The Impact of Cellular Networks on Disease Comorbidity

Molecular Systems Biology 5; Article number 262; doi:10.1038/msb.2009.16
Citation: ***Molecular Systems Biology*** **5:262**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

The impact of cellular networks on disease comorbidity

Juyong Park[1,2,*], Deok-Sun Lee[1,2,4], Nicholas A Christakis[3] and Albert-László Barabási[1,2,5,*]

[1] Departments of Physics, Biology, and Computer Science, Center for Complex Network Research, Northeastern University, Boston, MA, USA, [2] Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA, USA, [3] Department of Health Care Policy, Harvard Medical School, Boston, MA, USA, [4] Department of Natural Medicine Sciences, Inha University, Incheon, Korea and [5] Department of Medicine, Harvard Medical School, Boston, MA, USA
* Corresponding authors. J Park or L Barabási, Departments of Physics, Biology, and Computer Science, Center for Complex Network Research, Northeastern University 360 Huntington Ave, Boston, MA 02115, USA. Tel.: +1 617 373 7774; Fax: +1 617 373 4385; E-mails: perturbation@gmail.com or barabasi@gmail.com

Received 24.7.08; accepted 25.2.09

The impact of disease-causing defects is often not limited to the products of a mutated gene but, thanks to interactions between the molecular components, may also affect other cellular functions, resulting in potential comorbidity effects. By combining information on cellular interactions, disease–gene associations, and population-level disease patterns extracted from Medicare data, we find statistically significant correlations between the underlying structure of cellular networks and disease comorbidity patterns in the human population. Our results indicate that such a combination of population-level data and cellular network information could help build novel hypotheses about disease mechanisms.
Molecular Systems Biology 7 April 2009; doi:10.1038/msb.2009.16
Subject Categories: bioinformatics; molecular biology of disease
Keywords: cellular networks; comorbidity; database; population-level statistics

Introduction

Most cellular functions are carried out by a complex network of genes, proteins, and metabolites that interact through biochemical and physical interactions (Gerstein *et al*, 2002; Barabási and Oltvai, 2004; Albert, 2005; Basso *et al*, 2005; Almaas, 2007; Alon, 2007; Yildrim *et al*, 2007). Therefore, disease-causing defects may initiate cascades of failures that trigger the co-emergence of multiple diseases in a patient, such as diabetes and obesity. Yet, given the environmental, lifestyle, or treatment-related factors that all contribute to comorbidity, it is not obvious whether these cellular network-based interdependencies manifest themselves at the individual or at the population level. Discovering such systematic correlations between cellular networks and disease patterns could potentially open new avenues for understanding the human interactome, and may help uncover hitherto unknown disease mechanisms (Ergun *et al*, 2007; Loscalzo *et al*, 2007; Braun *et al*, 2008).

The possibility that there may be systematic links between hereditary diseases, thanks to their common genetic origins, was postulated recently by Goh *et al*, who created a Human Disease Network (HDN) by connecting all hereditary diseases that share a disease-causing gene according to the Online Mendelian Inheritance in Man (OMIM) database (Goh *et al*, 2007; Feldman *et al*, 2008). Although some of the diseases connected in the HDN captured well-known comorbidity patterns, the functional relevance of the links in the network remains to be demonstrated, leaving open the question whether most diseases connected in the HDN exhibit significant comorbidity. Interestingly, the most disconnected disease class in the HDN is that of metabolic diseases. However, Lee *et al* recently showed that metabolic diseases can be also organized in a metabolic disease network if the enzymes and their associated diseases are linked through metabolic pathways (Lee *et al*, 2008). Most importantly, the study found that metabolic diseases connected through shared pathways tend to show significant comorbidity, suggesting that information encoded in the structure of the metabolic network is amplified, becoming discernible at the population level as comorbidity patterns.

Metabolic networks represent only one of the several networks functionally relevant to our understanding of cellular activity. Indeed, when it comes to cellular interactions of potential importance to human diseases, we need to consider protein–protein interaction (PPI) and coexpression

networks as well as the links between diseases generated by shared genes. Therefore, earlier research raises an important question: are the cellular-level relationships encoded by PPIs, coexpression, and shared genes amplified at the population level? That is, should we expect statistically significant comorbidity patterns for disease pairs that share a gene, whose proteins interact, or whose genes show high coexpression patterns? To answer these questions, we analyzed the large-scale comorbidity pattern extracted from the US Medicare claims database and the gene–disease association network from OMIM (McCusick, 1998). We find that cellular interaction links indeed manifest themselves at the population level, resulting in statistically significant comorbidity patterns. We quantify the relative magnitude of these correlations and discuss the current difficulties in mapping population- and cellular-level data into each other, as well as the benefits of such an approach toward elucidating disease mechanisms.

Results and discussion

The starting point of our study is the Medicare claims database containing the diagnoses that C_{ij} led to the hospitalization of N=13 039 018 elderly patients, each disease or condition identified by an ICD-9-CM code. We denote the incidence of disease i with I_i, and the number of patients who were simultaneously diagnosed with diseases i and j with C_{ij}. The comorbid tendency between the two diseases can be quantified using either the relative risk, $RR=C_{ij}/C_{ij}^*$, where $C_{ij}^*=I_iI_j/N$ is the expectation value of C_{ij} when the two diseases are independent, or the ϕ-correlation defined as $\phi = (NC_{ij} - I_iI_j)/\sqrt{I_iI_j(N-I_i)(N-I_j)}$. When two diseases co-occur more frequently than expected by chance, we have $RR>1$ and $\phi>0$. Note, however, that although RR and ϕ are not independent of each other, each carries unique biases that are complementary. Therefore, we use both measures of comorbidity to ensure the robustness of our findings (see Supplementary information (SI) for further details). The disease–gene associations used in the study were obtained from the OMIM database, which contains >4900 such associations as of October 2008. Although the disease–gene record is far from complete, OMIM is currently the most complete repository of all known disease genes and their associated disorders.

It is important to note that disease names used in the Medicare database by the medical and the insurance communities (the ICD-9-CM scheme) and those used in the OMIM database by geneticists are not identical. Therefore, we enlisted a professional ICD-9-CM coder to manually map the OMIM disease names into ICD-9-CM codes and established connections between the genetic associations and the comorbidity measures (see Box 1 and SI sections S1 and S2 for more detail). On account of the discrepancies in disease names and the complex, hierarchical nature of the ICD-9-CM scheme, we recognize that the mapping is not perfect, and may contain debatable and occasionally erroneous ICD-9-CM-to-OMIM correspondence. Therefore, we are providing the mapping used by us in the SI, offering a chance for the community to improve on it in future studies.

As OMIM comprises the set of hereditary or complex diseases with validated gene–disease associations, it is anticipated that only a subset of ICD-9-CM codes would correspond to the diseases in the OMIM. Indeed, we find that, of the >12 000 available ICD-9-CM codes, 763 unique ICD-9-CM codes can be mapped to OMIM diseases. The fact that our analysis is limited to 5% of possible diagnosis codes contained in the Medicare database, could limit our population (patient) coverage. We find, however, that this is not the case: as Figure 1A shows, 90% of patients in the Medicare database are diagnosed with at least one disease whose ICD-9-CM code is contained in our mapping to the OMIM database.

We use the following three quantities to capture the cellular network-level relationship between diseases i and j, as illustrated in Figure 1B for the case of breast cancer (ICD-9-CM 174) and cancer of bone and cartilage (ICD-9-CM 170.9, see also SI):

(i) n_{ij}^g, the number of shared genes associated with both diseases i and j, which quantifies the potential common genetic origin of the two diseases (Goh *et al*, 2007);

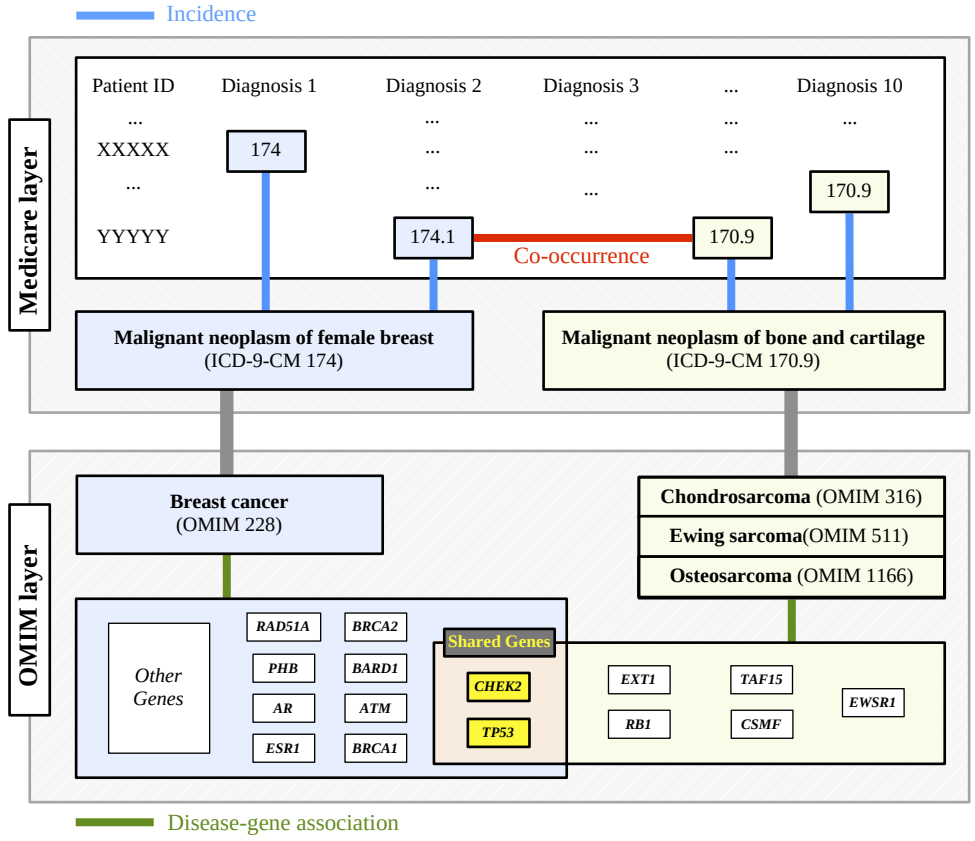
(ii) n_{ij}^p, the number of PPIs between the proteins of diseases i and j capturing the PPI network-level relationships between them (Rual *et al*, 2005; Stelzl *et al*, 2005).

(iii) $\bar{\rho}_{ij}$, the average Pearson correlation of coexpression between pairs of genes from each disease, capturing the degree to which the genes associated with the two diseases are coexpressed (Ge *et al*, 2005).

The main question can be formulated as follows: does the existence of these cellular-level links (i.e., $n_{ij}^g>0$, $n_{ij}^p>0$, $\bar{\rho}_{ij}>0$) between the two diseases increase the likelihood that individuals simultaneously develop both conditions? We start our investigation by measuring the Pearson correlation between the cellular variables (n_{ij}^g, n_{ij}^p, $\bar{\rho}_{ij}$) and comorbidities (RR and ϕ) for 83 924 disease pairs. Of these, 2239 pairs are linked through either shared genes ($n_{ij}^g \geqslant 1$) or PPIs ($n_{ij}^p \geqslant 1$; 658 with shared genes, and 1873 with PPIs). In Figure 2A and Table I we present the Pearson correlation coefficients (PCCs) between the comorbidity measures and the genetic variables. Although n^g, in general, has the highest correlation with comorbidity, we do observe positive PCC with all three variables.

There are numerous factors that determine whether two diseases co-occur in a patient, some of which are environmental, lifestyle-related or treatment-induced. Our study captures only the role of the cellular network on comorbidity. The small magnitude of the correlations observed by us suggests that the cellular network offers only a small contribution to the observed comorbidity. Note, however, that placing significant emphasis on the magnitude of these correlations is premature, as two known effects limit the correlations observed by us. First, the magnitude of the correlation is limited by the predictive power of specific genetic mutations catalogued in the OMIM database, and the likelihood of a patient developing a particular disease. Indeed, it is known that genetic mutations result in an increase of at most a few percentage points in the likelihood of an individual developing a specific complex disease (Loscalzo, 2007) and the correlations observed by us cannot exceed the known disease–gene correlations. Second, the correlations are further limited by the noise in the mapping between the OMIM diseases and the ICD-9-CM codes. As we noted, there is inherent ambiguity both in the mapping as well as in the process of assigning a

Box 1 Didactic Box



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 Figure 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). Because of 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). See Supplementary information for detail.

particular diagnosis to specific ICD-9-CM codes in hospitals. Each instance of such misdiagnosis or mapping ambiguity decreases the magnitude of the observed correlations. Therefore, at this point it is not the magnitude, but the statistical significance of the correlations that we can rely on. As summarized in Table I, the observed correlations are statistically significant.

To quantify the degree of comorbidity caused by the observed correlations, we measured the average comorbidities $\langle RR \rangle$ and $\langle \phi \rangle$ for disease pairs that are connected at the cellular network level. Compared with the entire set of 83 924 pairs of hereditary diseases considered in our study, we find (see Figure 2B; Table II) a two- to four-fold increase in the average comorbidity in disease pairs that share genes ($n_{ij}^{g} \geqslant 1$), indicating that if a patient develops a particular disease associated with a gene or multiple genes in the HDN, then they have a two-fold higher chance of developing another disease mapped to one or more common genes in the HDN, compared with diseases that are not. An increased comorbidity is also observed for disease pairs linked through PPIs ($n_{ij}^{p} \geqslant 1$) and high coexpression ($\bar{\rho}_{ij} \geqslant 0.5$).

The observed correlations between the cellular links and comorbidities raise a related question: would disease pairs that are more interconnected than others (i.e., have larger n_{ij}^{g}, n_{ij}^{p}, or $\bar{\rho}_{ij}$) show higher comorbidity? To address this, in Figure 2C we show that comorbidity increases rapidly with the number of shared genes: sharing two or more genes ($n_{ij}^{g} \geqslant 2$) results in nearly a five-fold increase in comorbidity compared with hereditary disease pairs that do not share genes. An increase in comorbidity is observed with increasing n_{ij}^{p} and $\bar{\rho}_{ij}$ as well (Figure 2D and E), although the effect is weaker than that observed for n_{ij}^{g}, which is not unexpected given the smaller impact that n_{ij}^{p} and $\bar{\rho}_{ij}$ have on comorbidity in comparison with n_{ij}^{g}.

Note that the average comorbidity measured between all pairs of diseases is >1 (Figure 2B), indicating that many patients develop multiple disorders, whether or not the specific diseases are linked at the cellular level. Such correlations have been observed in other studies focused on

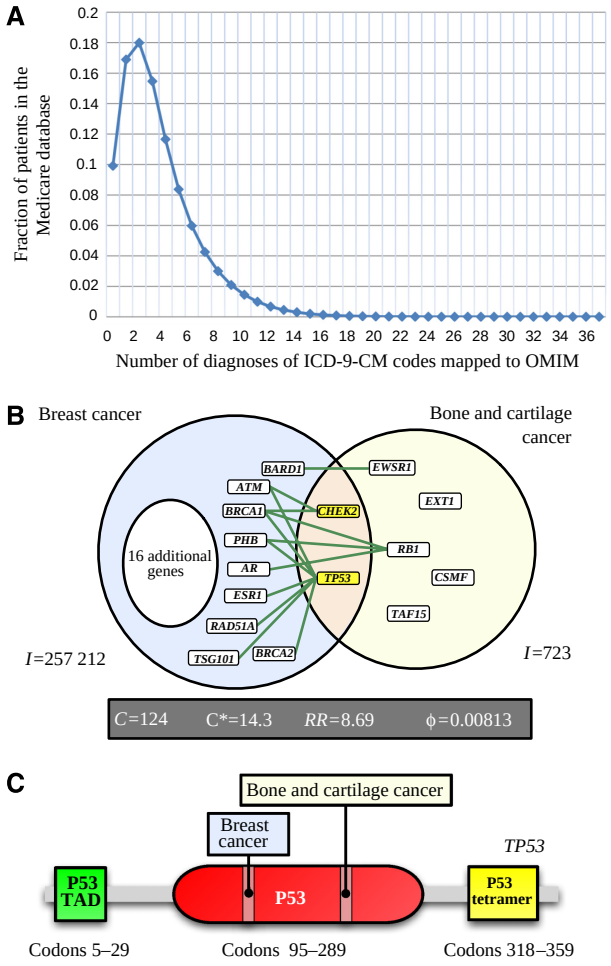


Figure 1 (**A**) The fraction of patients in the Medicare database diagnosed with ICD-9-CM codes mapped to OMIM diseases. Although they represent fewer than 6% of all ICD-9-CM codes, 90% of the patients were diagnosed with at least one. (**B**) Breast cancer and cancer of bone and cartilage offer an example of a disease pair linked on the cellular-network level. They share two genes (*CHEK2* and *TP53*), and their proteins interact through 13 protein–protein interactions (green lines). The average coexpression $\bar{\rho}$ between the genes of each disease is 0.103. Their comorbidity between the two diseases are $RR=8.69$, indicating that the number of patients who simultaneously develop both diseases shows a sevenfold increase compared with random expectation, and $\phi=0.00813$ ($P\approx 6\times 10^{-71}$). (**C**) The functional domains of the *TP53* protein. Breast cancer and cancer of bone and cartilage shown also constitute a domain-sharing disease pair: mutations on the *TP53* protein associated with breast cancer and cancer of bone and cartilage take place on the same P53 domain.

comorbidity patterns (Rzhetsky *et al*, 2007; Hidalgo *et al*, 2009). These overall comorbidity patterns are not particularly surprising considering that the Medicare population is 65 years of age or older, the age at which individuals do develop multiple disorders. Thus, the overall comorbidity represents the baseline against which we can assess the impact of the genetic and cellular networks. It is reassuring, therefore, that hereditary diseases that are linked in the HDN (and thus at the cellular level) show comorbidity higher than the baseline $\langle RR\rangle=1.92\pm0.01$ and $\langle\phi\rangle=(1.84\pm0.02)\times10^{-3}$ observed for the set of all disease pairs.

Despite the significant increase in $\langle RR\rangle$ and $\langle\phi\rangle$, there are many disease pairs that share genes yet fail to show significant comorbidity. We hypothesize that this is, in part, because of pleiotropy, which in this context means that different mutations on the same gene can have different pathological effects on a protein (Dudley *et al*, 2005), thereby predisposing an individual to different disorders. In general, we expect that disease pairs associated with mutations on the same functional domain of the shared protein show higher comorbidity than disease pairs whose mutations occur in different functional domains. To test this hypothesis, we identified the functional domains of disease-causing mutations on shared genes using the Pfam database (Finn *et al*, 2006). In agreement with our hypothesis, we find higher $\langle RR\rangle$ and $\langle\phi\rangle$, for disease pairs whose mutations are on the same domain of the shared gene, compared with disease pairs whose mutations are in distinct functional domains (Figure 2B).

The observed correlations suggest that a combination of disease data and cellular network information may assist us in identifying new comorbidity patterns alongside their potential genetic origin. Indeed, upon inspection of the 2239 disease pairs that are genetically linked (i.e., $n_{ij}^{g}\geqslant1$ or $n_{ij}^{p}\geqslant1$), we find several disease pairs whose comorbidity patterns are already well known to the medical community, such as diabetes and obesity (Evans *et al*, 2002), or breast cancer and osteosarcoma (Knowling and Basco, 1986). At the same time, due to the aforementioned mismatch between disease names used by clinicians (within the ICD-9 coding scheme) and by geneticists (within the OMIM tabulation), several highly comorbid disease pairs are readily anticipated (such as diabetes and hypoglycemia, as hypoglycemia is a common side effect of the treatment of diabetes) or cases in which one disease is a broader version of the other (such as mononeuritis and hereditary peripheral neuropathy). Such mapping limitations notwithstanding, we find several interesting disease pairs that are linked at the cellular level and also show significant comorbidity. For example, consider Alzheimer's disease (ICD-9-CM 331) and myocardial infarction (ICD-9-CM 410.9), for which earlier comorbidity studies were either inconclusive or contradictory (Bursi *et al*, 2006). As Figure 3A shows, we not only find statistically significant comorbidity ($P\approx10^{-5}$) between the two, but the figure suggests that the shared ACE and APOE genes may contribute to the observed effect. Similarly, we observe significant comorbidity ($P\approx10^{-148}$) between autonomic nervous system disorder (ICD-9-CM 337.9) and carpal tunnel syndrome (ICD-9-CM 354, Figure 3B). A known mechanism is L-chain amyloidosis, which may affect the autonomic nervous system and causes carpal tunnel syndrome when the amyloid infiltrates the flexor retinaculum of the patient's wrist (Haan and Peters, 1994). Figure 3B, however, suggests that a PPI between the associated genes of each disorder may also play a role in the observed effect. Although there may be additional possible physiological or social explanations for some of the observed comorbidities (see SI), the method described above has the potential to offer new, testable hypotheses about the biological basis of disease interrelationships. These examples were selected only to demonstrate the potential of the combined investigation of the network and population-level data in identifying potentially interesting disease pairs worthy of further study. A more detailed description of these disease pairs, along with the complete list of the 2239 genetically linked disease pairs and their genetic associations are provided in the SI.

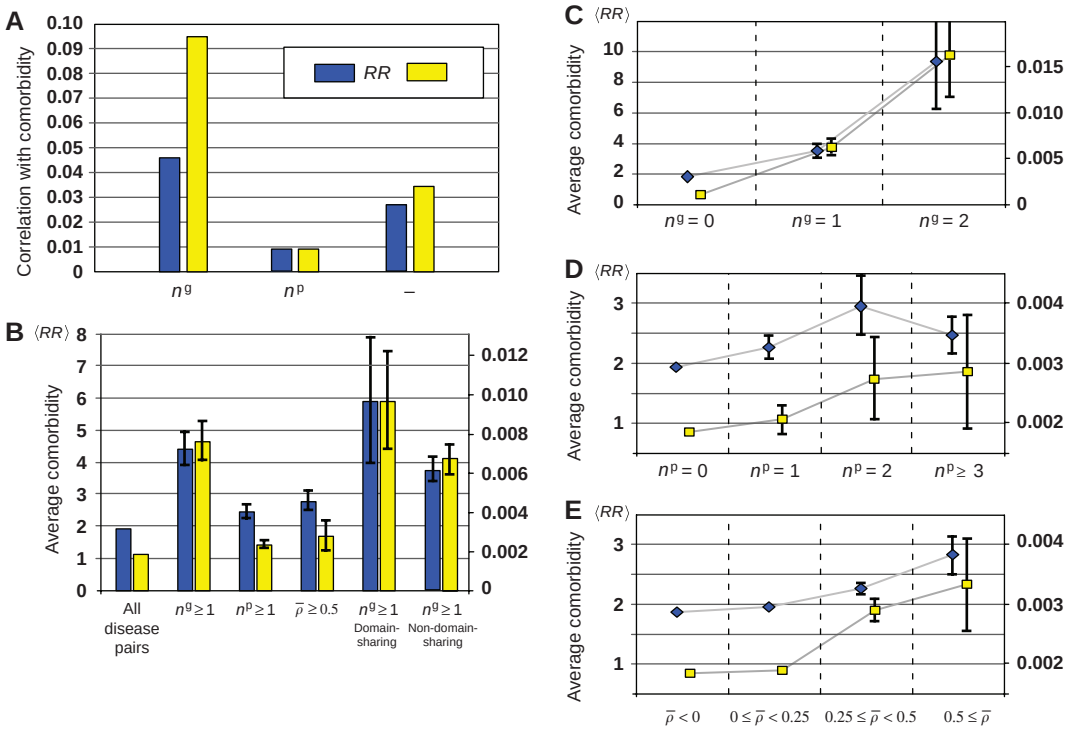


Figure 2 (**A**) The Pearson correlation between comorbidity and the three quantities (n^g, n^p, $\bar{\rho}$) that capture cellular-level links between diseases. See also Table I. (**B**) Average comorbidity for disease pairs satisfying the cellular constraints discussed in the text. See also Table II. (**C**–**E**) Average comorbidity for disease pairs with increasing values of n^g, n^p, and $\bar{\rho}$.

Table I The Pearson correlation between relative risk, ϕ-correlation and the three genetic variables (see Figure 2A)

Genetic variables	Pearson correlation with relative risk	Pearson correlation with ϕ-correlation
n^g	0.0469 ($P \approx 3.85 \times 10^{-4}$)	0.0902 ($P \approx 1.48 \times 10^{-4}$)
n^p	0.00948 ($P \approx 1.65 \times 10^{-2}$)	0.00941 ($P \approx 1.49 \times 10^{-2}$)
$\bar{\rho}$	0.0272 ($P \approx 1.07 \times 10^{-3}$)	0.0334 ($P \approx 3.41 \times 10^{-4}$)

All correlations are positive, with *P*-values shown in parentheses.

Table II The average comorbidity of disease pairs satisfying various criteria, quantifying the strength of the cellular and genetic links connecting the pairs

	Number of pairs	*RR* (error)	ϕ (error)
All diseases	83 924	1.92 (0.01)	1.84 (0.02) $\times 10^{-3}$
$n^g \geqslant 1$	658	4.35 (0.60)	7.57 (0.96) $\times 10^{-3}$
$n^p \geqslant 1$	1873	2.35 (0.16)	2.28 (0.23) $\times 10^{-3}$
$\bar{\rho} \geqslant 0.5$	215	2.79 (0.32)	3.44 (0.82) $\times 10^{-3}$
$n^g \geqslant 1$ Domain-sharing	182	5.98 (1.86)	9.77 (2.85) $\times 10^{-3}$
$n^g \geqslant 1$ Non-domain-sharing	476	3.73 (0.38)	6.72 (0.84) $\times 10^{-3}$

The main finding of this paper is that health care and treatment data on a large number of individuals offer information useful to systems biology that can complement the information from the well-established genomic studies. Indeed, Medicare and insurance databases already collect the health care history of millions of individuals, allowing us to uncover the correlations in the occurrence of various diseases. In parallel, increasing knowledge about the molecular origin of disease indicates that many disorders are rooted in defects in gene products that are part of the same cellular network, raising the possibility that these diseases should co-occur in the same individual. Admittedly, much of the currently available network data are incomplete and probably noisy. We may be approaching a tipping point, however, where we have acquired sufficient knowledge of human cellular networks to begin understanding the way a disturbance in the networks may contribute to the development of a disease and suggest potential disease-modifying factors.

To test the validity of this hypothesis, here we correlated cellular level information for human cells, namely data on shared genes, PPIs, and coexpression patterns, with comorbidity data obtained from the Medicare database. Despite the aforementioned limitations of the mapping and the data collection process, we found statistically significant correlations between cellular interactions and comorbidity patterns. We also found that disease pairs with higher correlations tend to be linked more strongly in the cellular network.

Although our work was mainly driven by the desire to uncover evidence that cellular information is amplified in the human population and thus can be detected from patient data, our results point to the potential usefulness of our approach in uncovering disease mechanisms. Indeed, we discuss two disease pairs in which the network-based information offers a plausible mechanism for statistically significant comorbidity patterns. These results suggest that Medicare and other insurance databases could play an increasing role in future studies of the systems biology of human cells and diseases.

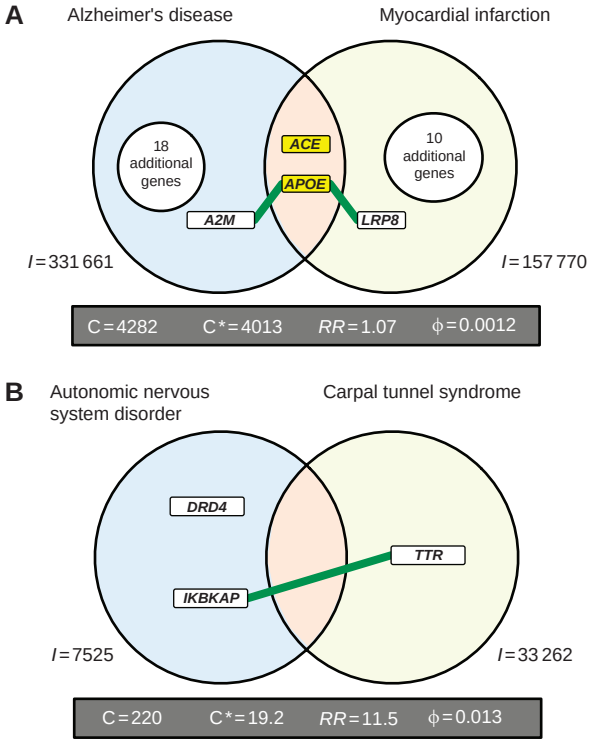


Figure 3 Two examples of disease (disorder) pairs with significant comorbidity that are connected at the cellular level through either shared genes (**A**) or protein–protein interactions (A and B). (A) Alzheimer's disease and myocardial infarction ($P \approx 10^{-5}$). (**B**) Autonomic nervous system disorder and carpal tunnel syndrome ($P \approx 10^{-148}$).

Materials and methods

Data sets

We used the HDN from Goh *et al* for disorder–gene associations (Goh *et al*, 2007), updated based on the version of the Morbid Map from OMIM at the time of the study. The most up-to-date version can be found at http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim. The PPI data were taken from Rual *et al* (2005) and Stelzl *et al* (2005). The genetic coexpression levels were calculated on the basis of an Affeymetrix microarray data (Ge *et al*, 2005) (see www.affymetrix.com). Protein domain information is available on UnitProt (http://www.uniprot.org/) and Pfam (http://www.sanger.ac.uk/Software/Pfam/).

Estimating *P*-values and errors

The *P*-values for the PCCs shown in Figure 2A and Table I were calculated by a Monte Carlo sampling method. We generate a randomized sequence of the genetic variables and calculate its PCC with comorbidity. After 2 million randomizations, the *P*-value is the fraction of the total trials that resulted in a PCC that is larger than what was observed.

As *RR* and ϕ are monotonically increasing functions of C_{ij}, their one-sided *P*-value is equal to the sum of probabilities that the co-occurrence C_{ij} is larger than the actual value. It can be obtained using standard computational software such as Mathematica (www.wolfram.com) by approximating the binomial distribution generated from the number of patients *N* and $C_{ij}^* = Np = I_i I_j / N$ as a Poisson distribution, and therefore

$$P = \sum_{k=C_{ij}}^{N} \frac{\exp(-C_{ij}^*) \times (C_{ij}^*)^k}{k!}$$

The errors on the comorbidity values (Figure 2) were calculated using the bootstrap method based on resampling (Efron, 1979; Newman and Barkema, 1998).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Quan Zhong, Cesar Hidalgo, Nick Blumm, and Marc Vidal for useful discussions. This research was supported by JSMF 220020084, NSF ITR DMR-0426737, NIH CEGS-1P50HG4233/CFDA #93.172, NIH U01 A1070499-01/111620-2, and NIH U56 CA113004/sub MGH.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Albert R (2005) Scale-free networks in cell biology. *J Cell Sci* **118:** 4947–4957

Almaas E (2007) Biological impacts and context of network biology. *J Exp Biol* **210:** 1548–1558

Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Gen* **8:** 450–461

Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Gen* **5:** 101–113

Basso K, Margolin AA, Stolovitzky G, Klein U, Della-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37:** 382–390

Braun P, Rietman E, Vidal M (2008) Networking metabolites and diseases. *Proc Natl Acad Sci* **105:** 9849–9850

Bursi F, Rocca WA, Kilian JM, Weston SA, Knopman DS, Jacobsen SJ, Roger VL (2006) Heart disease and dementia: a population-based study. *Am J Epidemiol* **163:** 135–141

Dudley AM, Janse DM, Tanay A, Shamir R, Church GM (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol* **1:** 1

Efron B (1979) Computers and the theory of statistics: thinking the unthinkable. *SIAM Rev* **21:** 460–480

Ergun A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ (2007) A network biology approach to prostate cancer. *Mol Syst Biol* **3:** 82

Evans JMM, Newton RW, Ruta DA, MacDonald TM, Morris AD (2002) Socio-economic status, obesity and prevalence of Type 1 and Type 2 diabetes mellitus. *Diabet Med* **17:** 478–480

Feldman I, Rzhetsky A, Vitkup D (2008) Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci* **105:** 4323–4328

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006) Pfam: clans, web tools and services. *Nucl A Res* **34:** D247–D251

Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, Aburatani H (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth-of-expression in normal tissues. *Genomics* **86:** 127–141

Gerstein M, Lan N, Jansen R (2002) Interacting interactomes. *Science* **295:** 284–286

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) The human disease network. *Proc Natl Acad Sci* **104:** 8685–8690

Haan J, Peters WG (1994) Amyloid and peripheral nervous system disease. *Clin Neuro Neurosurg* **96:** 1–9

Hidalgo CA, Blumm N, Barabási A-L, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. *PLOS Comp Bio*; doi:pcbi 1000353 (in press)

Knowling MA, Basco VE (1986) Breast-cancer after treatment for osteosarcoma. *Med Pediatr Oncol* **14:** 51–53

Lee D-S, Park J, Kay K-A, Christakis N-A, Oltvai ZN, Barabási A-L (2008) The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci* **105:** 9880–9885

Loscalzo J (2007) Association studies in an era of too much information: clinical analysis of new biomarker and genetic data. *Circulation* **116:** 1866–1879

Loscalzo J, Kohane I, Barabasi A-L (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol* **3:** 124

McCusick VA (1998) *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, 12 edn Baltimore: Johns Hopkins University Press

Newman MEJ, Barkema GT (1998) *Monte Carlo Methods in Statistical Physics*. Oxford (UK): Clarendon Press

Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437:** 1173–1178

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

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B *et al* (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122:** 957–968

Yildrim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) Drug-target network. *Nat Biotech* **25:** 1119–1126