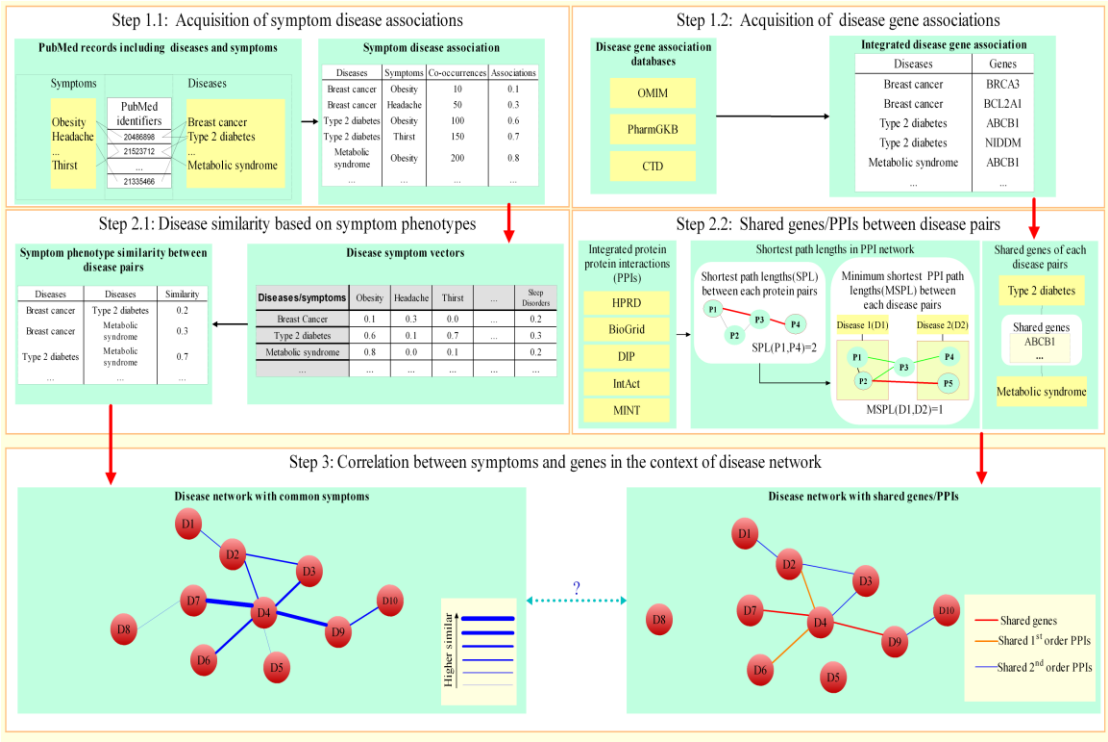
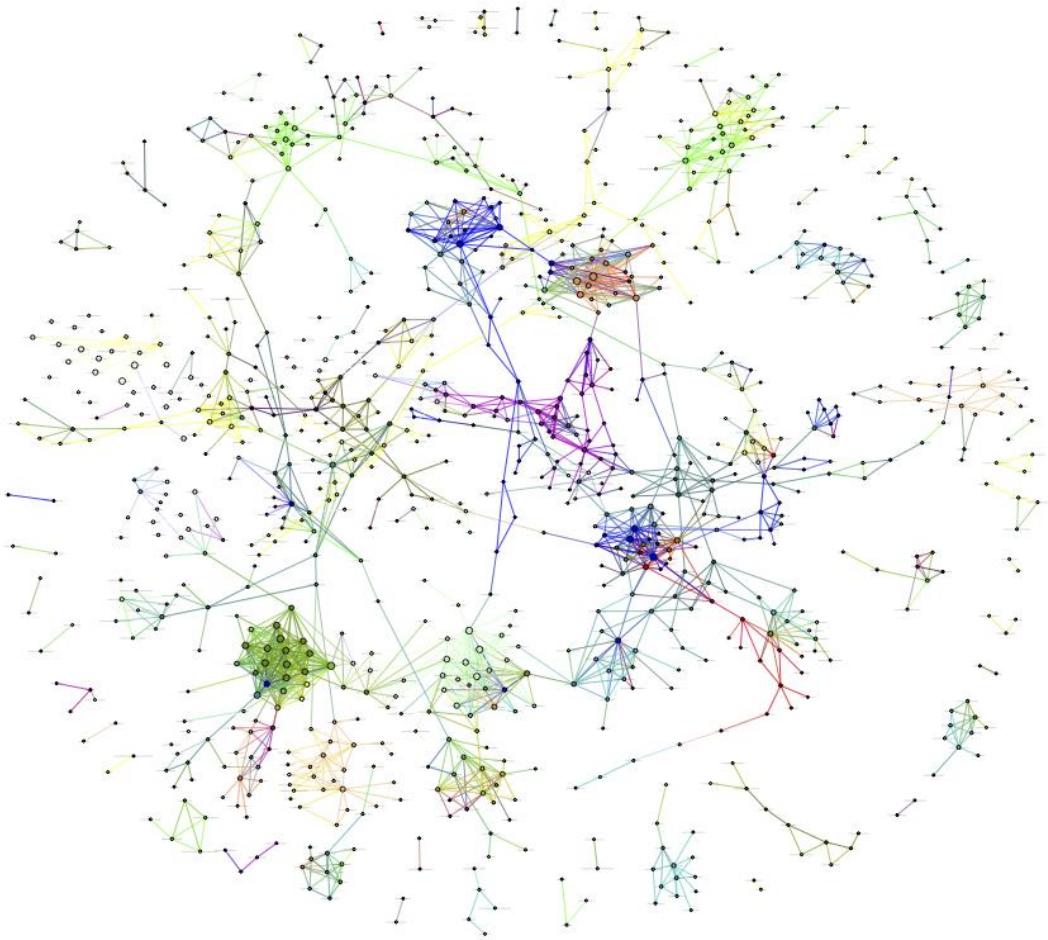
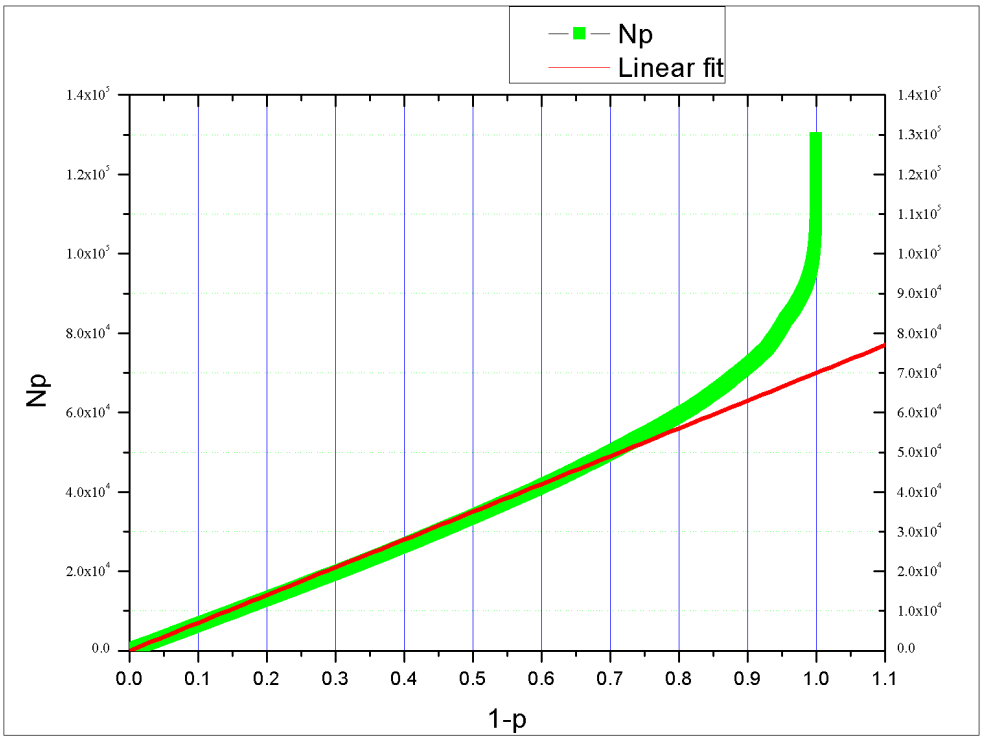
Supplementary Figures



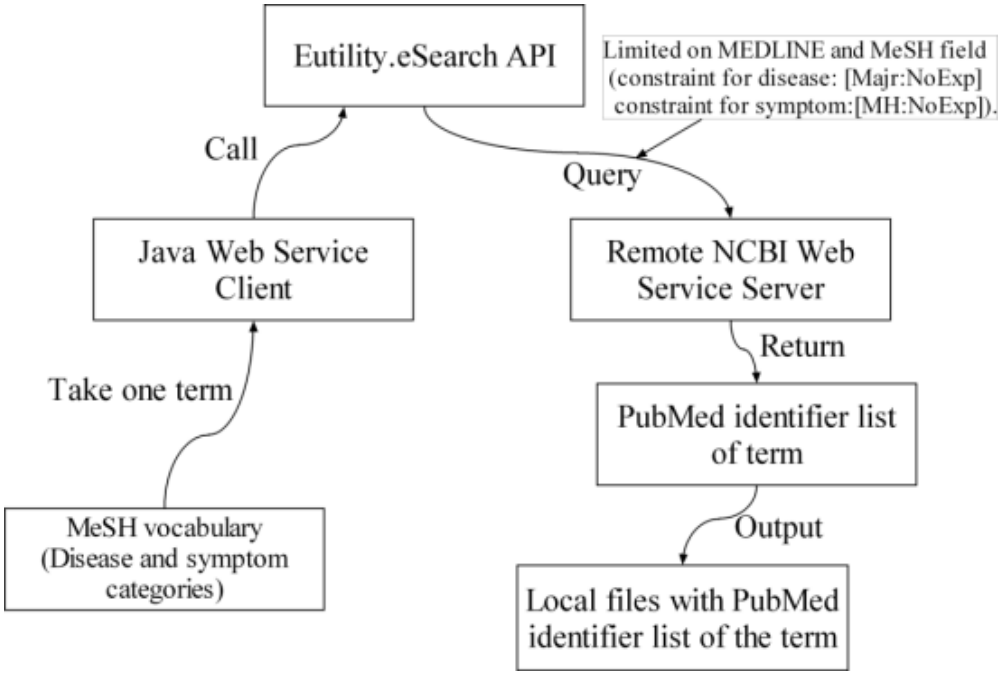
Supplementary Figure 1. **Illustration of the workflow for the construction of the different networks.** The workflow includes three main steps: We first obtain the basic association of diseases with symptoms and genes, respectively: (1.1) Disease-symptom associations are measured by a statistical analysis of the co-occurrence in the MeSH fields of the PubMed database, as acquired by an automated search protocol. (1.2) Disease-gene associations are obtained by integrating three publicly available databases (OMIM Morbid Map, PharmGKB and GAD). In a second step, we calculate the symptom-based similarity for each pair of diseases using the cosine measure (2.1). In (2.2), we integrate five publicly available protein-protein interaction (PPI) databases to one binary PPI network and compute the shortest path lengths for all protein pairs and the minimum shortest PPI path length between each disease pair. In a last step, we finally measure the correlation between symptoms and genes by comparing the overlapping links between the symptom-based disease network and the gene-based disease network.



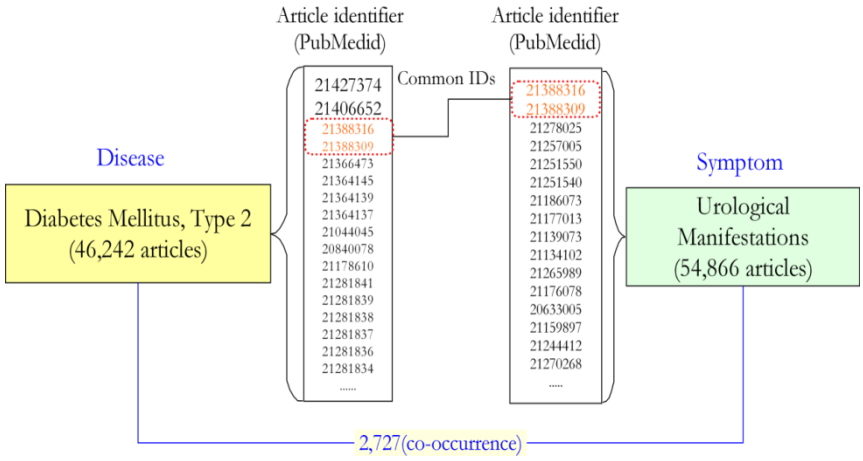
Supplementary Figure 2**. The backbone of disease network**. Nodes represent diseases, coloring is according to their respective broad MeSH disease category, see Section 5 for details.



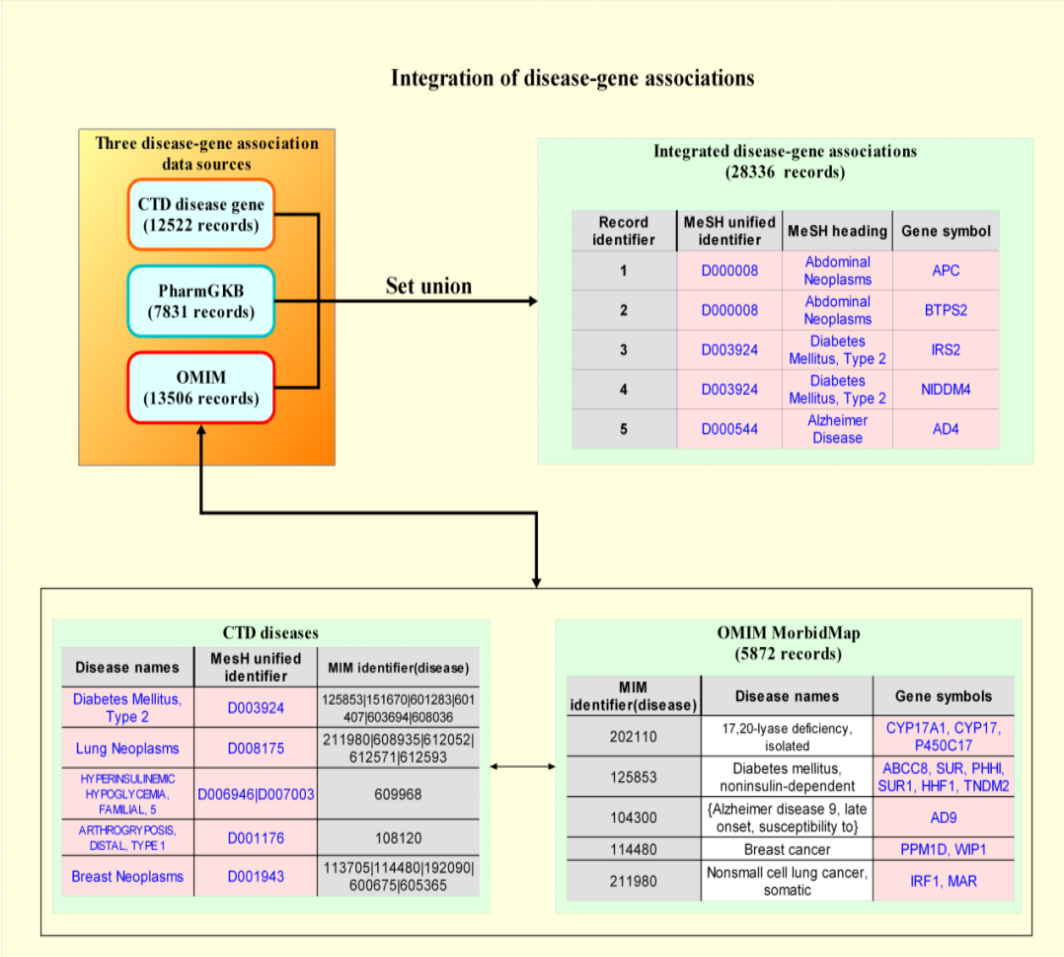
Supplementary Figure 3. **Plot of *p*-values for symptom-disease literature data with 147,380 records.** N_p denotes the number of P-values larger than a given p. The linear fit suggests that that there are approximately 70,000 true null hypotheses, resulting in P-value threshold of 0.13 and a Chi-Square value of 2.3.



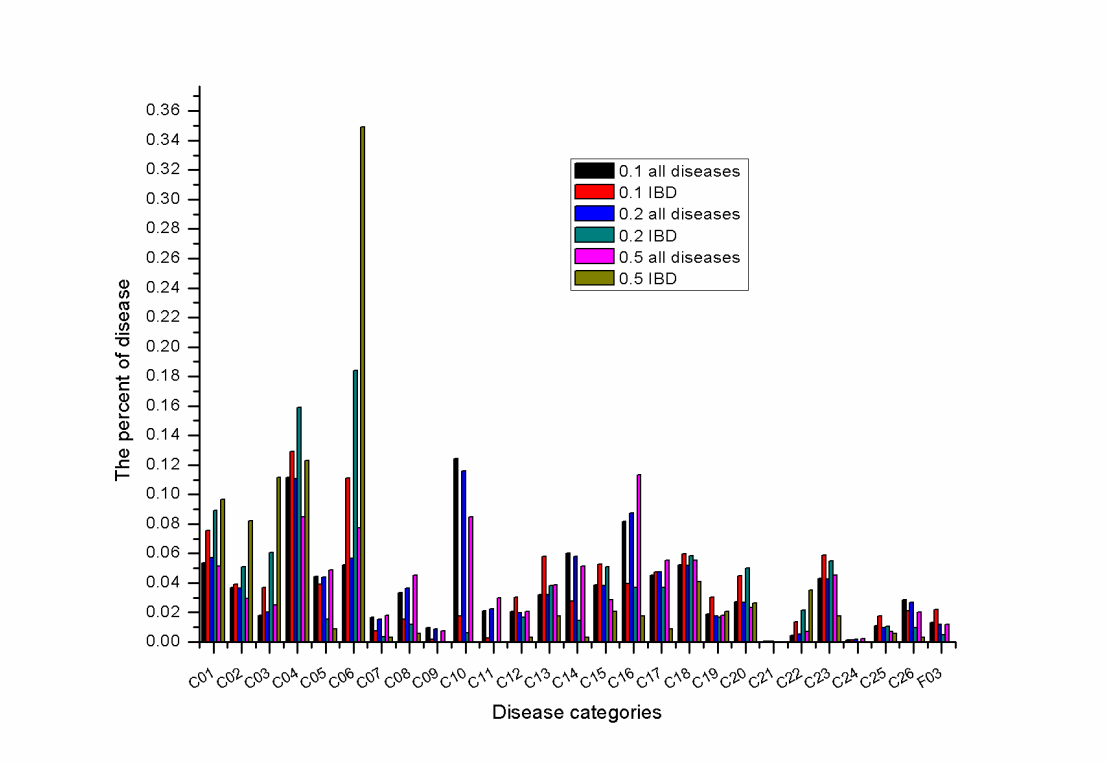
Supplementary Figure 4. **Illustration of the automated protocol to obtain disease and symptom related bibliographic data**.



Supplementary Figure 5. **Example for a disease-symptom co-occurrence.** We found 46,242 citations in MEDLINE containing the MeSH term “Diabetes, Mellitus, Type 2” and 54,866 citations with the term “Urological Manifestations”. 2,727 citations include both terms (e.g. the citations with the identifiers 21388316 and 21388309).



Supplementary Figure 6. **Integration of three disease-gene association databases.** After manually mapping the OMIM diseases to MeSH disease terms, we generated 13,506 associations between a disease (given by MeSH disease codes) and a gene (given by gene symbol). In total, the three databases yield 28,336 distinct records.



Supplementary Figure 7. **Percentage of different disease categories (MeSH) in the neighborhood of all diseases as compared to inflammatory bowel diseases (IBD) in the HSDN01, HSDN02 and HSDN05 disease network.** The occurrence ratio of the disease categories, such as C01, C03, C06 and C22, are clearly higher in the IBD neighborhood than in the neighborhood of any other diseases. The statistical evaluation using Fisher's exact test (Table S4) confirms these results.

Supplementary Tables

Disease categories	Root MeSH tree codes
Bacterial Infections and Mycoses	C01
Virus Diseases	C02
Parasitic Diseases	C03
Neoplasms	C04
Musculoskeletal Diseases	C05
Digestive System Diseases	C06
Stomatognathic Diseases	C07
Respiratory Tract Diseases	C08
Otorhinolaryngologic Diseases	C09
Nervous System Diseases	C10
Eye Diseases	C11
Male Urogenital Diseases	C12
Female Urogenital Diseases and Pregnancy Complications	C13
Cardiovascular Diseases	C14
Hemic and Lymphatic Diseases	C15
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16
Skin and Connective Tissue Diseases	C17
Nutritional and Metabolic Diseases	C18
Endocrine System Diseases	C19
Immune System Diseases	C20
Disorders of Environmental Origin	C21
Pathological Conditions, Signs and Symptoms	C23(Excluding C23.888)
Occupational Diseases	C24
Substance-Related Disorders	C25
Wounds and Injuries	C26
Mental Disorders	F03

Supplementary Table 1. **The main MeSH disease categories.**

Disease category (MeSH)	All diseases (HSDN01)	IBD (HSDN01)	All diseases (HSDN02)	IBD (HSDN02)	All diseases (HSDN05)	IBD (HSDN05)
C01	304,425	139	147,034	75	21,988	33
C02	207,580	72	93,730	43	12,554	28
C03	102,304	68	51,540	51	10,739	38
C04	632,378	238	284,439	134	36,295	42
C05	250,955	72	112,637	13	20,730	3
C06	294,595	205	145,332	155	33,035	119
C07	93,962	14	39,731	3	7,758	1
C08	189,668	28	93,733	10	19,333	2
C09	53,917	3	22,211	0	3,129	0
C10	703,468	32	298,782	5	36,223	0
C11	118,050	5	57,948	0	12,793	0
C12	116,349	56	51,214	14	8,709	1
C13	180,269	107	82,440	32	16,473	6
C14	339,553	51	148,625	12	22,038	1
C15	218,217	97	97,948	43	12,236	7
C16	460,628	73	225,112	31	48,336	6
C17	255,820	87	122,087	31	23,638	3
C18	296,584	110	133,449	49	23,649	14
C19	106,621	56	45,134	14	7,702	7
C20	154,185	82	69,375	42	9,972	9
C21	1,022	1	449	0	65	0
C22	25,739	25	13,072	18	2,939	12
C23	243,867	108	109,921	46	19274	6
C24	7,972	2	3,865	0	908	0
C25	61,422	32	25,151	9	2,991	2
C26	162,381	39	68,989	8	8,676	1
F03	75,015	40	30,055	4	5,124	0
Total count	5,656,946	1,842	2,574,003	842	427,307	341

Supplementary Table 2. **The distribution of disease categories in different**

symptom similarity scores of all diseases and inflammatory bowel diseases (IBD). We calculate the occurrence counts of the MeSH disease categories in the neighborhood of all distinct disease nodes and IBD in the disease network. The first column gives the occurrence count of each MeSH disease categories (i.e. C01-C26, F03) in any node neighborhood of the HSDN01 network. The second column gives the occurrence count of each MeSH disease categories in the neighborhood of IBD in HSDN01 network and so forth The data in this table is the basis for the Fisher's exact test, see Table S3.

Disease category	0.1 disease similarity score		0.2 disease similarity score		0.5 disease similarity score	
(MeSH)	P-value	Odds ratio	P-value	Odds ratio	P-value	Odds ratio
C01	0.0001	1.44	0.0002	1.61	0.0005	1.98
C02	0.58	1.07	0.027	1.42	1.5E-06	2.96
C03	9.7E-08	2.08	8.4E-12	3.16	3.7E-14	4.87
C04	0.020	1.18	2.1E-05	1.52	0.015	1.51
C06	2.2E-16	2.28	2.2E-16	3.77	2.2E-16	6.40
C20	2.9E-05	1.66	0.0002	1.90	0.71	1.13
C22	2.5E-06	3.01	5.8E-07	4.28	6.0E-06	5.27

Supplementary Table 3. **Fisher exact test results of the correlation between inflammatory bowel diseases (IBD) and seven MeSH disease categories (C01-C04, C06, C20 and C22).**

Supplementary Methods

1. Building the Human Symptom Disease Network

Acquisition of symptom and disease related bibliographic records

We use the Medical Subject Headings (MeSH)[1] terminology to generate symptom-disease relationships from the metadata extracted from PubMed[2] bibliographic records. PubMed is currently the most comprehensive literature database on biomedical sciences. It includes MEDLINE[3] and uses MeSH for each citation to facilitate information retrieval. MeSH is a controlled thesaurus that is used for the annotation of published articles, resulting in a high quality representation of their main topics and contributions. The MeSH terms are assigned manually by trained indexers and have been used in numerous biomedical text mining and literature-based discovery studies[4-7].

We downloaded the 2011 ASCII version of MeSH[8] that contains 26,142 distinct terms and their unified identifiers. The MeSH vocabulary is structured as a hierarchical tree with 16 top nodes, representing general categories, such as '*Anatomy*', '*Diseases*' and '*Phenomena and Processes*.' The broad category '*Diseases*' contains the sub-category '*Symptoms and Signs*' (MeSH tree code C23.888) that incorporates terms related to clinical manifestations observed by physicians or perceived by patients. We used all terms contained in the '*Disease*' category (Table S1), excluding '*Animal diseases*', as well as twenty terms, which only represent unspecific disease information, such as '*Diseases*' itself, '*Syndrome*', '*Chronic diseases*' and '*Infection*'. In total, we obtained 4,442 distinct MeSH disease terms and 327 distinct MeSH symptom terms to be used for the PubMed query. To ensure that we only retrieve records with the corresponding indexed disease terms as a major topic, we search MEDLINE with the constraint "[Majr:NoExp]", which filters for bibliographic records with the study of a specific disease as a main contribution. Using the E-Utility API web service interface of the National Center for

Biotechnology Information, we developed a JAVA program to automatically search all MEDLINE bibliographic records published between 1966 and October 2011 (Figure S4). The total number of corresponding PubMed records was 7,109,429, of which 6,553,494 included a disease and 1,405,038 a symptom term. The number of records that contain both a disease, as well as a symptom term was 849,103. They included all 4,442 MeSH disease terms and almost all (322, i.e. 98%) symptom terms.

Manual evaluation of the retrieved co-occurrences

For a basic validation of our literature-mining approach we have performed an extensive manual quality check of the core data. We randomly selected 1,000 PubMed identifiers from the 849,103 PubMed records and manually evaluated them with the aid of seven medical experts (Data S5). The full texts of the respective publications were consulted, unless the information contained in title, abstract and MeSH identifiers were conclusive. The distributions of diseases and symptoms within the random set were comparable with the ones from the full data, indicating that the sample is representative. Our evaluation focused on three key issues: (i) whether the relation between symptom and disease is direct and not mediated by other factors or simply coincidental co-occurrence; (ii) whether the relation is positive, i.e. it does not contain a negation as in "disease X is NOT related to symptom Y"; (iii) whether the relation is specific or whether a large number of diseases/symptoms are mentioned.
We find that (i) the vast majority of the identified relations between diseases and symptoms are medically meaningful and direct. The only notable (5.5%) confounding factor we identified were symptoms related to drug treatment instead of the immediate disease. (ii) The automated process yields very few false positives, as only 0.8% of the cases contained a negation that our text mining approach could not capture. (iii) The reported symptom-disease relations are very specific: 57.3% of the records contain only a single disease, 28.4% in whole data set contain two and only 14.3% more than two.

Statistical evaluation of the co-occurrences

The relation between any disease term and any symptoms term is assessed by a statistical analysis of their co-occurrence in the bibliographic records, see Figure S5 for an example. Similar methods have been used to extract disease-gene associations[4,7], gene-gene associations[9,10] and disease-drug associations[11]. For the basic test for a significant association between a given symptom-disease pair we use a Chi-Square test. The systematic identification of meaningful associations among items in large datasets is a common and challenging problem. On one side, the applied statistical test may be underpowered and fail to identify statistically weak, yet real associations, while on the other side a large number of tests may also result in many spurious associations. A priori we cannot be sure whether our dataset leans in either direction, even though it is reasonable to assume that many co-occurences are indeed meaningful, given the manual curation process of the MeSH metadata. In order to rationalize the otherwise somewhat arbitrary choice of a significance threshold, we use a method that has successfully been applied in similar cases[81]. The method is based on the fact that for repeated test with no true underlying associations there is a linear relation between the P-values and N_p (the number of test statistics with a P-value greater than p). We can therefore use a linear regression on a plot of the P-values to determine at which point the observed data deviate from the line predicted for random associations and thereby reliably identify non-random associations. This analysis yields a threshold of P-value ≈ 0.13, indicating that there are indeed relatively many false null hypotheses, i.e. true associations. For our subsequent analysis, we have nevertheless chosen to proceed with the more conservative and commonly used threshold of P-value = 0.05.

Construction of the backbone of the disease network

The subnet of the SGPDN disease network in Figure 1E of the main manuscript was extracted from the full network using the multi-scale backbone algorithm[12]. For a strict filtering of the significant edges we used a threshold parameter of α=0.05, resulting in remaining 2,159 edges out of all 133,106 edges in the full network.

Visualization was done using the software Gephi[13]. The nodes are colored according to the broad MeSH category of the corresponding diseases, e.g. red for diseases in the C01 category. For diseases that belong to multiple categories we chose the lowest root MeSH tree code. For example, in the MeSH terminology system, asthma belongs to immune system diseases (C20), as well as respiratory tract diseases (C08), in which case we chose the purple color of the C08 disease category. A full version of the network is shown in Figure S2, which includes the name of the disease and the corresponding broad MeSH classes for each node.

2. Data integration and disease term mapping

Data integration of disease-gene associations and protein-protein interactions

To investigate the molecular level regularities of shared symptoms in diseases, we use disease-gene associations and protein-protein interaction (PPI) databases to construct disease-related PPI networks. Due to the explosive development of high-throughput experimental technologies, the capacity of PPI databases has increased substantially. We integrate five publicly available human PPI databases[14], namely HPRD[15], BioGrid[16], DIP[17], IntAct[18] and MINT[19], resulting in 104,522 interactions among 14,212 proteins. Three disease-gene association databases, namely OMIM Morbid Map[20], PharmGKB[21] and comparative toxicogenomics database (CTD)[22] yield a total of 28,336 records (Figure S6) with 1,825 distinct diseases and 10,652 distinct gene symbols. Using the HUGO vocabulary[23] for the conversion between gene symbols and gene IDs, we finally obtained 4,594 disease related genes with documented interactions in the PPI network.

Mapping of MeSH disease terms to OMIM disease phenotype names

In order to combine the different disease terminologies we used the disease terminology of CTD to map from MeSH to OMIM identifiers. OMIM identifiers with no mappings in CTD were assigned manually, resulting in a total of 7,908 mappings between 4,847 OMIM disease identifiers and 1,604 MeSH disease identifiers. Since MeSH disease terms are typically more general than OMIM disease names, a single

MeSH identifier often contains mappings from several more specific OMIM identifiers. The other databases for disease-gene associations (PharmGKB and CTD) already used the MeSH terminology, so no mapping was necessary.

3. Comparison of the HSDN with related ontologies

Construction of the benchmark disease network

The Human phenotype ontology (HPO)[24] is a manually curated database initially derived from OMIM with the goal of covering all phenotypic abnormalities that are commonly encountered in human monogenic diseases[25]. We downloaded the HPO data (http://www.human-phenotype-ontology.org/, 17 March 2011) that include phenotype annotations (56,439 records) and their mappings to UMLS (Unified Medical Language System[26]) concepts (33,670 records). In order to extract those records that correspond to symptoms, we use the respective semantic type of UMLS ('*Sign or Symptom*', T184) The semantic types of the HPO phenotypes are collected from UMLS *via* its terminology services API (https://uts.nlm.nih.gov//home.html#apidocumentation, Sep 2011). The resulting subset of the HPO consists of 5,099 records that contain 2,111 distinct OMIM disease identifiers and 409 distinct symptom terms. The OMIM disease identifiers were then mapped to 940 distinct MeSH disease terms (see SM Section 2.2). Note that MeSH disease terms are typically more general and therefore several specific OMIM identifiers may map to one MeSH term. The final HPO network that can be used to benchmark our HSDN contains 940 diseases and 121,945 links indicating shared symptoms. As a well-defined and manual curated phenotype ontology, HPO has covered wide range of phenotypes. However, the majority (95%) of the 9,330 phenotype features in HPO correspond to *diseases*, *body parts* and *congenital abnormalities*. For a more detailed view on the semantic types of the HPO phenotypes, we use UMLS 2012AA to generate a comprehensive HPO mapping to UMLS with semantic types. This map includes 33,977 records with 9,256 distinct HPO phenotype identifiers, yet only 463 of which are symptom phenotypes (213 UMLS concept identifiers, see Data S6).

Link overlap between HPO and HSDN

From the 940 diseases in the benchmark HPO network, N_{HSDN} = 898 are present in the HSDN. The subset of the HSDN consisting of these 898 diseases contains L_{HSDN} = 372,509 links between the P = N_{HSDN} × (N_{HSDN} - 1) / 2 = 402,753 possible disease pairs, the respective subset of the HPO network contains L_{HPO} = 111,923 links. The probability in each of the two networks for a randomly chosen disease pairs to be connected is therefore $p_{\mathrm{SDN,HPO}} = L_{\mathrm{SDN,\ HPO}} / \mathrm{P}$. If the two networks were completely independent from each other, the expected number of disease pairs that are connected by both HPO and HDN links would be $p_{\mathrm{HSDN}} \times p_{\mathrm{HPO}} \times P$ = 103,518. The P-Value of the observed number of 107,098 shared links can be computed from the binomial distribution to be P-value = 2.2×10^{-16}.

Comparison of the HSDN with medical terminology systems

A detailed knowledge of symptom-disease relationships is the basis for any clinical diagnosis, especially for clinical syndromes and traditional disease classifications. In order to systematically record and process such clinical data, several medical terminologies have been developed. The most prominent example is SNOMED-CT (clinical terms)[27], which contains a comprehensive set of concepts from clinical practice concepts and their relationships. The main application of SNOMED-CT is the administration of clinical records, but it can also be used for bioinformatics applications[28]. Since the relationships within SNOMED-CT are directly derived from clinical practice, they would provide an ideal basis to explore symptom-disease relationships. We have therefore explored this possibility using the SNOMED-CT data as included in the UMLS system (UMLS 2012 AA). As the semantic types of 'Sign and Symptom' and 'Disease or Syndrome' in UMLS are coded as T184 and T047 respectively, we filtered the relationship records with constraints of these two semantic types. In total, we only obtained 2,340 relationships between 1,623 diseases and 817 symptoms (we include all kinds of relationships between diseases and symptoms in addition to the manifestation associations) from SNOMED-CT (Data S7), We find that 1,250 (77.0%) diseases have only one related symptoms and 236 (14.5%) diseases have two related symptoms. *Migraine disorder*, for example has the

UMLS concept unified identifier (CUI) C0149931 and the SNOMED identifier 37796009. There is only one relationship record in the data, in which the symptom is '*headache (finding)*' (CUI: C0018681). While *headache* represents the most dominant symptom of *migraine*, it is certainly not the only one, but the disease has various prodromes, accompanying symptoms and postdromes like *nausea*, *dizziness*[29], *vomiting*, *photophobia* and *fatigue*[30]. We further find that the subtypes like '*Migraine with aura*' (CUI: C0154723), '*Migraine without aura*' (CUI: C1827190) and '*Abdominal migraine*' (CUI: C0270858) have '*headache(finding)*' as their single symptom as well, so they cannot be differentiated by their related symptoms in SNOMED-CT. We conclude that the disease-symptom relationships contained in SNOMED-CT are far too limited to be useful for our purposes.

There are several other medical terminology systems like that include symptoms, such as ICD 9CM[31], Disease Ontology[32] and Symptom Ontology[33]. However, none of them includes a substantial number of disease-symptom relations that is nearly as comprehensive as the one we were able to extract from PubMed records.

4. Exploring the correlations between inflammatory bowel diseases (IBD) and broad MeSH disease categories

Filtering the disease network

The full disease network based on shared symptoms contains 4,172,423 edges. For a more detailed investigation of the correlation between IBD and the broad MeSH categories we generated three networks with different levels of minimal similarity: After filtering for interactions with symptom similarity score larger than 0.1 (HSDN01) we are left with 1,121,899 edges;similarity cutoffs of 0.2 and 0.5 yield 536,272 (HSDN02) and 86,941 edges (HSDN05), respectively.

Mapping diseases to MeSH tree root codes

We mapped each disease term in the HSDN to its broadest disease category (e.g. root MeSH tree code C01 for 'bacterial infections and mycoses', see Table S1) using the 2011 version of MeSH tree code ('mtrees2011.bin' with 52,546 records, downloaded from http://www.nlm.nih.gov/mesh/meshhome.html). For each network, HSDN01,

HSDN02 and HSDN05, we constructed a list of all diseases in the neighborhood of IBD and calculate the occurrence count for all 27 broad disease categories. For comparison with random expectation we repeated the procedure for all other diseases, see Table S3 and Figure S7. The data of Table S2 was used for calculating the correlation between IBD and each MeSH disease categories.

Fisher exact test

We use the 'fisher.test()' function of R (version 2.13.0) to calculate the correlation between IBD and MeSH disease categories.

Discussion

In particular we evaluated the correlations between IBD and seven specific disease categories, namely *bacterial infections and mycoses* (C01), *virus diseases* (C02), *parasitic diseases* (C03), *neoplasms* (C04), *digestive system diseases* (C6), *immune system diseases* (C20) and *animal diseases* (C22, only including the related diseases that are also manifested in human). The results (Table S4) show clear positive correlations between IBD and the categories C01, C03, C06 and C22. Furthermore, IBD correlates with C02 in the HSDN05 network, meaning that most of the correlated diseases in the C02 category have high symptom similarity with IBD. Conversely, C20 diseases correlate with IBD in the network including low similarity scores, indicating that most of the correlated C20 diseases have a low degree of shared symptoms with IBD.

Supplementary References

1. Lowe, H.J. & Barnett, G.O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* **271**, 1103-8 (1994).
2. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35**, D5-12 (2007).
3. Dee, C.R. The development of the Medical Literature Analysis and Retrieval System (MEDLARS). *J Med Libr Assoc* **95**, 416-25 (2007).
4. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat Genet* **31**, 316-9 (2002).
5. Bhattacharya, S., Ha-Thuc, V. & Srinivasan, P. MeSH: a window into full text for document summarization. *Bioinformatics* **27**, i120-8 (2011).
6. Swanson, D.R., Smalheiser, N.R. & Torvik, V.I. Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology* **57**, 1427-1439 (2006).
7. Jensen, L.J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* **7**, 119-29 (2006).
8. NLM. http://www.nlm.nih.gov/mesh/2011/download/termscon.html. 2011 edn Vol. 2011 The download page of 2011 MeSH (2011).
9. Hoffmann, R. & Valencia, A. A gene network for navigating the literature. *Nat Genet* **36**, 664 (2004).
10. Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28**, 21-8 (2001).
11. Chen, E.S., Hripcsak, G., Xu, H., Markatou, M. & Friedman, C. Automated acquisition of disease drug knowledge from biomedical and clinical

documents: an initial study. *J Am Med Inform Assoc* **15**, 87-98 (2008).

12. Serrano, M.A., Boguna, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci U S A* **106**, 6483-8 (2009).
13. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. in *International AAAI Conference on Weblogs and Social Media* 361-362 (AAAI, California,USA, 2009).
14. De Las Rivas, J. & Fontanillo, C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* **6**, e1000807 (2010).
15. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-71 (2003).
16. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).
17. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91 (2000).
18. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**, D452-5 (2004).
19. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**, D572-4 (2007).
20. Hamosh, A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **30**, 52-5 (2002).
21. Hewett, M. *et al.* PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* **30**, 163-5 (2002).
22. Davis, A.P. *et al.* The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res* **39**, D1067-72 (2011).
23. Eyre, T.A. *et al.* The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res* **34**, D319-21 (2006).
24. Robinson, P.N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**, 610-5 (2008).

25. Robinson, P.N. & Mundlos, S. The human phenotype ontology. *Clin Genet* **77**, 525-34 (2010).
26. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32**, D267-70 (2004).
27. Cote, R.A. & Robboy, S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA* **243**, 756-62 (1980).
28. Cimino, J.J. High-quality, standard, controlled healthcare terminologies come of age. *Methods Inf Med* **50**, 101-4 (2011).
29. Bisdorff, A. Migraine and dizziness. *Curr Opin Neurol* **27**, 105-10 (2014).
30. Roger P, A., David A, S. & Michael J, G. in *Clinical neurology* 85–88 (Lange Medical Books/McGraw-Hill, New York, 2009).
31. NCHS. http://www.cdc.gov/nchs/icd/icd9cm.htm. Vol. 2014 ICD-9-CM home page (CDC, 2013).
32. Schriml, L.M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* **40**, D940-6 (2012).
33. Baclawski, K., Matheus, C.J., Kokar, M.M., Letkowski, J. & Kogut, P.A. Towards a Symptom Ontology for Semantic Web Applications. in *The Semantic Web-ISWC 2004* Vol. 3298 650-667 (Lecture Notes in Computer Science, 2004).