HelmholtzZentrum münchen
German Research Center for Environmental Health

A network-based approach for the identification of multi-omics modules associated with complex human diseases

Maria Wörheide[1], Jan Krumsiek[3], Gabi Kastenmüller[1,4], Matthias Arnold[1,2]

(1) Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
(2) Department of Psychiatry and Behavioral Sciences, Duke University, Durham, NC, USA
(3) Weill Cornell Medicine, New York City, USA
(4) German Center for Diabetes Research (DZD), Neuherberg, Germany

Abstract

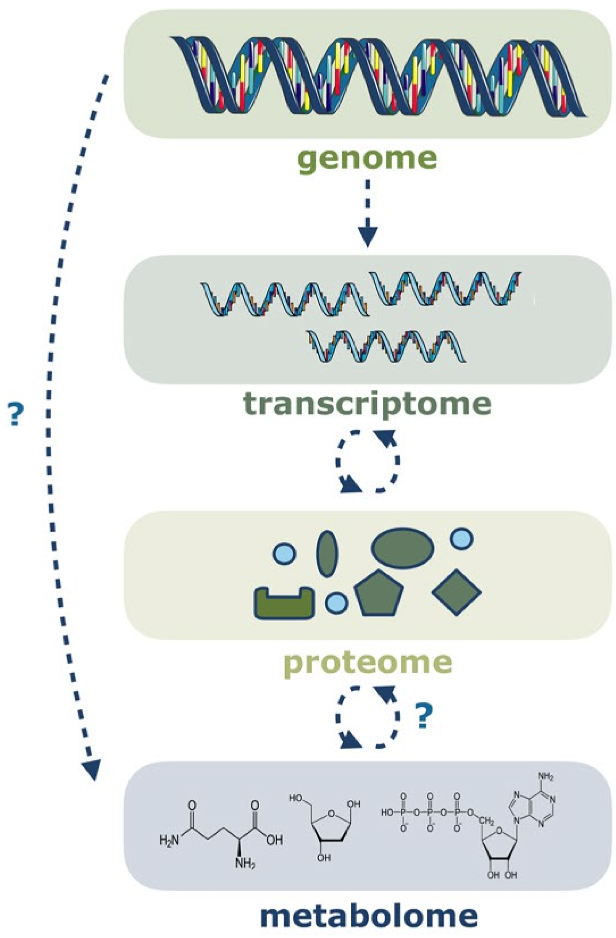
Application of advanced high-throughput omics technologies have provided us with vast amounts of quantitative, highly valuable data. For complex, heterogeneous, and untreatable diseases such as Alzheimer's disease (AD), the integration of different omics levels and their interconnections is desperately needed to understand the underlying molecular pathomechanisms and identify potential therapeutic targets. However, integrated, multivariate analyses of cross-omics data are not straightforward, and even if successfully applied, they often lack a human comprehensible representation.

Graph databases, such as Neo4j, provide an intuitive and mathematically well-defined framework to store and interconnect diverse biological domains in accessible network structures. Here, we propose a network-based, multi-omics framework developed with Neo4j that allows the large-scale integration and analysis of data on biological entities across omics, as well as results from association analysis with specific (endo)phenotypes. The backbone of this framework comes from known biological relationships such as gene-transcript-protein relations and functional/pathway annotations available in public databases.

This backbone is augmented with experimental, quantitative data across omics (e.g. eQTLs) derived in population-based studies. To identify modules within this network that are potentially relevant to a disease such as AD, we extend the framework using large-scale association data for AD (e.g. from case-control GWASs). We mined this comprehensive catalogue of biological information using established graph algorithms to identify potentially disease-related modules of tightly interlinked entities, and were able to obtain several subnetworks significantly enriched for AD-associations.

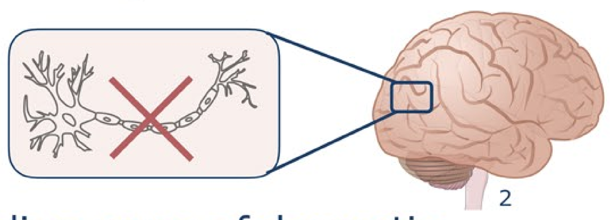
Motivation & Methods

Multi-omics data
from large international cohorts



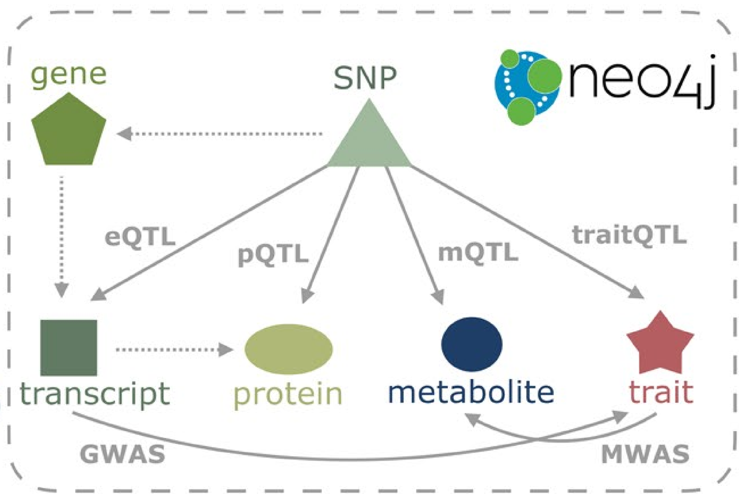
- complex interplay
- mainly analysed separately

Alzheimer's disease (AD)
a neurodegenerative disorder



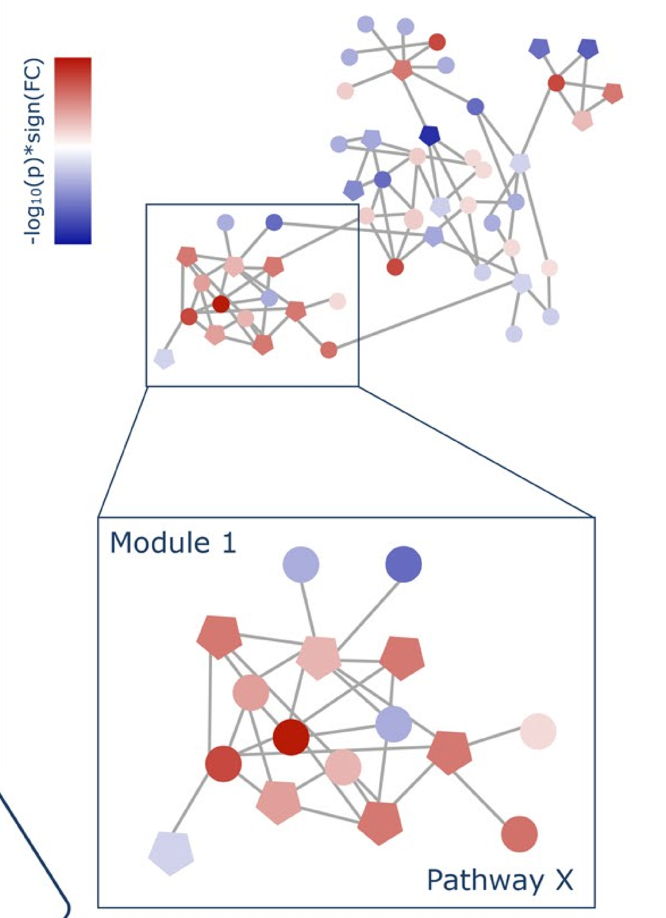
- leading cause of dementia
- no treatment or early intervention

Build molecular atlas
results from previous and novel studies



- multi-omics integration framework
- qualitative network of influences on AD

Analyse network
find interesting modules/entities



- tightly linked or dysregulated
- gain novel insights into AD
- aid prioritisation of targets

Data sources

Alzheimer's Disease Metabolomics Consortium

ADNI
- GGM with 127 metabolites
- metabolite-trait associations (mWAS)

AMP-AD
- gene co-expression for 7 brain regions

Genome-wide association studies (GWAS)
- cohorts incl. ADNI, UKBiobank, IGAP, ROS/MAP
- 36 AD related phenotypes across 6 studies
- 20,467,124 traitQTLs with p≤0.05

Population-based studies
- including KORA, TwinsUK
- 6 GWAS with metabolic traits (1,059,270 mQTL p≤0.05)
- GGM with 326 metabolites

GTEx v7 - Genotype-Tissue Expression project
- 337,621 eQTLs with q≤0.05
- 48 tissues (13 from the brain)

Ensembl v76
- 69,854 genes (23,848 protein-coding)
- 112,039 proteins and 234,475 transcripts
- 54,662,710 SNPs (variant annotation using SNiPA[1])

measured metabolites (platform-specific)
- 167 (Biocrates) and 870 (Metabolon)

Table 1: Database summary statistics

	biological entities	statistical results	all*	types
nodes	55,081,849	29,571,637	>85 million	51
edges	-	-	>170 million	39

* including knowledge databases such as HMDB, CheBI,...

Results

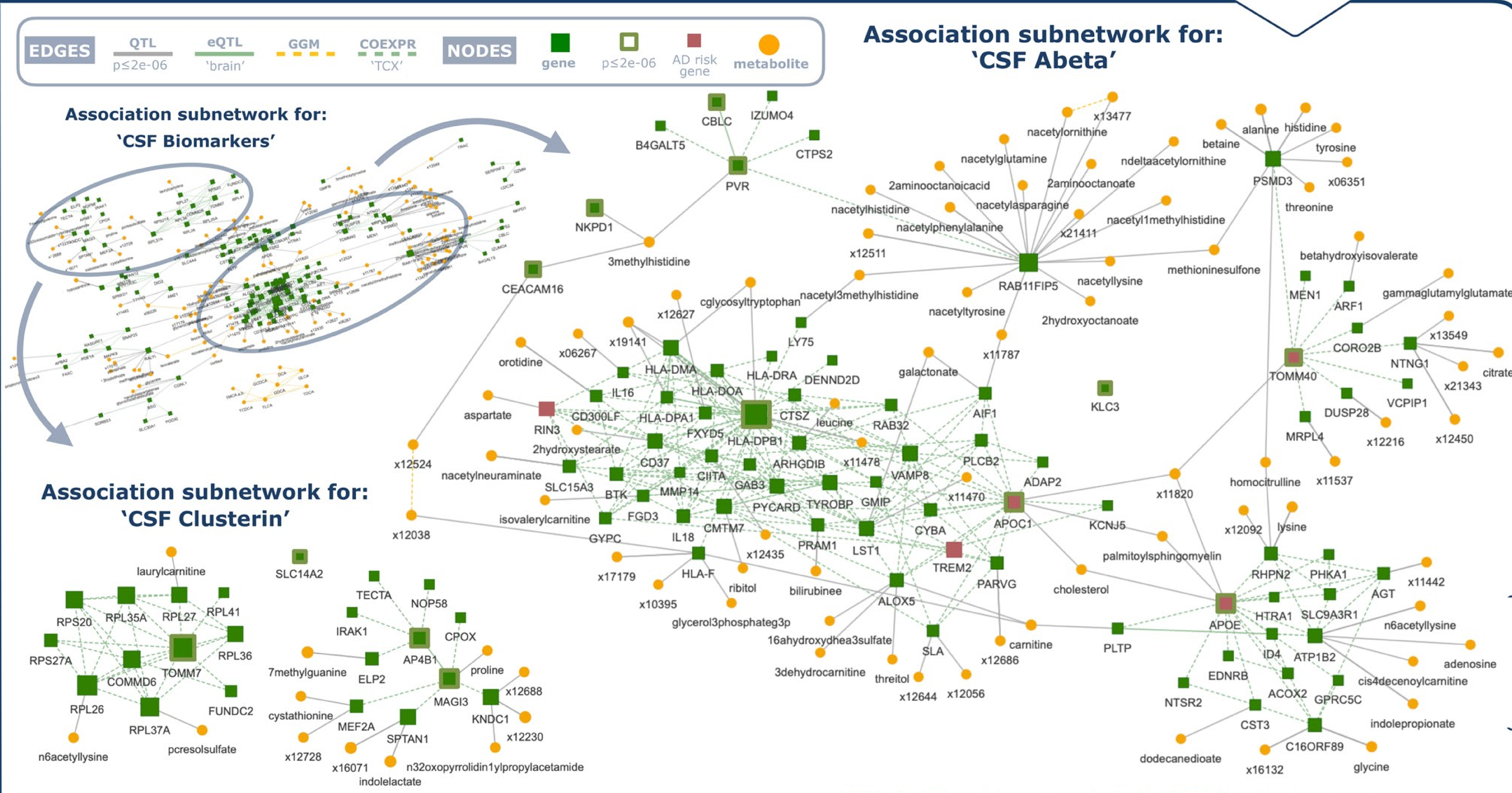


Table 2: GO enrichment analysis for 'CSF Clusterin' subnetwork

GO:Term	Term	P-value (Bonf.)
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	$4,84 \times 10^{-09}$
GO:0006613	cotranslational protein targeting to membrane	$6,16 \times 10^{-09}$
GO:0045047	protein targeting to ER	$7,48 \times 10^{-09}$
GO:0072599	establishment of protein localization to ER	$9,90 \times 10^{-09}$
GO:0070972	protein localization to ER	$2,86 \times 10^{-08}$
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	$3,08 \times 10^{-08}$
GO:0001775	protein targeting	$1,03 \times 10^{-07}$

Table 3: GO enrichment analysis for 'CSF Abeta' subnetwork

GO:Term	Term	P-value (Bonf.)
GO:0050865	regulation of cell activation	$8,14 \times 10^{-05}$
GO:0019882	antigen processing and presentation	$9,62 \times 10^{-05}$
GO:0002694	regulation of leukocyte activation	$2,44 \times 10^{-04}$
GO:0048002	antigen processing and presentation of peptide antigen	$3,63 \times 10^{-04}$
GO:0002376	immune system process	$7,40 \times 10^{-04}$
GO:0045321	leukocyte activation	$1,92 \times 10^{-03}$
GO:0001775	cell activation	$2,96 \times 10^{-03}$

Next steps

- application of established graph algorithms
- include proteomics and pQTL results
- transform qualitative network into quantitative network
- develop network-based scoring methods to identify potential disease drivers
- build user-friendly web-interface to query and analyse network

Contact

maria.woerheide@helmholtz-muenchen.de

@mwoerhe

For an online version of this poster:

(1) Arnold M, Raffler J, Pfeufer A, Suhre K, and Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics (2014).
(2) Sidiropoulos, Konstantinos, et al. "Reactome enhanced pathway visualization." *Bioinformatics* 33.21 (2017): 3461-3467.

IBIS Institute of Bioinformatics and Systems Biology · Duke University School of Medicine · Weill Cornell Medicine · HELMHOLTZ RESEARCH FOR GRAND CHALLENGES