Editorial

For reprint orders, please contact reprints@future-science.com

Future Medicinal Chemistry

Minimizing bias in target selection by exploiting multidisciplinary Big Data and the protein interactome

"Here, we argue the case for minimizing bias in target selection by exploiting multidisciplinary Big Data and assessing targets based on their biological, chemical and physical properties, as well as their role in the cellular protein interactome."

First draft submitted: 26 June 2016; Article accepted: 28 July 2016; Published online: 1 September 2016

Keywords: Big Data • cancer • cancer networks • drug combinations • drug discovery • drug resistance • knowledgebase • target selection

Bissan Al-Lazikani
Author for correspondence:
Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, London, SM2 5NG, UK
Bissan.al-lazikani@icr.ac.uk

Paul Workman
Author for correspondence:
Cancer Research UK Cancer Therapeutics Unit, The Institute of Cancer Research, London, SM2 5NG, UK
Paul.workman@icr.ac.uk

Despite insights from large-scale genomic studies, there has been limited success in reversing the declining approval rate for new drugs, across all types of human diseases [1]. Even in oncology, where a significant number of personalized medicines that target the genetic drivers of cancers have been approved, the number of new drugs is dwarfed by the volume of driver genes identified [2].

Attrition rates in the clinic remain unacceptably high, with only 10% of drugs transitioning from Phase I to approval [3]. Some studies show that oncology has the lowest success rate at around 7% [3] and the highest rate of late-stage failures in the clinic [4]. The major cause of failure is lack of efficacy [4] which can often be attributed to inadequate patient stratification for the pivotal trial or, significantly earlier, poor preclinical target validation. Apocryphal tales and increasingly more detailed reports of a lack of reproducibility of biological data fill the scientific airways [5,6]. With the cost of developing a single drug estimated in excess of US$2.5 billion, it is not surprising that the Pharma industry has focused heavily on well-established targets and pathways – commonly sacrificing innovation for the sake of risk mitigation [7]. Indeed, our analysis shows that most approved cancer drugs target only a small part of cancer's intricate cellular networks [2,8].

Yet, our need for mechanistically innovative cancer drugs has never been greater. The genetic and epigenetic heterogeneity of cancer means that oncology's previous one-size-fits-all approach to treatment is no longer valid. Moreover, we are fighting a constant battle against adaptive biochemical and transcriptional responses, clonal evolution and the incessant emergence of drug resistance to previously successful targeted therapies – making new approaches essential [9–13].

Therefore as a community, we face clear and urgent challenges: How do we expand our repertoire of innovative, mechanistically distinct drugs against robustly validated targets for cancer? How do we overcome the inexorable march of drug resistance?

A key underlying solution is making the best possible choice of new, innovative drug targets. Here, we argue the case for minimizing bias in target selection by exploiting multidisciplinary Big Data and assessing targets based on their biological, chemical and physical properties, as well as their role in the cellular protein interactome.

Objective assessment of targets

To make the best decisions one must be armed with as much relevant knowledge as possible.

FUTURE SCIENCE part of fsg

Rather than assess the suitability of a target for drug discovery using only narrow biological criteria and/or a simple glance at 'druggability', we argue that one should examine all known data from different disciplinary domains in an objective fashion for a range of potential targets. This is particularly important when selecting targets from a large functional screen [14] or omic [15] analysis. A major challenge comes when these campaigns generate large lists of potential targets, about many of which very little is known. The natural tendency would be to select from these gene lists those that are more familiar to us or those that belong to one of the handful of well-studied and validated cancer pathways. Such human bias negates the value of the unbiased screen or omic analysis. Unbiased target discovery surely needs to be complemented by similarly unbiased, optimally evidence-based target comparison, prioritization and selection.

> "The genetic and epigenetic heterogeneity of cancer means that oncology's previous one-size-fits-all approach to treatment is no longer valid."

To address this, we developed and made publicly available an approach for data-driven, unbiased prioritization of cancer targets. We have brought together, within a single knowledgebase, vast multidisciplinary data on genomics (including integrated linkage to The Cancer Genome Atlas data [16]) and the biology, pharmacology and chemistry of genes and proteins, incorporating integrated linkage to the ChEMBL database [17] along with relevant clinical information and multiple alternative views of druggability. Using this approach we have demonstrated the ability of our integrated and unbiased large-scale approach to identify otherwise ignored but biologically important and druggable cancer targets [2,18–19].

canSAR: Integrating Big Data for drug discovery

To enable such objective, data-driven, unbiased and multidisciplinary therapeutic target assessment and selection, we reasoned that the community needs a user-friendly resource to integrate and crystallize all relevant data from the diverse domains. As well as presenting the data in an integrated and meaningful way, such a resource needs powerful data analytics to 'learn' from these diverse data so as to help scientists select targets and generate new hypotheses. Not only is the sheer scale of the Big Data involved a challenge, but so too are the multiple different 'languages' – we needed to integrate biological, structural, chemical and clinical data, and to allow information flow between these different disciplinary domains.

To achieve these goals we built canSAR [20–23] as a Big Data, multidisciplinary knowledgebase for cancer drug discovery (Figure 1). It brings together and integrates genomic, protein sequence and structural data, chemical, pharmacological, biochemical and phenotypic data, as well as clinical annotations and data from clinical trials. In all, canSAR contains 1 billion experimental data points. Data and annotations span the whole human proteome, 12,000 cancer cell lines, >10,000 cancer patient samples, >1.2 million biologically active small molecules and >200,000 clinical trials. canSAR has been utilized by >180,000 users from >180 countries (Google Analytics) and is accessed daily by around 400 researchers from major cancer research centers, universities and the pharmaceutical industry.

Most valuably, canSAR incorporates sophisticated data analytics and machine learning to facilitate the objective and rapid prioritization of targets for further validation. In addition to individual or systematic target selection, canSAR is particularly powerful for hypothesis generation. It enables users to easily identify model systems and chemical tools for experiments and alerts them to potential cancer associations based on patient genomics or data from cellular, genetic or drug screens. Recently, we have incorporated significant amounts of new data into canSAR in two major areas: 3D protein structures and cellular protein wiring [22].

3D protein structures inform target tractability

The 3D structure of proteins is a powerful tool that is invaluable to small-molecule drug discovery. It aids hit identification and structure-based drug design. Importantly for the present discussion, an assessment of the druggability of observed cavities observed on the 3D structure should be incorporated into multidisciplinary target selection. This ideally requires a sophisticated combination of the different kinds of physicochemical and geometric analysis of cavities based on many different physicochemical and geometric properties – beyond simply just volume and enclosure. Moreover, as proteins are mobile flexible molecules, examining only one or two snapshots for druggability can be misleading, as one may fail to detect a druggable cavity because it appears in only one specific protein conformation of the protein [2]. Also, apparently druggable cavities may be the result of experimental artifacts. Thus multiple structures are incorporated in canSAR wherever feasible. canSAR not only combines the different available approaches but importantly also utilizes machine learning to identify and assess the druggability of >110,000 structures comprising >310,000 individual protein chains. canSAR already evaluates >2,119,000 cavities of which >95,000 are predicted to be druggable. The availability of these data at a researcher's fingertips allows the rapid and

objective assessment of large numbers of known or potential protein targets – especially when integrated with the other means of target analyses available through canSAR. Particularly important are the published chemical landscape and pharmacology data for each target – allowing the rapid objective evaluation of large numbers of proteins. Moreover, because the methodology in canSAR assesses druggability based on 30 different physicochemical and geometric site properties, it can identify novel druggable regulatory cavities on proteins that would be missed by more simple assessments [2,22].

The social life of cancer proteins

Like most humans, cancer proteins do not act in isolation. Cancer results from complex, aberrant communication networks downstream of the initial oncogenic drivers and further deregulated by subsequent adaption, evolution and malignant progression. Moreover, drug resistance commonly arises from the rewiring of cellular communication, both within the cancer cell and between the cell and its environment [24]. Therefore it is surprising that cellular wiring and molecular communication are rarely used as criteria for assessing targets for drug discovery. To address this, we set out to establish whether there are communication behaviors that are unique to cancer drug targets when compared with the remainder of the proteome and targets from other disease areas [8].

We compiled and analyzed a high-confidence, curated network representing the experimentally defined human protein interactome [8] and integrated this into canSAR [22]. The resulting map contained 90,000 interactions between 13,345 proteins representing a complex interaction network. By applying network biology theory and social networking analysis – akin to a molecular Facebook – we derived a communication map for this network. Our analysis showed that cancer proteins have distinctive 'social' behaviors when compared with targets from other disease areas and to the human proteome as a whole [8]. Identification of these distinct behaviors has allowed us to develop novel predictors for target selection using cellular interactions [8] which are now provided through the latest version of canSAR [22].

Significantly, our analysis demonstrated that the majority of cancer drug targets occupy especially highly connected local environments. Although these proteins are important for communication, strong local connectivity allows rapid rewiring should they become disabled. This is consistent with the pathway remodeling detected after the emergence of drug resistance [10,24]. Moreover, it is consistent with resistance emerging when drugs acting on targets that are adjacent in a pathway are combined, such as BRAF and MEK inhibitors in the treatment of melanoma [25]. Interestingly, we find that the targets of immunotherapies – currently of major clinical and commercial interest – have different social environments and may therefore be less susceptible to compensatory pathway rewiring when compared with typical cancer drug targets [22].

An important prediction we make is that targeted therapies may be of most benefit when combined with drugs that act on targets from different 'social environments'. Our observations highlight the importance of considering these complex issues at the stage of target selection.

> "…the canSAR knowledgebase enables a smart, objective and user-friendly assessment of potential drug targets from a large pool of candidates…"

Comprehensive, continuous & live analysis of targets

In summary, the canSAR knowledgebase enables a smart, objective and user-friendly assessment of potential drug targets from a large pool of candidates – using integrative multidisciplinary Big Data to extend beyond the confines of well-trodden ground. The integration of diverse multidisciplinary information through canSAR enables hypothesis generation and facilitates the further experimental exploration of novel targets. Moreover, as the only public resource of its particular kind – incorporating sophisticated data analytics and machine learning – canSAR empowers the cancer drug discovery community to make objective decisions informed by multidisciplinary evidence. We argue that the vital step of selecting a drug target – from among the very many options available now in cancer – should only be taken after examining all that is known of a target's 3D structure, biology, chemistry and pharmacology as well as clinical data. Moreover, we propose that a target's communication pattern is crucial to its role in disease and behavior as a therapeutic target and canSAR now enables users to factor in such social network considerations. We suggest that this might be especially important in selecting targets to minimize potential for facile drug resistance which is a major challenge in treating cancer patients. Examining data from all these distinct domains together, in an unbiased manner, empowers the cancer drug discovery community, both public and private, to make more informed decisions. Most of the target selection tools and data in canSAR apply across human disease beyond oncology.

This is not automated target selection by computer, but rather provision of a powerful approach and tools to facilitate maximally informed target selection. Transla-

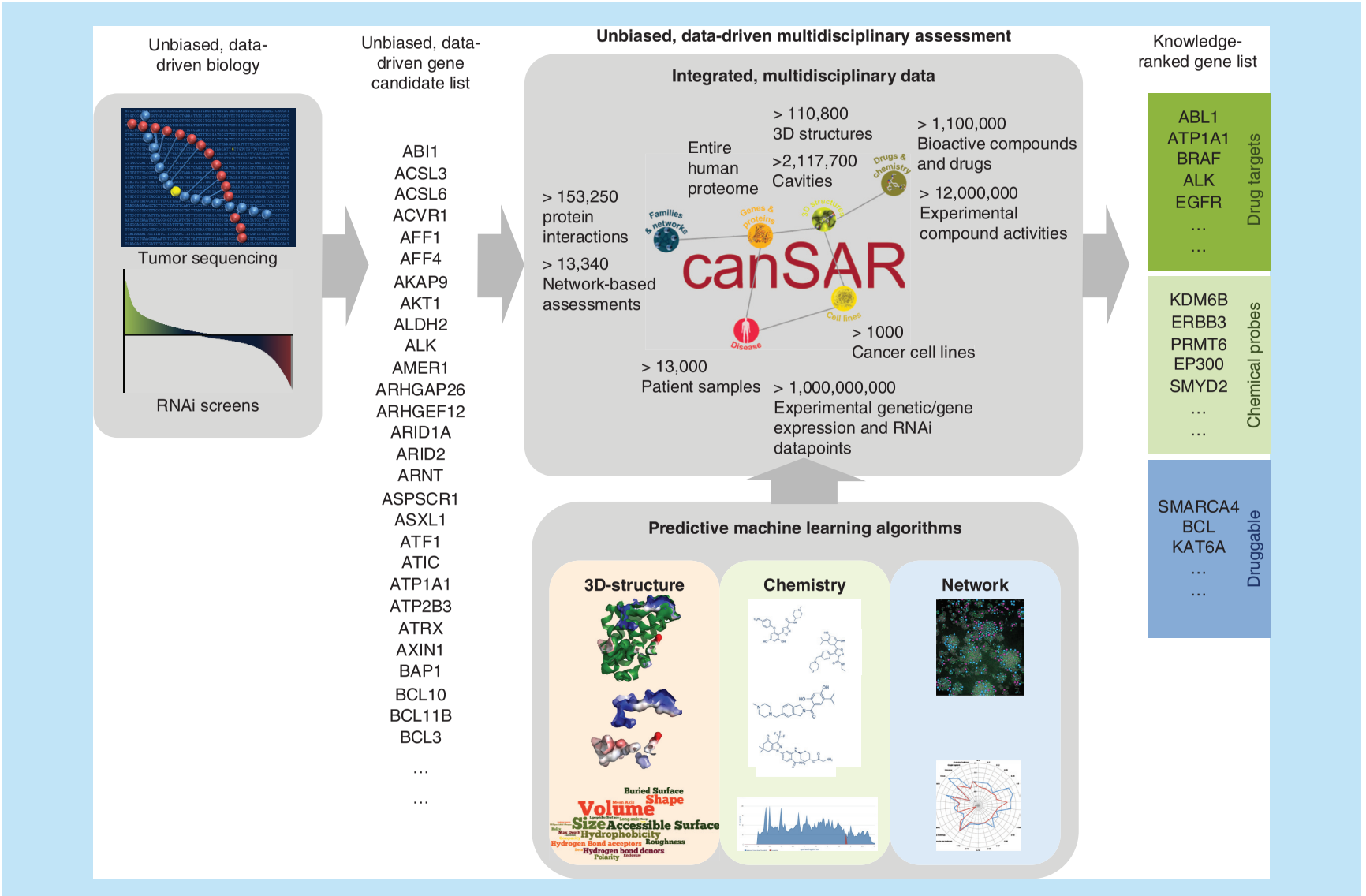

Figure 1. The canSAR knowledgebase is an integrated, multidisciplinary, Big Data resource and a suite of sophisticated machine-learning and assessment algorithms. Together, these allow automated annotation of large lists of genes with integrative drug discovery-relevant information from multiple domains (genomics, biology, structural, pharmacology, chemistry and clinical) that can be used for multidisciplinary objective target prioritization – incorporating druggability and protein network predictors as discussed in the text. For example, large gene lists identified through patient tumor sequencing, large-scale RNAi screens, or any other systematic profiling data can be uploaded into canSAR for target evaluation. The canSAR tools generate user-specified annotations, including multiple orthogonal druggability predictions. These analyses can be combined with the researcher's own data and experience for target assessment and prioritization.

tional scientists and drug discoverers are now empowered by easy access to all relevant information and integrative data mining tools – adding as they wish to their own data, experience and skills. With support from our funders (see the Acknowledgements section), and incorporating user feedback [26,27], we will enhance the intuitive user interface and the content and analytical power of canSAR, including links to the cBioPortal for Cancer Genomics [28] and the Open Targets platform [29]. The frequent automated updates available through canSAR enable users to regularly and rapidly refresh their view of the changing knowledge about our present and future drug targets. In this way we can constantly maintain a very current and broad as well as detailed comparative perspective of the rapidly changing challenges and opportunities in cancer drug discovery – and thus help improve success rates in clinical development and the defeat of drug resistance in cancer.

Acknowledgements

The authors thank N Evans, The Institute of Cancer Research for editorial assistance.

Financial & competing interests disclosure

The authors work at The Institute of Cancer Research, London, UK, which has a commercial interest in the discovery and development of anticancer drugs for different cancer types and has research and development interactions with multiple industry partners (http://www.icr.ac.uk/working-with-industry/about-the-enterprise-unit). Both the authors and The Institute of Cancer Research may benefit from this. The authors thank Cancer Research UK (grant number C309/A11566) and The Institute of Cancer Research for funding canSAR. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript apart from the editorial assistance mentioned above.

References

1 [No authors listed]. A decade in drug discovery. *Nat. Rev. Drug Discov.* 11(1), 3 (2012).

2 Patel MN, Halling-Brown MD, Tym JE, Workman P, Al-Lazikani B. Objective assessment of cancer genes for drug discovery. *Nat. Rev. Drug Discov.* 12(1), 35–50 (2013).

3 Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* 32(1), 40–51 (2014).

4 Arrowsmith J, Miller P. Trial watch: Phase II and Phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.* 12(8), 569 (2013).

5 Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10(9), 712 (2011).

6 Nature special: challenges in irreproducible research. www.nature.com/news/reproducibility-1.17552

7 Booth B. Cancer drug targets: the march of the lemmings. www.forbes.com

8 Mitsopoulos C, Schierz AC, Workman P, Al-Lazikani B. Distinctive behaviors of druggable proteins in cellular networks. *PLoS Comput. Biol.* 11(12), e1004597 (2015).

9 Greaves M, Maley CC. Clonal evolution in cancer. *Nature* 481(7381), 306–313 (2012).

10 Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* 30(7), 679–692 (2012).

11 Workman P, Clarke PA, Al-Lazikani B. Blocking the survival of the nastiest by HSP90 inhibition. *Oncotarget* 7(4), 3658–3661 (2016).

12 Schmitt MW, Loeb LA, Salk JJ. The influence of subclonal resistance mutations on targeted cancer therapy. *Nat. Rev. Clin. Oncol.* 13(6), 335–347 (2015).

13 Gonzalez de Castro D, Clarke PA, Al-Lazikani B, Workman P. Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. *Clin. Pharmacol. Ther.* 93(3), 252–259 (2013).

14 Cowley GS, Weir BA, Vazquez F *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* 1, 140035 (2014).

15 Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45(10), 1113–1120 (2013).

16 The cancer genome atlas. http://cancergenome.nih.gov/

17 Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107 (2012).

18 Workman P, Al-Lazikani B. Drugging cancer genomes. *Nat. Rev. Drug Discov.* 12(12), 889–890 (2013).

19 Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities within the DNA damage response. *Nat. Rev. Cancer* 15(3), 166–180 (2015).

20 Halling-Brown MD, Bulusu KC, Patel M, Tym JE, Al-Lazikani B. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res.* 40, D947–D956 (2012).

21 Bulusu KC, Tym JE, Coker EA, Schierz AC, Al-Lazikani B. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.* 42, D1040–D1047 (2014).

22 Tym JE, Mitsopoulos C, Coker EA *et al.* canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.* 44(D1), D938–D943 (2016).

23 canSAR. http://cansar.icr.ac.uk

24 Ramos P, Bentires-Alj M. Mechanism-based cancer therapy: resistance to therapy, therapy for resistance. *Oncogene* 34(28), 3617–3626 (2015).

25 Gowrishankar K, Snoyman S, Pupo GM, Becker TM, Kefford RF, Rizos H. Acquired resistance to BRAF inhibition can confer cross-resistance to combined BRAF/MEK inhibition. *J. Invest. Dermatol.* 132(7), 1850–1859 (2012).

26 Kotz J. Cancer target selection pressure. *SciBX: Science-Business eXchange* 6(6), doi:10.1038/scibx.2013.128 (2013) (Epub ahead of print).

27 Chau CH, O'Keefe BR, Figg WD. The canSAR data hub for drug discovery. *Lancet Oncol.* 17(3), 286 (2016).

28 Cerami E, Gao J, Dogrusoz U *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2(5), 401–404 (2012).

29 Open targets. www.targetvalidation.org