, Wang Z (2016) Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations. *PLoS One* **11**, 1–15.

[28] Ittner LM, Götz J (2011) Amyloid-β and tau--a toxic pas de deux in Alzheimer's disease. *Nat. Rev. Neurosci.* **12**, 65–72.

[29] Murray ME, Lowe VJ, Graff-Radford NR, Liesinger AM, Cannon A, Przybelski SA, Rawal B, Parisi JE, Petersen RC, Kantarci K, Ross OA, Duara R, Knopman DS, Jack CR, Dickson DW (2015) Clinicopathologic and 11C-Pittsburgh compound B implications of Thal amyloid phase across the Alzheimer's disease spectrum. *Brain* **138**, 1370–81.

[30] Brier MR, Gordon B, Friedrichsen K, McCarthy J, Stern A, Christensen J, Owen C, Aldea P, Su Y, Hassenstab J, Cairns NJ, Holtzman DM, Fagan AM, Morris JC, Benzinger TLS, Ances BM (2016) Tau and A imaging, CSF measures, and cognition in Alzheimers disease. *Sci. Transl. Med.* **8**, 338ra66-338ra66.

[31] Karran E, De Strooper B (2016) The amyloid cascade hypothesis: are we poised for success or failure? *J. Neurochem.* **139**, 237–252.

[32] Bloom GS (2014) Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis. *JAMA Neurol.* **71**, 505–8.

[33] Herrup K (2010) Reimagining Alzheimer's disease--an age-based hypothesis. *J. Neurosci.* **30**, 16755–62.

[34] Baumgart M, Snyder HM, Carrillo MC, Fazio S, Kim H, Johns H (2015) Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective. *Alzheimers. Dement.* **11**, 718–26.

[35] Lunnon K, Keohane A, Pidsley R, Newhouse S, Riddoch-Contreras J, Thubron EB, Devall M, Soininen H, Kłoszewska I, Mecocci P, Tsolaki M, Vellas B, Schalkwyk L, Dobson R, Malik AN, Powell J, Lovestone S, Hodges A (2017) Mitochondrial genes are altered in blood early in Alzheimer's disease. *Neurobiol. Aging* **53**, 36–47.

[36] Nunomura A, Perry G, Aliev G, Hirai K, Takeda A, Balraj EK, Jones PK, Ghanbari

H, Wataya T, Shimohama S, Chiba S, Atwood CS, Petersen RB, Smith MA (2001) Oxidative damage is the earliest event in Alzheimer disease. *J. Neuropathol. Exp. Neurol.* **60**, 759–67.

[37] Zhao Y, Zhao B (2013) Oxidative Stress and the Pathogenesis of Alzheimer's Disease. *Oxid. Med. Cell. Longev.* **2013**, 1–10.

[38] Craig LA, Hong NS, McDonald RJ (2011) Revisiting the cholinergic hypothesis in the development of Alzheimer's disease. *Neurosci. Biobehav. Rev.* **35**, 1397–1409.

[39] Goetzl EJ, Boxer A, Schwartz JB, Abner EL, Petersen RC, Miller BL, Kapogiannis D (2015) Altered lysosomal proteins in neural-derived plasma exosomes in preclinical Alzheimer disease. *Neurology* **85**, 40–47.

[40] Wolfe DM, Lee J-H, Kumar A, Lee S, Orenstein SJ, Nixon RA (2013) Autophagy failure in Alzheimer's disease and the role of defective lysosomal acidification. *Eur. J. Neurosci.* **37**, 1949–61.

[41] Vest RS, Pike CJ (2013) Gender, sex steroid hormones, and Alzheimer's disease. *Horm. Behav.* **63**, 301–7.

[42] Berridge MJ (2014) Calcium regulation of neural rhythms, memory and Alzheimer's disease. *J. Physiol.* **592**, 281–293.

[43] Calsolaro V, Edison P (2016) Neuroinflammation in Alzheimer's disease: Current evidence and future directions. *Alzheimers. Dement.* **12**, 719–32.

[44] Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, Jacobs AH, Wyss-Coray T, Vitorica J, Ransohoff RM, Herrup K, Frautschy SA, Finsen B, Brown GC, Verkhratsky A, Yamanaka K, Koistinaho J, Latz E, Halle A, Petzold GC, Town T, Morgan D, Shinohara ML, Perry VH, Holmes C, Bazan NG, Brooks DJ, Hunot S, Joseph B, Deigendesch N, Garaschuk O, Boddeke E, Dinarello CA, Breitner JC, Cole GM, Golenbock DT, Kummer MP (2015) Neuroinflammation in Alzheimer's disease. *Lancet. Neurol.* **14**, 388–405.

[45] Franco Bocanegra DK, Nicoll JAR, Boche D (2017) Innate immunity in Alzheimer's disease: the relevance of animal models? *J. Neural Transm.* 1–20.

[46] Heneka MT, Golenbock DT, Latz E (2015) Innate immunity in Alzheimer's disease. *Nat. Immunol.* **16**, 229–36.

[47] De Felice FG, Lourenco M V., Ferreira ST (2014) How does brain insulin resistance develop in Alzheimer's disease? *Alzheimers. Dement.* **10**, S26-32.

[48] Bedse G, Di Domenico F, Serviddio G, Cassano T (2015) Aberrant insulin signaling in Alzheimer's disease: current knowledge. *Front. Neurosci.* **9**, 204.

[49] Moreno-Treviño MG, Castillo-López J, Meester I (2015) Moving away from amyloid Beta to move on in Alzheimer research. *Front. Aging Neurosci.* **7**, 2.

[50] Herrup K (2015) The case for rejecting the amyloid cascade hypothesis. *Nat. Neurosci.* **18**, 794–9.

[51] Garrett MD (2016) Complexity Theory and Alzheimer's disease: A Call to Action. *Psychol. Today*.

[52] Calcoen D, Elias L, Yu X (2015) What does it take to produce a breakthrough drug? *Nat. Rev. Drug Discov.* **14**, 161–162.

[53] Cummings J (2010) What Can Be Inferred from the Interruption of the Semagacestat Trial for Treatment of Alzheimer's Disease? *Biol. Psychiatry* **68**, 876–878.

[54] Cummings J, Aisen PS, DuBois B, Frölich L, Jack CR, Jones RW, Morris JC, Raskin J, Dowsett SA, Scheltens P (2016) Drug development in Alzheimer's disease: the path to 2025. *Alzheimers. Res. Ther.* **8**, 39.

[55] Joanette Y, Hirsch EC, Goldman M (2014) The global fight against dementia. *Sci. Transl. Med.* **6**, 267ed22.

[56] International AD (2013) *The Global Impact of Dementia 2013 – 2050 Policy Brief for Heads of Government*.

[57] Cummings J, Morstorf T, Lee G (2016) Alzheimer's drug-development pipeline: 2016. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2**, 222–232.

[58] Mullane K, Williams M (2013) Alzheimer's therapeutics: continued clinical failures

question the validity of the amyloid hypothesis-but what lies beyond? *Biochem. Pharmacol.* **85**, 289–305.

[59] Langley GR (2014) Considering a new paradigm for Alzheimer's disease research. *Drug Discov. Today* **19**, 1114–1124.

[60] Citron M (2010) Alzheimer's disease: strategies for disease modification. *Nat. Rev. Drug Discov.* **9**, 387–98.

[61] Graham WV, Bonito-Oliva A, Sakmar TP (2017) Update on Alzheimer's Disease Therapy and Prevention Strategies. *Annu. Rev. Med.* **68**, 413–430.

[62] Carroll J (2017) *Another Alzheimer's drug flops in pivotal clinical trial*.

[63] Selkoe DJ, Hardy J (2016) The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8**, 1–14.

[64] Karran E, Hardy J (2014) A critique of the drug discovery and phase 3 clinical programs targeting the amyloid hypothesis for Alzheimer disease. *Ann. Neurol.* **76**, 185–205.

[65] De Strooper B (2014) Lessons from a failed γ-secretase Alzheimer trial. *Cell* **159**, 721–6.

[66] Castellani RJ, Perry G (2012) Pathogenesis and disease-modifying therapy in Alzheimer's disease: the flat line of progress. *Arch. Med. Res.* **43**, 694–8.

[67] Geerts H (2009) Of mice and men: bridging the translational disconnect in CNS drug discovery. *CNS Drugs* **23**, 915–26.

[68] Selkoe DJ (2011) Resolving controversies on the path to Alzheimer's therapeutics. *Nat. Med.* **17**, 1060–1065.

[69] Perry D, Sperling R, Katz R, Berry D, Dilts D, Hanna D, Salloway S, Trojanowski JQ, Bountra C, Krams M, Luthman J, Potkin S, Gribkoff V, Temple R, Wang Y, Carrillo MC, Stephenson D, Snyder H, Liu E, Ware T, McKew J, Fields FO, Bain LJ, Bens C (2015) Building a roadmap for developing combination therapies for Alzheimer's disease. *Expert Rev. Neurother.* **15**, 327–333.

[70] Mudher A, Lovestone S (2002) Alzheimer's disease-do tauists and baptists finally shake hands? *Trends Neurosci.* **25**, 22–6.

[71] Harrison JR, Owen MJ (2016) Alzheimer's disease: the amyloid hypothesis on trial. *Br. J. Psychiatry* **208**, 1–3.

[72] Becker RE, Greig NH (2012) Increasing the success rate for Alzheimer's disease drug discovery and development. *Expert Opin. Drug Discov.* **7**, 367–70.

[73] Hampel H, O'Bryant SE, Durrleman S, Younesi E, Rojkova K, Escott-Price V, Corvol J-C, Broich K, Dubois B, Lista S (2017) A Precision Medicine Initiative for Alzheimer's disease: the road ahead to biomarker-guided integrative disease modeling. *Climacteric* **0**, 1–12.

[74] Schultz T (2013) Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle. *Bull. Am. Soc. Inf. Sci. Technol.* **39**, 34–40.

[75] Machado CM, Rebholz-Schuhmann D, Freitas AT, Couto FM (2013) The semantic web in translational medicine: Current applications and future directions. *Brief. Bioinform.* **16**, 89–103.

[76] Moreau Y, Tranchevent L-C (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* **13**, 523–36.

[77] Derry JMJ, Mangravite LM, Suver C, Furia MD, Henderson D, Schildwachter X, Bot B, Izant J, Sieberts SK, Kellen MR, Friend SH (2012) Developing predictive molecular maps of human disease through community-based modeling. *Nat. Genet.* **44**, 127–130.

[78] Galasko D, Golde TE, Altar C, Amakye D, Bounos D, Bloom J, Clack G, Dean R, Devanarayan V, Fu D, Furlong S, Hinman L, Girman C, Lathia C, Lesko L, Madani S, Mayne J, Meyer J, Raunig D, Sager P, Williams S, Wong P, Zerba K, Koyama A, Okerere O, Yang T, Blacker D, Selkoe D, Grodstein F, Soares H, Potter W, Pickering E, Kuhn M, Immermann F, Shera D, Ferm M, Dean R, Simon A, Swenson F, Siuciak J, Kaplow J, Thambisetty M, Zagouras P, Koroshetz W, Wan H, Trojanowski J, Shaw L, Doecke J, Laws S, Faux N, Wilson W, Burnham S, Lam C,

Mondal A, Bedo J, Bush A, Brown B, Ruyck K De, Ellis K, Fowler C, Gupta V, Head R, Macaulay S, Pertile K, Rowe C, Rembach A, Rodrigues M, Rumble R, Szoeke C, Taddei K, Taddei T, Trounson B, Ames D, Masters C, Martins R, Hu W, Holtzman D, Fagan A, Shaw L, Perrin R, Arnold S, Grossman M, Xiong C, Craig-Schapiro R, Clark C, Pickering E, Kuhn M, Chen Y, Deerlin V Van, McCluskey L, Elman L, Karlawish J, Chen-Plotkin A, Hurtig H, Siderowf A, Swenson F, Lee V, Morris J, Trojanowski J, Soares H, Mehta P, Pirttila T, Patrick B, Barshatzky M, Mehta S, Fagan A, Head D, Shah A, Marcus D, Mintun M, Morris J, Holtzman D, Oijen M van, Hofman A, Soares H, Koudstaal P, Breteler M, Graff-Radford N, Crook J, Lucas J, Boeve B, Knopman D, Ivnik R, Smith G, Younkin L, Petersen R, Younkin S, Lambert J, Schraen-Maschke S, Richard F, Fievet N, Rouaud O, Berr C, Dartigues J, Tzourio C, Alpérovitch A, Buée L, Amouyel P, Schrijvers E, Koudstaal P, Hofman A, Breteler M, Butterfield L, Potter D, Kirkwood J, Yi J, Craft D, Gelfand C, Lista S, Faltraco F, Hampel H, Rifai M, Gillette M, Carr S, Burg S van der, Kalos M, Gouttefangeas C, Janetzki S, Ottensmeier C, Welters M, Romero P, Britten C, Hoos A (2013) Biomarkers for Alzheimer's disease in plasma, serum and blood - conceptual and practical problems. *Alzheimers. Res. Ther.* **5**, 10.

[79] Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712.

[80] Ashish N, Bhatt P, Toga AW (2016) Global Data Sharing in Alzheimer Disease Research. *Alzheimer Dis. Assoc. Disord.* **30**, 160–168.

[81] Pillai PS, Leong T-Y, Alzheimer's Disease Neuroimaging Initiative (2015) Fusing Heterogeneous Data for Alzheimer's Disease Classification. *Stud. Health Technol. Inform.* **216**, 731–5.

[82] Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, Scott Marshall M, Prud'hommeaux E, Hassanzadeh O, Pichler E, Stephens S (2011) Linked Open drug data for pharmaceutical research and development. *J. Cheminform.* **3**, 19.

[83] Fowler KD, Funt JM, Artyomov MN, Zeskind B, Kolitz SE, Towfic F (2015) Leveraging existing data sets to generate new insights into Alzheimer's disease biology in specific patient subsets. *Sci. Rep.* **5**, 14324.

[84] Chen JY, Shen C, Sivachenko AY (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. In *Pacific Symposium on Biocomputing*, pp. 367–78.

[85] Soler-López M, Zanzoni A, Lluís R, Stelzl U, Aloy P (2011) Interactome mapping suggests new mechanistic details underlying Alzheimer's disease. *Genome Res.* **21**, 364–76.

[86] Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15148–53.

[87] Liu B, Jiang T, Ma S, Zhao H, Li J, Jiang X, Zhang J (2006) Exploring candidate genes for human brain diseases from a brain-specific gene network. *Biochem. Biophys. Res. Commun.* **349**, 1308–1314.

[88] Caberlotto L, Nguyen T-P (2014) A systems biology investigation of neurodegenerative dementia reveals a pivotal role of autophagy. *BMC Syst. Biol.* **8**, 65.

[89] Caberlotto L, Lauria M, Nguyen TP, Scotti M (2013) The central role of AMP-kinase and energy homeostasis impairment in Alzheimer's disease: A multifactor network analysis. *PLoS One* **8**,.

[90] Schenone M, Dančík V, Wagner BK, Clemons P a (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **9**, 232–40.

[91] Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P (2007) Data integration and genomic medicine. *J. Biomed. Inform.* **40**, 5–16.

[92] Willighagen EL, Alvarsson J, Andersson A, Eklund M, Lampa S, Lapins M, Spjuth O, Wikberg JE (2011) Linking the Resource Description Framework to cheminformatics and proteochemometrics. *J. Biomed. Semantics* **2 Suppl 1**, S6.

[93] Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y,

Cho H, Standley DM, Nakagawa A, Nakamura H (2012) Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.* **40**, 453–460.

[94] Neumann E (2005) Finding the critical path: Applying the semantic web to drug discovery and development. *Drug Discov. World* **6**, 25–33.

[95] Glimm B, Stuckenschmidt H (2016) 15 Years of Semantic Web: An Incomplete Survey. *KI - Künstliche Intelligenz* **30**, 117–130.

[96] Klapsing R, Neumann G, Conen W (2001) Semantics in Web engineering: Applying the resource description framework. *IEEE Multimed.* **8**, 62–68.

[97] Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. *Sci. Am.* **284**, 34–43.

[98] Miller E (2005) An Introduction to the Resource Description Framework. *Bull. Am. Soc. Inf. Sci. Technol.* **25**, 15–19.

[99] Holford ME, Khurana E, Cheung K-H, Gerstein M (2010) Using semantic web rules to reason on an ontology of pseudogenes. *Bioinformatics* **26**, i71-8.

[100] Haarslev V, Möller R (2000) Consistency Testing: The RACE Experience. In *Automated Reasoning with Analytic Tableaux and Related Methods: International Conference, TABLEAUX 2000, St Andrews, Scotland, UK, July 3-7, 2000 Proceedings*, Dyckhoff R, ed. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 57–61.

[101] Tsarkov D, Horrocks I (2006) FaCT++ Description Logic Reasoner: System Description. In *Lecture Notes in Computer Science* Springer, pp. 292–297.

[102] Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y (2007) Pellet: A practical OWL-DL reasoner. *Web Semant. Sci. Serv. Agents World Wide Web* **5**, 51–53.

[103] Mishra RB, Kumar S (2011) Semantic web reasoners and languages. *Artif. Intell. Rev.* **35**, 339–368.

[104] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A,

Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA., Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018.

[105] Hoehndorf R, Schofield PN, Gkoutos G V. (2015) The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* **16**, 1069–1080.

[106] Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A, Gehant S, Laibe C, Redaschi N, Wimalaratne SM, Martin M, Le Novère N, Parkinson H, Birney E, Jenkinson AM (2014) The EBI RDF platform: Linked open data for the life sciences. *Bioinformatics* **30**, 1338–1339.

[107] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* **41**, 706–716.

[108] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C (2005) Relations in biomedical ontologies. *Genome Biol.* **6**, R46.

[109] Smith AK, Cheung K-H, Yip KY, Schultz M, Gerstein MK (2007) LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. *BMC Bioinformatics* **8 Suppl 3**, S5.

[110] Luciano JS (2005) PAX of mind for pathway researchers. *Drug Discov. Today* **10**, 937–42.

[111] Lam HYK, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong GT, Liu N, Crasto C, Morse T, Stephens S, Cheung K-H (2007) AlzPharm:

integration of neurodegeneration data using RDF. *BMC Bioinformatics* **8 Suppl 3**, S4.

[112] Luciano JS, Andersson B, Batchelor C, Bodenreider O, Clark T, Denney CK, Domarew C, Gambet T, Harland L, Jentzsch A, Kashyap V, Kos P, Kozlovsky J, Lebo T, Marshall SM, McCusker JP, McGuinness DL, Ogbuji C, Pichler E, Powers RL, Prud'hommeaux E, Samwald M, Schriml L, Tonellato PJ, Whetzel PL, Zhao J, Stephens S, Dumontier M (2011) The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *J. Biomed. Semantics* **2 Suppl 2**, S1.

[113] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung K-H (2007) Advancing translational research with the Semantic Web. *BMC Bioinformatics* **8 Suppl 3**, S2.

[114] Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis G, Larkin J, Russell B (2015) The NIH Big Data to Knowledge (BD2K) initiative. *J. Am. Med. Informatics Assoc.* **22**, 1114–1114.

[115] Gardner D, Shepherd GM (2004) A Gateway to the Future of Neuroinformatics. *Neuroinformatics* **2**, 271–274.

[116] Cohen PR (2015) DARPA's Big Mechanism program. *Phys. Biol.* **12**, 45008.

[117] Muldoon SF (2013) *Encyclopedia of Computational Neuroscience*.

[118] Yi Z, Dongsheng W, Tielin Z, Bo X (2014) Linked Neuron Data (LND): A Platform for Integrating and Semantically Linking Neuroscience Data and Knowledge. *Front. Neuroinform.* **8**,.

[119] Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B (2012) Open PHACTS: Semantic interoperability for drug discovery. *Drug Discov. Today* **17**, 1188–1198.

[120] Okano H, Yamamori T (2016) How can brain mapping initiatives cooperate to

achieve the same goal? *Nat. Rev. Neurosci.* **17**, 733–734.

[121] Nielsen FÅ (2014) Brede tools and federating online neuroinformatics databases. *Neuroinformatics* **12**, 27–37.

[122] Clark T, Kinoshita J (2007) Alzforum and SWAN: The present and future of scientific web communities. *Brief. Bioinform.* **8**, 163–171.

[123] Rigden DJ, Fernández-Suárez XM, Galperin MY (2016) The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic Acids Res.* **44**, D1-6.

[124] Horgan RP, Kenny LC (2011) "Omic" technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet. Gynaecol.* **13**, 189–195.

[125] Buguliskis JS (2016) The Epigenetic Insights of RNA-Seq. *Clin. Omi.* **3**, 10–13.

[126] Pedrotty DM, Morley MP, Cappola TP (2012) Transcriptomic Biomarkers of Cardiovascular Disease. *Prog. Cardiovasc. Dis.* **55**, 64–69.

[127] Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

[128] Valdivia-granda W a, Dwan C (2006) *Chapter 6 MICROARRAY DATA MANAGEMENT*.

[129] Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210.

[130] ArrayExpress Database.

[131] Hitzemann R, Bottomly D, Darakjian P, Walter N, Iancu O, Searles R, Wilmot B, McWeeney S (2013) Genes, behavior and next-generation RNA sequencing. *Genes, Brain Behav.* **12**, 1–12.

[132] Thompson JA, Tan J, Greene CS (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ* **4**, e1621.

[133] Janes KA, Yaffe MB (2006) Data-driven modelling of signal-transduction networks.

Nat. Rev. Mol. Cell Biol. **7**, 820–828.

[134] Panigrahi PP, Singh TR (2013) Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: Revealed association with aging and other diseases. *J. Theor. Biol.* **334**, 109–121.

[135] Liang D, Han G, Feng X, Sun J, Duan Y, Lei H (2012) Concerted perturbation observed in a hub network in Alzheimer's disease. *PLoS One* **7**, e40498.

[136] Sudmant PH, Alexis MS, Burge CB (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16**, 287.

[137] Walsh C, Hu P, Batt J, Santos C (2015) Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays* **4**, 389–406.

[138] Tseng GC, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40**, 3785–99.

[139] Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J (2009) Data Integration in Genetics and Genomics: Methods and Challenges. *Hum. Genomics Proteomics* **2009**, 1–13.

[140] Taminau J, Lazar C, Meganck S, Nowé A (2014) Comparison of Merging and Meta-Analysis as Alternative Approaches for Integrative Gene Expression Analysis. *ISRN Bioinforma.* **2014**, 1–7.

[141] Pepe MS, Feng Z (2011) Improving Biomarker Identification with Better Designs and Reporting. *Clin. Chem.* **57**, 1093–1095.

[142] Konstantinopoulos PA, Cannistra SA, Fountzilas H, Culhane A, Pillay K, Rueda B, Cramer D, Seiden M, Birrer M, Coukos G, Zhang L, Quackenbush J, Spentzos D (2011) Integrated Analysis of Multiple Microarray Datasets Identifies a Reproducible Survival Predictor in Ovarian Cancer. *PLoS One* **6**, e18202.

[143] Xu L, Tan A, Winslow RL, Geman D (2008) Merging microarray data from separate

breast cancer studies provides a robust prognostic test. *BMC Bioinformatics* **9**, 125.

[144] Liu C-C, Hu J, Kalakrishnan M, Huang H, Zhou X (2009) Integrative disease classification based on cross-platform microarray data. *BMC Bioinformatics* **10**, S25.

[145] Ramasamy A, Mondry A, Holmes CC, Altman DG (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **5**, 1320–1332.

[146] Rudy J, Valafar F (2011) Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinformatics* **12**, 467.

[147] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9309–14.

[148] Zintzaras E, Ioannidis JPA (2008) Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Comput. Biol. Chem.* **32**, 38–46.

[149] Choi JK, Yu U, Kim S, Yoo OJ (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19 Suppl 1**, i84-90.

[150] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109-14.

[151] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–40.

[152] Domingo-Fernández D, Kodamullil AT, Iyappan A, Naz M, Emon MA, Raschka T, Karki R, Springstubbe S, Ebeling C, Hofmann-Apitius M (2017) Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics* **33**, 3679–3681.

[153] Wilson JL, Hemann MT, Fraenkel E, Lauffenburger DA (2013) Integrated network

analyses for functional genomic studies in cancer. *Semin. Cancer Biol.* **23**, 213–218.

[154] Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–2.

[155] Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297.

[156] Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* **4**, e1000140.

[157] Aytes A, Mitrofanova A, Lefebvre C, Alvarez MJ, Castillo-Martin M, Zheng T, Eastham JA, Gopalan A, Pienta KJ, Shen MM, Califano A, Abate-Shen C (2014) Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy. *Cancer Cell* **25**, 638–651.

[158] Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, Basso K, Beltrao P, Krogan N, Gautier J, Dalla-Favera R, Califano A (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* **6**, 377.

[159] Li Y, Jackson SA (2015) Gene Network Reconstruction by Integration of Prior Biological Knowledge. *G3 (Bethesda).* **5**, 1075–9.

[160] Giorgi FM, Del Fabbro C, Licausi F (2013) Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics* **29**, 717–24.

[161] Wang S, Yin Y, Ma Q, Tang X, Hao D, Xu Y (2012) Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis. *BMC Plant Biol.* **12**, 138.

[162] Chiang J-H, Yu H-C (2003) MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics* **19**, 1417–22.

[163] Nikitin A, Egorov S, Daraselia N, Mazo I (2003) Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics* **19**, 2155–7.

[164] Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* **21**, 2933–42.

[165] Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis O V, Wingender E (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**, 3576–9.

[166] Slater T (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today* **19**, 193–8.

[167] Lee WP, Tzou WS (2009) Computational methods for discovering gene networks from expression data. *Brief. Bioinform.* **10**, 408–423.

[168] Markowetz F, Spang R (2007) Inferring cellular networks – a review. *BMC Bioinformatics* **8**, S5.

[169] Vert J-P (2010) Reconstruction of Biological Networks by Supervised Machine Learning Approaches. In *Elements of Computational Systems Biology* John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 163–188.

[170] de la Fuente A What are Gene Regulatory Networks? In *Handbook of Research on Computational Methodologies in Gene Regulatory Networks* IGI Global, pp. 1–27.

[171] Khosravi P, Gazestani VH, Pirhaji L, Law B, Sadeghi M, Goliaei B, Bader GD (2015) Inferring interaction type in gene regulatory networks using co-expression data. *Algorithms Mol. Biol.* **10**, 23.

[172] Woolcock KJ, Stunnenberg R, Gaidatzis D, Hotz H-R, Emmerth S, Barraud P, Bühler M (2012) RNAi keeps Atf1-bound stress response genes in check at nuclear pores. *Genes Dev.* **26**, 683–92.

[173] Tallam A, Perumal TM, Antony PM, Jäger C, Fritz J V., Vallar L, Balling R, del Sol A, Michelucci A (2016) Gene Regulatory Network Inference of Immunoresponsive Gene 1 (IRG1) Identifies Interferon Regulatory Factor 1 (IRF1) as Its

Transcriptional Regulator in Mammalian Macrophages. *PLoS One* **11**, e0149050.

[174] Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 12182–6.

[175] D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16**, 707–726.

[176] Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 418–29.

[177] Allen JD, Xie Y, Chen M, Girard L, Xiao G (2012) Comparing statistical methods for constructing large scale gene networks. *PLoS One* **7**, e29348.

[178] Werhli A V, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics* **22**, 2523–31.

[179] Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559.

[180] Peng J, Wang P, Zhou N, Zhu J (2009) Partial Correlation Estimation by Joint Sparse Regression Models. *J. Am. Stat. Assoc.* **104**, 735–746.

[181] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7.

[182] Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform. Syst. Biol.* **2007**, 79879.

[183] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8.

[184] Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328.

[185] Kiani NA, Zenil H, Olczak J, Tegnér J (2016) Evaluating network inference methods in terms of their ability to preserve the topology and complexity of genetic networks. *Semin. Cell Dev. Biol.* **51**, 44–52.

[186] Altay G, Emmert-Streib F (2011) Structural influence of gene networks on their inference: analysis of C3NET. *Biol. Direct* **6**, 31.

[187] Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20**, 3594–3603.

[188] Perrin B-E, Ralaivola L, Mazurie A, Bottani S, Mallet J, D'Alché-Buc F (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics* **19 Suppl 2**, ii138-48.

[189] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, e12776.

[190] Sławek J, Arodź T (2013) ENNET: inferring large gene regulatory networks from expression data using gradient boosting. *BMC Syst. Biol.* **7**, 106.

[191] Haury A-C, Mordelet F, Vera-Licona P, Vert J-P (2012) TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.* **6**, 145.

[192] Guo S, Jiang Q, Chen L, Guo D (2016) Gene regulatory network inference using PLS-based methods. *BMC Bioinformatics* **17**, 545.

[193] de Matos Simoes R, Emmert-Streib F (2012) Bagging statistical network inference from large-scale gene expression data. *PLoS One* **7**, e33624.

[194] LEISERSON MDM, VANDIN F, Wu HT, Raphael BJ Heat diffusion based genetic network analysis. US 20170300614A1, 2017.

[195] Vandin F, Upfal E, Raphael BJ (2011) Algorithms for Detecting Significantly

Mutated Pathways in Cancer. *J. Comput. Biol.* **18**, 507–522.

[196] Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge J V, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ (2014) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114.

[197] Emmert-Streib F, Dehmer M, Haibe-Kains B (2014) Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. cell Dev. Biol.* **2**, 38.

[198] Simoes RDM, Dehmer M, Emmert-Streib F (2013) B-cell lymphoma gene regulatory networks: Biological consistency among inference methods. *Front. Genet.* **4**, 1–14.

[199] Rhinn H, Fujita R, Qiang L, Cheng R, Lee JH, Abeliovich A (2013) Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease. *Nature* **500**, 45–50.

[200] Swarup V, Geschwind DH (2013) Alzheimer's disease: From big data to mechanism. *Nature* **500**, 34–35.

[201] Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, Fluder E, Clurman B, Melquist S, Narayanan M, Suver C, Shah H, Mahajan M, Gillis T, Mysore J, MacDonald ME, Lamb JR, Bennett DA, Molony C, Stone DJ, Gudnason V, Myers AJ, Schadt EE, Neumann H, Zhu J, Emilsson V (2013) Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720.

[202] Forabosco P, Ramasamy A, Trabzuni D, Walker R, Smith C, Bras J, Levine AP, Hardy J, Pocock JM, Guerreiro R, Weale ME, Ryten M (2013) Insights into TREM2 biology by network analysis of human brain gene expression data. *Neurobiol. Aging* **34**, 2699–714.

[203] Miller J a, Horvath S, Geschwind DH (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12698–703.

[204] Ray M, Ruan J, Zhang W (2008) Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases. *Genome Biol.* **9**, R148.

[205] Zou D, Ma L, Yu J, Zhang Z (2015) Biological databases for human research. *Genomics. Proteomics Bioinformatics* **13**, 55–63.

[206] Kumari A, Kanchan S, Sinha RP, Kesheri M (2016) Applications of Bio-molecular Databases in Bioinformatics. In *Medical Imaging in Clinical Practice* InTech, pp. 329–351.

[207] Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pereira RP, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U (2013) ArrayExpress update-trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **41**, 987–990.

[208] Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73.

[209] Petryszak R, Burdett T, Fiorelli B, Fonseca N a., Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni JC, Malone J, Megy K, Rustici G, Tang AY, Taubert J, Williams E, Mannion O, Parkinson HE, Brazma A (2014) Expression Atlas update - A database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.* **42**, 926–932.

[210] Schaefer C, Meier A, Rost B, Bromberg Y (2012) Snpdbe: Constructing an nsSnp functional impacts database. *Bioinformatics* **28**, 601–602.

[211] Famiglietti ML, Estreicher A, Gos A, Bolleman J, Géhant S, Breuza L, Bridge A, Poux S, Redaschi N, Bougueleret L, Xenarios I, UniProt Consortium (2014) Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.* **35**, 927–35.

[212] Zhang Z, Sang J, Ma L, Wu G, Wu H, Huang D, Zou D, Liu S, Li A, Hao L, Tian M, Xu C, Wang X, Wu J, Xiao J, Dai L, Chen L-L, Hu S, Yu J (2014) RiceWiki: a

wiki-based database for community curation of rice genes. *Nucleic Acids Res.* **42**, D1222–D1228.

[213] Gundla NK, Chen Z (2016) Creating NoSQL Biological Databases with Ontologies for Query Relaxation. *Procedia Comput. Sci.* **91**, 460–469.

[214] Have CT, Jensen LJ, Wren J (2013) Are graph databases ready for bioinformatics? *Bioinformatics* **29**, 3107–3108.

[215] Brazas MD, Yim DS, Yamada JT, Ouellette BFF (2011) The 2011 bioinformatics links directory update: more resources, tools and databases and features to empower the bioinformatics community. *Nucleic Acids Res.* **39**, W3–W7.

[216] Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ, Desfeux A (2014) OMICtools: an informative directory for multi-omic data analysis. *Database* **2014**, bau069-bau069.

[217] NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7-19.

[218] SIB Swiss Institute of Bioinformatics Members (2016) The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res.* **44**, D27-37.

[219] Helmy M, Crits-Christoph A, Bader GD (2016) Ten Simple Rules for Developing Public Biological Databases. *PLoS Comput. Biol.* **12**, 1–8.

[220] Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V (2009) Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41**, 149–155.

[221] Rung J, Brazma A (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89–99.

[222] Fluck J, Hofmann-Apitius M (2014) Text mining for systems biology. *Drug Discov. Today* **19**, 140–144.

[223] Hearst MA (1999) Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - Association for Computational Linguistics, Morristown, NJ, USA, pp. 3–10.

[224] Leach SM, Tipney H, Feng W, Baumgartner WA, Kasliwal P, Schuyler RP, Williams T, Spritz RA, Hunter L (2009) Biomedical Discovery Acceleration, with Applications to Craniofacial Development. *PLoS Comput. Biol.* **5**, e1000215.

[225] Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B (2013) Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* **46**, 200–211.

[226] Rebholz-Schuhmann D, Oellrich A, Hoehndorf R (2012) Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.* **13**, 829–39.

[227] Lu Z (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford).* **2011**, baq036.

[228] Leser U, Hakenberg J (2005) What makes a gene name? Named entity recognition in the biomedical literature. *Brief. Bioinform.* **6**, 357–69.

[229] Ananiadou S, Kell DB, Tsujii J (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.* **24**, 571–579.

[230] Bagewadi S, Bobić T, Hofmann-Apitius M, Fluck J, Klinger R (2014) Detecting miRNA Mentions and Relations in Biomedical Literature. *F1000Research* **3**, 205.

[231] Bobić T, Klinger R, Thomas P, Hofmann-apitius M (2012) Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions. *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguist.* 35–43.

[232] Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J (2009) Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on BioNLP Shared Task - BioNLP '09* Association for Computational Linguistics, Morristown, NJ, USA, p. 1.

[233] WANG Y, KIM J-D, SÆTRE R, PYYSALO S, OHTA T, TSUJII J (2010)

IMPROVING THE INTER-CORPORA COMPATIBILITY FOR PROTEIN ANNOTATIONS. *J. Bioinform. Comput. Biol.* **8**, 901–916.

[234] Zhang Y, Tao C, Jiang G, Nair A a, Su J, Chute CG, Liu H (2014) Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network. *J. Biomed. Semantics* **5**, 33.

[235] Malhotra A, Younesi E, Bagewadi S, Hofmann-Apitius M (2014) Linking hypothetical knowledge patterns to disease molecular signatures for biomarker discovery in Alzheimer's disease. *Genome Med.* **6**,.

[236] Seymour E, Damle R, Sette A, Peters B (2011) Cost sensitive hierarchical document classification to triage PubMed abstracts for manual curation. *BMC Bioinformatics* **12**, 482.

[237] Davis AP, Wiegers TC, Murphy CG, Mattingly CJ (2011) The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database* **2011**, 1–12.

[238] Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-R, Simonis N, Rual J-F, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M (2009) Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46.

[239] Campos D, Lourenço J, Matos S, Oliveira JL (2014) Egas: a collaborative and interactive document curation platform. *Database (Oxford).* **2014**, 1–12.

[240] Rak R, Rowley A, Black W, Ananiadou S (2012) Argo: An integrative, interactive, text mining-based workbench supporting curation. *Database* **2012**, 1–7.

[241] Hirschman L, Burns GAPC, Krallinger M, Arighi C, Cohen KB, Valencia A, Wu CH, Chatr-Aryamontri A, Dowell KG, Huala E, Lourenço A, Nash R, Veuthey AL, Wiegers T, Winter AG (2012) Text mining for the biocuration workflow. *Database* **2012**, 1–10.

[242] Xie B, Ding Q, Han H, Wu D (2013) MiRCancer: A microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* **29**, 638–644.

[243] Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N (2013) Pharmaco-miR: linking microRNAs and drug effects. *Brief. Bioinform.*

[244] Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q (2014) HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* **42**, 1–5.

[245] Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.* **11**, R6.

[246] Bartel DP, Lee R, Feinbaum R (2004) MicroRNAs : Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes. *Cell* **116**, 281–297.

[247] Jiang W, Zhang Y, Meng F, Lian B, Chen X, Yu X, Dai E, Wang S, Liu X, Li X, Wang L, Li X (2013) Identification of active transcription factor and miRNA regulatory pathways in Alzheimer's disease. *Bioinformatics* **29**, 2596–602.

[248] Delay C, Hébert SS (2011) MicroRNAs and Alzheimer's Disease Mouse Models: Current Insights and Future Research Avenues. *Int. J. Alzheimers. Dis.* **2011**, 894938.

[249] Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stähler C, Lang CJ, Meder B, Bartfai T, Meese E, Keller A (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.* **14**, R78.

[250] Gupta S, Ross KE, Tudor CO, Wu CH, Schmidt CJ, Vijay-Shanker K (2016) miRiaD: A Text Mining Tool for Detecting Associations of microRNAs with Diseases. *J. Biomed. Semantics* **7**, 9.

[251] Li G, Ross KE, Arighi CN, Peng Y, Wu CH (2015) miRTex : A Text Mining System for miRNA- Gene Relation Extraction. 1–24.

[252] Doubal FN, Ali M, Batty GD, Charidimou A, Eriksdotter M, Hofmann-Apitius M, Kim Y, Levine DA, Mead G, Mucke HAM, Ritchie CW, Roberts CJ, Russ TC, Stewart R, Whiteley W, Quinn TJ (2017) Big data and data repurposing - using

existing data to answer new questions in vascular dementia research. *BMC Neurol.* **17**, 72.

[253] Berlyand Y, Weintraub D, Xie SX, Mellis IA, Doshi J, Rick J, McBride J, Davatzikos C, Shaw LM, Hurtig H, Trojanowski JQ, Chen-Plotkin AS (2016) An Alzheimer's Disease-Derived Biomarker Signature Identifies Parkinson's Disease Patients with Dementia. *PLoS One* **11**, e0147319.

[254] Satoh J, Kino Y, Niida S (2015) MicroRNA-Seq Data Analysis Pipeline to Identify Blood Biomarkers for Alzheimer's Disease from Public Data. *Biomark. Insights* 21.

[255] Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC (2014) A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* **137**, 2564–2577.

[256] Kassaar O, Pereira Morais M, Xu S, Adam EL, Chamberlain RC, Jenkins B, James T, Francis PT, Ward S, Williams RJ, van den Elsen J (2017) Macrophage Migration Inhibitory Factor is subjected to glucose modification and oxidation in Alzheimer's Disease. *Sci. Rep.* **7**, 42874.

[257] Kostoff RN, Zhang Y, Ma J, Porter AL, Buchtel HA (2017) Prevention and Reversal of Alzheimer's Disease. *Georg. Inst. Technol.*

[258] Salta E, De Strooper B (2017) Noncoding RNAs in neurodegeneration. *Nat. Rev. Neurosci.* **18**, 627–640.

[259] Sammartino JC, Krallinger M, Valencia A (2016) Annotation process, guidelines and text corpus of small non-coding RNA molecules: The MiNCor for microRNA annotations. In *CEUR Workshop Proceedings*, pp. 56–63.

[260] Lamurias A, Clarke LA, Couto FM (2017) Extracting microRNA-gene relations from biomedical literature using distant supervision. *PLoS One* **12**, e0171929.

[261] Patrick E, Rajagopal S, Wong H-KA, McCabe C, Xu J, Tang A, Imboywa SH, Schneider JA, Pochet N, Krichevsky AM, Chibnik LB, Bennett DA, De Jager PL (2017) Dissecting the role of non-coding RNAs in the accumulation of amyloid and

tau neuropathologies in Alzheimer's disease. *Mol. Neurodegener.* **12**, 51.

[262] Díez-Planelles C, Sánchez-Lozano P, Crespo MC, Gil-Zamorano J, Ribacoba R, González N, Suárez E, Martínez-Descals A, Martínez-Camblor P, Álvarez V, Martín-Hernández R, Huerta-Ruíz I, González-García I, Cosgaya JM, Visioli F, Dávalos A, Iglesias-Gutiérrez E, Tomás-Zapico C (2016) Circulating microRNAs in Huntington's disease: Emerging mediators in metabolic impairment. *Pharmacol. Res.* **108**, 102–110.

[263] Cheng W-C, Tsai M-L, Chang C-W, Huang C-L, Chen C-R, Shu W-Y, Lee Y-S, Wang T-H, Hong J-H, Li C-Y, Hsu IC (2010) Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. *BMC Bioinformatics* **11**, 421.

[264] Sasaguri H, Nilsson P, Hashimoto S, Nagata K, Saito T, De Strooper B, Hardy J, Vassar R, Winblad B, Saido TC (2017) APP mouse models for Alzheimer's disease preclinical studies. *EMBO J.* **36**, 2473–2487.

[265] Onos KD, Sukoff Rizzo SJ, Howell GR, Sasner M (2016) Toward more predictive genetic mouse models of Alzheimer's disease. *Brain Res. Bull.* **122**, 1–11.

[266] Alzheimer's Association Calcium Hypothesis Workgroup (2017) Calcium Hypothesis of Alzheimer's disease and brain aging: A framework for integrating new evidence into a comprehensive theory of pathogenesis. *Alzheimers. Dement.* **13**, 178–182.e17.

[267] Hofmann-Apitius M, Ball G, Gebel S, Bagewadi S, de Bono B, Schneider R, Page M, Kodamullil A, Younesi E, Ebeling C, Tegnér J, Canard L (2015) Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders. *Int. J. Mol. Sci.* **16**, 29179–29206.

[268] Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, Jenkins SL, Feldmann AS, Hu KS, McDermott MG, Duan Q, Clark NR, Jones MR, Kou Y, Goff T, Woodland H, Amaral FMR, Szeto GL, Fuchs O, Schüssler-Fiorenza Rose SM, Sharma S, Schwartz U, Bausela XB, Szymkiewicz M, Maroulis V, Salykin A, Barra CM, Kruth CD, Bongio NJ, Mathur V, Todoric RD, Rubin UE,

Malatras A, Fulp CT, Galindo JA, Motiejunaite R, Jüschke C, Dishuck PC, Lahl K, Jafari M, Aibar S, Zaravinos A, Steenhuizen LH, Allison LR, Gamallo P, de Andres Segura F, Dae Devlin T, Pérez-García V, Ma'ayan A (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* **7**, 12846.

[269] Shah N, Guo Y, Wendelsdorf K V., Lu Y, Sparks R, Tsang JS (2016) A crowdsourcing approach for reusing and meta-analyzing gene expression data. *Nat. Biotechnol.* **34**, 803–6.

[270] Cui B, Lin H, Yang Z (2009) Uncertainty sampling-based active learning for protein–protein interaction extraction from biomedical literature. *Expert Syst. Appl.* **36**, 10344–10350.

[271] Silva C, Ribeiro B (2007) On Text-based Mining with Active Learning and Background Knowledge Using SVM. *Soft Comput.* **11**, 519–530.

[272] Benferhat S, Eds MA, Goebel R (2017) *Advances in Artificial Intelligence: From Theory to Practice*.