G2Cdb: the Genes to Cognition database

Citation for published version:

Digital Object Identifier (DOI):
10.1093/nar/gkn700

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Nucleic Acids Research

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
G2Cdb: the Genes to Cognition database

Mike D. R. Croning¹, Michael C. Marshall¹, Peter McLaren¹, J. Douglas Armstrong² and Seth G. N. Grant¹,*

¹Genes to Cognition Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA and ²Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK

Received August 15, 2008; Revised September 25, 2008; Accepted September 26, 2008

ABSTRACT

Neuroscience databases linking genes, proteins, (patho)physiology, anatomy and behaviour across species will be valuable in a broad range of studies of the nervous system. G2Cdb is such a neuroscience database aiming to present a global view of the role of synapse proteins in physiology and disease. G2Cdb warehouses sets of genes and proteins experimentally elucidated by proteomic mass spectroscopy of signalling complexes and proteins biochemically isolated from mammalian synapse preparations, giving an experimentally validated definition of the constituents of the mammalian synapse. Using automated text-mining and expert (human) curation we have systematically extracted information from published neurobiological studies in the fields of synaptic signalling electrophysiology and behaviour in knockout and other transgenic mice. We have also surveyed the human genetics literature for associations to disease caused by mutations in synaptic genes. The synapse proteome datasets that G2Cdb provides offer a basis for future work in synapse biology and provide useful information on brain diseases. They have been integrated in a such way that investigators can rapidly query whether a gene or protein is found in brain-signalling complex(es), has a phenotype in rodent models or whether mutations are associated with a human disease. G2Cdb can be freely accessed at http://www.genes2cognition.org.

INTRODUCTION

Synapses are the fundamental unit of computation in the brain playing key roles in information processing, behaviour and disease. They not only transmit information between cells but also detect patterns of neural activity and process this information by activating intracellular biochemical signalling pathways, which subsequently changes the properties of the neuron. G2Cdb presents an integrated view of the role of synapses, focusing on large, high-quality datasets describing synaptic proteins and diseases of the nervous system, particularly those affecting cognition.

Since the year 2000, proteomic studies have increased the number of known synaptic proteins over 10-fold and provided lists of proteins that represent the draft synapse proteome (1,2 and other references therein). The synapse has different compartments, such as the post-synaptic proteome (PSP), comprising ~1100 proteins, and pre-synaptic vesicles with ~80 proteins (3). The high degree of complexity was unexpected and understanding the function of individual proteins and the overall organization of the molecular networks presents a major challenge.

G2Cdb aims to be the central database for warehousing data on the synaptic proteome. Other useful databases exist for molecular neuroscience of which the closest is Synapse DataBase (SynDB). G2Cdb differs fundamentally from SynDB both in terms of the data content and in the way it is constructed. SynDB employs keyword and ontology-term searching of protein sequence and motif databases to provide an informatic definition of the synapse (4). In contrast, G2Cdb uses data curated from published studies of synaptic protein profiling to provide an experimentally validated representation of the mammalian synapse. Both approaches, and thus databases, are highly complementary.

Building upon this proteomic definition of the synapse G2Cdb integrates mouse and human genomic annotation resources, forming the basis of a ‘molecular catalogue’ of mammalian synaptic genes. Information mined from the human genetics literature reporting associations between synaptic gene mutations and disease is included, as is our in-depth and on-going survey of the neurobiological phenotypes observed in published studies of knockout and other transgenic mice.

With the aim of presenting a global view of the role of synapses in physiology and disease, these datasets have been integrated in a gene-centric manner. The resulting database, G2Cdb, should be of interest to all

*To whom correspondence should be addressed. Tel: +44 1223 834244; Fax: +44 1223 494919; Email: sg3@sanger.ac.uk

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
neuroscientists, clinicians and geneticists with interests in disease, given the ever-increasing number of synaptic proteins that are involved in human brain diseases such as Alzheimer’s disease, autism, mental retardation and schizophrenia (5).

G2Cdb can be freely accessed at www.genes2cognition.org.

CONSTRUCTION AND CONTENTS

G2Cdb consists of a catalogue of experimentally validated mammalian synapse genes that we have integrated with genomic annotation resources, studies of naturally occurring gene mutations found in the human central nervous system (CNS) diseases, brain gene expression resources, as well as behavioural studies and electrophysiological studies using genetically modified mice. The database creation was split into six principle steps (i) compiling a catalogue of mammalian synapse genes from proteomic studies, (ii) attaching gene symbols, protein names and their synonyms to the genes, (iii) constructing gene lists to represent the constituents of synaptic protein complexes and organelles, (iv) linking to external genome and molecular resources, (v) linking to transcript and protein expression resources, and (vi) automated text-mining and then human (manual) curation of published literature. Details for each of these are presented in the following sections.

Compiling a catalogue of mammalian synapse genes

The proteomic study of synaptic organelles and receptor–protein complexes biochemically isolated from nervous tissue gives us the best current picture of the components of the mammalian synapse. These studies utilize a series of purification strategies, followed by liquid chromatography tandem mass spectroscopy and/or large scale immuno-blotting to identify proteins (6).

We have extracted the data from seven large-scale proteomic profiling studies to populate G2Cdb (Table 1), thus creating a comprehensive set of proteins found at the mammalian synapse. In total, more than 1100 proteins have been identified revealing an unexpected complexity to the synapse which is also reflected in its complex evolutionary origin (7). We refer to this set as the PSP (2) and this is the current focus of G2Cdb.

As the next step in creating a mammalian synapse gene catalogue, we took the set of 1124 proteins found in the PSP and mapped the protein sequences back to the mouse genome via the Ensembl genome annotation database. Next, we used all-versus-all protein and transcript FASTA (8) searches to ensure the resulting mouse gene set was non-redundant and then assigned unique identifiers to these genes (of the form ‘Gxxxxxxxx’ where x is a digit 0–9). We found unique identifiers were necessary as genome database identifiers changing regularly in mammalian genomes, particularly in difficult-to-predict, multi-exon genes. Having our own identifiers insulates G2Cdb from such changes.

We also store and display homologous gene information for all G2Cdb genes. To date, we have focused on orthology information, utilizing the automated Ensembl predictions as a starting point. We found it valuable to manually curate gene orthology information particularly where clear 1:1 relationships between the human and mouse genes could not be automatically obtained. Such orthologues were manually assigned by searching using the mouse and human gene names on the Mouse Genome Informatics (MGI; 9) and HUGO Gene Nomenclature Committee (HGNC; 10) database websites, respectively. Where this did not provide a match, sequence similarity was used to assign orthologues.

Gene symbols, protein names and their synonyms

Different research communities (e.g. neuroscience or genetics) for historical reasons have often favoured the use of different names or symbols when referring to the same biological entities. We address this by labelling genes with approved gene symbols and additionally stored terms and synonyms obtained from various sources including the gene nomenclature committees (MGI and HGNC) and standardized protein names (UniProt; 11). This means the user can select from either standard names or use common synonyms employed in their field when searching.

**Table 1.** Published synapse proteomic profiling datasets used in the production of G2Cdb

<table>
<thead>
<tr>
<th>Paper</th>
<th>Year</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2004</td>
<td>Jordan et al. (30)</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>Li et al. (31)</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>Peng et al. (32)</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>Yoshimura et al. (33)</td>
</tr>
<tr>
<td></td>
<td>2002</td>
<td>Satoh et al. (34)</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>Walikonis et al. (35)</td>
</tr>
</tbody>
</table>
**Representation of synaptic protein complexes and organelles**

The protein products of nervous system genes are far from evenly distributed at nerve terminals or synapses. Rather, they are integrated into large macromolecular protein complexes by scaffolding proteins, producing units that have distinct functional roles which create the ability of molecules to couple via adaptors activating various cell-signalling pathways (12). To capture this structural and organizational information in G2Cdb, genes whose protein products are associated may be grouped into one or more gene lists. Prime examples include the two glutamate receptor complexes, the NMDA receptor complex (NRC/MASC) and the AMPA receptor complex (ARC), both of which are key receptor complexes in synaptic transmission and synaptic plasticity at the post-synaptic membrane (1–2). A further example is the clathrin-coated vesicle containing proteins mediating receptor endocytosis (13).

Towards integrating information on the function, interactions and phylogeny of individual proteins within the NRC/MASC complex, we previously probed its organization and developed a model of its function using a combination of annotation, network and statistical approaches (14). We proposed a modular network with 13 distinct subcomponents with distributed functionality that could explain many of the features of synaptic signalling. To facilitate use of our prior analysis and to allow for expansion, we added the gene lists in G2Cdb for the 13 subcomponents of the NRC/MASC complex.

**External genome and molecular resources**

Links were established for each G2Cdb gene to publicly available genome and molecular resources. These include Ensembl (15), Vega (16), UniProt (11), Entrez Gene (17), GeneCards (18) and OMIM (19) and PubMed (20; see Table 2 for the full list). We used mapping information from Ensembl and/or the approved gene symbols from Mouse gene nomenclature MGI http://www.informatics.jax.org Human gene nomenclature HGNC http://www.genenames.org Protein knowledgebase UniProt http://www.uniprot.org Mouse and human genomic annotation NCBI Entrez Gene http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene Mouse and human genomic annotation Ensembl http://www.ensembl.org Human genes and genetic phenotypes OMIM http://www.ncbi.nlm.nih.gov/omim Human genes and genetic phenotypes Vega http://vega.sanger.ac.uk Human gene information GeneCards http://www.genecards.org Human gene information Allen Brain Atlas http://www.brain-map.org Mouse brain gene expression Allen Brain Institute, we provide the Allen Brain Atlas (21) data links to their database of gene expression patterns in the mouse brain. Other mouse gene expression resources were mined by automated means (BGEM, GenePaint, GENSAT, EMAGE; 22–25). Links to human protein expression profiles are provided for the Human Protein Atlas (26). Using automated cross-referencing these gene and protein expression links are regularly updated, maintaining synchronization with these external resources.

<table>
<thead>
<tr>
<th>Content</th>
<th>Database</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse and human genomic annotation</td>
<td>Ensembl</td>
<td><a href="http://www.ensembl.org">http://www.ensembl.org</a></td>
</tr>
<tr>
<td>Human-curated gene structures</td>
<td>Vega</td>
<td><a href="http://vega.sanger.ac.uk">http://vega.sanger.ac.uk</a></td>
</tr>
<tr>
<td>Human gene information</td>
<td>GeneCards</td>
<td><a href="http://www.genecards.org">http://www.genecards.org</a></td>
</tr>
<tr>
<td>Mouse nervous system gene expression</td>
<td>BGEM</td>
<td><a href="http://www.stjudebgem.org">http://www.stjudebgem.org</a></td>
</tr>
<tr>
<td>Mouse gene expression in development</td>
<td>EMAGE</td>
<td><a href="http://genex.hgu.mrc.ac.uk/Emage/database">http://genex.hgu.mrc.ac.uk/Emage/database</a></td>
</tr>
<tr>
<td>Mouse embryo gene expression</td>
<td>GenePaint</td>
<td><a href="http://www.genepaint.org">http://www.genepaint.org</a></td>
</tr>
<tr>
<td>Human protein distribution</td>
<td>Human Protein Atlas</td>
<td><a href="http://www.proteinatlas.org">http://www.proteinatlas.org</a></td>
</tr>
<tr>
<td>Protein knowledgebase</td>
<td>UniProt</td>
<td><a href="http://www.uniprot.org">http://www.uniprot.org</a></td>
</tr>
<tr>
<td>Human gene nomenclature</td>
<td>HGNC</td>
<td><a href="http://www.genenames.org">http://www.genenames.org</a></td>
</tr>
<tr>
<td>Mouse gene nomenclature</td>
<td>MGI</td>
<td><a href="http://www.informatics.jax.org">http://www.informatics.jax.org</a></td>
</tr>
<tr>
<td>Published scientific literature</td>
<td>PubMed</td>
<td><a href="http://www.ncbi.nlm.nih.gov/pubmed">http://www.ncbi.nlm.nih.gov/pubmed</a></td>
</tr>
</tbody>
</table>
(temperature, stimulating protocol, etc.). For the second list, the behavioural assessment of animals with a gene mutation was manually curated and classed, based on a controlled vocabulary and an anatomically based grouping of the behavioural tests. Presently, we have amassed information on a total of ∼340 genes, many of which show phenotypes, by extracting information from about 600 mouse synaptic plasticity and behavioural studies.

Using similar techniques we have also surveyed the human genetics literature looking for associations to disease caused by mutations in synapse genes in our catalogue. To date, we have curated information on 324 diseases, including 90 nervous system diseases, associated with mutations in about 90 synapse genes, from almost 800 articles. Mutations are classified as one of about 35 types, such as single nucleotide polymorphism (SNP), insertion, deletion, nonsense and copy number variation (CNV).

The Genes to Cognition research programme (29) is currently performing a large-scale study of mice carrying mutations in post-synaptic proteins, with extensive phenotyping in biochemical, electrophysiological and behavioural domains. This data is deposited as it becomes available in G2Cdb.

WEB INTERFACE

Public access to G2Cdb is provided through a web interface where users can interactively query and retrieve data. We have integrated the above datasets from disparate knowledge domains so that interrogation by a single text search is possible. Users can rapidly determine if a particular gene is found in brain-signalling complex(es), has been altered and studied in experimental paradigms of learning and memory, and whether a mutation in this gene can be associated with a human disease. As an example, we can take ‘SAP102’. Searching for this gene returns links to the G2Cdb ‘GeneView’ pages for both the mouse and human genes with their species-specific links to external resources. Also (at the time of writing) curated results from two human genetics studies that have linked mutations in the gene encoding SAP102 (DLG3) to X-linked mental retardation are retrieved, as is data extracted from a published knockout mouse study reporting electrophysiological characterization (synaptic transmission, short- and long-term plasticity) and results from a battery of behavioural tests (Figure 1).

The website provides an interactive comparison tool called ‘CompareGeneLists’ to fully exploit the growing number of gene lists loaded into G2Cdb. This tool allows users to select two gene lists and discover the

Figure 1. Sample screenshots of G2Cdb. Searching for SAP102 (A) returns links to the G2Cdb ‘GeneView’ pages for both the mouse (B) and human genes with their species-specific links to external resources. Also returned are curated results from two human genetics studies (C) that have linked mutations in the gene encoding SAP102 (DLG3) to X-linked mental retardation (D), as is data extracted from a published knockout mouse study. The gene list comparison tool is also shown, displaying the results of comparing the mouse PSD and PSP datasets (E).
common genes and those unique in each set. One can thus
pose the question, ‘Which are the common and differing
genes found to encode the proteins of the NRC/MASC
and post-synaptic density?’

Complementing the text-searching abilities of G2Cdb,
one can browse the contents of the database by clicking
the various links available under ‘Database’ (on the menu
present on the left of all the G2Cdb pages) and in the ‘Top
Searches’ links. Examples are ‘Gene Lists’, ‘Human
Disease’ and ‘Genes shown to affect long-term potentia-
tion (LTP)’.

SOFTWARE IMPLEMENTATION

G2Cdb is implemented in a relational database manage-
ment system (MySQL) in a schema with about 70 tables in
which all the biological information and related controlled
vocabularies (CVs) are stored. The database back-end is
accessed by an object-orientated Perl API, together with
an extensive set of Perl scripts for loading the database.
These automated loading scripts integrate our curated lit-
erature datasets and those obtained from external
resources and perform quality control (QC) checks. The
G2Cdb website is driven by a set of complementary CGI
scripts.

The principal objects stored by G2Cdb are genes,
alleles, genetically modified mouse lines, phenotyping
experiments and diseases. The API allows one to create,
link, store and retrieve these objects in a high-level manner
(independent of the underlying SQL) making it straight-
forward to implement quite sophisticated analyses.
Generic mechanisms exist to attach synonyms, textual
notes, literature citations, external databases’ cross-
references and binary files (such as media files or pdfs)
to the core database objects, and to validate object attrib-
utes with CVs. The website also supports secure login (by
single sign-on) should this be required.

The software package comprising G2Cdb will be useful
to other investigators working in research fields outside of
cognition who have the need to integrate similar types of
data from genomics and experimental genetics (genetically
altered mice and phenotyping experiments) with disease
and mutation information from human genetics studies.
In common with other Wellcome Trust-funded infor-
matics initiatives, the software is open source, and is avail-
able free of charge under the terms of the Perl Artistic
License (by request to webmaster@genes2cognition.org).
The majority of programmes and modules are docu-
mented with embedded Perl documentation (POD) and the
schema and API are designed very much in keeping with
those of Ensembl, so bioinformaticians familiar with this
project should find it straightforward to employ the
system.

DATA UPDATES

We currently update G2Cdb 2–4 times a year. At these
times we release newly incorporated data, make improve-
ments to the website to support retrieval of the new data
types and generally improve site usability. We plan to load
new synapse proteomic profiling datasets as they become
available and continue our text mining and literature cura-
tion, focussing on extending the set of genes for which we
have curated synaptic plasticity and behavioural studies in
genetically modified mouse models. Similarly, we will con-
tinue to survey the human genetics literature for mutations
in the human orthologues of these genes and their associa-
tion to disease. The G2C research programme and other
large-scale research programmes will provide major
volumes of data to G2Cdb. The longer term objective is
to provide comprehensive information on all synapse pro-
teins across a wide range of phenotypes, functions and
species.

DISCUSSION

Driven by the experimental progress in synapse proteo-
mics and functional studies of these proteins, we identified
a need for a specialist database for the organization and
function of the synapse proteome. This includes capturing
detailed information from a number of sources of experimen-
tally validated results, expert curated information
from the literature and links to related external databases.
Importantly, G2Cdb provides a mechanism for the effective
re-use and analysis of a range of datasets that are
expensive to curate/produce by the wider community.

The synapse proteome datasets that G2Cdb provides
thus offer a basis for future research in synapse biology
and provide useful information on brain diseases.

A major application of G2Cdb will be the assembly of
molecular networks of the synapse proteome. These
include transcriptional, protein interaction and phosphor-
ylation networks. Combining this with phenotypic data on
specific proteins in electrophysiology and behaviour from
mice, and human disease information will be useful for the
systems biology of the synapse.

Toward the public understanding of science and the
education of school and undergraduates, we are collaborat-
ing with the Dolan DNA Learning Centre at Cold
Spring Harbor Laboratory to provide G2Cdb in a
format of use to school and college students. An educa-
tional website (www.g2conline.org) will cover a broad
range of educational material on the subjects of genes
and behaviour including novel network representations
of the G2Cdb datasets.

ACKNOWLEDGEMENTS

The authors would like to thank Phil Elliot and Keri-Lee
Page for their contributions of data to G2Cdb and Lianne
Stanford for critical review of the article.

FUNDING

Genes to Cognition Programme (Wellcome Trust).
Funding for open access charges: Wellcome Trust.

Conflict of interest statement. None declared.
REFERENCES


