Supplementary Information

S1. Data sets

a. Gene-phenotype associations from OMIM

The OMIM database (http://www.ncbi.nlm.nih.gov/omim/) provides a downloadable text version of gene-phenotype mappings called "*morbidmap*" (available at http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html#download), which contains 4602 known gene-phenotype associations as of September 2007 [1]. Some phenotypes listed in the file with a minor difference in their names, however, may be similar enough to be treated as one phenotype, which was done in the work of Goh *et al.* (for instance, "Hemophilia A" and "Hemophilia B" being mapped to a single disease ID 663. This ID, introduced in Goh *et al*., is not to be confused with the ICD-9-CM code; in fact, we introduce a mapping between this ID into ICD-9-CM codes. See Section S2.) and available at http://www.pnas.org/content/104/21/8685/suppl/DC1 [2], which was also the basis of our list of OMIM diseases studied here (with minor updates based on the most recent *morbidmap* at the time of this study, and the addition of (1) and (2) tags for a more inclusive mapping).

Upon a closer inspection of the *morbidmap*, however, we discovered a slight incompletion in the mapping, spotting some phenotypes in *morbidmap* with identical six-digit OMIM codes, but whose given alphabetical names were different enough to be considered as separate phenotypes in Goh *et al*. For instance, in SI Table 1 of Goh *et. al*, the disease phenotype OMIM 607948 was mapped to two separate diseases, ID 1043 (when it was called "Mycobacterium tuberculosis" along with other mycobacterial infections) and ID 1533 (when called "Tuberculosis") in the Goh *et al*. mapping:

1043	Mycobacterium tuberculosis, susceptibility to infection by, **607948** (3)	NRAMP1
1533	Tuberculosis, susceptibility to, **607948** (3)	IFNG

A problem that results from this artifact is that in later stages of mapping the OMIM diseases into the ICD-9-CM coding scheme used in the Medicare database for calculating comorbidity (described in detail in Section S2), the gene NRAMP1, along with several other genes, may not be considered to be associated with Tuberculosis, while it clearly should. We corrected this problem by including all genes that correspond to the six-digit OMIM ID that was originally mapped to each disease; in the example above, now both diseases 1043 and 1533 correctly contain all genes mapped to six-digit OMIM code 607948 in the *morbidmap*. . We have found out that 4 OMIM codes were mapped to three ICD9 codes, while 43 OMIM codes were mapped to two ICD9 codes.

b. Protein-protein interaction and gene expression.

The protein-protein interaction data were taken from Rual *et al.* [3] and Stelzl *et al.* [4]. In order to calculate the average human gene coexpression, we used an Affymetrix (www.affymetrix.com) microarray data that lists expression levels of select genes on 36 human tissues [5] (also see Section 3).

c. The Medicare database

We obtained raw Medicare claims files directly from Centers for Medicare & Medicaid Services (CMS, www.cms.hss.gov) in the form Medicare Provider Analysis and Review (MEDPAR) files. These files are made available subject to a Data Use Agreement. At present, access to such files is via the CMS-designated Research Data Assistance Center (ResDAC, www.resdac.umn.edu) program. The data we used contain the complete hospitalization records of 13,039,018 Medicare patients, for a total of 32,341,348 visits over 4 years, from 1990 to 1993. In the Medicare database, each line corresponds to a hospitalization even of a patient, and has a record of up to ten diagnoses. The diagnoses are presented using the ICD-9-CM scheme (www.icd9data.com), where a disease is assigned a numeric code such as 174 for breast cancer (Fig. S1).

The Medicare records are comprehensive, and they are frequently used for epidemiological and demographic research [6,7]. The present sample was abstracted from a comprehensive set of hospital visits of all elderly patients (aged 65 and up) in the Medicare program, which comprises 96% of the entire elderly American population. Our sample of 13,039,018 hospitalized patients shows a mean age of 76.3 ± 7.4; 41.8% were female; and 89.9% where Caucasians. Most patients were diagnosed with multiple diseases during the observation period, a co-occurrence that is in some cases accidental, but is also often causal, i.e., one disease increases the likelihood of the development of other diseases[8,9].

S2. Mapping between the Medicare database and the OMIM disease-gene association.

The mapping from OMIM diseases to ICD-9-CM codes we used in this study is given in the Supporting Table 1 (Supp_Table1.xls). Each line represents a mapping given in the following format:

ID	*ICD-9-CM*	*Name in Gho et al.*	*Name in ICD-9-CM*
1166	170.9	Osteosarcoma	MALIGNANT NEOPLASM OF BONE AND ARTICULAR CARTILAGE

As noted explained in Section S1, ID is the unique number given to the OMIM diseases in Goh *et al*. In some cases a suitable mapping was not found. In Fig S1, we show a schematic diagram of the procedure of counting incidences and co-ocurrences of disease pairs using the OMIM and the Medicare database in conjuction with this mapping. For instance, in this mapping, ICD-9-CM code 170.9 represents three OMIM diseases, ID 316 (*“Chondrosarcoma”*), ID 511 (*“Ewing_sarcoma”*), and ID 1166 (*“Osteosarcoma”*). We accept a single ICD-9-CM 170.9 code as our unit of disease, whose associated genes are the union set of those of their corresponding OMIM diseases, as shown in Fig S1. Now, to find the incidence of ICD-9-CM 170.9 and its comorbidity with other diseases, we parsed the Medicare database and counted the number of patients whose records show the ICD9-CM code 170.9. (See Fig. S1[1]). We can show that, after performing this procedure for all codes appearing in the mapping, 90.0% of the patients were diagnosed with at least one of the diseases we considered (Fig 1A). Finally, in order to mitigate the effect of biases due to extremely rare diseases and disease pairs, we considered only the

[1] We also consider the hierarchical nature of ICD-9-CM codes. In Fig S1 for instance, when we meet a subcode of breast cancer (174) such as 174.1 – meaning more detailed, specific diagnoses -- we count it also as an incidence of 174.

diseases whose incidences were 10 or larger, and the pairs of those diseases whose randomly expected co-occurrence C_{ij}^{*} was 1.0 or larger, resulting in a total of 83,924 disease pairs for the study. In Section S4, we compare the various thresholds for C_{ij}^{*}.

S3. Definitions of the genetic variables

a. **Number of shared genes** (n_{ij}^{g}): the number of disorder genes associated with both diseases i and j. Denoting by G_i and G_j the sets of genes associated with diseases i and j respectively, $n_{ij}^{g} = |G_i \cap G_j|$.

b. **Number of shared protein-protein interactions** (n_{ij}^{p}) is the number of protein-protein interactions that link genes of disease i to those of disease j.

c. **Average gene coexpression** $(\bar{\rho}_{ij})$ is represented by the average of the coexpression levels between every pair of genes associated with each disease. If we denote as x_{at} as the expression level of gene a on tissue t $(t = 1, \ldots, 36)$ listed in the Affeymetrix database, the gene coexpression level ρ_{ab} between two genes a and b is defined as the Pearson correlation between the two (where $n_t = 36$):

$$\rho_{ab} = \frac{n_t \sum_t x_{at} x_{bt} - \sum_t x_{at} \sum_t x_{bt}}{\sqrt{(n_t \sum_t x_{at}^2 - [\sum_t x_{at}]^2)(n_t \sum_t x_{bt}^2 - [\sum_t x_{bt}]^2)}}.$$

The genes for which no entry exists in the data set was set to have no correlation (i.e. 0.0) with other genes. When multiple expression values exist for a gene a in tissue t, x_{at} was set to be the average of the expression levels for the given tissue.

S4. The ϕ-correlation and the robustness of comorbidity measures

In addition to relative risk RR, we used the ϕ-correlation as a comorbidity measure, which we discuss in more detail. First, it is defined as

$$\phi_{ij} = \frac{NC_{ij} - I_i I_j}{\sqrt{I_i I_j (N - I_i)(N - I_j)}},$$

which can be shown to be equivalent to the Pearson correlation between variables that take dichotomous values (for instance, 0 or 1), so that $-1 \leq \phi_{ij} \leq 1$ [10]. Note that RR and ϕ are not independent from each other: we can rewrite ϕ_{ij} as $\phi_{ij} = \left(RR_{ij} - 1\right)\sqrt{I_i I_j / (N - I_i)\left(N - I_j\right)}$, which clearly shows that the conditions $\phi_{ij} > 0$ and $RR_{ij} > 1$ are equivalent -- both indicate that two diseases occur together more often than expected by chance alone. Nevertheless, each variable carries with it unique advantages and disadvantages over the other, which prompted us to use both to show the robustness of our findings. The primary advantage of using RR is that its meaning is very clear and intuitive. However, RR can be biased in the sense that it could assume an abnormally large value when the random expectation C_{ij}^* is small – i.e., when i and j are very rare. Our method to overcome this issue is to introduce a threshold (TH) for C_{ij}^* and consider disease pairs whose expected co-occurrence equal or exceed it. In contrast, the advantage of using ϕ_{ij} as the comorbidity measure comes from the fact that it is bounded in the range [-1,1]. However, even when two diseases are highly comorbid ($RR_{ij} \gg 1$), ϕ_{ij} can have an apparently small numerical value, because $\phi_{ij} \approx \left(RR_{ij} - 1\right)\sqrt{I_i I_j / N^2}$, and in our database usually $I_i, I_j \ll N$. Furthermore, two diseases being “maximally” comorbid given the incidences (meaning $C_{ij} = I_i$, one disease always occurring with the other, assuming $I_i \leq I_j$) doesn’t imply $\phi_{ij} = 1$: rather, we have $\phi_{ij}^{\max} = \sqrt{I_i\left(N - I_j\right)/I_j\,(N - I_i)}$ which is always smaller than 1 unless $I_i = I_j$. In order to compensate for this, we could use the normalized variable $\phi_{ij}/\phi_{ij}^{\max}$ so that when ϕ_{ij} is at its maximum, $\phi_{ij}/\phi_{ij}^{\max} = 1$.

In Fig. S2 we study the effect of the aforementioned threshold TH imposed on C_{ij}^* on the average values of RR_{ij}, ϕ_{ij}, and $\phi_{ij}/\phi_{ij}^{\max}$ for disease pairs that satisfy various criteria presented in Fig. 1C. We compare the cases of TH equal to 0 (no threshold, i.e. all disease pairs are included), 1, 3, 4, 5, and 10. Fig. S2 demonstrates that, as pointed out earlier, introducing thresholds significantly changes the magnitude of $\langle RR \rangle$ by removing pairs of exceptionally large RRs that arise from from C_{ij}^* that are far too small. Most importantly, $\langle RR \rangle$ becomes stable and robust for any threshold larger than 0 (i.e., TH$\geq$ 1).

Still, the qualitative trend we observe in the zero-threshold (TH=0) curve remains unchanged even in the thresholded curves except for the case of $\langle RR \rangle$ against $\bar{\rho}$. We also observe a similar trend in ϕ and $\phi/\phi^{\max}$, which again demonstrates the robustness of our conclusions. Due to such strong stabilizing effect and the robustness resulting from the thresholds on C_{ij}^{*}, we chose to show the curves of TH=1 in the paper. However, had we chosen any other threshold, the conclusions would have been qualitatively unchanged.

S5. Determining the *P*-values and Errors

The P-values for relative risk RR_{ij} and ϕ can be obtained using standard numerical analysis tools such as Mathematica by approximating the binomial distribution generated from N and $p = I_i I_j / N^2$ as a Poisson distribution (since $N = 1.3 \times 10^7 \gg 1$). Since $C_{ij}^{*} = N \times p = I_i I_j / N$, the P-value of comorbidities between diseases i and j is given by the formula

$$P_{ij} = \sum_{k=C_{ij}}^{N} \frac{\exp(-C_{ij}^{*}) \times \left(C_{ij}^{*}\right)^{k}}{k!}.$$

The P-values for Pearson correlations (appearing in Fig. 2A) between comorbidity and genetic variables was calculated via a Monte Carlo method, where we randomized the ordering of genetic variables many times (For Fig. 2A, we performed one million iterations for each case) and directly counted how often we observed a correlation larger than the actual values.

S7. List of genetically linked pairs

In Supporting Table 2 (Supp_Table2.xls), we provide a list of the 2,239 disease pairs that are genetically linked, i.e. for which $n^{g} \geq 1$ or $n^{p} \geq 1$. The table lists the ICD-9-CM codes of the linked diseases, the incidences and co-occurrences I_1, I_2, and C_{12}, the comorbidities RR and ϕ, the genetic variables n^{g}, n^{p}, and $\bar{\rho}$, and shared genes and the

PPIs (if any exists between the pair) linking the two diseases. Supplementary Table 3 (Supp_Table3.xls) lists the ICD-9-CM codes and their associated genes. ICD-9-CM codes noting non-disease diagnoses conditions (codes starting with alphabets) are omitted.

S4. Identifying functional protein domains using the Pfam database from the codon information

The *morbidmap* file does not contain the specifics of the mutation on the gene associated with the disease phenotype, so we extracted the information from the text file (omim.txt) of the full OMIM database (available on http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html#download) using text mining technique. In this study, we only considered mutations in which the the codon information were readily available, such as when involving amino acid substitutions recorded in the form “XnY” in the OMIM database, meaning X being replaced by Y at the codon *n*. For instance, we can find in file *“omim.txt”* the following:

.0018
OSTEOSARCOMA
OSTEOSARCOMA, MULTIFOCAL, INCLUDED
TP53, **ARG282TRP**

.0023
BREAST CANCER
TP53, **ARG181HIS**

This tells us that the mutations on *TP53* related to osteosarcoma and breast cancer occur on its codons 282 and 181, respectively. To find the domain of this protein from Pfam database, first we find the Uniprot ID of TP53, which is P04637 (from www.uniprot.org). According to the Pfam, codons 282 and 181 of *TP53* belong to the same *P53* domain on the protein (thereby, we define this pair of diseases domain-sharing), known to be involved in DNA binding (http://pfam.sanger.ac.uk/protein?entry=P04637). Osteosarcoma and breast cancer are associated with more mutations on TP53, depicted in Fig 1C in the manuscript. Indeed, 177 out of 378 ICD-9-CM diseases that share a gene with other diseases are found to have mutations on multiple domains of their shared genes. We provide some examples in the following table. Each domain is represented by its first and last codons.

ICD-9-CM / OMIM ID	*Name*	*Gene*	*Pfam Domain* **[start codon-end codon]**	*Mutations*
277.39 / 102	**Amyloidosis**	*APOA1*	[1-42]	GLY26ARG
			[43-67]	LEU60ARG, TRP50ARG
			[68-263]	LEU90PRO, ARG173PRO, LEU174SER, ALA175PRO
277 / 399	**Cystic Fibrosis**	*CFTR*	[1237-1419]	SER1255TER, TRP1282TER, TRP1316TER, ASN1303LYS, GLN1291HIS, ARG1283MET, GLN1238TER, GLN1313TER, PHE1286SER, GLY1249GLU, SER1251ASN, SER1255PRO, ASN1303HIS, HIS1282TER, GLN1352HIS, GLY1244VAL
			[1420-1480]	SER1455TER
			[3-62]	TRP57TER, GLU7TER
			[451-622]	PHE508DEL, ILE507DEL, GLN493TER, ALA455GLU, GLY542TER, SER549ASN, SER549ILE, SER549ARG, GLY551ASP, ARG553TER, ALA559THR, ARG560THR, TYR563ASN, PRO574HIS, ILE506VAL, PHE508CYS, GLY458VAL, GLY551SER, VAL520PHE, CYS524TER, SER492PHE, ARG560LYS, ALA554GLU, GLY480CYS, ILE556VAL, GLN524HIS, GLU542TER, GLN552TER, ARG553GLN, ALA561GLU
			[81-350]	ASP110HIS, ARG117HIS, ARG347PRO, ARG334TRP, GLY85GLU, PHE311LEU, ARG347LEU, ALA349VAL, GLU92LYS, ARG347HIS, GLY91ARG, GLU92TER, LEU206TRP, THR338ILE, TYR109CYS, GLU217GLY
			[862-1147]	TYR913CYS, ARG1066HIS, ALA1067THR, ARG1066CYS, TRP1089TER, GLN890TER, HIS949TYR, LEU1065PRO, GLN1071PRO, HIS1085ARG, TYR1092TER, GLY1244VAL
334.3 / 289	**Cerebellar Ataxia**	*CACNA1A*	[136-359]	GLY293ARG
			[522-713]	THR666MET, ARG583GLN
			[1601-1810]	ILE1710THR
151.9 / 1449	**Stomach Cancer**	*CDH1*	[267-366]	GLU336ASP, THR340ALA
			[598-686]	ALA617THR, ARG598TER, ALA634VAL
			[687-708]	GLN699TER
			[732-879]	VAL832MET

S8. Discussions of three disease pairs

Here we discuss in more detail the three example disease pairs introduced in the main text. Full genetic associations of each disease can be found in the Supplementary Tables 1 and 2.

a. Breast cancer and cancer of bone and cartilage

Worldwide, breast cancer is the most common type of cancer and the fourth most common cause of death from cancer (http://www.cancer.org/downloads/STT/Global_Cancer_Facts_and_Figures_2007_rev.pdf). Cases of patients developing breast cancer after treatment of osteosarcoma with possible involvement of P53 genes have been reported[11,12]. Risk of breast cancer in mothers of children with osteosarcoma and chondrosarcoma have been reported, further indicating a genetic componentin the aetiology of these cancers.[13]

b. Alzheimer's disease and myocardial infarction

Among the various conditions leading to dementia, Alzheimer's disease is the leading cause, accounting for more than 60% of all dementia cases [14]. Myocardial infarction (heart attack) is the leading cause of death for men and women in the U.S. (http://www.americanheart.org/downloadable/heart/113535864858055-1026_HS_Stats06book.pdf) Population-based studies have been carried out to evaluate the comorbidity of myocardial infarction and dementia with inconsistent results [15-17]. In one study, the comorbidity was shown to be gender-dependent: women with a history of myocardial infarction were 5 times more prone to dementia than those without a history, an effect absent in men [15]. In others, unrecognized myocardial infarction was associated with an increased risk of dementia in men [17]. Also, low cardiac output, cerebral hypoperfusion, and microembolization have been believed to be responsible for developing cognitive impairment after a myocardial infarction [16]. Genetic effects are also known to some degree: the sequence variation of ACE and the following variation in the level of the angiotensin I converting enzyme in plasma are known to be involved in both diseases. Blood pressure is partly regulated by angiotensin II, formed from angiotensin I, and the Alzheimer disease risk may be related to blood pressure regulation [18-21]. Apolipoprotein E (APOE) is a ligand for low density lipoprotein (LDL) receptor, very low density lipoprotein (VLDL) receptor, *etc.*, and its polymorphism has an impact on plasma cholesterol levels. Therefore, the APOE polymorphism is suspected to modulate the coronary heart diseases risk, and its association with the myocardial infarction and also with the Alzheimer's disease has been identified by population-based studies [22,23].

c. Carpal Tunnel Syndrome and Autonomic Nervous System.

Carpal Tunnel Syndrome (CTS) occurs when the median nerve, which runs from the forearm into the hand, becomes pressed or squeezed at the wrist. The carpal tunnel, a narrow, rigid passageway of ligament and bones at the base of the hand, houses the median nerve and tendons. Thickened or swollen irritated tendons compress the median nerve, resulting in pain, weakness, or numbness in the hand and the wrist (http://www.ninds.nih.gov/disorders/carpal_tunnel/detail_carpal_tunnel.htm). CTS is often the result of a combination of factors that increase pressure on the median nerve and tendons in the carpal tunnel, rather than a problem with the nerve itself. Likely causes include congenital factors (small carpal tunnels) and injuries (sprain or fracture). Additionally, some cases of CTS are known to be related to autonomic nervous system dysfunctions caused by amyloidosis (a disorder characterized by deposits of amyloids, mainly of proteins, in the body) of light polypeptide chain (L-chain) of immunoglobulin molecules. The L-chain amyloidosis is caused when an excessive amount of L-chains are metabolized into amyloidogenic L-chains (amyloid fibers) [24]. Initial manifestations of the L-chain amyloidosis, such as lightheadedness, fixed heart rate, bladder dysfunction, impotence and gastrointestinal disturbances, arise when the autonomic nervous system is affected[24,25]. CTS occurs when the L-chain amyloids infiltrate the flexor retinaculum of the wrist [24,26].

Supplementary References

1. McKusick, V.A. *Mendelian Inheritance in Man. A Catalog of HUman Genes and Genetic Disorders.*, (Johns Hopkins University Press, 1998).
2. Goh, K.-I. et al. The human disease network. *Proc. Natl. Acad. Sci.* **104**, 8685-8690 (2007).
3. Rual, J.-F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).
4. Stelzl, U. et al. A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell* **122**, 957-968 (2005).
5. Ge, X. et al. Interpreting expression profiles of cancers by genome-wide survey of breadth-of-expression in normal tissues. *Genomics* **86**, 127-141 (2005).
6. Lauderdale, D., Furner, S.E., Miles, T.P. & Goldbert, J. Epidemiological uses of Medicare data. *Am J Epidemiol* **15**, 319-327 (1993).
7. Mitchell, J.B. et al. Using Medicare claims for outcomes research. *Medical Care* **32**, JS38-JS51 (1994).

8. Hidalgo, C.A., Blumm, N., Barabasi, A.-L. & Christakis, N.A. A dynamic network approach for the study of human phenotypes. *submitted to PLOS comp. bio.* (2008).

9. Rzhetsky, A., Wajngurt, D., Park, N. & Zheng, T. Probing genetic overlap among human phenotypes. *Proc. Natl. Acad. Sci.* **104**, 11694-11699 (2007).

10. Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*, (Lawrence Eribaum Associates, 2002).

11. Knowling, M.A. & Basco, V.E. Breast-cancer after treatment for osteosarcoma. *Med. Pedia. Oncol.* **14**, 51-53 (1986).

12. Russo, C.L. et al. Secondary Breast-Cancer in patients presenting with osteosarcoma - possible involvement of germline P53 mutations. *Med. Pedia. Oncol.* **23**, 354-358 (1994).

13. Hartley, A.L., Birch, J.M., Marsden, H.B. & Harris, M. Breast cancer risk in mothers of children with osteosarcoma and chondrosarcoma. *Br. J. Cancer* **54**, 819-823 (1986).

14. Pasternak, J.J. *An introduction to Human Molecular Genetics*, (Wiley, 2005).

15. Aronson, M.K. et al. Women, myocardial infarction, and dementia in the very old. *Neurology* **40**, 1102-1106 (1990).

16. Bursi, F. et al. Heart Disease and Dementia:A Population-Based Study. *Amer. J. Epidemiol.* **163**, 135-141 (2006).

17. Ikram, M.A. et al. Unrecognized Myocardial Infarction in Relation to Risk of Dementia and Cerebral Small Vessel Disease. *Stroke* **39**, 1421-1426 (2008).

18. Kehoe, P.G. et al. Variation in DCP1, encoding ACE, is associated with susceptibility to Alzheimer disease. *Nat. Genet.* **21**, 71-72 (1999).

19. Narain, Y. et al. The ACE gene and Alzheimer's disease susceptibility. *J. Med. Genet.* **37**, 695-697 (2000).

20. Butler, R., Morris, A.D. & Struthers, A.D. Angiotensin-converting enzyme gene polymorphism and cardiovascular disease. *Clin. Sci.* **93**, 391-400 (1997).

21. Katzov, H. et al. A cladistic model of ACE sequence variation with implications for myocardial infarction, Alzheimer disease and obesity. *Hum. Mol. Gen.* **13**, 2647-2657 (2004).

22. Sparks, D.L. Coronary artery disease, hypertension, ApoE, and cholesterol: a link to Alzheimer's disease? *Ann. N. Y. Acad. Sci.* **826**, 128-146 (1997).

23. Lambert, J.-C. et al. Independent association of an APOE gene promoter polymorphism with increased risk of myocardial infaraction and decreased APOE plasma concentrations - the ECTIM study. *Hum. Mol. Gen.* **9**, 57-61 (2000).

24. Haan, J. & Peters, W.G. Amyloid and peripheral nervous system disease. *Clin. Neuro. Neurosurg.* **96**, 1-9 (1994).

25. Kyle, R.A. & Greipp, P.R. Amyloidosis (AL): Clinical and laboratory features in 229 cases. *Mayo Clin. Proc.* **58**, 665-683 (1983).

26. Kelly, J.J. Peripheral neuropathies associated with monoclonal proteins: A clinical review. *Muscle Nerve* **8**, 138-150 (1985).

Supplementary Figures

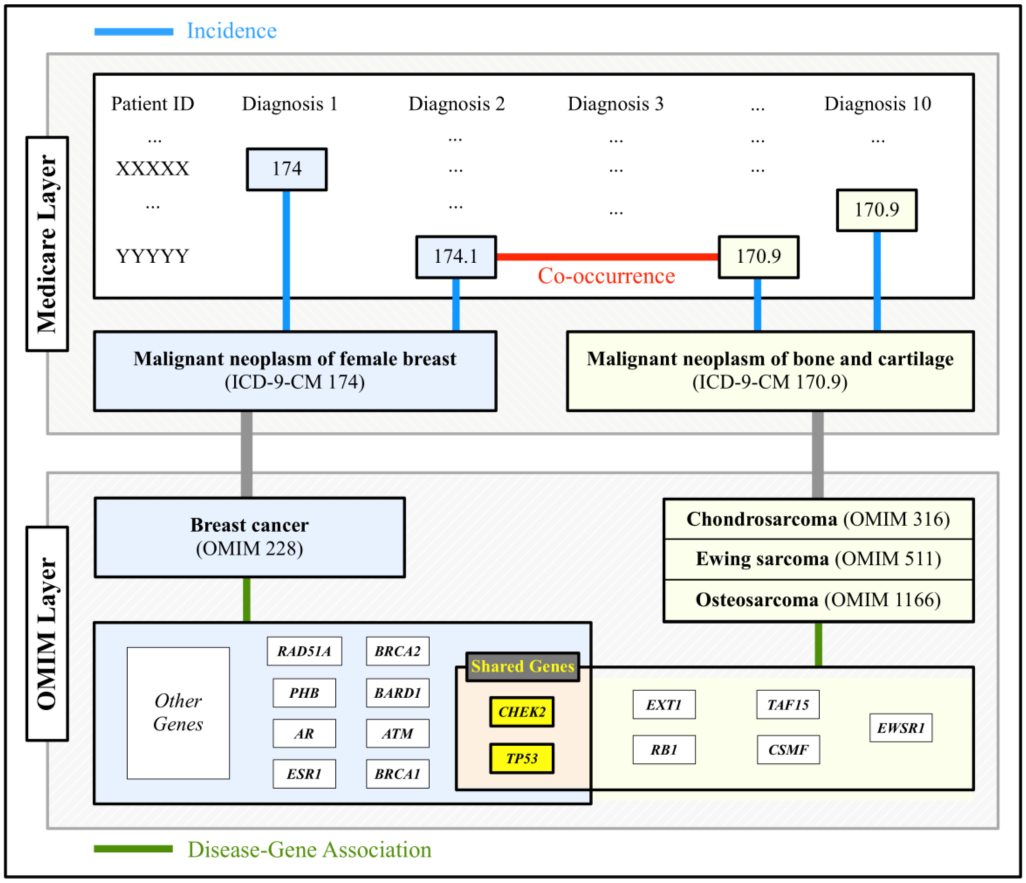

Fig S1. Schematic description of the procedure used to connect comorbidity (calculated in the Medicare Layer, top) and genetic associations (given in the OMIM Layer, bottom) between a pair of diseases. *Breast Cancer* and *Bone* and *Cartilage Cancer* are treated as the example here, also presented in Fig. 1B.

In the Medicare Layer (top), each disease is represented by an ICD-9-CM code, a widely used hierarchical disease diagnosis code system. The incidence I_i of each disease (represented by a blue line) is found by counting patients in the Medicare database diagnosed with the corresponding ICD-9-CM code and its sub-level codes (i.e. 174.1 is also counted as an incidence of 174 for breast cancer), while the co-occurrence C_{ij} (red line) of a disease pair is found by counting patients diagnosed with both codes. The comorbidity measures *RR* and ϕ can be calculated from these quantities and the total number of patients in the Medicare database (approximately 13 million). The associated genes of each disease are provided in the OMIM Layer (bottom, green lines). Due to differences in the disease-labeling schemes in the Medicare (ICD-9-CM) and the OMIM databases (the codes are as given in Goh *et al.* 2007), we manually constructed a mapping between the two (grey lines).

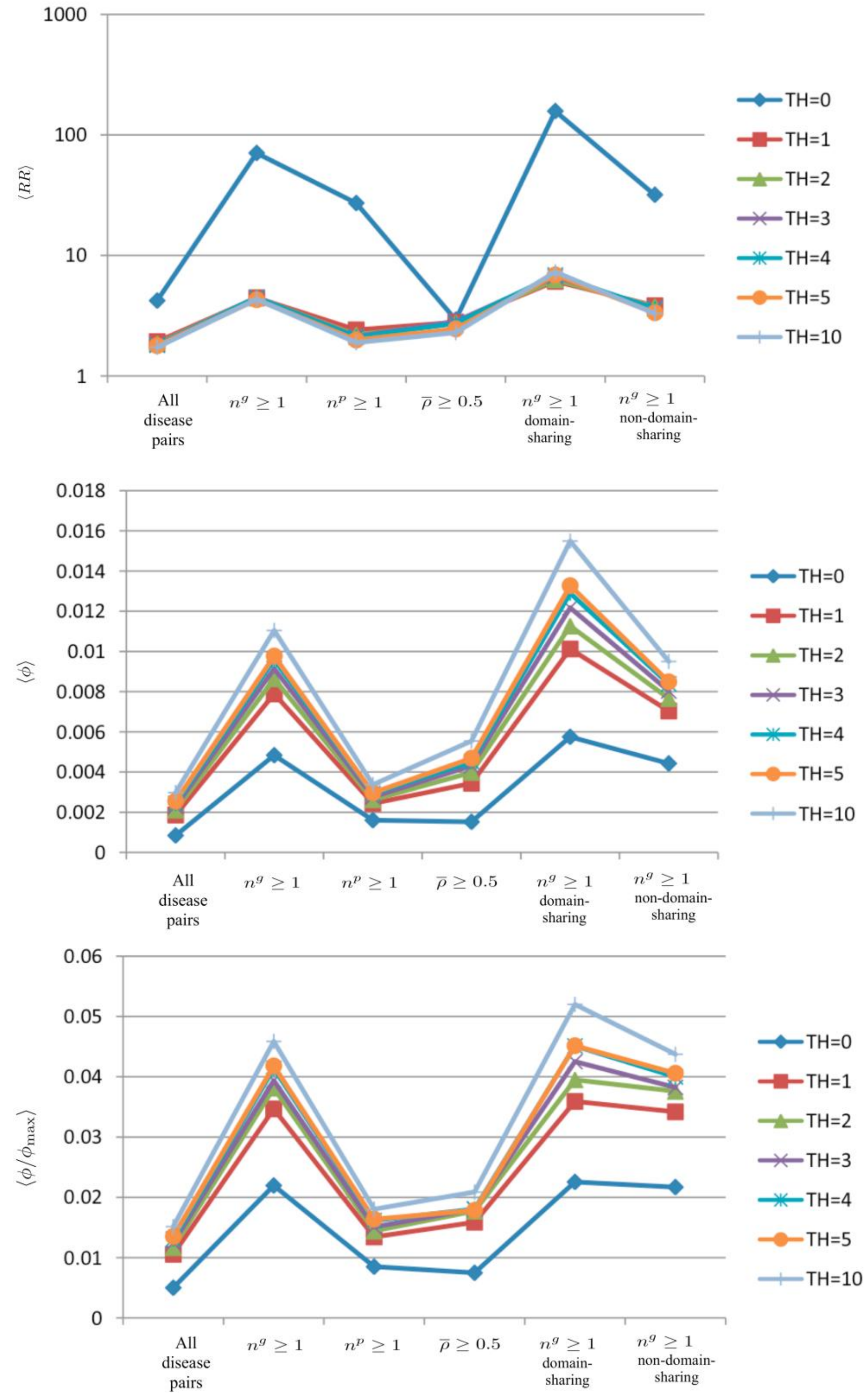

Fig S2. The average comorbidity values $\langle \boldsymbol{RR} \rangle$, $\langle \boldsymbol{\phi} \rangle$, and $\langle \boldsymbol{\phi}/\boldsymbol{\phi}^{\mathbf{max}} \rangle$ for disease pairs that satisfy genetic constraints and whose expected co-occurrence $\boldsymbol{C}_{\boldsymbol{ij}}^{*}$ equal or exceed a given threshold TH. The cases of $\langle \boldsymbol{RR} \rangle$ and $\langle \boldsymbol{\phi} \rangle$ for TH=1 are also shown in Fig. 1C. Imposing a threshold greatly reduces the magnitude of the comorbidity by removing pairs that exhibit unusually high $\boldsymbol{RR}$ values due to very small values of expected $\boldsymbol{C}_{\boldsymbol{ij}}^{*}$. Note that the curves with TH$\geq \mathbf{1}$ are stable over various thresholds.