



SWPM 2009

Workshop on Semantic Web
and Provenance Management

Westfields Conference Center, Washington D.C., USA.
October 25, 2009

Provenance information
in biomedical knowledge repositories
A use case



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Advanced Library Services project

◆ Biomedical Knowledge Repository

- Knowledge extracted from
 - Textual sources (e.g., biomedical literature) using Natural Language Processing (NLP) techniques
 - Structured knowledge bases (e.g., Entrez)
 - Terminological resources (e.g., UMLS)

◆ Support services including

- Enhanced information retrieval
- Multi-document summarization
- Question answering
- Knowledge discovery



Outline

- ◆ Examples of provenance information in biomedical knowledge bases
- ◆ Examples of applications requiring provenance information
- ◆ Issues and challenges

Examples of provenance information in biomedical knowledge bases

References for the examples

◆ Entrez System

National Center for Biotechnology Information (NCBI)

- Entrez Gene

- <http://www.ncbi.nlm.nih.gov/gene/7068>

- PubMed

- <http://www.ncbi.nlm.nih.gov/pubmed/17177139>

◆ Mouse Genome Informatics (MGI)

The Jackson Laboratory

- Mammalian Orthology

- http://www.informatics.jax.org/searches/homology_form.shtml




 Search Gene for Go Clear

Limits Preview/Index History Clipboard Details

Display Full Report Show 20 Sort by Relevance Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

 1: **THRB thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)** [*Homo sapiens*]

GeneID: 7068

updated 17-Oct-2009

Summary

Official Symbol THRB

 provided by [HGNC](#)
Official Full Name thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)

 provided by [HGNC](#)
Primary source [HGNC:11799](#)
See related [Ensembl:ENSG00000151090](#); [HPRD:07521](#); [MIM:190160](#)
Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)
Lineage *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as GRTH; PRTH; THR1; ERBA2; NR1A2; THRB1; THRB2; ERBA-BETA; MGC126109; MGC126110; THRB

NCB

All Database

Search Gene

Limits Prev

Display Full Re

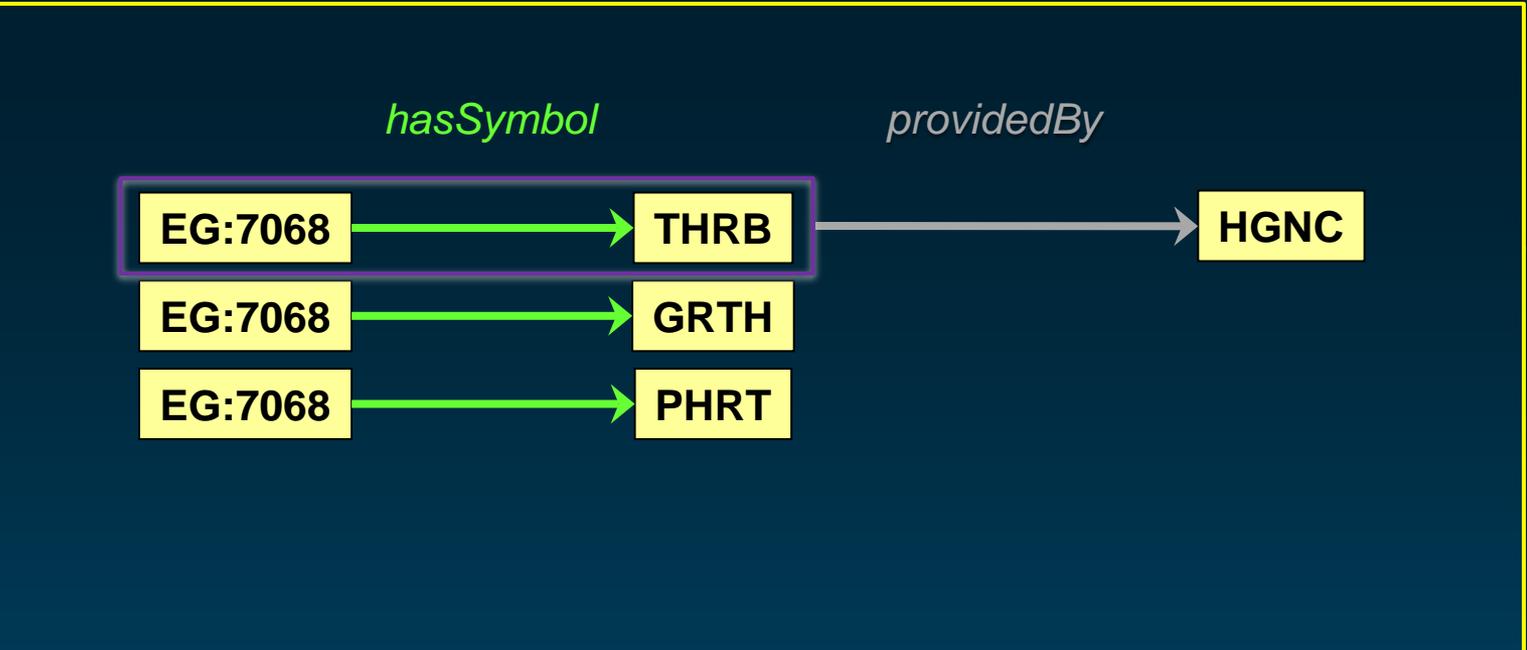
All: 1 Curre

1: THRB th

Homo sapiens

GeneID: 7068

Summary



Official Symbol THRB provided by [HGNC](#)

Official Full Name thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian) provided by [HGNC](#)

Primary source [HGNC:11799](#)

See related [Ensembl:ENSG00000151090](#); [HPRD:07521](#); [MIM:190160](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*

Also known as GRTH; PRTH; THR1; ERBA2; NR1A2; THR1; THR2; ERBA-BETA; MGC126109; MGC126110; THR



Interactions



Description

Product	Interactant	Other Gene	Complex	Source	Pubs
TR-beta interacts with pp32.					
NP_000452.2	NP_006296.1	ANP32A		BIND	PubMed
BTG1 interacts with THRB (TR-beta-1). This interaction was modeled on a demonstrated interaction between human BTG1 and TR-beta-1 from an unspecified species.					
NP_000452.2	NP_001722.1	BTG1		BIND	PubMed
NP_000452.2	Thyroid hormone receptor interactor 15	COPS2		HPRD	PubMed
TR interacts with CTCF.					
NP_000452.2	NP_006556.1	CTCF		BIND	PubMed
NP_000452.2	NP_003874.2	HDAC3		HPRD	PubMed
NP_000452.2	Thyroid hormone receptor interactor 8	JMJD1C		HPRD	PubMed
NP_000452.2	TRAP220	MED1		HPRD	PubMed
TR-beta interacts with hSrb7.					
NP_000452.2	NP_004255.2	MED21		BIND	PubMed
MPG interacts with THRB (TR-beta). This interaction was modeled on a demonstrated interaction between MPG from an unspecified species and THRB from an unspecified species.					
NP_000452.2	NP_002425.1	MPG		BIND	PubMed

Interactions

Description

Product	Interactant	Other Gene	Complex	Source	Pubs
TR-beta interacts with pp32.					
NP_000452.2	NP_006296.1	ANP32A		BIND	PubMed
BTG1 interacts with THRB (TR-beta-1). This interaction was modeled on a demonstrated interaction between human BTG1 and TR-beta-1 from an unspecified species.					
NP_000452.2	NP_001722.1	BTG1		BIND	PubMed
NP_000452.2	Thyroid hormone receptor interactor 15	COPS2		HPRD	PubMed

TR interacts

NP_000452.

NP_000452.

NP_000452.

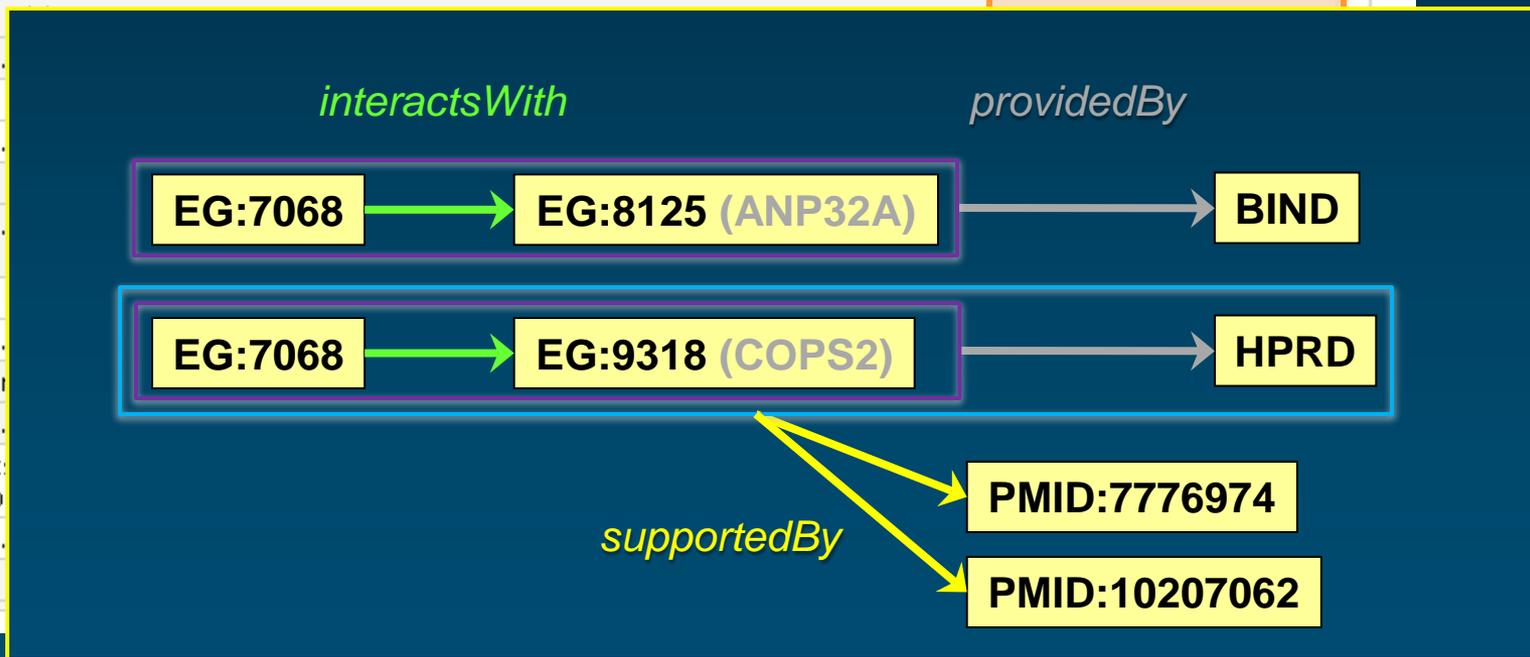
NP_000452.

TR-beta inter

NP_000452.

MPG interact from an unsp

NP_000452.



1: THRB thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian) [*Homo sapiens*]

GeneID: 7068

updated 17-Oct-2009

GeneOntology

Provided by [GOA](#)

Function	Evidence
metal ion binding	IEA
protein binding	IPI PubMed
sequence-specific DNA binding	IEA
steroid hormone receptor activity	IEA
thyroid hormone receptor activity	TAS PubMed
transcription corepressor activity	TAS PubMed
transcription factor activity	NAS PubMed
zinc ion binding	IEA
Process	Evidence
regulation of transcription, DNA-dependent	IEA
Component	Evidence
nucleus	TAS PubMed

1: **THRB thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)** [*Homo sapiens*]

GeneID: 7068

updated 17-Oct-2009

GeneOntology

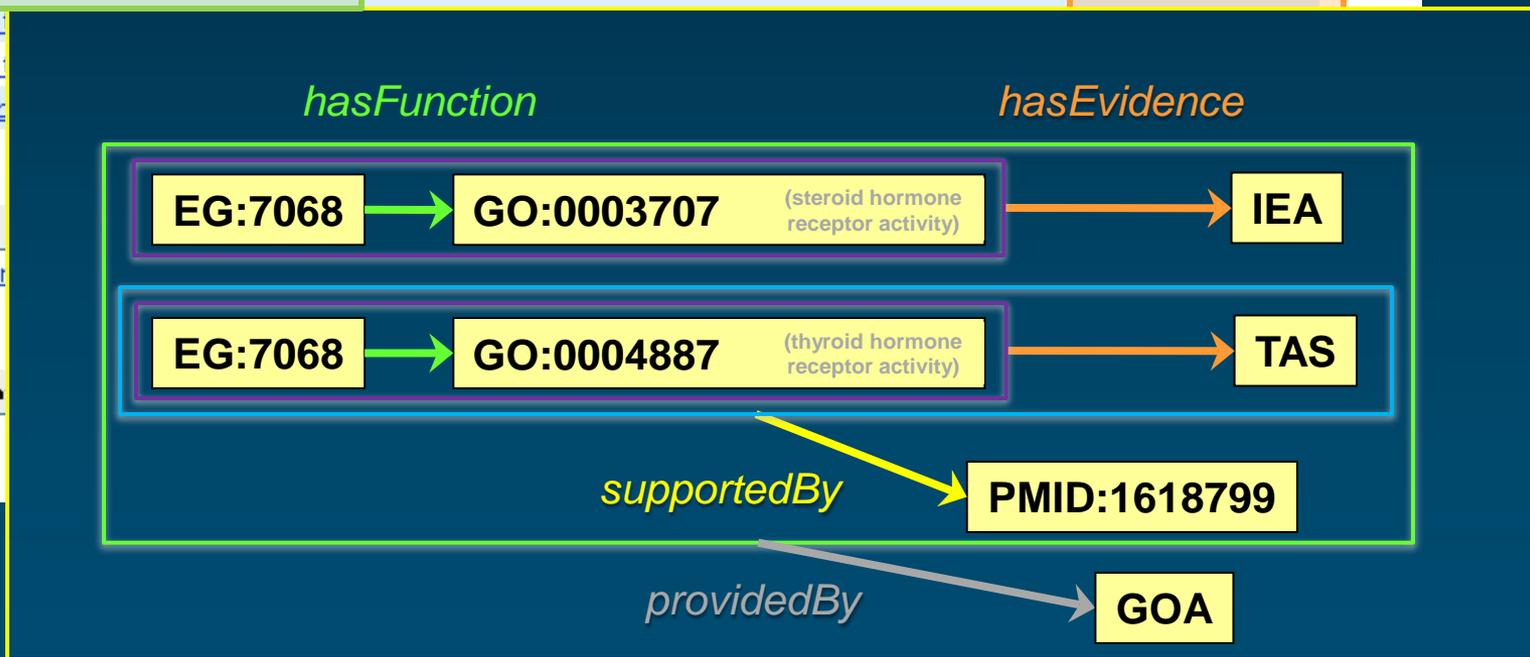
Provided by [GOA](#)

Function	Evidence
metal ion binding	IEA
protein binding	IPI PubMed
sequence-specific DNA binding	IEA
steroid hormone receptor activity	IEA
thyroid hormone receptor activity	TAS PubMed

[transcription](#)
[transcription](#)
[zinc ion b](#)

Process
[regulation](#)

Component
[nucleus](#)





Mammalian Orthology

Query Results

You searched for...

Organism(s): equals *mouse, laboratory*

Marker Symbol/Name: contains *thrb* searching current symbols/names and synonyms.

Comparison Organism: equals *human* searching only selected species.

Sort: by *Marker in primary species*

Display Limit: equals *500*

1 matching item displayed

You may select one or more sequences to download in FASTA format or forward to MouseBLAST.

Select all

Invert

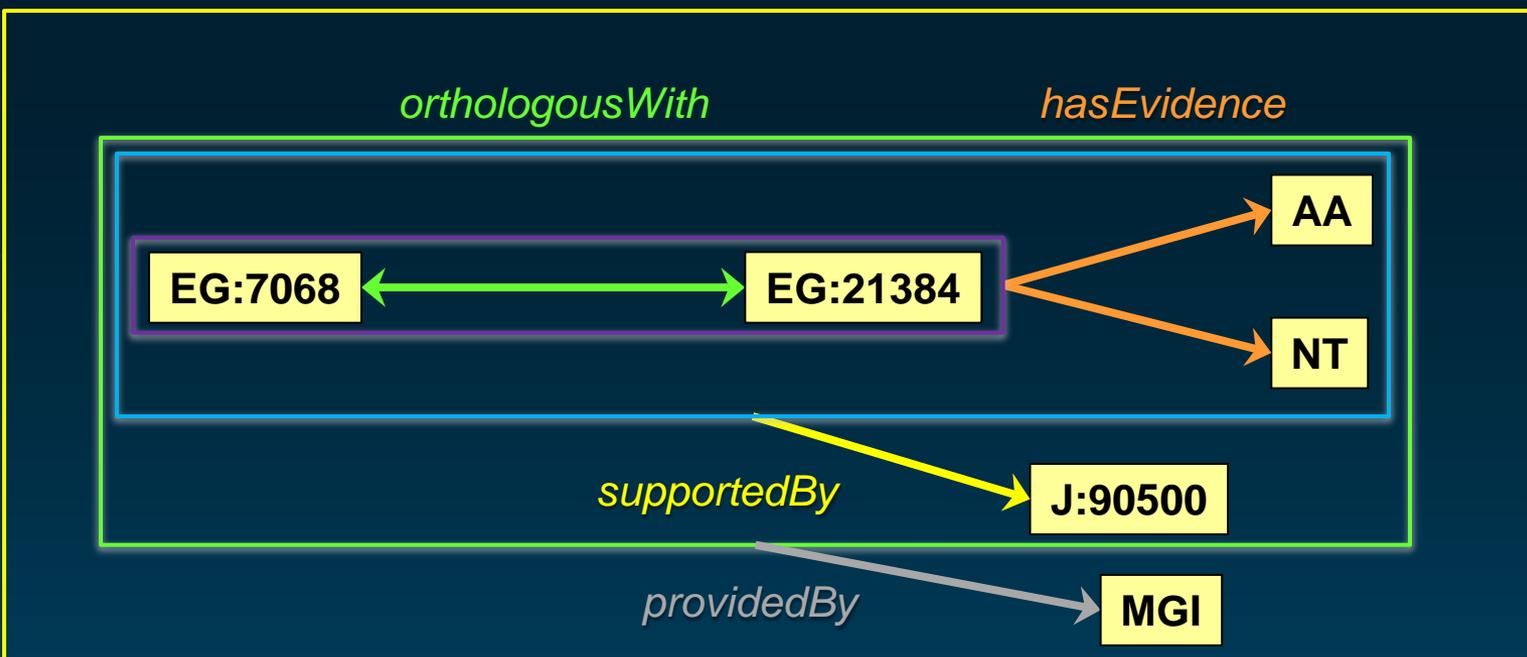
Reset

For the selected sequences

download in FASTA format

Go

Species	Symbol	Chr (Location) Coordinates	AccID (Source) <small>(Check box to select sequence)</small>	Evidence
human	THRB	3 (p24.2)	HGNC:11799 (HGNC) 7068 (Entrez Gene) <input type="checkbox"/> NM_000461 (RefSeq) <input type="checkbox"/> NP_000452 (RefSeq) 190160 (OMIM)	AA NT
mouse, laboratory	Thrb	14 (A3) 18493474-18870600(+) <small>NCBI Mouse Build 37</small>	MGI:98743 (MGI) 21834 (Entrez Gene) <input type="checkbox"/> BC089035 (GenBank) <input type="checkbox"/> P37242 (UniProt) <input type="checkbox"/> OTTMUSG00000023243 (VEGA Gene Model)	AA NT
References				1:90500



Species	Symbol	Chr (Location) Coordinates	AccID (Source) <small>(Check box to select sequence)</small>	Evidence
human	THRB	3 (p24.2)	<input checked="" type="checkbox"/> HGNC:11799 (HGNC) <input checked="" type="checkbox"/> 7068 (Entrez Gene) <input type="checkbox"/> NM_000461 (RefSeq) <input type="checkbox"/> NP_000452 (RefSeq) <input type="checkbox"/> 190160 (OMIM)	<div style="border: 1px solid orange; padding: 2px;">AA NT</div>
mouse, laboratory	Thrb	14 (A3) 18493474-18870600(+) <small>NCBI Mouse Build 37</small>	<input type="checkbox"/> MGI:98743 (MGI) <input checked="" type="checkbox"/> 21834 (Entrez Gene) <input type="checkbox"/> BC089035 (GenBank) <input type="checkbox"/> P37242 (UniProt) <input type="checkbox"/> OTTMUSG00000023243 (VEGA Gene Model)	<div style="border: 1px solid orange; padding: 2px;">AA NT</div>
References				J:90500



Mammalian Orthology Query Form

Sort by: Marker in primary species Chromosomal location in primary species

Include in results: only selected species all Orthologous species.

Max number of items returned: 100 500 No limit

Primary Species

ANY
mouse, laboratory (Mus musculus domesticus)
human (Homo sapiens)
cat, domestic (Felis catus)
cattle (Bos taurus)

Marker Symbol/Name:

NOT contains Search

Chromosome(s):

=

Cytogenetic Band:

begins

Author:

contains

Reference Accession ID: (from MGI, MEDLINE, etc.)

Marker Accession ID: (from MGI, GDB, etc.)

Comparison Species

ANY
mouse, laboratory (Mus musculus domesticus)
human (Homo sapiens)
cat, domestic (Felis catus)
cattle (Bos taurus)

Chromosome(s):

=

Cytogenetic Band:

begins



A novel mutation (E333D) in the thyroid hormone beta receptor causing resistance to thyroid hormone syndrome.

Maraninchi M, Bourcigaux N, Dace A, El-Yazidi C, Malezet-Desmoulins C, Krempf M, Torresani J, Margotat A.

UMR 476 INSERM/1260 INRA, Université de la Méditerranée, Faculté de Médecine, Marseille, France.

Resistance to thyroid hormone (RTH) is an inherited syndrome characterized by elevated serum thyroid hormones (TH), failure to suppress pituitary thyroid stimulating hormone (TSH) secretion, and variable peripheral tissue responsiveness to TH. The disorder is associated with diverse mutations in the thyroid hormone beta receptor (TRbeta). Here, we report a novel natural RTH mutation (E333D) located in the large carboxy-terminal ligand binding domain of TRbeta. The mutation was identified in a 22-year-old French woman coming to medical attention because of an increasing overweight. Biochemical tests showed elevated free thyroxine (T4: 20.8 pg/ml (normal, 8.5-18)) and triiodothyronine (T3: 5.7 pg/ml (normal, 1.4-4)) in the serum, together with an inappropriately nonsuppressed TSH level of 4.7 mU/ml (normal, 0.4-4). Her father and her brother's serum tests also showed biochemical abnormalities consistent with RTH. Direct sequencing of the TRbeta gene revealed a heterozygous transition 1284A>C in exon 9 resulting in substitution of glutamic acid 333 by aspartic acid residue (E333D). Further functional analyses of the novel TRbeta mutant were conducted. We found that the E333D mutation neither significantly affected the affinity of the receptor for T3 nor modified heterodimer formation with retinoid X receptor (RXR) when bound to DNA. However, in transient transfection assays, the E333D TRbeta mutant exhibited impaired transcriptional regulation on two distinct positively regulated thyroid response elements (F2- and DR4-TREs) as well as on the negatively regulated human TSHalpha promoter. Moreover, a dominant inhibition of the wild-type TRbeta counterpart transactivation function was observed on both a positive (F2-TRE) and a negative (TSHalpha) promoter. These results strongly suggest that the E333D TRbeta mutation is responsible for the RTH phenotype in the proposita's family.

PMID: 17177138 [PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances

Publication Types:

[Case Reports](#)

MeSH Terms:

[Adult](#)

[Amino Acid Substitution](#)

[DNA/genetics](#)

[Electrophoretic Mobility Shift Assay](#)

[Female](#)

[Gene Amplification](#)

[Humans](#)

[Male](#)

[Mutation](#)

[Pedigree](#)

[Thyroid Hormone Receptors beta/genetics*](#)

[Thyroid Hormone Resistance Syndrome/genetics*](#)

[Thyroid Hormones/blood](#)

A novel mutation (E333D) in the thyroid hormone beta receptor causing resistance to thyroid hormone syndrome.

Maraninchi M, Bourcigaux N, Dace A, El-Yazidi C, Malezet-Desmoulins C, Krempf M, Torresani J, Margotat A.

UMR 476 INSERM1260 INRA, Université de

Resistance to thyroid hormone (RTH) and variable peripheral tissue response to mutation (E333D) located in the large increasing overweight. Biochemical tests inappropriately nonsuppressed TSH of the TRbeta gene revealed a heterozygous TRbeta mutant were conducted. We found when bound to DNA. However, in trans elements (F2- and DR4-TREs) as we observed on both a positive (F2-TRE) proposita's family.

PMID: 17177139 [PubMed - indexed for MEDLINE]

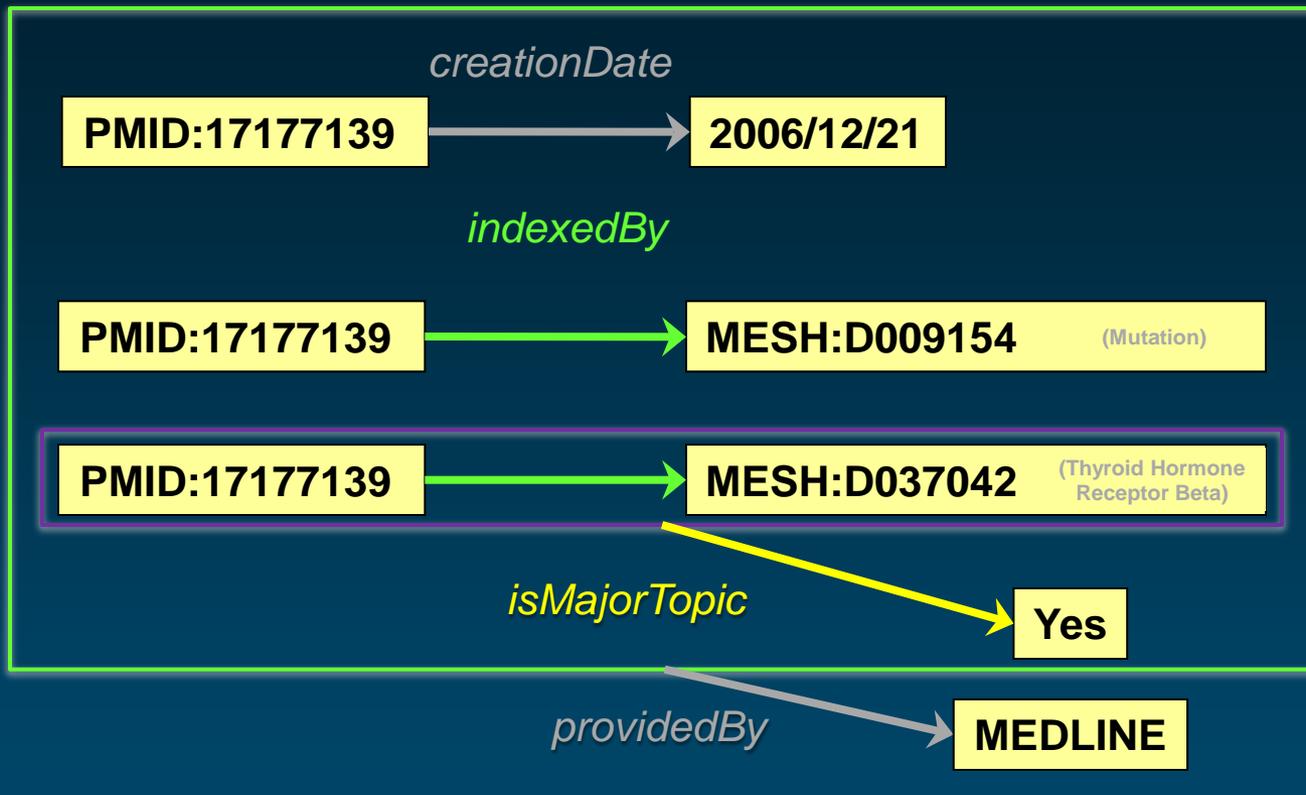
Publication Types, MeSH Terms

Publication Types:

Case Reports

MeSH Terms:

- Adult
- Amino Acid Substitution
- DNA/genetics
- Electrophoretic Mobility Shift Assay
- Female
- Gene Amplification
- Humans
- Male
- Mutation
- Pedigree
- Thyroid Hormone Receptors beta, genetics*
- Thyroid Hormone Resistance Syndrome, genetics*
- Thyroid Hormones/blood



Examples of applications requiring provenance information

Types of applications

- ◆ Information retrieval
- ◆ Multi-document summarization
- ◆ Question answering
- ◆ Knowledge discovery



Information retrieval

◆ Application

- Search by statements
e.g., find all documents asserting that
“IL-13 inhibits COX-2”

◆ Provenance information

- Publication date
- Origin of indexing
- ...
- (Similar to traditional search engines)



Multi-document summarization

◆ Application

- Extract and prioritize statements from multiple documents to create a summary

◆ Provenance information

- Level of confidence (e.g., for automatic extraction using NLP techniques)



Question answering

◆ Application

- Find answers to templated questions (e.g., “what genes does IL-13 exhibit?”)

◆ Provenance information

- Select reputable sources (provenance information associated with the documents: source)
- Select recent documents (provenance information associated with the documents: publication date)
- Select valid statements (provenance information associated with the statements: level of confidence)



Knowledge discovery

◆ Application

- Find path in a graph between entities of interest, using patterns of link types

◆ Provenance information

- Origin of the statements (not entities)
- Required for both asserted and inferred statements
 - Compute provenance information for inferred statements



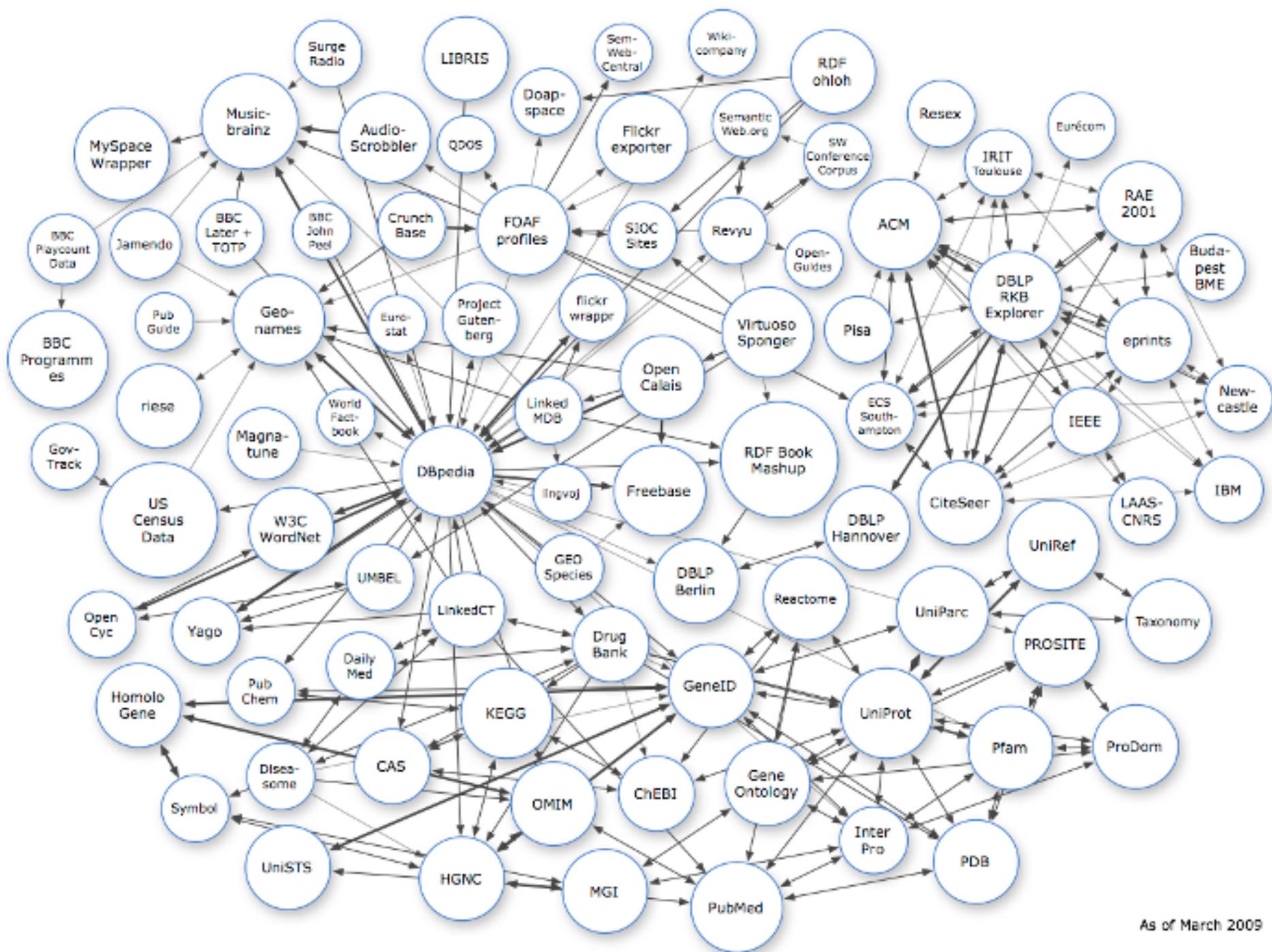
Issues and challenges

Limitations of naïve implementation

- ◆ Reification through blank nodes
 - Not intuitive to users
 - Further away from the domain model
 - Increases the complexity of queries
 - Inefficient
 - Increases the number of triples
 - Scalability issues

Lack of support for provenance

- ◆ No native support for provenance information in
 - Mainstream triple stores
 - Major query languages for triple stores
 - Many variants of SPARQL and RQL provide limited support
- ◆ Named graphs (supported in quad stores) do not offer the required level of granularity
- ◆ Standardization of emerging provenance models



Linked data vs. provenance

◆ Currently

- No provenance information in Linked Data
- Does Bio2RDF's "Banff manifesto" exclude provenance *de facto*? (no blank nodes allowed)
- Ability to link datasets outweighs absence of provenance information

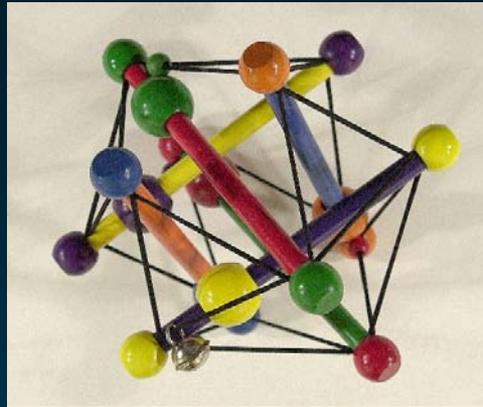
◆ Limitations

- Applications cannot select/exclude specific statements
- Navigation vs. knowledge discovery



Summary

- ◆ Need for systems handling provenance information
 - Transparently for the user
 - Directly in the triple stores / query languages
 - At different levels of granularity
 - e.g., resource vs. statement within a resource
 - For both asserted and inferred statements
 - Scalability
- ◆ Not exposing provenance information in Linked Data is a major limitation



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA