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


*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
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
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

Content produced

Short Head

Content produced

Long Tail

Contributors (sorted)

News reporting:
  Newspapers
  TV/Hollywood
  Consumer reports
  Food critics
  Olympics

Video:
  YouTube
  Amazon reviews
  Yelp
  American Idol

Product reviews:
  Blogs

Food reviews:
  Judging:
  Olympics
Wikipedia is reasonably accurate.

**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

<table>
<thead>
<tr>
<th></th>
<th>Articles</th>
<th>Words (millions)</th>
<th>Words/article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>20,000,000</td>
<td>1,000</td>
<td>435</td>
</tr>
<tr>
<td>Britannica Online</td>
<td>120,000</td>
<td>55</td>
<td>370</td>
</tr>
</tbody>
</table>

We can harness the **Long Tail of scientists** to directly participate in the gene annotation process.
Gene “stubs” seed community contributions

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

Role in brain pathology

Lissencephaly

Disruptions of the Reelin protein have been linked to lissencephaly, a brain disorder characterized by the absence of the cerebral cortex and the presence of abnormal, smooth brain surfaces (normocephaly). This condition is associated with severe developmental delays and intellectual disabilities.

Heparin

Reduced expression of heparin has been observed in the brains of individuals with schizophrenia, suggesting a potential role for heparin in the pathology of this disorder.

AMPK

AMPK activation and phosphorylation of the AMPK substrate is crucial for maintaining正常的 brain function and may be a target for future therapeutic interventions.

RNAi

RNAi technology has been used to investigate the role of specific genes in brain disorders, offering new insights into the molecular mechanisms underlying these conditions.

Hyperlinks to related concepts

References to the literature

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

Utility

Contributors

Users
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 $\approx 20$ 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

Novartis

Novartis International AG (NYSE: NVS) is a multinational pharmaceutical company based in Basel, Switzerland that manufactures drugs such as clozapine (Clozaril), diclofenac (Voltaren), carbamazepine (Tegretol), valsartan (Diovan), imatinib mesylate (Gleevec / Glivec), cyclosporin A (Neoral / Sandimmun), letrozole (Femara), methylphenidate hydrochloride (Ritalin), terbinfine (Lamisil), etc. The company name is derived from old Greek, and means "destroyer of birds". Novartis owns Sandoz, a large manufacturer of generic drugs. The company formerly owned the Gerber Products Company, a major infant and baby products producer, but announced in April 2007 it was selling Gerber to Nestlé. Legal responsibility for Gerber was transferred from Novartis to Nestlé on 1 September 2007.[1]

http://www.wikitrust.net/
Making the Gene Wiki more computable

### Structured annotations

<table>
<thead>
<tr>
<th>Function</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>metal ion binding</td>
<td>IEA</td>
</tr>
<tr>
<td>peptidase activity</td>
<td>IEA</td>
</tr>
<tr>
<td>protein tyrosine/threonine/tyrosine kinase activity</td>
<td>ISSS</td>
</tr>
<tr>
<td>serine/threonine peptidase activity</td>
<td>IEA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>axon guidance</td>
<td>ISSS</td>
</tr>
<tr>
<td>brain development</td>
<td>ISSS</td>
</tr>
<tr>
<td>cell adhesion</td>
<td>IEA</td>
</tr>
<tr>
<td>cell morphogenesis involved in differentiation</td>
<td>ISSS</td>
</tr>
<tr>
<td>central nervous system development</td>
<td>ISSS</td>
</tr>
<tr>
<td>cerebral cortex tangential migration</td>
<td>ISSS</td>
</tr>
<tr>
<td>glut cell differentiation</td>
<td>ISSS</td>
</tr>
<tr>
<td>multicellular organismal development</td>
<td>IEA</td>
</tr>
<tr>
<td>neuron migration</td>
<td>ISSS</td>
</tr>
<tr>
<td>peptidyl-tyrosine phosphorylation</td>
<td>ISSS</td>
</tr>
<tr>
<td>positive regulation of protein kinase activity</td>
<td>ISSS</td>
</tr>
<tr>
<td>positive regulation of small GTPase-mediated signal transduction</td>
<td>ISSS</td>
</tr>
<tr>
<td>response to pain</td>
<td>ISSS</td>
</tr>
<tr>
<td>spinal cord patterning</td>
<td>ISSS</td>
</tr>
</tbody>
</table>
Making the Gene Wiki more computable (Part I)

Generation → Representation → Utilization

Unstructured
Structured

“Semantic Wiki Links”

Data mining/statistics

See poster from Salvatore Loguercio
Making the Gene Wiki more computable (Part II)
Mining community-contributed content

Text, links, and citations
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...”
Snippet from article on 5-HT$_{1A}$ receptor:
“...5-HT$_{1A}$ 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 text from $5$-HT$_{1A}$ receptor

- Blood pressure
- Hypotension
- Receptor
- Agonists
- Heart rate
- Vasodilation
- Vagus nerve
Re-discovering common knowledge

**5-HT$_6$ receptor**

The 5-HT$_6$ receptor is a subtype of 5-HT receptor that binds the endogenous neurotransmitter serotonin 5-HT$_6$. It is a G protein-coupled receptor coupled to G$_{q/11}$ and mediates excitation.


*Cloning, characterization, and chromosomal localization of a human 5-HT$_6$ serotonin receptor.*

Kohen R, Metcalfe MA, Khan N, Druck T, Huebner K, Lachowicz JE, Meltzer HY, Sibley DR, Roth BL, Hamblin MW.

Department of Psychiatry and Behavioral Sciences, University of Washington, USA.

**Abstract**

We describe the cloning and characterization of a human 5-HT$_6$ serotonin receptor. The open reading frame is interrupted by two introns in positions corresponding to the third cytoplasmic loop and the third extracellular loop. The human 5-HT$_6$ cDNA encodes a 440-amino-acid polypeptide whose sequence diverges significantly from that published for the rat 5-HT$_6$.
Human muscle economy myoblast differentiation and excitation-contraction coupling use the same molecular partners, STIM1 and STIM2.

Darbellay B, Arnaudeau S, Ceroni D, Bader CR, König S, Bernheim L.

Department of Clinical Neurosciences, University Hospital of Geneva, CH-1211 Genève 4, Switzerland.

Abstract

Our recent work identified a critical role of Ca(2+) entry (COCE) as the critical Ca(2+) source required for the induction of human myoblast differentiation (Darbellay, B., Arnaudeau, S., König, S., Jousset, H., Bader, C., Demaursx, N., and Bernheim, L. (2009). J. Biol. Chem. 284, 5370-5380). The present work indicates that STIM2 silencing, similar to STIM1 silencing, reduces myoblast SOCE amplitude and differentiation. Because myoblasts in culture can be induced to differentiate into myotubes, which spontaneously contract in culture, we used the same molecular tools to explore whether the Ca(2+) mechanism of excitation-contraction coupling can be altered to affect differentiation.
Filling the gaps in gene annotation


Amyloid precursor-like protein 2 increases the endocytosis, instability, and turnover of the H2-K(d) MHC class I molecule.


Department of Biochemistry and Molecular Biology, University of Nebraska Medical Center, Omaha, NE 68198, USA.

Abstract

The defense against the invasion of viruses and tumors relies on the presentation of viral and tumor-derived peptides to CTL by cell surface MHC class I molecules. Previously, we showed that the ubiquitously expressed protein amyloid precursor-like protein 2 (APLP2) associates with the class I MHC molecules HLA-A2 and HLA-B27 and increases their stability and cell surface expression. Here, we report that APLP2 also increases the endocytosis, instability, and turnover of the H2-K(d) MHC class I molecule.
Associations mined from links and text

- Gene Wiki Articles (10,163)
- Filter out seeded text
- NCBO Annotator on text and links
- Matched GO terms (20,233)
- Compare to GO database
- 13,636 candidate annotations

- 9% match parent
- 24% exact match
- 3% match child
- 64% have no match

Prioritization
- Orthology
- Literature mining
- Computational prediction
Not just Gene Ontology mappings

- Physiology
  - Blood pressure
  - Hypotension
  - Disorders
- Molecular sequences
  - Receptor
- Chemicals and Drugs
  - Agonists
  - Heart rate
  - Vasodilation
- Anatomy

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.

\[ X = \begin{align*}
\text{Copy-editing} & \rightarrow \text{WP:MCB, Boghog} \\
\text{Figures} & \rightarrow \text{Artists and illustrators} \\
\text{Structure} & \rightarrow \text{Wiki links, infoboxes} \\
\text{Citations} & \rightarrow \text{DOI bot, CitationBot} \\
\text{Provenance} & \rightarrow \text{WikiTrust}
\end{align*} \]

Wikipedia

Domain expert

Information scientist
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

Group members

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

(*) See poster on WikiSyncing

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

Funding and Support

Genomics Institute of the Novartis Research Foundation

(BioGPS: GM83924, Gene Wiki: GM089820)