Bioinformatics, 35(21), 2019, 4509–4510
doi: 10.1093/bioinformatics/btz270
Advance Access Publication Date: 9 May 2019
Applications Note

OXFORD

Data and text mining

TractaViewer: a genome-wide tool for preliminary assessment of therapeutic target druggability

Neil Pearson[1,*], Karim Malki[1], David Evans[1], Lewis Vidler[1], Cara Ruble[2], James Scherschel[2], Brian Eastwood[1] and David A. Collier[1,*]

[1]Eli Lilly & Co. Ltd, Erl Wood Manor, Windlesham GU20 6PH, UK and [2]Eli Lilly & Co, Corporate Centre, Indianapolis, IN 46285, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 13, 2018; revised on March 26, 2019; editorial decision on April 7, 2019; accepted on April 7, 2019

Abstract

Summary: We present software to characterize and rank potential therapeutic (drug) targets with data from public databases and present it in a user-friendly format. By understanding potential obstacles to drug development through the gathering and understanding of this information, combined with robust approaches to target validation to generate therapeutic hypotheses, this approach may provide high quality targets, leading the process of drug development to become more efficient and cost-effective.

Availability and implementation: The information we gather on potential targets concerns small-molecule druggability (ligandability), suitability for large-molecule approaches (e.g. antibodies) or new modalities (e.g. antisense oligonucleotides, siRNA or PROTAC), feasibility (availability of resources such as assays and biological knowledge) and potential safety risks (adverse tissue-wise expression, deleterious phenotypes). This information can be termed ‘tractability’. We provide visualization tools to understand its components. TractaViewer is available from https://github.com/NeilPearson-Lilly/TractaViewer

Contact: pearson_neil@network.lilly.com or collier_david_andrew@lilly.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The wealth of information on disease aetiology from various omics provides an unprecedented opportunity to use data-driven approaches to nominate and rank potential drug targets for human disorders (Oprea *et al.*, 2018). Targets can be identified by genetic association with disease or disease-associated pathways, interactomes, networks or systems which identify points of intervention in pathological processes (Senger *et al.*, 2016). Multiple potential targets can be prioritized at the preclinical stage on the basis of target validation data, evidence for the proposed therapeutic hypothesis, molecular druggability, therapeutic modality and research feasibility, in order to prioritize the investment of critical resources (Oprea *et al.*, 2018). TractaViewer retrieves genome-wide information on tractability from public databases to inform decisions and identify gaps in our knowledge, in order to increase efficiency of drug discovery through greater probability of success (Paul *et al.*, 2010). Other resources currently exist for the assessment of drug targets, including Pharos (Nguyen *et al.*, 2016) and OpenTargets (Koscielny *et al.*, 2016). TractaViewer's automated mining is complimentary to these approaches, allowing users to draw conclusions about the state of current knowledge and visualize that information to reach a better understanding of target tractability.

2 Materials and methods

2.1 Target list import and cleanup

Input is supplied as a table or list of genes, identified by HGNC symbol, Ensembl ID and/or UniProt ID. Missing identifiers are automatically retrieved and disambiguated via queries to GeneNames (Gray *et al.*, 2014). Identifiers for homologues in model organisms are acquired from Ensembl's BioMart (Kinsella *et al*, 2011). Genes are also classified by their biotype (protein coding, RNA etc.) from UniProt.

2.2 Druggability

2.2.1 Small-molecule ligandability

We retrieve assessments of small-molecule druggability for protein targets, based on precedent of small-molecule activity (ChEMBL; Target Central Resource Database), known ligands, structure-based predictions, interactions with drugs from DGIdb (Griffith *et al.*, 2013), and a broad assessment of evidence (Pharos). Information from patents is not currently included. Targets with a default of $\geq$ 40% aligned protein sequence identity to any protein with precedent or structural indications of druggability are classed as potentially small-molecule druggable. Additionally, members of a druggable gene family may be scored as potentially ligandable. Existing approved drugs developed for a target (OpenTargets) may suggest repurposing opportunities or a mature target with limited potential for novelty. Targets with no precedent compounds and no evidence for druggable pockets may instead be classified as candidates for large molecule (antibody) or new modality approaches. The criteria for small-molecule druggability assessment are shown in Supplementary Figure S1.

2.2.2 Antibody targetability

Protein targets for therapeutic molecules (e.g. monoclonal antibodies) should be bioaccessible; therapeutic hypotheses calling for small-molecule drugs may also favour bioaccessible targets. We mine subcellular location data from the Human Protein Atlas (HPA) (Pontén *et al.*, 2008), membership in cell surface protein classes (e.g. GPCRs), secretome membership, cell surfaceome membership and extracellular matrix membership. Higher antibody targetability scores are assigned to targets belonging to any of these classes, with secreted proteins being preferred.

2.2.3 New modalities

Novel therapeutic options are indicated when the target has lower potential for small-molecule druggability or antibody targetability (e.g. is intracellular), or is a non-protein class (e.g. microRNA, long non-coding RNA). This includes instances where suppression of the protein or RNA species is desired by the therapeutic hypothesis, with options including antisense oligonucleotides (e.g. ASO, siRNA), antimirs or antagomirs, PROTAC or intrabody degradation, protein synthesis inhibition or gene editing approaches (e.g. CRISPR).

2.3 Potential risks

2.3.1 On-target toxicity

To assess potential on-target toxicity issues, we intersect withdrawn drugs (ChEMBL) with drug interaction data (DGIdb), producing a list of withdrawn drugs known to interact with the target.

2.3.2 Tissue expression localization

To mitigate on- and off-target toxicity, target expression restricted to the disease tissue or a targeted cell population is preferred; in addition, expression should be avoided in critical off-target tissues if possible—e.g. heart, kidney, liver or reproductive organs. Users may select target/off-target tissues fitting their therapeutic hypothesis; e.g. brain-expressed targets with low peripheral expression may be preferred for neurodegenerative diseases.

We assess expression at the tissue level to score potential off-target tissue safety hazards. The HPA classifies major tissues into 37 categories of tissue elevation. Target nominations are flagged as potential safety risks if they are tissue-elevated in categories not flagged as disease-relevant. We deprioritize targets expressed in tissues marked as conferring a higher risk of clinical toxicity, and give preference to tissue enhanced or tissue enriched (but not group enriched) targets in specified target tissues.

We also score targets for disease phenotype association (Human Phenotype Ontology, Köhler *et al.*, 2016) using toxicity-type classification from WITHDRAWN (Siramshetty *et al.*, 2015), from which we map toxicity categories of adverse effects associated with drug withdrawal to equivalent HPO terms. Genes tagged with these HPO terms are flagged for potential toxicity issues.

2.3.3 Cancer drivers and essential genes

Targets are checked against lists of known mutated cancer genes (Lawrence *et al.*, 2014) and mutational cancer driver genes (Tamborero *et al.*, 2013). We also flag essential genes (genes shown in CRISPR screens to be essential for survival in cell lines) (Chen *et al.*, 2017). Both associations indicate potential safety concerns for a target.

3 Results

Upon completion of data mining, data are displayed in a tabbed table, allowing users to rank and sort targets. Targets are classified in multiple dimensions, including small-molecule druggability, feasibility and safety. The decision criteria for these 'bucketing' processes are shown in help pages accessible within TractaViewer, and in Supplementary Figure S1. We provide a Shiny web app to facilitate a high-level overview of the acquired results.

4 Software availability

TractaViewer is available as source and as a precompiled binary for Windows (64 bit) at https://github.com/NeilPearson-Lilly/TractaViewer. At the time of writing, execution is supported on Windows platforms only; however, Linux support is in development.

Funding

This work was supported by Eli Lilly & Co. as part of the Accelerating Medicines Partnership-Alzheimer's Disease.

Conflict of Interest: none declared.

References

Chen,W.-H. *et al.* (2017) OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res.*, D940–D944.

Gray,K.A. *et al.* (2014) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, 1079–1085.

Griffith,M. *et al.* (2013) DGIdb: mining the druggable genome. *Nat. Methods*, **12**, 1209.

Kinsella,R.J. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011**.

Köhler,S. *et al.* (2016) The human phenotype ontology in 2017. *Nucleic Acids Res.*, **45**, 865–876.

Koscielny,G. *et al.* (2016) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*, **45**, 985–994.

Lawrence,M.S. *et al.* (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **7484**, 495.

Nguyen,D.-T. *et al.* (2016) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, 995–1002.

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

Paul,S.M. *et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.*, **9**, 203.

Pontén,F. *et al.* (2008) The Human Protein Atlas—a tool for pathology. *J. Pathol.*, **216**, 387–393.

Senger,M.R. *et al.* (2016) Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discov. Today*, **21**, 868–872.

Siramshetty,V.B. *et al.* (2015) WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.*, **44**, 1080–1086.

Tamborero,D. *et al.* (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**, 2650.