feature

Target 2035: probing the human proteome

Adrian J. Carter[1], adrian.carter@boehringer-ingelheim.com, **Oliver Kraemer**[1], **Matthias Zwick**[2], **Anke Mueller-Fahrnow**[3], **Cheryl H. Arrowsmith**[4,5] **and Aled M. Edwards**[4]

Biomedical scientists tend to focus on only a small fraction of the proteins encoded by the human genome despite overwhelming genetic evidence that many understudied proteins are important for human disease. One of the best ways to interrogate the function of a protein and to determine its relevance as a drug target is by using a pharmacological modulator, such as a chemical probe or an antibody. If these tools were available for most human proteins, it should be possible to translate the tremendous advances in genomics into a greater understanding of human health and disease, and catalyze the creation of innovative new medicines. Target 2035 is a global federation for developing and applying new technologies with the goal of creating chemogenomic libraries, chemical probes, and/or functional antibodies for the entire proteome.

Nature of the challenge

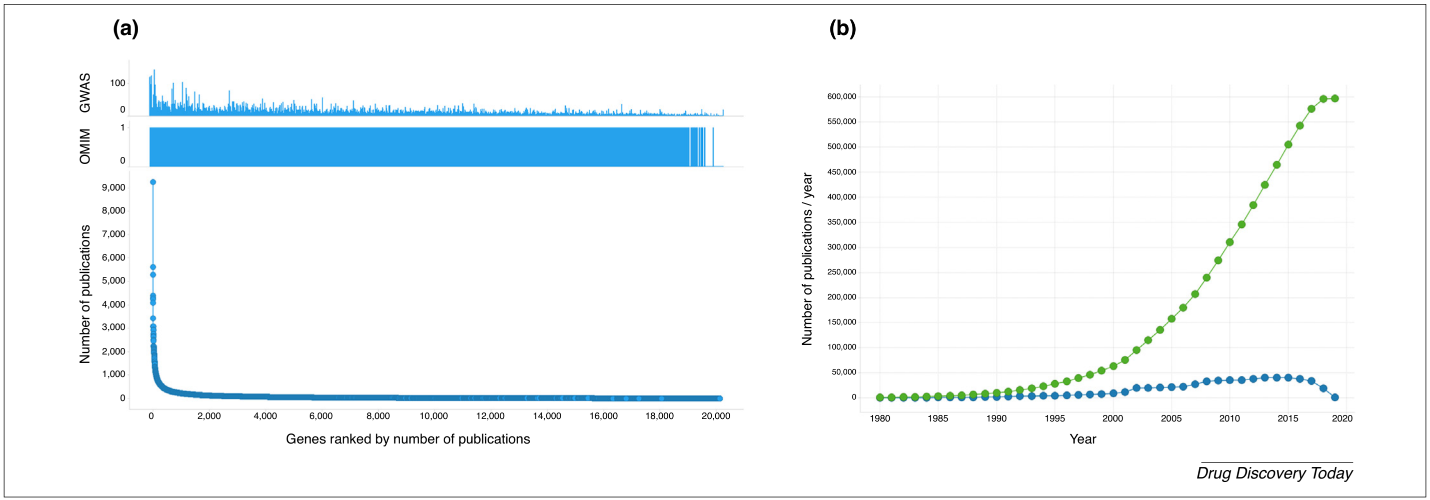
Proteins for which there is genetic evidence for a causal relationship with a disease are twice as likely to be therapeutically validated as drug targets compared with those with no clear genetic link to disease [1]. Accordingly, the sequencing of the human genome and the amazing progress in identifying mutations and genetic variations within populations correlated with disease raised great hope that many therapies would follow. Unfortunately, these largely correlative genomics studies have not yet translated to a sufficient mechanistic understanding of human biology to enable reliable development of new treatment strategies.

Developing more robust therapeutic hypotheses requires a deeper understanding of all disease-relevant genes. However, most newly sequenced genes (and their encoded proteins) are understudied and lie within the 'dark proteome' [2,3] and, as a corollary, most biomedical research is focused on only a small fraction of human genes [4]. We argue that this research ecosystem problem is the greatest impediment to our understanding of human biology and our ability to develop new medicines.

To design a solution, we first sought to quantify the extent of the problem. To interrogate the pattern of research activity across the human genome, we examined the distribution of publications for protein-coding genes and their products. Specifically, we used data sets provided by the NCBI [5] to evaluate the frequency patterns of 596 891 publications annotated to 20 122 human protein-coding genes (Fig. 1a; see also associated Data in Brief publication). The cumulative number of publications has increased annually from 1980 (Fig. 1b). This is despite a decrease in the annual number of publications in 2018 and 2019, an observation that probably reflects the backlog for entering and annotating the publications. The distribution of publications appears to resemble a power-law distribution of research activity across the human genome (Fig. 1a), which might reflect researchers obeying rules for scale-free networks in which cumulative interest begets further interest [6].

Over the past few years, investigators from a variety of fields have discovered that many networks (such as scientific papers linked by citations) are dominated by a relatively small number of nodes connected to many other nodes. Networks containing these important nodes, or hubs, are called 'scale-free', in the sense that some hubs have a seemingly unlimited number of links, whereas others have just a few, and no node is

https://doi.org/10.1016/j.drudis.2019.06.020



Drug Discovery Today

FIGURE 1

The natural tendency for scientists to focus on previously studied genes. (a) As described in the accompanying Data in Brief article, the number of publications annotated to each human protein-coding gene was plotted in order of decreasing number of publications, left to right. Since 1980, 596 891 publications in gene2pubmed have been annotated to 20 122 human protein-coding genes; 9 243 publications were linked to the gene encoding p53; and 623 genes were without a publication. For each protein, we also annotated results from the online catalog of human genes and genetic disorders in Online Mendelian Inheritance in Man (OMIM) as a binary response of either yes (1) or no (0) and the published genome-wide association study (GWAS) frequency data. (b) The number of publications on an annual basis that are annotated to human protein-coding genes (blue symbols), and the cumulative number of publications (green symbols) since 1980.

typical of the others [7,8]. Consequently, the pattern of publications in the gene2pubmed data set appears to be indicative of an underlying scale-free network with high numbers for only a few genes. This pattern has been dubbed the Matthew effect, where cumulative interest begets further interest [6]. Consistent with this mathematical model, we see that 623 genes do not have any publication records.

When we compared this distribution with the online catalog of human genes and genetic disorders captured in Online Mendelian Inheritance in Man (OMIM) and the NHGRI-EBI of published genome-wide association studies (GWAS), we found that the publication frequency exhibits an exponential distribution regarding their subject proteins, with the most frequently cited gene in the database having more than 9 243 publications (Fig. 1a). The result of this tendency is that many of the genes linked to disease phenotypes or those associated with specific disease traits by GWAS studies remain severely understudied: the so-called 'dark genome'. Thus, characteristics of a scale-free network appear to be driving the underlying interest in a particular gene or protein as measured by the publications, rather than the genetic information. What can we do to promote the exploration of currently understudied, but potentially disease-relevant, proteins?

The human proteome

Over the past 15 years, 'omics and systems biology approaches, including proteomics, transcriptomics, clustered regularly interspaced short palindromic repeat (CRISPR) approaches, small interfering RNA, and short hairpin RNA screens, have provided important, unbiased experimental characterization of gene products on a global scale. Large-scale programs, such as The Human Protein Atlas, are mapping all human proteins in cells, tissues, and organs by integrating antibody-based imaging, mass spectrometry-based proteomics, and transcriptomics (www.proteinatlas.org/) [9,10]. Together with the Human Proteome Project (https://hupo.org/human-proteome-project) established by the Human Proteome Organization (HUPO) [11–15], these groups are cataloging experimentally verified human proteins in healthy and diseased cells and tissues.

All of these efforts have been supported by the web-based, protein knowledge-based platform neXtProt (www.nextprot.org) [16]. A recent analysis using neXtProt reported 17 470 proteins with high confidence of protein-level evidence (PE1; i.e. 89% of the human proteome) [13]. However, the functional annotation of many of these proteins remains vague or simply predicted based on similarity to homologous proteins. Interestingly, the neXtProt analysis indicated that 1 937 proteins are lacking specific functional annotation, of which 1 260 are uncharacterized PE1 proteins (uPE1) and 677 uncharacterized missing proteins [13]. Indeed, these 1 937 dark proteins, which have no function as predicted by experimental characterization or homology to other proteins, have recently become the focus of a new HUPO C-HPP pilot initiative to demonstrate the feasibility of characterizing 50 of such dark uEP1 proteins, termed neXt-CP50 [17].

Other approaches to studying the function of proteins include systematically mapping protein–protein interactions and/or computational approaches [18–22]. Overall, genome- and proteome-scale approaches provide a global view of cellular organization and often predict protein functional annotations that contribute to a better understanding of normal and disease biology. However, inherent in such large-scale studies are false positive and false negative results that require detailed experimental validation of specific proteins before we can reach firm conclusions. Thus, we believe that combining such computational methodologies for studying protein–protein interactions, together with the case for pharmacological modulators we propose here, provides a route to a better understanding of disease biology.

Counteracting scientific bias

There is little scientific or reasoned basis for disease-related genes not to attract research interest [23,24], yet nearly two decades after completion of the human genome project the research patterns have not changed much. The incentive systems drive researcher behavior bias towards well-studied areas, and it is clear that the explanation for this phenomenon is mostly sociological and/or practical [3,25–28]. Therefore, the solutions must modify behavior. Looking to the past for clues as to what influences the choices of researchers, we searched for the scientific contributions that have had the most impact in shifting research to new understudied proteins. We believe that the best way to encourage or enable scientists to study a new protein is to make the cognate research tools available. High-quality tools that would enable researchers to study a new protein include cell lines lacking the encoding gene and specific and selective antibodies to facilitate cell biology studies, purified protein and structural information to facilitate biochemistry, and pharmacological modulators to facilitate functional studies and evaluation of the protein as a therapeutic target.

The idea that tools drive biology is not new. To paraphrase Sydney Brenner, ' . . . progress in science [and drug discovery] depends on new techniques [and new tools], new discoveries and new ideas, probably in that order' [29]. Systematic efforts to generate some of these tools are underway. The National Institutes of Health (NIH) and The Wellcome Trust are funding two programs to generate 'target enabling packages' for understudied disease-linked genes and/or proteins. These data packages include protocols to generate purified proteins, assays, and reagent antibodies. What we need now is a systematic effort to generate pharmacological modulators.

The case for pharmacological modulators

Historically, the availability of pharmacological modulators has had tremendous impact on research on the target protein. For almost every protein for which a cell-active pharmacological modulator has been made available, the paper describing the modulator is among the top-cited papers on that protein in the entire literature, and the modulator also served as the trigger for new studies (https://arxiv.org/abs/1102.0448). It is arguably the most impactful reagent for any protein. Regrettably, pharmacological modulators of the required potency and selectivity to support interpretable and reproducible science are both challenging and expensive to invent, and also require skills commonly found in industry.

In 2009, we created an open science, international public–private partnership (PPP) to test the idea that academia and industry scientists could collaborate to generate high-quality pharmacological probes for understudied proteins; the project focused specifically on generating chemical probes (www.thesgc.org/chemical-probes). We adopted organizational strategies to coordinate the research activities, established quantifiable quality criteria for the chemical probes, and created an independent evaluation system to ensure their quality. The project has proven this model works. It has produced more than 50 high-quality chemical probes for proteins linked to epigenetic and kinase signaling, many of which had previously been relatively understudied. All of these probes have been used not only to generate significant insight into biology, but also as seeds for drug discovery programs [30].

The initiative has also encouraged seven large pharmaceutical companies to enter into a precompetitive collaboration to make a large number of their innovative high-quality probes from their previous projects available to the broader scientific community, including all probe-associated data, control compounds, and recommendations on use in an unencumbered manner (https://opencienceprobes.sgc-frankfurt.de/) [31]. From the experience over the past decade, we conclude that there is a viable organizational structure in which academia and industry can collaborate to generate high-value tools for the public domain.

Therefore, we feel that it is time to consider a PPP that aims to create selective, openly available pharmacological modulators for the entire human proteome. The task will be challenging; at well over US$2 million per high-quality chemical probe, it will be prohibitively expensive and, for some technically challenging proteins, perhaps not even possible. However, we are in the midst of many technological revolutions that, if aligned and coordinated, can make the project feasible. These include advances in protein production and protein-based screening technologies [32,33], proteome-wide screening methods for protein–ligand interactions [34], target engagement and selectivity [35], increases in chemical space coverage by combining empirical and computational screens [36,37], new advances in synthetic chemistry [38], and DNA-encoded libraries [39], to name just a few.

Target 2035

With this goal in mind, scientific representatives from academia, industry, and public funders met in Berlin in July 2018 to discuss how and whether it was now time to tackle this problem. The consensus was that the goal was potentially feasible given our progress to date, the ability of an open science structure to facilitate participation by funders, industry, and scientists across borders, and the pace of technological improvement. The participants also agreed that the availability of these tools would have unimaginable benefits for understanding biology and disease.

Meeting participants agreed on the aspirational goal of generating cell-active, potent, well-characterized chemical probes or functional antibodies for nearly all human proteins by 2035, and agreed that it would only be possible and cost-effective with further technology development, and within a mission-focused international program involving both public and private sectors. Furthermore, to avoid redundant research and to minimize complications arising from economic self-interests of regions, universities, and companies, it should adhere to the tenets of open science and be coordinated as a federated group of technology and science projects.

Participants also agreed that the path to 2035 should have two phases. The first phase, from 2020 to 2025, would not necessarily be organized top-down, but rather would provide a broad roadmap around which international groups of like-minded scientists could self-assemble and federate. The key aims of the federation would include the following: (i) collecting, characterizing, and distributing existing pharmacological modulators for key representatives from all protein families in the current druggable genome and generation of chemical probes for additional family members; (ii) developing the crucial and centralized infrastructure to facilitate data collection, curation, dissemination, and mining that will empower the scientific community worldwide; and (iii) creating centralized facilities to provide quantitative genome-scale biochemical and cell-based profiling assays to the federated community.

The collective would create a forum to coordinate the development and testing of new technologies to extend the definition of druggability; this might include organizing grand challenges to the community. The tangible goal of the first phase might be to develop potent, well-characterized, functional modulators for a targeted set of proteins by 2025.

As technologies and methodologies develop, at some point the effort would transition to a more formalized federation to avoid duplication of effort, consolidate core activities, and share advances more quickly. This second phase, potentially from 2025 to 2035, would apply the new technologies and infrastructure to generate a complete set of pharmacological modulators (Box 1) for most of the 20 122 human protein-coding genes. Alternative approaches might

BOX 1

Classes of protein modulators for Target 2035Chemical probe

Chemical probes are compounds that selectively modulate a single protein (or small group of closely related proteins) in cells at reasonable concentrations (≤1 μmol/l).

Biological probe

Biological probes are monoclonal antibodies and/or other macromolecular affinity reagents shown to be selective and able to modulate the activity of a protein target.

Chemogenomic libraries

Chemogenomics describes the use of target family-directed chemical libraries with known selectivity profiles and broad annotation in several different assay panels (and not excessively promiscuous) that modulate a subset of proteins in a cell. They are used to identify possible target protein family members.

Protein Handle

Protein handles are small molecules that selectively bind to a single protein or small set of proteins, but might not necessarily modulate the activity of the target protein on their own. Ligands or handles can be used to create chimeric compounds that would recruit a second protein to modulate the target.

also emerge. For example, the recent demonstration of chemically mediated protein–protein interactions, such as proteolysis-targeting chimeras (PROTACs), highlights the enormous possibilities to create both positive and negative modulators of proteins [40].

Target 2035 is incredibly ambitious, but its concept and practicality is on firm ground. Conceptually, it is not unlike the approach of the Human Genome Project [41] and other large-scale genomic consortia, such as the International HapMap Project [42] and the Single Nucleotide Polymorphism Consortium [43]: start with what can be immediately achieved; develop new technologies; and collaborate across borders and with the private sector to access complementary expertise, experience, materials, and technologies. It is essential to establish clear and quantitative quality criteria for the output (target chemical tool profiles) to provide focus (Box 1). Other key concepts were agreed. First, there was consensus to organize the project initially around protein families because it is the most efficient, practical, and scientifically sound way to divide this large project into teams. Second, it was decided to establish clear open science principles to eliminate or reduce conflicts of interest, reduce legal encumbrances [44], and encourage participation by all interested scientists in the community.

Next steps

Several relevant initiatives are already underway, including the Structural Genomics Consortium [31,45], the RESOLUTE project on the solute carrier protein family (https://re-solute.eu), NIH's Illuminating the Druggable Genome program [46], and the Innovative Medicines Initiative call 17 proposal, 'Open access chemogenomics library and chemical probes for the druggable genome' (www.imi.europa.eu/apply-funding/open-calls/imi2-call-17). In the coming 5 years, we suggest that these early initiatives align their efforts under Target 2035 and call for others to join. These and other like-minded groups should self-organize to develop and share technologies and experiences, both positive and negative, toward Target 2035 goals. We also encourage public and philanthropic funders of science and biomedicine to support unbiased approaches and technology development to identify chemical and biologic modulators of the human proteome. Finally, we call on industry to join in the effort not only from a technological perspective, but also from an organizational and funding perspective, because eventual exploitation of the druggable genome is also in their interest.

Acknowledgments

The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada through Ontario Genomics Institute [OGI-055], Innovative Medicines Initiative (EU/EFPIA) [ULTRA-DD grant no. 115766], Janssen, Merck KGaA, MSD, Novartis Pharma AG, Ontario Ministry of Research, Innovation and Science (MRIS), Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and The Wellcome Trust.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2019.06.020.

References

1 Nelson, M.R. *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860

2 Hoffmann, R. and Valencia, A. (2003) Life cycles of successful genes. *Trends Genet.* 19, 79–81

3 Pfeiffer, T. and Hoffmann, R. (2007) Temporal patterns of genes in scientific publications. *Proc. Natl. Acad. Sci. U. S. A.* 104, 12052–12056

4 Oprea, T.I. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* 17, 317–332

5 Maglott, D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 35 (Database issue), D26–D31

6 Perc, M. (2014) The Matthew effect in empirical data. *J. R. Soc. Interface* 11, 20140378

7 Barabási, A.L. and Bonabeau, E. (2003) Scale-free networks. *Sci. Am.* 288, 60–69

8 Barabási, A.L. (2009) Scale-free networks: a decade and beyond. *Science* 325, 412–413

9 Uhlén, M. *et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 4, 1920–1932

10 Uhlén, M. *et al.* (2015) Tissue-based map of the human proteome. *Science* 347, 1260419

11 HUPO (2010) A gene-centric human proteome project. *Mol. Cell. Proteomics* 9, 427–429

12 Legrain, P. *et al.* (2011) The human proteome project: current state and future direction. *Mol. Cell. Proteomics* 10, M111.009993

13 Omenn, G.S. *et al.* (2018) Progress on identifying and characterizing the human proteome: 2018 metrics from the HUPO Human Proteome Project. *J. Proteome Res.* 17, 4031–4041

14 Paik, Y.-K. *et al.* (2012) The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* 30, 221–223

15 Marko-Varga, G. *et al.* (2013) A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* 12, 1–5

16 Gaudet, P. *et al.* (2013) neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* 12, 293–298

17 Paik, Y.-K. *et al.* (2018) Launching the C-HPP neXt-CP50 pilot project for functional characterization of identified proteins with no known function. *J. Proteome Res.* 17, 4042–4050

18 Rual, J.-F. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178

19 Rolland, T. *et al.* (2014) A proteome-scale map of the human interactome network. *Cell* 159, 1212–1226

20 Huttlin, E.L. *et al.* (2015) The BioPlex Network: a systematic exploration of the human interactome. *Cell* 162, 425–440

21 Cafarelli, T.M. *et al.* (2017) Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Curr. Opin. Struct. Biol.* 44, 201–210

22 Zhang, C. *et al.* (2018) Structure and protein interaction-based gene ontology annotations reveal likely functions of uncharacterized proteins on human chromosome 17. *J. Proteome Res.* 17, 4186–4196

23 Stoeger, T. *et al.* (2018) Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.* 16 e2006643

24 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262

25 Oprea, T.I. *et al.* (2018) Far away from the lamppost. *PLoS Biol.* 16 e3000067

26 Edwards, A. (2011) Too many roads not taken. *Nature* 470, 163–165

27 Fedorov, O. *et al.* (2010) The (un)targeted cancer kinome. *Nat. Chem. Biol.* 6, 166–169

28 Akinjiyan, F.A. *et al.* (2017) Lead discovery and chemical biology approaches targeting the ubiquitin proteasome system. *Bioorg. Med. Chem. Lett.* 27, 4589–4596

29 Robertson, M. (1980) Biology in the 1980s, plus or minus a decade. *Nature* 285, 358–359

30 Arrowsmith, C.H. *et al.* (2015) The promise and peril of chemical probes. *Nat. Chem. Biol.* 11, 536–541

31 Müller, S. *et al.* (2018) Donated chemical probes for open science. *eLife* 7, e34311

32 Patel, D. *et al.* (2014) Advantages of crystallographic fragment screening: functional and mechanistic insights from a powerful platform for efficient drug discovery. *Prog. Biophys. Mol. Biol.* 116, 92–100

33 O'Connell, T.N. *et al.* (2014) Solution-based indirect affinity selection mass spectrometry—a general tool for high-throughput screening of pharmaceutical compound libraries. *Anal. Chem.* 86, 7413–7420

34 Roberts, A.M. *et al.* (2017) Activity-based protein profiling for mapping and pharmacologically interrogating proteome-wide ligandable hotspots. *Curr. Opin. Biotechnol.* 43, 25–33

35 Jensen, A.J. *et al.* (2015) CETSA: a target engagement assay with potential to transform drug discovery. *Future Med. Chem.* 7, 975–978

36 Barelier, S. *et al.* (2014) Increasing chemical space coverage by combining empirical and computational fragment screens. *ACS Chem. Biol.* 9, 1528–1535

37 Lyu, J. *et al.* (2019) Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229

38 Campos, K.R. *et al.* (2019) The importance of synthetic chemistry in the pharmaceutical industry. *Science* 363, eaat0805

39 Goodnow, R.A., Jr *et al.* (2016) DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* 16, 131–147

40 Deshaies, R.J. (2015) Protein degradation: prime time for PROTACs. *Nat. Chem. Biol.* 11, 634–635

41 Collins, F.S. *et al.* (2003) The Human Genome Project: lessons from large-scale biology. *Science* 300, 286–290

42 The International HapMap Consortium (2003) The International HapMap Project. *Nature* 426, 789–796

43 Sachidanandam, R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933

44 Edwards, A. *et al.* (2017) A trust approach for sharing research reagents. *Sci. Transl. Med.* 9, eaai9055

45 Scheer, S. *et al.* (2019) A chemical biology toolbox to study protein methyltransferases and epigenetic signaling. *Nat. Commun.* 10 s41467-018-07905-4

46 Rodgers, G. *et al.* (2018) Glimmers in illuminating the druggable genome. *Nat. Rev. Drug Discov.* 17, 301–302

Adrian J. Carter[1,*]
Oliver Kraemer[1]
Matthias Zwick[2]
Anke Mueller-Fahrnow[3]
Cheryl H. Arrowsmith[4,5]
Aled M. Edwards[4]

[1]Discovery Research Coordination, Boehringer Ingelheim, 55216 Ingelheim am Rhein, Germany
[2]Computational Biology, Boehringer Ingelheim, 88400 Biberach an der Riß, Germany
[3]Lead Discovery, Bayer AG, 13342 Berlin, Germany
[4]Structural Genomics Consortium, University of Toronto, Toronto, Ontario, M5G 1L7, Canada
[5]Princess Margaret Cancer Centre, Toronto, Ontario, M5G 1L7, Canada

*Corresponding author.