SUPPLEMENTARY MATERIAL

System-wide pollution of biomedical Big Data: Consequences on knowledge.

Ankush Sharma [1,2,*] and Giovanni Colonna [3,*]

[1] Department of Biosciences University of Oslo, Oslo, Norway
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Medical Informatics, AOU-Vanvitelli, Università della Campania, Naples, Italy.
*Correspondence: ankush.sharma@ibv.uio.no, giovanni.colonna@unicampania.it

Received: date; Accepted: date; Published: date

BOX I - GLOSSARY OF TERMS

Entity is a thing with distinct and independent existence, for example, any **object** in the system that we want to model and store information about. Entities are recognizable **concepts**, either concrete or abstract, such as a surgeon, hospitals, things, or events that have relevance to the database. Some specific examples of entities in a biomedical database are gene, protein, disease, and so on.

Context is any information that can characterize the situation of an ***entity***.

Context-Aware System

A system that uses contexts to provide relevant information and services to users

Context and ontology

Context can be shared as an **ontology** (vocabularies of terms, events of the context) where ontology is important to define the **context** as a formal information. In **Artificial Intelligence**, ontology refers to an engineering artifact, made up by a specific vocabulary used to describe a certain reality, plus a set of explicit **assumptions** regarding the intended meaning of the vocabulary words.

Concept attributes

Intension - meaning of the concept.

Extension - the set of **objects** that the concept refers to.

Example: phrases "Non-coding RNA" and "Coding RNA" have different meanings (intension) but both refer to RNA molecule (extension).

Ontology - is a description of the concepts and relationships that can exist for an **agent** or a **community of agents**. It is a well-organized system of **human knowledge and information** made for machines to understand them easily and correctly through a shared understanding. Ontology is a common **framework** (vocabulary) that allows **data** to be shared and reused by humans and machines.

Ontology as Vocabulary - It is a specification of a conceptualization that refers to an engineering artifact, made up by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words, terms, lessics. It is a **Controlled Vocabulary** of Terms or Types of relations among terms, Rules, Axioms, and Constraints, where "Controlled" means a fixed set of (agreed upon) names used within a certain community to refer to subjects in certain domains.

Role of ontology: To make domain assumptions explicit. Easier to change domain assumptions (consider a genetics knowledge base). Easier to understand and update legacy data. To a separate

domain knowledge from operational knowledge. Re-use domain and operational knowledge separately (e.g., configuration based on constraints). To manage the combinatorial explosion of data/terms.

Organization of controlled vocabularies

Glossary - Controlled vocabularies + natural language explanation of the meaning of terms. We express meaning in a human readable form through Glossaries which help human not machines to understand the meaning of terms, often ambiguous.

Thesauri - Controlled vocabularies or glossary + some additional semantics. Synonyms / homonyms /antonyms relationships. Broader / narrower terms.

Index - Controlled vocabularies + references to the subject occurrences. Taxonomy and Classification Controlled vocabulary + hierarchic structure.

Ontology Components

Concepts / Class - Concepts of the domain or tasks, which are organized in taxonomies
Example: Genes, Proteins, Diseases.

Relations - The **interaction** between concepts of the domain.
Example: Protein A interacts with protein B, i.e., protein A binds protein B.

Functions - A special case of relations in which the n-the element of the relationship is unique for the n-1 preceding elements
Example: The sum of the prices

Axioms - Model sentences that are always true
Example: a + 0 = 0, if x > y, then x + a > y + a.

Instances/Individuals - To represent specific elements
Example: Tuna called Yellowfin tuna, the protein called myoglobin.

Semantics: The meaning of meaning. To share a common understanding of the structure of descriptive information among people, among software agents, between people and software. It is useful to enable the reuse of domain knowledge, introducing standards to allow interoperability.

Role of Computing as a Computational Intelligence

1. **Complex Systems Modeling**
- Discovering (and predicting) tempo-spatial transmission patterns
- Identifying underlying interactions

2. **Policy-level Decision Making**
- Active surveillance
- Strategy planning
- Resource deployment
- Policy assessment

3. **Process and Methods**
- Algorithms

Knowledge modeling is a process to create a computer interpretable model of knowledge. The resulting **knowledge model can only be computer interpretable** when it is expressed in some knowledge representation language or data structure that enables the knowledge to be interpreted by software and to be stored in a database or data exchange file.

Knowledge representation

Knowledge in biomedicine is often represented as a network or **graph. Graph** (or **network**) represents **semantic relations** between the **concepts**. It is made by vertices or nodes (represent concepts), connected by **edges** (represent relationships between nodes).

TABLE 1S

Set of 221 non-redundant genes extracted from a total of 337 hub-genes found in the literature. The acronyms of the genes are listed alphabetically.	**Redundancy of the 61 hub-genes. In parentheses the number of times that each gene has been found for a total of 177 times.**
A2M, ABAT, ACAA1, ACACA, ACADM, ACSM3, ACTB, AGXT, AHSG, AKT1, ALB, ALDH2, ALDH6A1, APOA1, APOC3, AR, ARHGAP39, ARPC4, ASPM, ATNXN1, AURKA, AXIN2, AZGP1, BCL2, BCLC, BHMT, BIRC5, BIRC5, BIRC5, BMP4, BUB1, BUB1B, C8ORF33, CALM3, CASP8, CCNA2, CCNB1, CCNB2, CCND1, CD8A, CDC20, CDC37L1, CDC45, CDCA8, CDH1, CDK1, CDKN1A, CDKN3, CENPA, CENPE, CENPF, CEP55, CHEK1, CKAP5, CLU, COL1A1, COL1A2, COL4A1, COMMD5, COPS5, CRYL1, CSNK2A1, CXCL1, CXCL12, CYCS, CYP2B6, CYP3A4, CYP4A11, DAO, DCAF13, DKK1, DNMT1, DSCC1, DTL, ECHDC2, ECHS1, EFNA4, EGFR, EGR1, EHHADH, ENO1, ERBB2, ESPL1, ESR1, EWSR1, F2, F8, FGF2, FN1, FOS, FOXO1, GCDH, GINS1, GLI1, GMPS, GNAO1, HBA1, HBA2, HBB, HBD, HDAC1, HLAB, HMGA1, HMMR, HRAS, HRG, HSF1, HSP, HSP90AA1, HSPA1A, IGF1, IKGKB, ILF3, INCENP, INTS8, ITGA2, JUN, KDM6B, KIF11, KIF20A, KIF23, KIF2C, KIF4A, KNG1, KRAS, KRT18, LRRC14, MAD2L1, MAPK1, MAPK8, MCM10, MCM2, MCM3, MCM4, MCM6, MDM2, MELK, MME, MMP2, MT2A, MUT, MYC, NCAPG, NCOR1, NDRG1, NEK2, NPM1, NSMCE2, NUDCD1, NUSAP1, OS, PBK, PCNA, PEG10, PHF20L1, PIK3CD, PIK3CG, PIK3R1, PINK6, PLCB1, PLK1, POP1, POTEF, POU3F4, PRC1, PRKCA, PRKDC,	CDC20(9), BUB1(6), CCNB2(6), TOP2A(6), CCNB1(5), CDK1(5), MAD2L1(5), MYC(5), HSP90AB1(4), ESR1(4), PRKDC(4), ARHGAP39(3), AURKA(3), BIRC5(3), BUB1B(3), C8ORF33(3), CCNA2(3), CDKN3(3), CSNK2A1(3), DSCC1(3), INTS8(3), NUDCD1(3), PCNA(3), POP1(3), PUSL1(3), RFC4(3), STIP1(3), TGD5(3), UBD(3), ACAA1(2), ACADM(2), ACSM3(2), AURKB(2), BCL2(2), CDKN1A(2), CEP55(2), CHEK1(2), CKAP5(2), COL1A1(2), DLGAP5(2), ERBB2(2), FOS(2), FOXO1(2), HMMR(2), HSF1(2), KIF20A(2), KIF2C(2), KNG1(2), KRAS(2), MCM2(2), MCM4(2), MELK(2), MMP2(2), MUT(2), PLK1(2), PRC1(2), RACGAP1(2), SPARC(2), VWF(2), YDJC(2), ZNF623(2).

PRKG2, PRR7, PTEN, PTGS2, PUSL1, PYCRL, Q12834, RACGAP1, RAP2A, RFC4, SCNN1A, SFN, SIRT1, SLA2547, SPARC, SPC24, SPP2, SRC, STAT3, STAU2, STIP1, SUCLA2, SUMO1, SUMO2, TAF12, TFRC, TGFB1, TIGD5, TK1, TOP1MT, TOP2A, TP53, TPX2, TTK, TTR, TXNRD1, UBD, UBE21, UBR5, UQCRC2, USL1, VCAM1, VEGFA, VIM, VTN, VWF, YDJC, YWHAZ, ZIC2, ZNF16, ZNF250, ZNF623, ZNF648, ZWINT	

Figure 1S : Interaction networks of 221 HUB genes found in the literature with a confidence score of 0.4, figure1A in main text

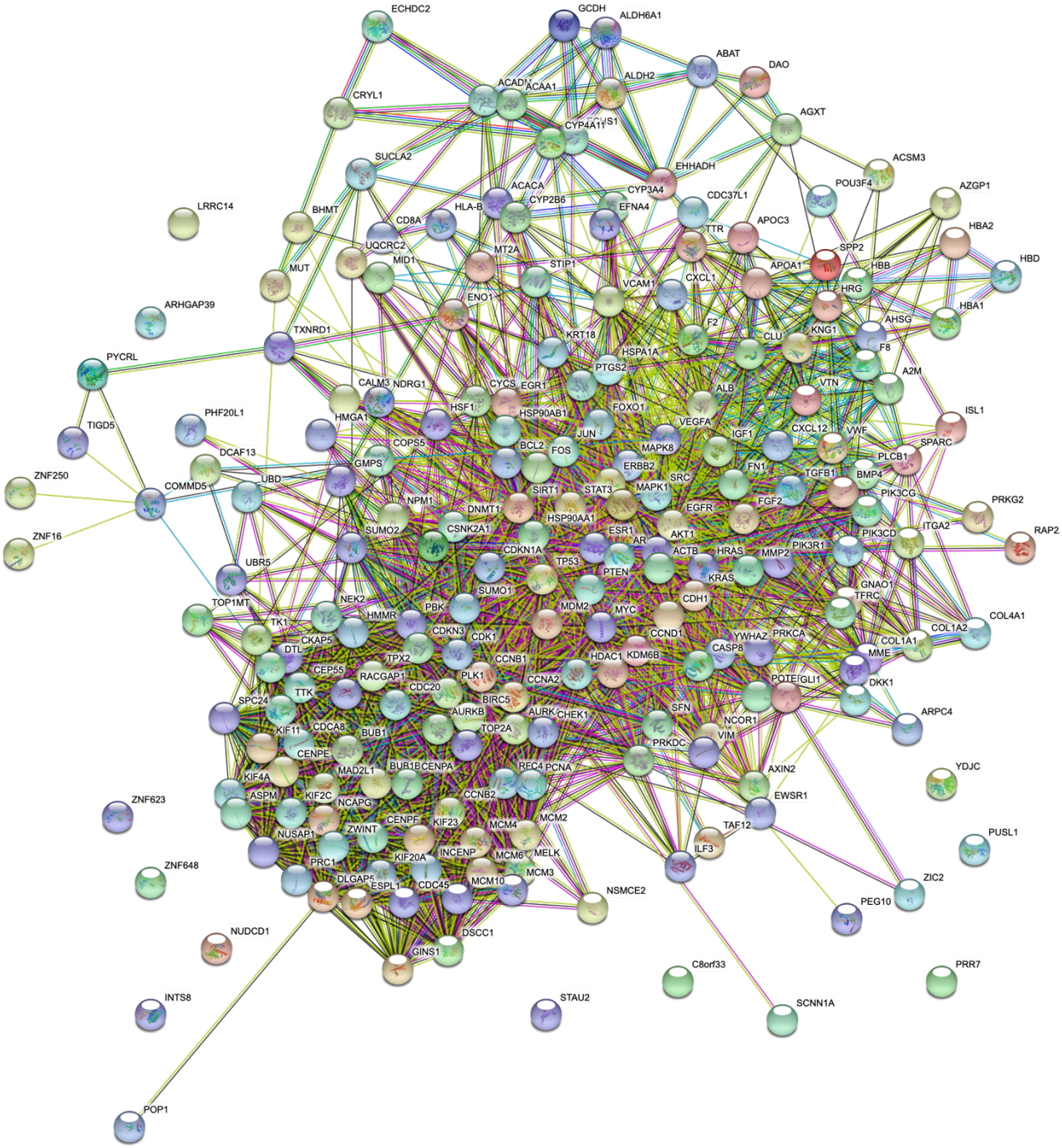

Figure2S. Interaction networks of 221 HUB genes with a confidence score of 0.9 . Figure 1b in main text

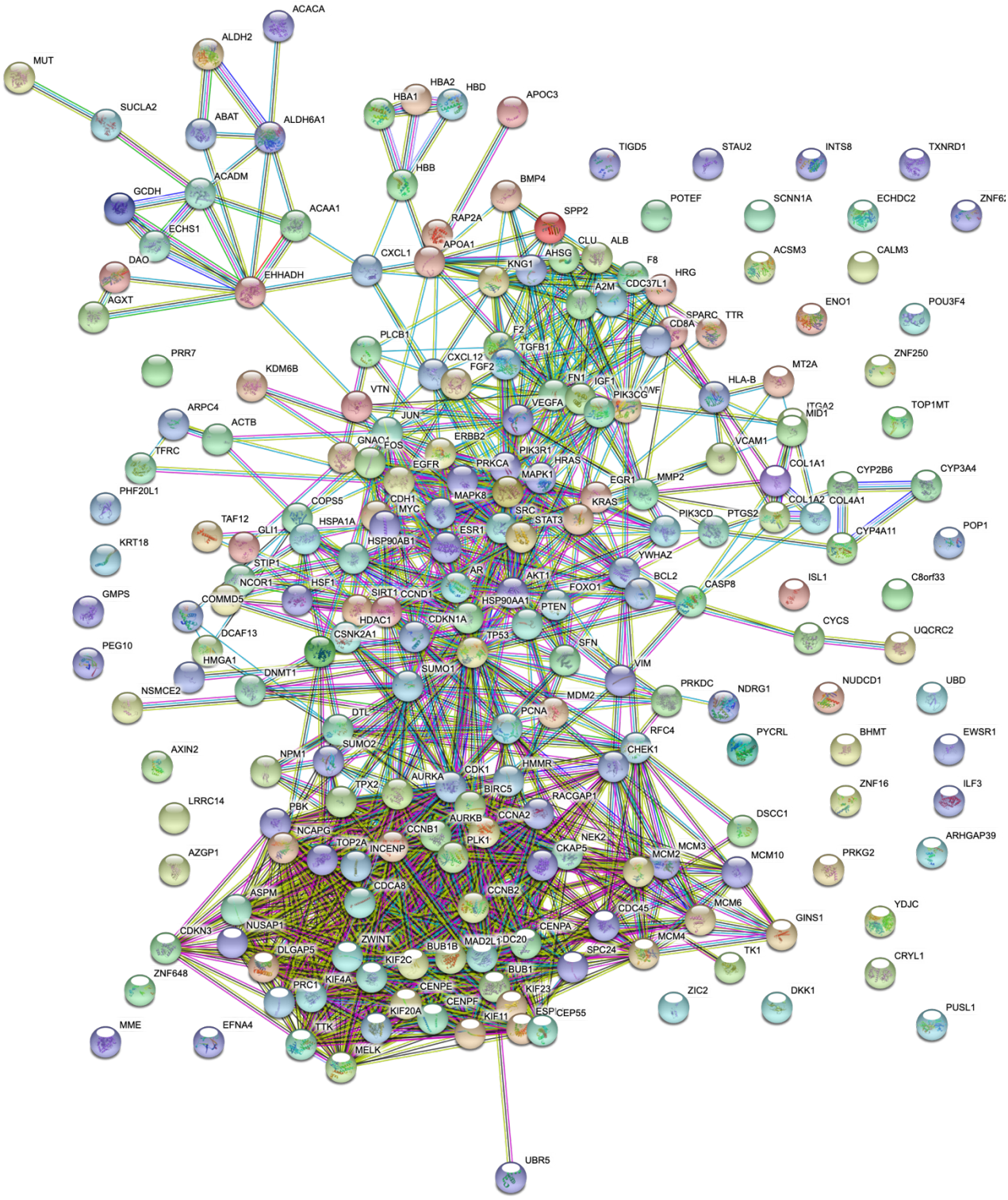

Figure3S: PPI networks of 221 hub proteins showing the experimentally validated interactions with. Confidence score of 0.4, Figure 1C in main text

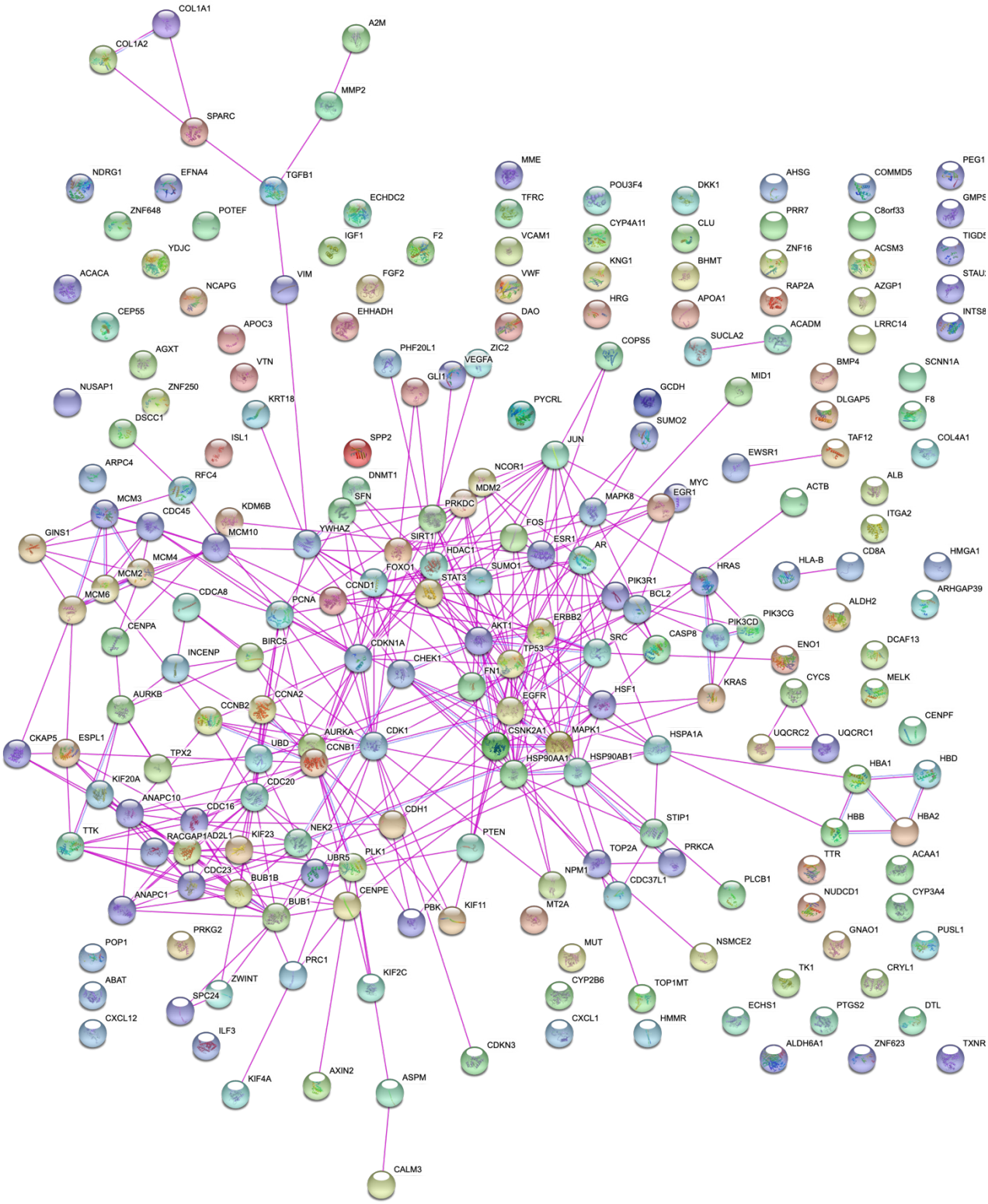

Figure 4S: PPI networks of 221 hub proteins showing the experimentally validated interactions with. Confidence score of 0.9, Figure 1D in main text