Molecular Systems Biology 7; Article number 494; doi:10.1038/msb.2011.29
Citation: ***Molecular Systems Biology*** **7:494**
© 2011 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/11
www.molecularsystemsbiology.com

Protein localization as a principal feature of the etiology and comorbidity of genetic diseases

Solip Park[1], Jae-Seong Yang[1], Young-Eun Shin[2], Juyong Park[3], Sung Key Jang[1,2,4,5] and Sanguk Kim[1,2,6,*]

[1] School of Interdisciplinary Bioscience and Bioengineering, Pohang University of Science and Technology, Pohang, Korea, [2] Division of Molecular and Life Science, Pohang University of Science and Technology, Pohang, Korea, [3] Physics Department, Kyung Hee University, Seoul, Korea, [4] Division of Integrative Bioscience and Biotechnology, Pohang University of Science and Technology, Pohang, Korea, [5] Biotechnology Research Center, Pohang University of Science and Technology, Pohang, Korea and [6] Division of IT Convergence Engineering, Pohang University of Science and Technology, Pohang, Korea
[*] Corresponding author. Division of Molecular and Life Science, Pohang University of Science and Technology, Pohang 790-784, Korea. Tel.: +82 54 279 2348; Fax: +82 54 279 2199; E-mail: sukim@postech.ac.kr

Received 16.11.10; accepted 19.4.11

Proteins targeting the same subcellular localization tend to participate in mutual protein–protein interactions (PPIs) and are often functionally associated. Here, we investigated the relationship between disease-associated proteins and their subcellular localizations, based on the assumption that protein pairs associated with phenotypically similar diseases are more likely to be connected via subcellular localization. The spatial constraints from subcellular localization significantly strengthened the disease associations of the proteins connected by subcellular localizations. In particular, certain disease types were more prevalent in specific subcellular localizations. We analyzed the enrichment of disease phenotypes within subcellular localizations, and found that there exists a significant correlation between disease classes and subcellular localizations. Furthermore, we found that two diseases displayed high comorbidity when disease-associated proteins were connected via subcellular localization. We newly explained 7584 disease pairs by using the context of protein subcellular localization, which had not been identified using shared genes or PPIs only. Our result establishes a direct correlation between protein subcellular localization and disease association, and helps to understand the mechanism of human disease progression.

Molecular Systems Biology **7**: 494; published online 24 May 2011; doi:10.1038/msb.2011.29
Subject Categories: metabolic and regulatory networks; molecular biology of disease
Keywords: cellular networks; comorbidity; human disease; subcellular localization

Introduction

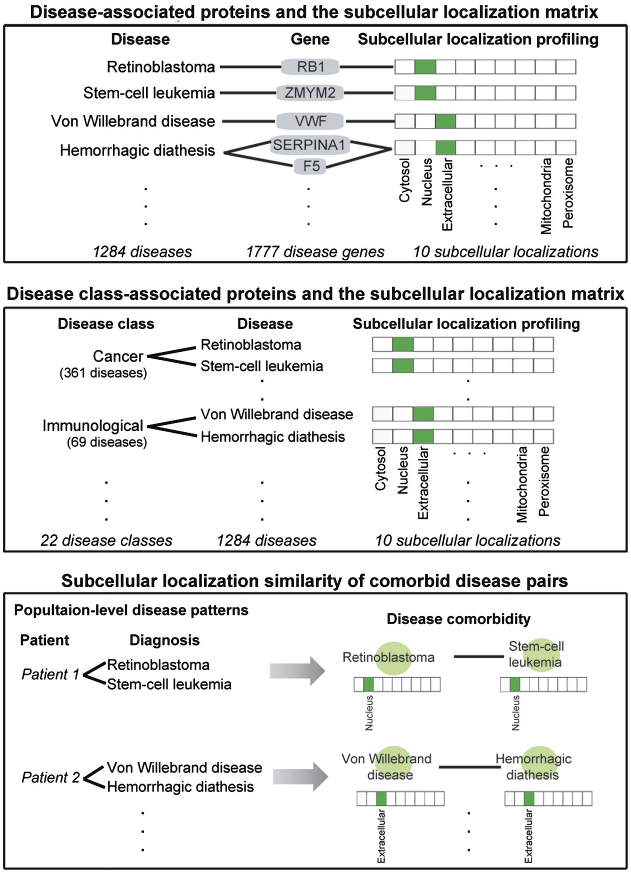
Establishing the interrelationship between the genotype and the phenotype is one of the most challenging yet pertinent problems in biomedical research (Lamb *et al*, 2006). Molecular and genetic studies of diseases have been devoted to identifying disease-causing mutations through diverse gene-based methods such as recombination mapping and genome-wide association studies (Botstein and Risch, 2003; Broeckel and Schork, 2004). Traditional gene-based approaches have been compiled into a list of disease-associated genes. In addition, the rapid accumulation of functional genomics and proteomics data provides information on the protein–protein interactome, an extensive map of metabolism, and regulatory networks that complement current gene-based approaches (Rual *et al*, 2005; Stelzl *et al*, 2005; Duarte *et al*, 2007; Shlomi *et al*, 2008).

Recently, it was shown that the emergence of phenotypically similar diseases are triggered as a result of molecular connections between disease-causing genes (Oti and Brunner, 2007; Zaghloul and Katsanis, 2010). From a genetics perspective diseases are associated with certain genes (Goh *et al*, 2007; Feldman *et al*, 2008), whereas from a proteomics perspective phenotypically similar diseases are connected via biological modules such as protein–protein interactions (PPIs) or molecular pathways (Lage *et al*, 2007; Jiang *et al*, 2008; Wu *et al*, 2008; Linghu *et al*, 2009; Suthram *et al*, 2010). These molecular connections between diseases were observed on the population level as well: diseases connected through molecular connections such as shared genes, PPIs, and metabolic pathways tend to show elevated comorbidity (Rzhetsky *et al*, 2007; Lee *et al*, 2008; Zhernakova *et al*, 2009; Park *et al*, 2009a). While these findings constitute a step toward improving our understanding of the mechanism of disease progression, there are still many more molecule-level connections between disease pairs that need to be explored in order to establish a firmer comorbidity association.

Subcellular localization provides spatial information of proteins in the cell; proteins target subcellular localizations to interact with appropriate partners and form functional complexes in signaling pathways and metabolic processes (Au *et al*, 2007). Mutations in disease-causing genes alter the synthesis of the gene product, or change the targeting process of proper subcellular localizations, which in turn perturb the cellular functions of the proteins. Abnormal protein localizations are known to lead to the loss of functional effects in diseases (Luheshi *et al*, 2008; Laurila and Vihinen, 2009). For example, mis-localizations of nuclear/cytoplasmic transport have been detected in many types of carcinoma cells (Kau *et al*, 2004). A proper identification of protein subcellular localization can hence be useful in discovering disease-associated proteins (Giallourakis *et al*, 2005; Calvo and Mootha, 2010). Also, we have previously demonstrated that proteins associated with the same disease tend to localize in the same subcellular compartments (Park *et al*, 2009b). With this understanding, we postulate that disease-associated proteins connected by subcellular localizations could also explain the phenotypic similarities between diseases. Furthermore, such connections may also couple to disease progressions that contribute to multiple disease manifestation, that is, comorbidity.

In this study, we investigated the interrelationship between diseases and subcellular localizations. Furthermore, we also explored the molecular connections between disease-associated proteins, and applied the subcellular localization similarity of disease pairs to understanding the human disease progression by analyzing comorbid disease pairs (Box 1 and described further in Materials and methods). We constructed, for the first time, a matrix of disease-associated proteins and their subcellular localization which describes the interrelationship between the two. From this matrix, we found that proteins associated with the same disease are likely enriched in particular subcellular localizations in the cell. We also observed that phenotypically similar diseases clustered in the same disease classes are associated with particular subcellular localizations. Furthermore, a positive correlation was found between subcellular localization similarity of disease pairs and comorbidity measures, which explains the molecular connections between comorbid disease pairs connected via subcellular localization. Subcellular localization furthermore enhanced the comorbid tendencies of disease pairs, and uncovered the hitherto-unknown molecular connections between 7584 disease pairs. This constitutes a novel approach to establishing the relationship between protein subcellular localization and the molecular connections of comorbid disease pairs, offering insight into previously unexplained mechanisms of disease progression.

Box 1 Schematic overview of the relationship between diseases and subcellular localizations



To build disease-associated proteins and subcellular localization matrix, 1284 diseases and 1777 disease-associated proteins were taken from OMIM database (Hamosh *et al*, 2005). Each disease-associated protein was mapped onto relevant subcellular localizations. Diseases were classified into 22 disease classes by the physiological system affected (Goh *et al*, 2007). (Middle) Subcellular localization information of the classified disease-associated proteins was attributed to the profile of disease classes. Disease progression was compiled from the hospitalization of 13 million patients from US Medicare database (Park *et al*, 2009a). Comorbid disease pairs were identified by calculating co-occurring disease pairs in individual patients. Subcellular localization similarity was calculated from the quantitative relationship between comorbid disease pairs and their subcellular localization profiles.

Results

Systematic construction of the atlas of human disease-associated proteins and their subcellular localizations

Protein subcellular localization has been extensively studied through various methods to determine a variety of protein functions. To the best of our knowledge, the connection between diseases and subcellular localizations are yet to be studied systematically. To resolve this, we constructed, for the first time, a human Disease-associated Protein and subcellular Localization (DPL) matrix (top panel in Box 1). For this purpose, we utilized the list of 1284 diseases representing the grouping of phenotypes (MIM record) based on disease names and their 1777 associated proteins available from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al*, 2005). This approach has been widely used in recent systematic disease analyses of shared molecular characteristics between disease subtypes (Lee *et al*, 2008; Park *et al*, 2009a; Li and Patra, 2010).

Disease-associated proteins were mapped to their encoded subcellular localizations based on the Swiss Prot annotation scheme and the consensus localization predictions we recently reported (Park *et al*, 2009b; see Supplementary File 1). We considered 10 different subcellular localizations (cytosol, endoplasmic reticulum (ER), extracellular, Golgi, peroxisome, mitochondria, nucleus, lysosome, plasma membrane, and others) for the localization mapping of disease-associated proteins, although minor localizations were considered simply as 'others' since the number of disease-associated proteins of such locations was too small to analyze (fewer than 10 proteins with confidence). We analyzed the covariance of a disease with a subcellular localization by identifying the number of disease-associated proteins by co-assigning diseases and subcellular localizations. Then, the DPL matrix was built by transforming the covariance into an association score (AS) between a disease and a subcellular localization (see Materials and methods and Supplementary File 2).

Diseases have their unique subcellular localization profiles

Our DPL matrix provides the '*cellular localization map of diseases*' that represents the spatial index of diseases in the cell. We found that each disease shows unique characteristics of subcellular localization profile in the DPL matrix. We were interested in determining whether subsets of 1284 human diseases exhibit distinct enrichment profiles across subcellular localizations. We calculated pairwise correlations and performed a hierarchical clustering of the enrichments of the 1284 diseases across 10 different subcellular localizations (Figure 1). To validate the reliability of ASs, we calculated their *Z*-values; the *Z*-value represents the significance of the subcellular localization enrichment of a disease. We observed that the *Z*-values and subcellular localization-disease association scores are indeed highly correlated (R^2=0.97), and we considered an AS $\geqslant$0.05 to be statistically significant ($P<0.01$; Supplementary Figure 1A). Specifically, diseases that are caused by molecular defects in specific organelles showed significant ASs (AS $\geqslant$0.2, *Z*-value $>$10, $P<1.00\times10^{-10}$) (Supplementary Figure 1B). For example, Mitochondria Complex I-III deficiency, a well-known mitochondrial disease (Pagliarini *et al*, 2008; Rotig, 2010), was significantly enriched within the mitochondria (*Z*-value=10.6, $P<1.00\times10^{-10}$) (Supplementary Figure 1B). Also, Adrenoleukodystrophy, a peroxisome biogenesis disorder (Wanders and Waterham, 2005), was significantly enriched within the peroxisome (*Z*-value=17, $P<1.00\times10^{-10}$).

Our DPL matrix revealed that 778 diseases ($\sim$62%, P=1.40×10^{-3}) are enriched in a single localization and 273 diseases ($\sim$21%, P=3.45×10^{-3}) are enriched in dual localizations. In the DPL matrix, certain disease-associated proteins are likely to be found in membrane-bounded organelles such as mitochondria, lysosome, and peroxisome, indicating that the mutations of proteins localized to these compartments are connected to the pathophysiological conditions of those organelles. For example, HMG-CoA synthase deficiency caused by the shortage of mitochondrial 3-hydroxy-3-methly-glutaryl-CoA synthase is enriched in mitochondria, whereas

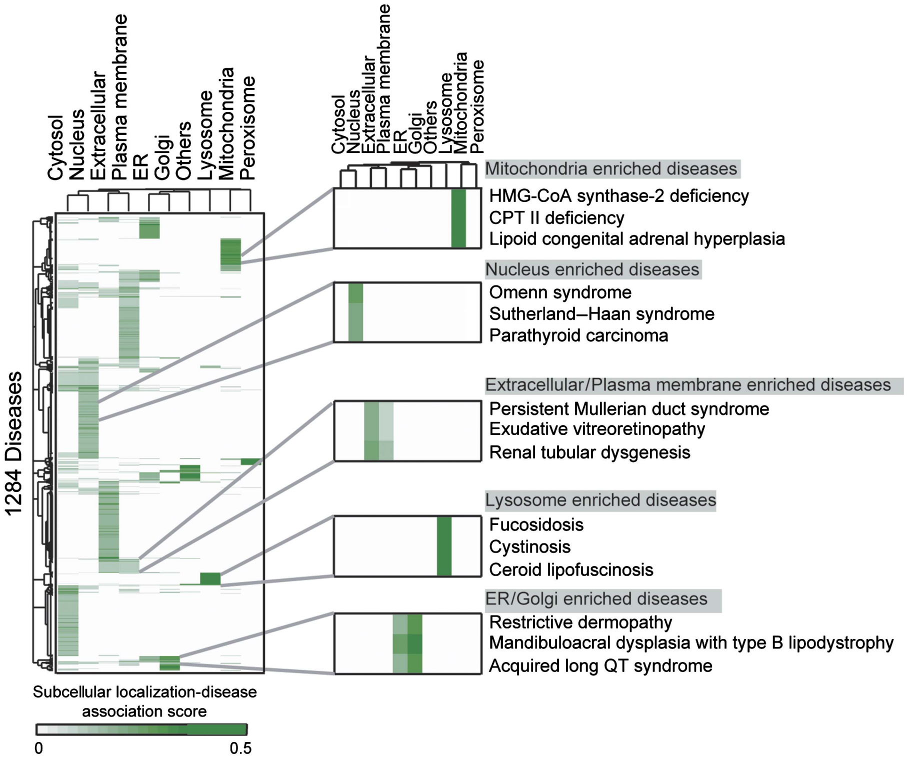


Figure 1 Hierarchical clustering demonstrating the intimate relationships between disease-associated proteins and their subcellular localizations. A two-dimensional hierarchical clustering was performed to organize and visualize the matrix of 10 different subcellular localizations and 1284 diseases. Enlarged portions represent clusters of highly enriched diseases in certain subcellular localizations (right panel).

genetic disorders belonging to lysosomal diseases caused by the dysfunction of lysosomal storage enzymes such as GM2-ganglinosidosis and sialidosis are enriched in lysosome (Parenti, 2009). Meanwhile, certain disease-associated proteins in the DPL matrix are enriched in dual localizations, such as extracellular/plasma membrane or ER/Golgi. Although these two pairs of subcellular localizations appear to be distinct compartments at first, they are functionally related compartments in close proximity during protein translocation process in the cell, and thus are likely to share interacting protein partners (Gandhi *et al*, 2006). Disease-associated proteins localizing in cytosol, interestingly, were not highly enriched when compared with other subcellular localizations. It might be related to the dynamic nature of many cytosolic proteins that are known to shuttle across subcellular compartments and interact with proteins in other localizations.

Although calculating the ASs of disease-subcellular localizations turned out to be rigorous (see Materials and methods), we note the existence of potential issues related to the coverage of OMIM database due to the fact that our matrix reflects only the curated disease-gene associations. For instance, diseases with a single associated protein might introduce bias into the enrichment profile in the DPL matrix. To test the validity of the DPL matrix against such bias, we used disease sets having two or more disease-associated proteins and reconstructed the matrix of disease-associated proteins and their subcellular localization (Supplementary Figure 2A). Even without diseases with only one associated proteins, we confirmed that most diseases (~63%, 307 diseases) were preferentially enriched in particular subcellular compartments when compared with random expectation (Supplementary Figure 2B, $P=1.00\times10^{-5}$).

Next, we applied the disease-associated protein complex data to test the variations in disease-protein associations (Lage *et al*, 2007). To reconstruct the DPL matrix in this case, 882 diseases were used along with the disease-associated proteins as the 'seed' from which disease-associated protein complexes were assembled from the physical interactions of disease-associated proteins in the human protein interaction network based on the study of Lage *et al* (2008). This matrix again confirmed that disease enrichments in particular subcellular localizations are strongly correlated in the DPL matrix based on the OMIM data set (Supplementary Figure 3A). To compare the similarity between subcellular localization profiles, we selected an identical disease set from the matrices based on the OMIM data and on the disease-associated protein complex data, and confirmed that there exists a significant correlation (Supplementary Figure 3B, Pearson's correlation coefficient (PCC)=0.78, $P=1.17\times10^{-100}$), indicating the robustness of the properties that the profiles of disease-associated proteins and their subcellular localizations against the variations in disease-protein association data sets.

Phenotypically related diseases have similar subcellular localization enrichment profile

Subcellular localization enrichments of diseases in the DPL matrix show that certain disease types display strikingly similar enrichment patterns across multiple subcellular localizations. Moreover, we found that in many cases phenotypically similar diseases were enriched in specific subcellular localizations. For instance, many diseases in the metabolic disease class including HMG-CoA synthase-2 deficiency and CPT II deficiency are co-enriched in mitochondria. This suggests that phenotypically similar diseases are clustered on the molecular level, and display similar subcellular localization profiles due to the proteins of same molecular pathway likely being located in the same compartment.

We grouped the manually determined classification of 1284 diseases to 22 human disease classes based on the physiological systems affected (Goh *et al*, 2007), and investigated whether phenotypically similar diseases share similar subcellular localization profiles. Here, we built the Disease Class-associated proteins and their subcellular Localization (DCL) matrix similar to the DPL matrix (middle panel in Box 1). Most disease classes (~80%) show statistically significant enrichments in particular subcellular localizations (Figure 2A, $P=1.00\times10^{-5}$). An interesting example is the class of cancers (Figure 2B, $P=1.00\times10^{-12}$)—known to be associated with genes that typically express themselves in a broad range of tissues (Lage *et al*, 2008)—which appear to be significantly enriched inside the nucleus. This tells us that the molecular connections between cancer-associated proteins in the oncogenic activation of transcription factors localized in the nucleus are key in the progression of cancer (Libermann and Zerbini, 2006). Meanwhile, the immunological disease class is significantly enriched in the extracellular region (Figure 2C, $P=1.00\times10^{-20}$) where cell communication and signal transduction take place. Extracellular proteins serve as transducers of extracellular signals into intracellular physiology, having important roles in the modulation of the immune response in disease processes (Lin *et al*, 2008). Connective tissue diseases are also found to be significantly enriched in the extracellular region (Supplementary Figure 4A, $P=1.75\times10^{-11}$); mutations in extracellular matrix proteins are known to cause a wide range of inherited connective tissue diseases (Bateman *et al*, 2009). Osteoarthritis, a common connective tissue disease, for example, is related to the expression of MATN3 located in the cartilage extracellular matrix that contributes to the development of cartilage (Klatt *et al*, 2009). In contrast to the disease classes highly enriched in a specific subcellular localization, several other disease class exhibits enrichment within multiple subcellular localizations in the DCL matrix (Supplementary Figure 4B), the developmental disease class being an example. These diseases are known to be related to diverse pathological changes in various cellular processes and signaling pathways (Tomancak *et al*, 2007; Zhang *et al*, 2010). Indeed, we found that the proteins associated with developmental diseases are located in diverse subcellular compartments such as the nucleus, the plasma membrane, and the extracellular region.

Comorbid disease pairs are connected by subcellular localization on molecular level

Comorbidity represents the co-occurrence of multiple diseases in the same individual (Lee *et al*, 2008; Hidalgo *et al*, 2009; Park *et al*, 2009a). Many comorbid disease pairs have been shown to share common genes in the human disease network.

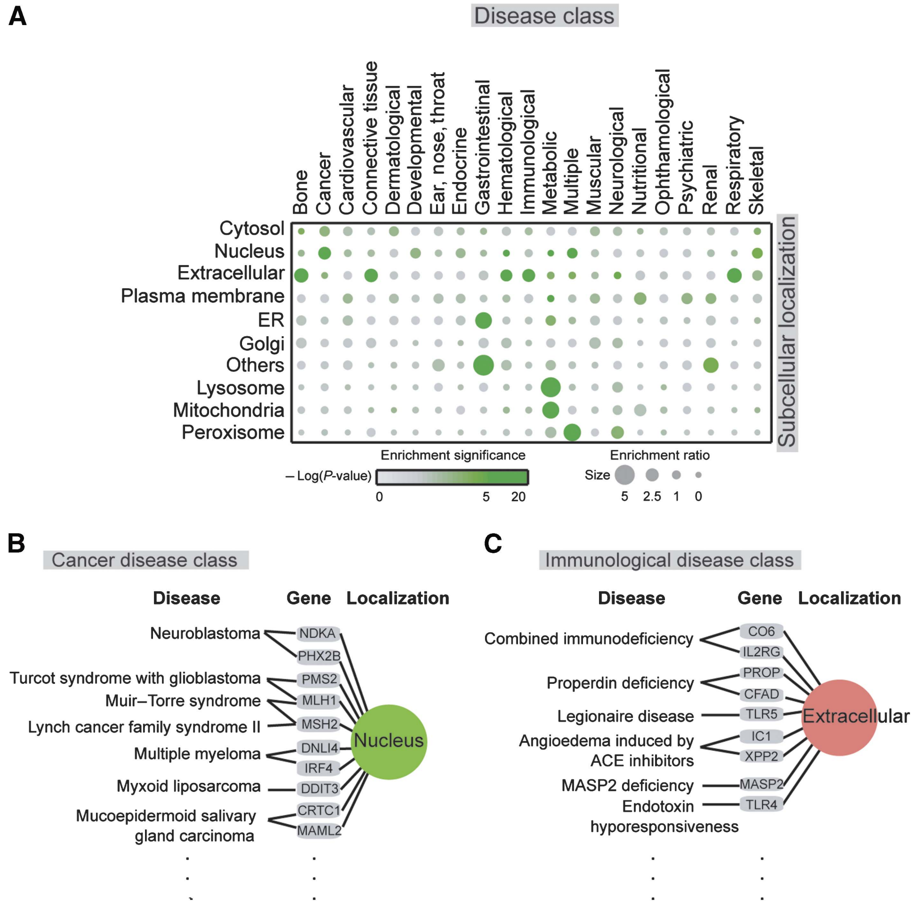


Figure 2 Correlation between disease classes and subcellular localizations. (**A**) The enrichment of disease-associated proteins in specific subcellular localization is evident in various disease classes. The enrichment ratio is proportional to the diameter of the circles: it indicates fold-enrichments calculated as the ratio of the number of observed to the expected disease class-associated proteins in the subcellular localization. Color saturation represents the statistical significance (the *P*-values) of the enrichment ratio. (**B**, **C**) Cancer and immunological disease classes offer examples of disease classes significantly enriched in particular subcellular localizations.

For example, Diabetes and Alzheimer's disease share a risk factor in angiotensin I converting enzyme, and frequently occur together in an individual. In such instances, comorbidity can be partially attributed to the disease connections on the molecular level. Such line of thinking has been applied to identifying the molecular connections of diseases such as shared genes, PPIs, co-expression, and metabolic pathways as potential causes of comorbidity (Rzhetsky *et al*, 2007; Lee *et al*, 2008; Park *et al*, 2009a). To explore the impact of protein subcellular localization on comorbidity, we hypothesized that certain disease pairs could also be connected via subcellular localization by the molecular connections between the disease-associated proteins (bottom panel in Box 1 and Supplementary Figure 5). Multiple myeloma and Glomerulopathy is an example of comorbid disease pairs associated with nuclear proteins, in which subcellular localization is likely to be the contributor of disease co-manifestation, not shared genes or PPIs (Figure 3A).

To explore whether the quantitative correlation between subcellular localizations can explain the comorbidity of disease pairs, we utilized the US Medicare database documenting diagnoses of 13 039 018 elderly patients between the years 1990 and 1993, which has also been successfully used in recent comorbidity studies (Lee *et al*, 2008; Hidalgo *et al*, 2009; Park *et al*, 2009a). Relative risk (*RR*) was used as a quantitative index of the comorbidity tendency, the degree of co-occurrences of disease pairs in patients (see Materials and methods).

We found a positive correlation between subcellular localization similarity and *RR* (Figure 3B, PCC between *RR* and subcellular localization similarity=0.81, $P=2.96\times10^{-5}$). The subcellular localization similarity represents the correlation of subcellular localization profiles between disease pairs. This result appears robust since comorbidity tendency depends neither on the number of disease-associated proteins nor the measurement of comorbidity indices (Supplementary Figure 6). We repeatedly observed positive correlations between *RR* and subcellular localization similarity when we considered only disease pairs with more than two associated proteins or used an alternative comorbidity index, the ϕ-correlation (Lee *et al*, 2008; Hidalgo *et al*, 2009; Park *et al*, 2009a). We discovered that many comorbid disease pairs are indeed connected via subcellular localization. Analbuminemia and Pneumonitis, for example, exhibit a statistically significant comorbidity relationship ($P=1.45\times10^{-12}$) and are both associated with extracellular proteins (Figure 3C).

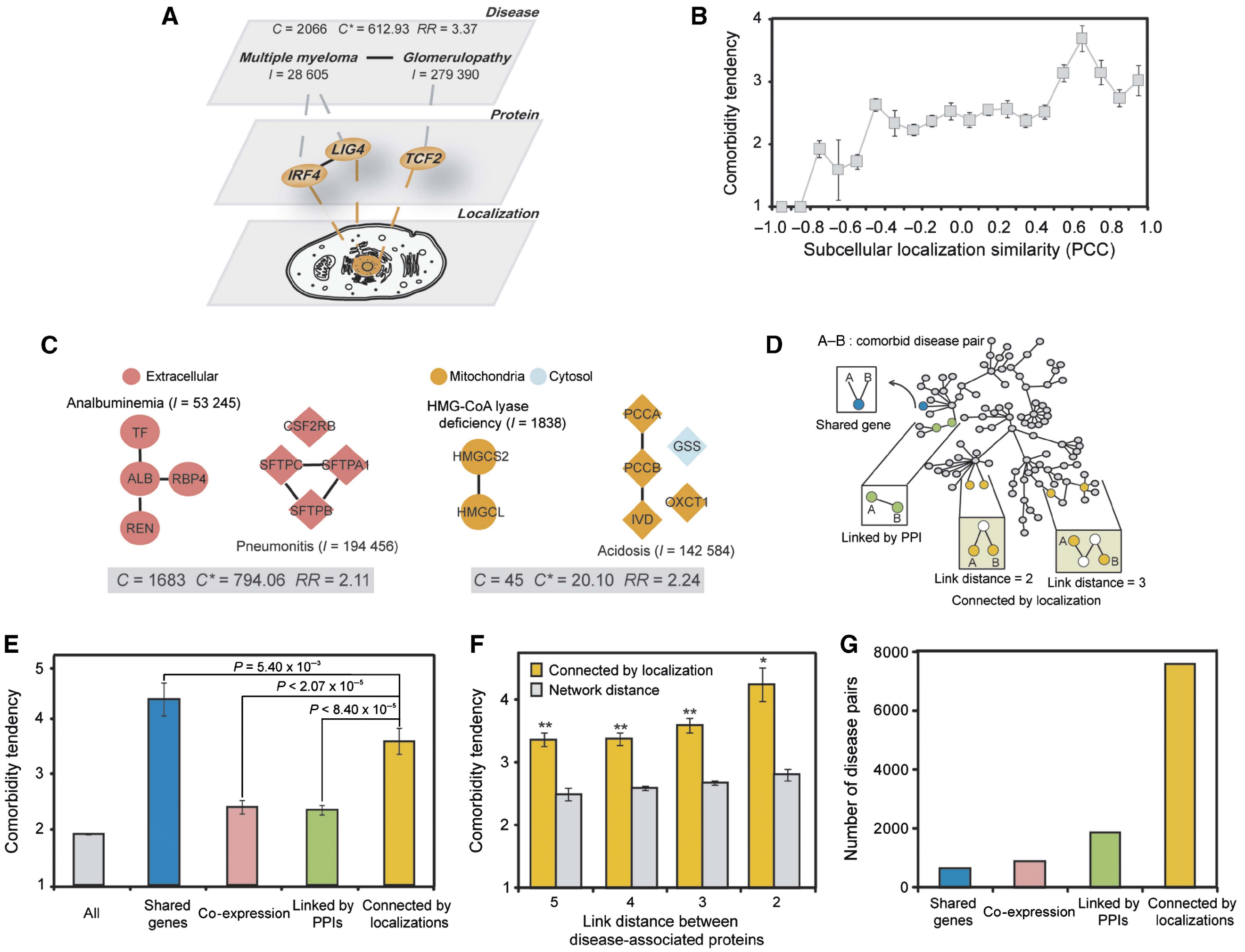


Figure 3 The implication of subcellular localization for disease comorbidity. (**A**) Multiple myeloma and Glomerulopathy is an example of a comorbid disease pair connected via subcellular localization, not via share genes or protein–protein interactions (PPI) (*upper panel*). PPIs are shown as solid lines (*middle panel*). Shared subcellular localization of the disease-associated proteins (nucleus) is highlighted using orange (*bottom panel*). (**B**) Average comorbidity tendencies (*RR*) for disease pairs with increasing subcellular localization similarities. The Pearson's correlation between average comorbidity tendencies and subcellular localization similarities is 0.8. (**C**) Examples of two comorbid disease pairs connected by subcellular localizations. (**D**) Comorbid disease pairs and their molecular connections are overlaid on the depicted on the PPI network. Molecular connections include shared genes, PPIs, and indirect links connected by subcellular localizations. (**E**) Average comorbidity tendencies of disease pairs by using shared genes, co-expression, linked by PPIs, and connected by subcellular localization are compared ($P<5.40\times10^{-3}$; Mann–Whitney test). (**F**) Average comorbidity tendencies were measured for disease pairs connected via subcellular localization and the link distances (*$P<0.4\times10^{-2}$, **$P<0.2\times10^{-2}$; Mann–Whitney test). (**G**) The numbers of disease pairs that share genes, co-expression, linked by PPIs, and connected via subcellular localization.

Analbuminemia is a genetic metabolic defect caused by an impairment in the syntheses of serum albumin (Koot *et al*, 2004) and Pneumonitis is known to be caused by low albumin concentration in the blood (Conde and Lawrence, 2008), suggesting that the similarity of the subcellular localization of associated proteins, in this case extracellular region, gives rise to the observed comorbidity. Similarly, HMG-CoA lyase deficiency and acidosis, both having associated proteins in mitochondria, also show significant comorbidity (P=2.73×10^{-6}) (Figure 3C). It is known that HMG-CoA lyase deficiency affects the metabolic processes of leucine and keratones that lead to the acidic condition of blood (Olpin, 2004).

Phenotypically similar diseases are known to be caused by functionally related modules either in a protein complex or in molecular pathways through direct or indirect protein interactions (Lage *et al*, 2008). From the analysis of human protein interaction network, we discovered that comorbid disease pairs are found to be not only sharing genes or linked by PPIs, but also connected by subcellular localization and indirect interactions in the network (Figure 3D). To our surprise, when we compared the *RR* of disease pairs linked via various molecular connections, we found that disease pairs connected by subcellular localization showed a near three-fold higher comorbidity tendency (with link distances equal to 2 or 3) when compared with random pairs (Figure 3E). Disease pairs that share genes still displayed the highest comorbidity tendency as expected: sharing genes themselves indicates a common genetic origin.

We then assessed quantitatively the impact of network distances and subcellular localizations on the comorbidity tendency of disease pairs. We expected the proteins associated with comorbid disease pairs to be located closely in the protein interaction network via fewer links compared with random disease pairs. Indeed, a higher comorbidity tendency was found when two disease-associated proteins were positioned within a shorter distance (gray plots in Figure 3F). Moreover, when subcellular localization information was combined with small network distances, the comorbidity tendency increased dramatically (orange plots in Figure 3F). It suggests that subcellular localization and close network distances, two conceptually distinct molecular connections, contributed synergistically to the comorbidity tendency. We also observed a similar synergistic effect to the comorbidity tendency when subcellular localization was combined with co-expression (Supplementary Figure 7). Indeed, such a combination also dramatically increased the coverage of disease pairs and allowed the explanation of the molecular connections between 7584 disease pairs (Figure 3G, the full list is provided in Supplementary File 3, http://sbi.postech.ac.kr/dpl). This increased coverage does not come at the expense of comorbidity strength; however, subcellular localization information uncovers a comparable or higher comorbidity tendency than shared genes, co-expression, or PPIs (Figure 3E and G).

Discussion

Here, we presented a systematic strategy to correlate diseases and subcellular localization enrichments of their associated proteins. We expect subcellular localization to be helpful in discovering novel disease-associated genes; when proteins are involved in a common biological pathway or process with disease-associated proteins, it is very plausible that they are themselves disease-associated proteins (Barabasi *et al*, 2011). For example, we present three disease modules representing the clusters of interacting proteins connected by subcellular localizations and sharing disease annotations in Supplementary Figure 8. For instance, a disease module of cerebral degeneration comprises eight mitochondrial proteins among which five are already known to be involved in the same disease. We expect that the other three proteins could be associated with the disease since they are connected by same localization and interact with the same disease-associated proteins.

We found that certain disease classes showed enrichment in particular subcellular localizations, such as connective tissue diseases in the extracellular region. Disease classes are generally related to tissue types because disease classes correspond to the physiological systems affected (Jiang *et al*, 2008), such as the neurological disease class in brain tissue and the immunological disease class in thyroid. Many diseases caused by defects in human genes also have tissue-specific pathology; and thus, tissue types provide another important layer of spatial information on human pathology (Winter *et al*, 2004; Lage *et al*, 2008). While a systematic understanding of the relationship between tissue and subcellular localization is still incomplete, it has been shown that genes highly expressed in a tissue-specific manner are localized in specific subcellular compartments (Kislinger *et al*, 2006). For example, tissue-specific expressions of extracellular matrix proteins are important for their function, and mutations of those proteins are known to cause various connective tissue diseases including Osteogenesis, Chondrodysplasias, and Ehlers–Danlos syndrome (Bateman *et al*, 2009). Therefore, it is evident that the connections between tissue types and subcellular localizations need to be explored further.

In a data-driven research as ours, the robustness of the databases themselves is undoubtedly paramount. Thus, an effort to cross-check and validate one's findings using similar yet distinct databases are clearly necessary, some of which we present and discuss here.

First, we note that a proper scheme of annotating subcellular localization annotation is key for our analysis. It is possible that different subcellular localization information may affect our result, that is, the relevance of the connection via subcellular localization to the comorbidity tendency. We therefore performed a test of mitochondrial localization by using three different subcellular localization annotation sets: the Swiss Prot annotation, ConLoc, and comprehensive localization annotation by using MitoCarta (Supplementary File 4; Pagliarini *et al*, 2008). We observed that, in general, MitoCarta covered more diseases and showed higher correlations (PCC) between subcellular localization similarity and comorbidity tendency (Supplementary Figure 9). Although MitoCarta gave a somewhat higher correlation (PCC=0.86), the present application of ConLoc showed a comparable coverage of diseases and correlation (PCC=0.83), demonstrating the robustness of our original analysis and conclusion. We also observe that MitoCarta improved our knowledge of the molecular connections of comorbidity derived from the most comprehensive and accurate molecular characterization of the mitochondrial proteins. Given that large-scale experiments such as Human Protein Atlas or MitoCarta (Pagliarini *et al*, 2008; Uhlen *et al*, 2010) have improved our ability to identify subcellular localizations in human proteome, we expect the process of uncovering the molecular connections between comorbid diseases to become expedited and more comprehensive.

Second, we combined disease subtypes into single diseases by disease names introduced in Goh *et al* (2007). To verify the effect of combining disease subtypes on the DPL matrix, we calculated the subcellular localization similarity between combined disease and their subtypes. We found that disease subtypes were enriched in the same subcellular localizations on the DPL matrix, as they were in the analysis of single diseases (Supplementary Figure 10). It suggests that disease subtypes tend to share their subcellular localizations as well. For example, Fanconi anemia subtypes are mostly enriched in the nucleus, whereas complement deficiency subtypes are enriched in the extracellular region.

Third, although OMIM stands as a reliable resource for Mendelian disease-gene association, its main focus is monogenic diseases and generally does not consider complex diseases affected by environmental factors. Since both genetic and environmental factors contribute to disease progression, our analysis leaves room for improvement regarding non-Mendelian diseases (Liu *et al*, 2009). We therefore performed an analysis of subcellular localization enrichments in non-Mendelian diseases using the Genetic Association Database (GAD) that covers common complex diseases (Becker *et al*, 2004). We reconstructed the matrix of disease-associated proteins and their subcellular localization of 427 diseases from GAD (Supplementary Figure 11A). From the matrix, we again observed that proteins associated with non-Mendelian diseases from GAD showed subcellular localization enrichments, as was the case for Mendelian diseases. For example, proteins associated with Bipolar Disorder, a complex disease, are enriched in the cytosol, whereas proteins associated with Type-2 Diabetes are enriched mostly in the plasma membrane (Supplementary Figure 11B).

Finally, we again note that the mapping between the OMIM and the ICD-9-CM codes was constructed by human experts for the purpose of merging the genetics data and the population-level comorbidity statistics, used in previous studies (Park *et al*, 2009a). It has been brought to our attention that, as our main analysis was complete, the Unified Medical Language System (UMLS) also aims to become a compendium of biomedical vocabularies including OMIM and ICD-9-CM (Butte and Kohane, 2006), and thus could be used in our context as well, presenting us with an opportunity to cross-validate the mappings as well as our results. Indeed, when we applied the UMLS-based OMIM-to-ICD mapping, we again observed that disease pairs connected by subcellular localizations show higher comorbidity than average over all disease pairs (Supplementary Figure 12). Furthermore, we also observed that comorbidity increases when subcellular localization information is combined with small network distances. There exist some subtle yet understandable disagreements between the two mappings notwithstanding. For instance, in the case of 'Achondroplasia (MIM ID: 100800)' the human experts of the original mapping chose to utilize 733.9 in ICD-9-CM while the UMLS resulted in it being mapped to 756.4 in ICD-9-CM. Most importantly, though, we observe the aforementioned similarity in the trends of our analyses based on the two mappings, and that we believe that they strongly indicate the robustness of our conclusions.

Disease progression is not restricted to the mutation of disease-causing genes, but also affected by molecular connections in '*disease modules*,' resulting in comorbidity (Fraser, 2006; Lee *et al*, 2008). Phenotypically similar diseases are caused by the perturbation of network modules such as shared genes, metabolic pathways, and PPIs. In this study, for the first time we applied subcellular localization information to elucidate the molecular connections between comorbid diseases. Furthermore, we demonstrated that integrating subcellular localization and network distances improved the identification of the molecular connections of disease pairs. We believe that, based on our finding, our approach helps to define the boundaries of '*disease modules*.' Taken together, integration of diverse molecular connections should improve the molecular level understanding of hitherto unexplained comorbid disease pairs and help us in expanding the scope of our knowledge of the mechanism of human disease progression. Finally, we believe that, as more sophisticated, large-scale databases are constructed and come to light, the issues arising from the distinct features or inconsistencies of data will need to be addressed in order to go forward in the growing field of molecular systems research, to which we hope our work have made a valuable contribution.

Materials and methods

Data sets

The OMIM database (http://www.ncbi.nlm.nih.gov/omim/) provides gene-disease associations between 2929 disease types in the Morbid Map (MM) and 1777 disease-associated genes. Some disease types listed in the MM with a minor difference in their names, however, may be similar enough to be clustered as on disease, which was done in the work of Goh *et al* (2007). Disease can be further grouped into 1340 distinct diseases by combining disease subtypes into a single disease, based on their given disease names. For example, the 11 Fanconi anemia subtypes were merged into the disease 'Fanconi anemia' as a single disease ID 523. First, the merge was done by running a string-match script. Then, each entry was verified manually. As a result, 2161 disease terms were grouped into unique 1228 diseases.

We used the hospitalization records from the US Medicare database used in recent comorbidity studies (Lee *et al*, 2008; Hidalgo *et al*, 2009; Park *et al*, 2009a). It contains the Medicare claims of 13 039 018 hospitalized patients during 4 years (from 1990 to 1993) recorded in the ICD-9-CM format (http://www.icd9data.com) where a disease is assigned a numeric code. By using the curated mapping of the ICD-9-CM codes based on the OMIM diseases by using an expert coder and standard coding procedures implemented in hospitals for assigning ICD-9-CM codes to prose description of disease (Lee *et al*, 2008; Park *et al*, 2009a), 83 924 pairs of hereditary diseases were considered in this study.

Subcellular localization mapping for disease-associated proteins

The subcellular localization of disease-associated proteins was first derived from the Swiss Prot annotation information. Subcellular localization information was available for 1168 proteins from the CC (Cellular Component) field of Swiss Prot. For the remaining 609 proteins which do not have subcellular localization annotations, ConLoc and Proteome Analyst were used for the prediction of subcellular localizations (Szafron *et al*, 2004; Park *et al*, 2009b). ConLoc predicts protein subcellular localization based on the optimization of prediction results from 13 localization predictors for 5 major localizations (cytosol, extracellular, mitochondria, nucleus, and plasma membrane) (Park *et al*, 2009b). It achieved the highest prediction accuracy of 0.96 and Matthew's correlation coefficient of 0.86 on the localization prediction of human proteins. ConLoc outperformed all the individual predictors and showed the highest sensitivity on the independent test set of 345 mitochondrial proteins. Moreover, ConLoc achieved the equivalent accuracy on the prediction of multi-localized proteins compared with that of single-localized proteins. Predictions of other subcellular localizations (ER, Golgi, peroxisome, mitochondria, and lysosome) are provided by Proteome Analyst.

DPL matrix

To investigate the correlation between disease-associated proteins and their subcellular localization, we calculated the number of co-assigned disease-associated proteins of a given disease to the subcellular localization. We used Ochiai's coefficient (OC) as a measure of similarity derived from the co-annotations (Lage *et al*, 2008), and calculated an AS as a percentage of the total normalized co-assigning of a given disease-associated proteins in subcellular localizations. When constructing the DCL matrix, the following definitions were used

$$\mathrm{OC}(kD, kL) = \sqrt{\frac{nDL^2}{nD \cdot nL}} \quad \mathrm{AS}(kD, kL) = 100\frac{\mathrm{OC}(kD, kL)}{\sum_i \mathrm{OC}(kD, kL_i)}$$

where *nD* is the total number of disease-associated proteins in a disease and *nL* is the total number of disease-associated proteins in a subcellular localization.

To validate the AS reliability, *Z*-value was calculated from 1000 randomly constructed DPL matrixes.

Comorbidity measure (*RR*)

We used the *RR* as the quantitative measure of comorbidity tendency of two disease pairs (Park *et al*, 2009a) and checked the robustness of our analysis using ϕ-correlation as well. *RR* and ϕ-correlation allow us to quantify the co-occurrence of different diseases compared with random. These are defined as

$$\text{Relative risk }(RR) = \frac{C_{ij}}{C^*_{ij}}$$

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

where N is the total number of Medicare patients; (13 039 018), I_i is the incidence of disease i, C_{ij} is the number of patients who had both diseases i and j, and C^*_{ij} is equal to $I_i I_j / N$, the random expectation. When a disease pair co-occurs more frequently than expected by chance, we have $RR > 1$ and $\phi > 0$ (Hidalgo *et al*, 2009; Park *et al*, 2009a).

Subcellular localization similarity of disease pairs

We analyzed the subcellular localization similarity of disease pairs using subcellular localization profiles in the DPL matrix. Denoting the AS of each disease in each subcellular localization by x_{il} where i is the disease index and l is the subcellular localization index running from 1 to N_l (=10), we calculated the PCC as the subcellular localization similarity measure for each pair of diseases i and j, given as

$$\mathrm{PCC}_{ij} = \frac{N_l \sum_l x_{il} x_{jl} - \sum_l x_{il} \sum_l x_{jl}}{\sqrt{N_t \sum_l x_{il}^2 - \left(\sum_l x_{il}\right)^2} \sqrt{N_t \sum_l x_{jl}^2 - \left(\sum_l x_{jl}\right)^2}}$$

Statistical significance

The *P*-values for the subcellular localization enrichments shown in Figures 1 and 2 and Supplementary Figures 1 and 2 were calculated using the Monte Carlo method (Metropolis and Ulam, 1949). We randomly assigned the subcellular localization annotation to the disease-associated proteins and after 100 000 randomizations, the *P*-values were taken to be the fraction of the total trials that resulted in subcellular localization enrichments larger than observed in data (Park *et al*, 2009a).

Interaction network construction

The human protein interaction network was compiled from eight existing interaction databases: the Biomolecular Interaction Network Database, the Human Protein Reference Database, the Molecular Interaction database, the Database of Interacting Proteins, IntAct, BioGRID, Reactome, and the Protein-Protein Interaction Database. We removed redundant interactions and filtered interactions so that low-confidence interactions were removed, similar to the work of Kenneth D *et al* (Bromberg *et al*, 2008). Specifically, protein interactions were excluded from high-throughput methods, orthologous interactions from lower organisms than human, or predicted by *in silico* methods. The final network comprises 65 135 interactions between 10 652 human proteins.

Co-expression analysis of disease pairs

To analyze the co-expression of disease pairs, we used the Novartis Research Foundation Gene Expression Database (GNF) tissue atlas that includes RNA expression experiments from 79 human tissues (Su *et al*, 2004). We normalized microarray data using MAS5 followed by Bossi and Lehner (2009). Average gene co-expression (ρ_{ij}) was calculated by the average of the co-expression levels between every pair of genes associated with each disease. Denoting the x_{at} as the expression level of gene a on tissue t (t=1, …, 79), the gene co-expression level ρ_{ab} between two genes a and b is defined as the Pearson's correlation between the two (where N_t=79):

$$\rho_{ab} = \frac{N_t \sum_t x_{at} x_{bt} - \sum_t x_{at} \sum_t x_{bt}}{\sqrt{N_t \sum_t x_{at}^2 - \left(\sum_t x_{at}\right)^2} \sqrt{N_t \sum_t x_{bt}^2 - \left(\sum_t x_{bt}\right)^2}}$$

Non-Mendelian diseases and genes association

To analyze non-Mendelian DPL enrichment, we used Gene Association Database (GAD) archive of human genetic association studies (Becker *et al*, 2004). The December 2010 version of GAD was downloaded from http://geneticassociationdb.nih.gov/. We selected only positive genetic associations, and collected 427 diseases and 167 disease-associated genes (Supplementary File 5).

Medicare diseases mapping to the genetic diseases

We used the BioPortal (http://bioportal.bioontology.org/) (Noy *et al*, 2009) to construct OMIM-to-ICD code mapping using the UMLS (Bodenreider, 2004). The ontologies of OMIM and ICD-9-CM were downloaded from the BioPortal, and then the disease terms in OMIM and ICD-9-CM were mapped to the concept unique identified (CUI) in UMLS taking disease synonyms into consideration (Yang *et al*, 2011). Through this procedure, we mapped 488 ICD-9-CM codes to 527 OMIM diseases with 524 CUIs (Supplementary File 6). We considered 250 ICD-9-CM codes to 284 OMIM diseases mapping that contain disease-associated proteins.

Supplementary information

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

Acknowledgements

We thank SBI and MV laboratory members for useful discussions and KOBIC for help in compiling the UMLS database. This work was supported in part by National Research Foundation grants (FPR08B1-220, R31-2009-000-10100-0, 2009-0091503 of the World Class University program, 20100028453, 20100004910, and 20100020528 of the NCRC program) funded by the Korean Ministry of Education, Science, and Technology, grant KHU2010-20100116 from Kyung Hee University, and the Future Internet NAP of Korean Research Council of Fundamental Science and Technology.

Author contributions: SP and SK conceived the project; SP developed the method; SP, JY, YS, JP, SJ, and SK analyzed the results; SP, JP, and SK wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Au CE, Bell AW, Gilchrist A, Hiding J, Nilsson T, Bergeron JJ (2007) Organellar proteomics to create the cell map. *Curr Opin Cell Biol* **19:** 376–385

Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12:** 56–68

Bateman JF, Boot-Handford RP, Lamande SR (2009) Genetic diseases of connective tissues: cellular and extracellular effects of ECM mutations. *Nat Rev Genet* **10:** 173–183

Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. *Nat Genet* **36:** 431–432

Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32:** D267–D270

Bossi A, Lehner B (2009) Tissue specificity and the human protein interaction network. *Mol Syst Biol* **5:** 260

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33**(Suppl)**:** 228–237

Broeckel U, Schork NJ (2004) Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping. *J Physiol* **554:** 40–45

Bromberg KD, Ma'ayan A, Neves SR, Iyengar R (2008) Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. *Science* **320:** 903–909

Butte AJ, Kohane IS (2006) Creation and implications of a phenome-genome network. *Nat Biotechnol* **24:** 55–62

Calvo SE, Mootha VK (2010) The mitochondrial proteome and human disease. *Annu Rev Genomics Hum Genet* **11:** 25–44

Conde M, Lawrence V (2008) Postoperative pulmonary infections. *Clin Evid (Online)* **2008**

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* **104:** 1777–1782

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

Fraser HB (2006) Coevolution, modularity and human disease. *Curr Opin Genet Dev* **16:** 637–644

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS *et al* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* **38:** 285–293

Giallourakis C, Henson C, Reich M, Xie X, Mootha VK (2005) Disease gene discovery through integrative genomics. *Annu Rev Genomics Hum Genet* **6:** 381–406

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. *Proc Natl Acad Sci USA* **104:** 8685–8690

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514–D517

Hidalgo CA, Blumm N, Barabasi AL, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol* **5:** e1000353

Jiang X, Liu B, Jiang J, Zhao H, Fan M, Zhang J, Fan Z, Jiang T (2008) Modularity in the genetic disease-phenotype network. *FEBS Lett* **582:** 2549–2554

Kau TR, Way JC, Silver PA (2004) Nuclear transport and cancer: from mechanism to intervention. *Nat Rev Cancer* **4:** 106–117

Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125:** 173–186

Klatt AR, Klinger G, Paul-Klausch B, Kuhn G, Renno JH, Wagener R, Paulsson M, Schmidt J, Malchau G, Wielckens K (2009) Matrilin-3 activates the expression of osteoarthritis-associated genes in primary human chondrocytes. *FEBS Lett* **583:** 3611–3617

Koot BG, Houwen R, Pot DJ, Nauta J (2004) Congenital analbuminaemia: biochemical and clinical implications. A case report and literature review. *Eur J Pediatr* **163:** 664–670

Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci USA* **105:** 20870–20875

Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25:** 309–316

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313:** 1929–1935

Laurila K, Vihinen M (2009) Prediction of disease-related mutations affecting protein localization. *BMC Genomics* **10:** 122

Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL (2008) The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci USA* **105:** 9880–9885

Li Y, Patra JC (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* **26:** 1219–1224

Libermann TA, Zerbini LF (2006) Targeting transcription factors for cancer gene therapy. *Curr Gene Ther* **6:** 17–33

Lin H, Lee E, Hestir K, Leo C, Huang M, Bosch E, Halenbeck R, Wu G, Zhou A, Behrens D, Hollenbaugh D, Linnemann T, Qin M, Wong J, Chu K, Doberstein SK, Williams LT (2008) Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. *Science* **320:** 807–811

Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol* **10:** R91

Liu YI, Wise PH, Butte AJ (2009) The 'etiome': identification and clustering of human disease etiological factors. *BMC Bioinformatics* **10**(Suppl 2)**:** S14

Luheshi LM, Crowther DC, Dobson CM (2008) Protein misfolding and disease: from the test tube to the organism. *Curr Opin Chem Biol* **12:** 25–31

Metropolis N, Ulam S (1949) The Monte Carlo method. *J Am Stat Assoc* **44:** 335–341

Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* **37:** W170–W173

Olpin SE (2004) Implications of impaired ketogenesis in fatty acid oxidation disorders. *Prostaglandins Leukot Essent Fatty Acids* **70:** 293–308

Oti M, Brunner HG (2007) The modular nature of genetic diseases. *Clin Genet* **71:** 1–11

Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134:** 112–123

Parenti G (2009) Treating lysosomal storage diseases with pharmacological chaperones: from concept to clinics. *EMBO Mol Med* **1:** 268–279

Park J, Lee DS, Christakis NA, Barabasi AL (2009a) The impact of cellular networks on disease comorbidity. *Mol Syst Biol* **5:** 262

Park S, Yang JS, Jang SK, Kim S (2009b) Construction of functional interaction networks through consensus localization predictions of the human proteome. *J Proteome Res* **8:** 3367–3376

Rotig A (2010) Genetic bases of mitochondrial respiratory chain disorders. *Diabetes Metab* **36:** 97–107

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S *et al* (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 complex human phenotypes. *Proc Natl Acad Sci USA* **104:** 11694–11699

Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol* **26:** 1003–1010

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, Toksoz 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

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101:** 6062–6067

Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol* **6:** e1000662

Szafron D, Lu P, Greiner R, Wishart DS, Poulin B, Eisner R, Lu Z, Anvik J, Macdonell C, Fyshe A, Meeuwis D (2004) Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res* **32:** W365–W371

Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* **8:** R145

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F (2010) Towards a knowledge-based human protein Atlas. *Nat Biotechnol* **28:** 1248–1250

Wanders RJ, Waterham HR (2005) Peroxisomal disorders I: biochemistry and genetics of peroxisome biogenesis disorders. *Clin Genet* **67:** 107–133

Winter EE, Goodstadt L, Ponting CP (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* **14:** 54–61

Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. *Mol Syst Biol* **4:** 189

Yang JO, Oh S, Ko G, Park SJ, Kim WY, Lee B, Lee S (2011) VnD: a structure-centric database of disease-related SNPs and drugs. *Nucleic Acids Res* **39:** D939–D944

Zaghloul NA, Katsanis N (2010) Functional modules, mutational load and human genetic disease. *Trends Genet* **26:** 168–176

Zhang SH, Wu C, Li X, Chen X, Jiang W, Gong BS, Li J, Yan YQ (2010) From phenotype to gene: detecting disease-specific gene functional modules via a text-based human disease phenotype network construction. *FEBS Lett* **584:** 3635–3643

Zhernakova A, van Diemen CC, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet* **10:** 43–55