NLM Indexing Initiative Tools for NLP: MetaMap and the Medical Text Indexer

Natural Language Processing: State of the Art, Future Directions

April 23, 2012

Alan R. Aronson

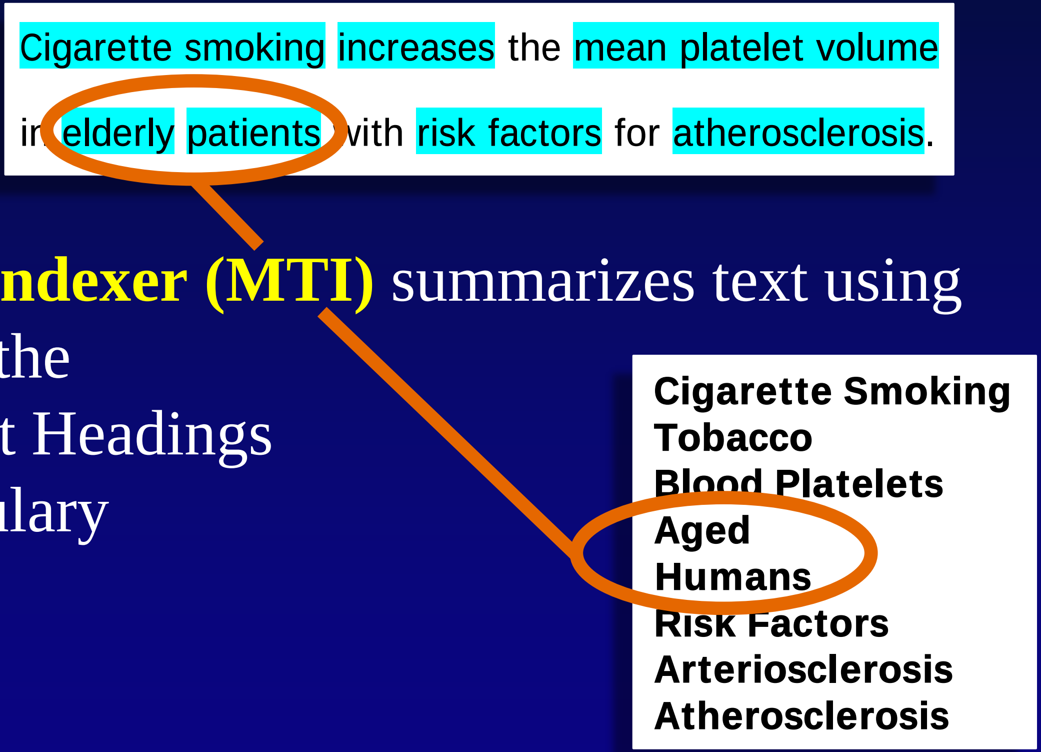
Outline

- Introduction
- MetaMap
    - Overview
    - Linguistic roots
    - Recent Word Sense Disambiguation (WSD) efforts
- The NLM Medical Text Indexer (MTI)
    - Overview
    - MTI as First-line Indexer (MTIFL)
    - Recent improvements
    - Gene indexing

MetaMap/MTI Example

- **MetaMap** identifies biomedical concepts in text

- **Medical Text Indexer (MTI)** summarizes text using **MetaMap** and the Medical Subject Headings (MeSH) vocabulary



Cigarette smoking increases the mean platelet volume in elderly patients with risk factors for atherosclerosis.

Cigarette Smoking
Tobacco
Blood Platelets
Aged
Humans
Risk Factors
Arteriosclerosis
Atherosclerosis

Outline

- Introduction
- MetaMap
  - Overview
  - Linguistic roots
  - Recent Word Sense Disambiguation (WSD) efforts
- The NLM Medical Text Indexer (MTI)
  - Overview
  - MTI as First-line Indexer (MTIFL)
  - Recent improvements
  - Gene indexing

MetaMap Overview

- Named-entity recognition program
- Identify UMLS Metathesaurus concepts in text
- Linguistic rigor
- Flexible partial matching
- Emphasis on thoroughness rather than speed

The MetaMap Algorithm

- Parsing
  - Using SPECIALIST minimal commitment parser, SPECIALIST lexicon, MedPost part of speech tagger
- Variant generation
  - Using SPECIALIST lexicon, Lexical Variant Generation (LVG)
- Candidate retrieval
  - From the Metathesaurus
- Candidate evaluation
- Mapping construction

MetaMap Evaluation Function

- Weighted average of
  - centrality (is the head involved?)
  - variation (average of all variation)
  - coverage (how much of the text is matched?)
  - cohesiveness (in how many pieces?)

Me

Infer

```
C0180860: Filters                    [mnob]
C0581406: Optical filter             [medd]
C1522664: filter information process [inpr]
C1704449: Filter (function)          [cnce]
```

Meta	Metathesau	Metathesauru	UMLS Semantic Type
909	C0080306:	Inferior Vena Cava Filter	[medd]
804	C0180860:	Filter	[mnob]
804	C0581406:	Filter	[medd]
804	C1522664:	Filter	[inpr]
804	C1704449:	Filter	[cnce]
804	C1704684:	Filter	[medd]
[illegible]	[illegible]	[illegible]	[medd]
[illegible]	[illegible]	[illegible]	[blor]
673	C0042460:	Vena caval	[bpoc]
637	C0038257:	Stent	[medd]
637	C1705817:	Stent	[medd]
637	C0447122:	Vena	[bpoc]

```
C0038257: Stent, device            [medd]
C1705817: Stent Device Component   [medd]
```

MetaMap Final Mappings

Inferior vena caval stent filter

Final Mappings (subsets of candidate sets):

```
Meta Mapping (911)
909  C0080306: Inferior Vena Cava Filter [medd]
637  C1705817: Stent                     [medd]

Meta Mapping (911):
909  C0080306: Inferior Vena Cava Filter [medd]
637  C0038257: Stent                     [medd]
```

Word Sense Disambiguation (WSD)

- Kids with *colds* may also have a sore throat, cough, headache, mild fever, fatigue, muscle aches, and loss of appetite.

- Candidate MetaMap mappings for *cold*

```
C0234192: Cold (Cold sensation)
C0009264: Cold (Cold temperature)
C0009443: Cold (Common cold)
```

Knowledge-based WSD

- Compare UMLS candidate concept profile vectors to context of ambiguous word
- Concept profile vectors' words from definition, synonyms and related concepts

Common cold	
Weight	**Word**
265	infect
126	disease
41	fever
40	cough

Cold temperature	
Weight	**Word**
258	temperature
86	hypothermia
72	effect
48	hot

- Candidate concept with highest similarity is predicted

Knowledge-based WSD

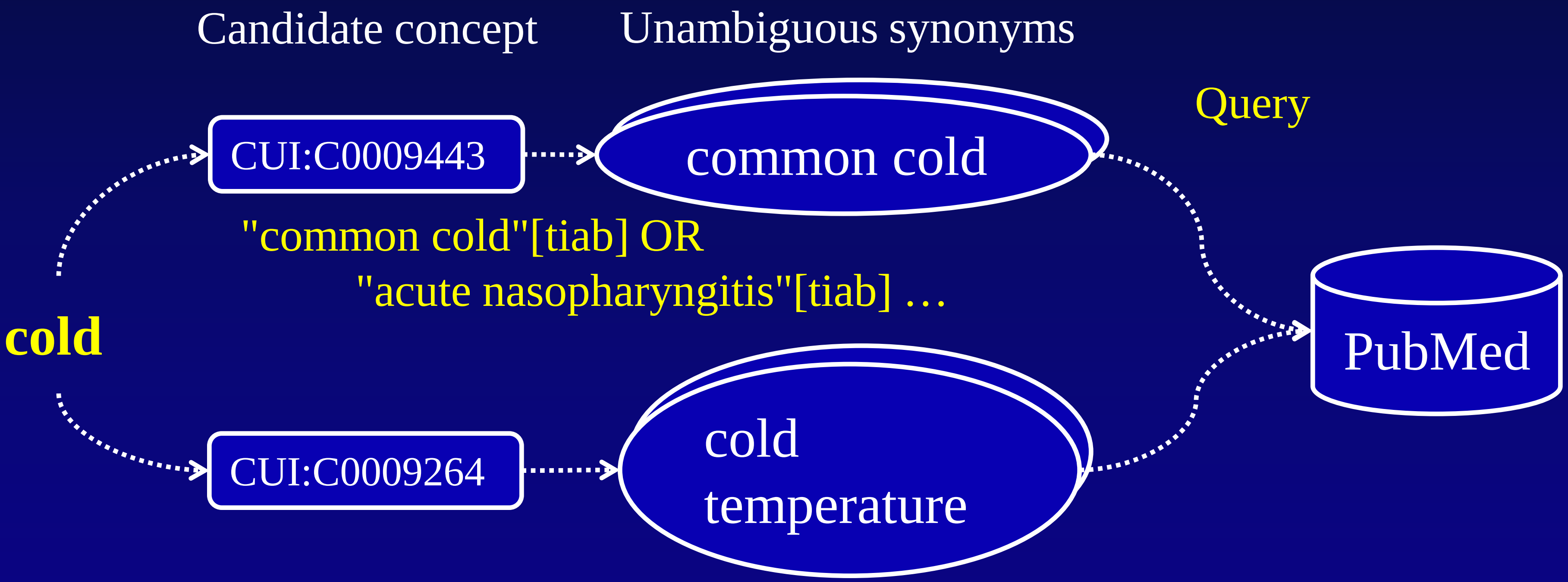
- Kids with *colds* may also have a sore throat, *cough*, headache, mild *fever*, fatigue, muscle aches, and loss of appetite.

Common cold	
Weight	**Word**
265	infect
126	disease
41	**fever**
40	**cough**

Cold temperature	
Weight	**Word**
258	temperature
86	hypothermia
72	effect
48	hot

Automatically Extracted Corpus WSD

- MEDLINE contains numerous examples of ambiguous words context, though not disambiguated



"common cold"[tiab] OR "acute nasopharyngitis"[tiab] …

"cold temperature"[tiab] OR "low temperature"[tiab] …

WSD Method Results

- Corpus method has better accuracy than UMLS method

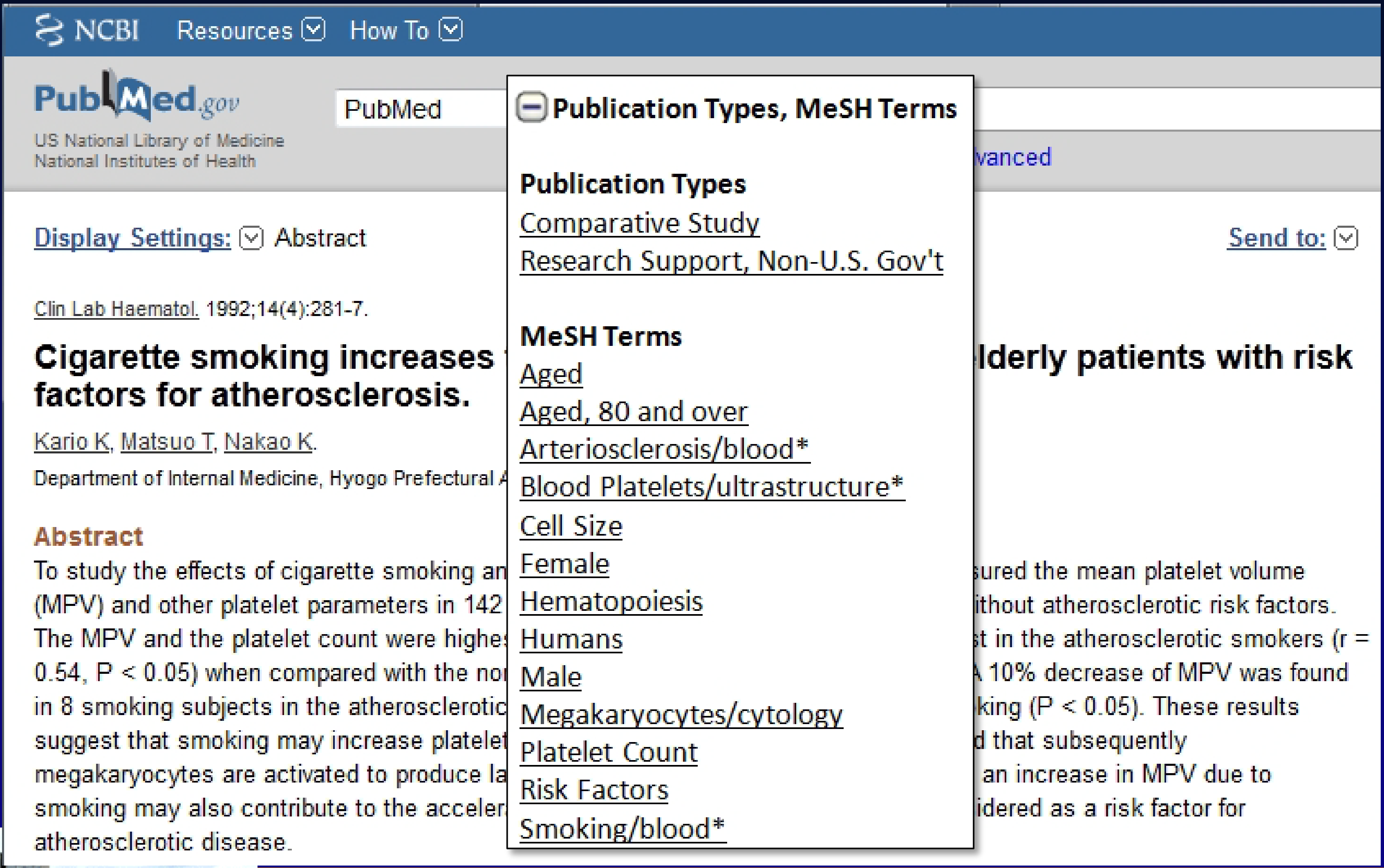
	UMLS	Corpus
NLM WSD	0.65	**0.69**
MSH WSD	0.81	**0.84**

- MSH WSD data set created using MeSH indexing
  - 203 ambiguous words
  - 81 semantic types
  - 37,888 ambiguity cases
- Indirect evaluation with summarization and MTI correlates with direct evaluation

Outline

- Introduction
- MetaMap
  - Overview
  - Linguistic roots
  - Recent Word Sense Disambiguation (WSD) efforts
- The NLM Medical Text Indexer (MTI)
  - Overview
  - MTI as First-line Indexer (MTIFL)
  - Recent improvements
  - Gene indexing

MEDLINE Citation Example



NCBI
Resources
How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
vanced

Display Settings: Abstract

Send to:

Clin Lab Haematol. 1992;14(4):281-7.

Cigarette smoking increases … **lderly patients with risk factors for atherosclerosis.**

Kario K, Matsuo T, Nakao K.

Department of Internal Medicine, Hyogo Prefectural A

Abstract

To study the effects of cigarette smoking an … sured the mean platelet volume
(MPV) and other platelet parameters in 142 … ithout atherosclerotic risk factors.
The MPV and the platelet count were highe … st in the atherosclerotic smokers (r =
0.54, P < 0.05) when compared with the no … A 10% decrease of MPV was found
in 8 smoking subjects in the atherosclerotic … king (P < 0.05). These results
suggest that smoking may increase platelet … d that subsequently
megakaryocytes are activated to produce la … an increase in MPV due to
smoking may also contribute to the acceler … idered as a risk factor for
atherosclerotic disease.

⊖ **Publication Types, MeSH Terms**

Publication Types

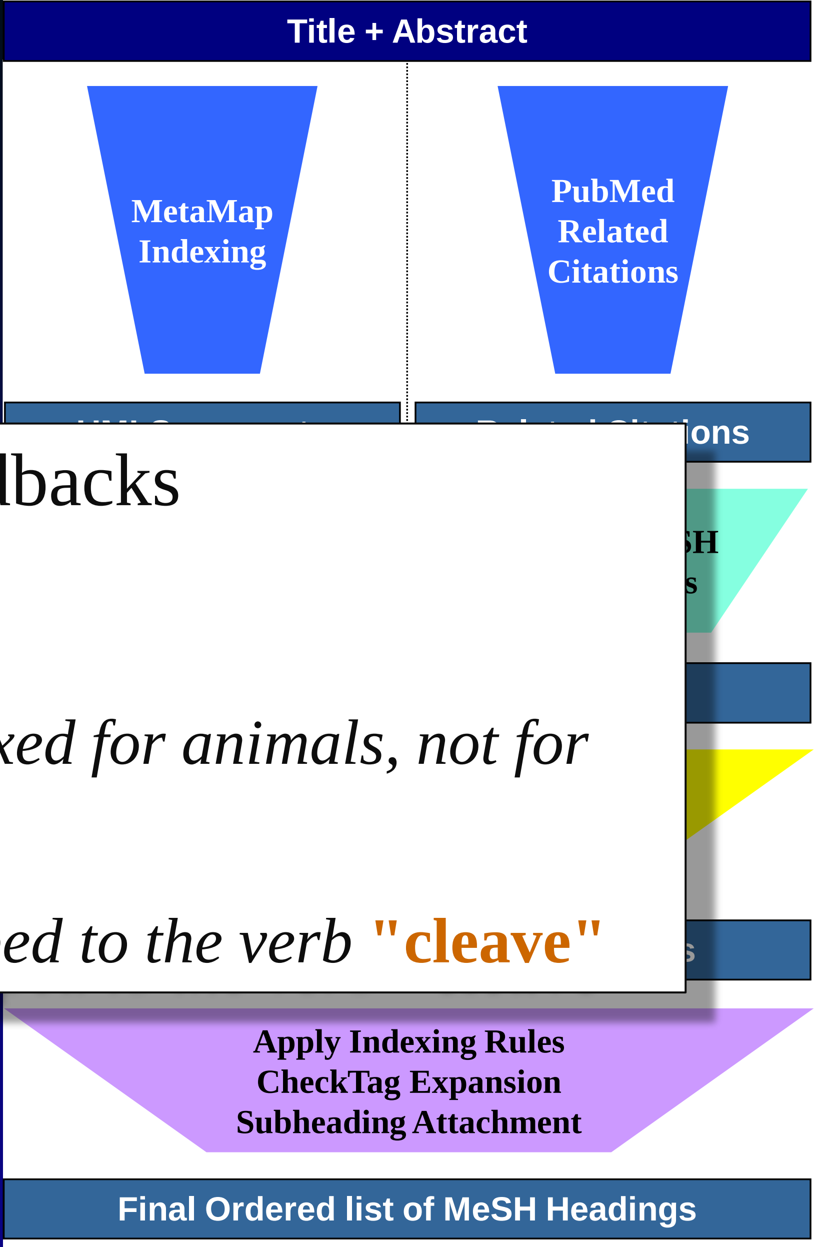
Comparative Study
Research Support, Non-U.S. Gov't

MeSH Terms

Aged
Aged, 80 and over
Arteriosclerosis/blood*
Blood Platelets/ultrastructure*
Cell Size
Female
Hematopoiesis
Humans
Male
Megakaryocytes/cytology
Platelet Count
Risk Factors
Smoking/blood*

MTI

- MetaMap Indexing – Actually found in text



Received **2,330** Indexer Feedbacks

Incorporated **40%** into MTI

March 20, 2012

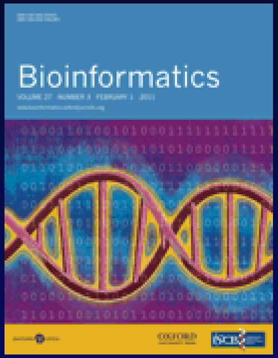
Hibernation *should only be indexed for animals, not for* **"stem cell hibernation"**

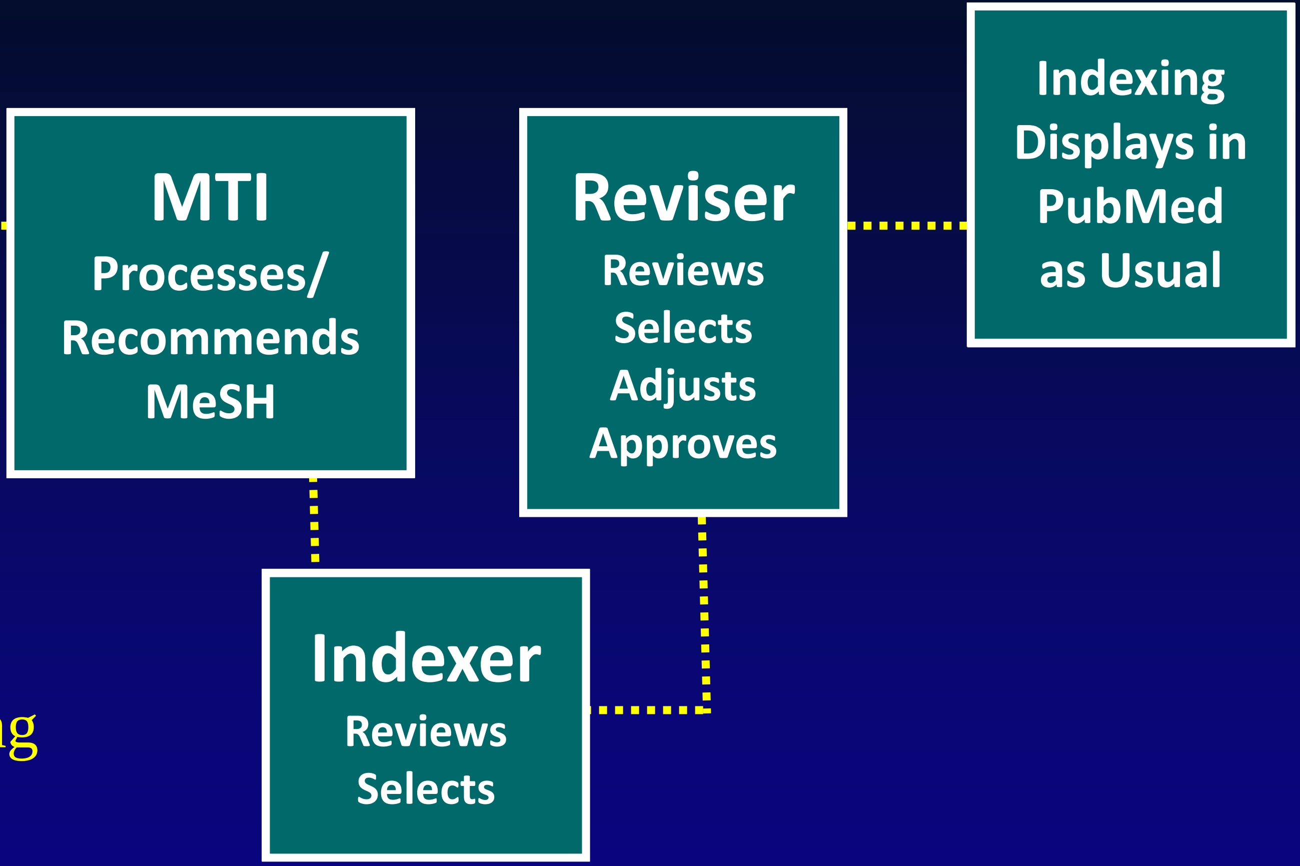
Clove *(spice) should not be mapped to the verb* **"cleave"**

MTI Uses

- Assisted indexing of MEDLINE by Index Section
- Assisted indexing of Cataloging and History of Medicine Division records
- Automatic indexing of NLM Gateway meeting abstracts
- First-line indexing (MTIFL) since February 2011

MTI as First-Line Indexer (MTIFL)





MTI
Processes/
Recommends
MeSH

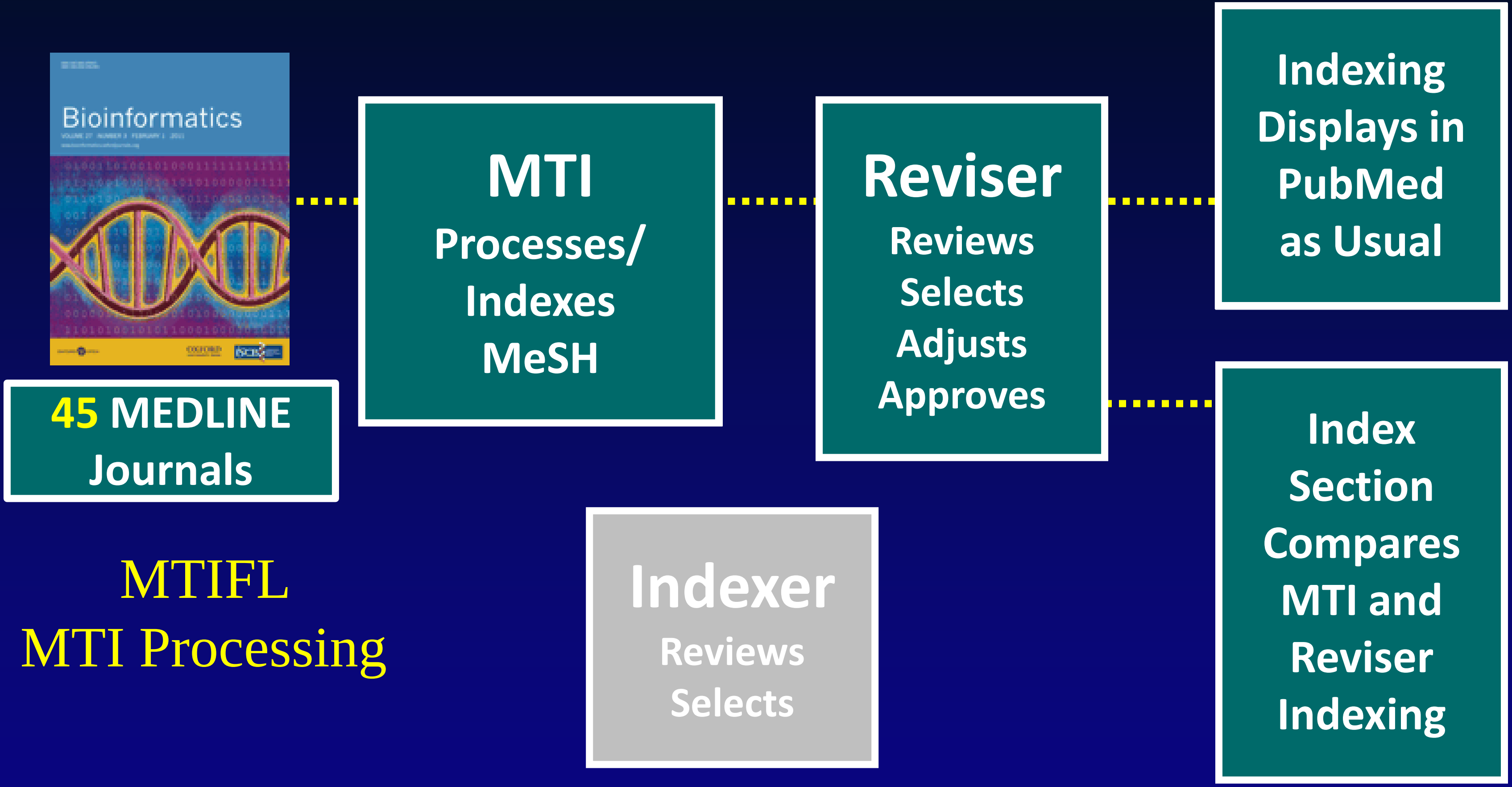
Indexer
Reviews
Selects

Reviser
Reviews
Selects
Adjusts
Approves

Indexing
Displays in
PubMed
as Usual

"Normal"
MTI Processing

MTI as First-Line Indexer (MTIFL)



45 MEDLINE Journals

MTI
Processes/
Indexes
MeSH

Reviser
Reviews
Selects
Adjusts
Approves

Indexing Displays in PubMed as Usual

Index Section Compares MTI and Reviser Indexing

Indexer
Reviews
Selects

MTIFL
MTI Processing

CheckTags Machine Learning Results

- 200k citations for training and 100k citations for testing

CheckTag	F_1 before ML	F_1 with ML	Improvement
Middle Aged	1.01%	59.50%	+58.49
Aged	11.72%	54.67%	+42.95
Child, Preschool	6.11%	45.40%	+39.29
Adult	19.49%	56.84%	+37.35
Male	38.47%	71.14%	+32.67
Aged, 80 and over	1.50%	30.89%	+29.39
Young Adult	2.83%	31.63%	+28.80
Female	46.06%	73.84%	+27.78
Adolescent	24.75%	42.36%	+17.61
Humans	79.98%	91.33%	+11.35
Infant	34.39%	44.69%	+10.30
Swine	71.04%	74.75%	+3.71

CheckTags Machine Learning Results

- 200k citations for training and 100k citations for testing

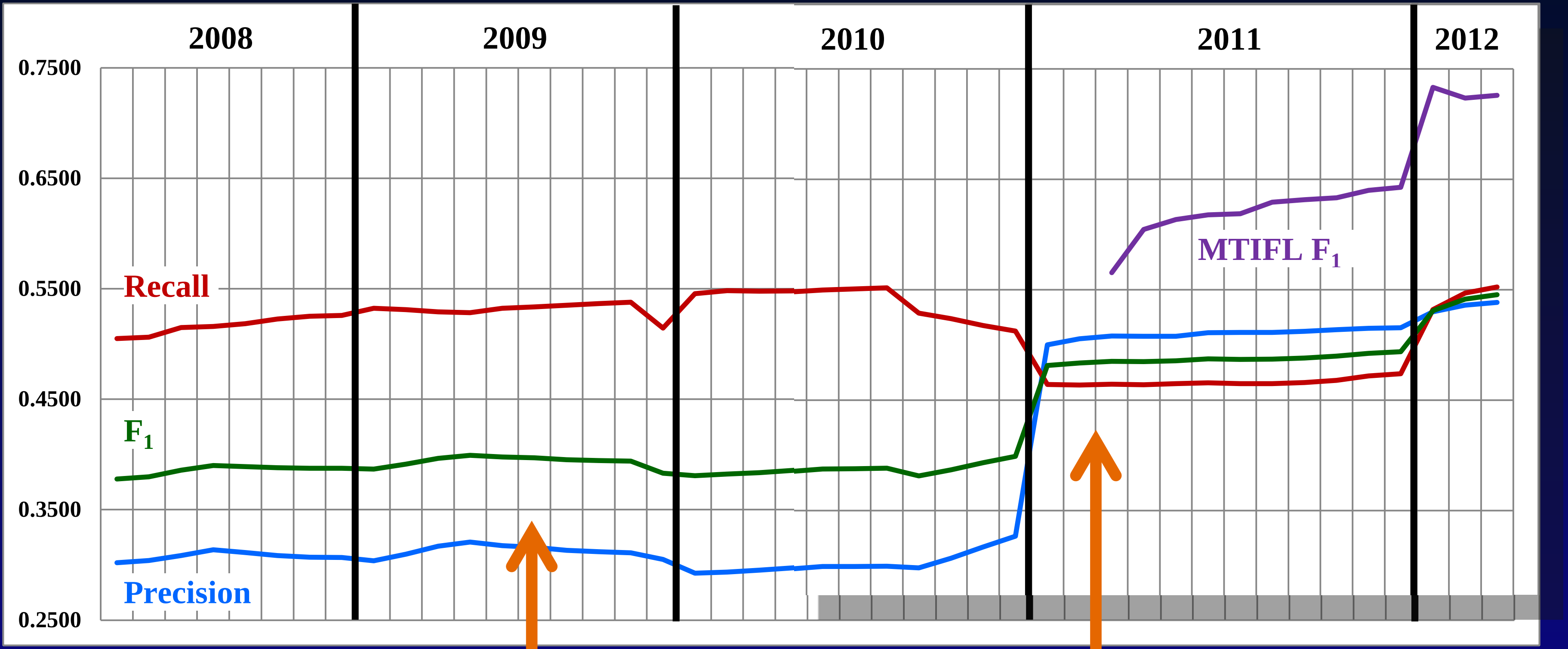
CheckTag	F_1 before ML	F_1 with ML	Improvement
Middle Aged	1.01%	59.50%	+58.49
Aged	11.72%	54.67%	+42.95
Child, Preschool	6.11%	45.40%	+39.29
Adult	19.49%	56.84%	+37.35
Male	**38.47%**	71.14%	+32.67
Aged, 80 and over	1.50%	30.89%	+29.39
Young Adult	2.83%	31.63%	+28.80
Female	**46.06%**	**73.84%**	**+27.78**
Adolescent	24.75%	42.36%	+17.61
Humans	**79.98%**	**91.33%**	**+11.35**
Infant	34.39%	44.69%	+10.30
Swine	71.04%	74.75%	+3.71

CheckTags Machine Learning Results

- 200k citations for training and 100k citations for testing

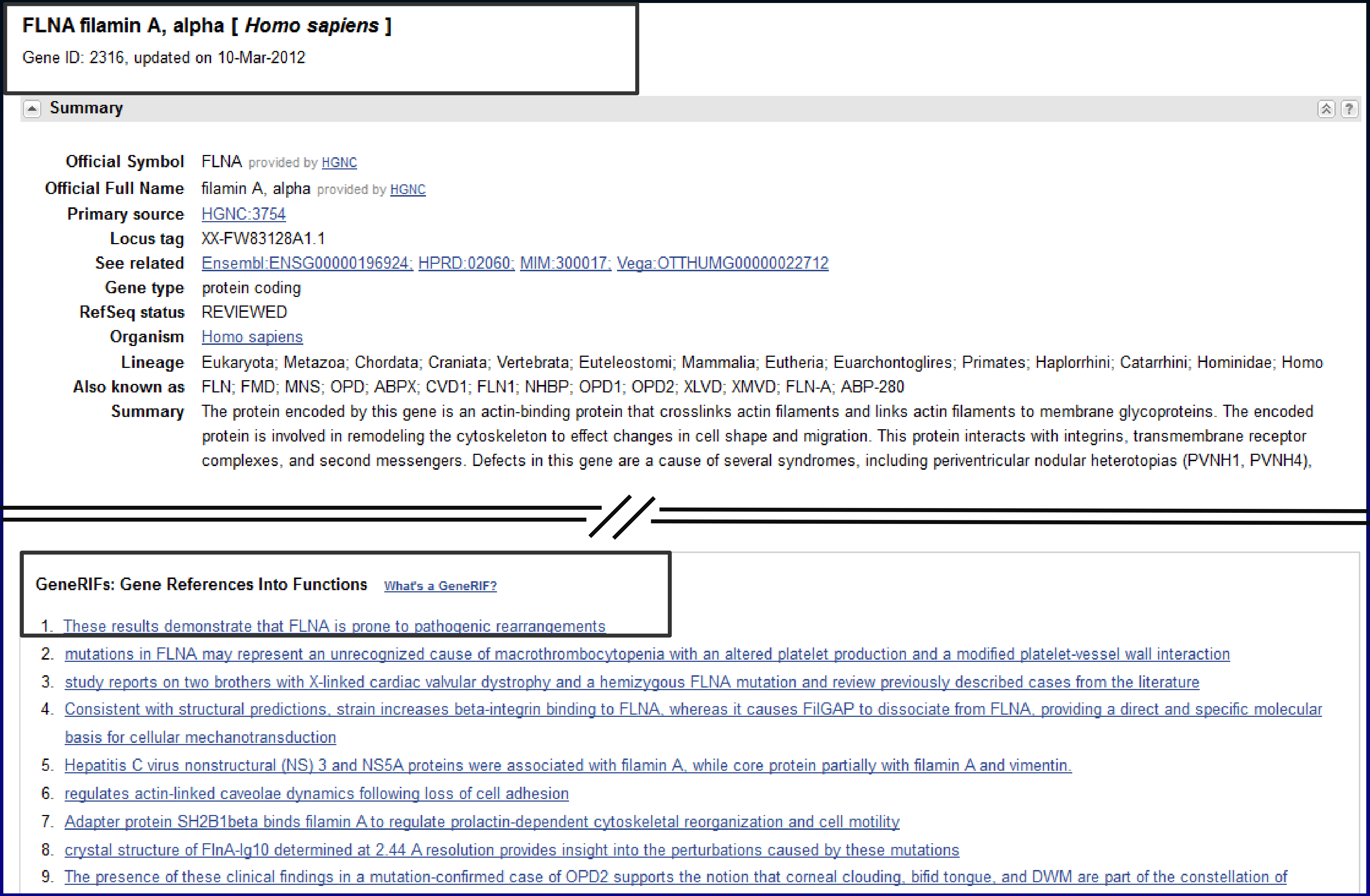
CheckTag	F_1 before ML	F_1 with ML	Improvement
Middle Aged	**1.01%**	**59.50%**	**+58.49**
Aged	11.72%	54.67%	+42.95
Child, Preschool	6.11%	45.40%	+39.29
Adult	19.49%	56.84%	+37.35
Male	38.47%	71.14%	+32.67
Aged, 80 and over	1.50%	30.89%	+29.39
Young Adult	2.83%	31.63%	+28.80
Female	46.06%	73.84%	+27.78
Adolescent	24.75%	42.36%	+17.61
Humans	79.98%	91.33%	+11.35
Infant	34.39%	44.69%	+10.30
Swine	71.04%	74.75%	+3.71

MTI - How are we doing?



Fruition of 2011 Changes

Focus on Precision versus Recall



FLNA filamin A, alpha [*Homo sapiens*]

Gene ID: 2316, updated on 10-Mar-2012

Summary

Official Symbol FLNA provided by HGNC

Official Full Name filamin A, alpha provided by HGNC

Primary source HGNC:3754

Locus tag XX-FW83128A1.1

See related Ensembl:ENSG00000196924; HPRD:02060; MIM:300017; Vega:OTTHUMG00000022712

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 FLN; FMD; MNS; OPD; ABPX; CVD1; FLN1; NHBP; OPD1; OPD2; XLVD; XMVD; FLN-A; ABP-280

Summary The protein encoded by this gene is an actin-binding protein that crosslinks actin filaments and links actin filaments to membrane glycoproteins. The encoded protein is involved in remodeling the cytoskeleton to effect changes in cell shape and migration. This protein interacts with integrins, transmembrane receptor complexes, and second messengers. Defects in this gene are a cause of several syndromes, including periventricular nodular heterotopias (PVNH1, PVNH4),

GeneRIFs: Gene References Into Functions

What's a GeneRIF?

1. These results demonstrate that FLNA is prone to pathogenic rearrangements
2. mutations in FLNA may represent an unrecognized cause of macrothrombocytopenia with an altered platelet production and a modified platelet-vessel wall interaction
3. study reports on two brothers with X-linked cardiac valvular dystrophy and a hemizygous FLNA mutation and review previously described cases from the literature
4. Consistent with structural predictions, strain increases beta-integrin binding to FLNA, whereas it causes FilGAP to dissociate from FLNA, providing a direct and specific molecular basis for cellular mechanotransduction
5. Hepatitis C virus nonstructural (NS) 3 and NS5A proteins were associated with filamin A, while core protein partially with filamin A and vimentin.
6. regulates actin-linked caveolae dynamics following loss of cell adhesion
7. Adapter protein SH2B1beta binds filamin A to regulate prolactin-dependent cytoskeletal reorganization and cell motility
8. crystal structure of FlnA-Ig10 determined at 2.44 A resolution provides insight into the perturbations caused by these mutations
9. The presence of these clinical findings in a mutation-confirmed case of OPD2 supports the notion that corneal clouding, bifid tongue, and DWM are part of the constellation of

The Gene Indexing Assistant (GIA)

- An automated tool to assist the indexer in identifying and creating GeneRIFs
  - Evaluate the article
  - Identify genes
  - Make links to Entrez Gene
  - Suggest geneRIF annotation

- Anticipated Benefits:
  - Increase in speed
  - Increase in comprehensiveness

The NLM Indexing Initiative Team

- Alan R. Aronson (Project Leader)
- James G. Mork (Staff)
- François-Michel Lang (Staff)
- Willie J. Rogers (Staff)
- Antonio J. Jimeno-Yepes (Postdoctoral Fellow)
- J. Caitlin Sticco (Library Associate Fellow)

http://metamap.nlm.nih.gov