Bioinformatics, 35(11), 2019, 1981–1984
doi: 10.1093/bioinformatics/bty913
Advance Access Publication Date: 5 November 2018
Applications Note

OXFORD

Databases and ontologies

DrugThatGene: integrative analysis to streamline the identification of druggable genes, pathways and protein complexes from CRISPR screens

Matthew C. Canver[1,2,†], Daniel E. Bauer[3,4,5], Takahiro Maeda[6] and Luca Pinello[1,2,7,*]

[1]Molecular Pathology Unit, Center for Computational and Integrative Biology, Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA 02129, USA, [2]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA, [3]Division of Hematology/Oncology, Boston Children's Hospital, Boston 02115, USA, [4]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Boston 02115, USA, [5]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA, [6]Center for Cellular and Molecular Medicine, Kyushu University Hospital, Fukuoka 812-0054, Japan and Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[*]To whom correspondence should be addressed.
[†]Present address: Department of Pathology and Laboratory Medicine, New York-Presbyterian Hospital, Weill Cornell Medicine, New York, NY 10065, USA

Associate Editor: Jonathan Wren
Received on August 16, 2018; revised on October 23, 2018; editorial decision on October 30, 2018; accepted on October 31, 2018

Abstract

Motivation: The clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) nuclease system has allowed for high-throughput, large scale pooled screens for functional genomic studies. To aid in the translation of functional genomics to therapeutics, we developed DrugThatGene (DTG) as a web-based application that streamlines analysis of potential therapeutic targets identified from functional genetic screens.
Results: Starting from a gene list as input, DTG offers automated identification of small molecules along with supporting information from human genetic and other relevant databases. Furthermore, DTG aids in the identification of common biological pathways and protein complexes in conjunction with associated small molecule inhibitors. Taken together, DTG aims to expedite the identification of small molecules from the abundance of functional genetic data generated from CRISPR screens.
Availability and implementation: DTG is an open-source and free software available as a website at http://drugthatgene.pinellolab.org. Source code is available at: https://github.com/pinellolab/DrugThatGene, which can be downloaded in order to run DTG locally.
Contact: lpinello@mgh.harvard.edu

1 Introduction

Recent advances in the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) nuclease system have allowed for high-throughput pooled screens for functional genomic studies. These large-scale screens are often conducted in models of human disease/phenotypes, which may potentially lead to the identification of therapeutic targets. While computational methods have been developed to automate the analysis of CRISPR-based screens (Li *et al.*, 2014, 2015; List *et al.*, 2016; Miles *et al.*, 2016; Winter *et al.*, 2016; Yu *et al.*, 2015), methods to streamline the identification of small molecules for identified

Table 1. List of databases utilized by DTG

Database	Data Contents	Website	Reference
Cancer Target Discovery and Development (CTD^2)	Database of drugs with known targets with a focus on cancer targets	https://ocg.cancer.gov/programs/ctd2	–
ClinVar	Database of relationships between human variation and phenotypes	https://www.ncbi.nlm.nih.gov/clinvar	Landrum *et al.* (2014)
CORUM	Database of mammalian protein complexes	http://mips.helmholtz-muenchen.de/corum/#	Ruepp *et al.* (2008)
Drug Gene Interaction Database (DGIdb)	Database of drugs with known targets	http://www.dgidb.org/	Griffith *et al.*, (2013) and Wagner *et al.* (2016)
Exome Aggregation Consortium (ExAC)	Database aggregating exome sequencing data. The database consists of >60 000 whole exome sequences from unrelated individuals	http://exac.broadinstitute.org	Lek *et al.* (2016)
gnomAD	Database aggregating exome and genome sequencing data. The database consists of >120 000 exomes and >15 000 whole genome sequences.	http://gnomad.broadinstitute.org	Lek *et al.* (2016)
Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways	Pathways derived from the KEGG database based on molecular interaction, reaction and relation networks	https://www.genome.jp/kegg/	Ogata *et al.* (2000)
Online Mendelian Inheritance in Man (OMIM)	Comprehensive catalog of human genes and genetic phenotypes	https://www.omim.org	Hamosh *et al.* (2002)
Pharos	Integrated database of targets within the druggable genome	https://pharos.nih.gov/idg/index	Nguyen *et al.* (2017)
STRING	Database of protein–protein interaction networks	https://string-db.org/	von Mering *et al.* (2005)

therapeutic targets remain elusive. A variety of studies have sought to address this question by designing screens focused on classically druggable targets, such as kinases or phosphatases (Housden *et al.*, 2015; Shi *et al.*, 2015; Wong *et al.*, 2016). Others have sought to reverse this pipeline and use CRISPR/Cas9 to identify the target(s) for a given small molecule (Deans *et al.*, 2016). However, no analytical methods exist at present to help bridge the gap between large-scale functional genetic screens and small molecule inhibitors available against putative therapeutic targets.

2 DrugThatGene (DTG)

We developed DrugThatGene (DTG) as a freely-available, web-based application to automate analysis of potential therapeutic targets identified from functional genetic screens. Starting from a gene list obtained through a screening strategy, DTG's analysis identifies druggable genes, pathways and protein complexes through comprehensive integration of several databases of human genetic variation, small molecules, biological pathways and protein–protein interactions (Table 1). Specifically, DTG streamlines analysis for users by automatically querying, collecting and aggregating data for each putative therapeutic target from relevant databases (Table 1). Further, DTG performs pathway and protein complex composition analysis, which allows for the identification of pathways and protein complexes enriched in the input gene list along with associated small molecule inhibitors. The synthesized data from DTG is displayed in easy to interpret tables, which dramatically reduces the time for users as compared to retrieving and manually combining data from other websites/databases and performing pathway/protein complex analysis on a one-by-one basis (from hours to a few seconds).

3 DTG implementation

DTG requires a list of HUGO Gene Nomenclature Committee (HGNC) genes symbols as input data (Yates *et al.*, 2017). While intended to receive the top genes identified from a pooled CRISPR screen, DTG will analyze any given gene list regardless of its origin. In addition to a gene list, users must specify the number of genes required to identify a common pathway and a common protein complex (the default value is two for both common pathways and common protein complexes). For example, a value of two in the common pathways field specifies that at least two genes from the input gene list must be in the same pathway to be identified by DTG as a common pathway. Users can perform more conservative analysis by requiring a larger number of input genes to be contained within a common pathway or common protein complex. A conservative threshold may be helpful for large input gene lists in an effort to identify the most highly enriched pathways and/or protein complexes. Alternatively, a more relaxed threshold can be used to encourage identification of a greater number of common pathways or common protein complexes. A relaxed threshold may be helpful for small input gene lists in order to offer an increased number of potential pathways and/or protein complexes for future investigation. Notably, the thresholds for common pathways and common protein complexes can be modulated independently. The webtool allows for up to 200 input gene symbols. The limit of 200 input gene symbols is imposed to avoid web server overburdening. If users want to perform DTG analysis on input gene lists of >200 genes, the source code can be downloaded from Github to run DTG locally with an increased input limit.

The output of DTG analysis consists of four tables: (i) A table of genes with small molecules available to target them from the DGIdb and CTD^2 databases as well as further druggability information available within the Pharos database (Table 1). In addition, KEGG pathways, protein complex involvement based on the CORUM database, and a protein–protein interaction plot from the STRING database are provided for each gene. Of note, this table does not offer common pathways or common protein complexes; every

relevant KEGG pathway and protein complex is provided for each gene. Finally, human genetic variation databases (OMIM, ClinVar, gnomAD and ExAC) are displayed to assist in the identification of druggable targets by providing loss of function phenotypes and implications on human disease/phenotypes (Table 1). The number of loss of function variants found in the ExAC database is displayed alongside the ExAC Missense Z/pLI scores, which offer quantitative assessments of a given gene's tolerance for mutation. The druggable genes table is displayed in the same order as the input gene list. The relevant links to the original entries of the queried databases are provided (if available). (ii) A table of common KEGG pathways containing the user-specified number of genes from the input list is displayed. This table offers a list of common biological pathways including small molecules available to target members of each respective pathway. The pathways are sorted by the fraction of input genes in each pathway (number of input genes in pathway/total number of genes in pathway) as these may offer more probable targets for validation. (iii) A table of common protein complexes based on the CORUM database containing the user-specified number of genes from the input list is displayed. This table offers a list of common protein complexes including small molecules available to target members of each respective protein complex. The protein complexes are sorted by the fraction of input genes in each complex (number of input genes in complex/total number of genes in complex) as these may offer more probable targets for validation. (iv) A fourth table lists all input genes not included in DTG's analysis because they were not found in any of the utilized databases, which typically results due to deviation from the HGNC gene symbol nomenclature. All four output tables can be downloaded as text files.

The 'Help' section included on the DTG website offers step-by-step instruction for using DTG along with sample input gene lists based on previously published CRISPR screens (Shalem *et al.*, 2014; Yamauchi *et al.*, 2018).

4 Example implementation

A recent report from Yamauchi *et al.* highlights the utility of DTG. In this study, a genome-wide CRISPR screen was performed to identify genes required for acute myeloid leukemia (AML) cell survival *in vitro* and *in vivo*. This work led to the identification of 130 genes, which included the mRNA decapping enzyme scavenger (*DCPS*) gene. The authors subsequently validated DCPS as a therapeutic target for AML and demonstrated that a DCPS inhibitor (RG3039) led to slowed AML cell proliferation and induced AML cell differentiation (Yamauchi *et al.*, 2018). This list of 130 essential genes for AML cell survival was input into DTG with the requirement of two genes for identification of common pathways and common protein complexes. DTG identified 15/130 (11.5%) druggable genes based on the databases used at the time of the writing of this manuscript. Of note, DTG identified two inhibitors for DCPS, which included the validated RG3039 inhibitor: (1) 5-{[1-(2-fluorobenzyl)piperidin-4-yl]methoxy}quinazoline-2,4-diamine (RG3909; CHEMBL251429) and (2) 5-[(1S)-1-(3-chlorophenyl)ethoxy]quinazoline-2,4-diamine (D156844; CHEMBL253976).

Yamauchi *et al.* demonstrated the role of DCPS in pre-mRNA metabolic pathways. In addition, the authors identified DCPS interactors such as spliceosomes, transcription-export complex (TREX), nuclear pore complex (NUP), the nucleosome remodeling and deacetylase (NuRD) complex and pre-rRNA processing complexes (Yamauchi *et al.*, 2018). DTG analysis of the essential gene list identified 164 common pathways with 66/164 (40.2%) pathways targetable by available drugs (when requiring two genes to identify a common pathway). Notably, 13 of the top 20 (13/20, 65%) pathways with the highest fraction of input genes in each pathway involved mRNA processing, which included 3/13 (23.1%) mRNA processing-related pathways targetable by available drugs. Interestingly, DTG identified 61 common protein complexes with 4/61 (6.6%) protein complexes targetable by available drugs (when requiring two genes to identify a common protein complex). The protein complexes identified by DTG included the TREX, NuRD, spliceosome and pre-rRNA processing complexes identified by Yamauchi *et al.*

Taken together, DTG analysis identified numerous druggable genes, pathways and protein complexes for validation as potential therapeutics for AML. Moreover, the druggable pathways and protein complexes with the highest fraction of input genes implicate high confidence therapeutic targets for investigators to study and offer insight into the biology related to AML pathogenesis. Of note, DTG analysis of the input gene list identified known genes/pathways/protein complexes for drug targeting as well as novel targets for future validation. The list of 130 essential genes for AML cell survival is available in the 'Help' section of the DTG website.

5 Conclusions

High-throughput, large scale functional genetic screening enabled by CRISPR/Cas systems has accelerated the ability to interrogate human diseases and phenotypes. DTG offers an integrative web-based application to exploit the knowledge from functional genetic screens to aid in the translation of functional genomics to therapeutics.

Funding

M.C.C. was supported by a National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Award (F30DK103359). D.E.B. was supported by K08DK093705, R03DK109232, DP2OD022716, P01HL032262, Burroughs Wellcome Fund, Doris Duke Charitable Foundation and an ASH Scholar Award, T.M. was supported by an NIH R01DK111455, an American Society of Hematology Bridge Grant Program and a JSPS Grant-in-Aid for Scientific Research (A) (17H01567). L.P. was supported by a National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399) and a Centers for Excellence in Genomic Science (CEGS) Award (RM1HG009490).

Conflict of Interest: none declared.

References

Deans,R.M. *et al.* (2016) Parallel shRNA and CRISPR-Cas9 screens enable antiviral drug target identification. *Nat. Chem. Biol.*, **12**, 361–366.

Griffith,M. *et al.* (2013) DGIdb: Mining the druggable genome. *Nat. Methods*, **10**, 1209–1210.

Hamosh,A. *et al.* (2002) Online Mendelian Inheritance in Man (OMIM): a directory of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Housden,B.E. *et al.* (2015) Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Sci. Signal*, **8**, rs9.

Landrum,M.J. *et al.* (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, 980–985.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60 706 humans. *Nature*, **536**, 285–291.

Li,W. *et al.* (2014) MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.*, **15**, 554.

Li,W. *et al*. (2015) Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol.*, **16**, 281.

List,M. *et al*. (2016) Comprehensive analysis of high-throughput screens with HiTSeekR. *Nucleic Acids Res.*, **44**, 6639–6648.

von Mering,C. *et al*. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, 433–437.

Miles,L.A. *et al*. (2016) Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens. *FEBS J.*, **283**, 3170–3180.

Nguyen,D.T. *et al*. (2017) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, D995–D1002.

Ogata,H. *et al*. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Ruepp,A. *et al*. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, 646–650.

Shalem,O. *et al*. (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.

Shi,J. *et al*. (2015) Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.*, **33**, 661–667.

Wagner,A.H. *et al*. (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.

Winter,J. *et al*. (2016) CaRpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. *Bioinformatics*, **32**, 632–634.

Wong,A.S.L. *et al*. (2016) Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. *Proc. Natl. Acad. Sci. USA*, **113**, 2544–2549.

Yamauchi,T. *et al*. (2018) Genome-wide CRISPR-Cas9 screen identifies leukemia-specific dependence on a pre-mRNA metabolic pathway regulated by DCPS. *Cancer Cell*, **33**, 386–400.e5.

Yates,B. *et al*. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.

Yu,J. *et al*. (2015) ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. *Bioinformatics*, **32**, 260–267.