The integrated disease network†

Kai Sun, a Natalie Buchan, b Chris Larminie* b and Nataša Pržulj* a

The growing body of transcriptomic, proteomic, metabolomic and genomic data generated from disease states provides a great opportunity to improve our current understanding of the molecular mechanisms driving diseases and shared between diseases. The use of both clinical and molecular phenotypes will lead to better disease understanding and classification. In this study, we set out to gain novel insights into diseases and their relationships by utilising knowledge gained from system-level molecular data. We integrated different types of biological data including genome-wide association studies data, disease–chemical associations, biological pathways and Gene Ontology annotations into an Integrated Disease Network (IDN), a heterogeneous network where nodes are bio-entities and edges between nodes represent their associations. We also introduced a novel disease similarity measure to infer disease–disease associations from the IDN. Our predicted associations were systemically evaluated against the Medical Subject Heading classification and a statistical measure of disease co-occurrence in PubMed. The strong correlation between our predictions and co-occurrence associations indicated the ability of our approach to recover known disease associations. Furthermore, we presented a case study of Crohn’s disease. We demonstrated that our approach not only identified well-established connections between Crohn’s disease and other diseases, but also revealed new, interesting connections consistent with emerging literature. Our approach also enabled ready access to the knowledge supporting these new connections, making this a powerful approach for exploring connections between diseases.

1 Introduction

Recent advances in ‘omics’ technologies have led to a wealth of biological data. These data, including genetic data (e.g., data obtained by genome-wide association studies, GWAS1–4), genomic data (e.g., microarray data5–7), proteomic data (e.g., yeast two-hybrid screening data8–11 or affinity capture mass spectrometry data12–14), metabolomic data (e.g., metabolite data derived from mass spectrometry-based studies15–17) and transcriptomic data (e.g., RNA-Sequencing data18–20), have been used to improve our understanding of the underlying mechanisms of human diseases. Since various types of biological data reside in diverse repositories, data integration has become crucial, as it not only provides a comprehensive view of data, but also enables further exploration of biological knowledge by efficiently mining and analysing data.

One of the major tasks of biological data integration is to collect and combine heterogeneous data from different repositories. Repositories such as Online Mendelian Inheritance in Man (OMIM)21 for phenotype–genotype associations, the Biological General Repository for Interaction Datasets (BioGRID)22 for protein–protein interactions (PPIs) and genetic interactions, the National Human Genome Research Institute (NHGRI)
molecular data, measuring the correlation between disease-relationships and to reposition drugs from multiple unequivocally. During the past decade, some studies have been clinical symptoms appear, and specificity in defining disease classification lacks sensitivity in identifying disease before phenotypes. However, it has been widely recognised that such diseases is mainly derived from the similarity of clinical biological data integration. Current classification of human logies such as GO, DO and Human Phenotype Ontology networks and comparing annotations in biomedical onto-

We proposed to uncover novel associations between diseases via biological data integration. Current classification of human diseases is mainly derived from the similarity of clinical phenotypes. However, it has been widely recognised that such classification lacks sensitivity in identifying disease before clinical symptoms appear, and specificity in defining disease unequivocally. During the past decade, some studies have been proposed to improve our understanding of the relationship between diseases using large-scale biological data. Disease–disease associations have been inferred by different computational approaches, including text mining the literature, analysing disease-associated SNPs from GWAS data, fusing systems-level molecular data, measuring the correlation between disease-related gene expression data, constructing disease–disease networks and comparing annotations in biomedical ontologies such as GO and Human Phenotype Ontology (HPO). We believe a network-based, integrative approach that takes advantage of various types of biological data will provide further insights into disease relationships.

In this study, we integrated different types of biological data collected from diverse repositories, including disease–gene associations, disease–chemical associations, biological pathways and GO annotations, to gain novel insights into human diseases and their relationships. In particular, we integrated these data into an Integrated Disease Network (IDN), a heterogeneous network where nodes are bio-entities and edges between nodes represent their associations. To uncover novel disease associations, we developed a novel computational approach that estimates the similarity between diseases. Our predicted associations were systematically evaluated against the hierarchical tree of Medical Subject Headings (MeSH) and a statistical measure of disease co-occurrence in the literature. Furthermore, we presented a case study of Crohn’s disease to illustrate how the integrated disease network could be used to improve our current understanding of disease relationships.

2 Results and discussions
2.1 Construction of the IDN
We constructed the IDN by integrating large-scale biological data, including disease–gene associations, disease–chemical associations, biological pathways and GO annotations. These data were collected from a number of repositories, as demonstrated in Fig. 1. All diseases annotated in the repositories were mapped to MeSH disease terms and all genes were mapped to Entrez gene IDs for the purpose of integration. Disease–gene associations were collected from three repositories: the Comparative Toxicogenomics Database (CTD), Functional Disease Ontology (FunDO) and GWAS catalog. All of these repositories contain curated disease–gene associations extracted from the literature. Note that in this study, we use a broad definition of disease–gene association: a gene is considered to be associated with a disease if their connection has been reported by the literature. Therefore, a disease–gene association is not necessarily genetic, as it might be derived from non-genetic approaches. We did not include the OMIM database because the CTD database collects disease–gene associations derived from OMIM. Among the three repositories, the GWAS catalog has fewer disease–gene associations compared to CTD and FunDO (Fig. 1), since it contains only associations derived from a subset of diseases for which genetic studies have been conducted. However, for a specific disease (e.g., Crohn’s disease, see Section 2.3 for details), the GWAS catalog may contain more associations than the others. In addition, associations in the GWAS catalog may be more robust and reflect causal associations between gene and disease. Since the three repositories focus on different aspects, the overlap among them is small: only 88 associations are supported by all three datasets (see Fig. S1, ESI† for details). Therefore, integrating the three datasets largely increases the coverage of disease–gene associations.

Disease–pathway and disease–GO term associations were inferred from disease–gene, gene–pathway, gene–GO term associations by using enrichment analysis (Fig. 1, also see Section 3.2 for details). A pathway (a GO term) is considered to be associated with a disease if it is significantly overrepresented within the set of genes associated with that disease. To investigate how the number of inferred associations deviates from random, we randomly rewired disease–gene associations, while keeping the number of associations of each disease and gene unchanged (also see Section 3.3 for details). The randomisation process

‡ http://www.nlm.nih.gov/mesh/
was repeated 100 times. The average number of disease–pathway (disease–GO term) associations obtained from random disease–gene associations is 10,378/\binom{C_6}{223} (2395/\binom{C_6}{57}), which is 2.84-fold (3.60-fold) lower than the actual number of significant disease–pathway (disease–GO term) associations (p-value < 0.01). These differences suggested that genes associated with the same disease are more likely to share common biological functions than those associated with different diseases.

In total the IDN consists 47,342 nodes representing different bio-entities and 432,660 edges representing the associations between them (Table 1). On average, a disease node is connected to 9.93 gene nodes, 26.49 chemical nodes, 26.09 pathway nodes and 8.62 GO term nodes in the IDN; a gene node is connected to 4.35 disease nodes, 14.22 pathway nodes and 9.09 GO term nodes. The largest connected component of the IDN contains 46,235 (97.66%) nodes, covering 5541 (91.02%) diseases. Table S3 (ESI†) lists the top 20 nodes that have the highest degrees in the IDN. The high degree nodes in the IDN are mainly general GO terms (e.g., ‘Protein binding’), common pathways (e.g., ‘Signal transduction’) and well-studied or complex diseases (e.g., ‘Neoplasms’). In the connected network, the average distance between a pair of disease nodes is 4.06 (note that the minimum distance between two diseases is 2). These results indicate the interconnectedness of human diseases.

2.2 Inferring disease–disease associations

We proposed a novel integration-based similarity measure to infer disease–disease associations from the IDN. In particular, we assigned a score ranging from 0 to 1 to qualify the strength of the association between a pair of diseases in the IDN. A similarity score of 0 suggests that the two diseases are not directly associated: either they have no common neighbours in the IDN, or their association score is not higher than expected at random (see Section 3.3). We only considered the 711 diseases that are associated with at least one gene, one chemical, one pathway and one GO term in the evaluation. A disease–disease network can be constructed based on the similarity scores. When considering only the top 1% associations as the edges in the disease–disease network, the network contains 711 nodes (including 101 isolated nodes) and 2,525 edges (Fig. S2, ESI†), and the average clustering coefficient of the network is 0.419. When considering the top 5% associations as the edges in the disease–disease network, the network contains 711 nodes (including only 5 isolated nodes) and 12,621 edges, and the
average clustering coefficient becomes 0.446. To compare the clustering coefficient of the disease–disease network to random networks, we generated 30 Erdős–Rényi random graphs with the same number of nodes and edges of the corresponding disease–disease network and computed their clustering coefficients. The average clustering coefficients of ER random graphs with the same size of the top 1% and 5% disease–disease networks are 0.014 ± 0.017 and 0.051 ± 0.001 respectively. The high clustering coefficient of the disease–disease network indicates that diseases having common neighbours in the network tend to be connected with each other, as shown in Fig. S2 (ESI†). Since diseases that belong to the same cluster are expected to have common underlying mechanisms, these results also suggest the potential use of the IDN to redefine disease classifications.

To assess the reliability of our similarity measure, we first evaluated our results against the MeSH classification§ by comparing similarity scores of diseases from the same MeSH tree branch to those from different MeSH branches. For example, Crohn’s disease and ulcerative colitis (UC) are both in the ‘Digestive System Diseases’ branch, while sarcoidosis (SA) is in the ‘Hemic and Lymphatic Diseases’ branch. We found that diseases from the same MeSH branch have substantially higher similarity scores (0.0174 ± 0.0368) than diseases from different MeSH branch (0.0066 ± 0.0199).

We further evaluated our similarity measure against the co-occurrence of disease MeSH terms in the literature. To obtain co-occurrence scores for all pairs of diseases we analysed, we used a statistical approach proposed by Li and Agarwal (2009). In this approach, the co-occurrence score of a pair of diseases is measured by the statistical significance of their co-occurrence in the PubMed¶ abstracts. We collected 15.29 million abstracts from PubMed covering articles published from January 1900 to October 2013, and identified disease MeSH terms that were associated with these abstracts as major MeSH Headings. Statistical significance of disease co-occurrence was computed using one-tailed Fisher’s exact test and adjusted by Benjamini–Hochberg correction (see ref. 52 for details). In total, we collected 8459 significant co-occurrence associations (i.e., disease pairs having a corrected p-value < 0.05) between the 711 diseases we evaluated.

We used ROC curves analysis to investigate the correlation between our similarity scores with disease co-occurrence scores. Fig. 2 and Table S1 (ESI†) show the ROC curves and AUC values obtained by evaluating our similarity measure. We found that the correlation between our integration-based similarity scores and disease co-occurrence scores is substantially higher than expected at random. In addition, our similarity measure tends to detect strong disease co-occurrence associations with small p-values (e.g., p-value < 10^{-20}). When setting the co-occurrence p-value threshold to 0.05 (i.e., all disease pairs having a co-occurrence p-value less than 0.05 are considered to be associated), 8459 pairs in total), we obtained an AUC value of 0.7580 (Table S1, ESI†). This value increased to 0.7646 with a p-value threshold of 0.01 (7713 co-occurrence associations in total), and 0.8282 with a p-value threshold of 10^{-20} (3330 co-occurrence associations in total). We also plotted the percentage of our inferred disease–disease associations versus the percentage of the co-occurrence associations recovered by our approach to assess the discovery rate (Fig. S3, ESI†). Among the 3330 strong co-occurrence associations (i.e., p-value < 10^{-20}), 2831 disease pairs (85.01%) have similarity scores higher than 0. When considering only the top 1% (5%) inferred associations, we are able to recover 14.65% (43.69%) of the co-occurrence associations. The top 35% inferred associations cover all 2831 co-occurrence associations. These results demonstrate the ability of our approach to recover known disease–disease associations.

More importantly, our results reveal the need of data integration for uncovering biological knowledge. We show that using the integration of disease-related data leads to better performance in recovering disease co-occurrence associations than solely using one type of data (Fig. 2 and Table S1, ESI†). Among the four types of disease-related data, disease–gene association seems to be the most effective in predicting disease associations, followed by disease–chemical association, disease–pathway association and disease–GO term association. However, this result may also suggest that analysing shared biological pathways and shared GO annotations between diseases may offer the opportunity to uncover previously undiscovered disease associations.

2.3 Crohn’s disease: a case study

To illustrate how our approaches can be used to improve our current understanding of human diseases and their relationships, we present a case study of Crohn’s disease. Crohn’s disease,
also known as regional enteritis, is a chronic relapsing inflammatory disorder that can affect any part of the gastrointestinal tract. It can cause a wide variety of clinical symptoms, including abdominal pain, diarrhoea, and fever.\textsuperscript{59} Recent studies suggested that Crohn’s disease can be considered as an abnormal immune response to the bacteria in genetically susceptible individuals.\textsuperscript{60,61} However, the exact underlying cause of Crohn’s disease is still unknown.

### 2.3.1 Identifying disease associations

In the IDN, Crohn’s disease is linked to 370 gene nodes (40 from CTD, 83 from FunDO, and 285 from GWAS Catalog). Among them only 5 genes are reported in all three repositories, including tumor necrosis factor (TNF), interleukin 10 (IL10), nucleotide-binding oligomerization domain containing 2 (NOD2), immunity-related GTPase family, M (IRGM) and cyclin Y (CCNY). These 370 Crohn’s disease associated genes are significantly enriched in 120 biological pathways (39 from KEGG, 17 from Reactome, 11 from WikiPathways and 53 from GeneGO) and 47 GO terms (5 cellular component terms, 39 biological process terms, and 3 molecular function terms). There are also 25 chemical nodes that are connected to Crohn’s disease in the IDN.

Fig. 3 shows the top 20 diseases associated with Crohn’s disease identified by our approach. 5 diseases among them are also associated with Crohn’s disease according to the literature co-occurrence measure (p-value < 0.05). After grouping these 20 diseases based on the MeSH tree classification, we found that 4 out of them are under the same MeSH branch ‘Digestive System Diseases’ with Crohn’s disease. Meanwhile, our results show that Crohn’s disease is closely related to 5 diseases classified as ‘Bacterial Infections and Mycoses’ (e.g., leprosy) and 2 diseases classified as ‘Immune System Diseases’ (e.g., multiple sclerosis). This finding supports the most widely held hypothesis that Crohn’s disease is caused by the dysfunctional interaction between the intestinal immune system and a subset of commensal enteric bacteria.\textsuperscript{60,61} Apart from the 5 known disease associations (i.e., diseases that are associated according to the MeSH classification or the co-occurrence measure), the remaining 15 associations are considered as novel associations inferred by our approach. High similarity scores and non-significant co-occurrence p-values of these novel associations suggest they may have been overlooked in the previous literature.

### 2.3.2 Investigating common mechanisms

We further investigated the common underlying mechanisms of Crohn’s disease and its top 1% associated diseases, including inflammatory bowel diseases (IBD), UC, SA, psoriatic arthritis (PA), Behçet syndrome (BS), leprosy (LE) and celiac disease (CE). We generated an induced subgraph from the IDN representing the associations between the 8 diseases and their common annotated genes (Fig. 4). We showed that our method successfully linked Crohn’s disease and UC to the umbrella term IBD. We also identified the associations between Crohn’s disease and CE, which is another gastrointestinal autoimmune disease. BS, an autoimmune disease with highly heterogeneous presentation that includes gastrointestinal inflammation which is sometimes histologically difficult to distinguish from Crohn’s disease,
was also strongly associated with Crohn’s disease in our results. PA might also be expected to be linked to Crohn’s disease as it is a closely related disorder, while linkage to infective disorders such as LE may reflect the importance of the gut microbiome in Crohn’s disease. There is also a hypothesis that LE may cause Crohn’s disease supported by striking overlap in the genetic control of these two diseases. These results highlight the potential of our method to identify and group similar diseases through different data types.

As can be seen in Fig. 4, our method identified key genes contributing to the associations between the diseases. Many of these serve as positive controls as they are established in the literature to be linked to Crohn’s disease and at least one other of the seven disease annotations, such as NOD2, TNF, IL12B, and IL23R. Indeed, anti-TNF antibodies are licensed for the treatment of Crohn’s disease, PA, UC within the list of seven diseases, as well as other autoimmune diseases and the anti-IL12/23 monoclonal antibody ustekinumab has also demonstrated preliminary efficacy in Crohn’s disease. The ability of our approach to recapitulate established associations between diseases and summarise the factors contributing to these connections (Fig. 4, Fig. S4 and S5, ESI†) gave us confidence to explore other genes identified by our approach.

We selected pseudouridylate synthase 10 (PUS10) for further investigation as the gene is not widely established in the literature as being involved in Crohn’s disease or the other diseases it connects (UC and CE). Seven papers contributed to the PUS10 linkage to Crohn’s and other gastrointestinal autoimmune diseases identified in our approach. These papers focus on the genetics of Crohn’s disease, CE and UC and include genetic association of PUS10 with these diseases (the PubMed IDs of these papers are: 23128233, 21297633, 20228799, 21298027, 20190752, 22412388 and 21102463). We first utilised I2E from Linguamatics Ltd, Cambridge, UK, a specialist text mining tool to explore the published literature space. This approach did not identify significantly more references than the method employed in this paper and suggests the search space was comprehensive. Next, we interrogated the European Bioinformatics Institute (EBI) Expression Atlas to determine the tissue distribution of PUS10 mRNA. We found that PUS10 is primarily expressed in the small intestine/duodenum with elevated expression also observed in lymph node, liver and testis consistent with a role in the gastrointestinal tract and immune mechanisms. However, no differential expression of PUS10 was observed in disease control studies of Crohn’s disease, UC or CE using Expression Atlas. In addition, we explored the pathways through which PUS10 could play a role in Crohn’s disease, but we were unable to identify any annotated pathways that contained PUS10. However, PUS10 (also known as DOBI) is known to modulate TNF-related apoptosis-inducing ligand (TRAIL)-induced apoptosis. This provides a plausible mechanism through which PUS10 might exert an effect in Crohn’s disease since TRAIL is known to play an important role in mediating cell apoptosis. Supporting this, apoptosis and necroptosis either of immune cells or epithelial cells is thought to have an important role in Crohn’s pathogenesis, and the TRAIL–TRAIL-receptor system shows altered expression in IBD sufferers. Interestingly, gastrointestinal bacterial pathogens are able to control apoptotic signalling in human intestinal cells to establish infection. This is achieved by virulence effector proteins that translocate to the host cell cytoplasm and target death receptor signalling cascades that may include blockade of TRAIL-induced cell death. Taken together, these observations suggest that the TRAIL death receptor pathway may play an important role in the pathogenesis of Crohn’s disease, and that this may be modulated by the gut microbiota.

The method described in this paper may be a very effective knowledge mining tool. The method is able to survey and summarise a large molecular knowledge space and still provide the user with direct access to the underlying evidence. Critically, the implementation of disease similarity measure was highly effective at drawing out and specifically presenting knowledge that is critical in driving the observed connectivity between the diseases drawn from a relatively small corpus of literature, as seen with PUS10. This is particularly important when our knowledge of biology, and that reported in the literature, is so variable and knowledge rich areas of biology have a tendency to bias/smother other salient factors.

2.3.3 Predicting drug repositioning opportunities. Disease–disease associations uncovered by our approach can be further used to explore drug repositioning opportunities. Drug repositioning
(i.e., identifying new disease indications for approved drugs) is considered as one of the most effective strategies in drug discovery. It can not only substantially reduce the costs and time of drug development, but also provide safer treatment for patients as it lowers the risk of unexpected toxicity and side effects.\(^{88}\) Recently, many computational approaches have been developed for drug repositioning. These approaches can be roughly classified as 'drug-based' or 'disease-based'.\(^{83}\) 'Drug-based' approaches infer repositioning opportunities from chemical or pharmaceutical perspectives, such as the chemical properties\(^{84,85}\) and molecular activities\(^{86}\) of drug compounds. 'Disease-based' approaches infer repositioning opportunities from clinical perspectives, such as drug indications\(^{48,49,87}\) and side effects.\(^{85,88–90}\) While most of these approaches have focused on either drug–target or drug–disease associations, only a few studies have provided an integrated approach to combine different types of associations for drug repositioning. Sanseau et al. (2012)\(^{91}\) presented an analysis that for the first time incorporated drug–target associations with GWAS data to identify potential drug repositioning opportunities. Another integrated approach for drug repositioning was proposed by Daminelli et al. (2012).\(^{40}\) They constructed a drug–target–disease network by integrating disease–chemical and drug–target associations, and mined the network for network motifs of bi-cliques in order to predict disease–drug associations. However, the major limitation of this approach is that the predictions may have a high false positive rate since they were inferred solely based on the network structure.\(^{40}\)

Here we present an integrated approach for computational drug repositioning. We obtained disease–drug–target associations from Drugbank 3.0\(^{92}\) and Informa Pipeline\(^*\) in November 2013 (data from Informa Pipeline were licensed). From all genes linked to Crohn's disease in the IDN, we focused on the 190 genes that are associated with at least one of the other 7 diseases. 39 out of these genes encode proteins that are reported as drug targets in Drugbank or Informa Pipeline. Note that a drug–target association reported in Drugbank might be inferred from the literature, thus a target in Drugbank is not necessarily druggable. We found that 6 of these 39 genes, including C–C motif chemokine 11 (CCL11), IL12B, interleukin 13 (IL13), interleukin 6 (IL6), toll-like receptor 4 (TLR4) and TNF, have been targeted for the treatment of Crohn's disease (Table 2). The drugs developed for these targets may provide repositioning opportunities for diseases that are closely associated with Crohn's disease. Meanwhile, drugs that have been developed for these associated diseases may be candidates for the treatment of Crohn's diseases (Table 3). For example, thalidomide, an immunomodulatory drug that inhibits the production of TNF, is approved by the U.S. Food and Drug Administration (FDA) for the treatment of leprosy. Several recent clinical studies (e.g.,\(^ {93,94}\)) have shown the positive effect of thalidomide on the remission of Crohn's disease, supporting the potential repositioning of thalidomide for Crohn's disease.

** http://sites.informahealthcare.com/pipeline/

<table>
<thead>
<tr>
<th>Drug</th>
<th>Target</th>
<th>Stage</th>
<th>Associated diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlimimab</td>
<td>CCL11</td>
<td>Preclinical</td>
<td>IBD, LE</td>
</tr>
<tr>
<td>Ustekinumab</td>
<td>IL12B</td>
<td>Phase III</td>
<td>All top 7 except SA</td>
</tr>
<tr>
<td>QAX-576</td>
<td>IL13</td>
<td>Phase II</td>
<td>IBD, UC</td>
</tr>
<tr>
<td>Clazakizumab</td>
<td>IL6</td>
<td>Phase II</td>
<td>IBD, UC</td>
</tr>
<tr>
<td>PF-4236921</td>
<td>IL6</td>
<td>Phase II</td>
<td>IBD, UC</td>
</tr>
<tr>
<td>Ololizumab</td>
<td>IL6</td>
<td>Phase II</td>
<td>IBD, UC</td>
</tr>
<tr>
<td>Naltrexone</td>
<td>TLR4</td>
<td>Phase II</td>
<td>UC, SA, BS, CE</td>
</tr>
<tr>
<td>VT-346</td>
<td>TNF</td>
<td>Preclinical</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>Ozoralizumab</td>
<td>TNF</td>
<td>Phase I</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>XPro-1595</td>
<td>TNF</td>
<td>Preclinical</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>Certolizumab pegol</td>
<td>TNF</td>
<td>Launched</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>ESBA-105</td>
<td>TNF</td>
<td>Phase II</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>Adalimumab</td>
<td>TNF</td>
<td>Launched</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>Infliximab</td>
<td>TNF</td>
<td>Launched</td>
<td>All top 7 diseases</td>
</tr>
<tr>
<td>PF-05230905</td>
<td>TNF</td>
<td>Phase I</td>
<td>All top 7 diseases</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug</th>
<th>Target</th>
<th>Indication</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI-201</td>
<td>IL12B</td>
<td>UC</td>
<td>Phase II</td>
</tr>
<tr>
<td>Tralokinumab</td>
<td>IL13</td>
<td>UC</td>
<td>Phase II</td>
</tr>
<tr>
<td>Anrakizumab</td>
<td>IL13</td>
<td>UC</td>
<td>Phase II</td>
</tr>
<tr>
<td>VT-384</td>
<td>IL18</td>
<td>IBD</td>
<td>Phase I</td>
</tr>
<tr>
<td>GSK-1070806</td>
<td>IL18</td>
<td>IBD</td>
<td>Phase I</td>
</tr>
<tr>
<td>Interleukin-2</td>
<td>IL2RA</td>
<td>LE</td>
<td>Launched</td>
</tr>
<tr>
<td>CCR-294</td>
<td>NO52</td>
<td>IBD</td>
<td>Phase II</td>
</tr>
<tr>
<td>DCAM-253</td>
<td>RIPK2</td>
<td>IBD</td>
<td>Preclinical</td>
</tr>
<tr>
<td>Thalidomide</td>
<td>TNF</td>
<td>LE</td>
<td>Launched</td>
</tr>
<tr>
<td>Golimumab</td>
<td>TNF</td>
<td>UC</td>
<td>Launched</td>
</tr>
<tr>
<td>Golimumab</td>
<td>TNF</td>
<td>SA</td>
<td>Phase</td>
</tr>
<tr>
<td>AG-014</td>
<td>TNF</td>
<td>IBD</td>
<td>Preclinical</td>
</tr>
<tr>
<td>HL-036</td>
<td>TNF</td>
<td>IBD</td>
<td>Preclinical</td>
</tr>
<tr>
<td>Etanercept</td>
<td>TNF</td>
<td>PA</td>
<td>Launched</td>
</tr>
<tr>
<td>AVX-470</td>
<td>TNF</td>
<td>UC</td>
<td>Phase I</td>
</tr>
</tbody>
</table>

### 3 Materials and methods

#### 3.1 Data collection

We collected disease–gene associations, disease–chemical associations, biological pathways and GO annotations from a number of repositories. Since different repositories may use different disease naming schemes, we mapped all diseases annotated in these data to MeSH terms for a correct integration. Details of data sources and mapping approaches are described below.

#### 3.1.1 Disease–gene associations

Disease–gene associations were collected from three repositories: CTD, FunDO and GWAS catalog.

CTD provides curated disease–gene associations that were extracted from the published literature by CTD curators, or were derived from the OMIM database. We downloaded disease–gene associations from CTD in July 2013, and collected 21 625 curated associations between 4286 diseases (annotated by MeSH terms) and 6940 genes.
FunDO contains Disease Ontology (DO)\footnote{http://disease-ontology.org/} annotations of the human genome extracted from the NCBI Gene Reference Into Function (GenRIF) database.\footnote{http://www.ncbi.nlm.nih.gov/gene/about-generif} We used the latest stable version of FunDO (released in October 2008), which contained gene annotations for 1854 diseases (annotated by DO IDs). We then mapped DO IDs to MeSH terms by using the mapping provided in DO version 3 (the latest stable version, released in May 2007). After mapping, the dataset contained 1255 diseases (mapped from 1528 DO IDs), 4886 genes and 22 879 disease–gene associations.

GWAS catalog associates collections between diseases (or traits) and single-nucleotide polymorphisms (SNPs) identified by GWAS studies. We downloaded GWAS catalog data in August 2013. Similar to Sanseau et al. (2012),\textsuperscript{91} we eliminated unreplicated associations and associations with p-value higher than 10\textsuperscript{-7} to minimize false-positives. Diseases annotated in GWAS catalog were manually mapped to MeSH terms. In total, 6430 associations between 370 diseases and 3762 genes were collected from the GWAS catalog.

The set of disease–gene associations used in this study was the union set of CTD, FunDO and GWAS catalog. It contained a total of 48 549 associations between 4891 diseases and 11 160 genes.

### 3.1.2 Disease–chemical associations

Disease–chemical associations used in this study were downloaded from the CTD database in July 2013. CTD contains both curated and inferred disease–chemical associations, and we used only curated associations extracted by CTD curators. We collected 80 225 associations between 3029 diseases and 8201 chemicals.

### 3.1.3 Biological pathways

We collected pathway data from four repositories: KEGG, Reactome, WikiPathways\textsuperscript{45} and GeneGO (i.e., MetaBase from Thomson Reuters). All of these repositories contain manually curated pathway maps representing molecular reactions and interactions in a cell. We downloaded pathway data from the four repositories in September 2013 (pathway data from GeneGO were licensed). Only human pathways were considered in our study. In total, we collected 2319 biological pathways (276 pathways from KEGG, 1416 pathways from Reactome, 201 pathways from WikiPathways and 426 pathways from GeneGO) associated with 9992 genes.

### 3.1.4 GO annotations

GO annotations based on human evidence were downloaded from the GO database\textsuperscript{46} in September 2013. We removed annotations with evidence code ‘Inferred from Electronic Annotation’ (IEAs), as IEAs were computationally inferred annotations which had not been reviewed by curators. In total, we collected 123743 annotations between 13 609 genes and 11 266 GO terms (1034 cellular component terms, 7480 biological process terms, and 2752 molecular function terms).

### 3.2 Enrichment analysis

We inferred disease–pathway associations and disease–GO term associations by identifying pathways and GO terms that were significantly enriched within the set of genes associated with each disease. The statistical significance (p-value) of the enrichment was calculated according to the hypergeometric distribution for sampling without replacement (identical to one-tailed Fisher’s exact test). For a given disease $D_i$ and a given pathway (or GO term) $P_j$, let $|GP_j|$ be the number of genes associated with $D_i$, $|GP_j|$ be the number of genes annotated with $P_j$, and $|GP|$ be the total number of genes in the pathway (or GO annotation) data. If there were $x$ genes annotated with pathway $P_j$ that were associated with $D_i$ (thus $x = |GP_j \cap GP_i|$), we computed the p-value of the enrichment (i.e., the probability of observing the same or higher enrichment purely by chance) as:

$$p\text{-value} = 1 - \sum_{k=0}^{x-1} \left( \frac{|GP_j|}{k} \right) \left( \frac{|GP|-|GP_j|}{|GP|-k} \right).$$

We used the Bonferroni correction to adjust p-values for multiple comparison. A pathway or a GO term is said to be associated with a disease if the adjusted p-value is less than 0.05. Through this approach, we identified 29 432 disease–pathway associations between 1128 diseases and 1491 pathways, and 8614 disease–GO term associations between 999 diseases and 1714 GO terms.

### 3.3 Disease similarity measure

To uncover associations between diseases from the IDN, we developed a novel measure to estimate disease similarity scores. The similarity measure is based on the vector space model,\textsuperscript{27,28} which is frequently used in the area of information retrieval. Four types of disease-related data were used for computing similarity scores, namely disease–gene associations, disease–chemical associations, disease–pathway associations and disease–GO term associations. Let $E$ be the set of non-disease elements (i.e., genes, chemicals, pathways and GO terms) in the IDN. For each disease $D_i$, we first constructed an $|E|$-dimensional vector $\hat{D}_i$, called $D_i$’s signature, to represent information we gained for $D_i$:

$$\hat{D}_i = (D_{i,g}, D_{i,c}, \ldots, D_{i,p}, \ldots, D_{i,s} | E_i),$$

The $k$th element of $\hat{D}_i$ is defined as:

$$D_{ik} = w_{E_k} \times f(D_i, E_k),$$

where $f(D_i, E_k)$ is a function that indicates the presence or absence of the association between the disease $D_i$ and the non-disease element $E_k$:

$$f(D_i, E_k) = \begin{cases} 1 & \text{if } E_k \text{ is associated with } D_i \\ 0 & \text{else} \end{cases}$$

and $w_{E_k}$ is a weight (also known as the inverse document frequency (idf)) assigned to $E_k$. Let $D$ be the set of diseases annotated in the data, and $D_{E_k}$ be the set of diseases associated with $E_k$. The idf of $E_k$ is defined as:

$$w_{E_k} = \text{idf}(E_k, D) = \log \frac{|D|}{|D_{E_k}|}$$

# References

\textsuperscript{91} Sanseau et al. (2012).
Assigning the idf as a weight to each non-disease element $E_k$ allows us to increase the importance of $E_k$ if $E_k$ is rare across all diseases (i.e., $E_k$ is associated with only a few diseases). Meanwhile, the logarithm in eqn (4) prevents our results from being dominated by the most rare non-disease elements, since the number of associated diseases may greatly differ from one to another. For a pair of diseases $D_i$ and $D_j$, their similarity can be measured by the similarity between their corresponding signatures $\vec{D_i}$ and $\vec{D_j}$. Thus we computed the raw similarity score between $D_i$ and $D_j$ as the cosine of the angle $\theta_{ij}$ between $\vec{D_i}$ and $\vec{D_j}$:

$$\cos \theta_{ij} = \frac{\vec{D_i} \cdot \vec{D_j}}{||\vec{D_i}|| \cdot ||\vec{D_j}||} \quad (5)$$

If considering only the overlaps of non-disease elements, $D_i$ may always have a higher similarity score with a complex disease (i.e., disease that is caused by a combination of genetic, environmental, and lifestyle factors) than a monogenic disease (i.e., disease that is caused by a single gene). The cosine similarity reduces this bias: it compares disease signatures $D_i$ and $D_j$ on a normalised space, as it considers only their orientation and not magnitude. This enable that complex diseases do not always come out on top in computation of disease similarities. Similarly, the cosine similarity also reduces the bias caused by well-studied diseases in our analysis.

To further ensure that our results could not be obtained by random chance, we computed a random similarity score for each pair of diseases $D_i$ and $D_j$ to adjusted their raw similarity score $\cos \theta_{ij}$. To do this, we first generated random disease–gene associations by rewiring the associations: for a disease–gene association $(D_x, G_y)$, we randomly chose another association $(D_z, G_y)$ and replaced the two associations by $(D_x, G_y)$ and $(D_y, G_z)$, if $D_x \neq D_y$, $G_x \neq G_y$, and $G_y \neq G_z$. This process was repeated until all disease–gene associations were rewired. Similarly, we generated random disease–chemical associations. Random disease–pathway associations and random disease–GO term associations were obtained by running enrichment analysis using random disease–gene associations.

We then used these random association data to compute a random similarity score for each pair of diseases based on eqn (2)–(5). This process was repeated 100 times to generate the average random similarity score $\bar{\gamma}_{ij}$ for each pair of diseases $D_i$ and $D_j$.

Finally we computed the adjusted similarity score for diseases $D_i$ and $D_j$ as:

$$\text{Similarity} (D_i, D_j) = \max (\cos \theta_{ij} - \bar{\gamma}_{ij}, 0) \quad (6)$$

This adjustment further reduces the potential bias caused by complex diseases and well-studied diseases.

4 Conclusions

We integrated different types of large-scale biological data collected from different repositories with the aim of gaining further understanding of how diseases associated with each other. We constructed a heterogeneous network, i.e., the IDN, to represent the integrated data, where nodes are bio-entities such as diseases, genes, chemicals, pathways or GO terms, and edges between nodes represent the associations between these bio-entities. In addition, we proposed a novel integration-based similarity measure to infer disease–disease associations from the IDN. Our similarity measure was systematically evaluated against the MeSH tree classification and a statistical measure of disease occurrence in the literature, confirming the ability of our approach to recover known disease–disease associations. Furthermore, the case study of Crohn’s disease well demonstrated the ability our approach to identify previously undiscovered disease associations, investigate common underlying mechanisms driving diseases, and infer drug repositioning opportunities. Our approach may also be a very effective knowledge mining tool, as it was able to survey and summarise a large molecular knowledge space and still provide direct access to the underlying evidence, which is particularly important for biological knowledge discovery.

Acknowledgements

We thank members of GlaxoSmithKline (GSK) Computational Biology group, especially Peter Woollard, Hannah Tipney, Bhushan Bonde, Joanna Betts, Enrico Ferrero and Pankaj Agarwal, for their helpful comments and suggestions. We thank John Stone, Martin Sims, Jennifer Wright and Soumitra Ghosh for useful discussions on Crohn’s disease. We also thank Matt Nelson for his help with GWAS catalog data. This work was supported by GSK Research & Development Ltd, the European Research Council (ERC) Starting Independent Researcher Grant 278212, the National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) OIA-1028394, the Serbian Ministry of Education and Science Project III44006 and the ARRIS Project J1-5454.

References
