SOFTWARE TOOL ARTICLE

Assessing drug target suitability using TargetMine [version 1; peer review: 2 approved]

Yi-An Chen¹, Erika Yogo², Naoko Kurihara², Tomoshige Ohno², Chihiro Higuchi¹, Masatomo Rokushima², Kenji Mizuguchi¹

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

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
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\(^1\). 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\(^1\), while the analysis results from GlaxoSmithKline suggest that selecting genetically supported targets could double the success rate in clinical development\(^1\). Several existing resources provide information about genetic evidences, such as DisGeNET\(^2\), Open Targets\(^3\), and Pharos\(^4\). 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\(^5\), a data warehouse for assisting target prioritization, and improved its functionalities for target assessment, particularly in small molecule drug discovery. TargetMine\(^5\) utilizes the InterMine framework\(^6\) 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.

Methods
Implementation
TargetMine\(^5\) is based on the InterMine framework, an open-source data warehouse system designed for biological data integration\(^7\). 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\(^8\). We had tried to associate genes to related diseases using the GWAS catalog in the former release of TargetMine\(^1\). To annotate disease terms to a trait or study, we first chose the disease ontology (DO)\(^9\)\(^,\)\(^10\) 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)\(^11\) to describe the curated GWAS traits in the recent implementation\(^12\). 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\(^13\)\(^,\)\(^14\). 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 haplotype, thus the

![Figure 1. Key factors to be considered in drug target selection.](image)

Linkage to disease
- Genetic disease associations
- Gene expression changes
- Functional assays in human cells
- KO mice phenotype

Tractability
- Target classes
- Protein 3D structures
- Existing small molecules

Potential side effects
- Genetic disease associations
- Tissue selectivity of expression
- KO mice phenotype
Figure 2. The new implemented data model. The colored lines indicate how the genes and diseases/phenotypes are associated in the post processing step.

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 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 (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.

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)\textsuperscript{18}, the integrative Japanese Genome Variation Database (1KJFN)\textsuperscript{19} (download from the archive in National Bioscience Database Center), the Exome Variant Server (EV5)\textsuperscript{20}, and the 1000 Genomes Project (1KGP)\textsuperscript{21,22}. At the moment, we only incorporate the population specific frequency for those SNPs stored in TargetMine.

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 clinical assertions to disease (MedGen), Gene to SNP to clinical assertions to disease (MedGen) terms (the green lines in Figure 2). 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\textsuperscript{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 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\textsuperscript{23}, and monoclonal antibodies to PCSK9 have been launched on the market as drugs for hypercholesterolemia with and without genetic predispositions\textsuperscript{24,25}.

Figure 3A demonstrates a schematic representation of the searching protocol for genetic disease associations with TargetMine. We first went to a gene report page by searching for the PCSK9 gene from the top page of TargetMine (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\textsuperscript{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\textsuperscript{27} (not shown), as mentioned above.

Querying target tractability for small molecule drugs with TargetMine
We performed another feasibility study to examine whether TargetMine\textsuperscript{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\textsuperscript{28}, indicating that PCSK9 is not a highly tractable target.

Figure 4A shows a schematic diagram of the procedure of querying tractability with TargetMine. 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, and identified several Protein Data Bank (PDB) entries for this protein (Figure 4C). Then, we moved to the “Protein Structure” page.
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.
of a specified PDB ID (2p4e in this case) and found that in the “DrugEBIliity” table (from the DrugEBIliity 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 DrugEBIliity may need to be further validated.

Collectively, we were able to confirm that the new version of TargetMine can quickly provide lines of evidences 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 for prioritizing candidate targets, we conducted a case study where we employed a list of genes associated with hypercholesterolemia in 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 Materials and Methods), (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 \leq 100 \text{nM}).

We first searched PubMed using the term “hypercholesterolemia” (from 2017/1/1 to 2018/9/10) and curated the resultant literature with the “Pubtator” text-mining tool, resulting in 510 human genes (Figure 5A). We then selected the genes meeting the requirements defined above using the “Query Builder” in TargetMine. Figure 5B shows an overview of the actual query, which aimed to extract the genes with gene evidences obtained from the GWAS catalog, where “Mapped Trait” contained “LDL cholesterol”, “total cholesterol”, or “low density lipoprotein cholesterol”, from ClinVar where “Reported phenotype Info” contains the term “Hypercholesterolemia”, and from dbSNP where “Mesh Terms Name” of related articles contains...
“Hypercholesterolemia”. 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. 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. This result suggests that the new version of TargetMine allows us to effectively prioritize target candidate genes in terms of linkage to disease, tractability and competitiveness. 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).

Conclusions
These use cases demonstrate that the updated version of TargetMine 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 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 also facilitates more flexible and more complex queries for advanced users.

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

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.

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


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 DrugEBillity 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