HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2019 March 20.

Published in final edited form as:

Pac Symp Biocomput. 2019 ; 24: 308–319.

Precision drug repurposing via convergent eQTL-based molecules and pathway targeting independent disease-associated polymorphisms

Francesca Vitali[†,1,2], **Joanne Berghout**[†,1,2,3], **Jungwei Fan**[†,1,2], **Jianrong Li**[1,2], **Qike Li**[1], **Haiquan Li**[*,1,2,4], and **Yves A. Lussier**[*,1,2,3,4,5,6,7]

[1]Center for Biomedical Informatics and Biostatistics (CB2), The University of Arizona, Tucson, AZ 85721, USA

[2]Department of Medicine COM-T, The University of Arizona, Tucson, AZ 85721, USA

[3]The Center for Applied Genetics and Genomics in Medicine, The University of Arizona, Tucson, AZ 85721, USA

[4]Department of Biosystems Engineering, The University of Arizona, Tucson, AZ 85721, USA

[5]BIO5 Institute, The University of Arizona, Tucson, AZ 85721, USA

[6]UA Cancer Center, The University of Arizona, Tucson, AZ 85721, USA

[7]UA Health Science (UAHS), The University of Arizona, Tucson, AZ 85721, USA

Abstract

Repurposing existing drugs for new therapeutic indications can improve success rates and streamline development. Use of large-scale biomedical data repositories, including eQTL regulatory relationships and genome-wide disease risk associations, offers opportunities to propose novel indications for drugs targeting common or convergent molecular candidates associated to two or more diseases. This proposed novel computational approach scales across 262 complex diseases, building a multi-partite hierarchical network integrating (i) GWAS-derived SNP-to-disease associations, (ii) eQTL-derived SNP-to-eGene associations incorporating both *cis*- and *trans*- relationships from 19 tissues, (iii) protein target-to-drug, and (iv) drug-to-disease indications with (iv) Gene Ontology-based information theoretic semantic (ITS) similarity calculated between protein target functions. Our hypothesis is that if two diseases are associated to a common or functionally similar eGene - and a drug targeting that eGene/protein in one disease exists - the second disease becomes a potential repurposing indication. To explore this, all possible pairs of independently segregating GWAS-derived SNPs were generated, and a statistical network of similarity within each SNP-SNP pair was calculated according to scale-free overrepresentation of convergent biological processes activity in regulated eGenes ($ITS_{eGENE\text{-}eGENE}$) and scale-free overrepresentation of common eGene targets between the two SNPs ($ITS_{SNP\text{-}SNP}$). Significance of $ITS_{SNP\text{-}SNP}$ was conservatively estimated using empirical scale-free permutation resampling

[*]Corresponding authors contributed equally to this work.
[†]Authors contributed equally to this work

keeping the node-degree constant for each molecule in each permutation. We identified 26 new drug repurposing indication candidates spanning 89 GWAS diseases, including a potential repurposing of the calcium-channel blocker Verapamil from coronary disease to gout. Predictions from our approach are compared to known drug indications using DrugBank as a gold standard (odds ratio=13.1, p-value=2.49×10^{-8}). Because of specific disease-SNPs associations to candidate drug targets, the proposed method provides evidence for future precision drug repositioning to a patient's specific polymorphisms.

Keywords

Drug repurposing; network analysis; drug repositioning; translational bioinformatics

1. Introduction

Drug repurposing is an approach that investigates an approved drug for its potential efficacy as a treatment for other diseases[1]. This strategy can be cheaper, faster, and more efficient than *de novo* drug discovery since many preclinical and safety studies have already been completed[2, 3].

Some reported repurposing successes have relied on serendipitous clinical observation (i.e., Sildenafil/Viagra repurposed from pulmonary arterial hypertension to erectile dysfunction)[4] while many others use disease-specific basic biology hypotheses where a single molecular factor has been independently associated with pathology in two or more diseases (i.e., FYN in solid tumor proliferation and Alzheimer's)[3]. Employing scalable computational methods offers great potential for finding credible, novel, and hypothesis-free repurposing opportunities[2, 5] by rapidly linking genetic risk factors and/or molecules perturbed during disease processes with known drug targets or other identified consequences of therapy[2, 5–7]. Several computational network analysis methods have been developed for drug repurposing, generally beginning from a seed set of well-described proteins or druggable targets. These then incorporate data from protein-protein and/or protein-drug biochemistry to propose new functional candidate molecules and drug activity based on presumptive physical interactions[8, 9]. Other methods examine gene expression changes to predict signature similarity between two diseases or between a disease and a drug exposure as a way to propose candidates[10, 11]. However, these methods are limited due to (i) typically relying on single-scale methodologies and (ii) focusing on coding DNA or their gene products. High-level integration of different data sources and knowledge are required to efficiently perform multiscale analysis for a more thorough approach to hypothesis-free drug repurposing, as well as integration of signals from noncoding areas of the genome.

Genome-wide association studies (**GWAS**) represent a large potential source of information on genetic factors associated with disease risk or severity. However, about 50% of associations detected by GWAS have mapped to intergenic or noncoding sequences, suggesting altered regulatory capacity that remains difficult to interpret[12]. Fortunately, massive amounts of new data have been generated to address questions of noncoding function. These include the Genotype-Tissue Expression (GTEx) resource which mapped

expression quantitative trait loci (eQTL) linking single nucleotide polymorphisms (**SNP**) to tissue-specific regulation of gene transcripts (**eGenes**)[13]. Colocalization of GWAS positional loci with these data[14, 15] and/or with additional computational integration of data in other knowledge bases (e.g., protein-protein interaction networks, Gene Ontology (**GO**)[16] annotations) shows that GWAS loci are enriched in putatively functional regions[13, 14]. In addition, non-scalable and rate-limited studies have led to the discovery and characterization of several new disease-gene and disease-biological pathway mechanistic candidates[17–21].

Motivation. We have previously designed a multiscale network approach where SNPs from GWAS are connected to gene products and their annotations via eQTL[22]. In that study, we demonstrated that pairs of independently segregating GWAS SNPs associated to the same disease were significantly more likely to be involved in similar biological processes, colocalized with binding sites for the same transcription factor(s), and involved in chromatin interactions with each other when compared to pairs of SNPs where each SNP mapped to a different disease[22]. This is consistent with the prevailing idea that heterogeneous risk factors for a given complex disease will display some form of coalescent properties and/or converge into a few non-random, key pathways involved in driving pathology, at least in many cases[23, 24].

In this study, we ***hypothesized*** that the downstream convergence of eQTL signals between highly similar SNP-SNP pairs can be leveraged to identify druggable molecular targets relevant to two diseases. Therapeutic modulation of that factor or the pathways it is involved with present a potential opportunity for drug repurposing. We computed similarity scores between risk factors (here, SNP-SNP pairs) based on information theoretic semantic (**ITS**) similarity of their associated gene ontology biological process terms ($\boldsymbol{ITS_{GENE\text{-}GENE}}$) and overrepresentation of shared or similar eGenes ($\boldsymbol{ITS_{SNP\text{-}SNP}}$). These data were integrated with drug targeting data[25, 26]. We further demonstrate that a scale-free resampling analysis of the resulting multiscale network discovers and prioritizes a significant number of known drug-to-indication relationships from our gold standard, i.e., known treatments for the network diseases. We also report a repurposing example with literature evidence confirming the plausibility of our findings. The drug repurposing approach we developed is different from the standard approaches (for a review refer to[5]) since, to our knowledge, no method has been yet published that integrates GWAS studies with eQTL associations as pairs, with gene ontology similarities leveraged to repurpose drugs across diseases incorporating both identical and similar pathological effectors and mechanisms.

2. Methods

2.1. Datasets

GWAS SNP-to-disease associations were obtained from the NHGRI-EBI GWAS Catalog[27] (11/20/2017) comprising 53,009 associations between 2,373 diseases/traits and 41,973 lead SNPs. *SNP-to-eGene associations*. A comprehensive secondary *cis*- and *trans*-eQTL analysis by Fagny et al[19] of the original raw data in the Genotype-Tissue Expression dataset[28] (GTEx vers. 6.0) was used for linking SNPs to eGenes (http://networkmedicine.org:3838/eqtl/; 19 tissues). Fagny et al[19] adjusted p-values for multiple testing using Benjamini-Hochberg correction for *cis*- and *trans*-eQTL separately, and

suggest retaining associations with False Discovery Rate (FDR)< 0.2. Sample genotypes were imputed by GTEx[29], providing comprehensive overlap with the GWAS SNP set. The entire dataset included 5,896,354 associations between 1,114,453 SNPs and 21,971 eGenes.

Molecular drug-to-indication and target-to-drug and associations were downloaded from DrugBank API Portal (v1, 02/01/2018)[25] and DrugBank (01/11/2017)[26] respectively. The database consisted of 4,943 associations linking 1,133 drugs with 2,622 unstructured indications (i.e., diseases), as well as 11,978 associations linking 2,515 molecular targets with 5,623 drugs.

Gene Ontology (GO)[30] (06/28/2016) provided 29,690 GO IDs in Biological Processes (GO-BP) and 120,779 associations involving 16,604 genes and 11,052 GO-BP IDs.

2.2. Building the drug repurposing network

Briefly, we constructed an integrated multiscale biomolecular network connecting (i) diseases to (ii) SNPs to (iii) eGenes (eQTL transcripts) and cognate proteins intersected with both (iv-a) GO biological processes annotations (GO-BP) and (iv-b) drugs acting on the protein molecular targets (Fig. 1). This network thus links each SNP to a set of eGenes and GO-BP terms. All possible SNPSNP pairs were created, filtered to remove those marking the same linkage locus, and SNP-SNP similarity was computed based on information theoretic semantic similarity of each eGene pair's GO-BP terms ($ITS_{eGENE\text{-}eGENE}$) and overrepresentation of the SNP-pair's shared or similar eGenes ($ITS_{SNP\text{-}SNP}$). Statistically prioritized SNP pairs within a disease were used for method and target validation (Fig. 1D). SNP pairs that spanned two diseases yet still showed an overrepresentation of shared and/or highly similar molecular downstream eGenes were suggested as repurposing candidates (Fig. 1C and **4B**).

Preprocessing the data was necessary for the integration of each element in the drug repurposing network. First, disease terms used by the GWAS Catalog and DrugBank required standardization into a formal representation (Methods 2.2.1), as well as an automated approach for match identical or highly similar diseases between these datasets (Methods 2.2.2). Next, we developed a method to establish the convergent biomolecular processes revealed by within-disease GWAS risk SNP-SNP pairs and compute similarity of these processes across diseases. We propose a nested information theoretic distance that considers the functional similarities between downstream eGenes of SNP pairs for prioritization of SNP pairs (Methods 2.2.3–5). Once the statistically significant eGene and SNP pairs are identified (FDR<0.05), we construct the biomolecular layer (Methods 2.2.6) and integrate this with the drug information (Methods 2.2.7) to create the **Drug Repurposing Network**.

2.2.1. Formal representation of disease terms (NHGRI GWAS and DrugBank).

—Multiple GWAS disease traits collected from the NHGRI GWAS Catalog were grouped into semantic disease bundles, each assigned to a SNOMED-Clinical Terms (**CT**) concept representation[31]. The GWAS curator-assigned Experimental Factor Ontology (EFO)[27] was used to filter out non-disease phenotypes (e.g., pharmacogenomics responses, etc.) by retaining those under the branch EFO0000408: disease, reducing the 2,373 GWAS traits to

533 diseases. Text mining scripts and cross-mapping were used to link SNOMED-CT concepts to the EFO diseases, which were checked and curated into 262 bundles and coded to SNOMED-CT IDs. These bundles are referred to as "**GWAS diseases**" hereafter. We similarly coded 1,936 out of the 2,622 unstructured text disease terms of "**DrugBank indications**" to 2,054 distinct SNOMED IDs (Fig.1C). Note that one DrugBank indication can map to multiple SNOMED IDs.

2.2.2. Disease similarity computation.—SNOMED-CT ontology was chosen because of its rich hierarchical relationships and high clinical coverage relevant to GWAS diseases and DrugBank indications. Disease-disease semantic similarity was determined by applying Lin's information-theoretic similarity (**ITS)** metric[32] with Sánchez et al.'s information content (**IC**) estimation[33] (Eq.1). By integrating these, ITS between diseases d_1 and d_2 ($ITS_{DISEASE\text{-}DISEASE}$) can be calculated through Eq. 2, based on the hierarchical structure of the SNOMED-CT ontology. ITS similarity scores range from 0 to 1, where 1 corresponds to identity and 0 to complete dissimilarity. Two disease concepts with ITS>0.7 were considered similar. Using SNOMED, similarity is computed between every disease pair within the GWAS disease list as well as across the GWAS disease list and the DrugBank indication list (Eq. 2). Of note, drug repurposing is predicted between independent GWAS disease(s)-associated SNPs with non-trivial convergent eQTL mechanisms (Sections 2.2.3–5), in which one of these GWAS diseases is similar or identical to a DrugBank indication (***ITSGWAS_Disease–DrugBank_Indication(d1,d2)***>0.7, applied Eq.2; Methods 2.2.7).

$$IC(c) = -log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \quad (1)$$

$$ITS_{DISEASE-DISEASE}(d_1,d_2) = \frac{2\times IC(lca(d_1,d_2))}{IC(d_1)+IC(d_2)} \quad (2)$$

where $|leaves(c)|$ is the number of leaf nodes under the concept c, $|subsumers(c)|$ is the number of ancestor nodes above the concept, max_leaves is the total number of leaves covered by the root node, d is a disease, and lca is the least common ancestor to d_1 and d_2.

2.2.3. Information theoretic similarity between two eGenes using GO Biological Processes.—We also applied the information-theoretic approach that we previously published[34] to calculate functional similarity between any pair of eGenes (Fig. 1A), i.e., $ITS_{eGENE\text{-}eGENE}$. In GO, each gene product (g_x), used here as the canonical cognate protein of an eGene transcript, can be annotated to a set of GO terms (T), denoted as $T(g_x)$. The similarity between *eGene 1* (g_1) and *eGene 2* (g_2) is calculated by semantic similarity between $T(g_1)$ and $T(g_2)$. For each GO-BP term (t_i) associated to g_1, the similarity score $ITS_{GO\text{-}GO}(t_i, t_j)$ is then calculated for every GO term (t_j) associated to g_2 ($t_i \in T(g_2)$) (Fig.1A) and use the maximum value among them (*max*); and vice-versa for g_2. The

similarity between two genes g_1 and g_2 is thus calculated as the average of all these maximum scores (Eq.3):

$$ITS_{eGENE-eGENE}(g_1, g_2) = \frac{\sum_{t_i \in T(g_1)} \max_{t_j \in T(g_2)} \left(ITS_{GO-GO}(t_i, t_j) \right) + \sum_{t_j \in T(g_2)} \max_{t_i \in T(g_1)} \left(ITS_{GO-GO}(t_i, t_j) \right)}{|T(g_1)| + |T(g_2)|} \quad (3)$$

where $|T(g_1)|$ is the cardinality of the set $T(g_1)$. The $ITS_{eGENE\text{-}eGENE}$ output has a range between 0 and 1, where 0 indicates two genes having no similar GO annotations and 1 indicates two genes having identical GO annotations.

2.2.4. Information theoretic similarity between SNPs.—The ITS of a SNP-SNP pair was calculated where both are (i) associated with at least one of the 262 GWAS diseases (**Methods; 2.1.1**) and (ii) regulate at least one eGene. Our previously published calculation[22] of similarity between a pair of SNPs (***ITS*$_{SNP\text{-}SNP}$**) is an extension of the $ITS_{eGENE\text{-}eGENE}$. Since every SNP can be associated with multiple eGenes and every eGene can be associated with multiple GO terms, the $ITS_{SNP\text{-}SNP}$ is a nested calculation that leverages the $ITS_{eGENE\text{-}eGENE}$ scores. It is based on the average similarity of the set of genes associated by eQTL with the two SNPs, as shown in Eq.4 below:

$$ITS_{SNP-SNP}(s_1, s_2) = \frac{\sum_{g_i \in G(s_1)} \max_{g_j \in G(s_2)} \left(ITS_{eGENE-eGENE}(g_i, g_j) \right) + \sum_{g_j \in G(s_2)} \max_{g_i \in G(s_1)} \left(ITS_{eGENE-eGENE}(g_i, g_j) \right)}{|G(s_1)| + |G(s_2)|} \quad (4)$$

where SNP s_1 was associated with a set of genes $G(s_1)$, and $|G(s_1)|$ is the cardinality of the set $G(s_1)$, similarly for s_2. The $ITS_{eGENE\text{-}eGENE}$ is the similarity of two genes computed with Eq.3. Likewise, the $ITS_{SNP\text{-}SNP}$ has a score ranging from 0 to 1; a value of 1 indicates two SNPs of perfect similarity, and 0 refers to two SNPs of null functional similarity.

2.2.5. Scale-Free permutation for FDR estimation of ITS.—10,000 and 100,000 conservative scale-free permutations were performed to estimate statistical significance of

the $ITS_{eGENE\text{-}eGENE}$ and $ITS_{SNP\text{-}SNP}$ scores (~500,000 core hours), respectively. In each permutation, the node degree of every node in the gene-GO annotation network was preserved (each specific gene retained the node degree of GO term associations and vice-versa). Multiplicity of prioritization was controlled by Benjamini-Hochberg with a cutoff of FDR≤0.05 (*p.adjust* for both $ITS_{eGENE\text{-}eGENE}$ and $ITS_{SNP\text{-}SNP}$).

2.2.6. Biomolecular network layer construction (Fig.1).—The drug repurposing network construction starts by defining its biomolecular layer. This level associates **GWAS diseases**, **SNPs**, and **molecular targets** (Fig.1A). Disease-to-SNP edges were obtained from GWAS lead SNPs, and SNP-to-regulated molecular target (eGene) edges were obtained from eQTL data as described in **Methods 2.1.2**. This produced a network of 9,750 associations between 8,955 SNPs and 235 unique diseases, where each of the retained SNPs was also associated with at least one eGene via eQTL. All SNP-SNP pairs were generated and filtered to remove SNP pairs (i) separated by less than 5Mb, (ii) in linkage disequilibrium with one another (r^2>0.01) according to HapMap and 1000 Genomes CEU data, and/or (iii) SNP pairs where both mapped within the Major Histocompatibility Complex (MHC; Chr6: 28,477,797–33,448,355, ±2 Mb; GRCh37). SNP-SNP pairs where only one SNP mapped to the MHC were retained. This was done to remove SNP pairs trivially marking the same locus. Similarity is computed ($ITS_{SNP\text{-}SNP}$) for each retained SNP pair according to Eq.4 (Methods 2.2.4). Focusing only on the SNP pairs that were statistically significant ($ITS_{SNP\text{-}SNP}$, FDR<0.05), $ITS_{eGENE\text{-}eGENE}$ is computed (Eq.3) to further filter. SNP pairs that satisfied both $ITS_{SNP\text{-}SNP}$ and $ITS_{eGENE\text{-}eGENE}$ at FDR<0.05 were considered as having convergent biological mechanisms and used to construct the final biomolecular network.

2.2.7. Construction of the drug repurposing network.—The final network construction step involves the integration of drug knowledge (Fig.1B) with the biomolecular level by matching protein-coding eGenes with the molecular targets of drugs acquired from DrugBank (**Methods 2.1.3**). In this step, the disease indications are included for these drugs, as they serve to validate our predictions when recapturing known indications (validation, Methods 2.3) and to identify novel opportunities predicted by our method that can be used for drug repurposing (Methods 2.4).

2.3. Validation of the drug repurposing network

Before analyzing potential drug repurposing candidates, we validated our drug repurposing network by determining whether known drug indications for the included GWAS diseases could be inferred from the network above the chance expectation (Fig.1C). To this end, a Fisher's Exact Test (FET) is performed considering: (i) all druggable molecular targets (**DMTs**) and (ii) all druggable diseases (**DD**). In this validation, a DMT was defined as any eGene that has at least one drug in DrugBank targeting the cognate protein, and that the drug is indicated for one or more of the 262 GWAS diseases defining our set (Methods 2.2). A DD is defined as any GWAS diseases in the network associated with at least one target eGene found in DrugBank, and therefore corresponds to all the GWAS diseases that could theoretically be validated using these databases. In this way, we can determine how many of the theoretical combinations of DMTs and DDs (DMTs*DDs) are predicted by analysis of

significant eGenes associated with prioritized SNP pairs with convergent mechanisms. The enrichment of gold standard drug indications among the predictions is conducted assuming that the GWAS disease-eGenes analysis can, in principle, discover any drug targets in DrugBank. We constructed the contingency table to perform the FET by counting the number of DMT-DD interactions (i) present/not present in Drugbank vs (ii) included/not included in our final ITS-filtered network (Fig.1C).

The validation procedure includes similarity between GWAS diseases and indications (Fig. 1D; $\boldsymbol{ITS_{GWAS\_Disease–GWAS\_Disease}}$, Eq.2; Methods 2.2.2). The network validation procedure is then conducted by applying two additional conditions, one stringent and one more relaxed (Fig.1D), using DrugBank as a gold standard. First, convergent mechanisms between two SNPs associated with the same GWAS are prioritized ($ITS_{GWAS\_Disease–GWAS\_Disease}>0.7$), i.e., similar SNP pairs ($ITS_{SNP\text{-}SNP}$ FDR<0.05) with eGene pairs ($ITS_{eGENE\text{-}eGENE}$ FDR<0.05), and the number of eGene-GWAS disease associations that were identical or similar ($ITS_{GWAS\_Disease–DrugBank\_Indication}>0.7$) to the related molecule-indication associations found in DrugBank were counted (Fig.1D). In the relaxed condition, the same procedure is applied, but without the constraint that both SNPs in the prioritized pair must map to the same disease (Fig.1D).

2.4. Drug repurposing pattern identification

Drug repurposing candidates are identified by analyzing specific network patterns as illustrated in Fig.1D. We prioritized all subnetworks involving pairs of GWAS diseases related to convergent mechanism in which at least one eGene was targeted by a drug known to treat one of the two GWAS diseases or a similar ($ITS_{GWAS\_Disease–DrugBank\_Indication} \leq 0.7$) disease. Thus, if the drug is prescribed as a treatment for two diseases dissimilar in the pair ($ITS_{GWAS\_Disease–GWAS\_Disease}>0.7$), then it is predicted as a repurposing candidate across the two GWAS diseases.

3. Results and discussion

3.1. Overall results and visualization

The drug repurposing network (Fig.2A) comprises 1,865 nodes and 15,655 edges (Fig.2B) and was obtained after considering the similarity of 479,896 SNP-SNP pairs. 74,803 SNP pairs are prioritized with significant convergent biomolecular mechanisms ($ITS_{SNP\text{-}SNP}$ with FDR<0.05, Eq.4; Methods 2.2.5). The list of similar SNP-pairs is further constrained to those with an association to at least one disease for which an indication is known in DrugBank, resulting in 9,418 retained SNP pairs, their associated significant eGene pairs ($ITS_{eGENE\text{-}eGENE}$ with FDR<0.05, Methods 2.2.3), and drug information (Methods 2.2.7). All retained SNP pairs marked two independently segregating disease loci, based on the positional and linkage filters applied in Methods 2.2.6. SNP pair similarity was driven by both *cis*- and *trans*-eQTL associations, with 8,329 SNP pairs prioritized through regulation of similar eGenes found in *cis* to each SNP, and 1,089 SNP pairs prioritized based on at least one *trans*-regulated eGene by one of the SNPs (Fig. 2A). Fig. 2B shows details of the network nodes and edges. While they remain a minority, having 12% of prioritized SNP pairs reliant on *trans*-eQTL relationships highlights the importance of including these

complex regulatory data, as these would have been overlooked by focusing exclusively on those genes near the GWAS SNP. The subnetwork relevant for drug repurposing comprises only the SNP-pairs for which their prioritized eGenes code for the protein target of an existing drug (Fig. 2C).

3.2. Network validation results

We validated our network by calculating the enrichment of drug targets predicted by our method (Methods 2.3) over drug targets reported in a curated database gold standard (DrugBank). First, identical or similar disease indications matched to any of the 262 GWAS diseases are extracted, which resulted in 127 "druggable" diseases (DD) together with their 1,336 associated druggable molecular targets (DMT). This yielded 169,672 eGene-disease combinations that could potentially be predicted (DMT*DD). Assuming the stringent criterion where DrugBank's annotated drug indication must be identical or similar to the GWAS disease and both SNPs in the prioritized SNP-SNP pair must be associated to that same GWAS disease, our method predicted 56 relationships involving DMTs and GWAS diseases. DrugBank included 2,783 DMT-DD associations with 10 overlapping (Fisher's Exact Test-FET $p=2.5\times10^{-8}$; odds ratio=13.1). When considering the more relaxed criterion of high similarity between gold standard diseases and predicted indications, we found 29 overlapping, from a total of 299 potential predictions (FET $p= 3.6\times10^{-14}$; odds ratio= 6.5).

Fig.3A illustrates a drug target for Rheumatoid Arthritis (RA) that was predicted by eQTL similarity of two distinct GWAS SNPs[35] ($ITS_{SNP\text{-}SNP}$ FDR=0.0007) and confirmed in DrugBank as the known target of Etanercept indicated for Polyarticular Juvenile Idiopathic Arthritis (PJIA)[36, 37]. These two RA SNPs (rs72717009 and rs4239702) affect the expression of *FCGR2C* and *CD40* respectively. The gene products of *FCGR2C* and *CD40* are annotated to highly similar biological processes ($ITS_{eGENE\text{-}eGENE}$ FDR=0.01), suggesting a convergent mechanism revealed by these two independently segregating factors. Since RA and PJIA are highly similar diseases ($ITS_{RA\text{-}PJIA}$=0.78), our approach could correctly predict Etanercept as a treatment for RA[37].

3.3. Drug repurposing results

Following the procedure in Methods 2.4, we extracted the GWAS diseases having convergent mechanisms ($ITS_{SNP\text{-}SNP}$ FDR<0.05 and $ITS_{GENE\text{-}GENE}$ FDR<0.05) with one of the GWAS diseases for which at least one gold standard indication was present in the network. In detail, we identified 181 distinct GWAS disease pairs involving 90 diseases. 19 of these diseases had a molecularly-targeted treatment indicated in DrugBank that matched the eGene-prioritized molecular targets (i.e., GWAS $disease_A$ shown in Fig.1D). 89 diseases had new drug candidates identified by our network, potentially allowing repurposing (i.e., GWAS $disease_B$ in Fig.1D). We extracted 1,288 patterns (Supplementary Material -Table S1) including 26 drug candidates relevant to at least one of the 89 GWAS diseases. The subnetwork obtained by considering the drug repurposing patterns is depicted in Fig.2B and comprises 628 nodes (90 GWAS diseases, 253 SNPs, 108 eGenes, 26 drugs and 151 indications) and 1,758 edges. Within the 391 SNP-SNP pairs (edges), 25 were prioritized based on at least one *trans*-eQTL association and 366 are driven exclusively by *cis*-eQTL associations. Tissue source of each eQTL association are provided in Table S1. As eQTL

detection power varied between tissues in our input and multi-organ pathologies are common in complex diseases, we chose not to restrict our results to only those with shared or overlapping tissue sources. However, as candidates are considered more closely, these filters may allow prioritization and/or a cleaner set of hypotheses.

Fig.3B illustrates Verapamil as a candidate drug target for gout repositioned from coronary artery disease that was predicted by eQTL similarity of their respective distinct GWAS SNPs ($ITS_{SNP\text{-}SNP}$ FDR=0.000039). The proposed method predicted that KCNH2 is involved in similar biological processes as KCKN7 (FDR $ITS_{eGENE\text{-}eGENE}$ FDR<10^{-4}). Verapamil is a calcium channel blocker and inhibitor of the protein Potassium voltage-gated channel subfamily H member 2 (*KCNH2*)[38]. It is a class IV anti-arrhythmia agent currently used to treat hypertension, angina, and cluster headache. The cross-disease prioritized SNP pair indicates that variation at rs13232179 (coronary artery disease[39]) modulates expression of KCNH2 in tibial artery and that variation at rs10791821 (gout[40]) modulates expression of KCNK7 in tibial artery, transverse colon, esophagus muscularis, and thyroid. Functional similarity between KCNH2 and KCNK7 suggests that effective pathway modifying medications may play a role in both conditions. Supporting this prediction, studies have demonstrated that other calcium channel blockers are associated with a lower risk of incident gout[41].

4. Limitations and future studies

Currently, our method cannot detect if the effect of the expression from eQTL studies is concordant; and so, the proposed method may predict adverse events as well as drug repurposing opportunities. For example, Adalimumab (Fig.2C), currently prescribed for inflammatory bowel disease, is predicted as a possible treatment for Multiple Sclerosis (MS). However, anecdotal cases report worsening of MS patients treated with this drug[42]. Regulation of eGenes in distinct tissues may also have important biological consequences. Future studies will focus on (i) experimental validation of select candidates, (ii) to provide the data with filtering and analysis tools as an online public repository, and (iii) the integration of directional eQTL information in the presence of specific SNP variants to determine if these cases can be predicted.

5. Summary and conclusion

Drug repurposing offers novel venues to use currently available or investigational drugs. We developed a computational drug repurposing approach leveraging several data and knowledge resources, by integrating GWAS studies, eQTL data, drug information, and GO similarities in a multi-partite hierarchical network. Our approach is anchored on the identification of convergent *cis*- and *trans*- eQTL targets across distinct disease-associated polymorphisms. These repurposings are distinct from previous approaches in that we integrate convergent downstream *cis*- and *trans*-eQTL signals from any polymorphism, inclusive of intergenic regions. This automatically suggests drug repurposing through shared molecular target candidates identified across diseases, beyond the straightforward “host” or “nearest” gene overlap (e.g., protein-interaction networks). Our study demonstrates that GWAS and eQTL-derived networks can predict a significant number of gold standard

indications and novel drug repurposing suggestions. Because of specific disease SNPs-associations to candidate drug targets, the proposed method provides evidence for future precision drug repositioning to a patient's specific polymorphisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Drs. M Fagny, JN Paulson, J Quackenbush and J Platig for providing early access to tissue specific eQTL associations.

This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, The UA Cancer Center, and NIH (U01AI122275)

References

1. Chong CR and Sullivan DJ, Jr., Nature, 2007, 448, 645–646. [PubMed: 17687303]
2. Cha Y, Erez T, et al., British journal of pharmacology, 2018, 175, 168–180. [PubMed: 28369768]
3. Strittmatter SM, Nat Med, 2014, 20, 590–591. [PubMed: 24901567]
4. Ban TA, Dialogues Clin Neurosci, 2006, 8, 335–344. [PubMed: 17117615]
5. Li J, Zheng S, et al., Briefings in bioinformatics, 2016, 17, 2–12. [PubMed: 25832646]
6. Lamb J, Crawford ED, et al., science, 2006, 313, 1929–1935. [PubMed: 17008526]
7. Sirota M, Dudley JT, et al., Science translational medicine, 2011, 3, 96ra77.
8. Cheng F, Desai RJ, et al., Nat Commun, 2018, 9, 2691. [PubMed: 30002366]
9. Luo Y, Zhao X, et al., Nat Commun, 2017, 8, 573. [PubMed: 28924171]
10. He X, Fuller CK, et al., Am J Hum Genet, 2013, 92, 667–680. [PubMed: 23643380]
11. Corsello SM, Bittker JA, et al., Nat Med, 2017, 23, 405–408. [PubMed: 28388612]
12. Visscher PM, Wray NR, et al., AJHG, 2017, 101, 5–22.
13. Consortium G, Science, 2015, 348, 648–660. [PubMed: 25954001]
14. Consortium GT, Laboratory DA, et al., Nature, 2017, 550, 204–213. [PubMed: 29022597]
15. Schaub MA, Boyle AP, et al., Genome research, 2012, 22, 1748–1759. [PubMed: 22955986]
16. Ashburner M, Ball CA, et al., Nature genetics, 2000, 25, 25–29. [PubMed: 10802651]
17. Lee Y, Li H, et al., J Am Med Inform Assoc, 2013, 20, 619–629. [PubMed: 23355459]
18. Yue Z, Arora I, et al., BMC bioinformatics, 2017, 18, 532. [PubMed: 29297292]
19. Fagny M, Paulson JN, et al., PNAS USA, 2017, 114, E7841–e7850. [PubMed: 28851834]
20. Zhang J, Jiang K, et al., PloS one, 2015, 10, e0116477. [PubMed: 25803826]
21. Sanseau P, Agarwal P, et al., Nature biotechnology, 2012, 30, 317.
22. Li H, Achour I, et al., NPJ Genom Med, 2016, 1.
23. Califano A, Butte AJ, et al., Nat Genet, 2012, 44, 841–847. [PubMed: 22836096]
24. Boyle EA, Li YI and Pritchard JK, Cell, 2017, 169, 1177–1186. [PubMed: 28622505]
25. Mullen J, Cockell SJ, et al., PloS one, 2016, 11, e0155811. [PubMed: 27196054]
26. Law V, Knox C, et al., Nucleic Acids Res, 2014, 42, D1091–1097. [PubMed: 24203711]
27. MacArthur J, Bowler E, et al., Nucleic acids research, 2017, 45, D896–D901. [PubMed: 27899670]
28. Consortium GT, Science, 2015, 348, 648–660. [PubMed: 25954001]
29. Battle A, Brown CD, et al., Nature, 2017, 550, 204–213. [PubMed: 29022597]
30. Gene Ontology C, Nucleic Acids Res, 2015, 43, D1049–1056. [PubMed: 25428369]
31. Lussier YA, Rothwell DJ and Cote RA, Methods Inf. Med, 1998, 37, 161–164. [PubMed: 9656658]

32. Lin D, Icml, 1998, 98, 296–304.
33. Sánchez D, Batet M and Isern D, Knowledge-Based Systems, 2011, 24, 297–303.
34. Tao Y, Li J, et al., Bioinformatics (Oxford, England), 2007, 23, i529–i538.
35. Okada Y, Wu D, et al., Nature, 2014, 506, 376–381. [PubMed: 24390342]
36. Lovell DJ, Giannini EH, et al., NEJM, 2000, 342, 763–769. [PubMed: 10717011]
37. Moreland LW, Schiff MH, et al., Annals of internal medicine, 1999, 130, 478–486. [PubMed: 10075615]
38. Tfelt-Hansen P and Tfelt-Hansen J, Headache, 2009, 49, 117–125. [PubMed: 19125880]
39. Lettre G, Palmer CD, et al., PLoS genetics, 2011, 7, e1001300. [PubMed: 21347282]
40. Matsuo H, Yamamoto K, et al., Annals of the rheumatic diseases, 2016, 75, 652–659. [PubMed: 25646370]
41. Choi HK, Soriano LC, et al., Bmj, 2012, 344, d8190. [PubMed: 22240117]
42. Matsumoto T, Nakamura I, et al., Clinical rheumatology, 2013, 32, 271–275. [PubMed: 23149905]

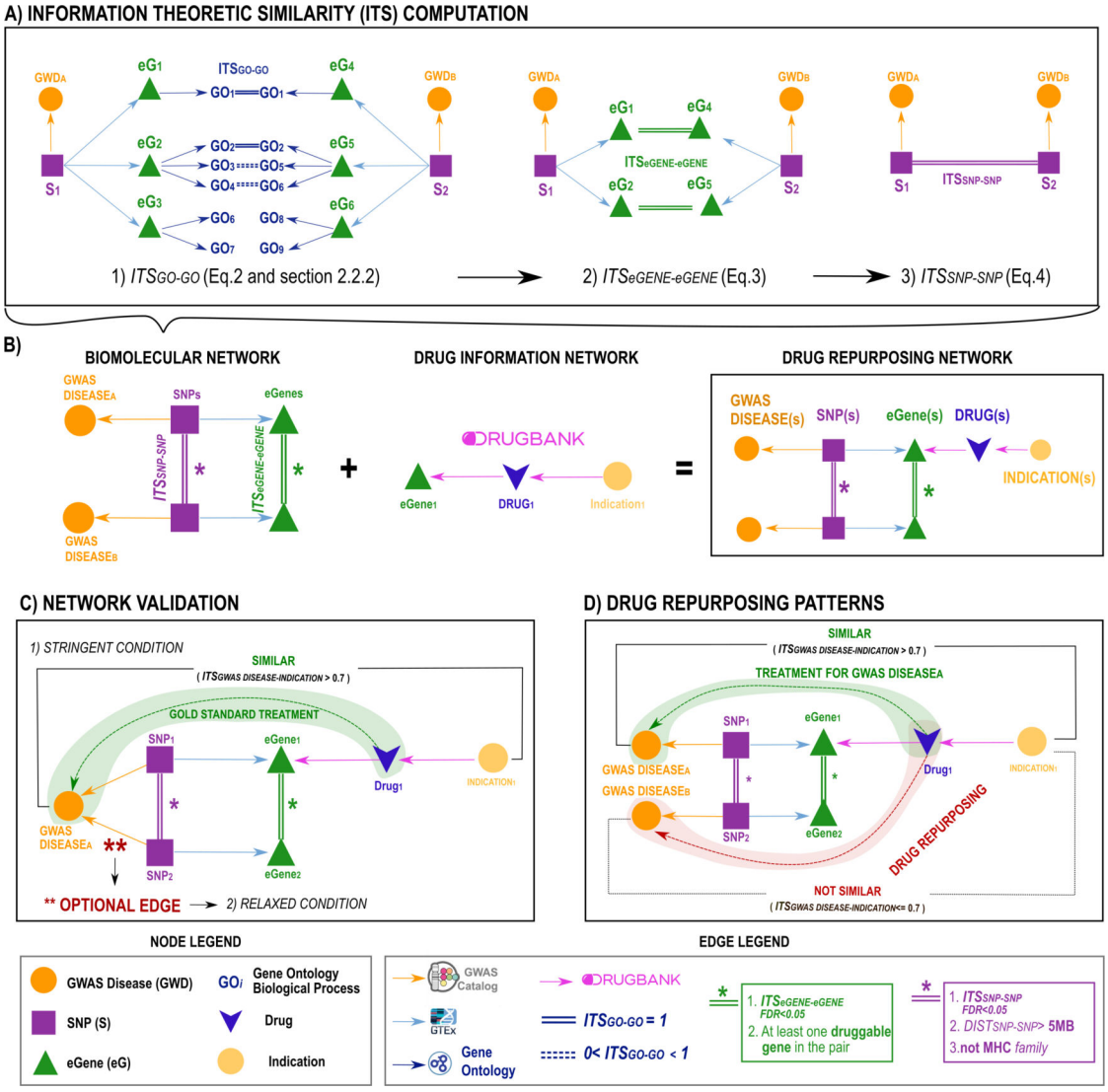
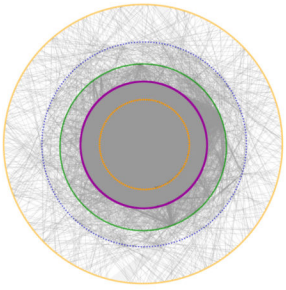


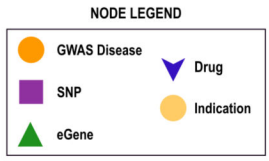
Fig. 1. Overview of the construction, computational prioritization, and validation of the drug repurposing network.

A) *ITS computation.* We applied ITS to compute the similarity between GO-BPs, SNPs, and genes through a cascade process as described in Methods 2.2.3–5. This began construction of the biomolecular network layer. **B)** *Integration of multiscale biomolecular associations* using GWAS diseases, SNPs, and eGenes as nodes. The associations (edges) between nodes were obtained by extracting GWAS disease-to-SNP, and SNP-to-eGene (SNPeG) relationships from the database resources described (Methods 2.1). The biomolecular network was then filtered to remove SNP-SNP pairs not meeting the introduced criteria (Edge Legend, Methods 2.2.6.). $ITS_{SNP\text{-}SNP}$ is computed as in Eq.4 considering all the eGenes extracted from eQTL data and the network was further refined to include only significantly similar eGene-eGene pairs, i.e. $ITS_{SNP\text{-}SNP}$ and $ITS_{eGENE\text{-}eGENE}$ (Eq.3) False Discovery Rate (FDR) <0.05. Drug-eGene and Drug-indication associations extracted from Drugbank (drug information layer) are included to obtain the final drug-repurposing network. **C)** *Network validation.* The drug repurposing network is validated by querying if

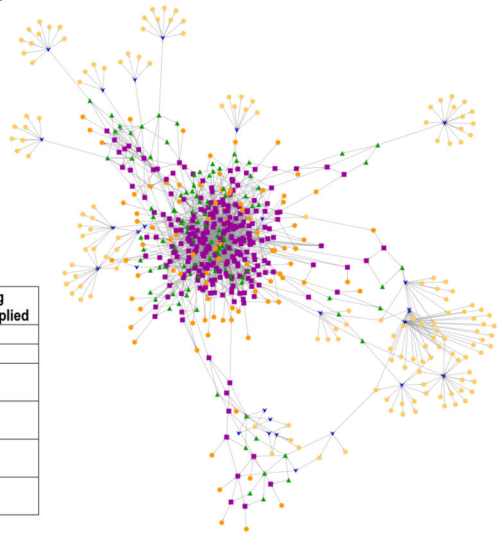
the network predicted a significantly high number of gold standard treatments for GWAS diseases. Two conditions of validation are proposed, one stringent and one more relaxed (**). **D)** *Drug repurposing patterns*. We extracted GWAS disease pairs and the related convergent mechanisms where at least a gold standard treatment was predicted for one of the two GWAS diseases. The approach predicts new candidate therapies by repositioning drugs across these GWAS disease pairs.

A)





C)



B)

Nodes	#
GWAS diseases	128
SNPs	586
eGenes	347
Drugs	231
Indications	573
	1,865

Edges	#	Mean node outdegree in integrated drug repurposing network after ITS filters applied	
SNP-SNP pairs	9,418	3.1	SNP
eGene -eGene pairs	536	3.1	eGene
GWAS disease-SNP	2,703	36.9 1.2	SNPs per GWAS disease GWAS diseases per SNP
SNP-eGene	1,771	3.0 5.1	eGenes per SNP SNPs per eGene
eGene-drug	361	4.3 1.2	Drugs per eGene target Targets per drug
Drug-indication	866	3.8 1.5	Disease indications per drug Drugs per disease indication
	15,655		

Fig. 2. Drug Repurposing Network.

A) Comprehensive biomolecular network comprising significant convergent *cis*- and *trans*-eQTL mechanisms between GWAS disease-associated SNPs ($ITS_{SNP\text{-}SNP}$ FDR<0.05; $ITS_{eGENE\text{-}eGENE}$ FDR<0.05), for which there exists indications in DrugBank (i.e., the molecular target of $Drug_i$ is the protein transcribed by at least one eGene associated by eQTL to SNP-SNP $Pair_x$; Fig.1B; Methods 2.2). **B)** Tables summarizing the number of nodes and edges of the network shown in panel A; for each edge type we also reported the mean node outdegree. **C)** Prioritized subset of the network in panel A relevant for drug repurposing because it satisfies one additional criteria: the disease indication of a $Drug_i$ is identical or similar to the GWAS disease associated to the SNP-SNP $Pair_x$ related to the eGene targeted by the $Drug_i$ (Fig.1C; Methods 2.4; $ITS_{GWAS\_Disease\text{–}DrugBank\_Indication}$>0.7). In Supplementary Material -Figure S1, we reported a high-resolution version of this network with labeled network node names.

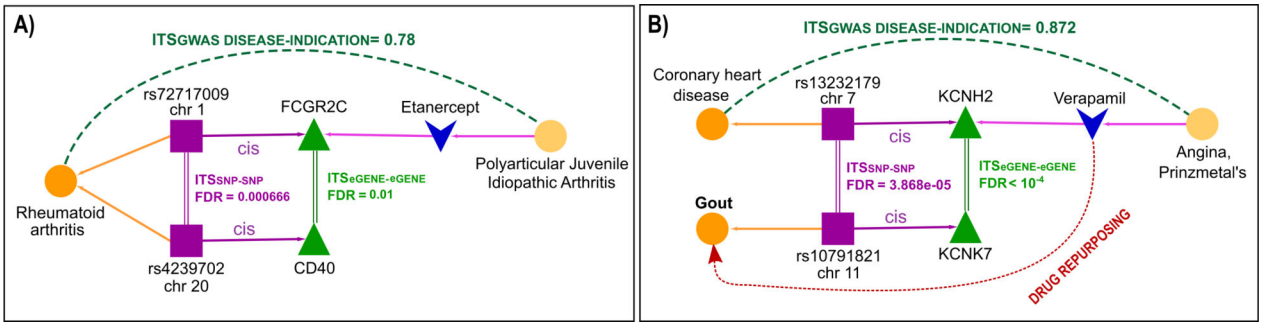


Fig. 3. Examples of prediction by eQTL signal convergence across distinct chromosomes.

A) *Gold standard validation.* In the drug repurposing network, we could confirm Etanercept as standard treatment for Rheumatoid arthritis. **B)** *Drug repurposing.* Our approach was able to predict Verapamil as a new potential treatment for gout, for which a retrospective study reports lower incidence of gout.