Check for updates

SOFTWARE TOOL ARTICLE

REVISED Assessing drug target suitability using TargetMine [version 2; peer review: 2 approved]

Yi-An Chen [1], Erika Yogo[2], Naoko Kurihara[2], Tomoshige Ohno[2], Chihiro Higuchi[1], Masatomo Rokushima[2], Kenji Mizuguchi [1]

[1]National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka, 5670085, Japan
[2]Shionogi Pharmaceutical Research Center, Shionogi & Co., Ltd., Toyonaka, Osaka, 5610825, Japan

v2 **First published:** 28 Feb 2019, **8**:233 (https://doi.org/10.12688/f1000research.18214.1)
Latest published: 28 May 2019, **8**:233 (https://doi.org/10.12688/f1000research.18214.2)

Abstract
In selecting drug target candidates for pharmaceutical research, the linkage to disease and the tractability of the target are two important factors that can ultimately determine the drug efficacy. Several existing resources can provide gene-disease associations, but determining whether such a list of genes are attractive drug targets often requires further information gathering and analysis. In addition, few resources provide the information required to evaluate the tractability of a target. To address these issues, we have updated TargetMine, a data warehouse for assisting target prioritization, by integrating new data sources for gene-disease associations and enhancing functionalities for target assessment. As a data mining platform that integrates a variety of data sources, including protein structures and chemical compounds, TargetMine now offers a powerful and flexible interface for constructing queries to check genetic evidence, tractability and other relevant features for the candidate genes. We demonstrate these features by using several specific examples.

Keywords
disease, drug assessment, genetic variation, tractability

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 28 May 2019		
version 1 published 28 Feb 2019	✓ report	✓ report

1 **Rachel Lyne**, University of Cambridge, Cambridge, UK

2 **Anne Hersey**, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Yi-An Chen (chenyian@nibiohn.go.jp), Kenji Mizuguchi (kenji@nibiohn.go.jp)

Author roles: Chen YA: Conceptualization, Data Curation, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Yogo E**: Conceptualization, Investigation, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kurihara N**: Data Curation, Investigation, Methodology; **Ohno T**: Data Curation, Investigation; **Higuchi C**: Data Curation; **Rokushima M** : Conceptualization, Investigation, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Mizuguchi K**: Conceptualization, Funding Acquisition, Investigation, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was in part supported by JSPS KAKENHI (grant number 17K07268).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Chen YA *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Chen YA, Yogo E, Kurihara N *et al.* **Assessing drug target suitability using TargetMine [version 2; peer review: 2 approved]** F1000Research 2019, **8**:233 (https://doi.org/10.12688/f1000research.18214.2)

First published: 28 Feb 2019, **8**:233 (https://doi.org/10.12688/f1000research.18214.1)

REVISED **Amendments from Version 1**

To respond to the reviewers' comments, in this revision, we enhance the explanation of the use cases. We also incorporate an application for the pathway analysis in the use case section. In addition, we share the details of our analysis as extended data which can be found at OSF.

Figure 5 has been updated, and a new Figure 6 shows an example of pathway enrichment analysis.

See referee reports

Introduction

A drug discovery project typically begins with the identification of a target molecule. In evaluating potential drug targets, several factors must be taken into account: linkage to disease, tractability (the possibility of finding small molecule compounds with high affinity), potential side effects, novelty, as well as the competitiveness in the market (Figure 1). Among these factors, the linkage to disease and the tractability are particularly important in terms of the drug efficacy, and become key factors in whether or not the pharmaceutical research and development (R&D) is successful when selecting drug targets[1,2]. The most important part of the linkage to disease is genetic associations for the disease or relevant traits. According to analyses reported by AstraZeneca and GlaxoSmithKline, the success rate of such R&D is increased when the choice of the selected target is supported by genetic evidence. The report from AstraZeneca shows that 73% of projects with some genetic linkage of the target to the disease indication in Phase II were active or successful compared to 43% of projects without such data[3], while the analysis results from GlaxoSmithKline suggest that selecting genetically supported targets could double the success rate in clinical development[4]. Several existing resources provide information about genetic evidence, such as DisGeNET[5], Open Targets[6], and Pharos[7]. However, a simple list of genes with genetic linkage to the disease is often insufficient for evaluating the disease rationale fully, and additional information and analysis such as pathway enrichment analysis will be needed to assess other aspects of target suitability (e.g. drug mechanisms and safety). In addition, few resources provide tractability information, with the recent update of Open Targets being an exception.

To address these issues, we have updated TargetMine[8], a data warehouse for assisting target prioritization, and improved its functionalities for target assessment, particularly in small molecule drug discovery. TargetMine[8] utilizes the InterMine framework[9] and facilitates flexible query construction spanning a wide range of integrated data sources including those relevant for evaluating linkage to disease and tractability. More specifically, we have integrated new data sources for genetic disease associations including ClinVar, dbSNP, and 1000 Genome Project, incorporated more details of the genome wide association studies from the GWAS catalog, and improved the data model overall to enable more efficient data mining. The new version provides a user-friendly and yet powerful interface to explore the disease rationale for human genes and helps prioritize the candidate genes in terms of both the genetic evidence and target tractability. In addition, with the assistance of InterMine APIs, repeated analysis and queries can be processed efficiently.

Methods

Implementation

TargetMine[8] is based on the InterMine framework, an open-source data warehouse system designed for biological data integration[9]. In this update, we added a few customized data sources by defining new data models and implementing new data parsers. Details of how we designed the data models are described in the following sub-sections.

GWAS catalog

The GWAS catalog, founded by NHGRI, is a curated archive of the published genome wide association studies[10]. We had tried to associate genes to related diseases using the GWAS catalog in the former release of TargetMine[11]. To annotate disease terms to a trait or study, we first chose the disease ontology

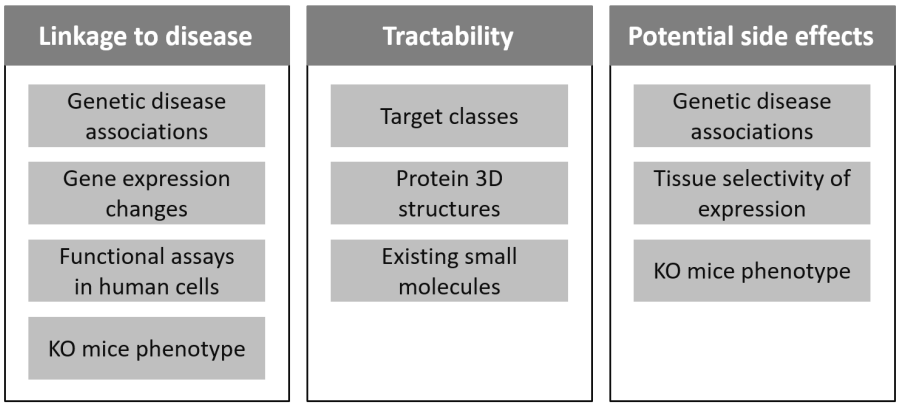


Figure 1. Key factors to be considered in drug target selection. Linkage to disease, tractability and adverse event risk are among the major factors to assess the suitability of novel target candidates. Much of the evidence regarding these factors is available in public domain resources.

(DO)[12,13] and then manually assigned the terms with the assistance of some text matching approaches. However, this process required some knowledge and involved a lot of manual examinations. Thus, it became difficult to keep updating regularly. Fortunately, the curation team started to use experiment factor ontology (EFO)[14] to describe the curated GWAS traits in the recent implementation[15]. EFO covers several domain-specific ontologies that facilitate easier data integration. In our new implemented model, we replace DO terms with EFO terms and also incorporate some more information from each study (Figure 2). SNP annotations and details of EFO terms are retrieved from the dbSNP database and EFO, respectively.

ClinVar

ClinVar is a public archive of the relation between human variations and phenotypes[16,17]. As defined by ClinVar, a "Variation" could be a single variant, a compound heterozygote, or a complex haplotype. If a haplotype consists of multiple alleles, each allele is assigned with an independent identifier. On the other hand, the same allele could be the member of a different

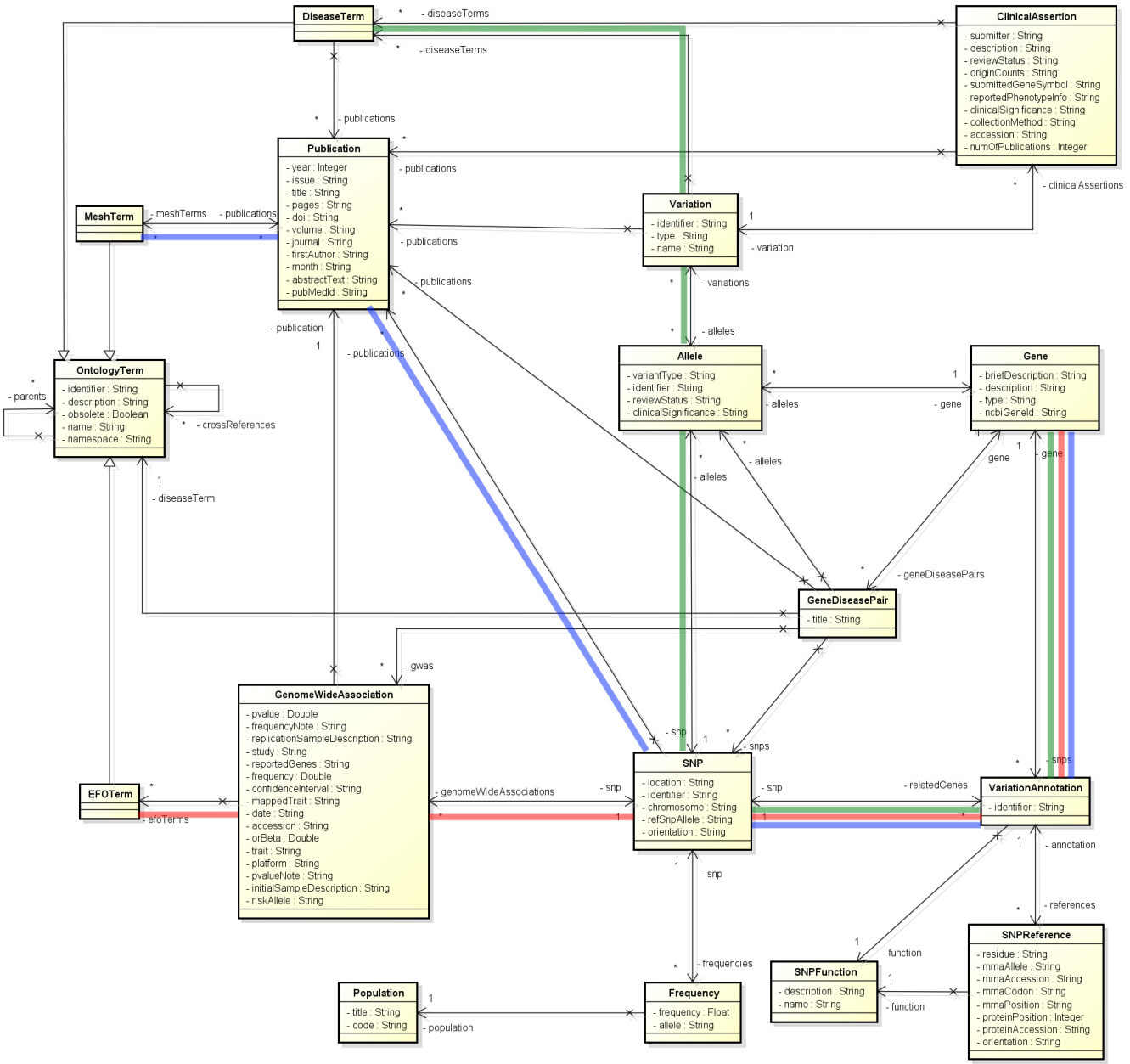


Figure 2. The new implemented data model. The colored lines indicate how the genes and diseases/phenotypes are associated in the post processing step.

haplotype, thus the relation between the "Variation" and "Allele" is a many-to-many association. An "Allele" is supposed to describe a specific change of a variation, e.g. G>A. However, the SNP entries in dbSNP sometimes merge different combinations of variations (alleles) together if the variations occur at the same genomic position. Thus, an "SNP" entity may contain multiple "Allele" entries in the data model (Figure 2). Here, we only retrieve the SNP identifier, and the rest of the annotations are integrated from the dbSNP database. The structural variations which reference the dbVar records are not included in the current version. In addition, those alleles which were not assigned with any dbSNP or dbVar identifiers were treated as SNP entities and were stored in TargetMine[8] using the information provided by ClinVar. Most of the data were processed from tab delimited files, while some information that were not available in the tab delimited files were processed from XML files. MedGen terms, which are used to integrate the human medical genetic information at NCBI (https://www.ncbi.nlm.nih.gov/medgen/), were adopted to describe diseases and phenotypes.

dbSNP

dbSNP is a database which archives short human genetic variations. We first performed a whole data dump to a relational database, and then made queries to extract the necessary information into a flat table. These data include genomic position (based on genome assembly GRCh38), reference mRNA, nucleotide variation, reference protein, and amino acid variation, if available. SNP to gene is a many-to-many relationship, thus we introduce an intermediate class named "VariationAnnotation" to associate them together (Figure 2). Although the InterMine framework is capable of incorporating whole SNP entries in dbSNP, the integration takes a few days to finish. Considering the frequency that we update TargetMine[8] (once a month), it is not very practical to spend a few days doing the integration. As a tradeoff, we decided to store only a subset of SNPs. Only those SNPs which are related with GWAS associations or clinical assertions, or those where there is an associated publication, are selected for storage in TargetMine[8].

Frequency data

Population specific genetic variation frequency is important for evaluating drug efficacy. We preprocessed the frequency data from several data sources, including the Human Genetic Variation Database (HGVD)[18], the integrative Japanese Genome Variation Database (1KJPN)[19] (download from the archive in National Bioscience Database Center), the Exome Variant Server (EVS)[20], and the 1000 Genomes Project (1KGP)[21,22]. At the moment, we only incorporate the population specific frequency for those SNPs stored in TargetMine[8].

Post-processing the integrated data

Our implementation allows us to associate the genetic phenotype (disease) and the gene via the GWAS or ClinVar dataset, or moreover the relation that is implied from the disease related MeSH (Medical Subject Headings, https://www.ncbi.nlm.nih.gov/mesh) terms assigned to the correlated publications of the SNPs. In order to make a shortcut and to summarize the available information, we perform post-processing and store the results using a new class named "GeneDiseasePair". At the moment, there are three types of shortcuts. Gene to SNP to GWAS to EFO terms for GWAS catalog data (the red lines in Figure 2). Gene to SNP to clinical assertions to disease (MedGen) terms (the green lines in Figure 2). And Gene to SNP to publication to MeSH terms (the blue lines in Figure 2). The "GeneDiseasePair" class also includes correlated information including ontology terms, studies, SNPs and publications. These improvements in the data model facilitate quick access from a gene to the associated diseases, annotated by different data sources.

Operation

TargetMine[8] is a Java-based web application that runs on Apache Tomcat. The user interface communicates with the integrated data stored in PostgreSQL, a relational database.

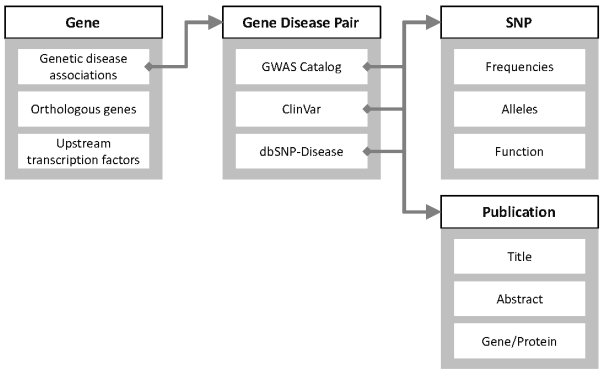
Use cases

Querying linkage to disease with TargetMine

To demonstrate the effectiveness of the new version of TargetMine[8] in evaluating linkage to disease, we conducted a feasibility study, taking human PCSK9, proprotein convertase subtilisin/kexin type 9, as a typical case. The *PCSK9* gene encodes a protein that promotes degradation of low-density lipoprotein (LDL) receptors in hepatocytes, thereby elevating or maintaining LDL cholesterol levels in the blood. Mutations in this gene are shown to be associated with familial hypercholesterolemia[23], and monoclonal antibodies to PCSK9 have been launched on the market as drugs for hypercholesterolemia with and without genetic predispositions[24,25].

Figure 3A demonstrates a schematic representation of the searching protocol for genetic disease associations with TargetMine[8]. We first went to a gene report page by searching for the *PCSK9* gene from the top page of TargetMine[8] (not shown). From the gene report page, we got information of genetic disease associations (Figure 3B) as well as many other basic or advanced characteristics such as orthologous genes and upstream transcription factors. The results table of genetic disease associations for *PCSK9* enabled us to confirm that a number of SNPs relevant to this gene have been reported to be associated with plasma LDL cholesterol levels, hypercholesterolemia, or coronary artery disease. By clicking the record of association between "low density lipoprotein cholesterol measurement" and *PCSK9* in the GWAS catalog section (Figure 3B), we moved to a "gene disease pair" page and checked the details of the GWAS record, including the information on samples, statistical significance and publications (Figure 3C). Clicking on the SNP identifier (e.g., rs2479409) redirected us to an SNP report page containing the individual SNP basic information (allele, function, literature) and allele frequencies of different human populations (from 1000 Genome Project[26] and others, not shown in the figure). Similarly, we examined the associations between "Hypercholesterolemia, autosomal dominant, 3" and *PCSK9* from the ClinVar section in the table (Figure 3B) and got the details of the ClinVar record such as clinical assertions and publications (Figure 3D). The publications here reported mutations in *PCSK9* as a cause of autosomal dominant hypercholesterolemia[23] (not shown), as mentioned above.

(A)



(B) Genetic disease associations

61 Genetic disease associations

GWAS catalog	p-value	Number of publications	Number of SNPs
ischemic cardiomyopathy - PCSK9	3.0E-10	1	1
LDL cholesterol change measurement - PCSK9	5.0E-9	1	1
angina pectoris - PCSK9	3.0E-10	1	1
percutaneous transluminal coronary angioplasty - PCSK9	3.0E-10	1	1
total cholesterol measurement - PCSK9	4.0E-24, 2.0E-39, 2.0E-17, 4.0E-6, 1.0E-32, 3.0E-8, 1.0E-23	5	4
coronary artery bypass - PCSK9	3.0E-10	1	1
blood metabolite measurement - PCSK9	3.0E-61	1	1
low density lipoprotein cholesterol measurement - PCSK9	2.0E-28, 3.0E-50, 2.0E-44, 2.0E-92, 8.0E-7, 3.0E-42, 9.0E-9, 4.0E-20	6	4
coronary artery disease - PCSK9	3.0E-10, 2.0E-22, 2.0E-25, 2.0E-25	2	1
lipoprotein measurement - PCSK9	3.0E-61	1	1
osteoarthritis, knee - PCSK9	9.0E-6	1	1
myocardial infarction - PCSK9	3.0E-10	1	1
ClinVar	Clinical significant		
Hypercholesterolemia, autosomal dominant, 3 - PCSK9	Benign, Benign/Likely benign, Conflicting interpretations of pathogenicity, Conflicting interpretations of pathogenicity, association, Likely benign, Pathogenic, Pathogenic/Likely pathogenic, Uncertain significance	54	91
Hypocholesterolemia - PCSK9	Benign, Conflicting interpretations of pathogenicity, association, Pathogenic	8	4
Familial hypercholesterolemia - PCSK9	Benign, Benign/Likely benign, Conflicting interpretations of pathogenicity, Conflicting interpretations of pathogenicity, association, Likely benign, Likely pathogenic, Pathogenic, Pathogenic/Likely pathogenic, Uncertain significance	54	201
Familial hypobetalipoproteinemia - PCSK9	Benign/Likely benign, Conflicting interpretations of pathogenicity, Likely benign, Uncertain significance	17	73
Low density lipoprotein cholesterol level quantitative trait locus 1 - PCSK9	Benign, Conflicting interpretations of pathogenicity, association, Pathogenic, association	11	4
Familial hypercholesterolemias - PCSK9	Pathogenic/Likely pathogenic	5	1
dbSNP-PubMed-MeSH			
Coronary Artery Disease - PCSK9		9	8
Hypercholesterolemia - PCSK9		12	14
Hyperlipoproteinemia Type II - PCSK9		16	21

(C) GWAS Catalog (8/8 records)

8 GWAS

Snp . Identifier	Risk Allele	Frequency	Pvalue	Or Beta	Confidence Interval	Trait	Initial Sample Description	Replication Sample Description	Study	Accession	Publication . Pub Med Id
rs2479409	rs2479409-G	0.3	2.0E-28	2.01	[1.58-2.44] mg/dL increase	LDL cholesterol	95,454 European ancestry individuals	NA	Biological, clinical and population relevance of 95 loci for blood lipids.	GCST000759	20686565
rs2479409	rs2479409-G	0.32	3.0E-50	0.064	[NR] unit increase	LDL cholesterol	94,595 European ancestry individuals	93,982 European ancestry individuals	Discovery and refinement of loci associated with lipid levels.	GCST002222	24097068
rs11591147	rs11591147-T	0.01	2.0E-44	0.47	[0.41-0.53] unit decrease	LDL cholesterol	2,758 European ancestry individuals	18,544 European ancestry individuals	Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.	GCST000134	18193044
rs11591147	rs11591147-G	0.98	2.0E-92	0.528	[0.46-0.58] s.d increase	LDL cholesterol	up 62,166 European ancestry individuals	NA	The impact of low-frequency and rare variants on lipid levels.	GCST002898	25961943
rs505151	rs505151-A	0.94	8.0E-7	0.11	[0.067-0.153] unit decrease	LDL cholesterol levels	22,526 East Asian ancestry individuals	37,842 East Asian ancestry individuals	Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels.	GCST004236	28334899
rs2479409	rs2479409-A	0.667	3.0E-42	0.0642	[0.056-0.072] unit decrease (EA Beta value)	LDL cholesterol levels	32,285 East Asian ancestry individuals, 173,082 European ancestry individuals	8,478 Chinese ancestry individuals	Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels.	GCST004233	28334899
rs505151	rs505151-A	0.94	9.0E-9	0.107	[0.072-0.142] unit decrease	LDL cholesterol levels	32,285 East Asian ancestry individuals, 173,082 European ancestry individuals	8,478 Chinese ancestry individuals	Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels.	GCST004233	28334899
rs12136600	rs12136600-?		4.0E-20	0.07746	[0.061-0.094] unit decrease novel	Low density lipoprotein cholesterol	72,866 Japanese ancestry individuals	NA	Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases.	GCST006004	29403010

(D) ClinVar (13/87 records)

SNPs Snp . Identifier	Function Name	Disease Terms Name	Alleles Clinical Significance	Publications PubMed ID
rs137852912	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	10205269
rs137852912	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	10764678
rs28942111	intron	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	12730697
rs28942112	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic	12730697
rs137852912	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	14727179
rs28942111	intron	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	15166014
rs137852912	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	15772090
rs137852912	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	18250299
rs28942111	intron	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	18250299
rs28942112	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Pathogenic	18250299
rs1057519691	intron	Hypercholesterolemia, autosomal dominant, 3	Pathogenic/Likely pathogenic	25741868
rs1057519692	intron	Hypercholesterolemia, autosomal dominant, 3	Uncertain significance	25741868
rs143275858	ncRNA	Hypercholesterolemia, autosomal dominant, 3	Conflicting interpretations of pathogenicity	25741868

Figure 3. Searching information about linkage to disease with TargetMine. (**A**) Outline of the procedure for searching. (**B**) A screenshot of the summary of Genetic disease associations of PCSK9. (**C**) GWAS records of a pair of PCSK9 and low density lipoprotein cholesterol measurement. (**D**) ClinVar records of a pair of PCSK9 and hypercholesterolemia, autosomal dominant, 3.

Querying target tractability for small molecule drugs with TargetMine

We performed another feasibility study to examine whether TargetMine[8] provides informative evidence to assess target tractability for small molecules. In this case we also used PCSK9 as an example because no potent small molecule inhibitors for this protein have been reported so far in spite of the intensive research activities of many laboratories[27], indicating that PCSK9 is not a highly tractable target.

Figure 4A shows a schematic diagram of the procedure of querying tractability with TargetMine[8]. We first went to the protein report page of PCSK9 and found the bioactive compounds targeting this protein. As we expected, it was revealed that no potent compounds could be found in the ChEMBL database, and the lowest IC50 value was 440 nM (CHEMBL3923422) (Figure 4B). On the PCSK9 protein report page, we also checked the experimentally determined 3D structures, referred to as "protein structure regions" in TargetMine[8], and identified several Protein Data Bank (PDB) entries for this protein (Figure 4C). Then, we moved to the "Protein Structure" page of a specified PDB ID (2p4e in this case) and found that in the "DrugEBIlity" table (from the DrugEBIlity database), some domains of the PCSK9 protein had positive ensemble scores (Figure 4D), which are not ligand-based, but structure-based tractability scores. This result indicates that PCSK9 protein may contain some sites/pockets that can bind small molecules, although ensemble scores of DrugEBIlity may need to be further validated.

Collectively, we were able to confirm that the new version of TargetMine[8] can quickly provide lines of evidence to assess linkage to disease and target tractability of PCSK9, and that the gathered data correctly reflected the real world situation; namely, it has been a challenge to obtain potent small molecule inhibitors for PCSK9, whereas antibody drugs for this protein have been successfully developed and marketed recently.

Gathering and prioritizing candidate drug target genes

To assess the utility of the new update of TargetMine[8] for prioritizing candidate targets, we conducted a case study where we employed a list of genes associated with hypercholesterolemia in the literature. We tentatively defined three key properties of a novel drug target suitable for small molecules as follows: (1) being associated with hypercholesterolemia via SNPs (GWAS catalog, ClinVar, or dbSNP-Pubmed; see below), (2) having greater than or equal to 50% of protein 3D structures with positive ensemble scores (DrugEBIlity), and (3) having no reported (ChEMBL) potent small molecule inhibitors (IC50 or EC50 ≤ 100 nM), for selecting relatively novel targets. It should be noted that the third criterion here depends on the prioritization purpose; if we attempted to adopt a so-called "me-too" approach, we should select target candidates with potent small molecule inhibitors instead.

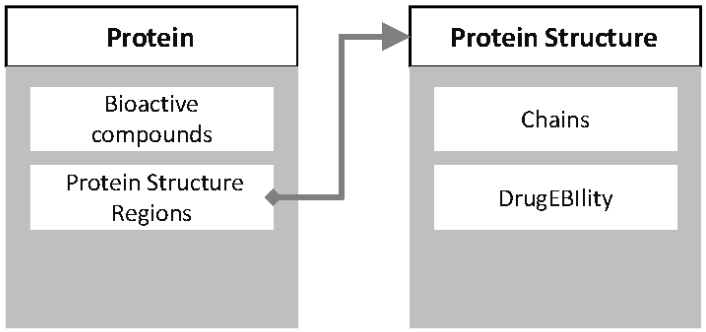
We first searched PubMed using the term "hypercholesterolemia" (from 2017/1/1 to 2018/9/10) and curated the obtained hits with the "Pubtator" text-mining tool[28], resulting in 510 human genes (Figure 5A). We then selected the genes meeting the requirements defined above using the "Query Builder" in TargetMine[8]. Figure 5B shows an example of actual query, which aimed to extract the genes with genetic evidence obtained from the GWAS catalog, where "Mapped Trait" contained "LDL cholesterol", "total cholesterol", or "low density lipoprotein cholesterol". Similarly, genes with genetic evidence obtained from ClinVar and dbSNP-Pubmed, with potent small molecules, and those predicted to be tractable in DrugEBIlity database were also extracted using the query builder. (The relevant data are shared as the supplementary materials on the TargetMine website; see "Data Availability" for the URL.) Thus, the new implementation enabled us to filter objects on complex conditions with a user-friendly, intuitive graphical interface.

Genes that satisfied all three requisites above are presented in Figure 5C (*CYP7A1*, *FABP2*, *LDLR*, *MYLIP*, *PCSK9*, *SREBF2* and *STAP1*). Among the seven genes we found *MYLIP* and *STAP1*. MYLIP is an E3-ubiquitin ligase that degrades LDL receptors in the liver, which are therefore considered to be a potential therapeutic target for dyslipidemia[29]. Similarly, the *STAP1* gene has been recently annotated as a fourth locus associated with autosomal-dominant hypercholesterolemia, and might be a novel target for therapeutic development of hypercholesterolemia[30]. This result suggests that the new version of TargetMine[8] allows us to effectively prioritize target candidate genes in terms of linkage to disease, tractability and competitors. On the other hand, the list includes intractable targets such as PCSK9 and LDLR, indicating the need for improvement of the data and/or the thresholds with which tractable proteins are selected (in this study, ≥50% of protein 3D structures have positive ensemble scores in DrugEBIlity database).

Evaluating candidate drug target genes in the context of pathways

TargetMine also allows us to seamlessly conduct pathway analysis, which identifies biological processes or pathways that are statistically enriched in a list of genes. From the perspective of prioritizing target candidates, pathway analysis is useful for highlighting the genes on pathways targeted by existing drugs or pathways with concerns for adverse events when blocked. Figure 6 shows the result of pathway enrichment analysis of the 510 human genes, which were obtained by PubMed search in Figure 5A using the term "hypercholesterolemia". As expected, the most highly enriched pathway is "Cholesterol metabolism", followed by "Lipoprotein metabolism". It is also readily recognizable that PCSK9 and MYLIP (mentioned above), both of which are involved in the degradation of LDL receptors, are on the same pathway (i.e., cholesterol metabolism) (Figure 6B). This observation may suggest that discovering MYLIP inhibitors should not be so attractive because antibody drugs for PCSK9 have already been marketed, although small molecules against MYLIP may still have some advantages, such as lower cost and better compliance (oral availability), against anti-PCSK9 antibodies.

(A)



(B) Bioactive compounds

Protein(s) --> Compounds with bioactivities (64 rows)

Showing 1 to 25 of 64 rows

Rows per page: 25 page 1

Protein DB identifier	Protein Primary Accession	Compounds Compound . Identifier	Activities Type	Activities Concentration(nM)
PCSK9_HUMAN	Q8NBP7	ChEMBL:CHEMBL3923422	IC50	440
PCSK9_HUMAN	Q8NBP7	ChEMBL:CHEMBL3952343	IC50	520
PCSK9_HUMAN	Q8NBP7	ChEMBL:CHEMBL3906851	IC50	592
PCSK9_HUMAN	Q8NBP7	ChEMBL:CHEMBL3915651	IC50	640
PCSK9_HUMAN	Q8NBP7	ChEMBL:CHEMBL3946218	IC50	690

(C) Protein Structure Regions

55 Protein Structure Regions

Protein . Primary Accession	Start	End	PDB Sequence Start	PDB Sequence End	PDB ID	Chain
Q8NBP7	1	692	None	None	2p4e	A
Q8NBP7	1	692	None	None	2p4e	P
Q8NBP7	31	152	None	152	2pmw	A
Q8NBP7	153	692	153	None	2pmw	B
Q8NBP7	29	152	None	152	2qtw	A
Q8NBP7	153	692	153	None	2qtw	B
Q8NBP7	153	451	153	None	2w2m	A
Q8NBP7	53	152	None	152	2w2m	P
Q8NBP7	153	451	153	None	2w2n	A
Q8NBP7	53	152	None	152	2w2n	P

(D) DrugEBIlity

3 DrugEBIlities

Primary Identifier	Ensembl	Tractable	Druggable	Protein Structure . Pdb Id
3170434	-0.31	true	false	2p4e
3170433	0.79	true	true	2p4e
4159286	0.7	true	true	2p4e

Figure 4. Searching information about target tractability for small molecule drug with TargetMine. (**A**) Outline of the procedure for searching. (**B**) Protein structure regions and their ensemble scores calculated by DrugEBIlity. (**C**) Compounds with bioactivity for PCSK9.

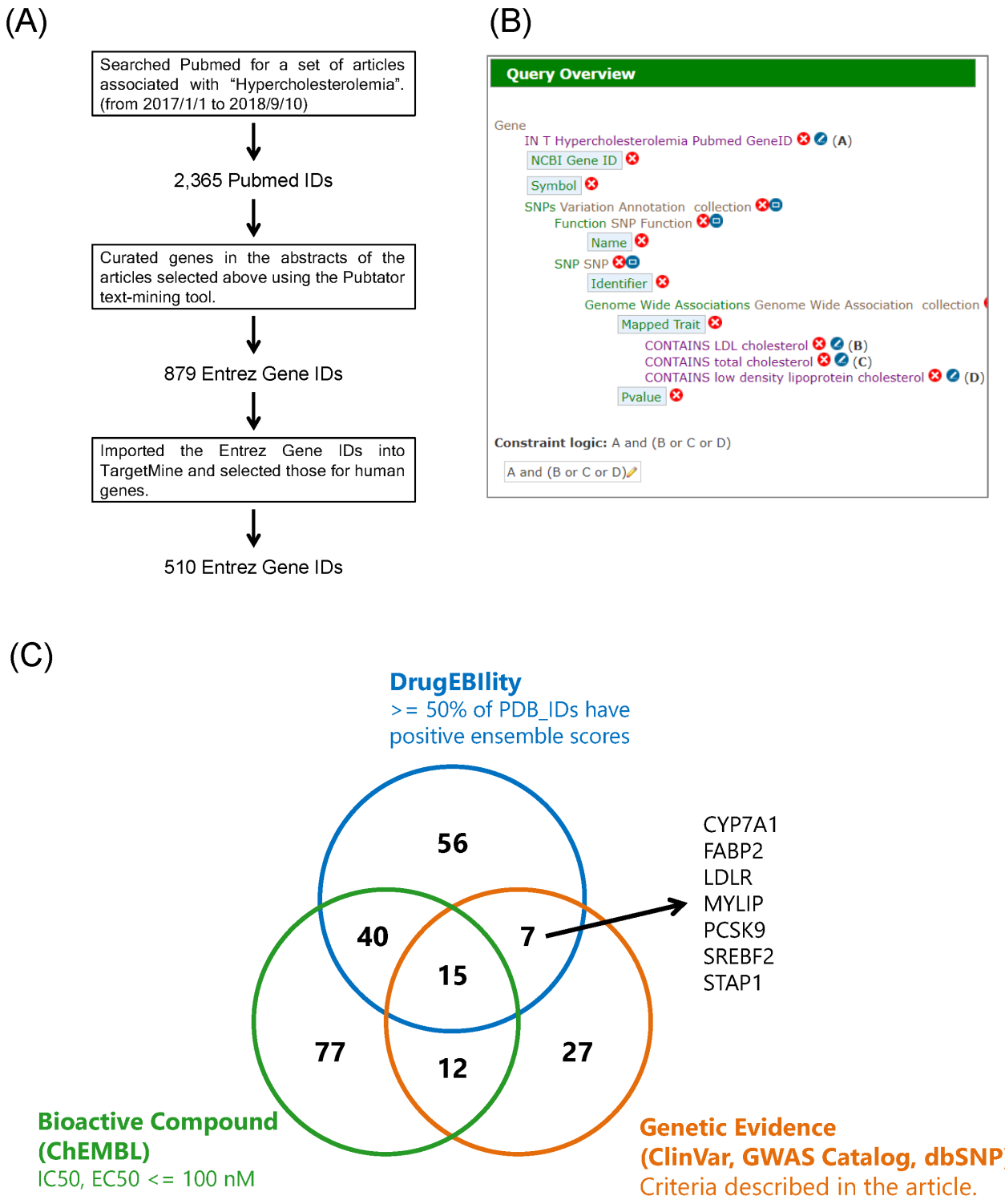


Figure 5. Gathering and prioritizing candidate drug target genes for hypercholesterolemia. (**A**) Gathering hypercholesterolemia-related genes from article information in PubMed. (**B**) The screenshot of the query builder in TargetMine to extract the genes with genetic evidence from GWAS catalog. Other query screenshots used in this prioritization process can be found in the extended data. (**C**) Prioritizing hypercholesterolemia-related genes with TargetMine to identify novel targets for small molecule drugs. Top prioritized genes were defined as those that met all of the following three requirements: 1) more than or equal to 50% of protein 3D structures (PDB IDs) having positive ensemble scores, 2) no potent bioactive compounds (EC50 or IC50 ≤ 100 nM in ChEMBL) and 3) having genetic associations with hypercholesterolemia (for more details, see the Use Cases section).

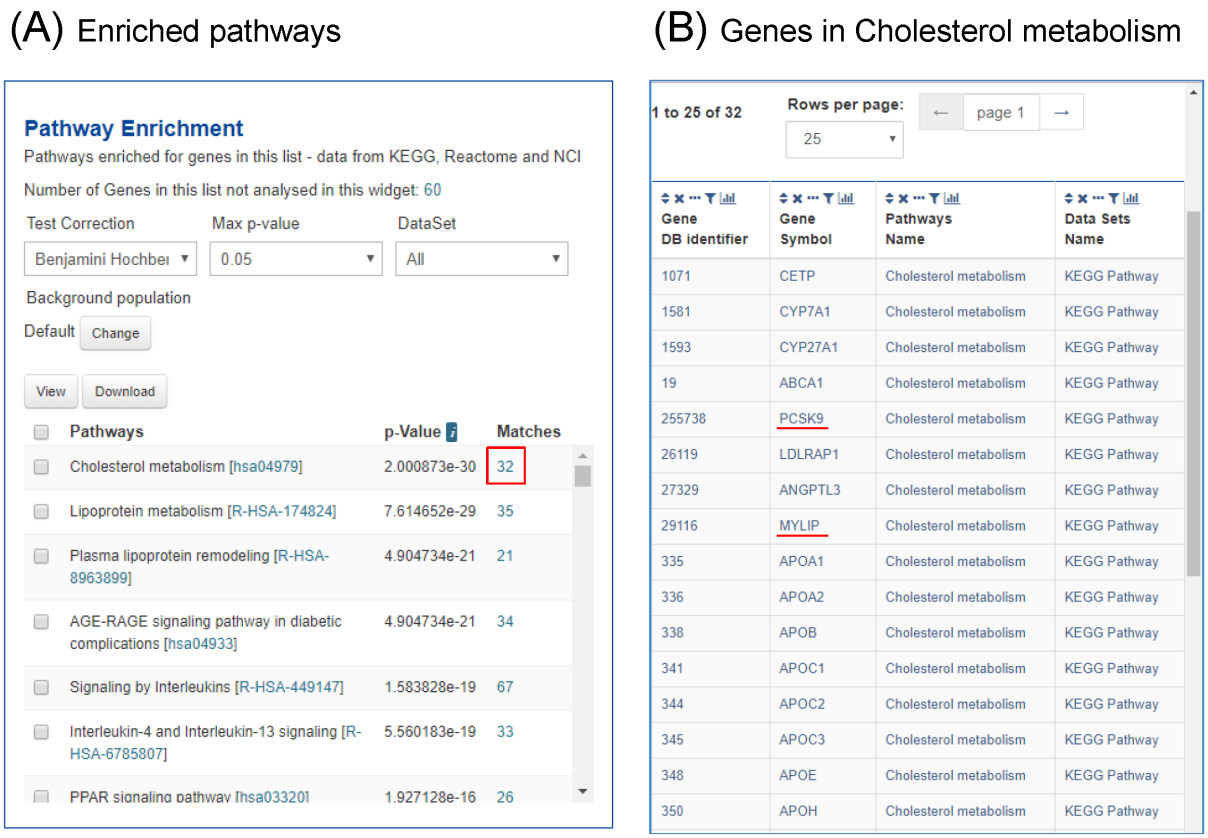


Figure 6. An example of pathway enrichment analysis. (**A**) Result of pathway enrichment analysis of the 510 human genes, which were described in Figure 5A. (**B**) Genes overlapped between the 510 human genes and those included in "Cholesterol metabolism" pathway.

Conclusions

These use cases demonstrate that the updated version of TargetMine[8] can be applied in pharmaceutical R&D, from the aspect of understanding the linkage to disease, examining the tractability of targets and prioritizing candidates. The recent update of the Open Targets platform[31] also starts to cover "DrugEBIlity" data and protein structural information, suggesting that an integrated resource containing gene-disease associations and tractability information is indispensable for the pharmaceutical R&D. In addition, taking advantage of the features of the InterMine framework, TargetMine[8] also facilitates more flexible and more complex queries for advanced users.

Data availability

Underlying data

The TargetMine data warehouse is publicly available at https://targetmine.mizuguchilab.org.

Extended data

Open Science Framework: Hypercholesterolemia related genes and TargetMine queries. https://doi.org/10.17605/OSF.IO/FUW53[32]

This project contains the following extended data:

- Hypercholesterolemia_PubMed_gene_list.tsv (The list of the 510 genes used in the use case, described in Figure 5A.)
- q_Hypercholesterolemia_ClinVar.xml (The query for extracting the genes with genetic evidence from ClinVar.)
- q_Hypercholesterolemia_dbSNP_PubMed.xml (The query for extracting the genes with genetic evidence from the associated publications in dbSNP.)
- q_Hypercholesterolemia_Bioactive_Compound.xml (The query for retrieving the genes with bioactive compounds.)
- q_Hypercholesterolemia_DrugEBIlity.xml (The query for retrieving genes with the DrugEBIlity scores.)
- q_Hypercholesterolemia_GWAScatalog.xml (The query for extracting the genes with genetic evidence from GWAS catalog.)

Extended data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Software availability

Source code available from: https://github.com/chenyian-nibio/targetmine

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2573565[8].

License: MIT License.

Grant information

This work was in part supported by JSPS KAKENHI (grant number 17K07268).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to thank Andrew Myers' help for improving the quality of writing.

References

1. Brown KK, Hann MM, Lakdawala AS, *et al.*: **Approaches to target tractability assessment - a practical perspective.** *Medchemcomm.* 2018; **9**(4): 606–613.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Bunnage ME: **Getting pharmaceutical R&D back on target.** *Nat Chem Biol.* 2011; **7**(6): 335–339.
PubMed Abstract | Publisher Full Text
3. Cook D, Brown D, Alexander R, *et al.*: **Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework.** *Nat Rev Drug Discov.* 2014; **13**(6): 419–431.
PubMed Abstract | Publisher Full Text
4. Nelson MR, Tipney H, Painter JL, *et al.*: **The support of human genetic evidence for approved drug indications.** *Nat Genet.* 2015; **47**(8): 856–860.
PubMed Abstract | Publisher Full Text
5. Piñero J, Bravo À, Queralt-Rosinach N, *et al.*: **DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.** *Nucleic Acids Res.* 2017; **45**(D1): D833–D839.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Koscielny G, An P, Carvalho-Silva D, *et al.*: **Open Targets: a platform for therapeutic target identification and validation.** *Nucleic Acids Res.* 2017; **45**(D1): D985–D994.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Nguyen DT, Mathias S, Bologa C, *et al.*: **Pharos: Collating protein information to shed light on the druggable genome.** *Nucleic Acids Res.* 2017; **45**(D1): D995–D1002.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Chen YA, Tripathi LP, Mizuguchi K: **TargetMine, an Integrated Data Warehouse for Candidate Gene Prioritisation and Target Discovery.** *PLoS One.* 2011; **6**(3): e17844.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Kalderimis A, Lyne R, Butano D, *et al.*: **InterMine: extensive web services for modern biology.** *Nucleic Acids Res.* 2014; **42**((Web Server issue): W468–472.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Welter D, MacArthur J, Morales J, *et al.*: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1001–1006.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Chen YA, Tripathi LP, Mizuguchi K: **An integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework.** *Database (Oxford).* 2016; **2016**: pii: baw009.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Kibbe WA, Arze C, Felix V, *et al.*: **Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data.** *Nucleic Acids Res.* 2015; **43**(Database issue): D1071–1078.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Schriml LM, Arze C, Nadendla S, *et al.*: **Disease Ontology: a backbone for disease semantic integration.** *Nucleic Acids Res.* 2012; **40**(Database issue): D940–946.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Malone J, Holloway E, Adamusiak T, *et al.*: **Modeling sample variables with an Experimental Factor Ontology.** *Bioinformatics.* 2010; **26**(8): 1112–1118.
PubMed Abstract | Publisher Full Text | Free Full Text
15. MacArthur J, Bowler E, Cerezo M, *et al.*: **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).** *Nucleic Acids Res.* 2017; **45**(D1): D896–D901.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Landrum MJ, Lee JM, Benson M, *et al.*: **ClinVar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Res.* 2016; **44**(D1): D862–868.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Landrum MJ, Lee JM, Riley GR, *et al.*: **ClinVar: public archive of relationships among sequence variation and human phenotype.** *Nucleic Acids Res.* 2014; **42**(Database issue): D980–985.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Higasa K, Miyake N, Yoshimura J, *et al.*: **Human genetic variation database, a reference database of genetic variations in the Japanese population.** *J Hum Genet.* 2016; **61**(6): 547–553.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Nagasaki M, Yasuda J, Katsuoka F, *et al.*: **Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals.** *Nat Commun.* 2015; **6**: 8018.
PubMed Abstract | Publisher Full Text | Free Full Text
20. NHLBI GO Exome Sequencing Project (ESP), S., WA. Vol. 2017.
21. 1000 Genomes Project Consortium, Auton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015; **526**(7571): 68–74.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Sudmant PH, Rausch T, Gardner EJ, *et al.*: **An integrated map of structural variation in 2,504 human genomes.** *Nature.* 2015; **526**(7571): 75–81.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Abifadel M, Varret M, Rabès JP, *et al.*: **Mutations in PCSK9 cause autosomal dominant hypercholesterolemia.** *Nat Genet.* 2003; **34**(2): 154–156.
PubMed Abstract | Publisher Full Text
24. Kiyosue A, Honarpour N, Kurtz C, *et al.*: **A Phase 3 Study of Evolocumab (AMG 145) in Statin-Treated Japanese Patients at High Cardiovascular Risk.** *Am J Cardiol.* 2016; **117**(1): 40–47.
PubMed Abstract | Publisher Full Text
25. Teramoto T, Kobayashi M, Tasaki H, *et al.*: **Efficacy and Safety of Alirocumab in Japanese Patients With Heterozygous Familial Hypercholesterolemia or at High Cardiovascular Risk With Hypercholesterolemia Not Adequately Controlled With Statins - ODYSSEY JAPAN Randomized Controlled Trial.** *Circ J.* 2016; **80**(9): 1980–1987.
PubMed Abstract | Publisher Full Text
26. Clarke L, Fairley S, Zheng-Bradley X, *et al.*: **The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data.** *Nucleic Acids Res.* 2017; **45**(D1): D854–D859.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Liu ZP, Wang Y: **PCSK9 Inhibitors: Novel Therapeutic Strategies for Lowering LDLCholesterol.** *Mini Rev Med Chem.* 2019; **19**(2): 165–176.
PubMed Abstract | Publisher Full Text
28. Wei CH, Kao HY, Lu Z: **PubTator: a web-based text mining tool for assisting biocuration.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W518–522.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Zhang CP, Tian Y, Zhang M, *et al.*: **IDOL, inducible degrader of low-density lipoprotein receptor, serves as a potential therapeutic target for dyslipidemia.** *Med Hypotheses.* 2016; **86**: 138–142.
PubMed Abstract | Publisher Full Text
30. Day IN: ***FH4=STAP1*. Another gene for familial hypercholesterolemia? Relevance to cascade testing and drug development?** *Circ Res.* 2014; **115**(6): 534–536.
PubMed Abstract | Publisher Full Text
31. Carvalho-Silva D, Pierleoni A, Pignatelli M, *et al.*: **Open Targets Platform: new developments and updates two years on.** *Nucleic Acids Res.* 2019; **47**(D1): D1056–D1065.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Chen YA: **Hypercholesterolemia related genes and TargetMine queries**. 2019.
https://www.doi.org/10.17605/OSF.IO/FUW53

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 28 March 2019

https://doi.org/10.5256/f1000research.19924.r45061

© 2019 Hersey A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Anne Hersey**
Chemogenomics Team, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

The authors have produced a nice tool (TargetMine) that can be used for prioritizing targets for drug discovery. They have done this by integrating data from a range of other resources that have the potential to identify the relevance of targets for diseases of interest as well as the likelihood that a small molecule can be found for that target (tractability). The authors have also provided all the code in the form of a github repository and the data model for TargetMine is shown in the article.

They describe adequately how to use the resource with the aid of a couple of specific use cases that they work through in the article. I have followed through the first example and can obtain the same results as they do. The 2nd example starts from a series of PubMed articles so is less easy to reproduce oneself although the logic is sensible. TargetMine also has links to tutorials that can be used to guide a user through using the resource although I haven't tried these. In the introduction the authors mention not just using genetic data to assess the relevance of targets to disease but additional information such as pathway analysis although this isn't exemplified in their use cases; perhaps their use cases could be expanded to include this? In the 2nd use case it would be good to describe at the start why they were prioritizing targets which had no reported inhibitors with IC50s <100nM. I assume this was because they were looking for novelty but this is not explained. To me the 15 targets at the centre of the Venn diagram (figure 5) would also be relevant and tractable targets.

I think it would be useful to be able to link back to the source databases that the information comes from. For example, when I select a UniProt ID I would expect to link back to UniProt itself. Likewise, it would be good to be able to link back to PDB where you can view the 3D structure of the protein. I couldn't find a way of doing this.

As the authors mention in their conclusion the Open Targets Platform has also recently started to use the DrugEBIlity algorithms. The authors might be interested to know that this is not currently being further updated and will soon be replaced with an alternative tractability prediction method that will also be free and openly available. A minor but important point: the authors use the words "Ensembl" and "ensemble"

to refer to the ensemble of prediction models that are used as one of the measures of tractability. “Ensemble” is the correct word which is very different from “Ensembl” the genome browser (https://www.ensembl.org/). I think it would be useful if the authors could correct this throughout the article.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Chemoinformatics, drug discovery, bioactivity databases

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 27 March 2019

https://doi.org/10.5256/f1000research.19924.r45533

© 2019 Lyne R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Rachel Lyne
Cambridge Systems Biology Centre, Department of Genetics, University of Cambridge, Cambridge, UK

The authors describe data updates to the already established TargetMine database. The updates allow candidate genes to be checked for genetic evidence for disease and target tractability. The authors clearly demonstrate the utility in integrating the new data through a set of use cases. The sophisticated web interface allows for powerful exploration and advanced querying of the data, making the resource a valuable tool for pharmaceutical research and development. The following minor points should be addressed:

1. In the introduction it is mentioned that additional information and analysis, such as pathway enrichment, are needed to assess aspects of target suitability, but such analysis is not shown in

the use-cases. It would be useful to demonstrate features that TargetMine provides that are not available in similar systems such as Open Targets and Pharos. In addition the authors describe how the new data can be used but do not tie this in with the wealth of data already available in TargetMine which could also be used for tractability studies, such as uniprot and interpro protein domains.

2. TargetMine includes an extensive API but this is not mentioned in the paper. Access to this data through an API could have clear advantages for anyone wishing to set up a workflow or provide a more automated analysis.

3. The use case “Gathering and prioritising candidate drug target genes” would be difficult for someone unfamiliar with searching an interMine-based database to reproduce. First, the set of genes (or the filtered set of human genes) should be provided as supplemental material. An overview of a set of queries constructed using the TargetMine query builder is provided. However, to reproduce this set of queries using the query builder is not immediately obvious for a naive user. A more detailed set of screenshots as supplementary material could help, or provision of the set of queries as a series of template searches.

4. A note on how often the data will be updated should be included.

5. Plans for further data additions could be included.

6. The reference for TargetMine (8) should be updated to contain the correct author list.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Databases, Data integration, data analysis

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research