Journal of Biomedical Informatics 56 (2015) 348–355

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

PhenoPredict: A disease phenome-wide drug repositioning approach towards schizophrenia drug discovery

Rong Xu [a], QuanQiu Wang [b]

[a] Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, United States
[b] ThinTek, LLC, Palo Alto, CA 94306, United States

ARTICLE INFO

Article history:
Received 8 October 2014
Revised 26 June 2015
Accepted 29 June 2015
Available online 4 July 2015

Keywords:
Drug repositioning
Drug discovery
Systems biology
Disease phenotype
Schizophrenia

ABSTRACT

Schizophrenia (SCZ) is a common complex disorder with poorly understood mechanisms and no effective drug treatments. Despite the high prevalence and vast unmet medical need represented by the disease, many drug companies have moved away from the development of drugs for SCZ. Therefore, alternative strategies are needed for the discovery of truly innovative drug treatments for SCZ. Here, we present a disease phenome-driven computational drug repositioning approach for SCZ. We developed a novel drug repositioning system, PhenoPredict, by inferring drug treatments for SCZ from diseases that are phenotypically related to SCZ. The key to PhenoPredict is the availability of a comprehensive drug treatment knowledge base that we recently constructed. PhenoPredict retrieved all 18 FDA-approved SCZ drugs and ranked them highly (recall = 1.0, and average ranking of 8.49%). When compared to PREDICT, one of the most comprehensive drug repositioning systems currently available, in novel predictions, PhenoPredict represented clear improvements over PREDICT in Precision-Recall (PR) curves, with a significant 98.8% improvement in the area under curve (AUC) of the PR curves. In addition, we discovered many drug candidates with mechanisms of action fundamentally different from traditional antipsychotics, some of which had published literature evidence indicating their treatment benefits in SCZ patients. In summary, although the fundamental pathophysiological mechanisms of SCZ remain unknown, integrated systems approaches to studying phenotypic connections among diseases may facilitate the discovery of innovative SCZ drugs.

1. Introduction

Mental illness causes enormous personal and societal burdens [1]. In fact, mental illnesses, such as schizophrenia (SCZ), bipolar disease, depression and other psychiatric disorders, is the leading cause of impairment and disability in the United States and world-wide, accounting for around one-third of the disabilities in the world [2,3]. SCZ is arguably the most intractable among all psychiatric disorders [4]. SCZ has a life-time prevalence of 1%, typically beginning before age 25 years and persisting throughout the life of the individual [3]. Currently, effective drugs do not exist for treating SCZ [5]. Despite the vast unmet medical need, many drug companies have moved away from SCZ drug development, in part because of the high costs, high failure rates in clinical trials, lengthy development processes, and a poor understanding of underlying mechanisms of the disease [5–7].

In this study, we present a computational drug repositioning approach towards discovering innovative drug candidates for the treatment of SCZ. Psychiatric drug discovery has traditionally come from repositioning existing drugs based on serendipitous clinical observations [8]. For example, lithium, originally approved as a sedating agent, is now used in the treatment of mania [9]. Chlorpromazine, originally approved as an antihistamine, is now used in the treatment of schizophrenia [10]. Iproniazid, originally approved as an anti-tuberculosis agent, is now used in the treatment of depression [11]. Ketamine, originally approved as an anesthetic agent, has rapid antidepressant effects in patients with major depression [12]. Computation-based repositioning approaches that automatically reason over vast amounts of genetic, genomic, chemical, and phenotypic data can greatly speed up the timeline of traditional serendipity-based psychiatric drug discovery process and facilitate the identification of truly innovative drug treatments for SCZ and other psychiatric disorders [13–15]. However, computational drug repositioning approaches for identifying novel drug candidates for SCZ has not been fully explored.

Computational drug repositioning approaches can be classified as either drug-based or disease-based [14,15]. Drug-based

E-mail addresses: rxx@case.edu (R. Xu), qwang@thintek.com (Q. Wang)

http://dx.doi.org/10.1016/j.jbi.2015.06.027
1532-0464/

approaches leverage on the known molecular structures or functions of drugs, such as chemical structures and properties, molecular docking, gene expression, drug treatment indications, and drug side effects [16–24]. In the past 50 years, psychiatric drug discovery has been largely drug-based and has focused on identifying molecules with which existing drugs interact. Consequently, all current antidepressants, antipsychotics, and anti-anxiety drugs developed and marketed from the 1950s to the current day have targeted the same molecular pathways in the brain as their prototypes [5]. it has been recognized that drug-based discovery, with its focus on finding drug candidates based on existing drugs, might by definition fail to identify new therapeutic mechanisms [25]. An alternative approach is disease-based discovery, which puts less emphasis on existing drugs and focuses more on disease mechanisms and interrelationships. Because disease-based approaches look for similarities and interrelationships among diseases, these approaches are able to identify innovative drugs. Compared to drug-based repositioning approaches, disease-based approaches are surprisingly less explored and mainly used disease gene expression data [19,20].

We hypothesize that higher-level phenotypic overlaps among diseases reflect underlying biological commonalities and that insights from one disease may be used to inform our developing knowledge of others. We developed a phenotype-driven drug repositioning system, PhenoPredict, to exploit drug repositioning opportunities rendered by disease phenotype data captured in the Human Phenotype Ontology (HPO) and a comprehensive drug-disease treatment relationship knowledge base (TreatKB) that we recently constructed [26–28]. HPO is a standardized vocabulary of phenotypic abnormalities encountered in human disease [29]. HPO contains phenotypic descriptions of 7529 diseases, the majority of which were derived from the Online Mendelian Inheritance in Man (OMIM) [30]. Studies of phenotypic abnormalities in HPO have advanced our understanding of the genetic bases of diseases [31–33]. In a recent study, Gottlieb et al. used disease phenotypic similarities defined in HPO and drug-drug similarities from other databases to construct a classifier (PREDICT) and then used it to determine treatment associations between 593 drugs and 313 diseases, including SCZ [34]. Different from PREDICT, PhenoPredict used a network-based approach to systematically exploit phenotypic interrelationships among diseases as defined in HPO. More importantly, PhenoPredict used a novel drug prioritization algorithm to exploit treatment connections among diseases as defined in TreatKB, which is a key component of PhenoPredict. Compared to PREDICT, our study included significantly more drugs and diseases (2482 drugs and 24,511 unique disease concepts). We compared PhenoPredict to PREDICT in novel drug predictions using multiple evaluation datasets and demonstrated that PhenoPredict achieved consistently better performances.

2. Data and methods

The experiment framework for PhenoPredict is depicted in Fig. 1 and consists of four phases: (1) We constructed a phenotypic disease network (PDN) using disease-disease similarity measures from HPO. We then developed a network-based ranking algorithm to find diseases that are phenotypically related to SCZ; (2) In order to validate the network construction and ranking algorithms of PhenoPredict and to better understand SCZ-related diseases, we analyzed disease class distributions among diseases at different ranking cutoffs and investigated what kinds of diseases were enriched among top-ranked SCZ-related diseases; (3) We developed a novel drug repositioning algorithm to systematically identify drug repositioning candidates from SCZ-related diseases. We evaluated PhenoPredict using FDA-approved SCZ drugs. We compared PhenoPredict to PREDICT in novel predictions; and (4) In order to better understand top-ranked drug candidates, we examined drug class distributions among both top- and intermediate-ranked drug candidates.

2.1. Construct the phenotypic disease network (PDN) and find SCZ-related diseases from PDN

2.1.1. Construct phenotypic disease network (PDN)

PDN was constructed by directly using the disease-disease similarity matrix obtained from HPO. In HPO, individual diseases are often associated with multiple phenotypic terms. Similarity measures for any two given phenotypic terms were calculated based upon shared information content (frequency among annotations of all diseases) in the set of their common-ancestor nodes. The similarity between two diseases was then calculated by matching each phenotypic term of one disease with the most similar term of the other disease; the average was taken over all pairs

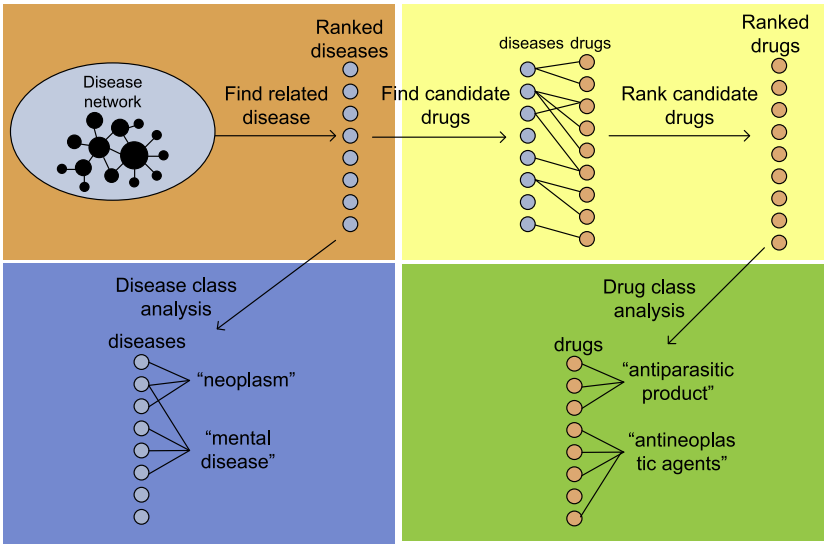

Fig. 1. The overall experimental flow chart for PhenoPredict.

Table 1
Sixteen disease chapters (classes) and the number of diseases (synonym expanded) in each chapter.

Disease class	Diseases (n)	Disease classes	Diseases (n)
Certain infectious and parasitic diseases	11,598	Diseases of the circulatory system	5544
Neoplasms	14,158	Diseases of the respiratory system	3156
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	3264	Diseases of the digestive system	5960
Endocrine, nutritional and metabolic diseases	5438	Diseases of the skin and subcutaneous tissue	4390
Mental and behavioural disorders	6162	Diseases of the musculoskeletal system and connective tissue	11,520
Diseases of the nervous system	5258	Diseases of the genitourinary system	5247
Diseases of the eye and adnexa	3735	Congenital malformations, deformations and chromosomal abnormalities	9064
Diseases of the ear and mastoid process	1815	Certain conditions originating in the perinatal period	3454

of phenotypic terms [29]. We downloaded the HPO disease-disease similarity matrix and mapped disease terms to the Unified Medical Language System (UMLS) [35] Concept Unique Identifiers (CUIs) in order to facilitate the subsequent linking to drug-disease treatment pairs in TreatKBs that were constructed from other data sources. A total of 5708 out of 7529 disease terms in HPO were mapped to UMLS CUIs. Instead of excluding unmapped terms, we used the term names as their unique identifiers. In total, we obtained 17,523,509 disease-disease pairs, representing 7210 unique disease concepts. The similarity scores from the matrix were used as the edge weights of PDN. We also generated ten random PDNs by randomly shuffling edges of the real PDN.

2.1.2. Develop network-based ranking algorithm for finding SCZ-related diseases

Recently, we developed network-based approaches to prioritize genes for a given disease [36] and to prioritize diseases for a given microbial metabolite [37]. In this study, we applied these network-based algorithms to prioritize diseases for SCZ. The iterative network-based ranking algorithm is defined as: $p^{t+1} = (1-r)Mp^t + rp^0$, wherein M is the column-normalized adjacency matrix of PDN, γ is a preset probability of restarting from the initial seed node ($\gamma = 0.1$ in this study), and p^t is a vector in which the i_{th} element holds the normalized ranking score of disease i at t_{th} iteration. The initial probability vector p^0 contains normalized probability values for input. In our study, p^0 contains SCZ, with a probability of 1.0. Diseases are ranked according to values in the steady-state probability vector, which is obtained by iterating the algorithm until the change between p^{t+1} and p^t is less than 10^{-6}.

2.2. Analyze disease class distribution at different ranking cutoffs

To better understand ranked diseases, we analyzed disease class distribution at ten different ranking cutoffs. Using SCZ as the seed, we retrieved a ranked list of 7204 diseases from PDN. We classified these diseases into sixteen categories using the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10), a disease classification scheme designated by the World Health Organization (WHO) [38]. The ICD10 includes 22 highest-level disease classes (or chapters) such as "Neoplasms" and "Diseases of the nervous system". We used sixteen chapters and excluded the other six non-specific disease classes such as "Codes for special purposes" and "Injury, poisoning and certain other consequences of external causes". Because the terms used in ICD10 may differ from those in PDN, we expanded disease terms in ICD10 to their synonyms through UMLS CUIs. Disease chapters and the numbers of diseases in each chapter are listed in Table 1.

At ten ranking cutoffs (10%, 20%, ..., 100%), we calculated percentages of these sixteen disease classes among retrieved diseases. For example, at the 100% cut-off (all 7204 retrieved diseases), 3.89% of the diseases were classified as "Mental, behavioural disorders". At the 10% cutoff (top 720 diseases), 87 out of the 720 diseases (12.05%) were classified as "Mental, behavioural disorders," representing a 209.8% increase as compared to the 100% cut-off ((12.05–3.89)/3.89 = 209.8%). This means that top-ranked diseases on average included more "Mental, behavioural disorders" than lower-ranked diseases. While this was expected and demonstrates the validity of the disease ranking algorithm, we found that certain other disease classes such as "Endocrine, nutritional and metabolic diseases" were enriched among top-ranked diseases.

2.3. Reposition drugs

2.3.1. Drug repositioning algorithm

We developed an approach to systematically identify drug repositioning candidates from SCZ-related diseases. We ranked drugs based on the number of SCZ-related diseases that they are currently approved to treat as well as the ranking scores of these diseases. For example, if drug 1 treats 25 top-ranked diseases, it would be ranked higher than drug 2, which treats only one or two lower-ranked diseases. The drug ranking algorithm is defined as: $R_{drug} = \sum_{i=1}^{n} R\_disease\_i$, wherein n is the number of SCZ-related diseases that are currently approved to treat and $R\_disease\_i$ is the disease ranking score (output from the network-based disease ranking algorithm). During the experiment, we found that certain drugs were consistently ranked highly for both the real PDN and random PDNs. For example, the drug "chlordiazepoxide" was ranked at top 0.32% for the real PDN and on average at top 0.36% for andom PDNs. We designed our reprioritization strategy by accounting for rankings of a drug for random PDNs. A drug was ranked highly if and only if it was ranked highly based on the real PDN and the ratio of its ranking for the real PDN to that for random PDNs is at least 2-fold.

2.3.2. Comparison of four TreatKBs in a de novo validation setting using 18 known SCZ drugs as evaluation dataset

In order to systematically reposition drug treatments from one disease to another, it is critical to have a comprehensive drug treatment knowledge base. In our recent studies, we constructed four large-scale drug-disease treatment knowledge bases (TreatKBs) from multiple heterogeneous and complementary data sources using advanced computational techniques including natural language processing, text mining, and data mining [26–28]. The databases included 9216 drug-disease treatment pairs extracted from FDA drug labels, 111,862 pairs extracted from the FDA Adverse Event Reporting System (FAERS), a database supporting the FDA's post-marketing drug safety surveillance efforts, 34,306 pairs extracted from 22 million published biomedical literature abstracts, and 69,724 pairs extracted from 171,805 clinical trials. The combined TreatKB consists of 208,330 unique drug-disease treatment pairs, representing 2484 drugs and 24,511 unique disease concepts.

We evaluated PhenoPredict using all 18 FDA-approved SCZ drugs by comparing its performance across four TreatKBs. Since SCZ and its associated drug treatment pairs were removed from

the inputs to the repositioning algorithm (SCZ-related diseases and drug-disease treatment pairs), the evaluation is in fact a *de novo* validation. We calculated the rankings of the 18 FDA-approved SCZ drugs among all retrieved drugs and used them as our gold standard. We assumed that the higher these gold standard drugs were ranked, the better the ranking algorithm was. We compared the performances (recall and average rankings) across four TreatKBs separately and in combination.

2.3.3. Compare PhenoPredict to PREDICT in novel predictions

We compared PhenoPredict to PREDICT in novel predictions using three evaluation datasets: (1) 195 drugs that had been tested in SCZ clinical trials; (2) 50 drugs that were in ongoing SCZ clinical trials initiated in 2012 and after. These drugs may represent newer SCZ drugs; and (3) 114 drugs that the literature implies have been used to treat varying symptoms of SCZ. These three evaluation datasets were derived from TreatKBs, which was constructed from multiple data resources including 22 million published biomedical literature abstracts and 171,805 clinical trials [26–28]. The 18 FDA-approved drugs were removed from these three evaluation datasets. Note that all SCZ-related drug treatment information were removed from TreatKB before PhenoPredict made predictions for SCZ.

We used Precision-Recall (PR) curves instead of Receiver Operator Characteristic (ROC) curves to evaluate and compare PhenoPredict to PREDICT. PR curves are often used to evaluate ranked classification results in information retrieval [39]. ROC curves are commonly used to evaluate binary classification problems in machine learning and data mining [40]. A PR space is defined as precision (fraction of examples classified as positive that are truly positive) and recall (true positive rate) as x and y axes, respectively. An ROC space is defined by FPR (false positive rate) and TPR (the same as recall) as x and y axes, respectively. Studies have shown that in domains where the number of negatives greatly exceeds the number of positives, such as in drug repositioning and most other biomedical classification domains, ROC curves can present an overly optimistic view of an algorithm's performance as compared to PR curves [41,42]. Davis et al. proved that a curve dominates in ROC space if and only if it dominates in PR space and algorithms that optimize the ROC curve are not guaranteed to optimize the PR curve [42]. Therefore, in our study, we used PR curves even though most biomedical classification studies use ROC curves.

PREDICT utilizes multiple drug-drug and disease-disease similarity measures for the prediction task [34]. PREDICT first trains a logistic regression classifier using known drug-disease associations. It then classifies additional drug-disease associations based on their similarity to the known associations. We compared PhenoPredict to PREDICT in novel predictions. A total of 593 drugs were included in PREDICT, among which 79 drugs were classified as positives for SCZ. The 79 drugs along with their corresponding probabilities (ranging from 0.543 to 0.994) are publicly available [34]. The remaining 524 drugs were predicted by PREDICT as negatives for treating SCZ. We assigned each negative prediction a value that was randomly picked from 0.0 to 0.499. We repeated this process of assigning values to negatives for ten iterations and generated ten datasets for PREDICT. PR curves for these ten datasets were similar, therefore we did not generate more datasets for PREDICT. The PR curves for PREDICT were then averaged across the ten datasets that we generated. The output from PhenoPredict is a ranked list of 2484 drugs. Using each of the three evaluation datasets as gold standard, we calculated precisions at 10 different recall cutoffs (0.1, 0.2, . . . , 1.0) for both PhenoPredict and PREDICT and plotted the PR curves. The area under curves (AUC) was used to compare the two approaches.

2.4. Analyze repositioned drug candidates

It is important to the current study, as well as to future work in the fields of computational drug repositioning, to better understand the nature of identified drug repositioning candidates. In order to facilitate such an understanding, we examined the class distributions of drug repositioning candidates. Drug classes were defined by the Anatomical Therapeutic Chemical (ATC) classification system [43]. The ATC system consists of 13 first-level codes, 94 s-level codes, 267 third-level codes, 882 fourth-level codes, and 4580 fifth-level codes. The fifth-level codes are individual drugs. In our study, we used the third level ATC codes for the analysis. We examined top ranked drug classes for drug candidates ranked in the range of top 0–15% and in the range of top16–30% separately.

3. Results

3.1. Disease class analysis

Using SCZ as the seed, we retrieved a ranked list of 7204 diseases from PDN. We calculated percentages of sixteen disease classes among these retrieved diseases at ten different ranking cutoffs (10%, 20%, . . . , 100%). Among the sixteen disease classes, three disease classes were enriched among top-ranked diseases: "Mental, behavioural disorders", "Diseases of the nervous system", and "Endocrine, nutritional and metabolic diseases" (Fig. 2). The increase for the disease class "Mental, behavioural disorders" was particularly pronounced, with a 209.8% increase for the top 10% diseases as compared to all retrieved diseases. The increases for the other two classes are similar but less prominent. In summary, the enrichment of "Mental, behavioural disorders", to which SCZ belongs, among top-ranked diseases, demonstrated the validity of our phenotype-driven network-based disease ranking algorithm.

3.2. FDA-approved SCZ drugs were ranked highly

When the TreatKB containing only FDA-approved drug-disease treatments was used, PhenoPredict achieved a recall of 0.33 and an average ranking of 30.9%. When the other three TreatKBs were

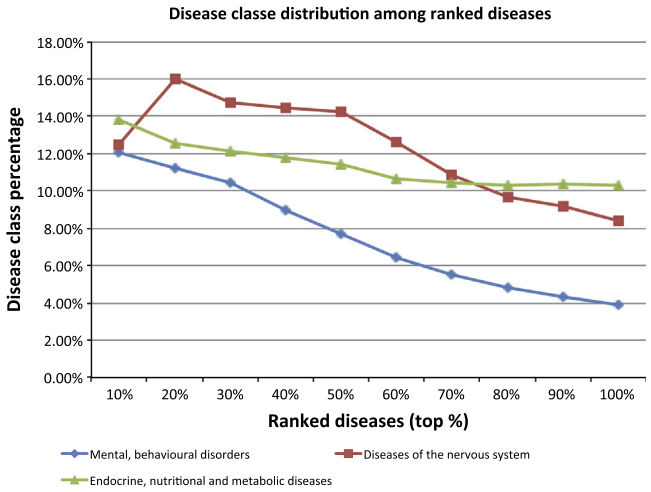

Fig. 2. Percentages of three disease classes among 7204 diseases retrieved from HPON at ten ranking cutoffs (top 10%, 20%, . . . , 100% (all diseases)). For example, ranked diseases at top 10% cutoff (top 720 diseases) contain 12.05% diseases from the class "Mental, behavioural disorders". Thirteen unenriched disease classes are not shown.

Table 2
Comparing recalls and average rankings of 18 FDA-approved SCZ drugs for four TreatKBs individually and combined.

TreatKB	Recall	Average ranking (%)
FDA-approved	0.33	30.9
Post-market	1.00	10.48
ClinicalTrials	0.67	21.65
Literature	0.83	10.97
Combined	**1.00**	**8.49**

used, PhenoPredict achieved a significantly better performance in terms of both recalls and rankings (Table 2). Significantly, when all fourTreatKB were combined, PhenoPredict achieved a recall of 1.00 and an average ranking of 8.49%. These results demonstrate that a comprehensive TreatKB is critical component of PhenoPredict.

3.3. PhenoPredict performed better than PREDICT in novel predictions

We plotted PR curves for PhenoPredict and PREDICT using the 195 drugs extracted from SCZ clinical trials as the evaluation set. The PR curve for PhenoPredict clearly dominates that for PREDICT. The area under the curve (AUC) for PhenoPredict is 0.489, representing a 98.8% improvement as compared to the AUC of 0.246 for PREDICT (Fig. 3).

When evaluated with 50 drugs extracted from ongoing SCZ clinical trials, PhenoPredic achieved an AUC of 0.128, representing an 81.1% improvement as compared to the AUC of 0.071 for PREDICT (Fig. 4).

The PR curves determined using the 114 drugs that the literature implies have been used to treat varying symptoms of SCZ as the evaluation set are shown in Fig. 5. PhenoPredict achieved of an AUC of 0.289, representing a 41.2% improvement as compared to the AUC of 0.208 for PREDICT. In summary, PhenoPredict consistently showed improved PR curves compared to those for PREDICT across three different evaluation datasets.

Table 3 shows the top 20 repositioned drug candidates, all of which are implicated as promising candidates through evidence from sources other than our experiment, such as FDA drug labels, clinical trials, or biomedical literature. Among these 20 drugs, 8 are FDA-approved drugs. These specific examples further demonstrate the potential of PhenoPredict in identifying promising drug repositioning candidates for SCZ.

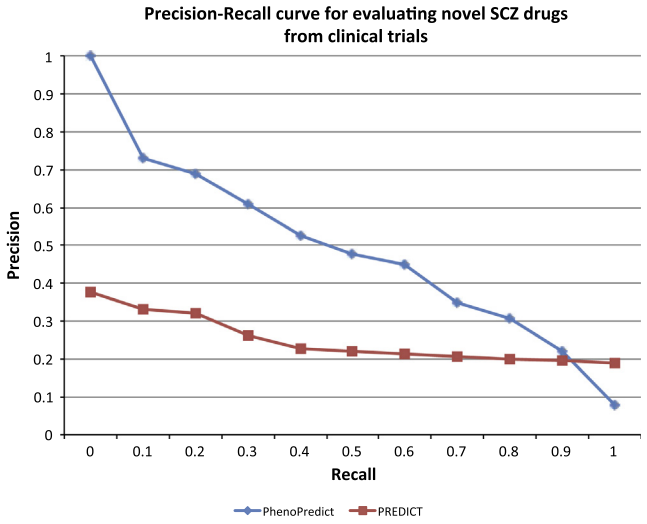

Fig. 3. Precision-Recall curves for PhenoPredict and PREDICT using 195 drugs from SCZ clinical trials as gold standard.

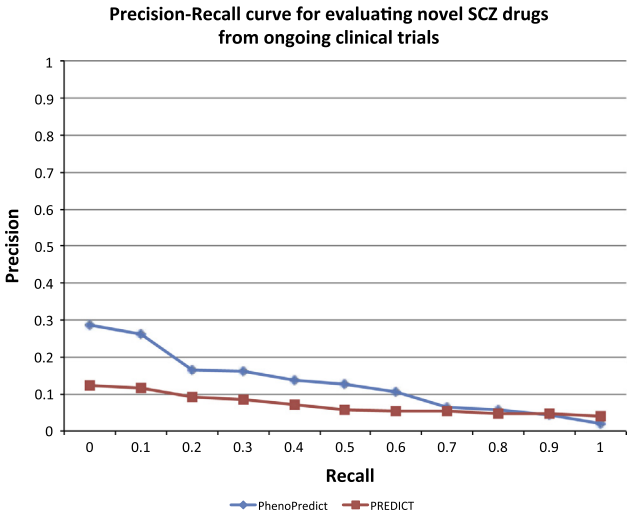

Fig. 4. Precision-Recall curves for PhenoPredict and PREDICT using 50 drugs from ongoing SCZ clinical trials as gold standard.

3.4. Analysis of repositioned drug candidates offers insights to common mechanisms of action

The top drug candidates, those ranked in the 0–15% range, were associated with a total of 95 third-level ATC codes. Fig. 6 shows the top 15 drug classes, among which 13 classes are related to antipsychotics, including antidepressants, antiepileptics, and dopaminergic agents. We have shown in the disease class analysis that mental diseases were highly enriched among top-ranked diseases; therefore, it is not surprising that most of the top ranked drug candidates are typical antipsychotics. This result also demonstrates that common pathophysiologic mechanisms are shared among phenotypically related psychiatric disorders and that traditional psychiatric drug discovery may have fully exploited this commonality (i.e. the same drugs are used among related diseases).

While top ranked drugs are mainly antipsychotics, drugs with intermediate rankings may provide opportunities for discovering innovative drugs. Fig. 7 shows the top 15 drug ATC codes for drugs ranked in the range of 16–30%. The majority of these top ATC codes are not related to antipsychotics. Evidence gleaned from the published biomedical literature shows that these drug classes may have treatment potential in SCZ patients. For example, two ATC codes "immunodepressants" and "antiinflammatory and antirheumatic" were ranked highly. Studies have shown that immune dysfunction and inflammation are involved in patients with SCZ [44,45]. Therefore, anti-inflammatory drugs may represent promising treatments for SCZ. In a randomized controlled study, celecoxib, a widely used anti-inflammatory agent, was shown to improve symptoms experienced by SCZ patients without major side effects [46]. Recent genetic findings from genome-wide association studies (GWAS) also point to possible common genetic connections between SCZ and immune disorders [47]. Beta-blockers were also ranked highly. Beta blockers are commonly used to treat hypertension and cardiovascular diseases. Studies have shown that they may reduce anxiety and extrapyramidal symptoms in SCZ and have been suggested as adjunctive therapies to antipsychotics in SCZ or similar severe mental disorders [48,49]. The output of PhenoPredict also suggests the potential use of angiotensin antagonists as an atypical SCZ treatment. Angiotensin antagonists are primarily used in the treatment of hypertension, congestive heart failure, and heart attacks.

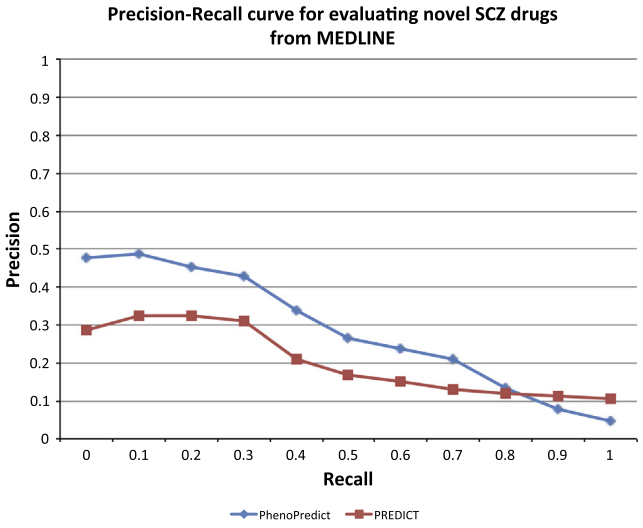

Fig. 5. Precision-Recall curves for PhenoPredict and PREDICT using 114 drugs extracted from biomedical literature (MEDLINE) as gold standard.

Interestingly, angiotensin has been shown to regulate the central nervous system activity [50,51]. Neurochemical and anecdotal reports suggest that angiotensin antagonists may have mood-elevating and cognitive enhancing functions in patients, however mechanisms of actions by which these inhibitors modify cognitive performance remain unknown [52].

4. Discussion

We developed a drug repositioning system, PhenoPredict, to exploit the phenotypic connections among diseases and applied it to identify drug repositioning candidates for the treatment of SCZ. PhenoPredict ranked many traditional antipsychotic drugs highly, demonstrating the validity of the algorithms. In addition, we discovered many drug repositioning candidates with mechanisms of action fundamentally different from traditional antipsychotics, each of which has substantial literature-based evidence implicating its potential benefits in the treatment of SCZ patients. However, PR curves for PhenoPredict are not optimal and can certainly be improved upon with future research efforts.

Table 3
Top 20-ranked repositioned drug candidates. NCT**: SCZ drugs from clinical trials. PMID**: SCZ drugs from biomedical literature. The FDA-approved SCZ drugs are highlighted.

R	Drug	Evidence	R	Drug	Evidence
1	Risperidone	**FDA-approved**	11	Memantine	NCT02001103 NCT00757978 NCT00097942
2	Methylphenidate	NCT00794040	12	Buspirone	NCT00178971
3	Quetiapine	**FDA-approved**	13	Paliperidone	**FDA-approved**
4	Citalopram	NCT00893256 NCT00047450 NCT01032083 NCT01032083	14	Haloperidol	**FDA-approved**
5	Olanzapine	**FDA-approved**	15	Lithium	NCT00202306 NCT00183443 NCT00202293
6	Sertraline	NCT00169988, NCT00531518	16	Amantadine	NCT00999505 NCT00975611 NCT00401973
7	Aripiprazole	**FDA-approved**	17	Levodopa	NCT01636037
8	Ziprasidone	**FDA-approved**	18	Atomoxetine	NCT00420498 NCT00222794 NCT00488163 NCT00628394 NCT00161031, NCT00089869
9	Clozapine	**FDA-approved**	19	Clomipramine	PMID9659874 PMID7635998 PMID7903293
10	Valproic acid	NCT00194025 NCT01094249 NCT02011750	20	Prednisone	PMID17245324 PMID23738211

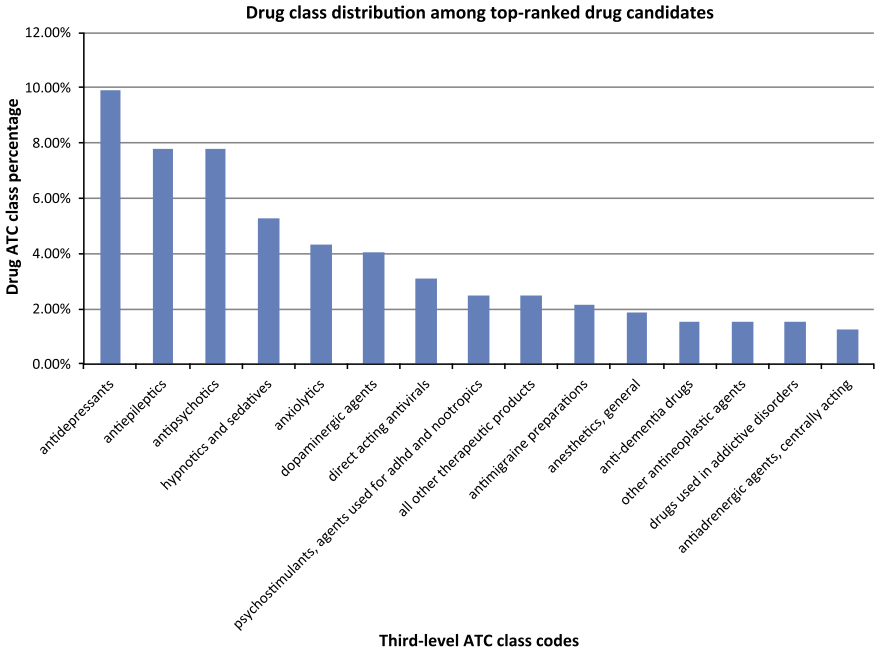

Fig. 6. Top 15 third level ATC codes (out of 95 codes) and their percentages for drug candidates ranked in the range of 0–15%. For example, 9.94% of drugs ranked in the range of 0–15% belong to the class "antidepressants".

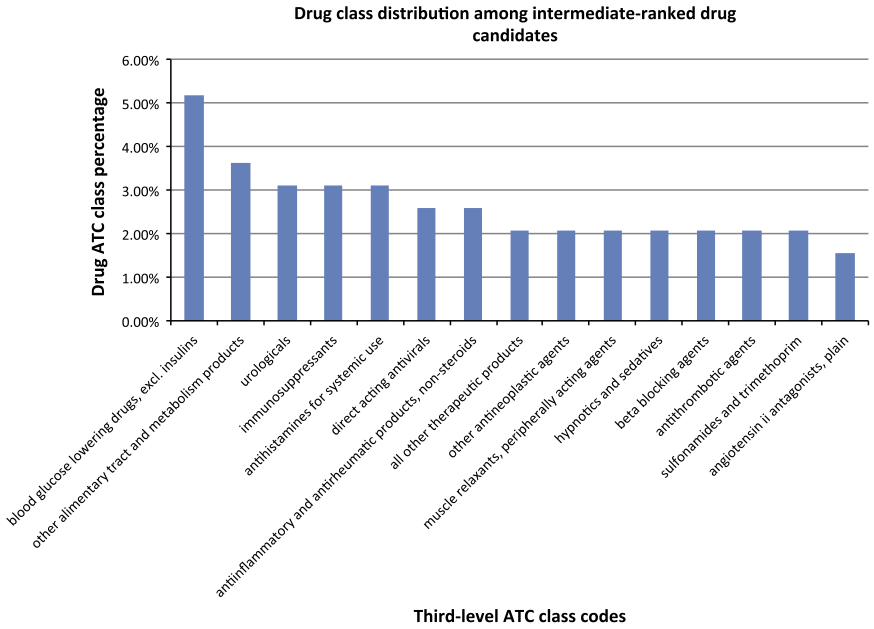

Fig. 7. Top 15 third level ATC codes (out of 93 codes) and their percentages for repositioned drug candidates ranked in the range of 16–30%. For example, 3.31% of drugs ranked in the range of 16–30% belong to the class "immunosuppressants."

First, it will be interesting to test the generalizability of PhenoPredict for other diseases. Currently, PhenoPredict included drug-disease treatment relationships for a total of 24,511 diseases and 2484 drugs. In theory, PhenoPredict can rank the 2484 drugs for each of the 24,511 diseases or vice versa.

Second, it will be interesting to investigate why PhenoPredict outperformed PREDICT. Such knowledge can offer insight into how to further improve both systems. Since the algorithms as well as the datasets included in both PhenoPredict and PREDICT are integral parts of these two systems, it is unclear which (algorithms or datasets or both) contributed to the PhenoPredict's advantage over PREDICT in finding drug candidates for schizophrenia. It will be interesting to investigate whether integrating datasets from both PhenoPredict and PREDICT can further improve the performances for each system.

Third, a limitation of using disease phenotypes in HPO for drug repositioning is that HPO mainly includes rare Mendelian disorders, the majority of which themselves have no available drug treatments. Therefore, the success of PhenoPredict in identifying drug repositioning candidates from similar diseases to a given input disease largely depends on the input disease as well as the treatment availability for top-ranked diseases.

Fourth, disease genetics and genomics, in combination with disease phenotypes, may further facilitate the discovery of truly innovative drug candidates for SCZ. Psychiatric disorders are among the most heritable of all common complex diseases. Human genomics and genetics studies have recently identified a large number of genetic risk factors for psychiatric disorders [47]. Although nearly all of the identified SCZ loci are nonspecific and not fully penetrant, recent GWAS studies have demonstrated shared genetic loci among phenotypically related psychiatric disorders including SCZ and bipolar disorder. While this justifies our approach of using disease phenotype data for drug repositioning, disease genetics and genomics may provide additional information not captured by disease phenotypes. However, the task of how to combining different level of evidence, including genetics, genomics, and phenomics, in order to build compassing predictive models for drug repositioning is challenging. We are actively exploring options for how to best accomplish this task.

Last but not least, incorporating other types of disease-phenotype relationships such as disease comorbidities and disease risk factors may offer additional drug repositioning opportunities for SCZ. Recently, we constructed three large-scale disease phenotypic knowledge bases, including a disease comorbidity knowledge base, a disease-risk relationship knowledge base, and a disease-manifestation knowledge base [27,53,54]. Unlike HPO, which includes exclusively Mendelian genetic disorders, these disease phenotype knowledge bases contain not only Mendelian disorders but also many common complex diseases. Currently, we are developing approaches to integrate disease phenotype knowledge from these complementary and heterogeneous data resources in an effort to further improve PhenoPredict.

Funding

RX was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under the NIH Director's New Innovator Award number DP2HD084068 and the Training grant in Computational Genomic Epidemiology of Cancer (CoGEC) (R25 CA094186-06). QW was partially supported by ThinTek LLC.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgements

Xu and Wang have jointly conceived the idea, designed and implemented the algorithms. All the authors have participated in study discussion and manuscript preparation.

References

[1] P.Y. Collins, V. Patel, S.S. Joestl, D. March, T.R. Insel, A.S. Daar, et al., Grand challenges in global mental health, Nature 475 (7354) (2011) 27–30.

[2] R.C. Kessler, W.T. Chiu, O. Demler, E.E. Walters, Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication, Arch. Gen. Psychiat. 762 (6) (2005) 617–627.
[3] C. Mathers, D.M. Fat, J.T. Boerma, The Global Burden of Disease: 2004 Update, World Health Organization, 2004.
[4] P.F. Sullivan, Puzzling over schizophrenia: schizophrenia as a pathway disease, Nat. Med. 18 (2) (2012) 210–211.
[5] S.E. Hyman, Time for new schizophrenia Rx, Science 343 (6176) (2014) 1177.
[6] S.E. Hyman, Revolution stalled, Sci. Transl. Med. 4 (155) (2012). 155cm11–155cm11.
[7] G. Miller, Is pharma running out of brainy ideas, Science 329 (5991) (2010) 502–504.
[8] S.E. Hyman, Psychiatric Drug Development: Diagnosing a Crisis, in: Cerebrum: The Dana Forum on Brain Science, vol. 2013, Dana Foundation.
[9] P.B. Mitchell, D. Hadzi-Pavlovic, Lithium treatment for bipolar disorder, Bull. World Health Organ. 78 (4) (2000) 515–517.
[10] F. Lopez-Munoz, C. Alamo, E. Cuenca, W.W. Shen, P. Clervoy, G. Rubio, History of the discovery and clinical introduction of chlorpromazine, Ann. Clin. Psychiatry 17 (3) (2005) 113–135.
[11] R.A. Maxwell, S.B. Eckhardt, Drug Discovery: A Casebook and Analysis, Humana Press, Clifton, NJ, (p. xvii)
[12] E. Dolgin, Rapid antidepressant effects of ketamine ignite drug discovery, Nat. Med. 19 (1) (2013) 8.
[13] T.T. Ashburn, K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, Nat. Rev. Drug Discovery 3 (8) (2004) 673–683.
[14] J.T. Dudley, T. Deshpande, A.J. Butte, Exploiting drugdisease relationships for computational drug repositioning, Briefings Bioinformatics 12 (4) (2011) 303–311.
[15] M.R. Hurle, L. Yang, Q. Xie, D.K. Rajpal, P. Sanseau, P. Agarwal, Computational drug repositioning: from data to therapeutics, Clin. Pharmacol. Ther. 93 (4) (2013) 335–341.
[16] M.J. Keiser, V. Setola, J.J. Irwin, C. Laggner, A.I. Abbas, S.J. Hufeisen, B.L. Roth, Predicting new molecular targets for known drugs, Nature 462 (7270) (2009) 175–181.
[17] S.L. Kinnings, N. Liu, N. Buchmeier, P.J. Tonge, L. Xie, P.E. Bourne, Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis, PLoS Comput. Biol. 5 (7) (2009) e1000423.
[18] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, T.R. Golub, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, Science 313 (5795) (2006) 1929–1935.
[19] M. Sirota, J.T. Dudley, J. Kim, A.P. Chiang, A.A. Morgan, A. Sweet-Cordero, A.J. Butte, Discovery and preclinical validation of drug indications using compendia of public gene expression data, Sci. Transl. Med. 3 (96ra) (2011) 77.
[20] J.T. Dudley, M. Sirota, M. Shenoy, R.K. Pai, S. Roedder, A.P. Chiang, A.J. Butte, Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease, Sci. Transl. Med. 3 (96) (2011). 96ra76–96ra76.
[21] A.P. Chiang, A.J. Butte, Systematic evaluation of drugdisease relationships to identify leads for novel drug uses, Clin. Pharmaco. Ther. 86 (5) (2009) 507–510.
[22] M. Campillos, M. Kuhn, A.C. Gavin, L.J. Jensen, P. Bork, Drug target identification using side-effect similarity, Science 321 (5886) (2008) 263–266.
[23] M. Duran-Frigola, P. Aloy, Recycling side-effects into clinical markers for drug repositioning, Genome Med. 4 (3) (2012).
[24] E. Lounkine, M.J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J.L. Jenkins, L. Urban, Large-scale prediction and testing of drug activity on side-effect targets, Nature 486 (7403) (2012) 361–367.
[25] E.J. Nestler, S.E. Hyman, Animal models of neuropsychiatric disorders, Nat. Neurosci. 13 (2010) 1161–1169.
[26] R. Xu, Q. Wang, Large-scale extraction of drug-disease treatment pairs from biomedical literature for drug repurposing, BMC Bioinformatics 14 (1) (2013) 181.
[27] R. Xu, L. Li, Q. Wang, Towards building a disease-phenotype relationship knowledge base: large scale extraction of disease-manifestation relationship from literature, Bioinformatics (2013), http://dx.doi.org/10.1093/bioinformatics/btt359.
[28] R. Xu, Q. Wang, Automatic signal prioritizing and filtering approaches in detecting post-marketing cardiovascular events associated with targeted cancer drugs from the FDA Adverse Event Reporting System (FAERS), J. Biomed. Inf. (2014) 171–177.
[29] P.N. Robinson, S. Kohler, S. Bauer, D. Seelow, D. Horn, S. Mundlos, The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, Am. J. Human Genet. 83 (5) (2008) 610–615.
[30] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (suppl 1) (2005) D514–D517.
[31] P.N. Robinson, S. Kohler, A. Oellrich, K. Wang, C.J. Mungall, S.E. Lewis, D. Smedley, Improved exome prioritization of disease genes through cross-species phenotype comparison, Genome Res. 24 (2) (2014) 340–348.
[32] X. Zhou, J. Menche, A.L. Barabasi, A. Sharma, Human symptomsdisease network, Nat. Commun. (2014) 5, http://dx.doi.org/10.1038/ncomms5212.
[33] Y. Moreau, L.C. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nat. Rev. Genet. 13 (8) (2010) 523–536.
[34] A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, Mol. Syst. Biol. 7 (1) (2011) 55.
[35] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (suppl 1) (2004) D267–D270.
[36] Y. Chen, R. Xu, Network-based gene prediction for Plasmodium falciparum malaria towards genetics-based drug discovery, in: International Conference on Intelligent Biology and Medicine (ICIBM 2014), December 4–6, San Antonio, TX.
[37] R. Xu, Q. Wang, L. Li, Genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine N-oxide (TMAO), a gut microbial metabolite of dietary meat and fat, In: International Conference on Intelligent Biology and Medicine (ICIBM 2014), December 4–6, San Antonio, TX.
[38] ICD-10: International Statistical Classification of Diseases and Related Health Problems: 10th Revision, World Health Organization, 1992.
[39] C.D. Manning, Foundations of Statistical Natural Language Processing, in: H. Schutze (Ed.), MIT Press, 1999.
[40] F.J. Provost, T. Fawcett, R. Kohavi, The case against accuracy estimation for comparing induction algorithms, in: International Conference on Machine Learning (ICML), vol. 98, 1998, pp. 445–453.
[41] J. Davis, E.S. Burnside, I. de Castro Dutra, D. Page, R. Ramakrishnan, V.S. Costa, J.W. Shavlik, View learning for statistical relational learning: with an application to mammography, in: International Joint Conference on Artificial Intelligence (IJCAI), 2006, pp. 677–683.
[42] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2005, pp. 233–240.
[43] World Health Organization: The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD), WHO, Norway.
[44] B. Dean, Understanding the role of inflammatory-related pathways in the pathophysiology and treatment of psychiatric disorders: evidence from human peripheral studies and CNS studies, Int. J. Neuropsychopharm. 14 (07) (2011) 997–1012.
[45] U. Meyer, Anti-inflammatory signaling in schizophrenia, Brain Behav. Immun. 25 (8) (2011) 1507–1518.
[46] N. Muller, M.J. Schwarz, S. Dehning, A. Douhe, A. Cerovecki, B. Goldstein-Muller, M. Riedel, The cyclooxygenase-2 inhibitor celecoxib has therapeutic effects in major depression: results of a double-blind, randomized, placebo controlled, add-on pilot study to reboxetine, Mol. Psychiatry 11 (7) (2006) 680–684.
[47] P.F. Sullivan, M.J. Daly, M. O'Donovan, Genetic architectures of psychiatric disorders: the emerging picture and its implications, Nat. Rev. Genet. 13 (8) (2012) 537–551.
[48] K. Wahlbeck, M.V. Cheine, S. Gilbody, J. Ahonen, Efficacy of Beta-blocker supplementation for schizophrenia: a systematic review of randomized trials, Schizophr. Res. 41 (2) (2000) 341–347.
[49] E. Shek, S. Bardhan, M.V. Cheine, J. Ahonen, K. Wahlbeck, Beta-blocker supplementation of standard drug treatment for schizophrenia, Schizophrenia Bull. (2010) (sbq089).
[50] M. van den Buuse, T.W. Zheng, L.L. Walker, D.A. Denton, Angiotensin-converting enzyme (ACE) interacts with dopaminergic mechanisms in the brain to modulate prepulse inhibition in mice, Neurosci. Lett. 380 (1) (2005) 6–11.
[51] I.M. Phillips, Functions of angiotensin in the central nervous system, Annu. Rev. Physiol. 49 (1) (1987) 413–433.
[52] A.M. Domeney, Angiotensin converting enzyme inhibitors as potential cognitive enhancing agents, J. Psychiatry Neurosci. 19 (1) (1994) 46.
[53] R. Xu, L. Li, Q. Wang, dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text, BMC Bioinformatics 15 (1) (2014) 105.
[54] Y. Chen, X. Zhang, G.Q. Zhang, R. Xu, Comparative analysis of a novel disease phenotype network based on clinical manifestations, J. Biomed. Inform. 53 (2015) 113–120, http://dx.doi.org/10.1016/j.jbi.2014.09.007.