Open access and UKPMC

D. Rebholz-Schuhmann
“Open Access im Forschungsbereich Gesundheit“, 13-November-2009, DKFZ, Heidelberg
An alternative pathway to beta-carotene formation in plant chondrobiota discovered by map-based cloning of beta and old-gold color mutations in tomato.

Authors: Rosetti, G., Galbiati, G., Zeller, A. G.

Affiliation: Department of Genetics, The Life Sciences Institute, and Faculty of Agriculture, The Hebrew University of Jerusalem, Jerusalem, Israel.

Language: English


Publication Type: Journal Article (Research Support, Non-U.S. Govt.)

Full text article: [link]

Abstract:
Centiferated pigments in plants fulfill indispensable functions in photosynthesis. Carotenoids that accumulate as secondary metabolites in higher plants provide distinct coloration to flowers and fruits. In this work, we investigated the genetic mechanisms that regulate accumulation of carotenoids as secondary metabolites during tomato fruit growth. We identified two mutations that affect betacarotene biosynthesis in tomato. (Beta) (beta) is a single dominant gene that increases betacarotene in the fruit, and (beta) (beta) is a recessive mutation that abolishes betacarotene and increases lycopene. Using a map-based cloning approach, we isolated the genes (beta) and (beta). Molecular analysis revealed that (beta) encodes a new type of leucine zipper protein, an enzyme that converts lycopene to betacarotene. The amino acid sequence of (beta) is similar to other leucine zipper proteins, suggesting that it produces proteins that accumulate in fruits of peppers (Capsicum annuum). Our results show that betacarotene is synthesized in vivo during tomato fruit development by the (beta) gene. (Beta) (beta) is expressed at low levels during the later stages of tomato fruit development, whereas in the (beta) mutant its transcript is dramatically increased. Null mutations in the (beta) gene are responsible for the phenotype of (beta), indicating that (beta) is an allele of (beta). These results confirm that developmentally regulated transactivation is the major mechanism that governs lycopene accumulation in tomato fruits. The (beta) genes are conserved in various plant species, and they may be involved in various regulatory functions during plant development and enhance nutritional value of plant foods.
Standardization of Document Formats: IeXML, SciXML

Standardization of Content:
- Genes
- Chemical Entities
- Medical terms
- MeSH, GO terms

Performance assessment on a very large corpus (FP07, support action)

Bioinformatics user: Analytical pipelines

Research to drive standards
This overview: Open access

- Different kinds of open access: author pays / subscriber pays / hybrid / embargo period
- What infrastructure is required for open access?
- What is the benefit for the authors?
- What is the UKPMC infrastructure?
UK PubMed Central funding organisations
UK PubMed Central mission

To become the information resource of choice for the UK biomedical and health research community by:

- Establishing a comprehensive sustainable repository for UK-funded research outputs
- Improving information retrieval and knowledge discovery through the development of text and data-mining solutions
- Providing access to additional content that integrates seamlessly into the UKPMC website
- Creating comprehensive analysis and reporting tools for researchers and Funders to inform strategy and policy making
The UK PubMed Central funding organisations expect:

- Research outputs arising from research that we fund to be made freely and readily available;

- Electronic copies of any biomedical research papers that have been accepted for publication in a peer-reviewed journal, and are supported in whole or in part by funding from any of the UKPMC Funders, to be made available through PubMed Central (PMC) and UK PubMed Central, as soon as possible and in any event within six months of the journal publisher’s official date of final publication

- Authors and publishers, if an open access fee has been paid, to license research papers such that they may be freely copied and re-used for purposes such as text and data mining, provided that such uses are fully attributed. This is also encouraged where no fee has been paid.

For further information, please refer to Funder’s individual open access policies
Background to UK PubMed Central

- Originally UK PubMed Central was a mirror site of PubMed Central.
- PubMed Central was set up in February 2000 by the National Center for Biotechnology Information, Bethesda, United States.
- More information about PubMed Central can be found here: [http://www.pubmedcentral.nih.gov/about/intro.html](http://www.pubmedcentral.nih.gov/about/intro.html)
- Launched as a mirror site of PubMed Central in January 2007.
- Phase three development work (the current phase) began in July 2008, with British Library as lead partner.
- Closed Beta test to run from July 22nd – September 15th 2009.
- Open Beta to commence January 2009. We commit to collating and acting upon feedback for the continuation of phase three funding (up until 2011).
- Re-launch of UK PubMed Central is January 2010.
Figures and statistics

- From July 2008, a £1.3 million development programme commenced. This is supported by the UK’s Eight principle funders of biomedical and health research.
- There are more than 18 million citations and abstracts in PubMed, extracted from thousands of journals.
- PubMed does NOT link through to full-text articles, just abstracts and citations.
- UK PubMed Central DOES link through to full-text articles – in May 2009 there were over 1.6 million.
- UK PubMed Central is supported by the UK’s eight principle funders of biomedical and health research – Arthritis Research Campaign, Biotechnology and Biological Sciences Research Council, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Department of Health – National Institute for Health Research, Medical Research Council, The Wellcome Trust.
- 99.5% submissions to UK PubMed Central archive are made by journal publishers.
- 0.5% are managed by the authors or Principal Investigators themself.
Development activities

UK PubMed Central will achieve its aims by:

- Developing ingenious ways to search, retrieve and link research papers to relevant biomedical and health research knowledge
- Identifying and providing access to a wide-range of additional, valuable content
- Creating tools to enable users to track research grants: what they are, the research that arose from them, and who they were provided by

Development work undertaken by programme consortium
Programme consortium

Lead contractor: programme management, technical development, engagement and marketing, long term preservation, management of grants database

Initially hosts the service, builds small scale developments, input to future development

Text mining and data linking, developing discovery interface
UK PubMed Central – points of contact

Manuscript submission and feedback

UK PubMed Central Help Desk (for manuscript submission queries)
ukpmc@bl.uk

Promotion, engagement, news, events

UK PubMed Central Engagement Manager (for promotional details, news, events etc.)
ukpmc-engagement@bl.uk
UKPMC

• Over ~1.5 million full text articles
• Based on PubMed Central (PMC) at the NIH: PMCi
• Mandated deposition of articles by UK Funders
• 2006-2008: Basic set-up and grant reporting
• 2008-2011 timeframe, adding value to search and retrieval

Funded by the Wellcome Trust, BBSRC, MRC, British Heart Foundation, Cancer Research, Arthritis Research Campaign, NIHR
The Collaboration

The University of Manchester
  MIMAS: host PMCi and Grant Reporting
  NaCTeM: text mining

The British Library
  Content Addition
  Interface Development

The European Bioinformatics Institute
  Web services
  Metadata and full text indexing
  Text mining
The image depicts a diagram illustrating the architecture of a system involving multiple data sources and services. The diagram includes the following components:

- **NCBI**
  - Data in
  - PMC: data
  - Rendering: value

- **Manchester MIMAS**
  - Data in
  - PMCi: data + rendering

- **British Library**
  - Article
  - Added value

- **UKPMC architecture**
  - Full text + abstracts (citeXplore)

- **EBI**
  - Text mining, data integration
  - Manchester NaCTeM
  - EBI

- **Users**

The diagram shows the flow of data and services, indicating how data from NCBI and other sources are processed and made available to users through UKPMC and EBI services.
UKPMC: 1,495,000

PMC only: 315,000

 Distribution of Articles in PMC and UKPMC (June 2009)

515,000 XML
146,000 OPEN ACCESS
980,000 SCANNED

“free access”

~ 9 % the size of PubMed

ukpmc.ac.uk
CiteXplore: overview

- More than 22 million abstracts
  - PubMed: 19 million; patents: 1.88 million

- Website and web services
  - URL below, SOAP

- Basic search, some advanced search features
  - Lucene

- Added value: citations, database links, text mining
  - Citations: over 9 million PubMed articles cited from our UKPMC and CrossRef dataset

http://www.ebi.ac.uk/citexplore/
Text mining in CiteXplore via Whatizit

ADAMTS-7, a metalloproteinase that belongs to ADAMTS family, is important for the degradation of cartilage extracellular matrix proteins in arthritis. Herein we report that ADAMTS-7 is upregulated during chondrocyte differentiation and endochondral bone formation, and this inhibition depends on its proteolytic activity. The cysteine-rich domain of ADAMTS-7 is required for its interaction with the extracellular matrix, and the C-terminal fourthrombospondin motifs are necessary for its full proteolytic activity and inhibition of chondrocyte differentiation. ADAMTS-7 is an important target of canonical PTHrP signaling, since (i) PTHrP induces ADAMTS-7, (ii) ADAMTS-7 is downregulated in PTHrP null mutant (PTHrP -/-) growth plate chondrocytes, and (iii) blockade of ADAMTS-7 almost abolishes PTHrP.

OBJECTIVE: To identify risk factors that may predispose California sea lions (Zalophus californianus) to development of cutaneous poxvirus nodules during hospitalization in a rehabilitation center. DESIGN: Retrospective case-control study. ANIMALS: 90 California sea lions admitted to a rehabilitation center. PROCEDURE: Hospital records were reviewed. California sea lions admitted to the rehabilitation center between January 1 and December 31, 2002. All California sea lions (n = 18) that developed > or = 1 cutaneous poxvirus nodule during hospitalization were considered cases. Seventy-two California sea lions that did not develop poxvirus lesions during hospitalization were randomly selected (control group). The frequencies of various exposure factors prior to admission, at admission, and during hospitalization for

We previously demonstrated, using fluorescence recovery after photobleaching, that clathrin in clathrin-coated pits at the plasma membrane exchanges with free clathrin in the cytosol, suggesting that clathrin-coated pits are dynamic structures. We now investigated whether clathrin at the trans-Golgi network as well as the clathrin adaptors AP2 and AP1 in clathrin-coated pits at the plasma membrane and trans-Golgi network, respectively, also exchanges back to the cytosol. We found that when the budding of clathrin-coated vesicle is blocked without significantly affecting the structure of clathrin-coated pits, both clathrin and AP2 at the plasma membrane and clathrin and AP1 at the trans-Golgi network exchange rapidly with free proteins in the cytosol. In contrast, when budding of clathrin-coated vesicles was blocked at the plasma membrane or trans-Golgi network by hypertonic sucrose or K(+) depletion, conditions that markedly affect the structure of clathrin-coated pits, clathrin exchange was blocked but AP2 at the plasma membrane and both AP1 and the GGA1 adaptor at the trans-Golgi network continue to rapidly exchange. We conclude that clathrin-coated pits are dynamic structures with rapid exchange of both clathrin and adaptors and that adaptors are able to exchange independently of clathrin when clathrin exchange is blocked.
Added value: Anatomy of a CiteXplore record

Pubmed Id: 15192113
Title: ADAMTS7B, the full-length product of the ADAMTS7 gene, is a chondroitin sulfate proteoglycan containing a mucin domain.
Authors: Somerville RP, Longpré JM, Apel ED, Lewis RM, Wang LW, Sanes JR, Leduc R, Arte SS
Affiliation: Department of Biomedical Engineering, Lerner Research Institute, Orthopedic Research Center, Cleveland Clinic Foundation, Ohio 44195, USA.
Language: English

Journal: J. Biol. Chem. (ISSN: 0021-9298) (EISSN: 1083-329X)
Publication type: Journal Article, Research Support, Non-U.S. Govt, Research Support, U.S. Govt, P.H.S;
Full text: XML

Proteins
- **Seq id**: COG1104
- **Name/Info**: ADAMTS-7 precursor (EC 3.4.24.-) (A disintegrin and metalloproteinase with thrombospondin motifs 7)
- **Taxonomy**: Homo sapiens
- **Uniprot id**: O68210
- **Name/Info**: ADAMTS7B
- **Taxonomy**: Mus musculus
- **Uniprot id**: Q8TVN4
- **Name/Info**: source: EMBL manual annotation
- **Uniprot id**: Q6VY94
- **Name/Info**: source: EMBL manual annotation

Nucleotide sequences
- **EMBL id**: AY511000
- **Description**: Mus musculus ADAMTS7B (Adams7) mRNA, complete cds.
- **Seq length**: 5046

Abstract
We have characterized ADAMTS7B, the authentic full-length protein product of the ADAMTS7 gene. ADAMTS7B has a domain organization similar to that of ADAMTS12, with a total of eight thrombospondin type I repeats in its ancillary domain. Of these, seven are arranged in two distinct clusters that are separated by a mucin domain. Unlike the ADAMTS family, ADAMTS7B is modified by attachment of the glycosaminoglycan chondroitin sulfate within the mucin domain, thus rendering it a proteoglycan. Glycosaminoglycan addition has potentially important implications for ADAMTS7B cellular localization and for substrate recognition. Although not an integral membrane protein, ADAMTS7B is retained near the cell surface of HEK293F cells via interactions involving both the ancillary domain and the protease. ADAMTS7B undergoes removal of the prodomain by a multistep furin-dependent mechanism. At least part of the final processing event, i.e., cleavage following Arg1203 (mouse sequence annotation), occurs at the cell surface. ADAMTS7B is an active metalloproteinase as shown by its ability to cleave alpha(v) beta(3)-integrin, but it does not cleave specific peptide bonds in vimentin and aggrecan attacked by ADAMTS proteases. Together with ADAMTS12, whose primary structure also predicts a mucin domain, ADAMTS7B constitutes a unique subgroup of the ADAMTS family.

Grants

Cited by

Keywords (Mesh)

Chemicals
ADAMTS7B, the full-length product of the ADAMTS7 gene, is a chondroitin sulfate proteoglycan containing a mucin domain.

(Somerville RP, Longpré JM, Apel ED, Lewis RM, Wang LW, Sanes JR, Leduc R, Apte SS)

Department of Biomedical Engineering, Lerner Research Institute, Orthopedic Research Center, Cleveland Clinic Foundation, Ohio 44195, USA.

The Journal of Biological Chemistry [2004 Aug, 279(34):35159-75]

Type: Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, P.H.S.

DOI: 10.1074/jbc.M402380200

Abstract

We have characterized ADAMTS7B, the authentic full-length protein product of the ADAMTS7 gene. ADAMTS7B has a domain organization similar to that of ADAMTS12, with a total of eight thrombospondin type 1 repeats in its ancillary domain. Of these, seven are arranged in two distinct clusters that are separated by a mucin domain. Unique to the ADAMTS family, ADAMTS7B is modified by attachment of the glycosaminoglycan chondroitin sulfate within the mucin domain, thus rendering it a proteoglycan. Glycosaminoglycan addition has potentially important implications for ADAMTS7B cellular localization and for substrate recognition. Although not an integral membrane protein, ADAMTS7B is retained near the cell surface of HEK293F cells via interactions involving both the ancillary domain and the prodomain. ADAMTS7B undergoes removal of the prodomain by a multistep furin-dependent mechanism. At least part of the final processing event, i.e. cleavage following Arg(220) (mouse sequence annotation), occurs at the cell surface. ADAMTS7B is an active metalloproteinase as shown by its ability to cleave alpha(2)-macroglobulin, but it does not cleave specific peptide bonds in versican and aggrecan attacked by ADAMTS proteases. Together with ADAMTS12, whose primary structure also predicts a mucin domain, ADAMTS7B constitutes a unique subgroup of the ADAMTS family.
Addition of content to CiteXplore

- PubMed: 19 million
- PubMed Central: 1.8 million

~ 0.5 million

PMC will be a true subset at CiteXplore and UKPMC
Text mining component of UKPMC?

Named Entity Recognition
Organisms, GO Terms, Genes/Proteins,
Acc. Numbers, Diseases, Drugs, Chemicals

Fact Extraction (to 2011)
Protein-protein interactions, gene-phenotype relationships, drugs-proteins

→ Applications

Different uses for different applications
e.g. “human” vs. “ABCA1” vs. “inhibits”
Progress so far

1,551,533 documents: OCR, PDF, XML

Genes/Proteins: 5,110,489, Species: 3,892,466, GO terms: 4,493,691

<table>
<thead>
<tr>
<th>Top 10 Genes/Proteins</th>
<th>Top 10 Organisms</th>
<th>Top 10 GO Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4</td>
<td>Human</td>
<td>Binding</td>
</tr>
<tr>
<td>IFN</td>
<td>Mouse, mice</td>
<td>Membrane(s)</td>
</tr>
<tr>
<td>Insulin</td>
<td>Escherichia coli</td>
<td>Development</td>
</tr>
<tr>
<td>Actin</td>
<td>Animals</td>
<td>Transcription</td>
</tr>
<tr>
<td>P53</td>
<td>Rat(s)</td>
<td>Death</td>
</tr>
<tr>
<td>TNF</td>
<td>Bacteria</td>
<td>Host</td>
</tr>
<tr>
<td>GST</td>
<td>Yeast</td>
<td>Chromosome</td>
</tr>
<tr>
<td>GFP</td>
<td>Rabbit</td>
<td>Phosphorylation</td>
</tr>
<tr>
<td>LacZ</td>
<td>HIV</td>
<td>Intracellular</td>
</tr>
<tr>
<td>EGF</td>
<td>viruses</td>
<td>Transport</td>
</tr>
</tbody>
</table>

Many improvements that could be made
Cytochrome b, reproduction and mammals are frequently mentioned in this article.

<table>
<thead>
<tr>
<th>Genes/Proteins</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cytochrome b (3)</td>
<td></td>
</tr>
<tr>
<td>Nef (2)</td>
<td></td>
</tr>
</tbody>
</table>

GO Terms

| reproduction (2) |   |
| W chromosome (1) |   |

Organsims

| mammals (27) |   |
| rodents (7)  |   |
| 30 more      |   |

PTCH1, chromosome and humans are frequently mentioned in this article.

<table>
<thead>
<tr>
<th>Genes/Proteins</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PTCH1 (46)</td>
<td></td>
</tr>
<tr>
<td>SSH (5)</td>
<td></td>
</tr>
<tr>
<td>11 more</td>
<td></td>
</tr>
</tbody>
</table>

GO Terms

| chromosome (10) |   |
| development (8) |   |
| 20 more         |   |

Organsims

| human (10)      |   |
| mice (5)        |   |
| 13 more         |   |
Challenges

Technical

• Precision and recall
• OCR (see example)
• Experience in mining full text content; section identification
• Integrating text mining utility into regular workflows
• Provision of functions in a production environment

Political/Social

• Business Rules (see next slide), open and free access
• Content growth (compliance; new funding agencies)
• Biologists are unforgiving of TM “errors”
Application Goals of Text mining in UKPMC

High quality, integrated functions

Integrated browse functions
  Highlighting, article summaries
  e.g. articles with similar named entity profiles to this one

Integration with underlying databases
  e.g. UniProt, Array Express, PDB
  Highlighting, article summaries

Integrated search functions
  e.g. search for gene symbol & find co-occurring diseases
  facts are the ultimate in co-occurrence

Many of these kinds of functions are demonstrated in existing stand-alone apps
Who did the work?

EBI Literature Services
Peter Stoehr, Sharmila Pilia, Alan Horne, Mark Rijnbeek

EBI Text mining
Dietrich Rebholz-Schuhmann, Ian Lewin, Jee-Hyub Kim

Collaborators: UKPMC
NaCTeM: Sophia Ananiadou, CJ Rupp, Chikashi Nobata
MIMAS: Dave Chapman, Vic Lyte, Ross McIntyre
British Library: Ernie Ong, Phil Vaughan, Sandy Chevuru, Rob Rowbotham, Paul Davey, Heather Rosie