Building a European literature hub for the life sciences

current status and future directions
Data-literature relationships aid discovery
Big Data and Publishing: EBI Context

• Data deposition required for publication
  • Sequences - Structures - Gene Expression
  • Organic, “grass roots”

• Curation by professionals
  • InterPro
  • UniProt

• Next-generation sequencing
  • 1000 Genomes - Cancer Genomes Project
  • Beyond 1-to-1 data-paper links
Percentage of articles in PMC that have associated supplemental data
Literature at the EBI

- Primary submissions linked to publications
- Collaboration with the EPO
- Curated Resources e.g. UniProt

- CiteXplore (2006)
- UKPubMed Central (2008)
UK PubMed Central

A collaboration between:

• EBI
• British Library
• University of Manchester
PMC, UKPMC, and PMC Canada: PMC International
Association of SUMO1 and UBC9 genotypes with tumor response in non-small-cell lung cancer treated with irinotecan-based chemotherapy.

(PMID:19859084)
Han JY, Lee GK, Yoo SY, Yoon SJ, Cho EY, Kim HT, Lee JS
The Pharmacogenomics Journal [2010, 10(2):86-93]
Irinotecan induces small ubiquitin-like modifier (SUMO)-1 conjugation to topoisomerase-I, leading to enhanced sensitivity to irinotecan. In this study, we genotyped SUMO1... More »

Sumoylation of eIF4E activates mRNA translation.

(PMID:20224576)
Xu X, Vatsyayan J, Gao C, Bakkenist CJ, Hu J
Eukaryotic translation initiation factor 4E (eIF4E) is the cap-binding protein that binds the 5' cap structure of cellular messenger RNAs (mRNAs). Despite the obligatory... More »

N4BP1 is a newly identified nucleolar protein that undergoes SUMO-regulated polyubiquitylation and proteasomal turnover at promyelocytic leukemia nuclear bodies.

(PMID:20233849)
Journal of Cell Science [2010, 123(PT 8):1227-34]
A number of proteins can be conjugated with both ubiquitin and the small ubiquitin-related modifier (SUMO), with crosstalk between these two post-translational modifications... More »

SUMO proteins are involved in the stress response during spermatogenesis and are localized to DNA double-strand breaks in germ cells.

(PMID:20385780)
Shrivastava V, Pekar M, Grosser E, Im J, Vigodner M
Reproduction (Cambridge, England) [2010]
SUMO proteins have been implicated in cellular stress response in different tissues, but whether sumoylation has a similar role during spermatogenesis is currently unknown... More »
Patents in CiteXplore

CiteXplore - search results

insect resistance src:pat "THOMAS MEADE"

No highlighting | highlight

First Previous Result set size: 7 Page: 1 of 1 Page size: 25 Next Last

More

A METHOD TO CONTROL INSECTS RESISTANT TO COMMON INSECTICIDES
MEADE THOMAS, HUANG JIM X, ZHU YUANNING, LOSO MICHAEL R, ORR NAILAH, ROGERS RICHARD B, GIFFORD JAMES M, SPARKS THOMAS C

ID:CA2653186
Score: 421

Combinations of Cry1Ab and Cry1Fa as an insect resistance management tool
LANG BRUCE A, BING JAMES W, BACOCK JONATHAN M, MEADE THOMAS, STORER NICHOLAS P, SIEGFRIED BLAIR D, GUEDES PEREIRE ELISEN J

ID:US200706340
Score: 404

Combinations of Cry1Ab and Cry1Fa as an insect resistance management tool
LANG BRUCE A, BING JAMES W, BACOCK JONATHAN M, MEADE THOMAS, STORER NICHOLAS P, SIEGFRIED BLAIR DAVID, GUEDES PEREIRE ELISEN JOS

ID:US2008311096
Score: 404

INSECTICIDAL TOXIN COMPLEX FUSION PROTEINS
HEY Timothy D, CAI QIHUA, MEADE Thomas, MOON Haley, BURTON Stephanie, SHEETS Joel J, MERLO Donald J, WOOSLEY Aaron T

ID:WO2005084355
Score: 344

INSECTICIDAL TOXIN COMPLEX FUSION PROTEINS
HEY Timothy D, CAI QIHUA, MEADE Thomas, MOON Haley, BURTON Stephanie, SHEETS Joel J, MERLO Donald J, WOOSLEY Aaron T

ID:EP1756148
Score: 344

Insecticidal toxin complex fusion proteins
HEY TIMOTHY D, MEADE THOMAS, BURTON STEPHANIE L, MERLO DONALD J, CAI QIHUA, MOON HALEY J, SHEETS JOEL J, WOOSLEY AARON T, HEY TIMOTHY DENVER, BURTON STEPHANIE LOVE, MERLO DONALD JOSEPH, MOON HALEY 1D, SHEETS JOEL JAY, WOOSLEY AARON TODD

ID:US2006168683
Score: 302

A method to control insects resistant to common insecticides
MEADE THOMAS, ZHU YUANNING, SPARKS THOMAS C, HUANG JIM X, LOSO MICHAEL R, ORR NAILAH, GIFFORD JAMES M, ROGERS RICHARD B

ID:US2007261706
Score: 183
Combinations of Cry1Ab and Cry1Fa as an insect resistance management tool
(Patent: US2008311096) Free resource
LANG BRUCE A, BING JAMES W, BABCOCK JONATHAN M, MEADE THOMAS, STORER NICHOLAS P, SIEGFRIED BLAIR DAVID, GUEDES PEREIRA ELISEN JOSE
Compositions for controlling lepidopteran insects use Cry1Fa and Cry1Ab core toxin containing proteins in combination to delay or prevent development of resistance.

A method to control insects resistant to common insecticides
(Patent: AU2007261706) Free resource

A METHOD TO CONTROL INSECTS RESISTANT TO COMMON INSECTICIDES
(Patent: CA2653186) Free resource
N-Substituted sulfonamides are effective at controlling insects resistant to common insecticides.

INSECTICIDAL TOXIN COMPLEX FUSION PROTEINS
(Patent: EP1756148) Free resource
HEY Timothy D, CAI Qihuia, MEADE Thomas, MOON Haley, BURTON Stephanie, SHEETS Joel J, MERLO Donald J, et al.
The subject invention relates to insecticidal toxin complex (Δ2TCΔ2) fusion proteins and to polynucleotides that encode these fusion proteins. The subject invention also... More >

Combinations of Cry1Ab and Cry1Fa as an insect resistance management tool
(Patent: US2007006340) Free resource
LANG BRUCE A, BING JAMES W, BABCOCK JONATHAN M, MEADE THOMAS, STORER NICHOLAS P, SIEGFRIED BLAIR DAVID, GUEDES PEREIRA ELISEN JOSE
How European is UKPMC today?

- Telethon Italy
- Science Foundation Ireland
- Austrian Sciences Research Fund
- Health Research Board, Ireland
August 2008: Through FP7, the Commission is investing, between 2007 and 2013, over € 50 billion in research and expects to get the best possible return on this investment.

The OA pilot covers approximately 20% of the FP7 budget and will apply to specific areas of research:

- Health
- Energy
- Environment
- Information and Communication Technologies
- Research Infrastructures
- Socio-economic Sciences and Humanities
- Science in Society
OpenAIRE

European Commission want a reporting mechanism on their funded research

- 38 organizations representing all EU countries except Luxembourg
- Extension of DRIVER project (D-NET software)
- Develop Plug-in for DRIVER-compliant repositories
- Develop “orphan article” repository
- Help desk function

EBI role: report on data requirements for health research
The mission of ELIXIR is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society.
A Virtual Elixir Node

EU- and ERC-funded research

- Open AIRE plug-in
- EU IR
- EU IR
- EU IR
- UK IR

UK-funded research

- Adapt MSS
- Bulk Deposit
- UKPMC

Repository Junction
PubMed Central Article Request Frequency - calendar 2009
Available: 1.9 Million Articles
Used: 98%, Used > 10 times: 69%
Green Open Access

- Self archiving
- Institutional Repositories
- “free”
- Harder to find
- A bit messy
- Distributed, local
- Institutional mandates

Gold Open Access

- Journal OA options
- PMC, UKPMC
- Upfront cost
- Easier to find
- Formalized
- Centralized
- Funder mandates

~750 institutional repositories in Europe (~500? active)
The Open Access Citation Advantage

Alma Swan, Key Perspectives Ltd

Studies finding a positive Open Access citation advantage: 27
Studies finding no Open Access citation advantage: 4

Size of OA citation advantage (% increase in citations with OA:

- Biology: -5 to 36
- Comp Sci: 157
- Medicine: 300 to 450
- Agricultural Sci: 200 to 600

http://eprints.ecs.soton.ac.uk/18516/
Application-Driven Text Mining

• In the world of big data, text mining is a key integration mechanism between the literature and data

• Core of UKPMC added value; new feature development

• Technology transfer: transitioning from demonstration to application
  • Quality, service, standards
Spectrum of Text Mining Activities

Biology
- Primer Sequence Links
- Biological Organization
- Named Entity Recognition

Development
- Figure Search
- Reference Extraction
- Ontology/Vocabulary Maintenance

Research
- Acronym Disambiguation
- Topic Categorization
- Fact Extraction
- Hedging Recognition
- Automated Language Translation

Linguistics
**Text Mining in UKPMC (September 2010)**

<table>
<thead>
<tr>
<th>Semantic Type</th>
<th>Unique Terms</th>
<th>Articles</th>
<th>Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene/Protein</td>
<td>205,480</td>
<td>1,022,853</td>
<td>10,952,366</td>
</tr>
<tr>
<td>GO Terms</td>
<td>30,117</td>
<td>1,486,201</td>
<td>11,625,726</td>
</tr>
<tr>
<td>Organism</td>
<td>153,410</td>
<td>1,399,168</td>
<td>10,098,689</td>
</tr>
<tr>
<td>Disease</td>
<td>157,414</td>
<td>1,435,551</td>
<td>12,982,609</td>
</tr>
<tr>
<td>Accession No.</td>
<td>232,293</td>
<td>56,512</td>
<td>310,321</td>
</tr>
<tr>
<td>Chemical</td>
<td>71,171</td>
<td>1,370,740</td>
<td>18,328,952</td>
</tr>
</tbody>
</table>
A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis.

PMCID:PMC1976626


Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands.


Type: Journal Article, Research Support, Non-U.S. Gov't, Comparative Study

DOI: 10.1371/journal.pmed.0040278

Abstract

BACKGROUND: Rheumatoid arthritis (RA) is a chronic autoimmune disorder affecting approximately 1% of the population. The disease results from the interplay between an individual's genetic background and unknown environmental triggers. Although human leukocyte antigens (HLAs) account for approximately 30% of the heritable risk, the identities of non-HLA genes explaining the remainder of the genetic component are largely unknown. Based on functional data in mice, we hypothesized that the immune-related genes complement component 5 (Cs) and/or TNF receptor-associated factor 1 (TRAF), located on Chromosome 9q33-34, would represent relevant candidate genes for RA. We therefore aimed to investigate whether this locus would play a role in RA.

METHODS AND FINDINGS: We performed a multiregional case-control study using 40 single-nucleotide polymorphisms (SNPs) from the TRAF1 and C5 (TRAF1/C5) region in a set of 290 RA patients and 254 unaffected participants (controls) of Dutch origin. Stepwise replication of significant SNPs was performed in three independent studies.

Identified 13 unique Genes/Proteins in the Full Text

- TRAF1 (37)
- TNF (8)
- TNF receptor-associated factor 1 (14)

Identified 13 unique Diseases in the Full Text

- Rheumatoid Arthritis (28)
- Inflammatory arthritis (19)
- Arthritis
- Rheumatic diseases (3)

Identified 1 unique Accession Numbers in the Full Text

- Q09472 (1)

10 million articles cited at least once
Integration of citations, bibliographic, TM data
Desirable outcomes: better leads on difficult questions

Who else has published articles on these genes?

Which are the best recent reviews on angiogenesis inhibitors?

Who has expertise in using FluoroNanogold?

Which diseases are associated ADRB3?

Get me all the articles on ticks.

How strong is the evidence that p53 interacts with MDM2?
Credits

Core EBI Staff
Peter Stoehr
Sharmila Pillai
Alan Horne
Norman Cobley
Paula Buttery

Content Providers
PubMed
Agricola
EPO
Chinese Biological Abstracts

Collaborators
Dietrich Rebholz-Schuhmann (EBI text mining)
NacTeM, MIMAS, British Library (UKPMC)
European Patent Office