IDENTIFICATION OF NOVEL THERAPEUTICS FOR COMPLEX DISEASES FROM GENOME-WIDE ASSOCIATION DATA

M. P. GROVER

School of Medicine, Deakin University, Geelong, Victoria, Australia.
mgrover@deakin.edu.au

S. BALLOUZ

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States.
sballouz@cshl.edu

K. A. MOHANASUNDARAM

School of Medicine, Deakin University, Geelong, Victoria, Australia.
kmohanas@deakin.edu.au

R. A. GEORGE

Victor Chang Cardiac Research Institute, 405 Liverpool St, Darlinghurst, 2010, NSW, Australia.
r.george@victorchang.edu.au

C .D. H. SHERMAN

Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia.
craig.sherman@deakin.edu.au

T. M. CROWLEY

School of Medicine, Deakin University, Geelong, Victoria, Australia.

Australian Animal Health Laboratory, CSIRO Animal, Food and Health Sciences, Portarlington Road, Geelong, Victoria,
tam syn.crowley@deakin.edu.au

M. A. WO UTERS

School of Medicine, Deakin University, Geelong, Victoria, Australia.
m .wouters@deakin.edu.au

Human genome sequencing has enabled the association of phenotypes with genetic loci, but our ability to effectively translate this data to the clinic has not kept pace. Integration of drug-target data with candidate gene prediction systems can identify novel phenotypes which may benefit from current therapeutics. Such a drug repositioning tool can save valuable time and money spent on preclinical studies and phase I clinical trials.

We adopted a simple approach to integrate drug data with candidate gene predictions at the systems level. We previously used Gentrepid as a platform to predict 1,497 candidate genes for seven complex diseases considered in the Wellcome Trust Case Control Consortium genome wide association study. Using the publicly available drug databases, Therapeutic Target Database, PharmGKB and DrugBank as sources of drug-target association data, we identified a total of 428 (29%) candidate genes as novel therapeutic targets for the phenotype of interest and 2,130 drugs feasible for repositioning against the predicted targets.

By integrating genetic, bioinformatic and drug data, we have demonstrated that currently available drugs may be repositioned as novel therapeutics for the seven diseases studied here, quickly taking advantage of prior work in pharmaceutics to translate ground breaking results in genetics to clinical treatments.

Keywords: *Complex disease, Candidate gene, Inherited human disease, Drug database, Drug Target, Drug repositioning*

Background

The development of new therapeutics is essential to improve the human condition and lower the burden of disease. Due to our limited knowledge of the molecular basis of complex diseases, comparatively few gene targets for therapeutics have been identified to date. The standard approach to developing therapeutics involves testing many thousands of compounds against a known target in order to identify a lead compound. The lead compound can then be further refined *in silico* and *in vitro* before heading into the lengthy and costly clinical trials pipeline. This process, which consists of phases I, II, III and IV before final drug approval, involves 10-17 years of drug development from target identification until FDA/EMEA approval, with only a 10% probability of success [1]. As a result, the pharmaceutical industry spends an average of about 1.2 billion US dollars to bring each new drug to market [2]. There is also a high risk associated with *de novo* drugs due to unforeseen adverse side effects, as seen in the case of Thalidomide, a drug used to treat morning sickness which resulted in devastating birth defects [3].

A novel approach to therapeutic development is to identify new applications for drugs that have already been approved, or have successfully completed phase I clinical trials [4, 5]. This process of "drug repositioning" aims not to develop drugs *de novo,* but associate existing therapeutics with new phenotypes. Here, we attempted to reposition existing drugs to treat common complex diseases using recently acquired Genome Wide Association Study (GWAS) data.

Complex diseases are genetically intricate, polygenic and multifactorial [6]; and frequently arise as a consequence of interaction between genes and the environment. Recently, GWAS have begun to

unravel the complicated genetic basis of complex diseases. Sheer statistical power has allowed GWAS to successfully identify some associations between Single Nucleotide Polymorphisms (SNPs) and complex diseases [7]. Using these genetic loci, GWAS have enabled identification of novel drug targets for common diseases. Despite high investment, far fewer genes have been identified than can account for the heritable component of complex diseases, and the clinical benefit remains limited to date [8]. A factor that contributes to the missing heritability is likely to be noisy genotype-phenotype association signals [9]. Also, analysis of GWAS data using highly stringent thresholds for statistical significance, by testing multiple isolated SNPs, has limited the scope of gene discovery based on existing data [10]. As shown in Manhattan plots, GWAS data obviously contain far more information than the most significant peaks and more work needs to be done extracting data from these slightly less significant peaks [9, 11].

Currently available gene discovery platforms can enhance candidate gene identification from GWAS data [9]. Candidate gene prediction tools are designed to find the needle in the genetic haystack. These tools are based on the assumption that genes with similar or related functions cause similar phenotypes [12]. Specific candidate gene prediction tools differ in the strategy adopted for calculating similarity, and the databases utilized for the prediction [13, 14]. *Gentrepid* is one of the many bioinformatic tools developed to help geneticists predict and prioritize candidate genes [9, 15]. The salient features of the *Gentrepid* tool and its knowledge base are it utilizes two independent methods Common Pathway Scanning, a systems biology approach; and Common Module Profiling, a domain-based homology recognition approach, to prioritize candidate genes for human inherited disorders (See *Methods* for details). Compared to other prediction systems, *Gentrepid* is designed to make fewer, more conservative predictions which do not extensively extrapolate existing bioinformatic data i.e. it tends to be more specific than other systems [15].

Used in conjunction with public drug databases as major drug repositories, candidate gene prediction tools can facilitate the therapeutic drug-target discovery process. We have previously developed protocols to analyze GWAS data using a multilocus approach which combines bioinformatic and genetic data [9]. To demonstrate the usefulness of this protocol, we applied it to the well-studied Wellcome Trust Case Control Consortium (WTCCC) GWAS data for seven complex diseases [9]. Using a series of increasingly less conservative statistical thresholds, we attempted to discriminate the signal from the noise in the most statistically significant data ($p<10^{-4}$). By incorporating bioinformatic data, we were able to predict ~1,497 candidate genes by reanalyzing the well-studied data on the seven complex diseases, namely, Type 2 Diabetes (T2D), Bipolar Disorder (BD), Crohn's Disease (CD), Hypertension (HT), Type 1 Diabetes (T1D), Coronary Artery Disease (CAD) and Rheumatoid Arthritis (RA) [9].

Here, we extend this pipeline to identify potential novel drug targets among the predicted candidate genes by associating drug information extracted from publicly available drug databases. The three databases sourced in this study were DrugBank [16], PharmGKB [17] and the Therapeutic Target Database (TTD) [18]. The feasibility of this approach is illustrated for the seven complex diseases studied by the WTCCC [11]. This study allows identification of possible therapeutics for treatment of specific complex diseases by enabling association of predicted candidate genes with a complex disease and providing possible drug compound information towards cures. Thus, in combination with drug target information, *Gentrepid* can be utilized as a drug discovery tool to identify therapeutics which may be repositioned as novel treatments for seven complex diseases.

Methods

We implemented a computational workflow to enable repositioning of drugs by integrating two data sets (Figure 1)

1. Candidate gene dataset obtained by integration of genotype-phenotype data from the WTCCC GWAS study on seven complex phenotypes [11] and bioinformatic data on structural domains and systems biology to associate proteins that share common features or participate in the same complex or pathway [19];
2. Drug-Target association data set obtained from three drug database namely TTD, DrugBank and PharmGKB [16-18].

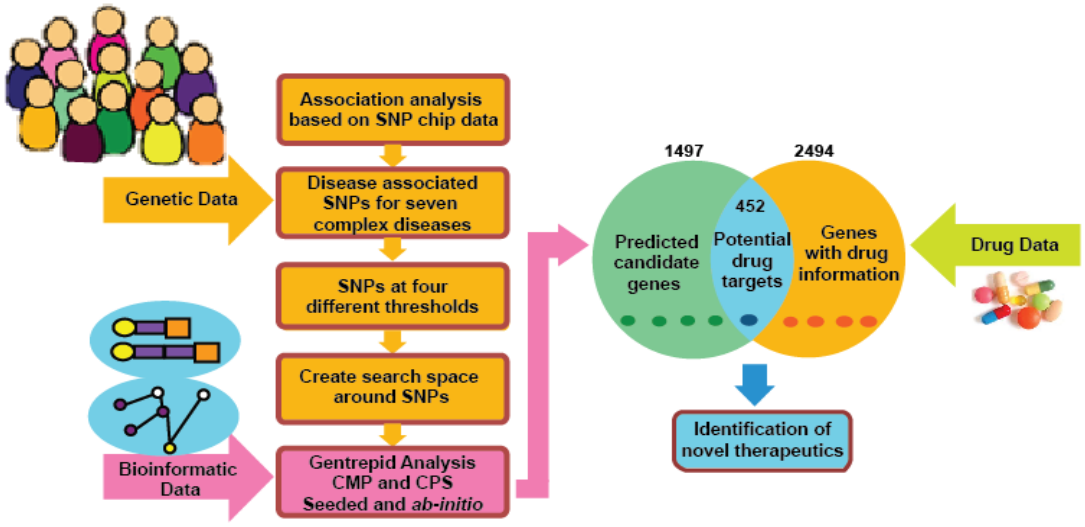


Figure 1. The complete workflow

We used *Gentrepid* as a gene discovery bioinformatics platform and drug databases implemented online as web based tools repository of drug data. In previous work, we predicted a total of ~1,497 candidate genes for seven complex diseases by careful reanalysis of the WTCCC GWAS data [11] using the *Gentrepid* candidate gene prediction system [9]. In the original analysis, a highly stringent significance threshold ($p<5 \times 10^{-7}$) was used in an attempt to correct for multiple testing. This conservative statistical approach, combined with the selection of the nearest-neighboring gene to the significant SNP, resulted in identification of only a small number of genes, with modest cumulative heritability, associated with each phenotype (Table 1). We specifically addressed these two issues in our reanalysis of this noisy data by:

1. Considering a series of four thresholds of decreasing stringency, starting with the highly significant threshold used in the original study and decreasing to weakly significant ($p< 10^{-3}$). This resulted in a series of four SNP sets containing up to ~804 SNPs being considered for each phenotype.

2. Creating six different search spaces around each SNP cluster, 3 of fixed-widths and 3 proximity-based, which were analyzed with our candidate gene prediction system. Twenty-four search spaces were constructed per phenotype using multiple SNP significance thresholds and gene selection methods. In total, 168 search spaces ranging in size from 2 to 4,431 genes (up to 10% of the genome) were analyzed using *Gentrepid*. *Gentrepid* is based on two modules:

The Common Pathway Scanning (CPS) module is based on the assumption that common phenotypes are associated with proteins that participate in the same complex or pathway [20]. System biology methods are currently favoured in candidate gene prediction because of the attractiveness of their underlying thesis. Their weakness is the lack of coverage of the underlying system biology knowledge bases. Many tools attempt to ameliorate the deficits of the knowledge base by extensive extrapolation of data from other species. Examples are GeneSeeker, ToppGene and Endeavour [13, 21-23]. *Gentrepid* CPS uses only human data to reduce the number of predicted false positives i.e. it makes fewer predictions which are more often correct compared to other prediction systems [15]. Common Module Profiling (CMP) is a novel sequence analysis approach based on the principle that candidate genes have similar functions to disease genes already determined for the phenotype [24].

Gentrepid differs from most candidate gene prediction systems which describe functional similarity via keywords, a procedure which also lacks good coverage of the human genome. In CMP, sequences are parsed at the domain level, linking them directly to function [19]. Although CMP's performance was disappointing in our original benchmark using a set of nine oligogenic diseases with Mendelian inheritance, it produced a surprising number of statistically significant results when confronted with the GWAS data on seven complex diseases [9]. This result was robust when compared with simulations using random SNPs, and may arise from an underlying role for homologous genes specific to complex diseases.

Drug-Gene Target dataset

We compiled drug-gene target dataset from three publicly available drug databases DrugBank [16] the Pharmacogenomics KnowledgeBase- PharmGKB [17] and the Therapeutic Target Database (TTD) [18] on June 2012.

Table 1. Average number of SNPs, loci and genes per phenotype used gene prediction with significant association p values.

Level		WS	MWS	MHS	HS
		$p \leq$ 1e-3	$p \leq$ 1e-4	$p \leq$ 1e-5	$p <$ 5e-7
SNPs		804.29	160.29	56.71	29.14
Loci		446.86	84.43	18.71	7.29
Total genes in search space -BY	**1Mbp**	3875.57	870.86	175.29	87.43
	0.5 Mbp	2140.00	477.29	106.00	57.29
	0.1 Mbp	654.57	148.43	43.71	23.00
Total genes in search space -NN	**Adjacent**	1412.14	292.43	62.29	26.14
	Nearest	452.86	91.0	22.29	10.14
	Smallest	198.71	42.57	11.43	5.43
Annotated candidate genes -BY	**1Mbp**	2285.29	528.86	116.43	61.57
	0.5 Mbp	1275.57	300.43	73.14	41.57
	0.1 Mbp	426.43	103.43	32.00	16.57
Annotated candidate genes -NN	**Adjacent**	803.14	172.00	40.71	17.57
	Nearest	285.71	59.00	15.57	6.14
	Smallest	155.29	33.43	8.86	5.57

DrugBank is a freely available online database that combines detailed drug data with comprehensive drug-target and indication information. In this study, we used the DrugBank drug IDs, drug generic and brand names, to represent drugs and the unique gene symbols to represent protein targets. We extracted ~6,711 drug entries active against the ~3,410 unique drug targets including ~2,022 human drug targets associated with ~3910 drugs. **PharmGKB** is another drug knowledge base that captures information about drugs, diseases/phenotypes and targeted genes. From this database, we extracted the "drug-associated genes" along with "description" which contains the disease information. We retrieved ~382 drugs for ~566 drug targets from the PharmGKB database because some drugs target multiple genes. **Therapeutic Target Database (TTD)** is also a freely available online drug database which integrates drug data with therapeutic targets. This database contains ~17,816 drugs against both human and non-human (bacterial and fungal protein) targets. We extracted "Drug names" along with "Disease" information and Uniprot accession numbers for "targets". UniProt accession numbers were replaced with official HUGO gene symbols using the G-profiler conversion tool [25]. Finally, we extracted ~2,960 drugs for ~544 unique human drug targets from this database.

Validation of predicted therapeutic targets

The predicted therapeutic targets were further validated using two benchmarks. In the first benchmark, the actual status of the gene was assessed by the existence or the non-existence of the abstract in the literature citing both the gene name and the phenotype. In the second benchmark, the actual status of genes based on whether they are designated as targets in the drug databases or not. This was repeated for all the six search spaces investigated for each phenotype (Table 1).

In the first benchmark work, all Pubmed IDs of literature related to Bipolar disorder, Type 1 diabetes, Type 2 diabetes, Crohn's disease, Coronary artery disease, Rheumatoid arthritis and Hypertension were first downloaded from Pubmed on Feb. 2013. For each target, we calculated the number of citations related to each disease by mapping the extracted Pubmed IDs to the gene citation information from Entrez Gene (ftp//ftp.ncbi.nih.gov/gene/), composed of genes and their corresponding cited literature. Further, Receiver Operation Characteristics Curves were created considering four thresholds of atleast one, five, ten, fifteen citations (see section validation of therapeutic targets in Result and Discussion).

In the second benchmark, genes present in six search spaces classified as "candidates" or "non-candidates" (Table 1). We considered targets already containing drugs for the phenotype of interest as "true positives". Targets with currently registered therapeutics for the phenotype of interest which were not predicted by *Gentrepid* but already present in the search space as "false negatives". Genes which were not predicted and not targetable by drugs as "true negatives" and predicted novel therapeutic targets were considered as "false positives" (see section validation of therapeutic targets in *Results and Discussion*).

Results and Discussion

Comparison of drug databases

The therapeutic drug-gene target association data was extracted from three following databases

1 DrugBank, a Cheminformatics/Bioinformatics database - 2,022 human targets associated with 3910 drugs [16];
2 PharmGKB, a Pharmacogenomics database - 382 drugs against 566 human targets [17] and;
3 TTD, a database with comprehensive information about drug targets - 2,960 drugs against 544 human targets [18]

It is estimated that ~3,000-5,000 genes are druggable which is ~10-17% of the entire genome The total number of unique targets, we retrieved from all the databases were ~2,494 genes which is ~8% of the entire genome (Figure 2). The gap between extracted targets and estimated druggable genes exist because of the fact that there are many druggable genes which are yet to be identified as drug targets. The total targets used in our study cover 50-83% of the possible druggable genes mentioned in previous studies [26-29]. We compared raw data such as drug targets across three drug databases to determine the redundancy of the information in these databases. With respect to drug targets, only ~4 % of drug target entries were common to all the three databases (Figure 3). When the databases were compared in a pairwise fashion, the proportion of common targets ranged from 5-12%. Each of the databases contains a significant amount of information that is unique to that database. TTD has the least number of unique targets (129), while DB and PharmGKB include ~1495 and ~326 respectively (Figure 3). The low proportion of similarity among the databases and the high number of unique targets in each database shows the databases are fairly complementary. In total, targets with drug information represents 8% of the human genome (Figure 2).

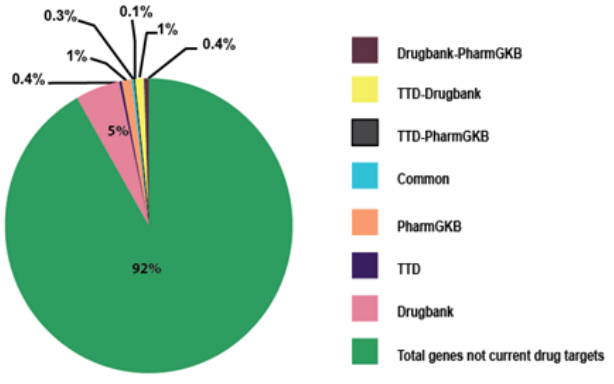


Figure 2. Coverage of human genome by annotated drug targets in drug databases

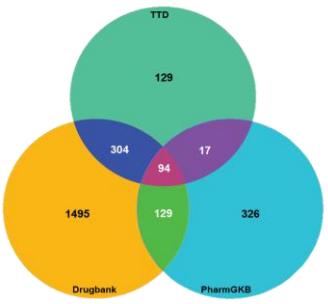


Figure 3. Comparison of extracted drug targets from three databases databases

Identification of Therapeutic Targets

We identified potential therapeutic targets for seven complex diseases by using lists of *Gentrepid*-predicted candidate genes generated by our reanalysis of the WTCCC data. In total, *Gentrepid* predicted ~1,497 candidate genes for all the seven diseases namely Type 2 Diabetes (291), Bipolar Disorder (212), Crohn's Disease (356), Hypertension (219), Type 1 Diabetes (280), Coronary Artery Disease (258) and Rheumatoid Arthritis (189) [3]. We searched for these candidate genes in the drug-gene target files obtained from all the three databases and found ~452 as potential therapeutic targets for the seven complex diseases (Table 1, Figure 4). This illustrates that almost ~30% of the total number of predicted candidate genes by *Gentrepid* are potential targets for therapeutic treatments using currently available drugs.

To drill a little further into the data, we assessed the therapeutic potential of each phenotype using currently available repositioned drugs, by calculating Targetability index (TI) i.e. the ratio of number of predicted targets to the number of predicted candidate genes for each phenotype (Table 2). CAD was the most targettable phenotype investigated (TI= 0.39), on the other hand, T1D was the least amenable phenotype with (TI= 0.27). The factor which likely to influence the targetability of the phenotype is our underlying knowledge of the phenotype. If the molecular pathways involved have been previously characterized, there is more likely to be drug-target information in the existing drug databases, even if the phenotype has not previously been associated with the molecular system. The low TI in case of T1D (0.27) and BD (0.28) likely arises from lack of knowledge of underlying pathways. More basic research in this area is required.

As shown in figure 4, all the three drug databases made significant contributions to target identification with the highest contribution from DrugBank (400) followed by TTD (156) and PharmGKB (61). DrugBank is a chemical as well as clinical drug database which contains broader coverage of drug targets compare to TTD, a chemical drug database and PharmGKB, a clinical drug database. Therefore, we obtained least number of therapeutic targets from PharmGKB and highest number of therapeutic targets from DrugBank. To summarize, the total coverage of the predicted targets from all the three databases was estimated to be 30% of the total candidate genes predicted by *Gentrepid* with the maximum contribution from DrugBank followed by TTD and PharmGKB (Figure 5).

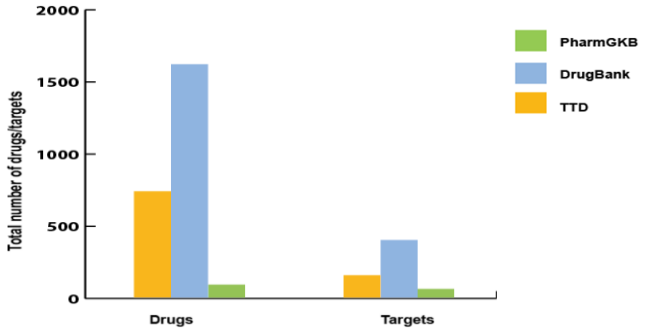


Figure 4. Comparison of number of predicted therapeutic targets and drugs

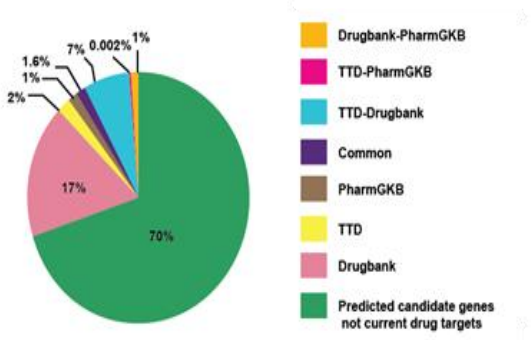


Figure 5. Predicted Therapeutic targets

Identification of Novel Therapeutic Targets

We classified 452 predicted targets to distinguish therapeutic targets which were "rediscovered" for the phenotype and those which might be novel therapeutic targets for the specific phenotype of interest from the seven diseases considered in our study. These targets contain registered therapeutics for other uses but have not registered for the phenotype of interest. We found 428 novel therapeutic targets accounting for almost 94% of the targets identified in the previous section. The remaining 24 targets either have therapeutics which are approved, in ongoing clinical trials or have been discontinued for the phenotype of interest (Table 3). We considered these 24 targets already containing drugs for the phenotype of interest (Table 3) as "true positives". Targets with currently registered therapeutics for the phenotype of interest which were not predicted by *Gentrepid* but already present in the search space as "false negatives". Genes which were not predicted and not targetable by drugs as "true negatives" and predicted novel therapeutic targets were considered as "false positives" (Table 4). Figure 6A, 6B and 6C show the individual number of novel targets obtained for each of the seven diseases from each drug database. The novelty of the predicted targets was assessed by calculating the ratio of the number of novel predicted targets to the number of targets predicted for each of the disease. The novelty ratio for all the diseases was between 1 and .92 (Table 2). We observed the highest ratio for CD (1.0) while the least for RA (0.92). The high ratio

of novel targets to predicted targets suggests repositioning could potentially have a large impact on clinical studies.

Identification of Novel Therapeutics

Furthermore, we attempted to identify novel drugs for our phenotype of interest. So, we compared our phenotype of interest (from the pool of seven diseases considered in our study) with phenotypes indications associated with the drug. In total, we retrieved ~7,252 drugs associated with human drug targets from all the drug databases. This resulted in retrieving 2,192 (~30%) unique drugs that target the 452 potential therapeutic targets.

As shown in Figure 4, maximum number of drugs were retrieved from DrugBank (~1,618) while the remaining were retrieved from TTD and PharmGKB (~735) and (91) respectively. Furthermore, T1D and CAD were predicted with the maximum number of novel therapeutics (Figure 7). Although CD was predicted with more number of novel targets, it had comparatively less number of novel therapeutics. BD had the least number of therapeutics as expected owing to the least number of therapeutic targets (Figure 7). In order to identify the novel drugs i.e. drugs not targeting our phenotype of interest, we filtered the above list of 2,192 drugs to retrieve 2,130 novel therapeutics. The total percentage of drugs that may be repositioned towards identified novel targets was estimated to be around ~30%of extracted drugs.

We identified both matches and mismatches between the current drug indication and the phenotype of our interest. The mismatches serve as the novel therapeutics whereas matches tend to relate to similar phenotypes. Table 3 shows 24 matches found in our study. For example, the drug "Aleglitazar" in phase II clinical trial for Diabetes Mellitus, Type 2 targets upon our predicted candidate gene named *PPARA* against Type II diabetes. It is seen that both the current phenotype associated with the drug and the phenotype of our interest are the same. Similar cases were observed with drugs like "Rosiglitazone" known to act upon target *PPARG* for diabetes mellitus, has a potential use in our phenotype of interest named Type I diabetes.

In case of mismatches (Table 4), we found novel therapeutics for the phenotype. For example, "Pirenzepine" is approved as a therapeutic drug for peptic ulcer disease which acts upon the *CHRM1* gene product. *CHRM1* is a predicted candidate gene for Type II diabetes, suggesting that the drug Pirenzepine may be repositioned as a novel therapeutic for Type II diabetes. Hence, the associated therapeutics for these novel targets may be repositioned against a phenotype of interest. This freely accessible easy identification of potential therapeutic targets, can accelerate the drug discovery process.

FDA-approved and clinical trial targets

We classified the predicted targets as FDA approved and clinical trial targets for seven complex diseases. An example depicted in figure 8 shows the comparison between T2D targets from TTD database and targets predicted by Gentrepid for T2D. Out of 84 targets predicted for T2D by Gentrepid (Table 2), 28 are from TTD (Figure 6A). Comparing these 28 targets with the 32 targets already present in TTD for T2D, we found three targets (*HSD11B1, PPARA, NR3C1*) targeted by drugs currently in clinical trials for T2D. In addition, *PPARA* is already targeted by FDA approved drugs. Hence, we predicted 25 novel drug targets from the TTD database for Type II diabetes. In total for seven diseases, we found 291 approved therapeutic targets and 95% of these as novel approved targets. We also found 334 clinical targets and 96% of these as novel clinical targets (Table 5). Both approved and clinical targets are potential drug targets however, approved targets will undoubtedly in the priority list for further experimental studies. To summarize, both approved and clinical novel targets are associated with therapeutics which can be repositioned as novel treatments towards cure of seven complex diseases.

Table 2. Table describing targetability index of seven diseases. Row abbreviations- PH- Phenotype of Interest; TT- Therapeutic Target; TI- Targetability Index; NTT-Novel Therapeutic Target; RTT-Replicated Therapeutic Target; T2D- Type 2 Diabetes; BD- Bipolar Disorder; CD- Crohn's Disease; HT-Hypertension; T1D- Type 1 Diabetes; CAD- Coronary Artery Disease; RA- Rheumatoid Arthritis; RN-Rank and NV- Novelty ratio

PH	TT	TI	RN	RTT	NTT	NV	RN
T2D	84	0.29	5	7	77	0.92	5
T1D	97	0.27	6	2	95	0.98	2
RA	77	0.38	2	6	71	0.92	5
HT	78	0.36	3	5	73	0.94	4
BD	59	0.27	6	1	58	0.98	2
CD	135	0.36	4	0	135	1	1
CAD	102	0.39	1	4	98	0.96	3

Validation of predicted therapeutic targets

To assess the validity of targets predicted by *Gentrepid* for each phenotype, we used two different benchmarks. In the first benchmark, the actual status of the gene was assessed by the existence or the non-existence of the abstract in literature citing both the gene name and the phenotype. In the second benchmark, the actual status of genes was based on whether they are designated as targets in the drug databases or not. This was repeated for all six search spaces investigated for each phenotype (Table1). The assessment was also based on Receiver Operation Characteristics (ROC) curves.

For the first benchmark, ROC curves for all the seven complex diseases were created by considering four thresholds for targets cited by at least one, five, ten or fifteen literature citations for the respective disease as true positives and targets without any citations or less than five, ten and fifteen citations as true negatives. Figure 9 contains all the ROC curves with Area Under Curve (AUC) values. The AUC values observed from the ROC curves were greater than or equal to 0.9 for each disease. The AUC values for all the seven diseases were obtained with 95% confidence interval and was significantly different from 0.5 (p - value .000) meaning that our results were significantly better than by chance. This suggests that our predictions of novel therapeutic targets for all the seven diseases are highly significant.

For the second benchmark, we performed a binary classification of genes in the six search spaces as "candidates" or "non-candidates". As described in Table 3, targets already containing drugs for the phenotype of interest considered as "true positives". Targets with currently registered therapeutics for the phenotype of interest which were not predicted by *Gentrepid* but already present in the search space as "false negatives". Genes which were not predicted and not targetable by drugs as "true negatives" and predicted novel therapeutic targets were considered as "false positives" (Table 6). This assessment was also based on ROC curves and AUC values considering six thresholds of search spaces described in table 1 from weakly significant (WS) candidate gene dataset (Figure 10). The AUC values observed from these ROC curves were also greater than or equal to 0.9 for each disease with 95% confidence interval. This also suggests that our predictions of novel therapeutic targets for all the seven diseases are highly significant.

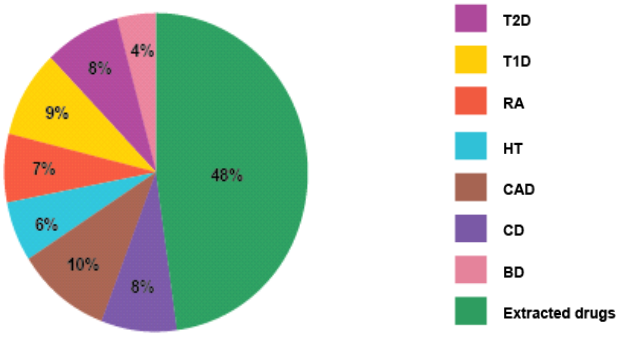


Figure 7. Predicted Therapeutics

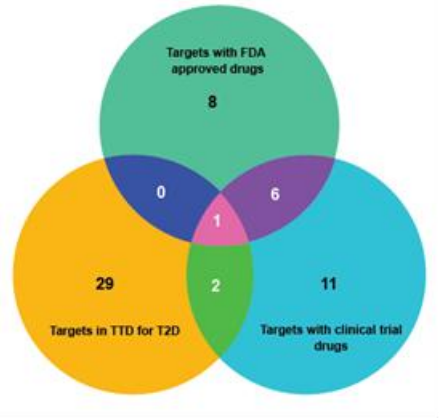


Figure 8. FDA approved and clinical trial targets for T2D

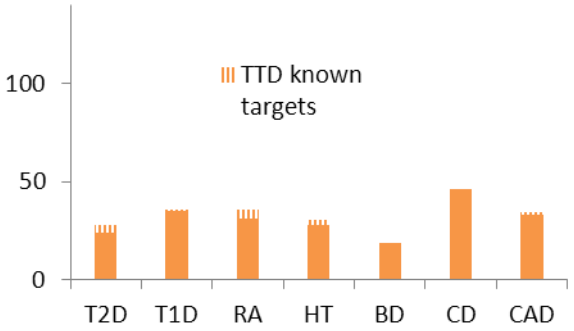


Figure 6A. Predicted known and novel therapeutic targets(TTD)

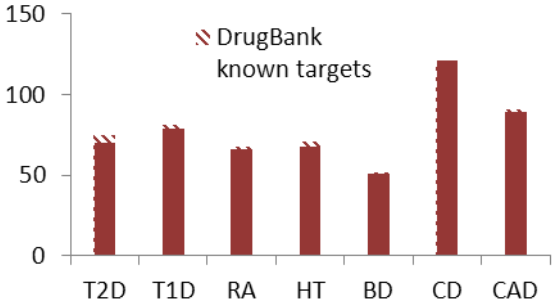


Figure 6B. Predicted known and novel therapeutic targets(DB)

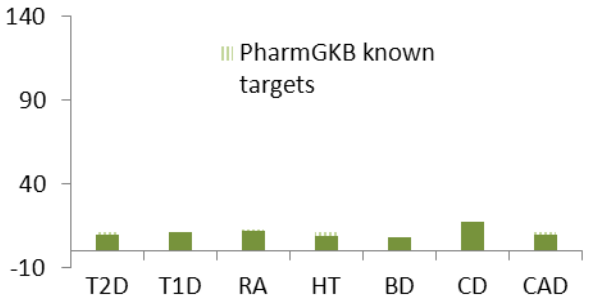
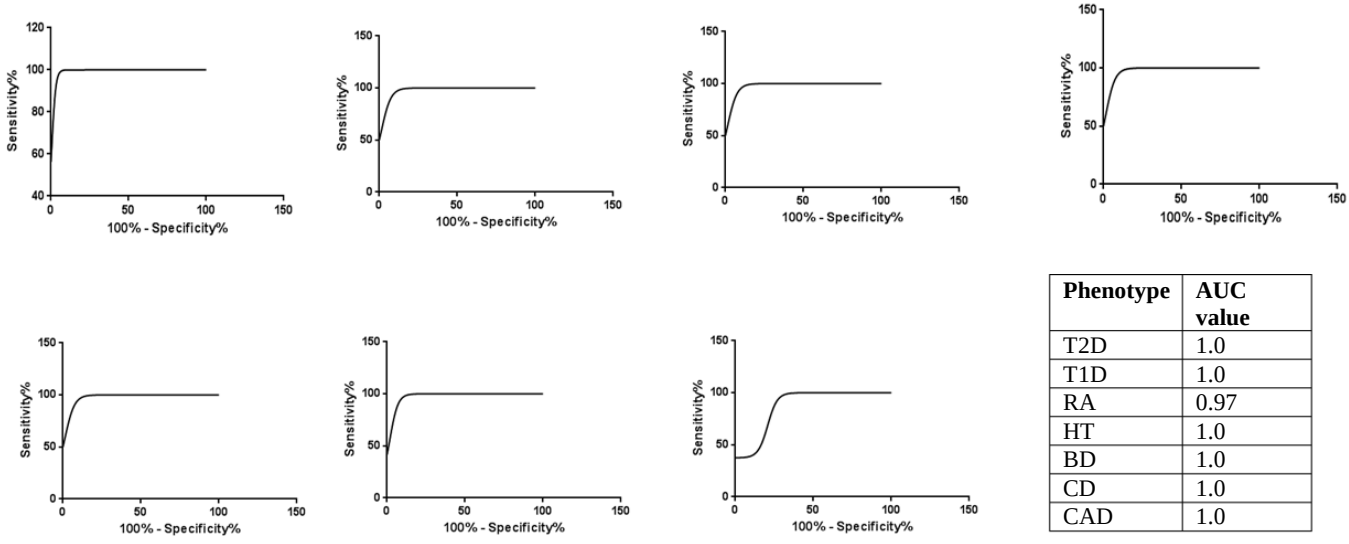


Figure 6C. Predicted known and novel therapeutic target (PharmGKB)



Phenotype	AUC value
T2D	1.0
T1D	1.0
RA	0.97
HT	1.0
BD	1.0
CD	1.0
CAD	1.0

Figure 9. ROC curve and AUC values for seven diseases based on Pubmed citations as benchmark

Table 3. Predicted known therapeutics - Row abbreviations- PH- Phenotype of Interest; T2D- Type 2 Diabetes; BD- Bipolar Disorder; CD- Crohn's Disease; HT- Hypertension; T1D- Type 1 Diabetes; CAD- Coronary Artery Disease and RA- Rheumatoid Arthritis

PH	Target	Drug name	Status	Database
T1D	*PPARG*	Rosiglitazone	Approved	TTD
	DGKA	Vitamin E	Approved	DrugBank
T2D	*CTSD*	Insulin recombinant	Approved	DrugBank
	PPARA	Aleglitazar	Phase II	TTD
	NR3C1	ISIS-GCCR	Preclinical	TTD
	TCF7L2	Repaglinide	Unknown	PharmGKB
	PPARD	Benzafibrate	Approved	DrugBank
	RB1	Insulin,procine	Approved	DrugBank
	HSD11B1	INCB13739	Phase IIa	TTD
RA	*TNF*	Infliximab	Approved	DrugBank
	ITGA4	R1295	Discontinued in phase I	TTD
	JAK2	INCB1824	Phase III	TTD
	IL15	AMG-714	Discontinued in phase I	TTD
	CCL2	MCP-1	Preclinical	TTD
	PRKCA	Vit E	Approved	DrugBank
HT	*DRD1*	Fenoldopam	Approved	TTD
	AGTR1	Valsartan	Approved	TTD
	CNR1	AZD1175	Approved	TTD
	AGT	Benazepril	Unknown	PharmGKB
	GUCY1A2	Isosorbide Mononitrate	Approved	DrugBank
BD	*SLC6A2*	Imipramine	Approved	DrugBank
CAD	*AGTR1*	Valsartan	Approved	DrugBank
	MYC	AVI4127	Phase I/II	TTD
	PLG	Abbokinase	Approved	DrugBank
	NOS3	ACCLAIM	Phase III	TTD

Table 4. Examples of novel therapeutics suitable for repositioning for the seven diseases.

PH	Target	Drug name	Status	Current Indication	Database
T1D	*RAR*	Alitretinoin	Approved	Kaposi's sarcoma	TTD
	GSK3B	Lithium	Unknown	Bipolar disorder	PharmGKB
T2D	*CHRM1*	Pirenzepine	Approved	Peptic ulcer disease	TTD
	LPL	Gemfibrozil	Approved	Hyperlipidemia	DrugBank
CAD	*FLT1*	Sorafenib	Launched	Advanced renal cell carcinoma	TTD
	KDR	Sunitinib	Launched	Advanced renal cell carcinoma	TTD
BD	*ESR1*	Trilostane	Approved	Cushing's syndrome	DrugBank
	ABCC1	Methotrexate	Unknown	Psoriasis	PharmGKB
HT	*TACR1*	GSK144814	Phase I	Schizophrenia	TTD
	NRP1	Palifermin	Approved	Oral mucositis	DrugBank
CD	*CRHR1*	CRF antagonist	Phase II completed	Irritable bowel syndrome	TTD
	INSR	Insulin detemir	Approved	Type I and II Diabetes	DrugBank
RA	*HLA-DRB1*	Glatiramer Acetate	Approved	Multiple sclerosis	TTD
	ACE	Ramipril	Approved	Hypertension	DrugBank

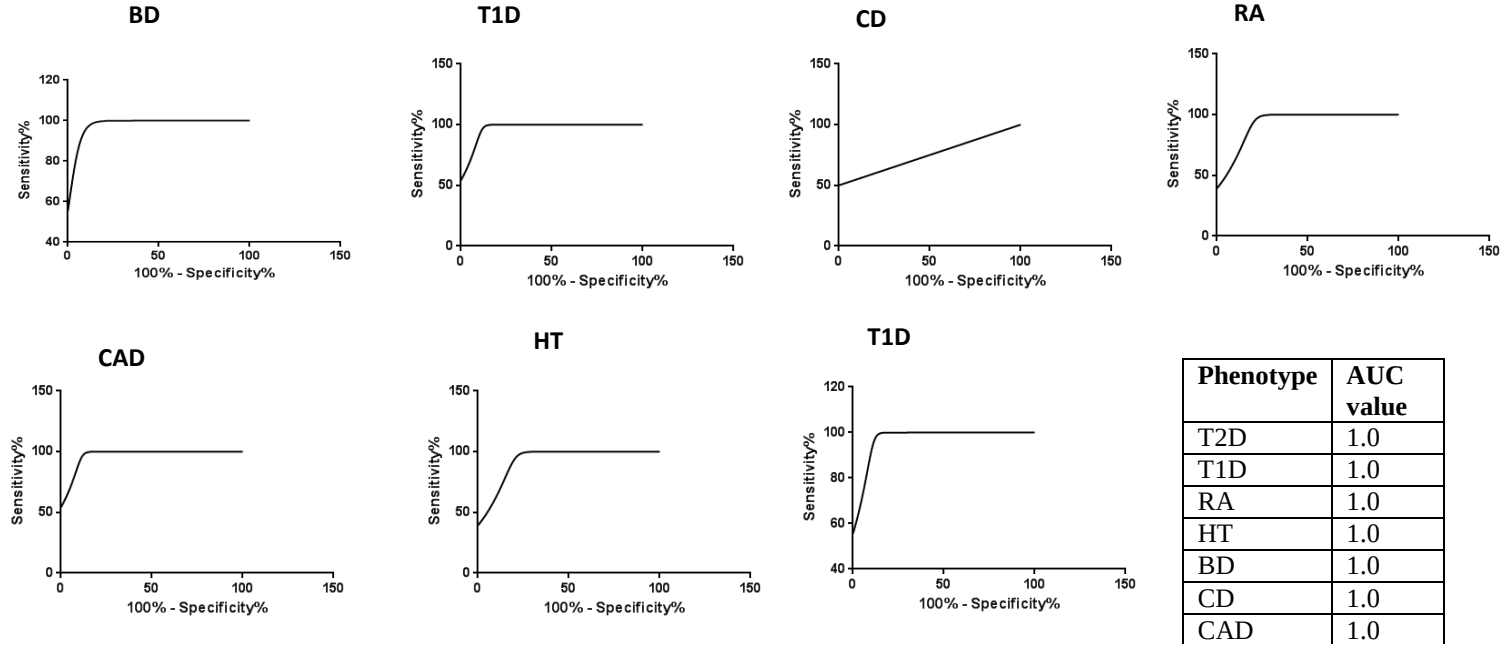
Table 6. A table describing binary classification genes and targets. Row abbreviations- TP - True Positives; FP - False Positives; TN - True Negatives, FN -False Negatives; TN True Negatives; Column abbreviation - PH - Phenotype

PH	Total genes in all search spaces	Binary classification	
		TP = 07	FP = 77
T2D	4,292	FN = 10	TN = 4,198
		TP = 02	FP = 95
T1D	5,340	FN = 10	TN = 5,233
		TP = 05	FP = 72
HT	8,427	FN = 32	TN = 8,318
		TP = 05	FP = 05
RA	4,970	FN = 10	TN = 4,815
		TP = 01	FP = 58
BD	5,667	FN = 08	TN = 5,605
		TP = 0	FP = 135
CD	5,644	FN = 02	TN = 5,512
		TP = 04	FP = 98
CAD	4,715	FN = 06	TN = 4,607

\\

Table 5. Table describing approved and clinical trial targets for seven complex diseases

Phenotype	Approved targets	Approved novel targets
T2D	45	41
T1D	57	55
HT	71	68
RA	55	53
CD	93	93
CAD	63	61
BD	37	36
Unique sum	291	277
Phenotype	**Clinical targets**	**Clinical novel targets**
T2D	65	62
T1D	73	72
HT	43	40
RA	59	54
CD	135	135
CAD	80	76
BD	44	44
Unique sum	334	318



Phenotype	AUC value
T2D	1.0
T1D	1.0
RA	1.0
HT	1.0
BD	1.0
CD	1.0
CAD	1.0

Figure 10. ROC curve and AUC values for seven diseases based on targets present in six search spaces obtained from weakly significant dataset

Evaluation of Predictions using Gene Function Annotations:

The Gene Ontology (GO) associations allow biologist to make inferences about group of genes instead of investigating each one individually. GO annotations describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species -independent manner. We used Gene Ontology functional annotations related to GO processes and molecular functions of predicted therapeutic targets. If the GO process of predicted target is similar to the biological process of respective disease, then it provides incidence of the close association between therapeutic target and the disease. Similarly, if GO annotation of molecular function for predicted target is similar to the function of molecular target of disease, it pinpoints strong relation between the target and the disease.

Gorilla [30] a gene enrichment analysis tool, was applied to identify enriched GO terms that appear for 452 potential therapeutic targets identified in our study for seven complex diseases. GOrilla is an interactive tool with running time of a few seconds (~7 seconds) using an efficient algorithm for computing the exact minimal hyper geometric (MHG) p -value, which circumvents the need for simulations, and an efficient software implementation. Gorilla [30], a gene ontology enrichment analysis tool, was applied to identify enriched GO terms that appear for 452 potential therapeutic targets identified in our study for seven complex diseases. We found that many of identified GO process terms are same or related to the biological processes of the seven diseases. The identified GO terms were evaluated with already available literature for the biological processes of seven diseases. Hypertension (HT) is a disease related to the nervous system [31] and most of the GO terms of predicted targets of HT are related to the nervous system which includes memory (GO:0007613), cognition (GO:0050890), learning or memory (GO:0007611), system process (GO:0003008), neurological system process (GO:0050877), regulation of ion homeostasis (GO:2000021), positive regulation of calcium ion transport into the cytosol (GO:0010524), and associative learning (GO:0008306). We also found some other GO terms such as carbohydrate transport (GO: 0008643), hexose transport (GO: 0008645), glucose transport (GO: 0015758), monosaccharide transport (GO: 0015749) which indicate that the predicted target for HT participate in multiple metabolic pathways. Bipolar disorder is a neuro- psychiatric diagnosis of mood disorder in which abnormalities occur in the structure and/or function of certain brain circuits [32]. In our study, predicted targets for Bipolar disorder (BD) were identified with the single GO process: response to stimulus (GO:0050896) which is a well-known biological process for neurological diseases [32]. Coronary artery disease is a common heart disease caused by plaque building up along the inner walls of the arteries of the heart. Predicted targets from Coronary artery disease (CAD) s how many known biological processes like cellular response to vascular endothelial growth factor stimulus (GO:0035924), regulation of cellular process (GO:0050794) and regulation of biological process (GO:0050789) as potential GO terms which are biological processes of CAD.

We also found some biological process related to new hallmarks of CAD, such as regulation of cell proliferation (GO:0042127), biological regulation (GO:0065007), response to stimulus (GO:0050896), negative regulation of developmental process (GO:0051093) [33]. Surprisingly, for Type 2 Diabetes which is a complex metabolic disorder of high blood glucose in the context of insulin resistance and relative insulin deficiency, we found only one GO term: response to radiation (GO:0009314). Rheumatoid arthritis (RA) is an autoimmune disease that results in a chronic, systemic inflammatory disorder that may affect many tissues and organs, but principally attack flexible (synovial) joints. For RA, we also found only one GO term: Organophosphate metabolic process (GO:0019637) which is related to inflammation, a well-known biological process for RA [34]. For Type1 Diabetes, a metabolic disorder characterized by absolute insulin deficiency due to destruction of islet cells in the pancreas. We found GO terms such as Response to alkaloid (GO: 0043279), drug (GO: 0042493) and stress (GO: 0006950) which are known biological processes. We also found some non-related terms such as negative regulation of transcription from RNA polymerase II promoter (GO:0000122), cellular response to stimulus (GO:0051716), cellular response to chemical stimulus (GO:0070887), response to chemical stimulus (GO:0042221), cellular developmental process (GO:0048869), responding to organic substance (GO:0010033), negative regulation of cell differentiation (GO:0045596), response to stress (GO:0006950), single-organism cellular process (GO:0044763), response to radiation (GO:0009314). For Crohn's Disease, a type of inflammatory bowel disease that may affect any part of the gastrointestinal tract. We found GO terms such as single-multicellular organism process (GO: 0044707), multicellular organismal process (GO:0032501) and GPCR signaling pathway (GO:0007186) [35]. These results also indicate that genes predicted as therapeutic targets for CAD, HT, CD and T1D are involved in multiple biological processes and pathways while targets for T2D, BD and RA are involved in the single biological process.

Molecular Function:

In addition, several types of molecules, such as signal transducers, trans-membrane receptors are often proposed as molecular targets in CAD [36] while neurotransmitter receptors HT [37] and transcription co-repressors and regulators in T1D [38]. We found that many receptors, and transcription factors related GO molecular function terms are enriched in predicted therapeutic targets for CAD, HT and T1D (Table 2). For example, neurotransmitter receptors (e.g. dopamine /serotonin receptors) are implicated in many neurological processes, including motivation, pleasure, cognition, memory, learning, fine motor control and modulation of neuroendocrine signalling. We did not retrieve any molecular function terms from GOrilla for BD, RA, CD and T2D. This might be because predicted targets for T2D, BD, RA and CD are involved in the limited number of metabolic processes and pathways. More basic research in required in this area.

Significance of the Work

The primary purpose of our work was to identify potential therapeutics and their targets by integrating publicly available genetic, bioinformatics and drug data. As the method involves repositioning of currently available drugs, it allows translational opportunities for drug testing [8]. Other bioinformatics tools which have been used to identify potential therapeutic targets for complex diseases and other diseases are designed to serve the purpose. For example, TARGET gene was used to identify and prioritize potential targets from hundreds of candidate genes for different types of cancer [39]. Another study identified potential drug targets for three neurological disorders- Alzheimer's disease, Parkinson's disease and Schizophrenia. This study involved the prediction of candidate genes using ToppGene and ToppNet prediction systems [22, 40]. The repositioning tools could be used as an initial screening tool for potential drugs which can be used for further evaluation [39]. It is important to note that not all repositioning opportunities will be successful as there are always some limitations.

Conclusions

There is a need to develop new approaches for the identification of therapeutic targets to accelerate the process of therapeutic discovery. In this study, our approach integrates detailed drug data with predicted candidate genes for seven complex diseases. This study enables people to efficiently identify possible novel therapeutic targets and alternative indication of existing therapeutics. We found 29% of predicted candidate genes as novel therapeutic targets from the candidate gene dataset and ~30% of drugs as novel therapeutics from the drug dataset for the seven complex diseases considered in our study. We have utilized both FDA approved drugs and drugs in clinical trials. Further investigation to verify action of these drugs is required for the discovery of drugs against potential targets. Hence, these drugs may be repositioned against seven phenotypes of interests. *Gentrepid*, thus can be utilized as a drug screening tool to save time and money spent on initial stages of drug discovery.

List of abbreviations:

GWAS: Genome-wide association studies; WTCCC: Wellcome Trust Case- Control Consortium; CPS: common pathway scanning; CMP: common module profiling; BD: Bipolar disorder; CAD: Coronary artery disease; CD: Crohn's disease; HT: Hypertension; RA: Rheumatoid arthritis; T1D: Type I diabetes; T2D: Type II diabetes; NN: Nearest neighbour approach; BY: Bystander approach; WS: Weakly significant set; MWS: Moderately-weak significant set; MHS: Moderately-high significant set; HS: Highly significant set, TTD: Therapeutic Target Database; PharmGKB: Pharmacogenomics Knowledge Base; DB: DrugBank; AUC: Area Under Curve; ROC: Receiver Operation Characterstics Curve.

Authors' contributions

MPG carried out the data mining and analysis, and worked on the design of the project. MAW conceived the study, participated in its design and reviewed the results from the data analysis. MPG, MAW, TMC, KAM and CDS helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Australian National Health and Medical Research Council [grant number 635512 to M.A.W].

References

1. Ashburn TT, Thor KB: **Drug repositioning: identifying and developing new uses for existing drugs**. *Nat Rev Drug Discov* 2004, **3**(8):673-683.
2. Pharmaceutical Research and Manufacturers of America: **PhRMA annual membership survey**. In. Washington, DC: PhRMA; 2012.
3. Lary J, Daniel K, Erickson J, Roberts H, Moore C: **The return of thalidomide: can birth defects be prevented?** *Drug Saf* 1999, **21**(3):161-169.
4. Chong CR, Sullivan DJ: **New uses for old drugs**. *Nature* 2007, **448**:645-646.
5. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, Bessarabova M: **Drug target prediction and repositioning using an integrated network-based approach**. *PLoS One* 2013, **8**(4):e60618.
6. Cooper RS: **Gene–Environment interactions and the etiology of common complex disease**. *Ann Intern Med* 2003, **139**(5_Part_2):437-440.
7. Altshuler D, Daly MJ, Lander ES: **Genetic mapping in human disease**. *Science* 2008, **322**(5903):881-888.
8. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V: **Use of genome-wide association studies for drug repositioning**. *Nat Biotechnol* 2012, **30**(4):317-320.
9. Ballouz S, Liu J, Oti M, Gaeta B, Fatkin D, Bahlo M, Wouters M: **Analysis of genome-wide association study data using the protein knowledge base**. *BMC Genet* 2011, **12**(1):98.
10. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS results: a review of statistical methods and recommendations for their application**. *Am J Hum Genet* 2010, **86**(1):6-22.
11. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.
12. Turner FS, Clutterbuck DR, Semple CAM: **POCUS: mining genomic sequence annotation to predict disease genes**. *Genome Biol* 2003, **4**(11):75-75.
13. Oti M, Ballouz S, Wouters MA: **Web tools for the prioritization of candidate disease genes**. *Methods Mol Biol* 2011, **760**:189-206.
14. Moreau Y, Tranchevent L-C: **Computational tools for prioritizing candidate genes: boosting disease gene discovery**. *Nat Rev Genet* 2012, **13**:523-536.
15. Teber E, Liu J, Ballouz S, Fatkin D, Wouters M: **Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies**. *BMC Bioinformatics* 2009, **10**(Suppl 1):S69.
16. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs**. *Nucleic Acids Res* 2011, **39**(suppl 1):D1035-D1041.
17. Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, Gong L, Owen R, Gong M, Gor W, Liu F, Truong C, Whaley R *et al*: **The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge**. *Nucleic Acids Res* 2008, **36**(suppl 1):D913-D918.
18. Zhu F, Shi Z, Qin C, Tao L, Liu X, Xu F, Zhang L, Song Y, Liu X, Zhang J: **Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery**. *Nucleic Acids Res* 2012, **40**(D1):D1128-D1136.
19. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction**. *Nucleic Acids Res* 2006, **34**(19):e130-e130.
20. Badano JL, Katsanis N: **Beyond Mendel: an evolving view of human genetic disease transmission**. *Nat Rev Genet* 2002, **3**(10):779-789.
21. van Driel MA, Cuelenaere K, Kemmeren PPCW, Leunissen JAM, Brunner HG, Vriend G: **GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases**. *Nucleic Acids Res* 2005, **33**(suppl_2):W758-761.
22. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene suite for gene list enrichment analysis and candidate gene prioritization**. *Nucleic Acids Res* 2009, **37**(suppl 2):W305-W311.
23. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y: **ENDEAVOUR update: a web resource for gene prioritization in multiple species**. *Nucleic Acids Res* 2008, **36**(suppl 2):W377-W384.
24. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes**. *Nature* 2001, **409**(6822):853-855.
25. Reimand J, Arak T, Vilo J: **g: Profiler—a web server for functional interpretation of gene lists (2011 update)**. *Nucleic Acids Res* 2011, **39**(suppl 2):W307-W315.
26. Overington JP, Al-Lazikani B, Hopkins AL: **How many drug targets are there?** *Nat Rev Drug Discov* 2006, **5**(12):993-996.
27. Hopkins AL, Groom CR: **The druggable genome**. *Nat Rev Drug Discov* 2002, **1**(9):727-730.
28. Sakharkar MK, Sakharkar KR, Pervaiz S: **Druggability of human disease genes**. *Int J Biochem Cell Biol* 2007, **39**(6):1156.
29. Zambrowicz BP, Sands AT: **Knockouts model the 100 best-selling drugs[mdash]will they model the next 100?** *Nat Rev Drug Discov* 2003, **2**(1):38-51.

30. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. *BMC Bioinformatics* 2009, **10**(1):48.

31. Raz N, Rodrigue KM , Acker JD: **Hypertension and the brain: vulnerability of the prefrontal regions and executive functions**. *Behav Neurosci* 2003, **117**(6):1169-1180.

32. Clark L, Goodwin GM : **S tate-and trait-related deficits in sustained attention in bipolar disorder**. *Eur Arch Psychiatry Clin Neurosci* 2004, **254**(2):61-68.

33. Vasa M , Fichtlscherer S, Aicher A, Adler K, Urbich C, M artin H, Zeiher AM , Dimmeler S: **Number and migratory activity of circulating endothelial progenitor cells inversely correlate with ri sk factors for coronary artery disease**. *Circ Res* 2001, **89**(1):e1-e7.

34. Weissmann G: **The role of lysosomes in inflammation and disease** . *Annu Rev Med* 1967, **18**(1):97-112.

35. Wehkamp J, Stange Ef: **A New Look at Crohn's Disease** . *Ann N Y Acad Sci* 2006, **1072**(1):321-331.

36. Schächinger V, Britten M B, Zeiher AM : **Prognostic impact of coronary vasodilator dysfunction on adverse long-term outcome of coronary heart disease** . *Circulation* 2000, **101**(16):1899-1906.

37. Hoyer D, Hannon JP, M artin GR: **Molecular, pharmacological and functional diversity of 5-HT receptors**. *Pharmacol Biochem Behav* 2002, **71**(4):533-554.

38. Habener J, Stoffers D: **A newly discovered role of transcription factors involved in pancreas development and the pathogenesis of diabetes mellitus**. *Proc Assoc Am Physicians* 1998, **110**(1):12.

39. Wu CC, D'Argenio D, Asgharzadeh S, Triche T: **TARGETgene: a tool for identification of potential therapeutic targets in cancer**. *PLoS One* 2012, **7**(8):e43305.

40. Kaimal V, Sardana D, Bardes EE, Gudivada RC, Chen J, Jegga AG: **Integrative systems biology approaches to identify and prioritize disease and drug candidate genes**. *Methods Mol Biol* 2011, **700**:241.