nature REVIEWS
DRUG DISCOVERY

Unexplored opportunities in the druggable human genome

The IDG Knowledge Management Center

Much of biomedical research and the development of therapeutics is focused on a small fraction of the human genome, ignoring many disease-relevant proteins and the associated scientific and commercial opportunities. The National Institutes of Health (NIH) Illuminating the Druggable Genome programme aims to catalyse research around understudied targets. As part of this effort, we categorized proteins based on available data, information and knowledge from various sources. Furthermore, we evaluated the commercial focus on therapeutic targets and associated disease areas, by processing global drug sales data for a 5-year period. Taken together, our quantification of available data and knowledge, as well as the financial analyses based on global drug sales and NIH funding for that same period, highlight a bias towards well-described proteins, particularly drug targets. Our analysis identifies a knowledge deficit, not just concerning experimental data, but also with respect to patents and publications. We believe this points towards novel research and commercial opportunities for drug discovery and development.

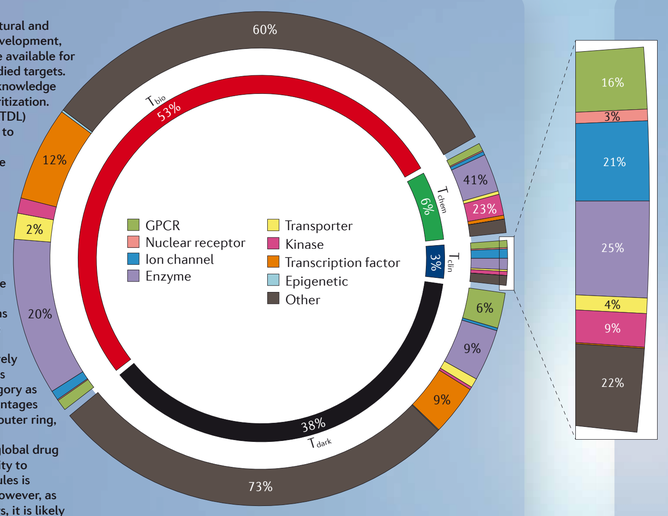
Supported by:
IDG KMC
Sales data provided by:
imshealth™
INTELLIGENCE APPLIED.

Target development levels

Most current protein classification schemes are based on structural and functional criteria. For target prioritization and therapeutic development, it is useful to understand how much and what types of data are available for a given protein, thereby highlighting well-studied and understudied targets. Thus a classification scheme based on the amount and types of knowledge that are available can lay the foundation for rational target prioritization.

To address this, we developed the target development level (TDL) classification scheme (see footnote*), which uses three criteria to classify a protein into one of four categories, according to the following ordering. Proteins annotated as drug targets by mode of action[1] were categorized as T_{clin}. Proteins for which small-molecule modulators (with reported bioactivity above specific cut-off values; see footnote†) are known were categorized as T_{chem}. Proteins categorized as T_{bio} either have confirmed OMIM disease phenotypes or Gene Ontology (GO) annotations[2] based on experimental evidence, or meet two of the following three conditions: a fractional count of PubMed abstracts mentioning the protein[3] above five, three or more Gene Reference Into Function (RIF) annotations[4], or more than fifty antibodies according to Antibodypedia[5]. The remaining proteins are classified as T_{dark} and are the proteins for which there is the least knowledge and the lowest number of molecular probes available. Moreover, research on these proteins receives relatively little funding, as indicated by the distribution of NIH R01 grants below. The figure to the immediate right shows each TDL category as percentage of the whole proteome in the inner ring, and percentages of each TDL category for selected major target families in the outer ring, with the T_{clin} category expanded.

Our analysis of commercial activity on the right is based on global drug sales, which stem from only 3% of the proteome (T_{clin}). Our ability to specifically and selectively perturb proteins using small molecules is apparently limited to less than 10% of the human proteome. However, as many of these proteins interact and re-arrange in multiple ways, it is likely that the proportion of targets that can be perturbed by small molecules exceeds 10%.



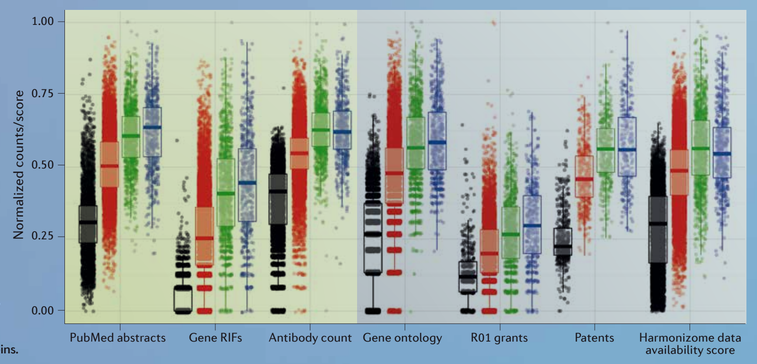
External validation of target development levels and the knowledge deficit

Different resources and data sets were used to examine the lack of annotation of proteins catalogued as T_{dark} compared with other proteins (see footnote*). From left to right, the first three normalized score distributions (on a green background) illustrate the criteria used to assign target development levels, namely PubMed abstracts, the Gene RIF counts and the antibody counts per protein.

These criteria were externally validated on four other sets of data (on a blue background) as follows:

- 'Gene ontology terms' illustrates the distribution of Gene Ontology[2] annotation counts per protein, using data from UniProt[6]. More than 90% of T_{dark} proteins, but less than 10% of T_{clin} proteins, have 10 Gene Ontology annotations or less.
- 'R01 grants' illustrates the distribution of text-mined NIH R01 grant counts detected for each protein, using data from NIH RePORTER[7]. Most T_{dark} proteins do not have research on them funded via the NIH R01 mechanism.
- 'Patents' illustrates the distribution of text-mined granted patents for each protein, using data from SureChEMBL[8]. T_{dark} proteins are included in a significantly lower number of patents compared with proteins in the other categories.
- 'Harmonizome data availability score' illustrates experimental information density per protein, processed from 70 genomic data sets[9]. T_{dark} proteins appear to have less information compared with the other categories.

Through the IDG KMC project, various sources of information related to biological assays, human and mouse phenotypes, disease associations, expression data from transcriptomic and proteomic experiments, pathways and other processes and functions have been integrated and harmonized. There is an uneven distribution of information for targets in multiple types of information, including research literature, patent literature and experimental data. In particular, there appears to be a substantial knowledge deficit for T_{dark} proteins, suggesting a considerable opportunity for novel discoveries, and ultimately new commercial opportunities, if further resources are directed towards investigating such proteins.



A target-centric analysis of global drug sales data

While evaluating the current level of knowledge about current and potential drug targets, we asked the question: "what are the most lucrative targets?" from a therapeutic perspective. To investigate this question, we analysed IMS Health data on drug sales from 75 countries, including Japan and those in Europe and North America, aggregated over a 5-year period (2011–2015). After excluding botanicals, traditional Chinese and homeopathic medicines and drugs perturbing non-human targets (such as some antiviral drugs), we identified 51,095 unique products. These were mapped to 1,069 active pharmaceutical ingredients (APIs) from DrugCentral[10], corrected by number of APIs per product and by number of T_{clin} targets per API. Sales organized by ATC level 2 codes[11] and by target class were normalized to percentage values, and depicted in a circular histogram (below); the value of the sales is indicated in the concentric circles around the histogram, with a scale in the bottom right. The top two ATC level 1 categories in terms of global sales of the associated APIs are 'antineoplastics and immunomodulators' followed by the 'nervous system'.

The commercial impact of the seven target classes discussed in this poster is summarized in the table on the top right. The most lucrative targets are G protein-coupled receptors (GPCRs), with ~27% of the overall sales. Enzymes (excluding kinases, which are considered separately given their importance), are the next class, followed by nuclear receptors and transporters. Sales of drugs that modulate these seven target classes, which include 269 T_{clin} targets and 1,027 APIs, reached nearly US$3 trillion for the 2011–2015 period. Cytokines, the latest addition (in the early 2000s) to the list, are currently the most lucrative target. The top 20 T_{clin} protein targets, ranked by aggregated API sales data, are listed in the table on the bottom right.

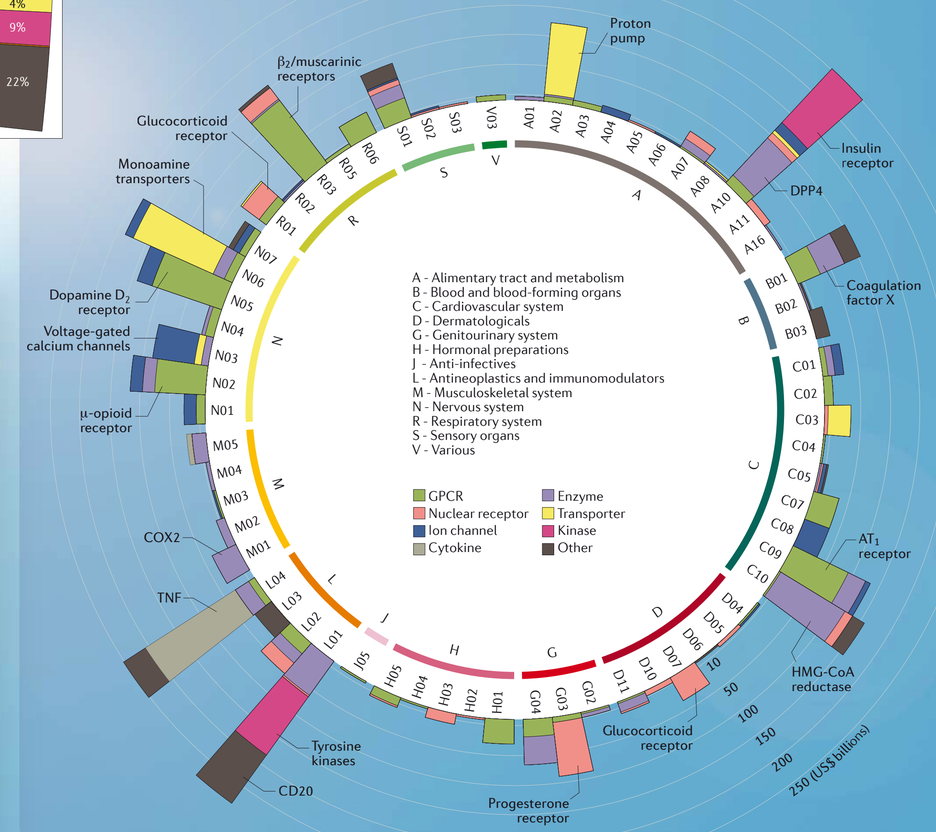


Table | **Commercial impact* of drugs according to target class**

Target class	Number of targets	Number of APIs	Sales (US$ billions)*	Market share
GPCR	72	372	889.17	27.42%
Enzyme	88	234	683.14	21.06%
Nuclear receptor	16	111	340.13	10.49%
Transporter	18	82	323.99	9.99%
Ion channel	23	167	281.11	8.67%
Kinase	43	49	240.46	7.41%
Cytokine	9	12	184.29	5.68%
Other	43	68	300.83	9.28%

*Sales are aggregated over the period 2011–2015.

Table | **Top 20 drug targets in terms of commercial impact***

Gene	Protein target	Top two APIs acting on the target by sales*	Drug category	Sales* (US$ billions)
TNF	Tumour necrosis factor	Adalimumab, etanercept	Immunosuppressants	163.39
INSR	Insulin receptor	Insulin and insulin analogues	Hypoglycaemic agents	143.55
NR3C1	Glucocorticoid receptor	Fluticasone, budesonide	Anti-inflammatory agents	142.75
HMGCR	HMG-CoA reductase	Rosuvastatin, atorvastatin	Hypolipidaemic agents	122.55
ATP4A and *ATP4B*	Proton pump (H^+/K^+-ATPase)	Esomeprazole, omeprazole	Anti-ulcer agents	118.16
AGTR1	AT_1 receptor	Valsartan, olmesartan	Antihypertensive agents	99.98
ADRB2	β_2 adrenoceptor	Salmeterol, salbutamol	Bronchodilators	90.02
OPRM1	μ-opioid receptor	Oxycodone, fentanyl	Analgesics	87.97
PTGS2	COX2	Celecoxib, paracetamol	Analgesics	84.04
DRD2	Dopamine D_2 receptor	Aripiprazole, quetiapine	Antipsychotics	74.91
CHRM1–5	Muscarinic acetylcholine receptor	Tiotropium bromide; solifenacin	Anticholinergics	64.16
SLC6A4	Sodium-dependent serotonin transporter	Escitalopram, duloxetine	Antidepressants	59.18
HTR2A	5-HT_{2A} receptor	Aripiprazole, quetiapine	Antipsychotics	57.58
CACNA1S/ CACNA1C/ CACNA1D/ CACNA1F	L-type calcium channel	Amlodipine, nifedipine	Antihypertensive agents	55.97
SLC6A2	Sodium-dependent noradrenaline transporter	Duloxetine, methylphenidate	Antidepressants and psychostimulants	55.72
VEGFA	Vascular endothelial growth factor A	Bevacizumab, ranibizumab	Antineovascularisation agents	55.15
HRH1	Histamine H_1 receptor	Olopatadine, cetirizine	Antihistamines	53.55
IFNAR1/IFNAR2	Type I interferon receptor	Interferon beta-1a, interferon beta-1b	Immunostimulants	51.40
SCN1A–5A and *SCN7A–11A*	Voltage-gated sodium channel	Lidocaine, lamotrigine	Antiarrhythmics and antiepileptics	50.64
ESR1	Oestrogen receptor	Ethinylestradiol, estradiol	Contraceptives/ oestrogen agonists	50.35

*Sales are aggregated over the period 2011–2015. 5-HT, 5 hydroxytryptamine. AT_1, angiotensin II type 1; COX2, cyclooxygenase 2; DPP4, dipeptidyl peptidase 4; HMG-CoA, 3-hydroxy-3-methylglutaryl-coenzyme A.

Illuminating the Druggable Genome Knowledge Management Center
Illuminating the Druggable Genome Knowledge Management Center (IDG KMC) provides an integrated informatics solution encompassing accrual, storage, analysis, visualization and dissemination of data and knowledge related to unannotated human proteins, focused on G-protein-coupled receptors, kinases, ion channels and nuclear receptors. IDG KMC offers informatics resources to assist investigators in uncovering the biological functions of previously uncharacterized proteins, by identifying knowledge gaps and prioritizing candidates for further experiment using multiple technology platforms, as outlined at the IDG website, http://targetcentral.ws/. These resources are disseminated via the IDG KMC portal, Pharos, https://pharos.nih.gov/, which supports efficient query, browsing and analysis of the IDG KMC database.

Footnotes and references
*The Target Central Resource Database, version 3.0.7 (September 2016), contains 20,186 human proteins (manually reviewed Uniprot entries only), which were classified into four target development levels (TDLs): T_{dark} (7,583 proteins), T_{bio} (10,759 proteins), T_{chem} (1,243 proteins), and T_{clin} (601 proteins). TDL levels were established by evaluating knowledge, as shown above. Each data point represents a protein, colour-coded by TDL; boxplots represent the interquartile range (25% to 75%) and the median for each knowledge availability metric. Raw counts and scores were first converted to log scale, then normalized to unit variance.
†The potency cut-offs for small-molecule modulators used for the target classes are: kinases: ≤30 nM; G protein-coupled receptors: ≤100 nM; nuclear receptors: ≤100 nM; ion channels: ≤10 μM; non-IDG family targets: ≤1 μM. T_{chem} proteins have bioactivities in ChEMBL[12] and DrugCentral and human curation for some targets.

1. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Disc.* doi: 10.1038/nrd.2016.230 (2016).
2. Ashburner, M. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
3. Text mining resources from JensenLab (CPR): http://jensenlab.org/
4. Gene Reference into Function (NCBI): http://www.ncbi.nlm.nih.gov/gene/about-generif
5. Antibodypedia website: http://www.antibodypedia.com/
6. UniProt Knowledgebase: http://www.uniprot.org
7. NIH RePORTER portal: https://projectreporter.nih.gov/reporter.cfm
8. SureChEMBL portal: https://www.surechembl.org/search/
9. Rouillard, A.D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* http://dx.doi.org/10.1093/database/baw100 (2016).
10. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucl. Acids Res.* doi: 10.1093/nar/gkw993; http://drugcentral.org
11. ATC classification system: http://www.whocc.no/atc_ddd_index
12. ChEMBL: https://www.ebi.ac.uk/chembl/

Authors and affiliations
- Cristian Bologa, Jayme Holmes, Subramani Mani, Steve L. Mathias, Tudor I. Oprea, Larry A. Sklar, Oleg Ursu, Anna Waller, Jeremy J. Yang and Gergely Zahoranszky-Kohalmi* are at the University of New Mexico, MSC09 5025, One University of New Mexico, 700 Camino de Salud NE, Albuquerque, New Mexico 87131, USA.
- Lars Juhl Jensen and Soren Brunak are at the Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3b, DK-2200 Copenhagen N, Denmark.
- Avi Ma'ayan and Joel Dudley are at the Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, New York 10029, USA.
- John Overington‡, Anne Hersey, Anna Gaulton, Anneli Karlson§ and George Papadatos§ are at the European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
- Rajarshi Guha, Ajit Jadhav, Dac-Trung Nguyen, Anton Simeonov and Noel Southall are at NIH-NCATS, 9800 Medical Center Drive, Rockville, Maryland 20852, USA.
- Stephan Schürer and Dušica Vidović are at Department of Pharmacology, Miller School of Medicine and Center for Computational Science, University of Miami, 1600 NW 10th Avenue, RMSB 6112B, Miami, Florida 33136, USA.
- Allen Campbell and Christian Reich are at IMS Health, 2200 Bridge Parkway, Redwood City, California 94065, USA.

Present addresses: *NIH-NCATS, 9800 Medical Center Drive, Rockville, Maryland 20852, USA. †GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK. ‡BenevolentAI, 40 Churchway, London, NW1 1LW, UK. Correspondence to T.I.O. toprea@salud.unm.edu

Acknowledgements
The authors are thankful to IMS Health for providing data for over 60,000 pharmaceutical formulations, representing aggregated 5-year sales for over 1,200 active pharmaceutical ingredients from 75 countries. Funding for this work was provided by NIH grant U54CA189205 (to T.I.O., A.S. and L.A.S.) and by NIH grant U54CA189201 (to A.M. & J.D). The production and distribution of this poster was funded by the Illuminating the Druggable Genome Knowledge Management Center, NIH grant U54CA189205. The poster content is peer reviewed and the responsibility of Macmillan Publishers Limited.
Edited by Peter Kirkpatrick; copy-edited by Carrie Hardy. http://www.nature.com/nrd/posters/druggablegenome/index.html