

Linking Bioinformatics Research data and Publications through Metadata and Knowledge Organization Systems

Jian Qin

Syracuse University

Marcia L. Zeng

Kent State University

KOS vocabularies for representing data and publication

Facts about organisms

Knowledge derived from studying the facts about organisms

Taxonomy

Data



Subject headings
Thesauri
Classification

Publication



LOCUS	SCU49845	5028 bp	DNA
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene (AXL2) and Rev7p (REV7) genes, complete genome.		
ACCESSION	U49845		
VERSION	U49845.1 GI:1293613		
KEYWORDS	.		
SOURCE	Saccharomyces cerevisiae (baker's yeast)		
ORGANISM	Saccharomyces cerevisiae Eukaryota ; Fungi ; Ascomycota ; Saccharomycetales ; Saccharomycetaceae ;		

Example: Taxonomic representation of a DNA sequence dataset in GenBank that **documents the organism** in the form of taxon lineage

Example: Subject representation of the publication related to the DNA sequence dataset (PubMed ID: 7871890), which strives to **provide as many and exhaustive access points as possible**

MeSH Terms

[Amino Acid Sequence](#)

[Base Sequence](#)

[Chromosomes, Fungal](#)

[Cloning, Molecular](#)

[DNA Damage*](#)

[DNA Replication](#)

[DNA, Fungal/biosynthesis](#)

[DNA, Fungal/secretion](#)

[DNA-Directed DNA Polymerase*](#)

[Fungal Proteins/chemistry](#)

[Fungal Proteins/genetics*](#)

[Genes, Fungal*](#)

[Genetic Complementation Test](#)

[Molecular Sequence Data](#)

[Mutagenesis*](#)

[Open Reading Frames](#)

[Saccharomyces cerevisiae/chemistry](#)

[Saccharomyces cerevisiae/genetics*](#)

[Saccharomyces cerevisiae Proteins*](#)

[Sequence Analysis, DNA](#)

[Sequence Homology, Amino Acid](#)

Substances

[DNA, Fungal](#)

[Fungal Proteins](#)

[REV7 protein, S cerevisiae](#)

[Saccharomyces cerevisiae Proteins](#)

[DNA-Directed DNA Polymerase](#)

Why should we care about linking data to publications?

- Evidence on which the publication is based, i.e., validity and verifiability
- Reproducibility of research
- Reuse and sharing of data more easily

Content representations for data and publications are different in terms of

1. Scope and coverage
2. Focuses or application practices
3. Ability and mechanisms for integrating biomedical research data

1. Scope and coverage

International
Disease (ICD)
cancer: primary
organs or sites

Malignant neoplasms
(C50-C50)

C50 Malig.
Incl.
Excl.

C50.0 Nipp

C50.1 Cent

C50.2 Upp

C50.3 Low

C50.4 Upp

C50.5 Lower-outer quadrant of breast

C50.6 Axillary tail of breast

C50.8 Overlapping lesion of breast

[See note 5 at the beginning of this chapter]

C50.9 Breast, unspecified

MeSH Heading	Breast Neoplasms	Source: https://www.nlm.nih.gov/cgi/mesh/2016/M_B.cgi?mode=&index=2282&view=concept
Tree Number	C04.588.180	
Tree Number	C17.800.090.500	
Annotation	human only; BREAST NEOPLASMS MAMMARY NEOPLASMS, ANATOMICAL EXPERIMENTAL: Manual 24.5+ neoplasm (IM)	
Concept 1 (Preferred)	Breast Neoplasms	
	Scope Note	Tumors or cancer
	Term	Breast Neoplasms
	Term	Breast Tumors
	Term	Neoplasms, Breasts
	Term	Tumors, Breast
Allowable Qualifiers	BL BS CF CH CI CL CN CO DH MO NU PA PC PP PS PX RA RH	
	Outer quadrant of breast	
	Lower-outer quadrant of breast	
	Axillary tail of breast	
	Overlapping lesion of breast	
	Breast, unspecified	

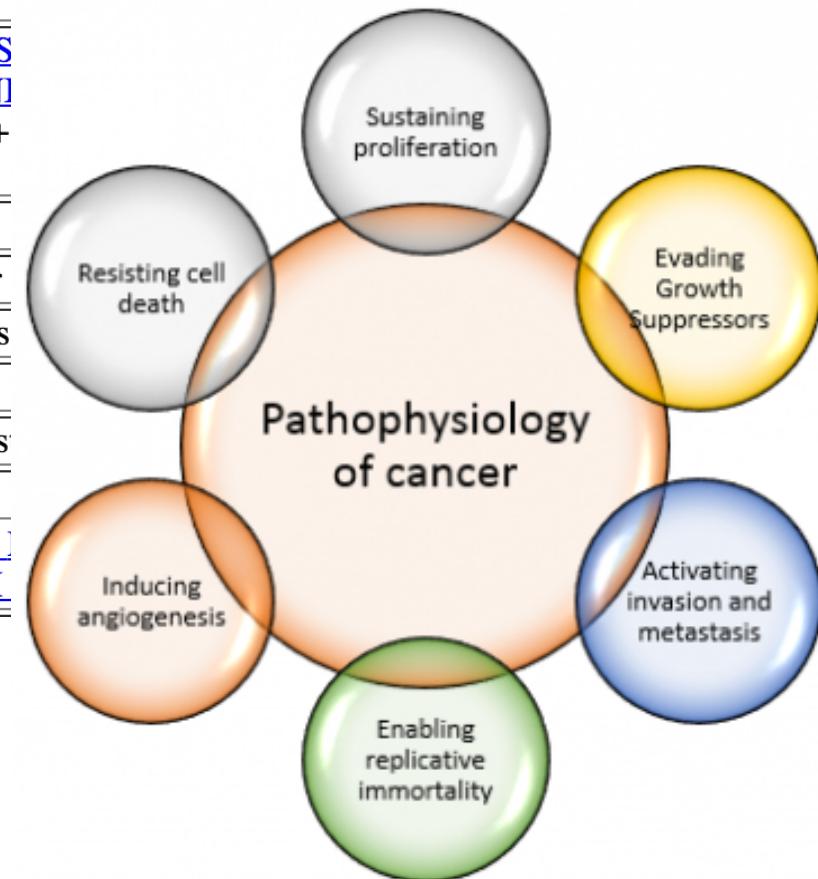


Image credit: http://www.physio-pedia.com/Physiotherapy_and_cancer_treatment

Constraints of conventional KOS vocabularies on coverage and scope

Coarse granularity on representing concepts and relationships

Covert relationships between concepts

Documentation of information about a concept

2. Focuses or application practices

Organizing knowledge of organisms

- Applying scientific taxonomy and nomenclature to
 - identify,
 - name, and
 - classify them
- in bioinformatics data, and
- in the metadata that describes such data.

Examples

- NCBI* Taxonomy
- NCBI Organismal Classification

*NCBI=National Center for Biological Information

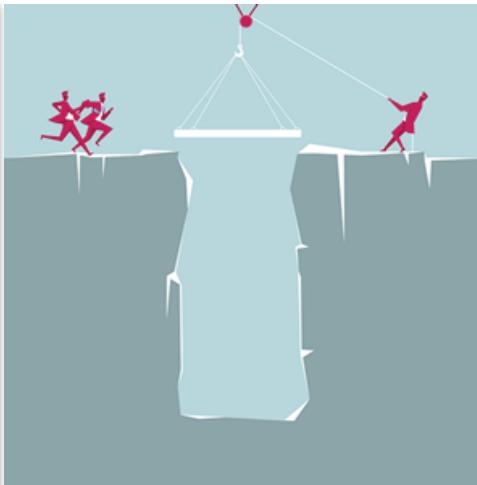


Image credit:
<https://blogs.cfainstitute.org/investor/files/2015/06/How-Financial-Advisers-Can-Help-Close-the-Behavior-Gap.png>

Organizing information and knowledge contained in research publications

- Applying thesauri and classifications to
 - index,
 - retrieve,
 - organize, and
 - connect
- the scholarly output from studying the organisms, and
- the scholarly output in regulation and guideline documents.

Examples

- *Medical Subject Headings (MeSH)*
- *NCI* Thesaurus (NCIt)*

*NCI = National Cancer Institute

3. Ability and mechanisms for integrating biomedical research data

• NCBI Organismal Classification

*NCBI=National Center for Biological Information

Hepatitis C virus

Taxonomy ID: 11103

Inherited blast name: viruses

Rank: species

Genetic code: [Translation table 1 \(Standard\)](#)

Host: vertebrates\ human

Other names:

synonym: post-transfusion hepatitis non A non B virus

synonym: human hepatitis virus C HCV

synonym: human hepatitis C virus HCV

synonym: human hepatitis C virus

synonym: hepatitis C virus HCV

acronym: HCV

misnomer: human hepatitis virus HCV

misnomer: Hepatitis C

Lineage (full)

[Viruses](#); [ssRNA viruses](#); [ssRNA positive-strand viruses](#),
[no DNA stage](#); [Flaviviridae](#); [Hepacivirus](#)

Attributes: term ID, inherited blast name, rank, genetic code, other name, host, and Lineage.

Source:

<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=hepatitis+C+virus&lvl=0&srchmode=1>

Entrez records		Genome Information					
Database name	Subtree links	TTT	F Phe	TCT	S Ser	TAT	Y Tyr
Nucleotide	201,670	TTC	E Phe	TCC	S Ser	TAC	V Tyr
Protein	175,322						
Structure	382						
Genome	1						
Popset	1,415	PCR		ALL			
Domains	6	BI - Broad Institute					
GEO Datasets	45	300,700		300,700			
PubMed Central	21,599						
Gene	14						
SRA Experiments	2,267						
Probe	464						
Assembly	7						
Bio Project	41						
Bio Sample	1,948						
PubChem BioAssay	3,249						
Taxonomy	192						
Trace records (raw single-pass reads of DNA sequence)							
Sequencing Center Name							
Record counts per type							
	PCR		ALL				
	BI - Broad Institute						
	300,700		300,700				
Totals per type							
	300,700		300,700				
External Information Resources (NCBI LinkOut)							
LinkOut	Subject	LinkOut Provider					
dryadb	supplemental materials	Dryad Digital Repository					
Hepatitis C virus	taxonomy/phylogenetic	Encyclopedia of Life					
GOLDCARD: Gc0065435	organism-specific	Genomes Online					
GOLDCARD: Gi0065295	organism-specific	Global Biocollections					
Show Biotic Interactions	taxonomy/phylogenetic	Global Biotic Interactions					
Related Immune Epitope Information	gene/protein/disease-specific	Immune Epitope Database					
Hepatitis C virus	taxonomy/phylogenetic	International Committee on Taxonomy of Viruses					
euHCVdb (Europe)	taxonomy/phylogenetic	NCBI taxonomy					
LANL HCV Database	taxonomy/phylogenetic	VirOligo Collection					
Hepatitis Virus Database (Japan)	taxonomy/phylogenetic	VirOligo Collection					
VirOligo	dna/protein sequence	VirOligo Collection					
VirOligo	dna/protein sequence	VirOligo Collection					

TGT C Cys
TGC C Cys
TGA * Ter
TGG W Trp

CGT R Arg
CGC R Arg
CGA R Arg
CGG R Arg

AGT S Ser
AGC S Ser
AGA R Arg
AGG R Arg

GGT G Gly
GGC G Gly
GGA G Gly
GGG G Gly

3. Ability and mechanisms for integrating biomedical research data

NCI Thesaurus (NCIt)

*NCI = National Cancer Institute

Hepatitis C Virus (Code C14312)

Terms & Properties Synonym Details Relationships Mappings

Table of Contents

- Terms & Properties
- Synonym Details
- Relationships
- Mapping Details

Terms & Properties

Preferred Name: Hepatitis C Virus

Definition: A small, enveloped, positive sense single strand RNA virus in the family Flaviviridae.

CDISC Definition: Any viral organism that can be assigned to the species Hepatitis C Virus.

Label: Hepatitis C Virus

NCI Thesaurus Code: C14312 ([Search for linked caDSR metadata](#)) ([search value](#))

NCI Metathesaurus Link: C0220847 ([see NCI Metathesaurus info](#))

Synonyms & Abbreviations: ([see Synonym Details](#))

HCV

Hepatitis C

HEPATITIS C VIRUS

Hepatitis C Virus

Virus-Hepatitis C

External Source Codes:

Source: C0220847

https://ncit.nci.nih.gov/ncitbrowser/pages/concept_details.jsf?dictionary=NCI_Thesaurus&version=16.09d&code=C14312&ns=NCI_Thesaurus&type=all&key=n1875063326&b=1&n=0&vse=null

Synonym Details

Term	Source	Type
HCV	NCI	AB
HCV	CDISC	SY
Hepatitis C	NCI	SY
HEPATITIS C VIRUS	CDISC	PT
Hepatitis C Virus	NCI	PT
Virus-Hepatitis C	NCI	SY

Relationships with other NCI Thesaurus Concepts

Parent Concepts:

[Hepacivirus](#)

[Hepatitis Virus](#)

Child Concepts: (none)

Role Relationships pointing from the current concept to other concepts: (none)

Associations pointing from the current concept to other concepts: (True for the current concept.)

Relationship

Concept_In_Subset	Value (qualifiers indented underneath)
Concept_In_Subset	CDISC SDTM Microorganism Terminology
Concept_In_Subset	CDISC SDTM Species Terminology
Concept_In_Subset	CDISC SDTM Terminology
Concept_In_Subset	

Concept_In_Subset

Incoming Role Relationships pointing from other concepts:

Incoming Associations pointing from other concepts:

Mapping relationships:

[see Mappings](#)

Mapping Details

Mapping through NCI Metathesaurus:

[C0220847](#)

Hepatitis C Virus (CUI C0220847)

Terms & Properties Synonym Details Relationships By Source View All

Synonym Details:

Term	Source	Type	Code
HCV - Hepatitis C virus	SNOMEDCT_US	SY	62944002
HCV	CDISC	SY	C14312
HCV	CSP	ET	3108-4622
HCV	CST	GT	HEPATITIS C
HCV	MDR	OL	10019183
HCV	MEDLINEPLUS	SY	1286
HCV	NCI	AB	C14312
HCV	MSH	PM	D016174
HCV	MDR	OL	10019752
HCV	SNOMEDCT_US	FN	62944002
HCV	NCBI	SY	11103
HCV	AOD	DE	0000016071
HCV	CDISC	PT	C14312
HCV	CSP	PT	3108-4622
HCV	CST	PT	HEPATITIS C
HCV	MDR	OL	10019751
HCV	MSH	PEP	D016174
HCV	MTHSP	SU	Q156415283
HCV	MTH	PN	NOCODE
HCV	NCBI	SCN	11103
HCV	NCI	PT	C14312
HCV	RXNORM	IN	1491963
HCV	SNOMEDCT_US	PT	62944002
HCV	MEDLINEPLUS	PT	1286
HCV	NCI	SY	C14312
human hepatitis C virus HCV	NCBI	SY	11103
human hepatitis C virus	NCBI	SY	11103
human hepatitis virus C HCV	NCBI	SY	11103
post-transfusion hepatitis non A non B virus	NCBI	SY	11103
Virus-Hepatitis C	NCI	SY	C14312

Ways of linking data to publication

- Identifiers



Object-to-object linking

- Semantic relationships

- KOS crosswalk

Concept-to-concept linking

- Co-indexing terms

Label-to-term linking

- Knowledge networks

Node-to-node linking

Object-to-object linking

LOCUS	SCU49845	5028 bp	DNA	PLN	21-JUN-1999
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p (AXL2) and Rev7p (REV7) genes, complete cds.				
ACCESSION	U49845				
VERSION	U49845.1 GI:1293613				
KEYWORDS	.				
SOURCE	Saccharomyces cerevisiae (baker's yeast)				
ORGANISM	Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.				
REFERENCE	1 (bases 1 to 5028)				
AUTHORS	Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.				
TITLE	Cloning and sequence analysis of the REV7 gene whose function is required for DNA damage-induced mutagenesis in <i>Saccharomyces cerevisiae</i>				
JOURNAL	Yeast 10 (11), 1503-1512				
PUBMED	7871890				
REFERENCE	2 (bases 1 to 5028)				
AUTHORS	Roemer,T., Madden,K.				
TITLE	Selection of axial genes from the yeast genome: Axl2p, a novel plasma membrane glycoprotein involved in the regulation of gene expression				
JOURNAL	Genes Dev. 10 (7), 771-783				
PUBMED	8846915				
REFERENCE	3 (bases 1 to 5028)				
AUTHORS	Roemer,T.				
TITLE	Direct Submission				
JOURNAL	Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New Haven, CT, USA				

A publication ID from PubMed is embedded in the dataset's metadata record

Hepatitis C Virus (CUI C0220847)

Many KOS vocabs already exist,
maybe mapped...

[Suggest ch](#)
[Terms & Properties](#)
[Synonym Details](#)
[Relationships](#)
[By Source](#)
[View All](#)
Synonym Details:
Term

HCV - Hepatitis C virus

HCV

HCV

HCV

HCV

HCV

Hepatitis C viruses

Hepatitis C virus (HCV)

Hepatitis C virus (organism)

hepatitis C virus HCV

hepatitis C virus

Hepatitis C

Hepatitis C

human hepatitis C virus HCV

human hepatitis C virus

human hepatitis virus C HCV

post-transfusion hepatitis non A non B virus

Virus-Hepatitis C

Concept-to-
concept linking

Source	Type	Code
SNOMEDCT_US	SY	62944002
CDISC	SY	C14312
CSP	ET	3108-4622
CST	GT	HEPATITIS C
MDR	OL	10019183
MEDLINEPLUS	SY	1286
NCI	AB	C14312
MSH	PM	D016174
MDR	OL	10019752
SNOMEDCT_US	FN	62944002
NCBI	SY	11103
AOD	DE	0000016071
CDISC	PT	C14312
CSP	PT	3108-4622
CST	PT	HEPATITIS C
MDR	OL	10019751
MSH	PEP	D016174
MTHSPL	SU	QI56415283
MTH	PN	NOCODE
NCBI	SCN	11103
NCI	PT	C14312
RXNORM	IN	1491863
SNOMEDCT_US	PT	62944002
MEDLINEPLUS	PT	1286
NCI	SY	C14312
NCBI	SY	11103
NCI	SY	C14312

<https://ncim.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI%20Metathesaurus&type=synonym&code=C0220847>

LOCUS	SCU49845	5028 bp	DNA
DEFINITION	Saccharomyces cerevisiae TCP1-beta gene (AXL2) and Rev7p (REV7) genes, complete		
ACCESSION	U49845		
VERSION	U49845.1	GI:1293613	
KEYWORDS	.		
SOURCE	Saccharomyces cerevisiae (baker's yeast)		
ORGANISM	Saccharomyces cerevisiae Eukaryota; Fungi; Ascomycota; Saccharomycetales; Saccharomycetaceae;		

Metadata for a DNA sequence dataset in the **GenBank data repository**

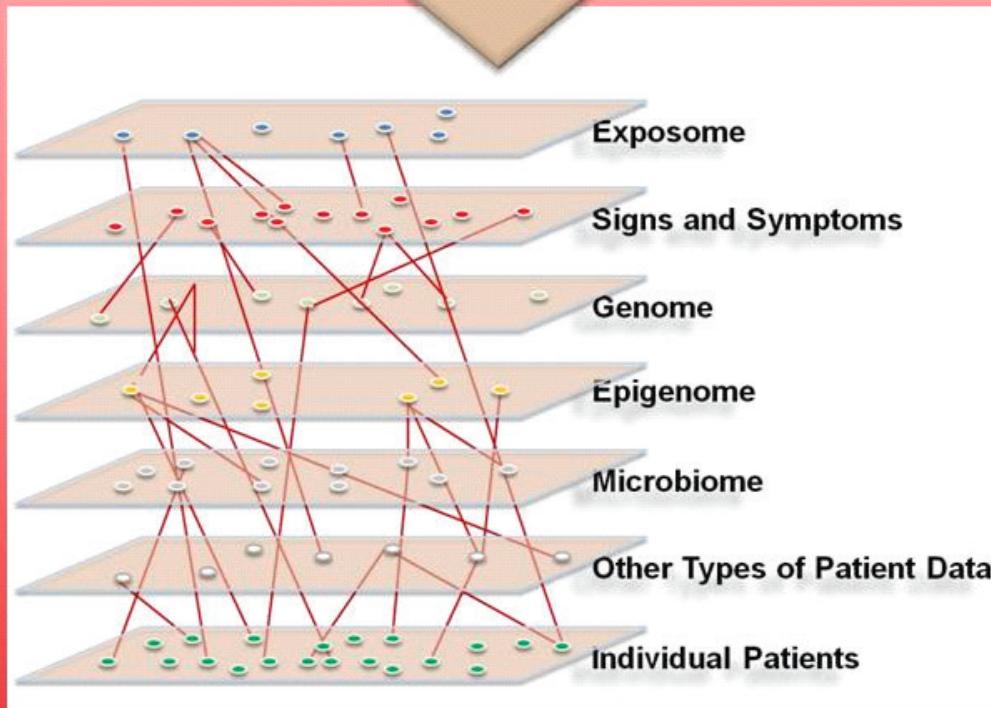
Label-to-term linking

Indexing terms in **PubMed** for the paper that resulted from studying the DNA sequence

MeSH Terms
Amino Acid Sequence
Base Sequence
Chromosomes, Fungal
Cloning, Molecular
DNA Damage*
DNA Replication
DNA, Fungal/biosynthesis
DNA, Fungal/secretion
DNA-Directed DNA Polymerase*
Fungal Proteins/chemistry
Fungal Proteins/genetics*
Genes, Fungal*
Genetic Complementation Test
Molecular Sequence Data
Mutagenesis*
Open Reading Frames
Saccharomyces cerevisiae/chemistry
Saccharomyces cerevisiae/genetics*
Saccharomyces cerevisiae Proteins*
Sequence Analysis, DNA
Sequence Homology, Amino Acid
Substances
DNA, Fungal
Fungal Proteins
REV7 protein, S cerevisiae
Saccharomyces cerevisiae Proteins
DNA-Directed DNA Polymerase

Information Commons

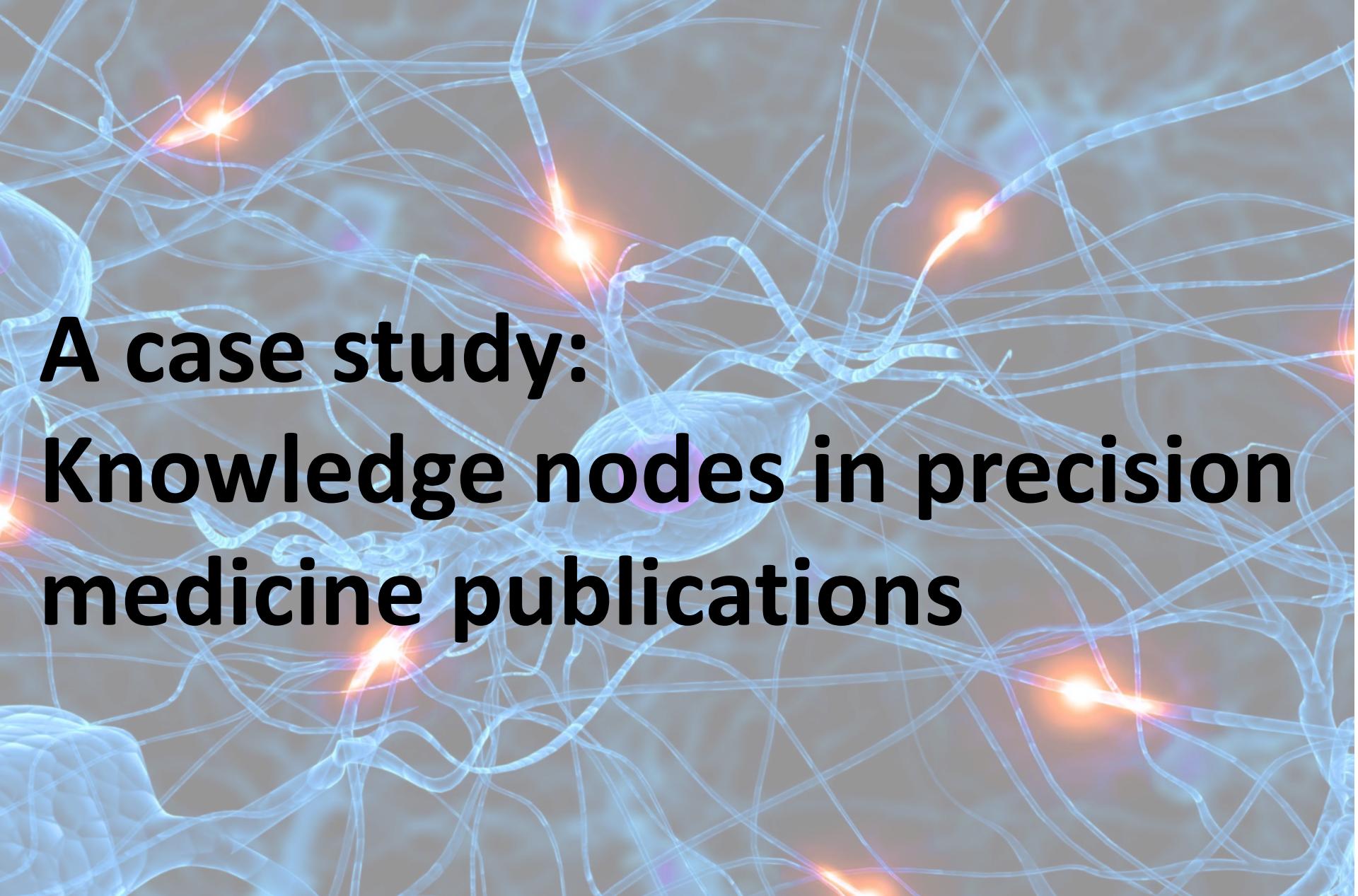
Node-to-node linking



(Source: NAS, 2011)

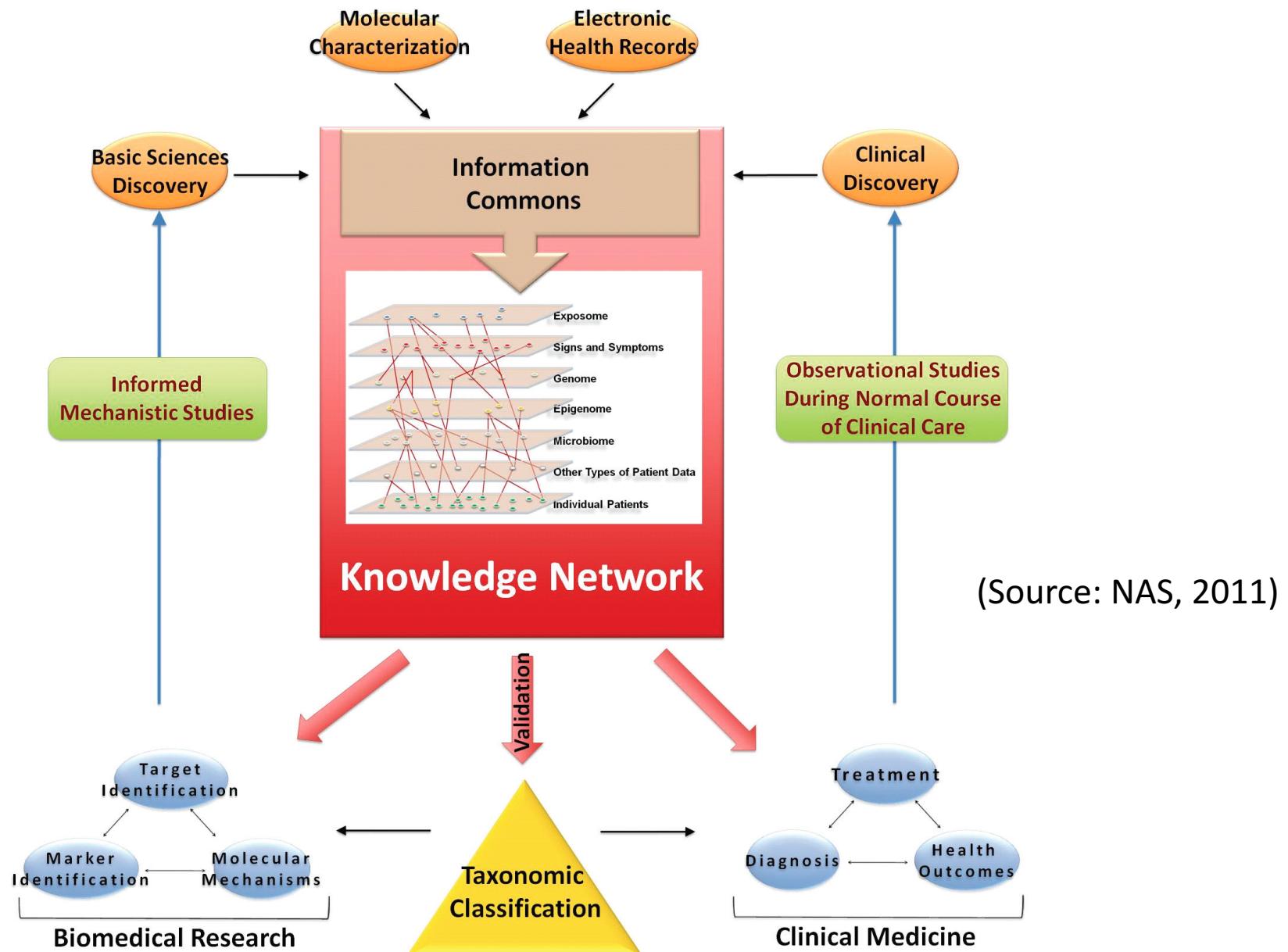
Knowledge Network

**The next question is:
How are we going to
create the links?**



A case study: Knowledge nodes in precision medicine publications

The vision of a Knowledge Network of Disease and Information Commons



Research problem

“Because **new information and concepts from biomedical research** cannot be **optimally incorporated** into the **disease taxonomy of today**, opportunities to define diseases more precisely and to **inform health-care decisions** are being missed.”

(Source: NAS, 2011)

Data from
basic research



Represented by
taxonomic classes

How can biomedical
research be optimally
incorporated into the
disease taxonomy of today?

Clinical practice



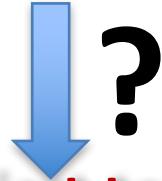
Coded by International
Classification of Diseases

Approach to address the research problem



Data

Attributes of data
(metadata)



Object-to-object linking
Concept-to-concept linking
Label-to-term linking
Node-to-node linking

New information and
discovery is reported in

Publication



Identify from publications

Knowledge nodes:

- Types?
- Attributes?

Relationships between nodes:

- Types?
- Attributes?

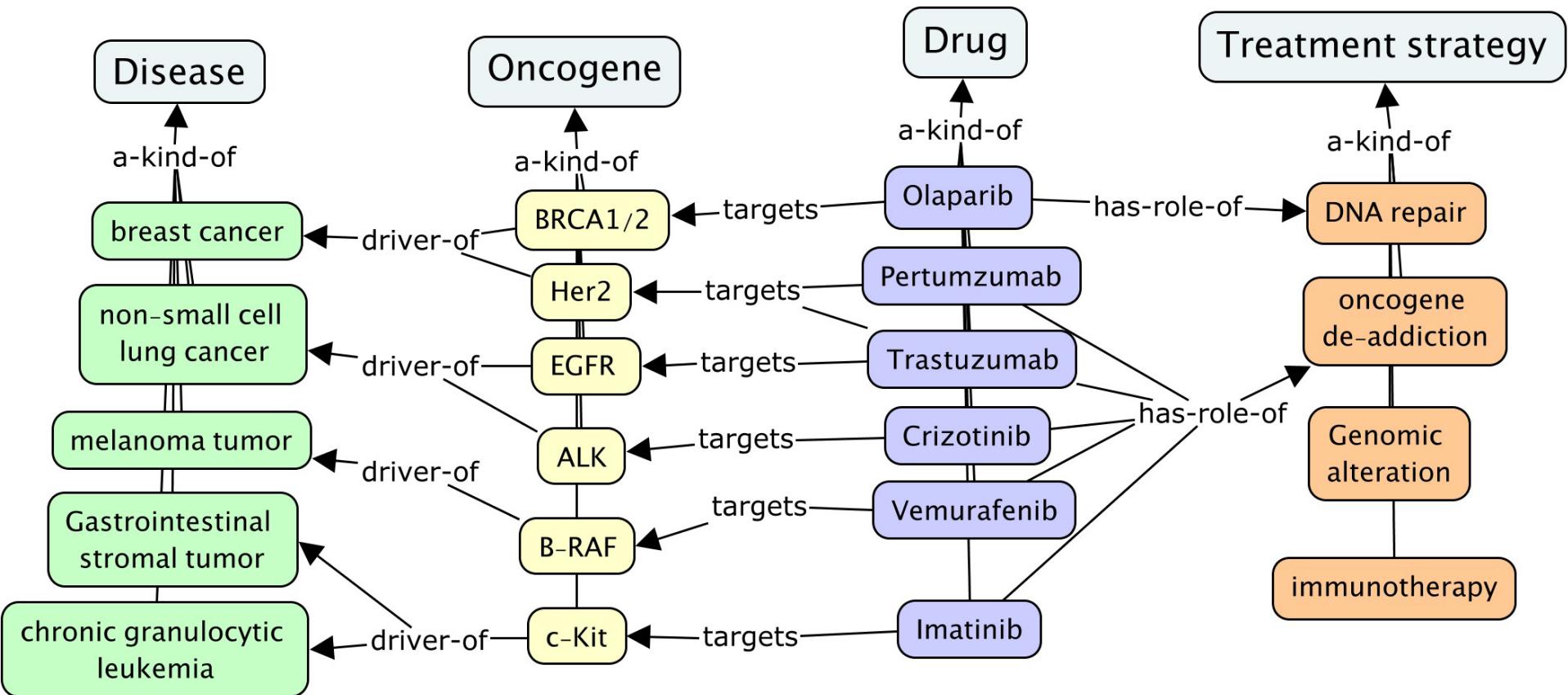
Pilot study: data

- A sample of 30 articles in precision medicine
 - Four in breast cancer
 - Five in diabetes
 - Eleven in oncology
- “Purposeful sampling”
 - To gain insights and in-depth understanding rather than empirical generalizations

Pilot study: Selecting knowledge nodes

- Molecular entities such as genes, proteins, genomes, etc.
- Disease names
- Names or terms related to treatments/therapies
- Methods, techniques, and types of decisions related to diagnosis
- Data sources used by the publication
- Types of relationships between potential knowledge nodes

Pilot study: Mapping knowledge nodes



A sample map of knowledge nodes and relationships from a research paper
(based on PubMed paper ID 25441102)

Pilot study: Preliminary results (1)

- Structural levels of nodes

Examples of knowledge nodes derived from the sample publications

Category	Atomic level (name of things)	Concept level	Cluster level
Gene	Her2, BRCA1, BRCA2, EGFR	Oncogenes	EGFR mutations in lung cancer
Disease	Non-squamous carcinoma, squamous cell carcinoma	Non-small cell lung cancer	Lung cancer
Drug	Pertumzumab, Lmatinib, Crizotinib	Tyrosine kinase inhibitor	Oncogene de-addiction

Pilot study: Preliminary results (2)

- Knowledge nodes by
 - Disciplinary field:
genetics, pathology, pathophysiology, oncology, virology, ...
 - Disease name and biomarker pairs:
 - Chronic myeloid leukemia (CML) with mutated gene **BCR-ABL**
 - Breast cancer with **positive estrogen receptor (ER)**, **BRCA1/2**, and **Her2**
 - **Non-small cell lung cancer** with mutations in multiple genes such as **epidermal growth factor receptor (EGFR)**, **excision repair-cross complementation group (ERCC)**, and **ribonucleotide reductase (RRM)**

Pilot study: Preliminary results (3)

- Knowledge nodes that blend clinical and basic research
 - clinically actionable mutations
 - phenotype of breast cancer
 - resistance to endocrine therapy
 - biomarkers predicting response to therapy
 - genomic drivers of cancer
 - predictive and prognostic biomarkers
 - intratumor heterogeneity
 - molecular classification of tumors

Pilot study: Preliminary results (4)

Major relationships types and patterns between knowledge nodes observed in the sample publications

Relationship	Pattern	Example
has-biomarker	Disease has-biomarker Gene	chronic myeloid leukemia has-biomarker BCR-ABL non-small cell lung cancer has-biomarker EGFR
is-driver-of	Gene is-driver-of Disease	Her2 is-driver-of breast cancer c-Kit is-driver-of chronic granulocytic leukemia
targets	Drug targets Gene	Crizotinib targets ALK Olaparib targets BRCA1/2
has-role-of	Drug has-role-of Treatment	Crizotinib has-role-of oncogene de-addiction Olaparib has-role-of DNA repair

Implications of preliminary results

- Knowledge nodes may be marked with different labels—structure, discipline, disease, gene or biomarker, treatment, ...
- Each label represents a dimension and the nodes in one dimension form a vector
- A node may reside in multiple dimensions at the same time
- The knowledge network of disease can be considered as the sum of nodes in all vectors, which becomes a data science research problem

Concluding remarks

- Linking between data and publications requires reexamining the data and knowledge landscape and renew our understanding of KOS in the context of data-intensive science
- New types of KOS need to be dynamic, flexible, and linkable
- Models, patterns, and computational algorithms will be needed to develop the knowledge network of disease that incorporates basic science with clinical practice

References

- NAS. (2011). Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease . Washington, D.C.: The National Academies Press. <https://www.nap.edu/catalog/13284/toward-precision-medicine-building-a-knowledge-network-for-biomedical-research>