Published online 12 September 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D251–D260***
doi:10.1093/nar/gkn568

OKCAM: an ontology-based, human-centered knowledgebase for cell adhesion molecules

Chuan-Yun Li[1,2], Qing-Rong Liu[1], Ping-Wu Zhang[1], Xiao-Mo Li[2], Liping Wei[2] and George R. Uhl[1,*]

[1]Molecular Neurobiology Branch, NIH-IRP (NIDA), Baltimore, MD 21224, USA and [2]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, China

Received June 24, 2008; Revised August 18, 2008; Accepted August 21, 2008

ABSTRACT

'Cell adhesion molecules' (CAMs) are essential elements of cell/cell communication that are important for proper development and plasticity of a variety of organs and tissues. In the brain, appropriate assembly and tuning of neuronal connections is likely to require appropriate function of many cell adhesion processes. Genetic studies have linked and/or associated CAM variants with psychiatric, neurologic, neoplastic, immunologic and developmental phenotypes. However, despite increasing recognition of their functional and pathological significance, no systematic study has enumerated CAMs or documented their global features. We now report compilation of 496 human CAM genes in six gene families based on manual curation of protein domain structures, Gene Ontology annotations, and 1487 NCBI Entrez annotations. We map these genes onto a cell adhesion molecule ontology that contains 850 terms, up to seven levels of depth and provides a hierarchical description of these molecules and their functions. We develop OKCAM, a CAM knowledgebase that provides ready access to these data and ontologic system at http://okcam.cbi.pku.edu.cn. We identify global CAM properties that include: (i) functional enrichment, (ii) over-represented regulation modes and expression patterns and (iii) relationships to human Mendelian and complex diseases, and discuss the strengths and limitations of these data.

INTRODUCTION

'Cell adhesion molecules' play central roles in much of the connection and communication between cells and their synapses (1). Cell adhesion-related communication is essential for many aspects of the proper development of a variety of organs and tissues (1). This cellular communication also plays substantial roles in the plasticity of cell recognition processes in the developed organism (2).

Cell adhesion molecules (CAM) may be especially important in the brain. The brain requires proper connections of many trillions of synapses to develop properly as well as substantial plasticity in many of these synapses to facilitate learning and memory. The dynamics of neuronal synaptic recognition, connection and disconnection appear to make substantial contributions to disorders that display mnemonic features, including addictions and autism (3,4). Current physiologic and cell biologic studies have implicated CAMs as good candidates to play important roles in synapse adhesion (1,5), neuronal connectivity and communication (1), signal transduction (5–8) and proper arrangement of pre-synaptic active zones and post-synaptic densities at classical synapses (9,10).

Current genetic studies have linked and/or associated variants in cell adhesion molecule genes with psychiatric, neurologic, neoplastic, immunologic and developmental phenotypes. The importance of CAMs in learning and memory-associated disorders is demonstrated in recent genome wide association studies (11). Vulnerabilities to addictions are associated with variants in CAM genes in studies of several independent samples (12–14). Genetic variants of the CAM genes NRXN1 and CNTNAP2 have been associated with autism (4,15). Variants in neuregulin have been associated with vulnerability to schizophrenia (16). Variants in an adhesion-like protein KIAA0319 have been associated with dyslexia (17,18).

These data underscore the importance of cell adhesion molecules in both Mendelian and complex disorders of brain and other organs and suggest that a more comprehensive view of these genes and molecules would be valuable. However, there is currently no systematic study that enumerates: (i) the number of genes and gene families that function as CAMs; (ii) common and/or global CAM functions, including those that might extend beyond their cell/cell recognition functions; (iii) common CAM genetic

*To whom correspondence should be addressed. Tel: +1 410 550 2843 x146; Fax: +1 410 550 1535; Email: guhl@intra.nida.nih.gov

variants that might provide individual differences in CAM structures and functions; (iv) over-represented regulation modes and expression patterns and (v) CAM associations with diseases, especially with brain disorders.

We now report compilation of a list of 496 human CAM genes and construction of corresponding cell adhesion molecule ontology (CAMO) to systematically address these questions. Detailed annotations on CAM genes are provided. Global properties of CAM genes, overrepresented types of variation, overrepresented regulation modes and expression patterns, and disease associations are identified. We report a knowledgebase for cell adhesion molecules (OKCAM) that provides ready access to these data and the associated ontologic system that we describe here.

IDENTIFICATION OF HUMAN CAM GENES AND RODENT HOMOLOGS

CAMs were identified based on compilation of data from manual curation of protein domain structures, Gene Ontology annotations, and 1487 annotation entries from keyword queries based on NCBI Entrez Gene annotations (Figure 1). First, we identified features of common protein domains for CAM families based on common motifs from cadherin, immunoglobulin/FibronectinIII (IgFn), integrin, neurexin, neuroligin and catenin families. Using these features, we developed Perl scripts to retrieve and standardize related InterPro domain architectures and the proteins that contain such architectures (19). After manual curation, 44 types of protein domains with 202 detailed domain architectures were identified. These included 532 human proteins that map onto 218 human genes. We used similar protocols to identify cell adhesion gene lists for rat and mouse; these genes were then further mapped to the human genome using Homologene (20). We next extracted CAMs using the Gene Ontology term 'cell adhesion' (GO:0007155) (21). We focus on curated entries; entries that are identified only by annotations that display Evidence Code IEA (Inferred from Electronic Annotation) are noted in Supplementary Table 7. Two hundred eighteen human proteins were identified, which mapped onto 196 human genes. Finally, we manually curated 1487 annotation entries selected from results of the Entrez Gene query 'adhesion AND Homo sapiens [organism]' (20). This approach added 136 more human genes to the list of cell adhesion molecules. In total, we thus identified 496 unique human CAM genes and their homologs in other species.

Meta-data about the domain architectures for CAMs in nonhuman species provided information about CAM evolutionary histories. Of the 113 types of protein domains assessed in our dataset, 705 detailed domain architectures were noted. Among these, only 44 domains with 202 domain architectures were identified in all of the three species, human, rat and mouse. For example, in the cadherin superfamily, there is only one human gene encodes a protein with enzymatic activity, though several dozen cadherins with enzymatic activities are found in bacteria and yeast. Several categories with large numbers of domain architectures that can be detected in lower species including *Caenorhabditis elegans*, *Drosophila melanogaster* and *Danio rerio*, are totally absent from human, rat and/or mouse. These categories include 'IgCAM-like cadherins' that display 29 such domain architectures, 'cadherins with Leucine-rich structures' that display two such domain architectures, 'toxin-related cadherins' that display such 36 domain architectures and 'cadherins with surface anchor structures' that display seven such domain architectures. In striking contrast, 119 of the 123 'cadherin' genes that can be identified in humans fall into the category of 'simple cadherins', that includes genes with only simple combinations of cadherin prodomains, cadherin domains and cadherin cytoplasmic domains. Although 79%, not all, of the proteins that we identify in this study display characteristic InterPro domains, the domain architecture patterns we identify do imply the specification of the CAMs in mammals.

DATA ANNOTATIONS

To elucidate the functions of CAMs, detailed annotations were given to each CAM gene. These data allow interpretation of features of each CAM at five levels: gene family and basic information, genetics, regulation, expression, and Mendelian or complex disease linkage/association.

Information about gene family and basic characterization comes from NCBI Entrez gene annotations (20), Gene Ontology (21), InterPro domains (19), protein interaction databases (22–24), knowledgebases for molecular pathways including KEGG (25), BioCarta and Pathway Interaction Database (PID) and the NCBI PubMed database (20). Genetic variations in these genes, including chromosome recombination hotspots (26), SNPs (20), insertion/deletions (27), chromosomal translocations (27) and CNVs (27), were retrieved from the UCSC Genome Browser Database (26), HapMap (28), NCBI dbSNP database (20) and Database of Genomic Variants (27), respectively. Information about potential or actual modes of regulation was annotated based on the presence of experimentally validated transcription factor binding sites (TFBS) (29), experimentally validated (30,31) and putative miRNA targets (32), noncoding RNA loci (33), *cis*/*trans*-natural antisense transcripts (NATs) (34,35), alternative splicing and post-translational modifications (36) from databases that included TransFac (29), Argonaute (31), TarBase (30), PicTar (32), NatsDB (34,35), NONCODE (33) and dbPTM (36). Information about mRNA expression levels came from: (i) integrated human expressed sequence tag profiles based on developmental stages and tissue distributions, as deposited in Unigene (20) and (ii) mouse brain region expression profiles described in the Allen Brain Atlas (37), with mapping of these data to human orthologs using Homologene (20). We integrated gene expression information at peptide/protein levels by collecting expressed proteins and peptides deposited in the PRIDE database (38). To assess potential disease linkages or associations, we integrated OMIM (20) and genome-wide association datasets (39), from public data deposited in the Genetic

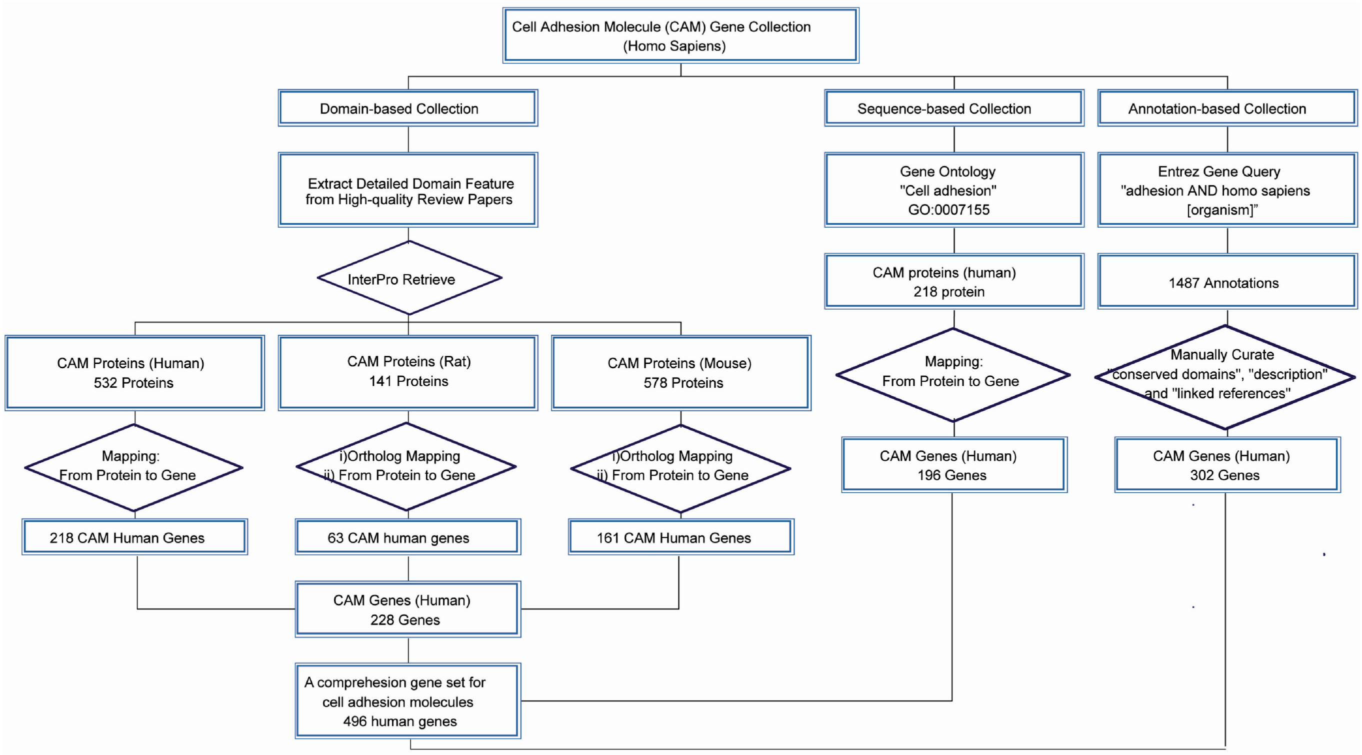


Figure 1. Collection of Human CAMs. CAMs were compiled by integrating Gene Ontology annotations, domain structure information and keywords query against NCBI Entrez Gene annotations. Four hundred and ninety-six unique human genes were identified as CAMs (additional genes that may also function in this way are identified in the supplement).

Association Database (39) and an additional 12 in-house genome wide association datasets.

Full descriptions of the annotation statistics are provided in Table 1. These annotations, extending from genome to post-translational modification, provide a novel avenue for studies of the global properties of CAM genes, overrepresented types of variation, overrepresented regulation modes and expression patterns, and disease associations, as we discuss in the following sections.

CONSTRUCTION OF A CAMO

We iteratively organized the information and knowledge for CAMs to construct a novel CAMO. CAMO was constructed as a directed acyclic graph (DAG) using DAG-Edit (40) to input, manage and update data, as shown in the screenshot (Supplementary Figure 1). We annotated each term with name, definition and source references. We added its relationship to other terms based on manual reviews of domain architecture and functional annotations at the five levels noted above.

If vertices represent terms and the relationships between terms are represented by edges, the terms in a DAG can be connected *via* a directed graph without cycles. CAMO thus provides a hierarchical description of functions and properties of CAMs with five top-level categories: CAM gene families, CAM genetics, CAM regulation, CAM expression and CAM diseases. Each top-level term is further divided into several categories to describe the functions in detail (Figure 2). *In toto*, CAMO has 850 terms with up to seven levels of depth. We mapped the 496 human genes that function in cell adhesion onto CAMO, providing a novel systematic description of CAMs (Figure 2). CAMO thus provides more specific, complete and resolved information about CAMs to scientists, especially to neuroscientists, than is available in general-purposed ontologies such as MeSH (41) and Gene Ontology (21).

OKCAM WEB INTERFACE DESIGN

We developed a PostgreSQL database termed 'OKCAM (Ontology-based Knowledgebase for Cell Adhesion Molecules)' to manage the CAM gene list, annotations and 'CAMO'. We implemented a web-based user interface of this database that uses PHP and PHP/SQL query scripts. Cross-references to key external databases were included to integrate functional information about CAM genes. These external databases provide annotations for CAM gene families, CAM genetics and genomics, CAM regulation modes and expression patterns, and relationships between CAMs and human diseases (Figure 3).

The information for each CAM gene is integrated and presented in a single graphical web page. For example, the OKCAM entry page for cadherin 1 (CDH1) (http://okcam.cbi.pku.edu.cn/entry-info.php?id = 999) shows that CDH1 is located on chromosome 16 in a chromosome

Table 1. Annotations for CAM genes

Description	Evidence entry no.	Annotated gene no.	Annotation coverage (%)	Reference
CAM gene families and basic information				
NCBI Entrez Gene annotations	496	496	100.0	(20)
Pubmed entries	18 371	490	98.8	(20)
Gene ontology annotations	4728	478	96.4	(21)
Protein interactions in BIND	230	77	15.5	(23)
Protein interactions in HPRD	2225	218	44.0	(24)
Protein interactions in BioGRID	2566	199	40.1	(22)
Enriched KEGG pathways	17	285	57.4	(25)
Enriched BioCarta pathways	19	259	52.2	NA
Enriched PID pathways (Pathway interaction database)	16	396	79.8	NA
CAM genetics				
Recombination hotspots	714	252	50.8	(26)
Chromosome insertion/deletion	493	159	32.1	(27)
GVD CNV	371	150	30.2	(27)
SNP in CDS regions	3756	418	84.3	(20)
SNP in UTR regions	4489	429	85.5	(20)
SNP in Intron regions	236 213	427	86.1	(20)
CAM expression				
Unigene expression profiles	24 480	451	90.9	(20)
Allen Brain Atlas expression Profiles (express in brain)	6205	355	71.6	(37)
Allen Brain Atlas expression Profiles (high expressed in brain)	1326	78	15.7	(37)
Proteomics evidence in PRIDE	15 331	277	55.8	(38)
CAM regulation				
Validated transcription factor binding sites in transfac	125	21	4.2	(29)
Validated transcription factor binding sites by chip-chip	189	140	28.2	(29)
Putative miRNA targets in PicTar	35 513	236	47.6	(32)
Validated miRNA targets in TarBase	5	5	1.0	(30)
Validated miRNA targets in Argonaute	419	109	22.0	(31)
Cis-NATs regulation	221	219	44.2	(34,35)
Trans-NATs regulation	145	36	7.2	(34,35)
Noncoding RNA loci	167	95	19.2	(33)
Alternative splicing	11 026	465	93.8	NA
Experiment validated PTMs	3080	373	75.2	(36)
Putative PTMs	15 829	401	80.8	(36)
Possible CAM diseases and disorders				
OMIM (with phenotype)	144	75	15.1	(20)
Vulnerable markers identified by GWA	647	80	16.1	(39)
In-house GWA vulnerable markers	121	64	12.9	NA

region that contains a recombination hotspot, copy number variations and insersion/deletions (‘CAM genetics information’). CDH1 transcripts are relatively highly expressed in adult (‘developmental stage’), mammary gland (‘tissue distribution’) and cerebral cortex (‘brain region’). Translation products are also expressed in placenta/blood serum (‘protein expression’). CDH1 is implicated in neoplasia by genomewide association studies and OMIM annotations (‘CAM disease’). Potential CDH1 regulatory modes include alternative splicing regulation, cis-NATs regulation, miRNA regulation as well as post-translational modifications (‘CAM regulation’). Links to the original databases and other resources facilitate information tracing.

We implemented four interactive browsing options in OKCAM to facilitate user queries. Users can browse cell adhesion genes by ‘CAMO’, displayed as hierarchical trees on the homepage. They can zoom in on a particular branch of the ontology by clicking the ‘+’ sign to expand the branch. For example, a user interested in ‘psychiatric disorders’ may expand this category, focus on ‘drug addiction’ and see the 49 CAM genes currently mapped on this term by clicking the number that follows this term (Figures 2 and 3). A ‘Chromosomal Overview’ browser supports browsing the CAM genes by clicks on chromosomal locations marked by ‘+++’ (Figure 3). A text search interface facilitates database queries that use either gene IDs or names. A fourth interface supports sequence searching based on BLAST nucleotide and amino acid sequence similarities. Each interactive browsing interface returns CAM gene/gene lists that meet query requirements. Users can then obtain further detailed annotation by clicking on the gene name (Figure 3). A download page makes all data, database schema and PostgreSQL commands available at http://okcam.cbi.pku.edu.cn/download.php.

APPLICATIONS OF OKCAM

The comprehensive annotations and ontology system of OKCAM facilitate studies of the global properties of the CAM genes, overrepresented types of variation, overrepresented regulation modes and expression patterns, and disease associations.

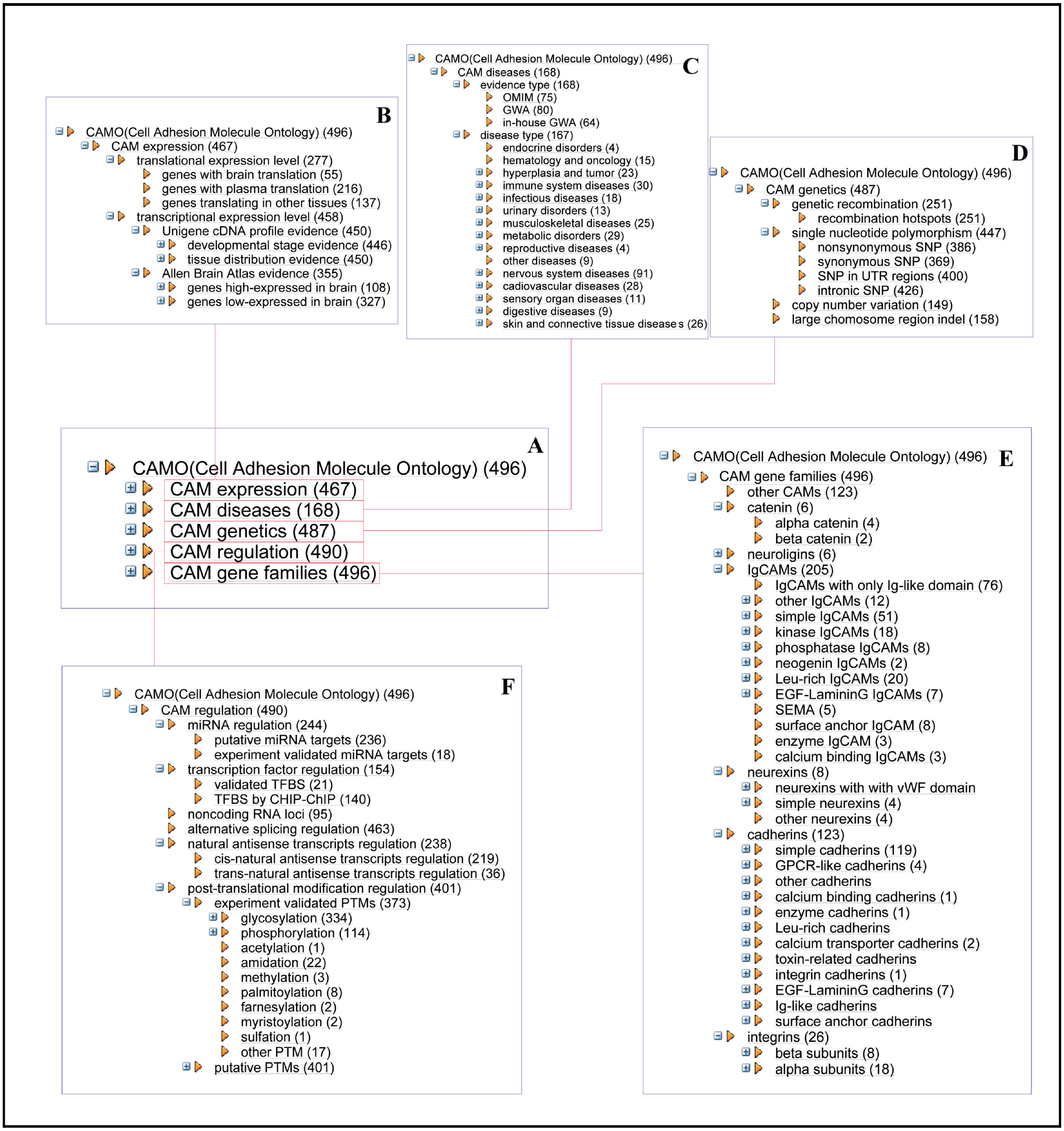
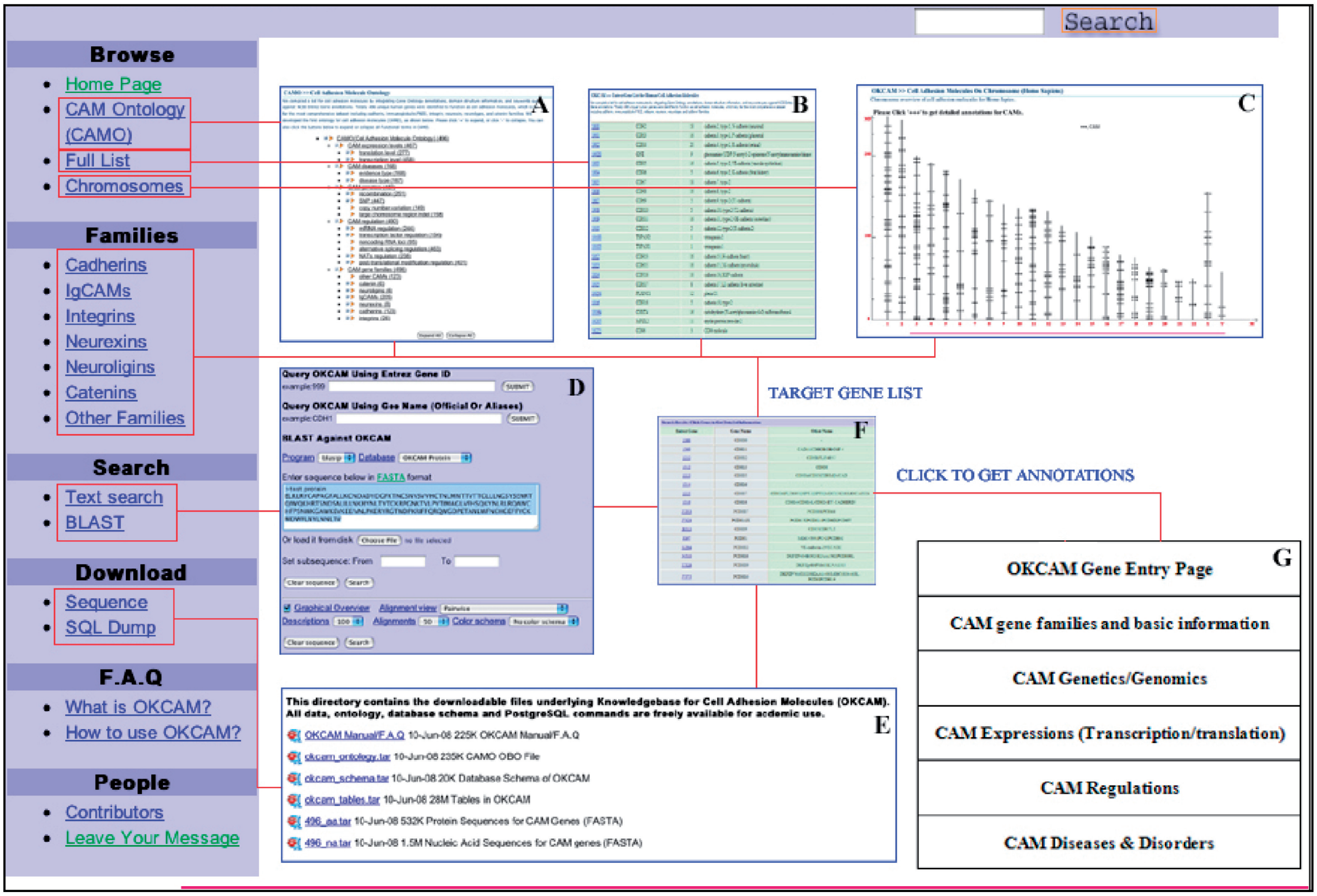

C
CAMO(Cell Adhesion Molecule Ontology) (496)
CAM diseases (168)
evidence type (168)
OMIM (75)
GWA (80)
in-house GWA (64)
disease type (167)
endocrine disorders (4)
hematology and oncology (15)
hyperplasia and tumor (23)
immune system diseases (30)
infectious diseases (18)
urinary disorders (13)
musculoskeletal diseases (25)
metabolic disorders (29)
reproductive diseases (4)
other diseases (9)
nervous system diseases (91)
cadiovascular diseases (28)
sensory organ diseases (11)
digestive diseases (9)
skin and connective tissue disease s (26)
B
CAMO(Cell Adhesion Molecule Ontology) (496)
CAM expression (467)
translational expression level (277)
genes with brain translation (55)
genes with plasma translation (216)
genes translating in other tissues (137)
transcriptional expression level (458)
Unigene cDNA profile evidence (450)
developmental stage evidence (446)
tissue distribution evidence (450)
Allen Brain Atlas evidence (355)
genes high-expressed in brain (108)
genes low-expressed in brain (327)
D
CAMO(Cell Adhesion Molecule Ontology) (496)
CAM genetics (487)
genetic recombination (251)
recombination hotspots (251)
single nucleotide polymorphism (447)
nonsynonymous SNP (386)
synonymous SNP (369)
SNP in UTR regions (400)
intronic SNP (426)
copy number variation (149)
large chomosome region indel (158)
A
CAMO(Cell Adhesion Molecule Ontology) (496)
CAM expression (467)
CAM diseases (168)
CAM genetics (487)
CAM regulation (490)
CAM gene families (496)
E
CAMO(Cell Adhesion Molecule Ontology) (496)
CAM gene families (496)
other CAMs (123)
catenin (6)
alpha catenin (4)
beta catenin (2)
neuroligins (6)
IgCAMs (205)
IgCAMs with only Ig-like domain (76)
other IgCAMs (12)
simple IgCAMs (51)
kinase IgCAMs (18)
phosphatase IgCAMs (8)
neogenin IgCAMs (2)
Leu-rich IgCAMs (20)
EGF-LamininG IgCAMs (7)
SEMA (5)
surface anchor IgCAM (8)
enzyme IgCAM (3)
calcium binding IgCAMs (3)
neurexins (8)
neurexins with with vWF domain
simple neurexins (4)
other neurexins (4)
cadherins (123)
simple cadherins (119)
GPCR-like cadherins (4)
other cadherins
calcium binding cadherins (1)
enzyme cadherins (1)
Leu-rich cadherins
calcium transporter cadherins (2)
toxin-related cadherins
integrin cadherins (1)
EGF-LamininG cadherins (7)
Ig-like cadherins
surface anchor cadherins
integrins (26)
beta subunits (8)
alpha subunits (18)
F
CAMO(Cell Adhesion Molecule Ontology) (496)
CAM regulation (490)
miRNA regulation (244)
putative miRNA targets (236)
experiment validated miRNA targets (18)
transcription factor regulation (154)
validated TFBS (21)
TFBS by CHIP-ChIP (140)
noncoding RNA loci (95)
alternative splicing regulation (463)
natural antisense transcripts regulation (238)
cis-natural antisense transcripts regulation (219)
trans-natural antisense transcripts regulation (36)
post-translational modification regulation (401)
experiment validated PTMs (373)
glycosylation (334)
phosphorylation (114)
acetylation (1)
amidation (22)
methylation (3)
palmitoylation (8)
farnesylation (2)
myristoylation (2)
sulfation (1)
other PTM (17)
putative PTMs (401)

Figure 2. Structure of CAMO. CAMO provides a hierarchical description of functions and properties of CAMs with five top-level categories (**A**): CAM expression (**B**), CAM diseases (**C**), CAM genetics (**D**), CAM gene families (**E**) and CAM regulation (**F**). Each top-level term is further divided into several categories that allow more detailed functional descriptions.

GLOBAL FEATURES OF CAMs

CAMs in our dataset were annotated using Gene Ontology (GO) (21) and the pathway databases KEGG (25), BioCarta and Pathway Interaction Database (PID). We can thus identify significantly enriched Gene Ontology terms and pathways using DAVID (42) and KOBAS (43,44), respectively. We selected the functional categories that were more likely to be biologically meaningful by calculating the statistical significance of each functional category in the input set of genes versus all annotated genes in the human genome. There was statistically significant enrichment for CAM genes in 16



This directory contains the downloadable files underlying Knowledgebase for Cell Adhesion Molecules (OKCAM). All data, ontology, database schema and PostgreSQL commands are freely available for acdemic use.

OKCAM Manual/F.A.Q 10-Jun-08 225K OKCAM Manual/F.A.Q

okcam_ontology.tar 10-Jun-08 235K CAMO OBO File

okcam_schema.tar 10-Jun-08 20K Database Schema of OKCAM

okcam_tables.tar 10-Jun-08 28M Tables in OKCAM

496_aa.tar 10-Jun-08 532K Protein Sequences for CAM Genes (FASTA)

496_na.tar 10-Jun-08 1.5M Nucleic Acid Sequences for CAM genes (FASTA)

Figure 3. Structure of OKCAM Web Server. Several interactive browsing options were implemented to facilitate user queries of OKCAM. These include ontology overview (**A**), full gene list overview (**B**), chromosomal overview (**C**), text search (**D**) and BLAST search (D). Each interactive browsing interface returns CAM gene/gene lists that meet query requirements (**F**). Users can then obtain further detailed annotations mentioned above by clicking on gene names (**G**). A download page makes all data, database schema and PostgreSQL commands available (**E**).

'molecular function' terms (Supplementary Table 1), 11 'subcellular localization' terms (Supplementary Table 2) and 45 'biological processes' terms (Supplementary Table 3), when compared to corresponding data for the whole genome.

Identification of functional enrichment for several of the 'molecular function' and 'subcellular localization' terms is reassuring. This identification provides relatively little additional information, however, since CAMs do function as 'adhesion molecules'. Most are well documented to sit within (or be anchored to) plasma membranes. However, there is also significant enrichment for other molecular functions that might not have been so readily anticipated, including calcium binding, protein kinase, and protein phosphatase activities (Supplementary Tables 1 and 4). The significant overrepresentation of CAM localizations within receptor complexes and extracellular matrix is also of interest (Supplementary Table 2). It is interesting that the CAMs identified in this work are overrepresented in not only 'cell adhesion' but also in biological processes that include signal transduction, responses to external stimuli, cell motility, migration, and nervous system development (Supplementary Table 3). Reassuringly, the molecular pathway enrichment analyses that used each of the three different pathway databases provided results that implicated their roles in largely similar functional pathways (Supplementary Table 5).

Data from OKCAM annotations for protein interactions allowed us to develop a molecular network based on proteins that could interact with the CAMs identified here (Supplementary Figure 2). As for other established biological networks (45,46), the connectivity distribution of the network that we nominated in this way appears to follow scale-free rules. CAMs appear to interact with each other to form a relatively tight 'core' that interrelates with hundreds of other signal transduction genes. Focus on the 'hub nodes' in this apparent network (Supplementary Figure 2) may even help to elucidate novel CAM roles in signal transduction that come from its partnerships with other signaling molecules.

CAM REGULATORY MODES

Mapping the CAMs in our dataset onto CAMO and detailed gene structural/regulatory terms allows us to identify specific potential regulatory modes for these CAMs.

We can then perform Monte Carlo analyses to test whether these structural/regulatory modes are overrepresented among CAMs. On human genomic level, both recombination 'hotspots' (Monte Carlo $P = 0.024$) and copy number variations (Monte Carlo $P < 0.0001$) are over-represented in chromosome regions that contain CAM genes. Indeed, 'cell adhesion molecule' is the GO category that is most enriched in the genes that overlap with 1447 copy number variants identified using Affymetrix 500 K and whole genome TilePath (WGTP) reagents (47). There is a more modest but still significant 1.42-fold enrichment for CAM genes in chromosomal regions that contain both copy number variations and recombination hotspots ($P = 0.07$). By contrast, we detected no significant difference for the densities of single nucleotide polymorphisms (SNP) distributions in chromosomal regions that contain CAM genes versus the whole genome ($P > 0.5$).

When we tested potential overrepresentation of transcriptional regulatory modes using hypergeometric tests, we found that the potential for miRNA regulation was significantly enriched for CAM genes when compared to the whole genome ($P < 0.0001$). In contrast, no over-represented transcription factor regulation for CAM genes were detected using either low scale experimentally validated ($P = 0.37$) or ChIP-chip data ($P = 0.51$). There was no significant over- or under-representation of CAMs among genes involved in either *cis*- or *trans*-NAT (35) regulation ($P > 0.5$ for each).

We can also seek overrepresentation of CAM alternative splicing by compiling the alternative splicing isoforms for each human gene mapped on CAMO and plotting the distributions of the numbers of isoforms for (i) CAMs versus (ii) all human genes (Supplementary Figure 3). The overall distributions appear similar. However, genes that utilize a wealth of alternative transcripts, those that encode ~40–50 alternatively spliced isoforms, are over-represented in the dataset that encodes CAMs. These genes provide an apparently distinct 'peak' in the distribution curve (Supplementary Figure 3). This analysis agrees with our previous work that has characterized multiple alternative splicing events in specific addiction-associated CAMs (13).

We integrated post-translational modification (PTM) data to identify possible contributions of this regulatory mode to CAM functions. On the basis of the experimentally validated PTM data deposited in dbPTM, the 496 CAM genes are candidates for involvement in glycosylation (334 genes), phosphorylation (114 genes), amidation (22 genes), palmitoylation (eight genes), methylation (three genes), farnesylation (two genes), myristoylation (two genes), sulfation (one gene) and acetylation (one gene). There is a highly significant enrichment for CAM N-linked glycosylation (331 genes, $P < 0.0001$), but not for O-linked glycosylation (10 genes). No significant over- or under-representation was detected for other modes of post translational modification.

On the basis of the OKCAM annotations and CAMO, we identified a list of regulatory modes for cell adhesion molecules. These analyses identified both expected and unexpected CAM regulatory modes. First, the data document the overrepresentation of CNVs within CAM genes, in ways that were suggested in even some of the initial descriptions of CNVs (48). Documenting a 1.4-fold enrichment for CAM genes in chromosomal regions that contain both copy number variations and recombination hotspots both supports these initial observations and provides a possible mechanism for the abundance of CNVs in CAM genes. Secondly, although many papers have described many alternative splicing isoforms for CAMs, it was somewhat surprising to note that the largest diversity of alternative transcripts (e.g. ~40–50) was selectively over-represented among CAM genes.

CAM EXPRESSION PATTERNS

Integration of data from human expressed sequence tags (EST) derived from brain libraries and mouse brain atlas expression profiles provided strong levels of agreement that support use of this comparative approach (Supplementary Table 6). We thus analyzed CAM expression patterns and levels in 17 mouse brain regions, based on Allen Brain Atlas profiles from murine brains. For each brain region, we used the program R to plot the density curves that illustrate the frequency distributions of expression levels for (i) CAMs and (ii) all human genes expressed in this brain region (Supplementary Figure 4). For 16 of the 17 brain regions, the expression distribution curves for the two datasets merged. In these brain regions, CAM genes taken as a group appear to be expressed in ways that are not markedly different from those of other brain-expressed genes. However, in the cerebral cortex, CAM genes with the highest expression levels appear to be over-represented. There is thus an additional peak in the CAM distribution curve that is not found when all other genes are examined (Supplementary Figure 4). While much prior data documents expression of many CAMs in cerebral cortex, the specificity of the relatively richer expression of CAMs in this brain region provides a novel observation.

CAM DISEASE ASSOCIATIONS

We assessed potential relationships between CAM variants and disease using data from OMIM, public GWAS data and our in-house datasets. These data nominate 167 human CAMs as likely to contain variants that could contribute to individual differences in vulnerability to disorders in brain and a variety of other organs (Figure 4). CAMs were identified by association and/or linkage findings in disorders of the nervous system (91 genes), immune system (30 genes), metabolism (29 genes), cardiovascular system (28 genes), skin and connective tissues (26 genes), musculoskeletal system (25 genes) and hyperplasia and/or tumors (23 genes). When assessed in relation to specific disorders or narrower classes of disorders, there were relatively large numbers of cell adhesion molecules implicated in substance dependence (49 genes), Alzheimer's disease (42 genes), tumors (21 genes), heart disease (20 genes), bipolar disorder (18 genes), autoimmune diseases (19 genes) and diabetes mellitus (17 genes). The number of CAMs whose variants are tentatively

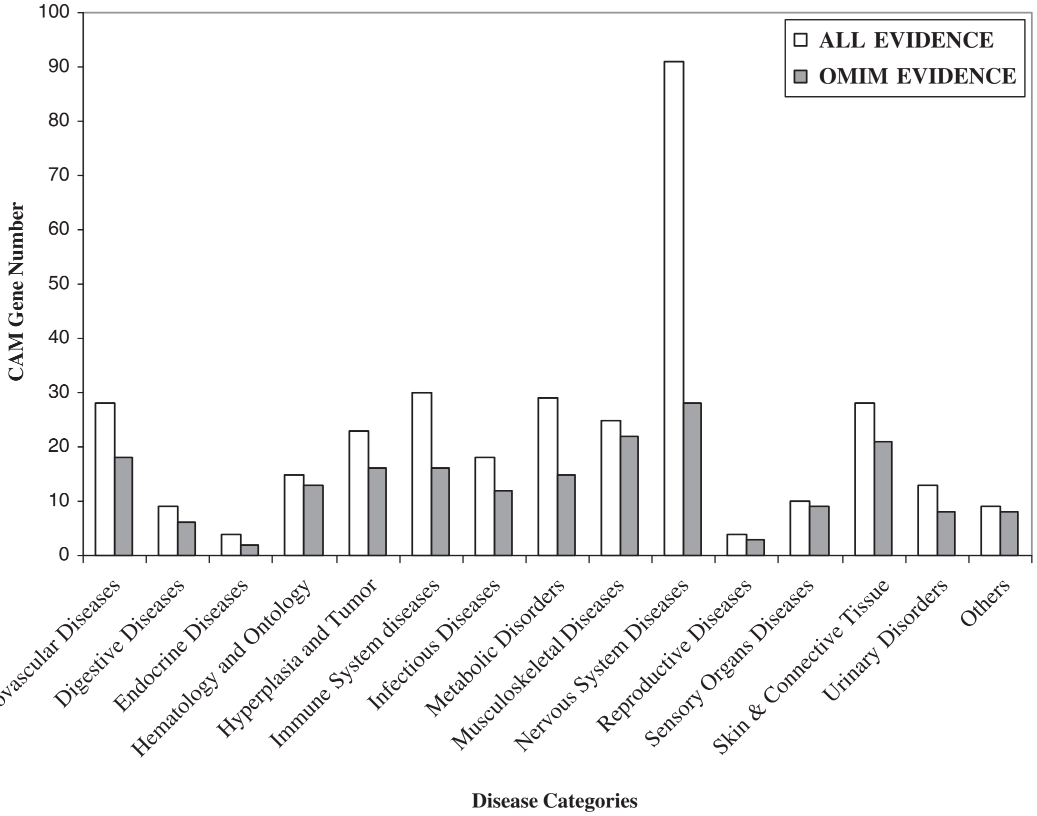


Figure 4. Distribution of CAM in OMIM and GWA. OMIM, GWA and/or our in-house GWA data implicates variants in at least 167 (of the 496) CAM genes in various diseases. Data from OMIM shares disease distribution patterns with that from GWA studies.

implicated in nervous system phenotypes is larger than anticipated by chance (Figure 4). The distribution of findings in other disorders is similar to that displayed by all genes, when comparing data from either OMIM or GWA datasets.

DISCUSSION

‘Cell adhesion molecules’ are increasingly recognized as ‘cell adhesion receptors’, since many of their functions are just ‘cell glue’ but rather are more consistent with roles in cell–cell and cell–matrix interactions and in molecular recognition events that transduce signals. The computational approaches that we use here to define and characterize a universe of ‘cell adhesion’ molecules provide both expected and unexpected results. These results should be assessed in light of the strengths and limitations of the approaches used here, and the strengths and limitations of the underlying datasets employed for these analyses. We also discuss details of the strengths and limitations of these data in Supplementary Text 1.

We have attempted to provide as comprehensive a list of human CAM genes, annotations and ontology-based CAM knowledgebase as possible. However, it is clear that there will be rapid progress in the study of these molecules and of cell adhesion mechanisms. The OKCAM database provides means for integrating new data and updating knowledge, in ways that should facilitate better and better understanding of the global and specific CAM properties. As CAM genomic features regulatory modes, expression patterns and disease associations become clearer, we thus hope that OKCAM should become even more comprehensive and useful.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Drs T. Drgon, A. Hishimoto, Y. Zhang and X. Yu for insightful suggestions. We are grateful to Shuqi Zhao, Xizeng Mao, Zhi-Yu Peng and Qi-Yao Li for assistance with OKCAM Web Server development.

FUNDING

National Institutes of Health Intramural Research Program (NIDA), NIH grants P50CA/DA84718; China Scholarship Council (C.Y.L.); China National High-tech 863 Programs (2006AA02A312, 2006AA02Z334); 973 Programs (2007CB946904). Funding for open access charge: NIDA/IRP grants P50CA/DA84718.

Conflict of interest statement. None declared.

REFERENCES

1. Yamada,S. and Nelson,W.J. (2007) Synapses: sites of cell recognition, adhesion, and functional specification. *Annu. Rev. Biochem.*, **76**, 267–294.
2. Takeichi,M. and Abe,K. (2005) Synaptic contact dynamics controlled by cadherin and catenins. *Trends Cell. Biol.*, **15**, 216–221.
3. Hishimoto,A., Liu,Q.R., Drgon,T., Pletnikova,O., Walther,D., Zhu,X.G., Troncoso,J.C. and Uhl,G.R. (2007) Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms. *Hum. Mol. Genet.*, **16**, 2880–2891.
4. Kim,H.G., Kishikawa,S., Higgins,A.W., Seong,I.S., Donovan,D.J., Shen,Y., Lally,E., Weiss,L.A., Najm,J., Kutsche,K. *et al.* (2008) Disruption of neurexin 1 associated with autism spectrum disorder. *Am. J. Hum. Genet.*, **82**, 199–207.
5. Shapiro,L., Love,J. and Colman,D.R. (2007) Adhesion molecules in the nervous system: structural insights into function and diversity. *Annu. Rev. Neurosci.*, **30**, 451–474.
6. Stoker,A.W. (2005) Protein tyrosine phosphatases and signalling. *J. Endocrinol.*, **185**, 19–33.
7. Salinas,P.C. and Price,S.R. (2005) Cadherins and catenins in synapse development. *Curr. Opin. Neurobiol.*, **15**, 73–80.
8. Hirano,S., Suzuki,S.T. and Redies,C. (2003) The cadherin superfamily in neural development: diversity, function and interaction with other molecules. *Front. Biosci.*, **8**, d306–355.
9. Song,J.Y., Ichtchenko,K., Sudhof,T.C. and Brose,N. (1999) Neuroligin 1 is a postsynaptic cell-adhesion molecule of excitatory synapses. *Proc. Natl Acad. Sci. USA*, **96**, 1100–1105.
10. Dityatev,A., Dityateva,G. and Schachner,M. (2000) Synaptic strength as a function of post- versus presynaptic expression of the neural cell adhesion molecule NCAM. *Neuron*, **26**, 207–217.
11. Butcher,L.M., Meaburn,E., Dale,P.S., Sham,P., Schalkwyk,L.C., Craig,I.W. and Plomin,R. (2005) Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. *Mol. Psychiatry*, **10**, 384–392.
12. Johnson,C., Drgon,T., Liu,Q.R., Walther,D., Edenberg,H., Rice,J., Foroud,T. and Uhl,G.R. (2006) Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **141B**, 844–853.
13. Liu,Q.R., Drgon,T., Johnson,C., Walther,D., Hess,J. and Uhl,G.R. (2006) Addiction molecular genetics: 639,401 SNP whole genome association identifies many 'cell adhesion' genes. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **141**, 918–925.
14. Uhl,G.R., Liu,Q.R., Drgon,T., Johnson,C., Walther,D., Rose,J.E., David,S.P., Niaura,R. and Lerman,C. (2008) Molecular genetics of successful smoking cessation: convergent genome-wide association study results. *Arch. Gen. Psychiatry*, **65**, 683–693.
15. Arking,D.E., Cutler,D.J., Brune,C.W., Teslovich,T.M., West,K., Ikeda,M., Rea,A., Guy,M., Lin,S., Cook,E.H. *et al.* (2008) A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am. J. Hum. Genet.*, **82**, 160–164.
16. Munafo,M.R., Attwood,A.S. and Flint,J. (2008) Neuregulin 1 genotype and schizophrenia. *Schizophr. Bull.*, **34**, 9–12.
17. Velayos-Baeza,A., Toma,C., da Roza,S., Paracchini,S. and Monaco,A.P. (2007) Alternative splicing in the dyslexia-associated gene KIAA0319. *Mamm. Genome*, **18**, 627–634.
18. Paracchini,S., Thomas,A., Castro,S., Lai,C., Paramasivam,M., Wang,Y., Keating,B.J., Taylor,J.M., Hacking,D.F., Scerri,T. *et al.* (2006) The chromosome 6p22 haplotype associated with dyslexia reduces the expression of KIAA0319, a novel gene involved in neuronal migration. *Hum. Mol. Genet.*, **15**, 1659–1666.
19. Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: Tools for Protein Sequence Classification and Comparison. *Methods Mol. Biol.*, **396**, 59–70.
20. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
21. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
22. Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al.* (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–640.
23. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
24. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–414.
25. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
26. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–779.
27. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
28. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
29. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites.10.1093/nar/24.1.238. *Nucleic Acids Res.*, **24**, 238–241.
30. Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.
31. Shahi,P., Loukianiouk,S., Bohne-Lang,A., Kenzelmann,M., Kuffer,S., Maertens,S., Eils,R., Grone,H.J., Gretz,N. and Brors,B. (2006) Argonaute–a database for gene regulation by mammalian microRNAs. *Nucleic Acids Res.*, **34**, D115–D118.
32. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
33. He,S., Liu,C., Skogerbo,G., Zhao,H., Wang,J., Liu,T., Bai,B., Zhao,Y. and Chen,R. (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, **36**, D170–D172.
34. Zhang,Y., Li,J., Kong,L., Gao,G., Liu,Q.R. and Wei,L. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res.*, **35**, D156–D161.
35. Zhang,Y., Liu,X.S., Liu,Q.R. and Wei,L. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.*, **34**, 3465–3475.
36. Lee,T.Y., Huang,H.D., Hung,J.H., Huang,H.Y., Yang,Y.S. and Wang,T.H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.
37. Lein,E.S., Hawrylycz,M.J., Ao,N., Ayres,M., Bensinger,A., Bernard,A., Boe,A.F., Boguski,M.S., Brockway,K.S., Byrnes,E.J. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.
38. Jones,P., Cote,R.G., Martens,L., Quinn,A.F., Taylor,C.F., Derache,W., Hermjakob,H. and Apweiler,R. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **34**, D659–D663.
39. Lin,B.K., Clyne,M., Walsh,M., Gomez,O., Yu,W., Gwinn,M. and Khoury,M.J. (2006) Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am. J. Epidemiol.*, **164**, 1–4.
40. Smith,B., Ceusters,W., Klagges,B., Kohler,J., Kumar,A., Lomax,J., Mungall,C., Neuhaus,F., Rector,A.L. and Rosse,C. (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
41. Lipscomb,C.E. (2000) Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.*, **88**, 265–266.

42. Huang da,W., Sherman,B.T., Tan,Q., Kir,J., Liu,D., Bryant,D., Guo,Y., Stephens,R., Baseler,M.W., Lane,H.C. *et al.* (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.

43. Wu,J., Mao,X., Cai,T., Luo,J. and Wei,L. (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34**, W720–W724.

44. Mao,X., Cai,T., Olyarchuk,J.G. and Wei,L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21**, 3787–3793.

45. Dunker,A.K., Cortese,M.S., Romero,P., Iakoucheva,L.M. and Uversky,V.N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.*, **272**, 5129–5148.

46. Han,J.D., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J., Cusick,M.E., Roth,F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

47. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

48. Jakobsson,M., Scholz,S.W., Scheet,P., Gibbs,J.R., VanLiere,J.M., Fung,H.C., Szpiech,Z.A., Degnan,J.H., Wang,K., Guerreiro,R. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.