Throwing down the gauntlet…

Bio-Ontologies SIG 2011

> If you believe *the hype* around wikis in biology, show us what you can do with 'wiki'-derived content…

When: July 15-16, 2011 (co-located with ISMB/ECCB)

Where: Vienna, Austria

Info: http://www.bio-ontologies.org.uk

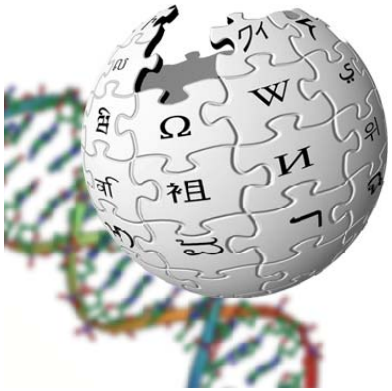
Organizers: Nigam Shah, Larisa Soldatova, Susie Stephens, Susanna Sansone

The Gene Wiki: Cultivating and mining community intelligence in biology

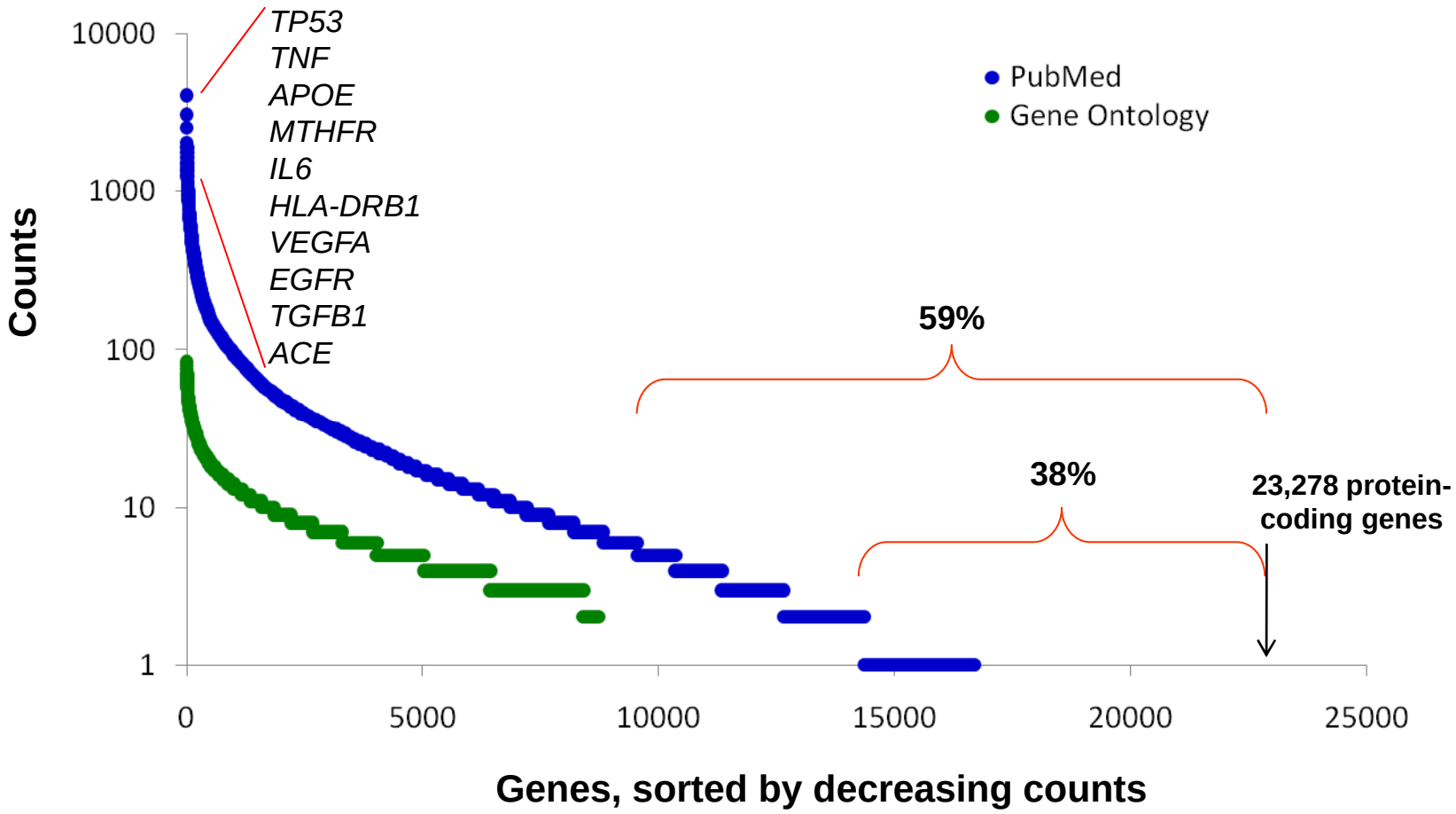
NETTAB 2010
November 30, 2010

Andrew Su

Genomics Institute of the
Novartis Research
Foundation

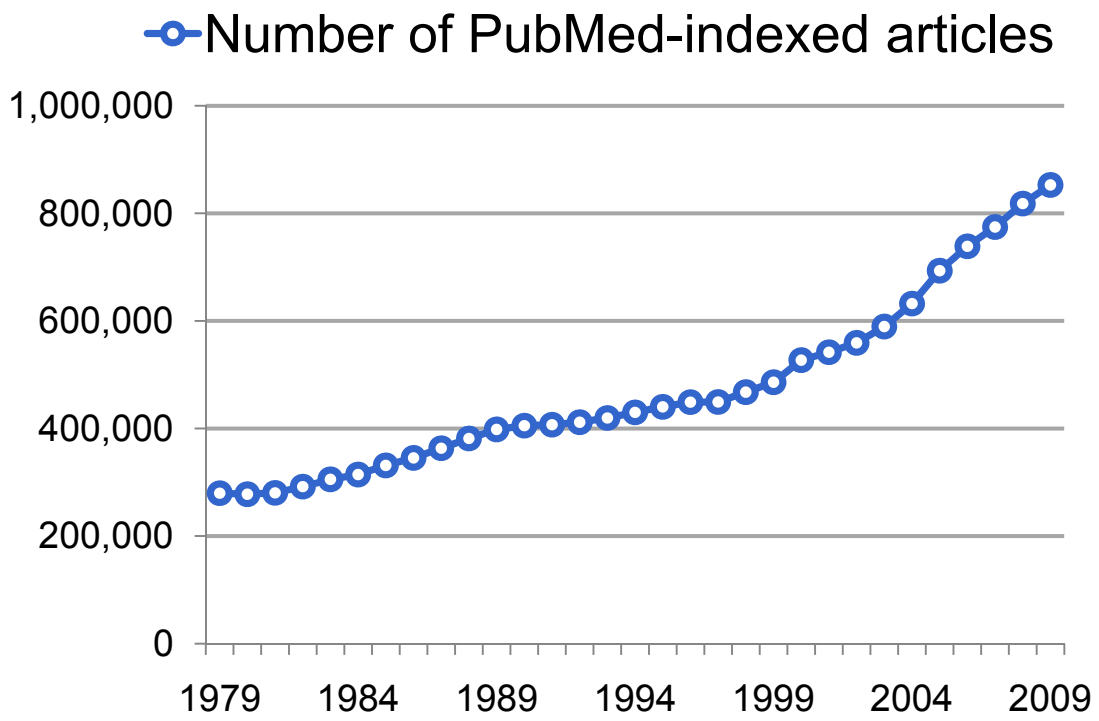


Few genes are well annotated…

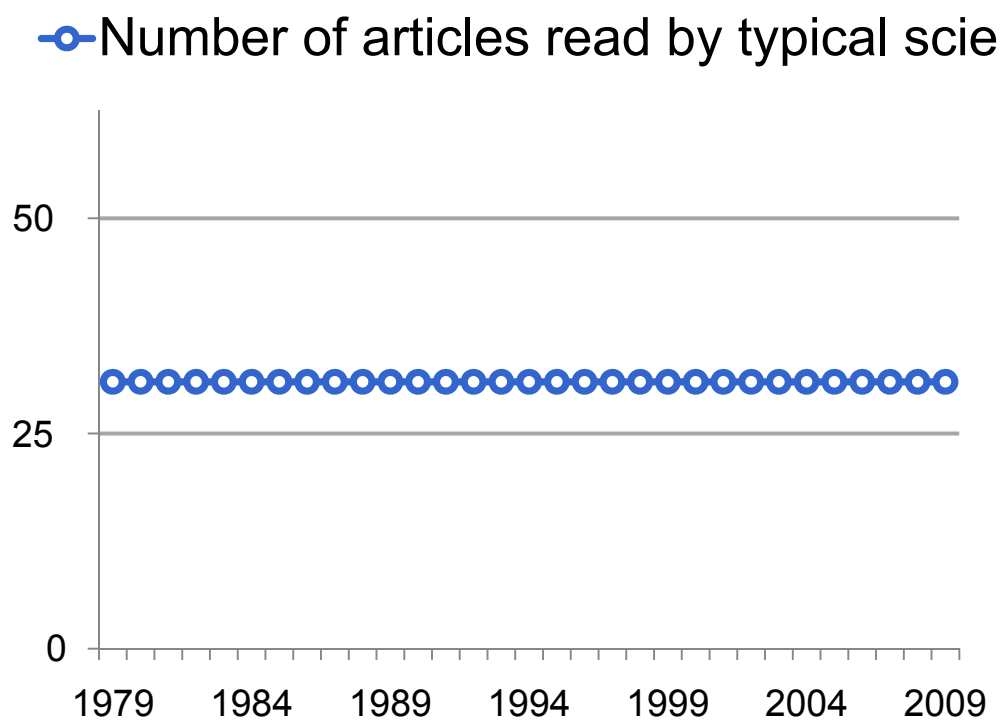


Data: NCBI gene2pubmed, August 2010

… because the literature is sparsely curated?



… because the literature is sparsely curated?




82,063 articles (0.4% of PubMed)
have been cited by GO annotations

Vol 455|4 September 2008 nature

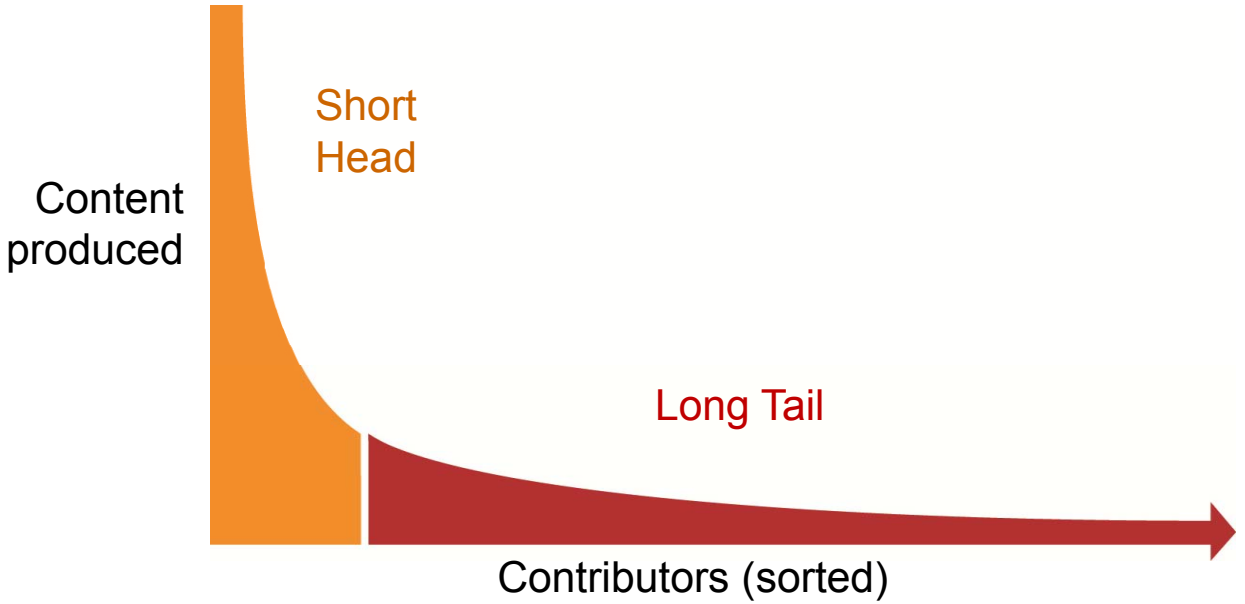
FEATURE

The future of biocuration

To thrive, the field that links biologists and their data urgently needs structure, recognition and support

> Sooner or later, the research community will need to be involved in the annotation effort to scale up to the rate of data generation.

The Long Tail is a prolific source of content



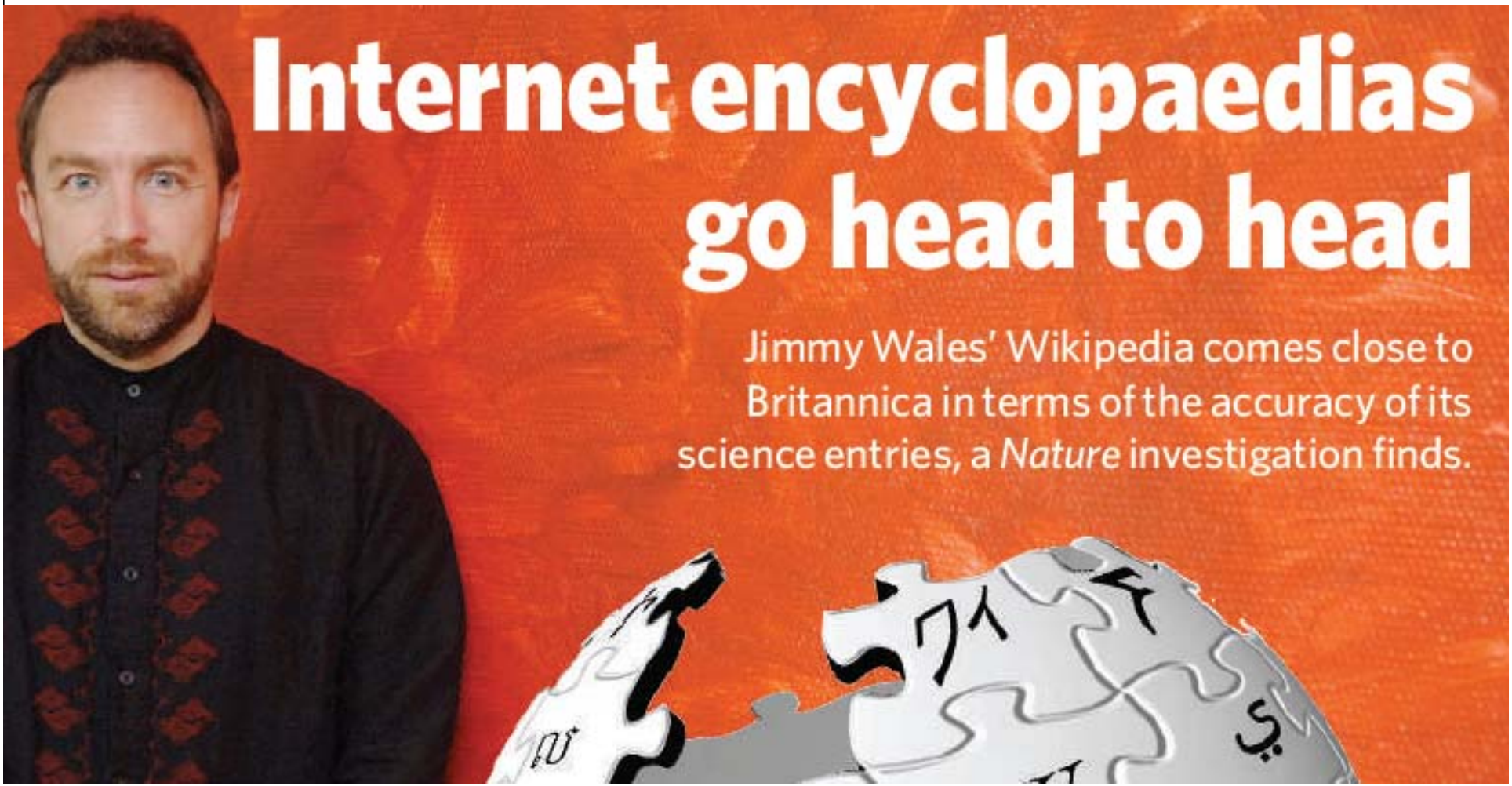
News reporting:	Newspapers	Blogs
Video:	TV/Hollywood	YouTube
Product reviews:	Consumer reports	Amazon reviews
Food reviews:	Food critics	Yelp
Judging:	Olympics	American Idol

Wikipedia is reasonably accurate

nature

Vol 438|15 December 2005

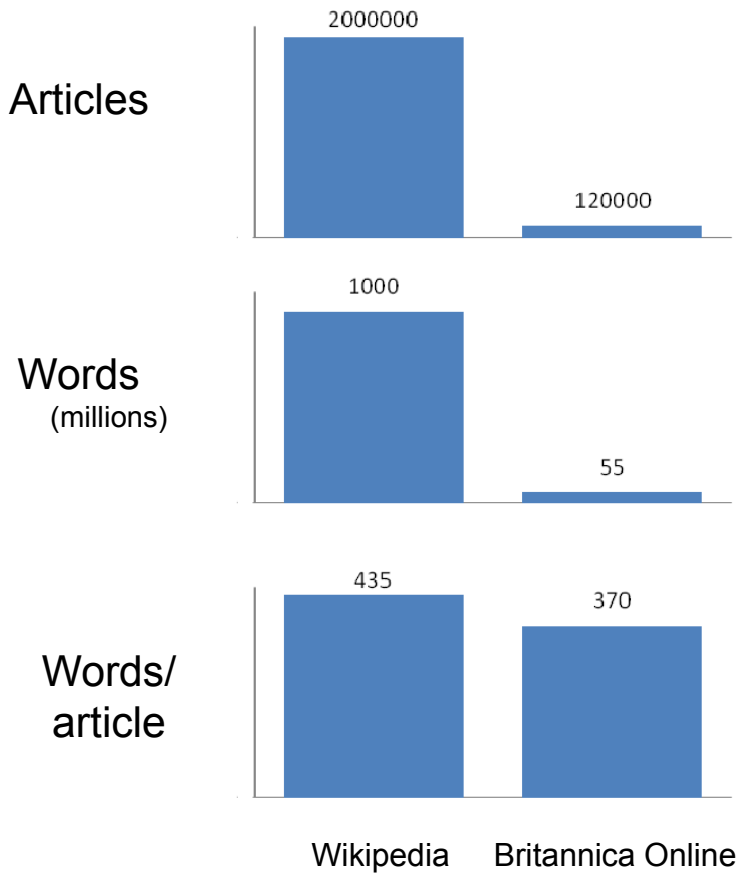
SPECIAL REPORT



Internet encyclopaedias go head to head

Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries, a *Nature* investigation finds.

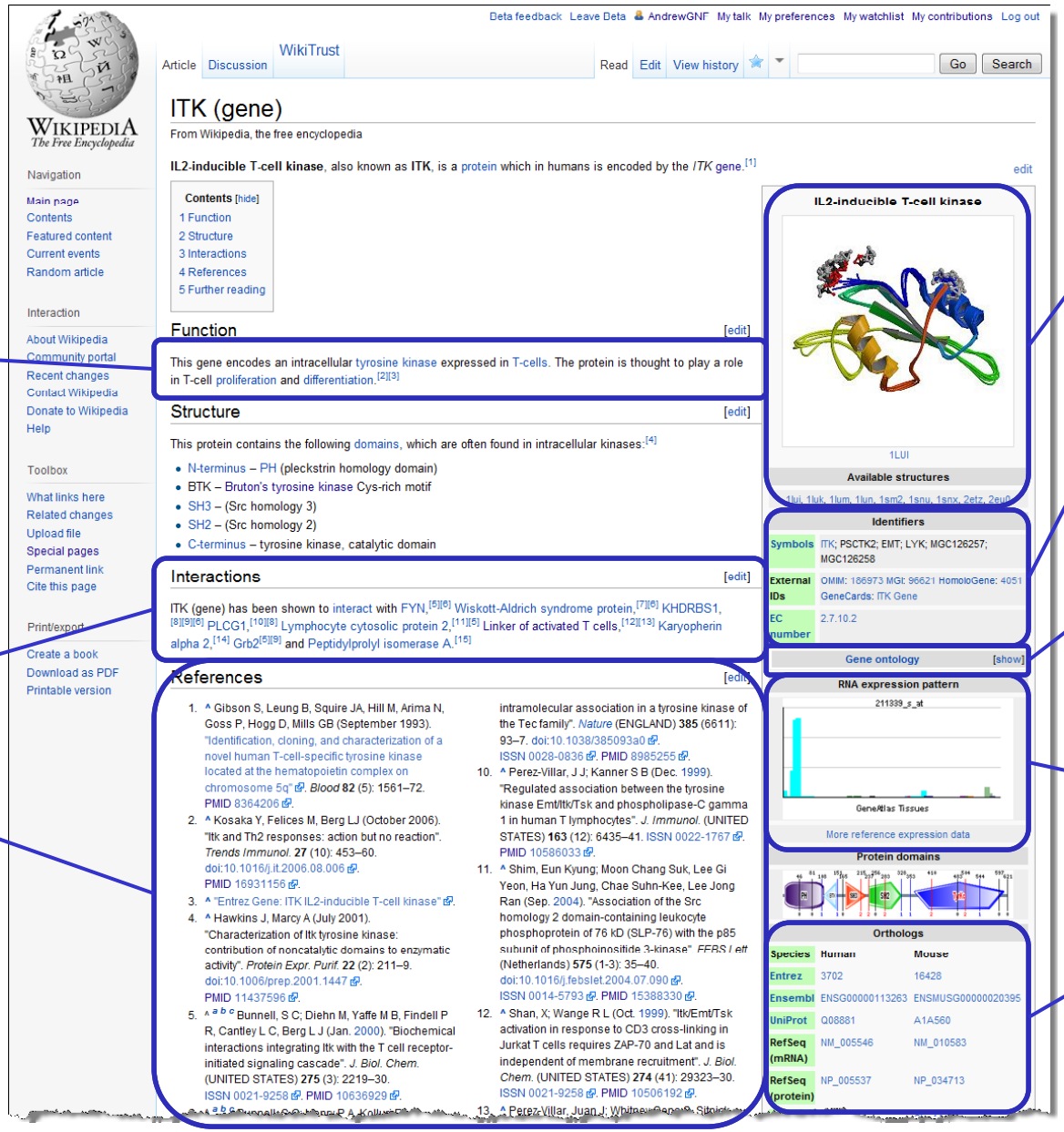
Wikipedia has breadth and depth



http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons, July 2008

We can harness the
Long Tail of scientists
to directly participate in
the gene annotation
process.

Gene “stubs” seed community contributions



Gene summary

Protein interactions

Linked references

Protein structure

Symbols and identifiers

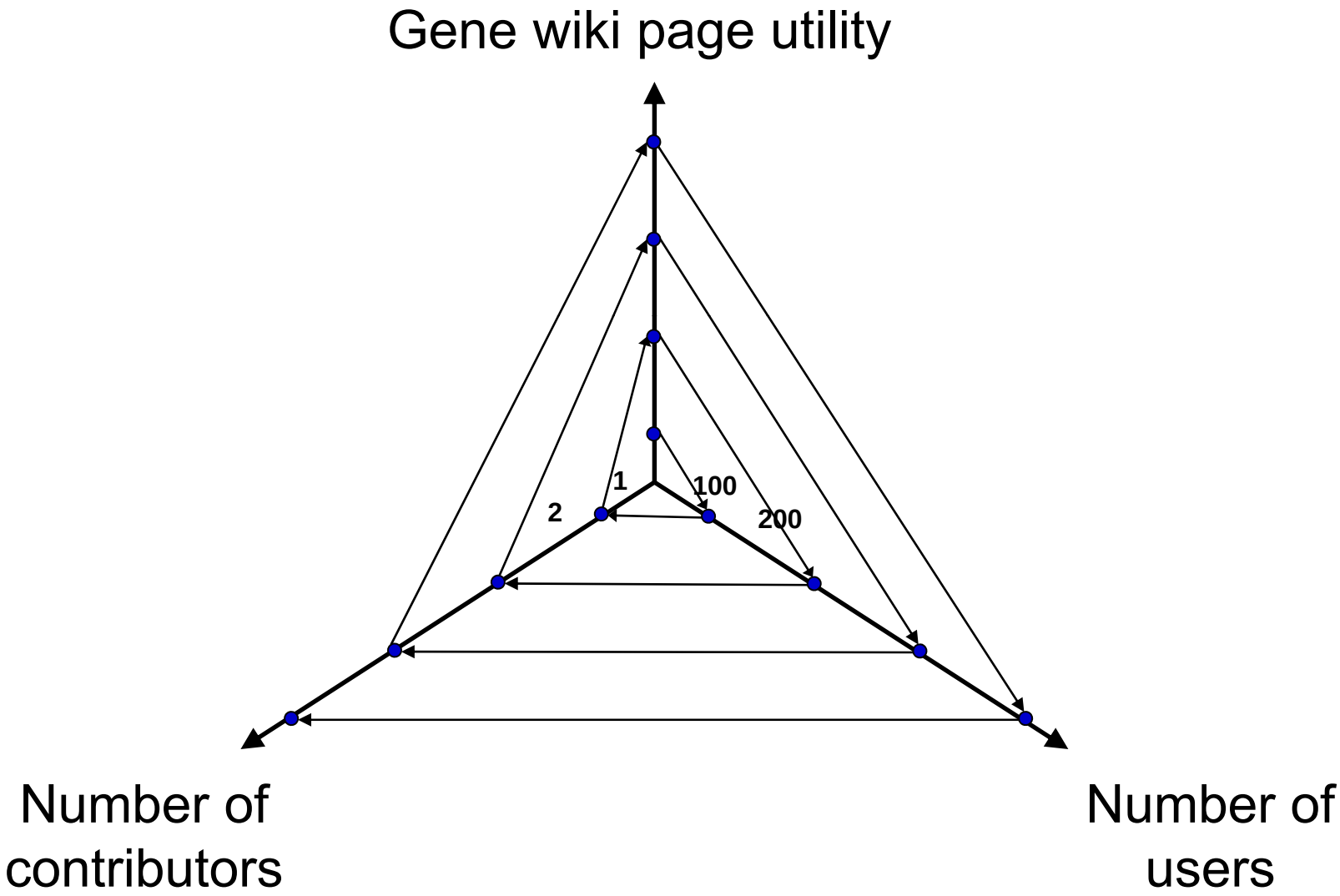
Gene Ontology annotations

Tissue expression pattern

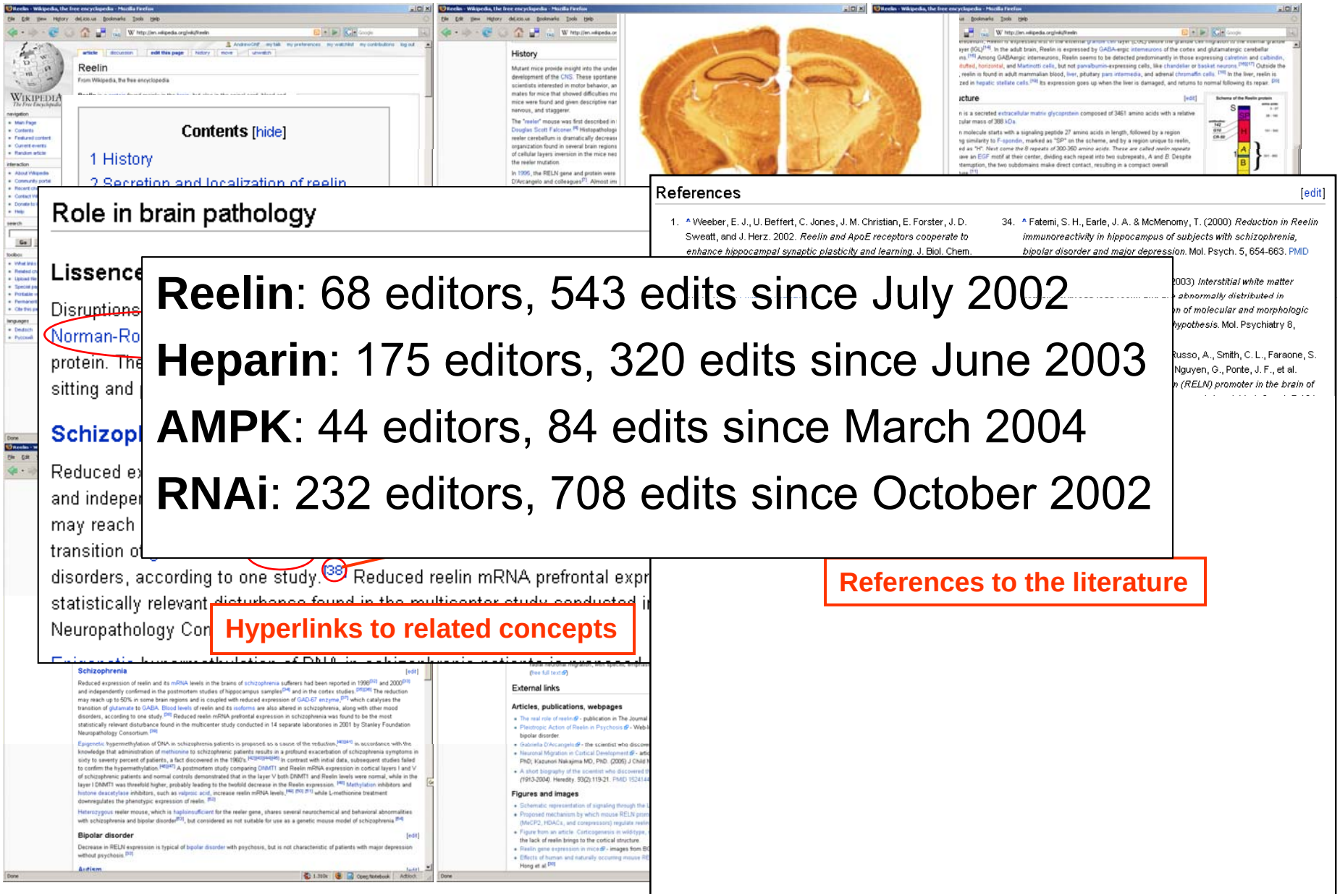
Links to structured databases

Idea: Jan 2007
Prototype: Mar 2007
Ver 1.0 bot: Nov 2007

Wiki success depends on a positive feedback



A review article for every gene is powerful



Filtering, extracting, and summarizing PubMed

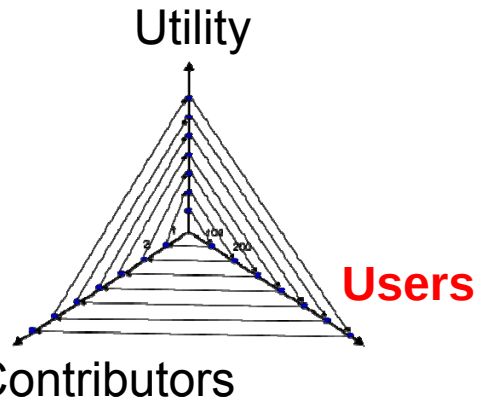
Gene Wiki has a diverse critical mass of readers

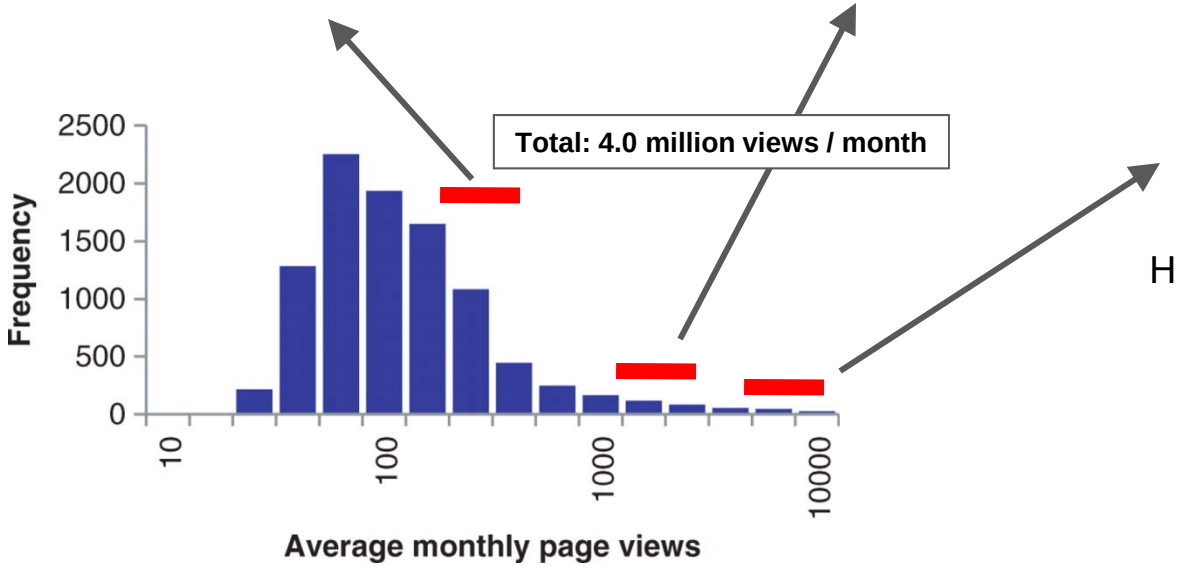
Rank 1001-1010: Specialists

CSDA
CNTNAP2
IGSF8
Adenosine A3 receptor
RYR1
ETV6
Small heterodimer partner
5-HT1D receptor
TRPC6
Interleukin-6 receptor

Rank 101-110: Scientists

Tau protein
Interleukin 10
APC
C-Met
Factor V
Interleukin 8
CD44
Histamine H1 receptor
Kappa Opioid receptor
Dihydrofolate reductase

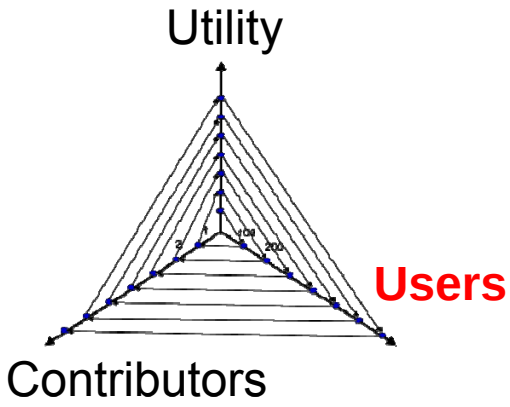


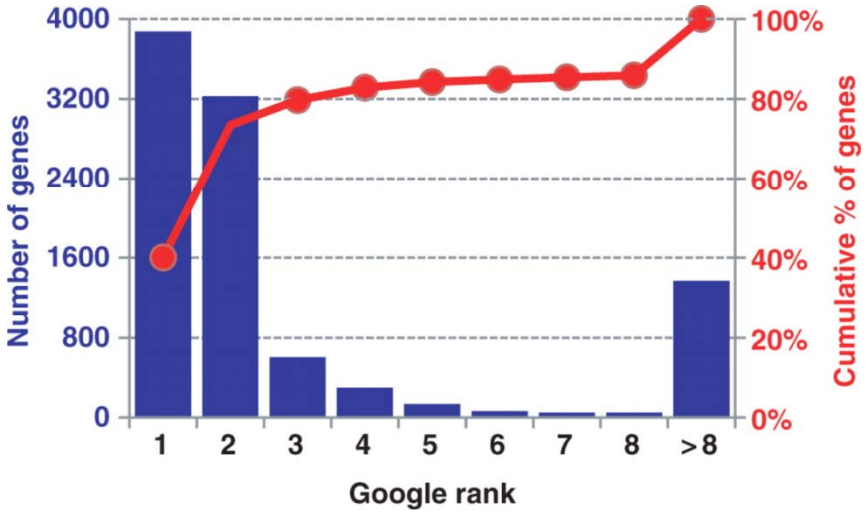


Rank 1-10: Laypeople

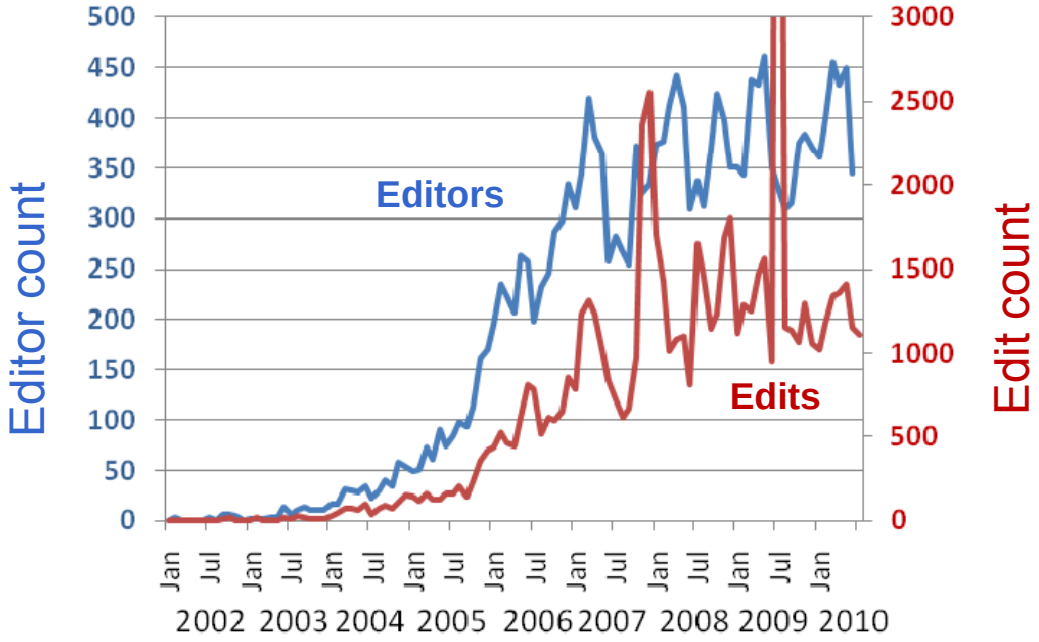
Insulin
Titin
Human chorionic gonadotropin
Vasopressin
ANKH
CLOCK
Catalase
Erythropoietin
Glucagon
Parathyroid hormone

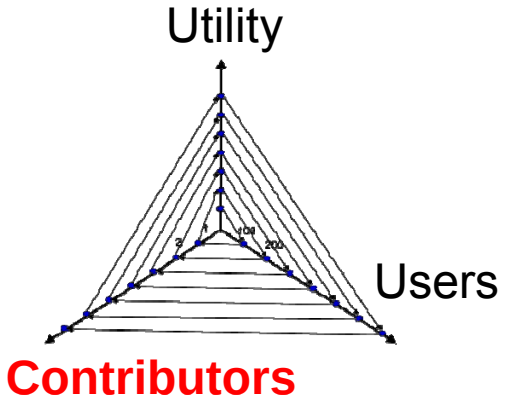
Readership is poised to grow





The Gene Wiki has a critical mass of editors





In Jan – Jun 2010 …

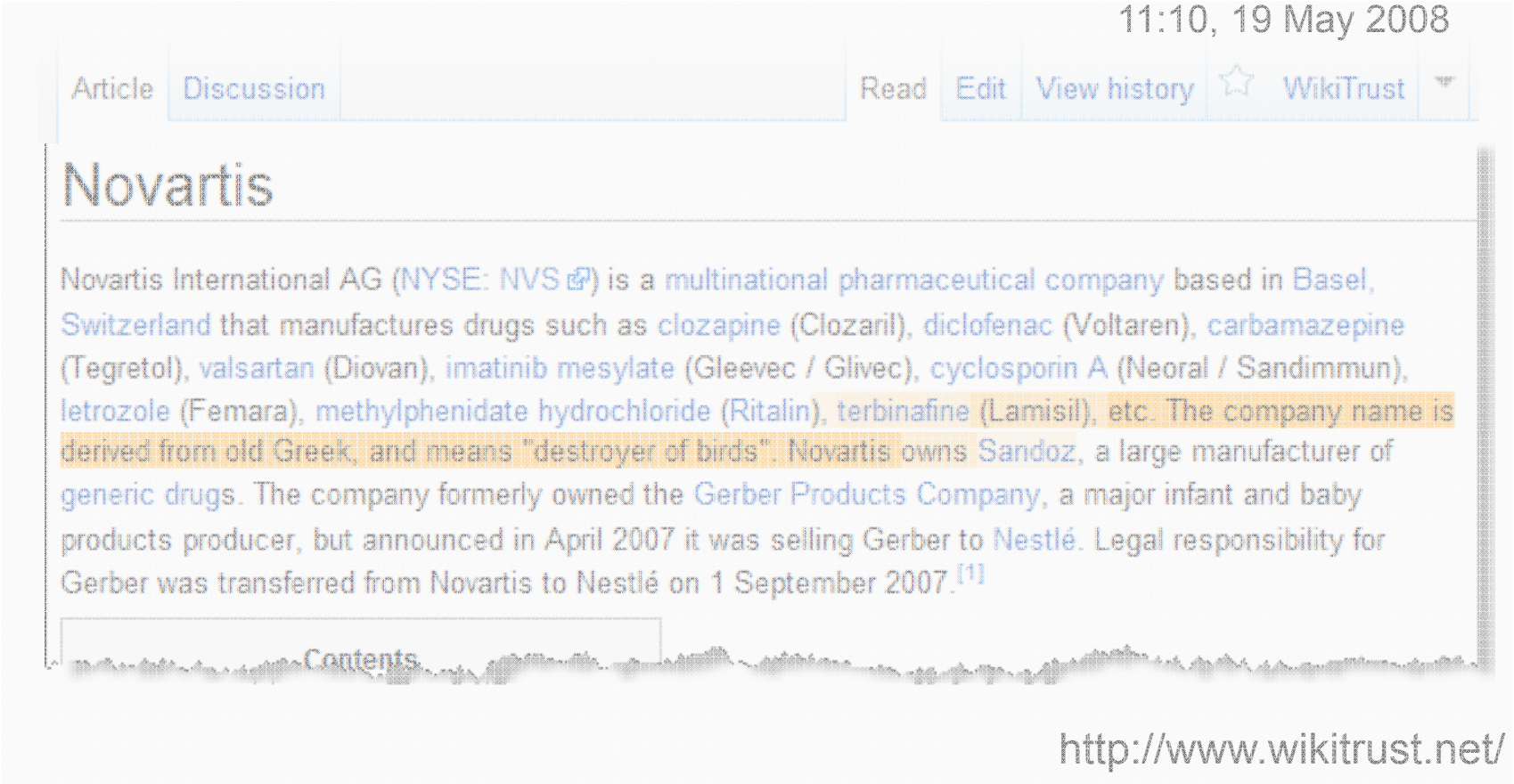
… 7474 edits were made by 2109 unique users

… total increase in text ≈ 20 *PLoS Biology* research articles

Making the Gene Wiki more reliable

WikiTrust principles:

- Authors who contribute long-lived content tend to be trustworthy
- Added content initially inherits the trust of its author
- Content persisting over time becomes more trusted
- Highlight untrusted text

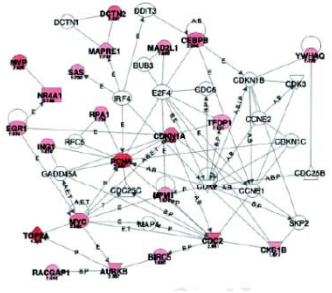


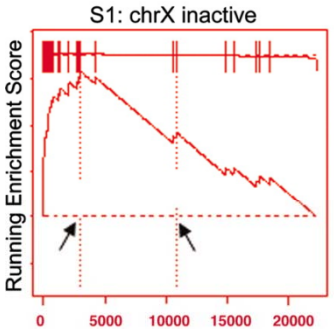
Making the Gene Wiki more computable

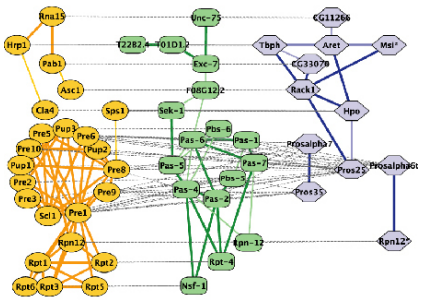
Structured annotations

Function	Evidence
metal ion binding	IEA
peptidase activity	IEA
protein serine/threonine/tyrosine kinase activity	ISS
serine-type peptidase activity	IEA

Process	Evidence
axon guidance	ISS
brain development	ISS
cell adhesion	IEA
cell morphogenesis involved in differentiation	ISS
central nervous system development	ISS
cerebral cortex tangential migration	ISS
glial cell differentiation	ISS
multicellular organismal development	IEA
neuron migration	ISS
peptidyl-tyrosine phosphorylation	ISS
positive regulation of protein kinase activity	ISS
positive regulation of small GTPase mediated signal transduction	ISS
response to pain	ISS
spinal cord patterning	ISS

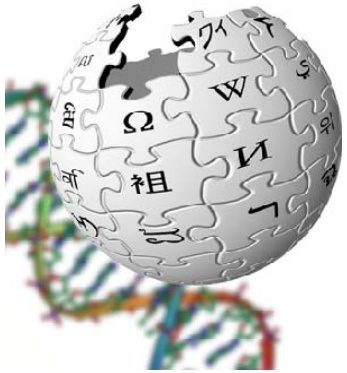






Making the Gene Wiki more computable (Part I)

Generation | **Representation** | **Utilization**



Unstructured

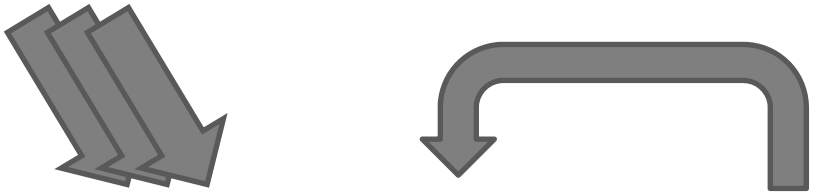
Structured

"Semantic Wiki Links"

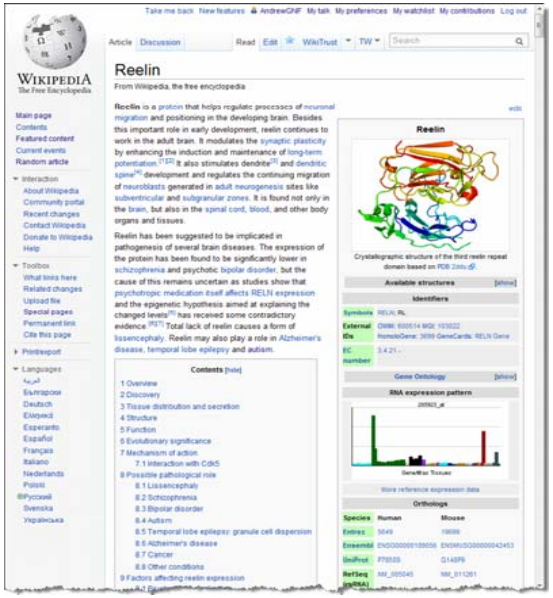
Semantic
Gene Wiki

See poster from
Salvatore Loguercio

Making the Gene Wiki more computable (Part II)



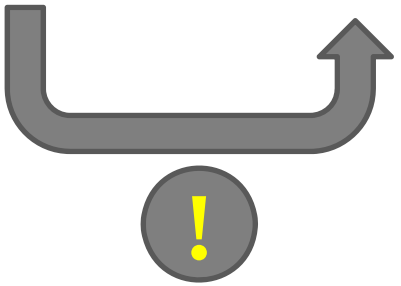
Free text



Structured annotations

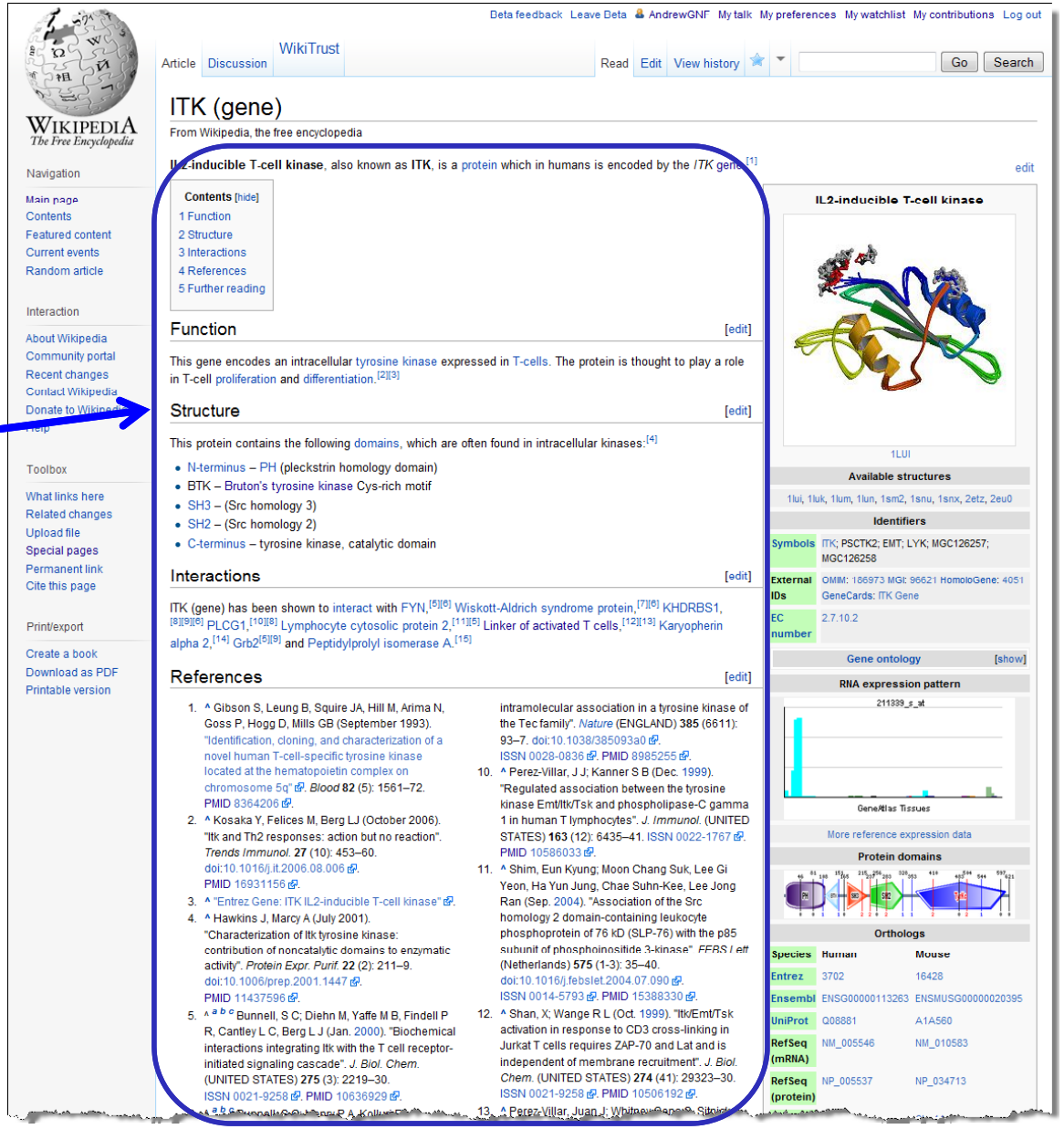
Function	Evidence
metal ion binding	IEA
peptidase activity	IEA
protein serine/threonine/tyrosine kinase activity	ISS
serine-type peptidase activity	IEA

Process	Evidence
axon guidance	ISS
brain development	ISS
cell adhesion	IEA
cell morphogenesis involved in differentiation	ISS
central nervous system development	ISS
cerebral cortex tangential migration	ISS
glial cell differentiation	ISS
multicellular organismal development	IEA
neuron migration	ISS
peptidyl-tyrosine phosphorylation	ISS
positive regulation of protein kinase activity	ISS
positive regulation of small GTPase mediated signal transduction	ISS
response to pain	ISS
spinal cord patterning	ISS



Mining community-contributed content

Text, links, and citations



Deta feedback Leave Deta AndrewGNF My talk My preferences My watchlist My contributions Log out

Article Discussion WikiTrust Read Edit View history Go Search

WIKIPEDIA The Free Encyclopedia

Navigation
Main page
Contents
Featured content
Current events
Random article

Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help

Toolbox
What links here
Related changes
Upload file
Special pages
Permanent link
Cite this page

Print/export
Create a book
Download as PDF
Printable version

ITK (gene)

From Wikipedia, the free encyclopedia

IL2-inducible T-cell kinase, also known as **ITK**, is a protein which in humans is encoded by the *ITK* gene.[1]

edit

Contents [hide]
1 Function
2 Structure
3 Interactions
4 References
5 Further reading

Function [edit]

This gene encodes an intracellular tyrosine kinase expressed in T-cells. The protein is thought to play a role in T-cell proliferation and differentiation.[2][3]

Structure [edit]

This protein contains the following domains, which are often found in intracellular kinases:[4]

- N-terminus – PH (pleckstrin homology domain)
- BTK – Bruton's tyrosine kinase Cys-rich motif
- SH3 – (Src homology 3)
- SH2 – (Src homology 2)
- C-terminus – tyrosine kinase, catalytic domain

Interactions [edit]

ITK (gene) has been shown to interact with FYN,[5][6] Wiskott-Aldrich syndrome protein,[7][6] KHDRBS1,[8][9][6] PLCG1,[10][8] Lymphocyte cytosolic protein 2,[11][5] Linker of activated T cells,[12][13] Karyopherin alpha 2,[14] Grb2[5][9] and Peptidylprolyl isomerase A.[15]

References [edit]

1. ^ Gibson S, Leung B, Squire JA, Hill M, Arima N, Goss P, Hogg D, Mills GB (September 1993). "Identification, cloning, and characterization of a novel human T-cell-specific tyrosine kinase located at the hematopoietin complex on chromosome 5q". *Blood* **82** (5): 1561–72. PMID 8364206.
2. ^ Kosaka Y, Felices M, Berg LJ (October 2006). "Itk and Th2 responses: action but no reaction". *Trends Immunol.* **27** (10): 453–60. doi:10.1016/j.it.2006.08.006. PMID 16931156.
3. ^ "Entrez Gene: ITK IL2-inducible T-cell kinase".
4. ^ Hawkins J, Marcy A (July 2001). "Characterization of Itk tyrosine kinase: contribution of noncatalytic domains to enzymatic activity". *Protein Expr. Purif.* **22** (2): 211–9. doi:10.1006/prep.2001.1447. PMID 11437596.
5. ^ a b c Bunnell, S C; Diehn M, Yaffe M B, Findell P R, Cantley L C, Berg L J (Jan. 2000). "Biochemical interactions integrating Itk with the T cell receptor-initiated signaling cascade". *J. Biol. Chem.* (UNITED STATES) **275** (3): 2219–30. ISSN 0021-9258. PMID 10636929.
10. ^ Perez-Villar, J J; Kanner S B (Dec. 1999). "Regulated association between the tyrosine kinase Emt/Itk/Tsk and phospholipase-C gamma 1 in human T lymphocytes". *J. Immunol.* (UNITED STATES) **163** (12): 6435–41. ISSN 0022-1767. PMID 10586033.
11. ^ Shim, Eun Kyung; Moon Chang Suk, Lee Gi Yeon, Ha Yun Jung, Chae Suhn-Kee, Lee Jong Ran (Sep. 2004). "Association of the Src homology 2 domain-containing leukocyte phosphoprotein of 76 kD (SLP-76) with the p85 subunit of phosphoinositide 3-kinase". *FEBS Lett.* (Netherlands) **575** (1-3): 35–40. doi:10.1016/j.febslet.2004.07.090. ISSN 0014-5793. PMID 15388330.
12. ^ Shan, X; Wange R L (Oct. 1999). "Itk/Emt/Tsk activation in response to CD3 cross-linking in Jurkat T cells requires ZAP-70 and Lat and is independent of membrane recruitment". *J. Biol. Chem.* (UNITED STATES) **274** (41): 29323–30. ISSN 0021-9258. PMID 10506192.

IL2-inducible T-cell kinase

1LUI

Available structures
1lui, 1luk, 1lum, 1lun, 1sm2, 1snu, 1snx, 2etz, 2eu0

Identifiers

Symbols	ITK; PSCTK2; EMT; LYK; MGC126257; MGC126258
External IDs	OMIM: 186973 MGI: 96621 HomoloGene: 4051 GeneCards: ITK Gene
EC number	2.7.10.2

Gene ontology [show]

RNA expression pattern

211339_s_at

GeneAtlas Tissues

More reference expression data

Protein domains

Orthologs

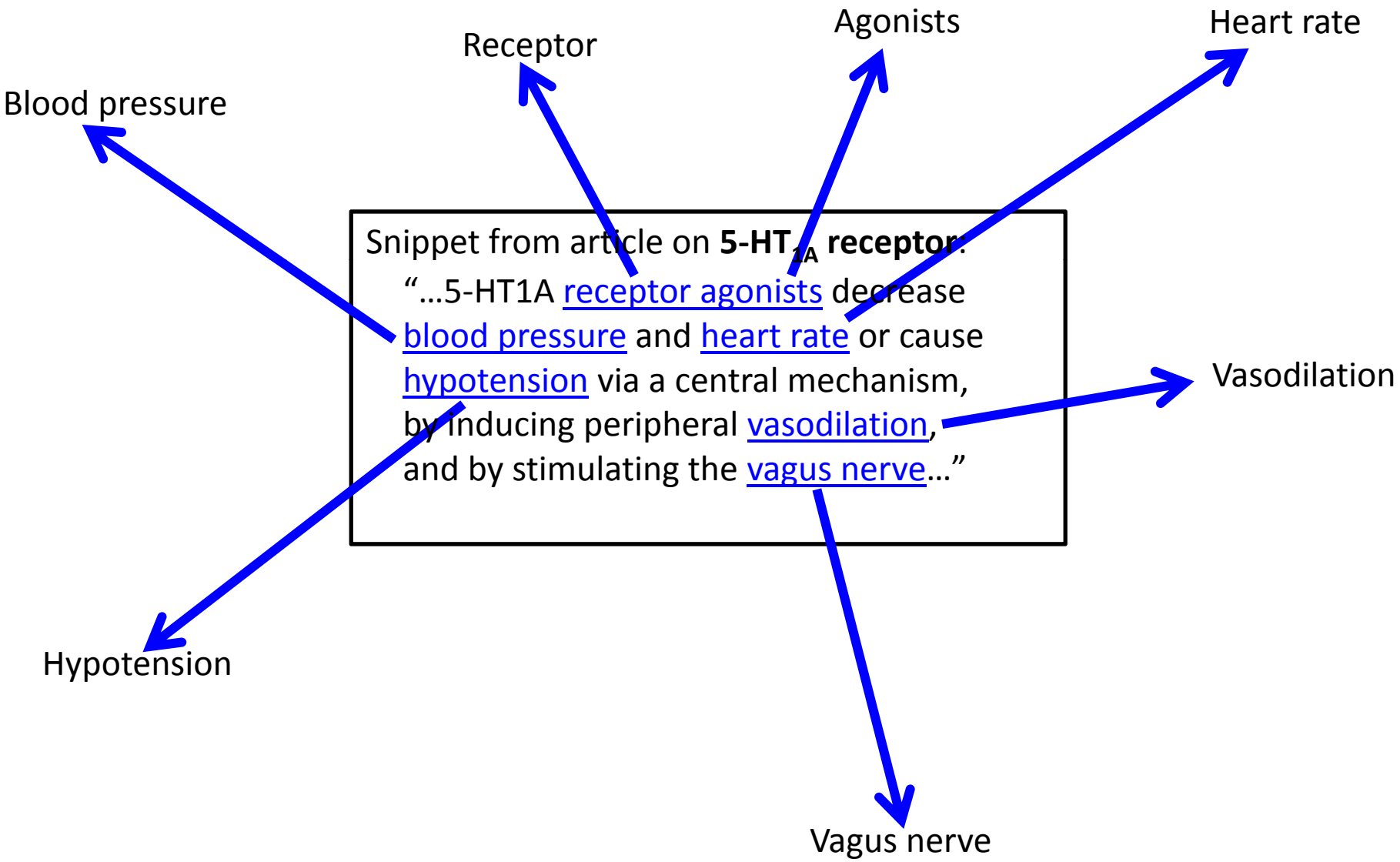
Species	Human	Mouse
Entrez	3702	16428
Ensembl	ENSG00000113263	ENSMUSG00000020395
UniProt	Q08881	A1A560
RefSeq (mRNA)	NM_005546	NM_010583
RefSeq (protein)	NP_005537	NP_034713

Example text from **5-HT$_{1A}$ receptor**

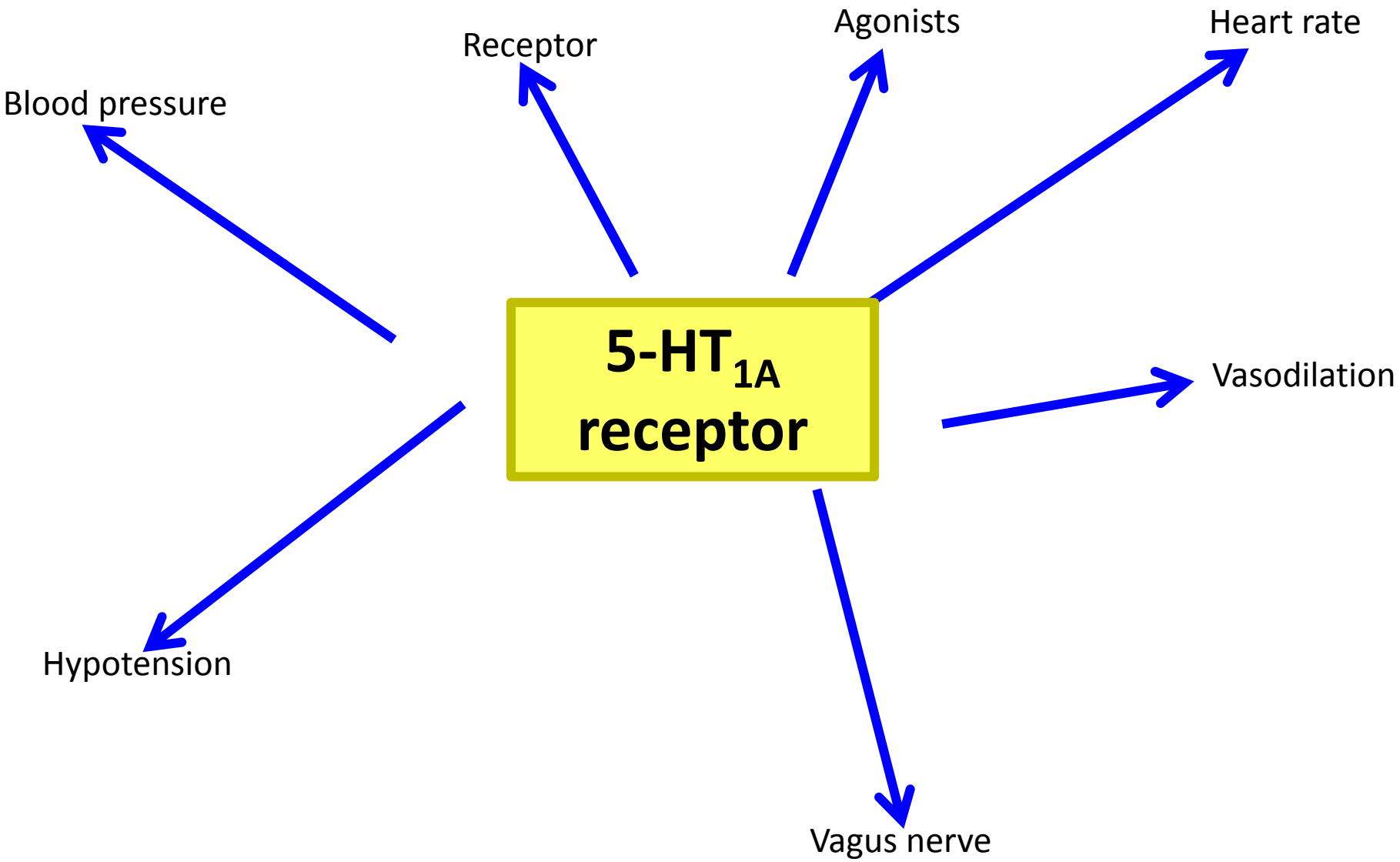
Snippet from article on **5-HT$_{1A}$ receptor**:

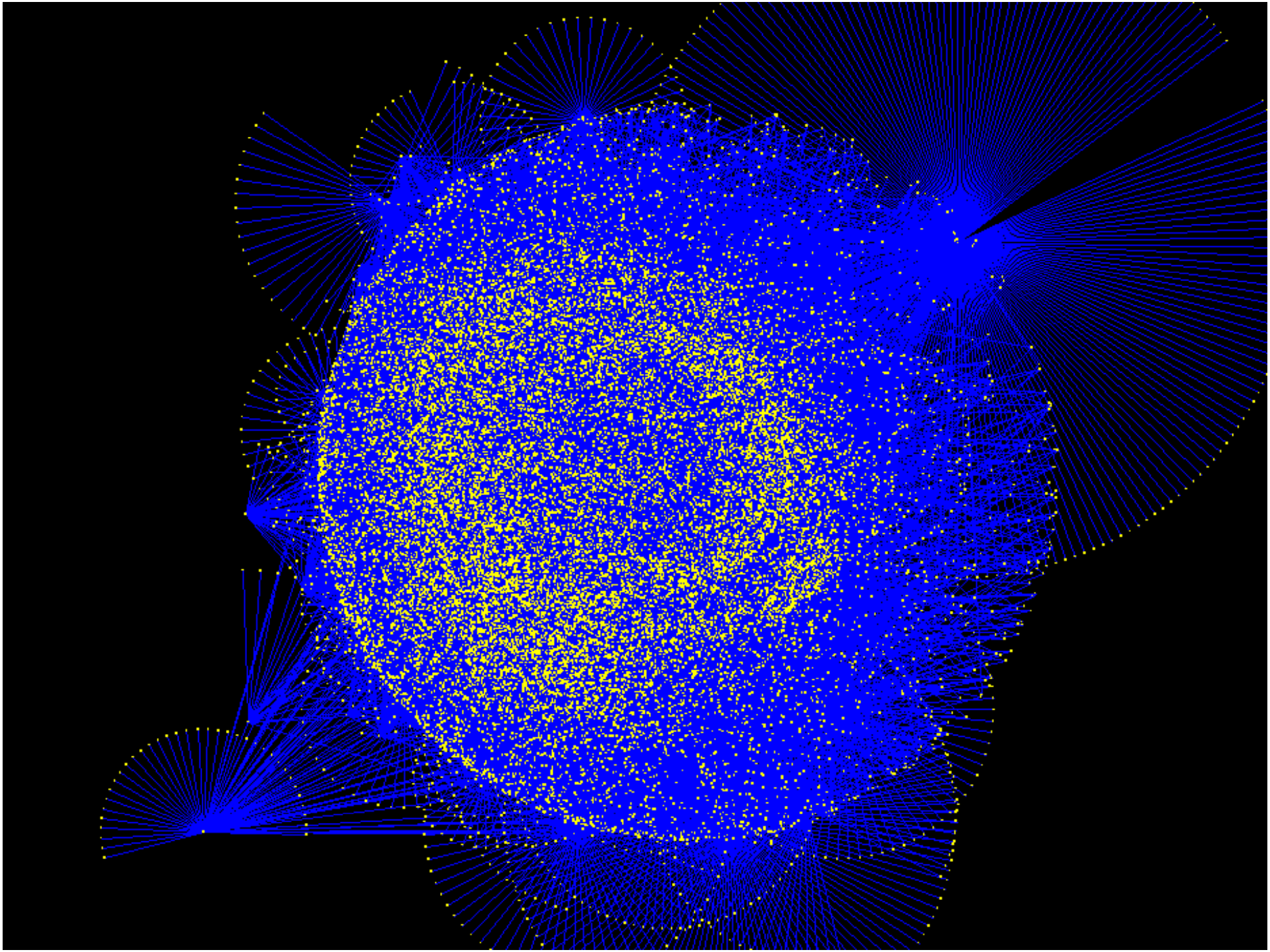
"…5-HT1A receptor agonists decrease blood pressure and heart rate or cause hypotension via a central mechanism, by inducing peripheral vasodilation, and by stimulating the vagus nerve…"

Example Hypertext from **5-HT$_{1A}$ receptor**

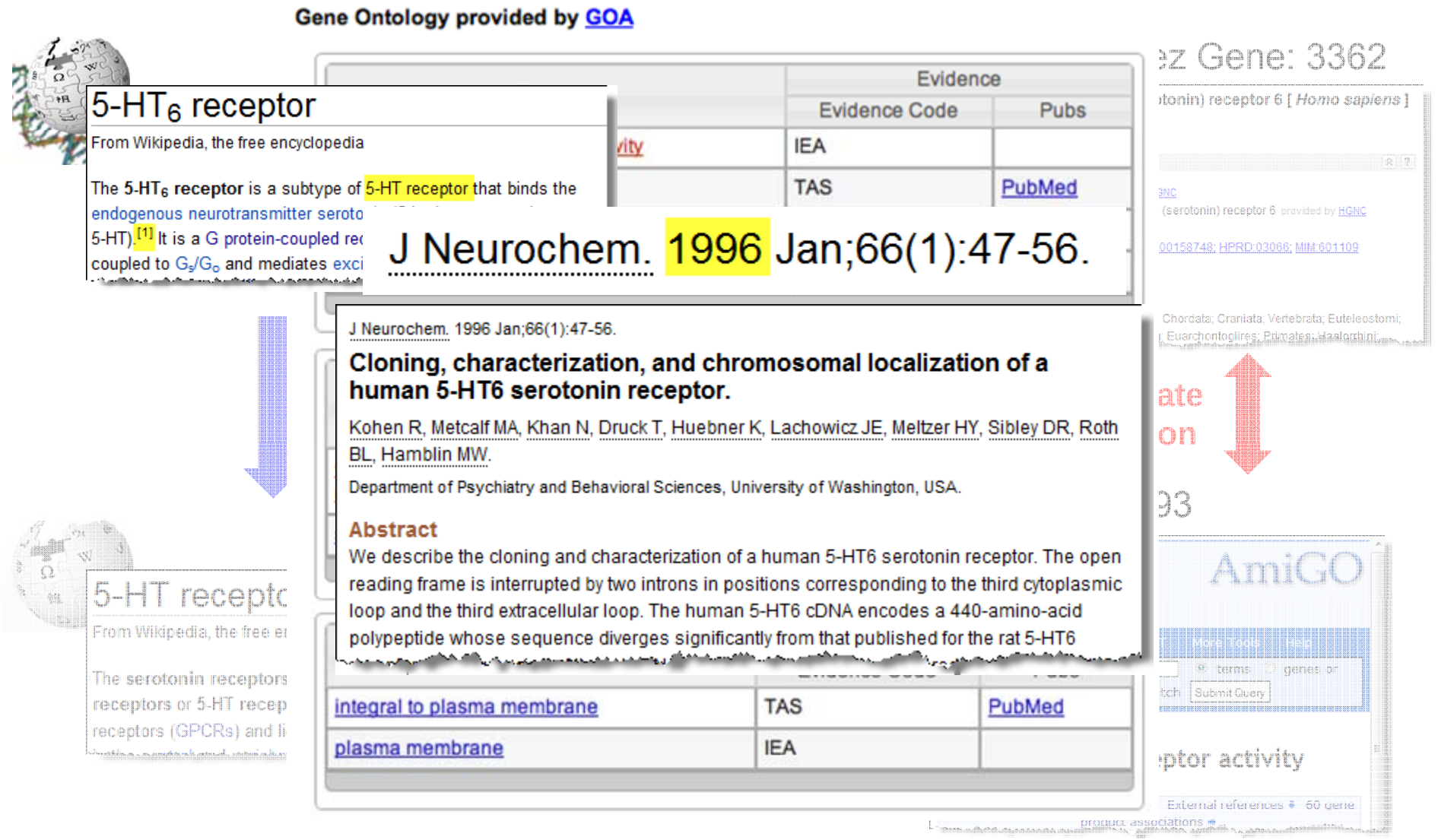


Example Hyper text from **5-HT$_{1A}$ receptor**

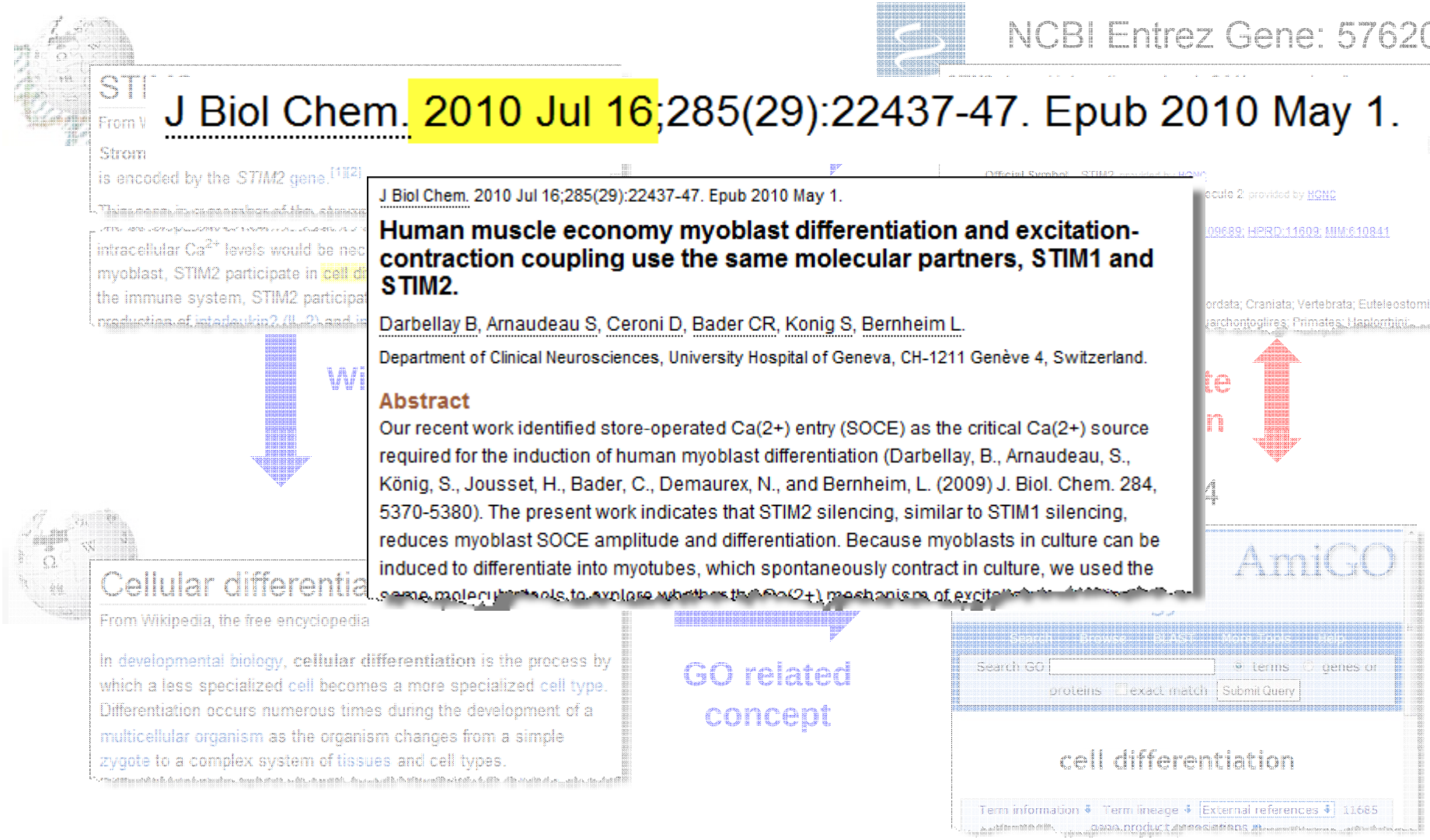




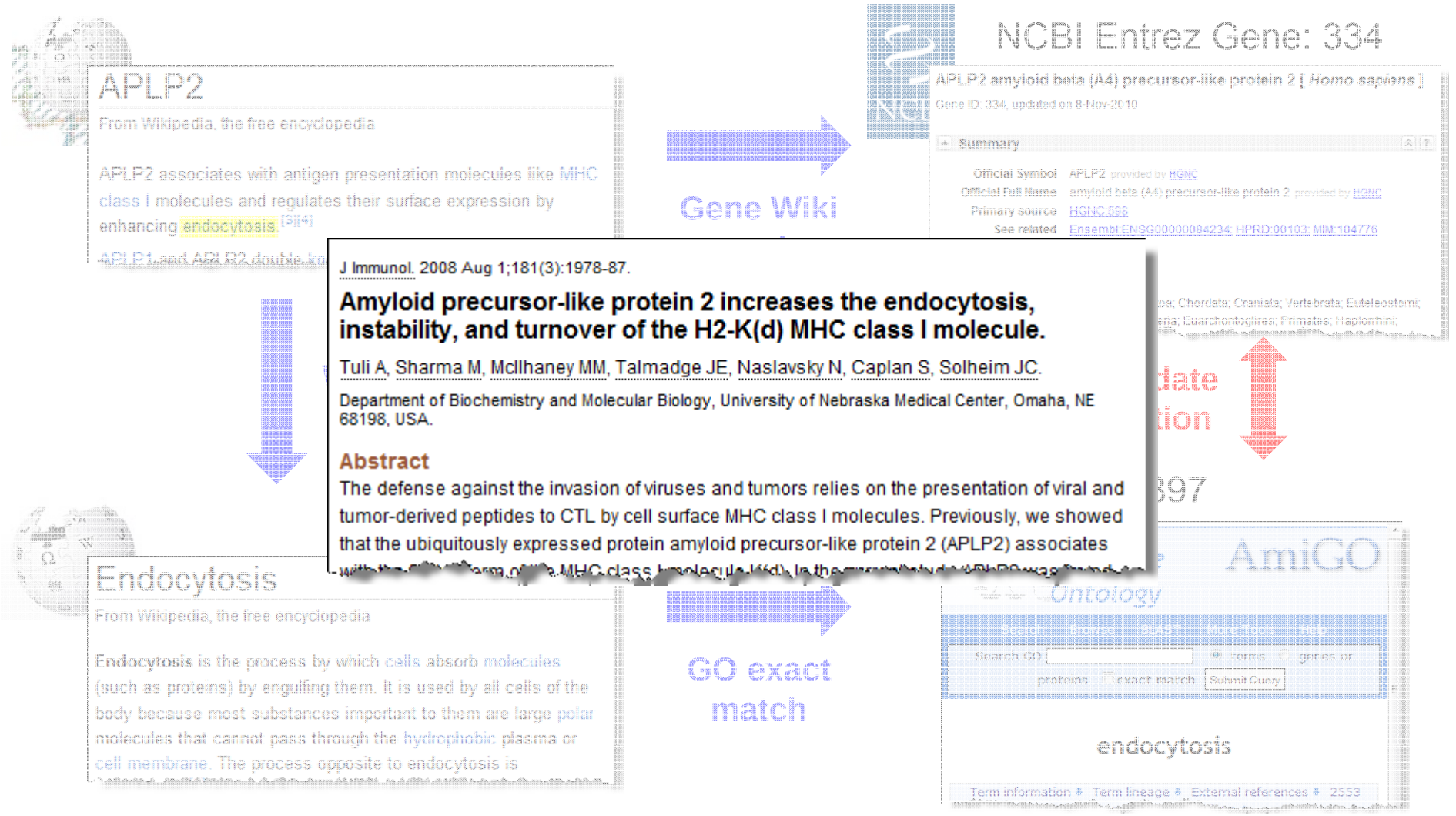
Re-discovering common knowledge



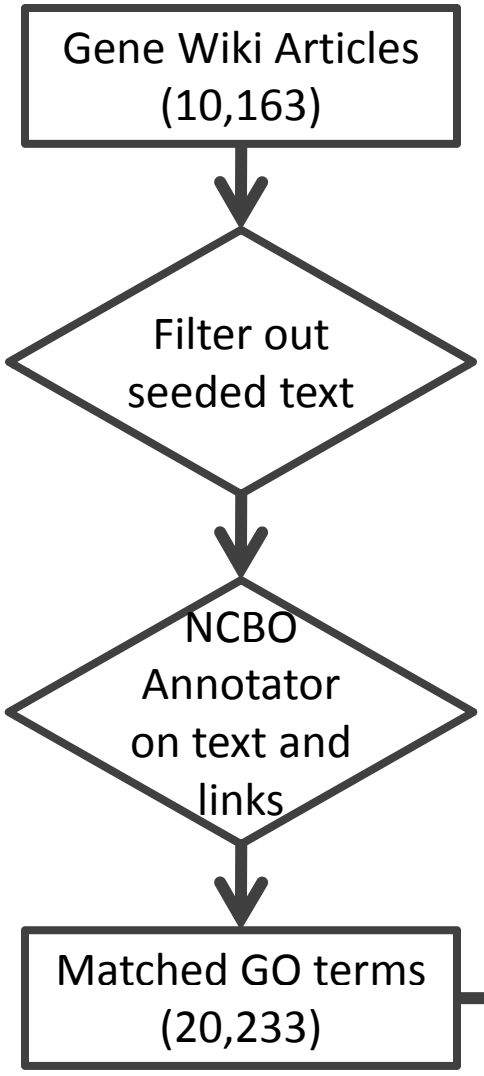
Mining the most recent literature



Filling the gaps in gene annotation

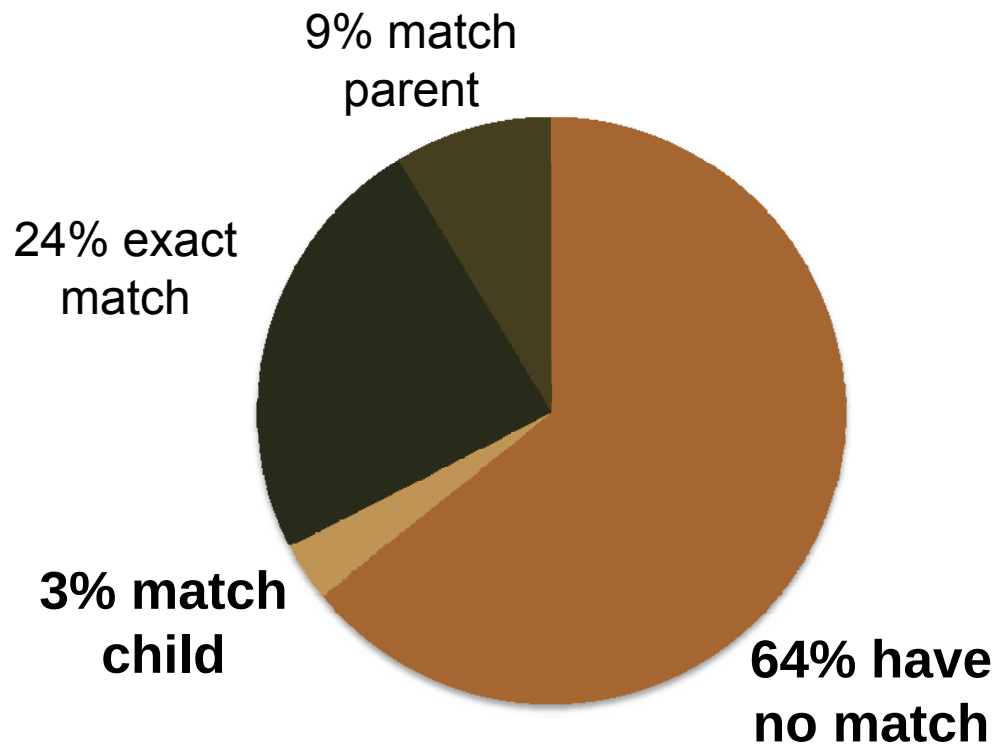


Associations mined from links and text



Compare to GO database

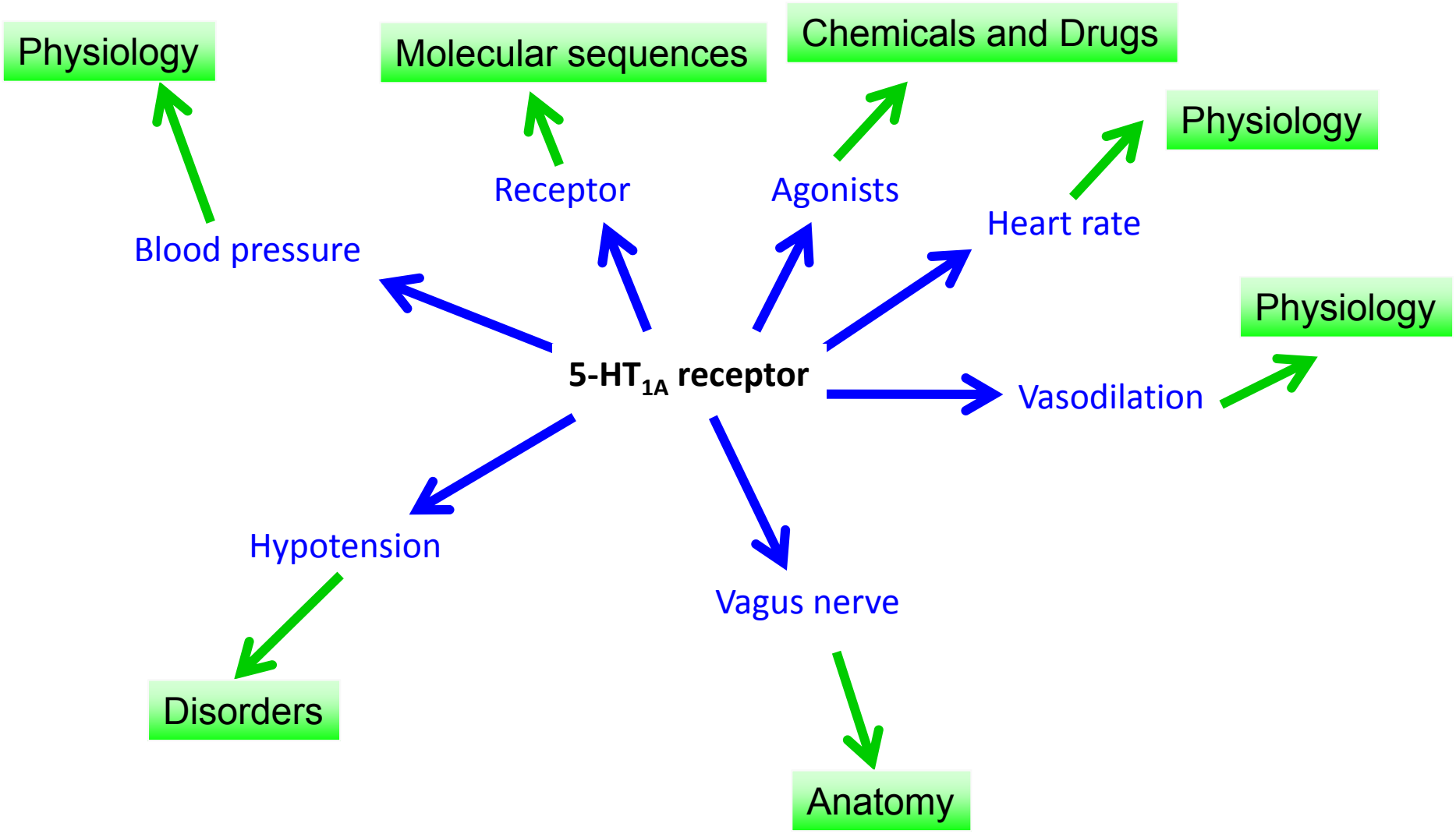
13,636 candidate annotations



Prioritization

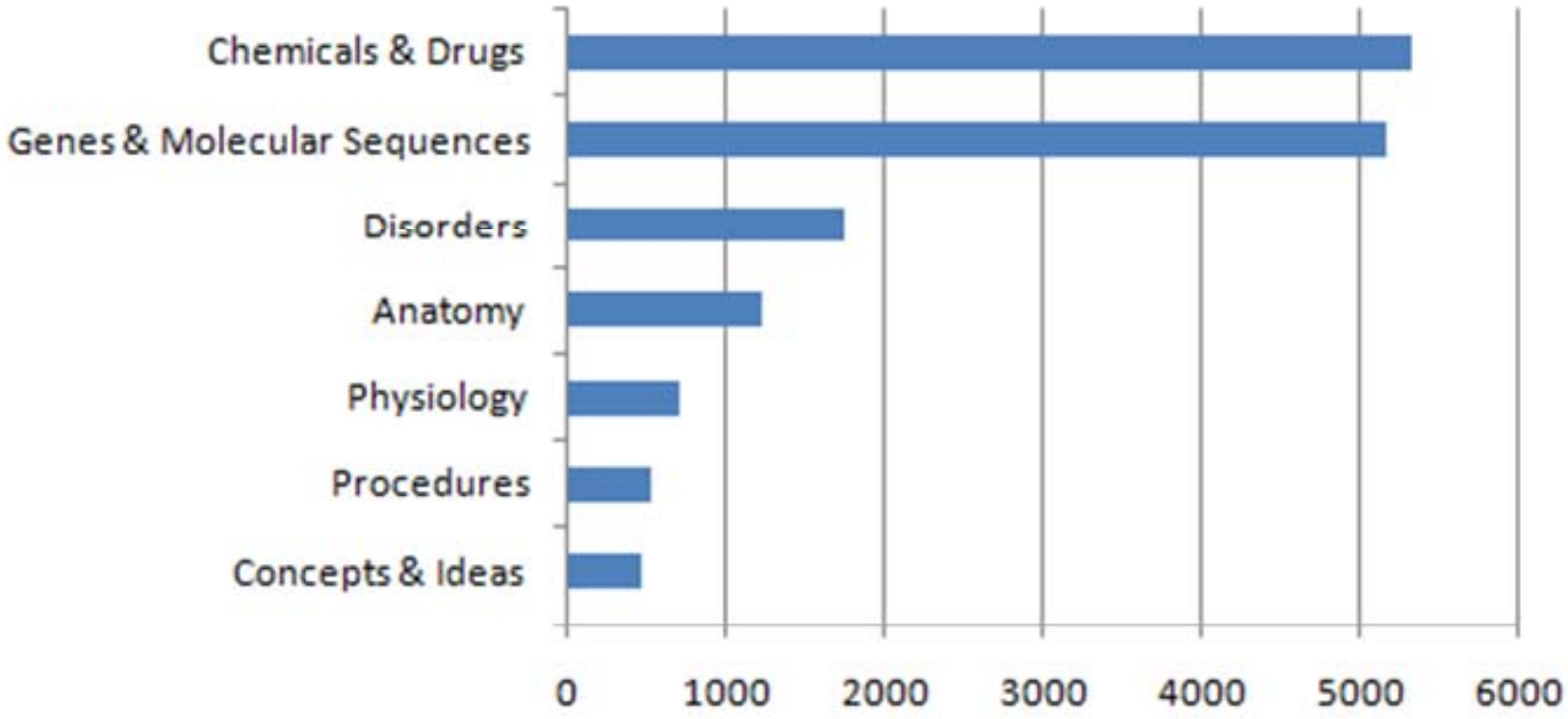
- Orthology
- Literature mining
- Computational prediction

Not just Gene Ontology mappings



UMLS (Unified Medical Language System) Semantic Groups

Mapping to many biomedical semantic groups



For community-based science, data is king

Data without **structure** is valuable,
but **structure** without data is not.

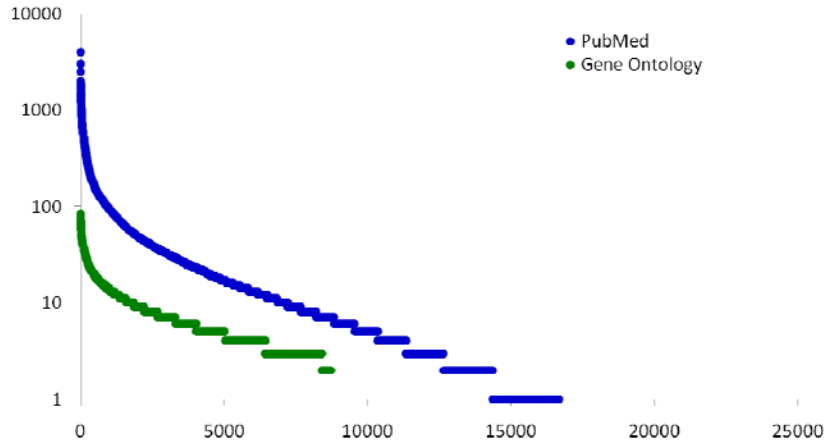
For community-based science, data is king

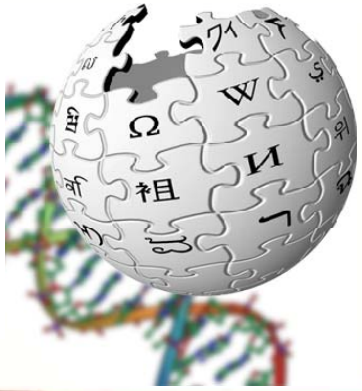
Data without **X** is valuable,
but **X** without data is not.

X =

		Wikipedia
Copy-editing	→	WP:MCB, Boghog
Figures	→	Artists and illustrators
Structure	→	Wiki links, infoboxes
Citations	→	DOI bot, CitationBot
Provenance	→	WikiTrust

The Gene Wiki
successfully harnesses the
Long Tail of scientists
for community annotation of
gene function





Collaborators

Doug Howe, ZFIN
Salvatore Loguercio (*), TU Dresden
John Hogenesch, UPenn
Angel Pizzaro, UPenn
Faramarz Valafar, SDSU
Pierre Lindenbaum,
Fondation Jean Dausset
Michael Martone, Rush
Konrad Koehler, Karo Bio
Many Wikipedia editors
WP:MCB Project

(*) See poster on WikiSyncing

Thanks also to:
Christopher Roeder, U.Col.
Conrad Plake, GoGene

Group members

Serge Batalov
Ghislain Bonamy
Ben Good
Jon Huss
Jeff Janes
Marc Leglise
Ian Macleod
Camilo Orozco
Julia Turner
Chunlei Wu

Funding and Support

Genomics Institute of the Novartis Research Foundation

NIGMS

(BioGPS: GM83924, Gene Wiki: GM089820)