A human phenome-interactome network of protein complexes implicated in genetic disorders

Kasper Lage[1,6], E Olof Karlberg[1,6], Zenia M Størling[1], Páll Í Ólason[1], Anders G Pedersen[1], Olga Rigina[1], Anders M Hinsby[1], Zeynep Tümer[2], Flemming Pociot[3,4], Niels Tommerup[2], Yves Moreau[5] & Søren Brunak[1]

We performed a systematic, large-scale analysis of human protein complexes comprising gene products implicated in many different categories of human disease to create a phenome-interactome network. This was done by integrating quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score, permitting identification of previously unknown complexes likely to be associated with disease. Using a phenomic ranking of protein complexes linked to human disease, we developed a Bayesian predictor that in 298 of 669 linkage intervals correctly ranks the known disease-causing protein as the top candidate, and in 870 intervals with no identified disease-causing gene, provides novel candidates implicated in disorders such as retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease, amyotrophic lateral sclerosis, Alzheimer disease, type 2 diabetes and coronary heart disease. Our publicly available draft of protein complexes associated with pathology comprises 506 complexes, which reveal functional relationships between disease-promoting genes that will inform future experimentation.

Several diseases with overlapping clinical manifestations are caused by mutations in different genes that are part of the same functional module. In such instances, the clinical overlap can be attributed to mutations in single genes rendering the complete module dysfunctional[1]. This concept has been applied to searches for disease genes by several computational methods, including, for example, schemes based on Gene Ontology annotations and gene expression data[2–12]. The advent of proteome-wide interaction screens in model organisms has revealed the modularity of the cellular interactome and that many genes exert their functions as components of protein complexes such as cellular machines, rigid structures, dynamic signaling or metabolic networks and post-translational modification systems[13].

[1]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark. [2]Wilhelm Johannsen Centre for Functional Genome Research, Department of Cellular and Molecular Medicine, The Panum Institute, University of Copenhagen, Blegdamsvej 3, DK-2200, Copenhagen N, Denmark. [3]Institute for Clinical Science, University of Lund, SE – 22100 Lund, Sweden. [4]Steno Diabetes Center, Niels Steensesvej 2, DK-2820 Gentofte, Denmark. [5]Department of Electrical Engineering, Faculty of Engineering, Katholieke Universiteit Leuven, B–3001 Heverlee, Belgium.[6]These authors contributed equally to this work. Correspondence should be addressed to S.B. (brunak@cbs.dtu.dk).

Published online 7 March 2007; doi:10.1038/nbt1295

Analyses involving model organisms, and more recently humans, show that direct and indirect interactions often occur between protein pairs responsible for similar phenotypes[14–22]. In humans this relationship can, for example, be observed in various inherited ataxias[20]. These findings hint at the widespread association of protein complexes with human disease and the likelihood that defects in several proteins, alone or in combination, can cause overlapping clinical manifestations. Systematic investigation of these complexes would help to elucidate cellular mechanisms underlying various disorders and prioritize positional candidates identified, for example, by linkage analysis or association studies.

Our strategy is predicated on the simple assumption that mutations in different members of a protein complex (predicted from protein-protein interaction data) lead to comparable phenotypes, the similarities of which can be automatically recognized by text mining. Computational integration of phenotypic data with a high-confidence interaction network of human proteins is required to perform such an analysis for many human diseases simultaneously. This creates a phenome-interactome network. However, there is no single standard vocabulary for phenotypic annotation in humans. Furthermore, protein interaction data are noisy, are scattered among different databases and contain many false positive interactions[23]. Additionally, only a few large-scale protein interaction studies have been finalized for the human proteome[24,25] rendering the coverage of human protein interaction data too low for a systematic study of protein complexes associated with human disease. Thus, extensive data integration, including conservative incorporation of protein interaction data from model organisms, streamlining of human phenotype data and thorough testing of the resulting method, is required for the systematic investigation of protein complexes associated with human disease.

RESULTS

Construction of a quality-controlled interaction network of human proteins and implementation of a thoroughly benchmarked computational phenotype similarity score allowed us to analyze a human phenome-interactome network. The results show that the 506 disease-associated protein complexes span a wide range of inherited disease categories. We furthermore trained a Bayesian predictor to prioritize candidates in 870 linkage intervals by assigning candidates to protein complexes and ranking these complexes based on the phenotypes associated with its members by text mining. The key steps in our approach are illustrated in **Figure 1**. Four disease-specific case studies are presented to illustrate how the complexes can be exploited to generate novel hypotheses, which directly suggest specific validation experiments involving particular patient-derived materials.

Measuring phenotype similarity scores

Text mining techniques are well suited for investigating phenotype-genotype relationships[8,11,12,14,26–28]. Inspired by such techniques, we created a scoring scheme that quantitatively measures the phenotypic overlap of Online Mendelian Inheritance in Man (OMIM)[29] records (**Supplementary Fig. 1** online). For every record we created a phenotype vector consisting of weighted medical terms present in the record, which represent the phenotype described in that particular record. The parsing of the OMIM records was done using MetaMap Transfer[30] (MMTx), a program that maps text to the Unified Medical Language System (UMLS)[31] metathesaurus (MTH) concepts. The pairwise phenotypic overlap between records was quantified by calculating the cosine of the angle between normalized vector pairs[32], which is a standard measure in such analyses. Essentially, the method amounts to detecting words (from the UMLS vocabulary) that are (i) common to the description of the two phenotypes and (ii) do not occur too frequently among all phenotype descriptions and thus are informative about the phenotype under consideration.

Even though our approach is comparable to successful methods reported in other contexts[28], there are a number of problems surrounding the use of MMTx and UMLS[33], and it is not obvious that the cosine distance between phenotype vectors can accurately capture and quantify the phenotypic overlap between record pairs. To evaluate the reliability of our method, we extracted a large set of ~7,000 OMIM record pairs, which had a high degree of phenotypic overlap. This assertion of phenotypic overlap was based on a combination of the opinion of expert OMIM curators and experts familiar with the diseases under consideration (**Supplementary Methods** online). To evaluate the phenotypic overlap of record pairs in this set, we manually curated 100 random record pairs. This evaluation showed that over 90% of the pairs consist of records with a high degree of phenotypic overlap (**Supplementary Table 1** online).

The reliability of the phenotype similarity score was then tested by fitting a calibration curve of the score against the overlap with the OMIM record pairs (that is, the percentage of the pairs with a given score found among the record pairs). This demonstrates their direct correlation (**Supplementary Fig. 2** online). The higher the phenotype similarity score between records measured by our text-mining scheme, the higher the probability that the records had been independently evaluated to have a phenotypic overlap by the OMIM curators, so that indeed the constructed phenotype vectors and scoring scheme produce a reliable measure of phenotypic overlap between OMIM records.

Constructing a scored network of human protein interactions

We created a human protein interaction network by pooling human interaction data from several of the largest databases and increased the coverage by transferring data from model organisms. We then devised and tested a network-wide confidence score for all interactions. This score relies on network topology and furthermore considers (i) that interactions from large-scale experiments generally contain more false positives than interactions from small-scale experiments[23], and (ii) that interactions are more reliable if they have been reproduced in more than one independent interaction experiment[23]. The reliability of this score as a measure of interaction confidence was confirmed by fitting a calibration curve of the score against overlap with a high-confidence set of about 35,000 human interactions (**Supplementary Fig. 3** online). The resulting network contains ~343,000 unique interactions between ~8,500 human proteins. Of these, ~62,000 are high-confidence interactions.

Testing the predictor on 1,404 linkage intervals

We trained a Bayesian predictor to rank known disease-causing proteins in linkage intervals, by assigning candidates to protein complexes and ranking these complexes based on the phenotypes assigned to their members by text mining. The predictor was validated by fivefold cross-validation on a total of 1,404 linkage intervals containing an average of 109 candidates and including one candidate known to be involved in the particular disease. For ranking candidates, the Bayesian predictor takes as input the patient phenotype (e.g., Leber congenital amaurosis) and a linkage interval, and the candidates are ranked by the following three steps (**Fig. 1**). First, a given positional candidate is queried for high-scoring interaction partners (termed a virtual pull-down of the protein). These interaction partners compose the candidate complex. Second, proteins known to be involved in disease are identified in the candidate complex, and pairwise scores of the phenotypic overlap between diseases of these proteins and the candidate phenotype are assigned. Third, based on the phenotypes represented in the candidate complex, the Bayesian predictor awards a posterior probability score to the candidate in the complex. All candidates in the linkage interval are ranked on the basis of this score. The biological interpretation of a high-scoring candidate is that this protein is likely to be involved in the molecular pathology of the disorder of interest, because it is part of a high-confidence candidate complex in which some proteins are known to be involved in highly similar (or identical) disorders.

Performance of the Bayesian model relying on phenomic scoring of protein complexes associated with disease

The results of prioritizing candidates in the 1,404 test linkage intervals show that the predictor has both good precision and recall (**Fig. 2a**). For each disease, we consider the known disease gene as the relevant gene. Our method makes a prediction for a disease if the top-scoring gene for this disease has a score above the threshold of 0.1. This threshold is chosen because predictions scoring below 0.1 approximate the chance of picking the correct gene randomly. The retrieved gene is then this top-scoring gene. Precision (at a given threshold) is the proportion of relevant genes among all retrieved genes (no. of relevant genes retrieved/no. of genes retrieved). Recall is the fraction of the relevant genes that have been retrieved at the same threshold (no. of relevant genes retrieved/no. of relevant genes). For the 1,404 linkage intervals, there are 669 different predictions with a score above 0.1. Among these, there were 298 correctly identified disease genes, so that the precision at this threshold is 45% (that is, 45% of the candidates that ranked number one with a score above 0.1 are correctly identified as genes causing disease) (**Fig. 2a**)—a level of precision far superior to random prediction. At this threshold, the recall is 21%. A plot of precision versus prediction score cutoff shows proportionality between the score and the chance that the candidate is correct. Candidates scoring above 0.9 are correct in more than 65% of the cases (**Fig. 2a**). Thus, high-scoring candidates are very likely to be correct, and the score awarded to a candidate is a direct indication of the chance that the gene contributes to the disease in question.

There were two main types of failures to identify the relevant genes. Either the proteins coded by the relevant genes do not have an interaction partner that is involved in a relevant phenotype (which applies to 59% of all intervals), or there is a gene in the region considered a better candidate by the predictor (which applies to 26% of all intervals). These 26% could in theory be correct predictions, as suggested by manual inspection of false predictions with high posterior probabilities. By far the most common failure is the lack of interaction partners involved in similar diseases. In 75% of such cases there were no candidates that scored above the threshold of 0.1. These failures could either be due to a lack of data or because some disease proteins do not interact with proteins involved in similar diseases. It seems most likely that the failures are due to a combination of both.

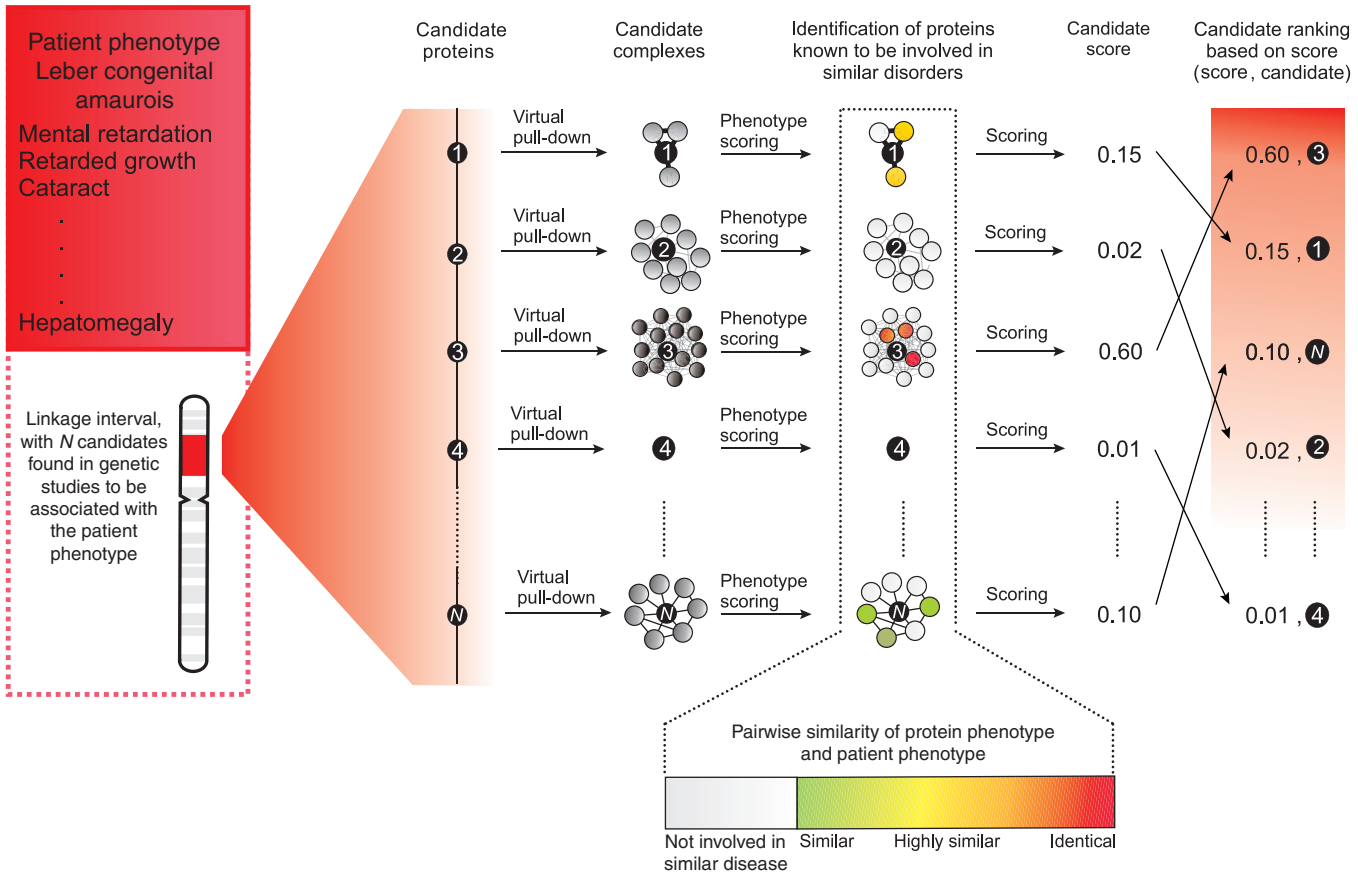

Figure 1 Steps in scoring each candidate in a linkage interval. First, a virtual pull-down of each candidate identifies putative protein complexes including the candidate. Each complex is named the candidate complex. Second, proteins responsible for promoting disease are identified in the candidate complex, and the pairwise similarity to the patient phenotype is measured by text-mining. In this case, proteins that are involved in different disorders comparable to Leber congenital amaurosis are colored according to the clinical overlap with this phenotype. The last step involves scoring and ranking the candidates by the Bayesian predictor. Each candidate is scored based on phenotypes associated with the proteins in the candidate complex, and all candidates in the interval are ranked based on this score.

We also tested a predictor trained on large-scale protein interaction data from which bias related to human diseases was eliminated (**Supplementary Methods** online). Here we observed a comparable precision to the predictor trained on the full protein interaction data set (**Fig. 2b**). Using these data, the precision above 0.1 is 25%, and above 0.9, it is 58%. Therefore, although the performance is slightly lower, it is still very high. These results illustrate the value of large-scale protein interaction data from model organisms, if subjected to stringent quality control. The much lower recall (2.3%) is to be expected with less data. This shows that it is possible to accurately identify disease genes using data from model organisms that were not produced specifically to investigate disease relationships.

Because mutational analysis of candidates in linkage intervals is extremely demanding in terms of resources, our method should be valuable for identifying highly likely candidates and thereby facilitating the discovery of novel genes involved in human disease.

Predicting novel disease gene candidates

OMIM contains 870 intervals linked to diseases for which there are no confirmed disease-causing genes. We ranked the genes in these intervals by the method depicted in **Figure 1**. The full set of predictions above the threshold of 0.1 can be seen in the **Supplementary Data**. We present the best-scoring candidates made by our predictor in **Supplementary Table 2** online. In each of the 91 represented intervals at least one candidate scores above 0.2. In some intervals there are also candidates scoring in the range 0.1–0.2, these are included for completeness, so the table contains a total of 113 candidates in 91 intervals.

All predictions in **Supplementary Table 2** were followed up by independent literature studies, where we investigated the distance of the predicted gene to the closest published high-resolution marker. Seven genes were located >20 Mb from such markers (labeled * in **Supplementary Table 2** online). We also investigated whether the candidates had previously been associated with the respective disorders, and whether there were inconsistencies between candidates we proposed and those proposed by other groups for the same diseases and intervals.

Twenty-four of the predictions point to genes that are most likely true positives, but where the causative mutation has not yet been identified (annotated with "2" or "2#" in **Supplementary Table 2** online). In these cases, our predictions should be seen as further evidence that the genes are involved in the respective diseases. Seven predictions point to genes where a causative mutation has been identified (annotated with "3" in **Supplementary Table 2** online). Together, these constitute 31 predictions most likely to be true. Of these, 25 are the best scoring in the interval, and 6 are scored second or lower. Sixteen predictions point to genes for which literature studies show that a different gene is strongly incriminated in the disease, most likely rendering the prediction wrong (annotated with "1#" in **Supplementary Table 2** online). Of these, 11 are the best-scoring candidate in the interval and 5 score second or lower. When considering

only the best-scoring candidate in each interval (as we have done in the benchmark), 25 are most likely true positives and 11 are most likely negatives. Thus, the precision is 69%—even better than the precision in the benchmark, where predictions above 0.2 have a precision of 49%. Sixty-six of the candidates belong to intervals where there is no evidence in the literature regarding a gene(s) that contributes to the pathology. We consider these as novel candidates. All complexes underlying the candidates scoring 0.1 or above are available for download from the database supporting this work.

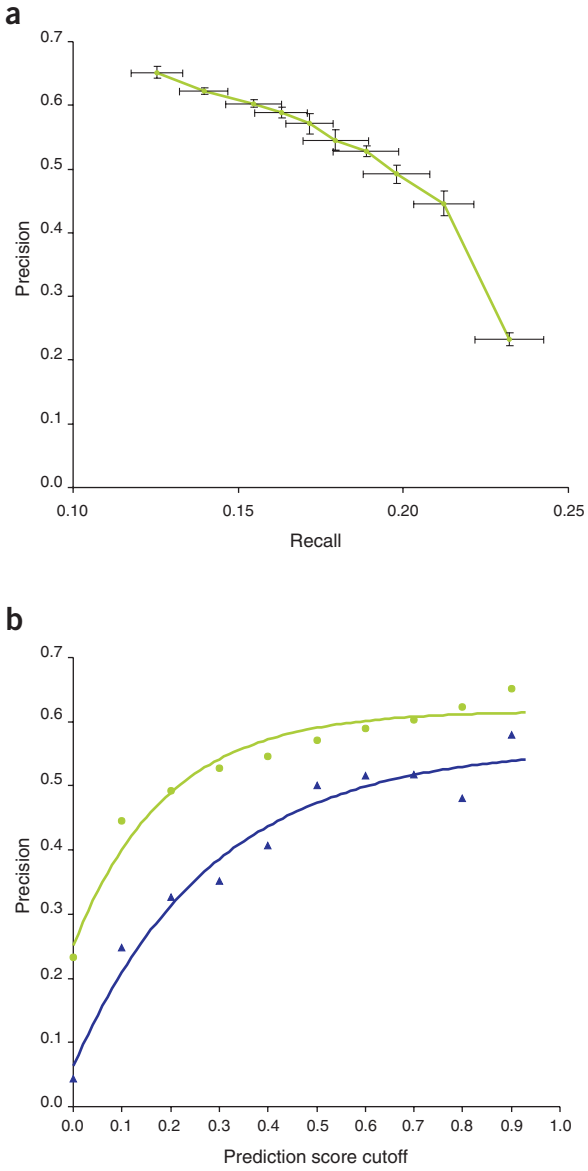

Figure 2 Performance of the Bayesian predictor. (**a**) A plot of recall of the predictor against precision shows that precision for high-scoring candidates can approach 65%. We also trained a predictor only on large-scale data where we had removed all data that were related to diseases that were represented in the test set. (**b**) Prediction score cutoff is plotted for both the predictor trained on all protein interaction data in our network (green line) and the predictor trained only on unbiased large-scale data (blue line). The precision of these two approaches is comparable, showing that it is possible to find disease genes with very high precision, even with unbiased large-scale data inferred from model organisms, if these data are scored correctly.

To exemplify the candidate protein complexes underlying the scoring of the Bayesian predictor, we present four case studies of the novel candidates from **Supplementary Table 2** online. Similar analysis can be carried out for all 506 complexes in the data set, pointing to specific approaches toward validation of the proposed relationships.

Case studies

Retinitis pigmentosa is a clinically and genetically heterogeneous group of disorders. Common traits are night blindness, constricted visual field and retinal dystrophy. In an associated interval on 2p15–p11 (ref. 34), the Bayesian predictor points to LOC130951 with a score of 0.5232. This protein is uncharacterized but evolutionarily conserved, and it is putatively involved in the disease based on an interaction with CRX[25,35] (**Fig. 3a**). CRX is a homeobox transcription factor known to be involved in retinitis pigmentosa and cone rod dystrophy[36]. The candidature of LOC130951 is not obvious, and because both interaction studies reporting the interaction to CRX are large scale, including thousands of interactions, it seems unlikely that LOC130951 would have been chosen as a suitable candidate by manual investigation of the interval.

Epithelial ovarian cancer arises as a result of genetic alterations in the ovarian surface epithelium. In an associated interval on 3p25–p22 (ref. 37), the Bayesian predictor points to Fanconi anemia group D2 protein (FANCD2) with a score of 0.9981. This protein is placed in a complex with breast cancer type 2 susceptibility protein (BRCA2), breast cancer type 1 susceptibility protein (BRCA1) and nibrin isoform 1 (NBN), all of which are involved in ovarian cancer, breast cancer or chromosomal instability disorders[38–41] (**Fig. 3b**). Furthermore, other proteins involved in cancer can be identified in the complex (**Supplementary Data** and **Supplementary Fig. 4** online). FANCD2 is part of the BRCA pathway in cisplatin-sensitive cells[42] and is known to be involved in different types of cancer[43]. However, to our knowledge, a mutation in this gene has never been demonstrated in epithelial ovarian cancer, and we consider it to be a likely candidate in epithelial ovarian cancer in families with linkage to 3p22–p25.

Inflammatory bowel disease is characterized by chronic, relapsing intestinal inflammation. In an associated interval on 6p[44,45], the Bayesian predictor points to receptor-interacting serine/threonine protein kinase (RIPK1) as the most likely candidate with a score of 0.9984 (**Fig. 3c**). The candidate complex includes the signaling proteins tumor necrosis factor receptor 2 (TNFRSF1B), tumor necrosis factor precursor (TNF) and tumor necrosis factor receptor precursor (TNFRSF1A), all known to be associated with inflammatory bowel disease or other inflammatory disorders. Furthermore, other proteins involved in inflammation and immune responses can be observed in the complex (**Supplementary Data** and **Supplementary Fig. 5** online). We thus identified a positional candidate, which is placed centrally in a complex of proteins known to be involved in inflammatory bowel disease and other types of inflammation. We note that RIPK1 lies 20.6 Mb from the closest high-resolution marker published. However, considering that all of 6q was screened for candidates, and that several genes lying far from the published markers are most likely true predictions in **Supplementary Table 2** online, we believe that RIPK1 is a very likely candidate involved in inflammatory bowel disease.

Amyotrophic lateral sclerosis (ALS) with frontotemporal dementia is a degenerative motor neuron disorder characterized by muscular atrophy, progressive motor neuron function loss and bulbar paralysis. In many families, hereditary ALS is associated with frontotemporal dementia and linkage has been shown to an area on 9q21–q22 (ref. 46). Here, the Bayesian predictor points to two likely candidates: bicaudal D homolog 2 (BICD2) and cytoplasmic isoleucyl-tRNA synthetase (IARS), scoring 0.4351 and 0.2154, respectively. Although BICD2 is scored highest,

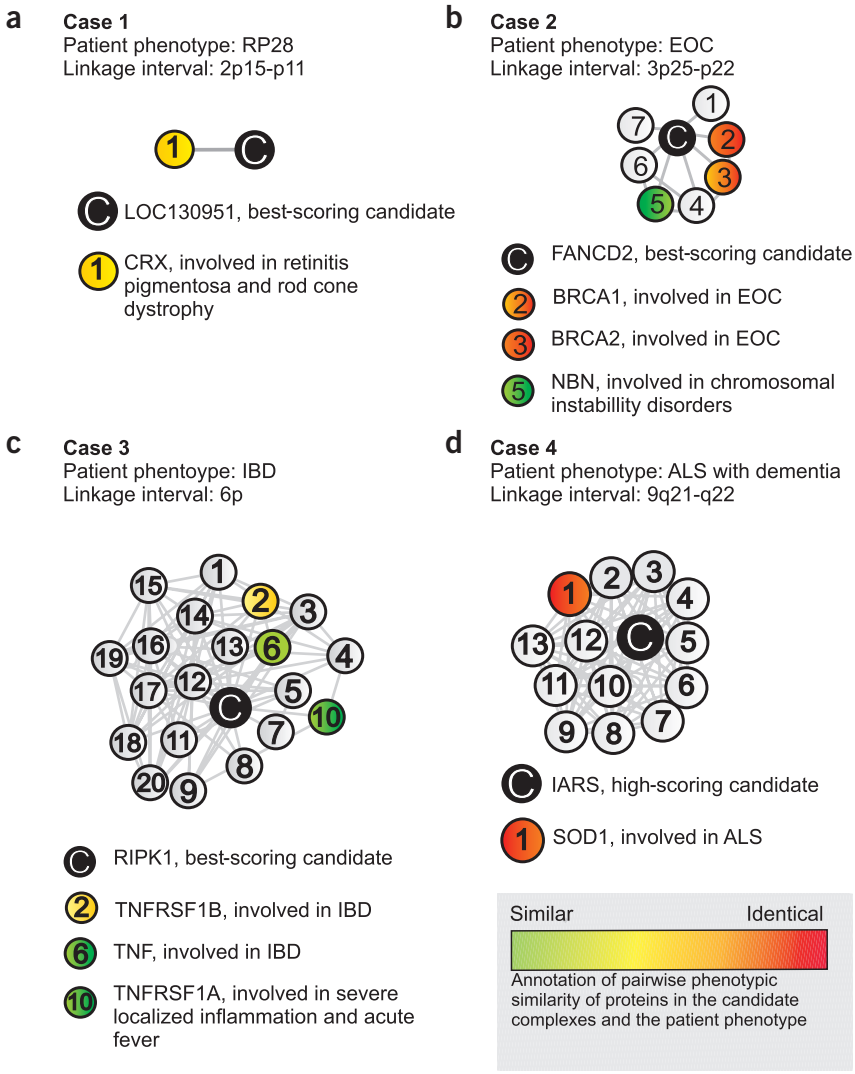

Figure 3 Case studies of four candidate complexes. (**a**–**d**) These candidate complexes are subjected to virtual pull-down with the best-scoring candidate in retinitis pigmentosa 28 (RP28) (**a**), epithelial ovarian cancer (EOC) (**b**), inflammatory bowel disease (IBD) (**c**) and a high-scoring candidate in amyotrophic lateral sclerosis (ALS) with frontotemporal dementia (**d**). Solid black circles (**c**) represent proteins that are the high-scoring candidates in the four disorders. Numbered circles are proteins that interact with the candidate proteins. Colored nodes are proteins identified by our phenotype-similarity scheme. Gray proteins are not predicted by our phenotype-similarity scheme to be implicated in any disease.

both candidates are awarded good scores and are plausible candidates for contributing to ALS associated with dementia. However, investigation of the candidate complexes suggests that BICD2 is more likely to be involved in nonfamilial ALS not associated with dementia, because it is part of a complex with dynactin, which is associated with ALS without dementia. IARS is in a complex with superoxide dismutase 1, a protein known to be involved in familial ALS[47] including dementia (**Fig. 3d**). Also, the IARS complex contains molecular chaperones and other proteins that have been connected to the disease and other types of dementias (**Supplementary Fig. 6** online), and the interaction data underlying the complex is highly reproducible (**Supplementary Data** online). Both candidates are likely, but the candidate complex underlying IARS is seemingly more relevant to familial ALS, and it is plausible that IARS could be involved in the disease in families with linkage to 9q21–q22. Because little is known about this disorder, the complex revealed here is an interesting new lead concerning its underlying causes.

These case studies indicate the value of data mining our phenome-interactome network and integrating interaction data across multiple organisms for positional candidate prioritization. In the case of retinitis pigmentosa and ALS with frontotemporal dementia, the predictor identifies nonobvious candidates in novel putative complexes supported by a network of reproducible interaction data from humans and multiple model organisms. In the cases of inflammatory bowel disease and epithelial ovarian cancer, we identify partly characterized complexes, where several members are known to be involved in the patient phenotype. However, because there are ~500 positional candidates in the case of inflammatory bowel disease, it would require extensive literature studies to reveal this network and candidate by manual data integration. We thus believe that RIPK1 would probably not have been identified as a good candidate despite prior knowledge of its involvement in a known network contributing to inflammatory responses.

DISCUSSION

We have recently witnessed the emergence of integrative methods for identifying probable disease genes in linkage intervals associated with disease based on data integration involving, for example, Gene Ontology categories and expression data[2–12]. Traditionally these methods are compared by measuring average fold enrichment of positional probability (**Supplementary Methods** online). If a method ranks the true candidate in the top 10% of all candidates in 50% of the linkage intervals, there is a tenfold enrichment in the successful predictions intervals and fivefold enrichment on average. We show that our method increases the probability 108.8 times for the successful predictions and 23.1 times on average, significantly outperforming the other computational methods for positional candidate prioritization, which report 5.6–31.2 times enrichment in the successful linkage intervals to 3.8 to 19.4 times enrichment on average (**Supplementary Table 3** online). The most common failure of our method to correctly identify the disease gene results from the inability to find interaction partners associated with a similar phenotype as the relevant protein. This could result from either a lack of data or the failure of these proteins to interact with proteins involved in similar phenotypes. In 75% of these cases, failure to identify another candidate scoring over 0.1 eliminates the possibility of an incorrect prediction.

Our ability to assign candidates to high-confidence protein complexes and rank these complexes in terms of phenomics has permitted us to present a first draft of 506 protein complexes associated with human disease. The success of our method can be attributed to a combination of factors. First, we integrate experimental protein interaction data with a phenotype similarity scheme, thereby taking advantage of the complete clinical spectrum of related human diseases. Also, we use

high-confidence protein complexes for identifying novel candidates, thus ensuring that we take advantage of the full protein network context of the candidate, which we show is well suited for functional association of proteins with diseases. Only three of the previously published methods use protein interaction data[3,21,22]. Whereas one[21] relies completely on unscored binary interaction pairs to identify candidates in identical diseases, others[3,22] incorporate unscored human protein interaction data as one of the weaker sources of information. The two latter methods do not take advantage of cross-species integration of interaction data and none of the three integrate phenotypic descriptions as we have done. Furthermore, two approaches[3,21] search only for candidates implicated in identical diseases and do not take advantage of information from different diseases with a phenotypic overlap. Another method[22] relies on provision of a training set and could theoretically be trained using proteins involved in nonidentical but overlapping phenotypes. These methods report 10.0–15.4 times enrichment in the successful linkage intervals and 5.0–10.0 times enrichment on average (**Supplementary Table 3** online). All three methods are innovative and of high quality, but the difference in performance can readily be explained by recalling that the use of high-confidence protein complexes and data about overlapping phenotypes is much better at inferring functional associations than the search for unscored single-interaction partners involved in identical phenotypes only. The complexes generated in the training and validation of the method provide a valuable resource for further investigations by researchers investigating these diseases, because the complexes place the disease-causing proteins in a functional context relative to other disease-associated proteins. We have created a database of these two data sets (available from http://www.cbs.dtu.dk/suppl/dgf/) providing a draft of 506 putative human disease complexes, determined by the current resolution of data. Our validation shows that the score associated with each complex can be used as a reliable indication of the quality of the data underlying the complex.

METHODS

Design choices of the Bayesian predictor. We have strived to make optimal design choices to guarantee the quality of the methodology. First, for the phenotype similarity score, we opted for the UMLS vocabulary, because it is a well-known resource for this type of analysis, and MMTx for the term mapping. There are some limitations when using MMTx and UMLS (see **Supplementary Methods** online), but we concluded that these are well suited for our analysis, and improvement of these resources is beyond the scope of this work. Second, we chose term frequency–inverse document frequency (tf-idf) as the term-weighting strategy. Compared with unweighted vectors and idf term weighting, tf-idf performed better (**Supplementary Fig. 7** online). Third, we used the cosine similarity measure between phenotype vectors, because it is a well-accepted similarity measure for weighted-term vectors. We demonstrate the robustness of this measure on phenotype vectors constructed from a different text source, weighting method and vocabulary (**Supplementary Fig. 8** online). Finally, for reporting likely candidates, a threshold of 0.1 on the Bayesian score was chosen on the basis of our benchmark. Using these design choices we created a Bayesian model that was trained and validated using fivefold cross-validation. Additionally, the model was thoroughly optimized to get the optimal separation of signal to noise from the phenotype similarity scheme, the protein interaction data and the other parameters in the model. This was done using a genetic algorithm (**Supplementary Methods** online).

Filtering irrelevant semantic types from UMLS. The UMLS vocabulary was manually checked for semantic types that were obviously not clinically relevant (for example, STY|T066|Machine Activity, STY|T068|Human-caused Phenomenon or Process, STY|T093|Health Care Related Organization, STY|T097|Professional or Occupational Group). Terms belonging to these semantic types were filtered out and do not appear in the phenotype vectors. This procedure helps in limiting the phenotype vectors to relevant medical terms to as large an extent as possible.

Phenotype similarity scores. Both the text and clinical synopsis parts of each OMIM record were parsed with MMTx (http://mmtx.nlm.nih.gov/) (for a discussion on the recall, precision and well documented problems of MMTx see **Supplementary Methods** online) to find the occurrence of medical terms in a subset of the UMLS vocabulary[31], where a number of obviously nonclinical semantic type categories had been removed. Phenotype vectors for each record were constructed so that the value of each dimension in the vector represents the number of occurrences of that term in that particular record. Because many relevant terms (for example, mental retardation) are very frequent in OMIM, we also assigned a weight to every extracted term in a phenotype vector. This was done by comparing the frequency with which the term was used in the record in question to its mention the term in all records (that is, all of OMIM). This weight is called tf-idf[48] (**Supplementary Methods** online) and markedly improves the predictive quality of the data (**Supplementary Methods** online). Furthermore, this procedure normalizes the term weight using the length of the specific record and the total length of all records. This normalization reduces negative bias in relation to short records, and positive bias in relation to long records. Once vectors for all records had been constructed, pairwise similarity was calculated as the cosine of the angle between the OMIM vectors after normalization[32]. We used the cosine measure as a natural similarity score for two vectors, because it is a standard measure used in this type of text-mining analysis and it is fast to calculate. We note a small bias against some of the phenotype vectors used to predict because of less well curated and described phenotype records in the prediction set than in the benchmarking set (**Supplementary Table 4** online). We believe this bias is largely caused by less extensive annotation by the OMIM curators of records describing loci where the disease gene has not been identified. The result is fewer predictions than expected from the benchmark. However, it is important to note that the predictions we do get are of equal quality to the benchmarking case, because the posterior probability score relies on the quality of the data used for the prediction.

Validating the phenotype similarity score. To investigate to what extent our phenotype vector cosine scores could correctly assign phenotype similarity between scored records, we fitted a curve of the score against the overlap in OMIM record pairs that had a high degree of phenotypic overlap (**Supplementary Methods** online). The curve shows that the computational phenotype similarity score is directly correlated to the probability of overlap with these record pairs (**Supplementary Fig. 2** online)

Constructing a scored human protein interaction network. Protein interaction data were downloaded from MINT[49], BIND[50], IntAct[51], KEGG annotated protein-protein interactions (PPrel), KEGG Enzymes involved in neighboring steps (ECrel)[52] and Reactome proteins involved in the same complex, indirect complex, reaction or neighboring reaction[53]. All human data were pooled, and to increase the coverage of interactions, interolog data (the transfer of protein interactions between orthologous protein pairs in different organisms)[54] were included by a method similar to that reported by Lehner and Fraser[55]. Interactions were transferred from 17 eukaryotic organisms and added to the network. Orthology was assigned using the Inparanoid database[56] with strict thresholds. To obtain a global interaction score for all interactions in the network, we constructed a probabilistic protein interaction score that took into account the topology of the interaction network surrounding the interaction, the experimental setup (large-scale vs. small-scale) and the number of different publications in which the interaction had been detected (**Supplementary Methods** online).

Making a virtual pull-down. A virtual pull-down of a given protein was done by querying the interaction network for all interactions of the protein (and subsequently all interactions between the interacting proteins) and only retaining the interactions over a given score threshold as defined by the genetic algorithm in the training steps of the Bayesian predictor. This means that the resulting interactions all are of high confidence and supported by network topology, different publications, reliable small-scale interaction experiments, reproducibility or a combination of these.

Identifying proteins involved in diseases in the candidate complexes. Ensembl Mart (http://dec2005.archive.ensembl.org/Multi/martview) was used to associate proteins to phenotypes and identify proteins involved in disease in the candidate complexes.

Making the benchmarking cases. A list of 3,256 disease genes was initially downloaded from the Disease Gene table in GeneCards (http://nciarray.nci.nih.gov/cards/). GeneCards mines several different databases, including OMIM, for text describing the disease genes in this table. For some of the disease genes the entries in GeneCards are sentences, originating from OMIM, specifically stating that defects in particular genes lead to particular diseases. To exclude genes associated to diseases by circumstantial evidence, and only include genes in which genetic defects were known to be causative in relation to the particular disorders, we included genes in the benchmarking set only if GeneCards had found such sentences in OMIM in relation to the gene. Because OMIM is a database manually curated by disease experts, we consider such statements from OMIM to be trustworthy. However, to double-check that no mistakes were made by GeneCards in the extraction procedure, or in the curation process by OMIM, we randomly selected 50 of these statements and manually checked (i) that such statements were actually present in the relevant OMIM files and (ii) that the statements were supported by cited literature. In these 50 cases no discrepancies were found, and this investigation led us to consider that all of the statements are correct. This procedure led to a subset of 963 genes and their corresponding proteins. These genes and proteins were associated with their respective phenotypes using GeneCards references to OMIM diseases. This showed that the 963 genes are involved in 1,404 distinct phenotypes, which were used for the training and validation of the Bayesian predictor. Benchmarking cases were made by associating the genes to distinct phenotypes using the annotation in GeneCards and by assigning the genes to artificial linkage intervals. This was done by including a random number of genes upstream and downstream of the known disease gene. The interval sizes were randomized so that they have a distribution similar to the intervals in OMIM morbidmap, for which no gene has been identified, leading to an average of 108.8 genes in each of the 1,404 linkage intervals.

Training and validating the Bayesian model. Training and benchmarking of the Bayesian model were done by fivefold cross-validation on the benchmarking set. The set of 1,404 benchmarking cases was split into five sets and the Bayesian model trained and optimized on four of these fractions (**Supplementary Methods** online). Subsequently, the optimized model was used to rank candidates in benchmarking cases made on the last fifth of the data set. This was done for all combinations of the five fractions. The benchmarking results can be seen in **Supplementary Table 5** online.

Bayesian disease gene predictor. The goal is to compute, for each candidate in a critical interval, the probability that this is the disease-related protein. High probabilities should be assigned to candidates that interact with one or more proteins involved in disorders that are phenotypically similar to the one being investigated. This logic is expressed in the form of a probabilistic model, and we use Bayes' theorem to compute the probabilities. The model includes parameters for (i) the probability that a candidate protein has any reported interaction partners, (ii) protein interaction score, (iii) the number of interaction partners that are involved in similar disorders and (iv) computational phenotype similarity score. All parameters are estimated from our data sets for both disease- and non-disease-associated genes, where we see that the parameter values are different in the two cases. The probability that protein number i (among N candidates) is the disease-associated one, is computed as follows:

$$P(\text{dis} = i \mid \text{DATA}) = \frac{P(\text{DATA} \mid \text{dis} = i) \times P(\text{dis} = i)}{\sum_{j=1}^{N} P(\text{DATA} \mid \text{dis} = j) \times P(\text{dis} = j)}$$

where $P(dis = i|\text{DATA})$ is the posterior probability that candidate number i is the disease-related protein after evaluating all the data. $P(\text{dis} = i)$ is the prior probability that candidate number i is the disease-causing protein, before evaluating any data. The prior value was set to $1/N$ for all candidates. The term $P(\text{DATA}|\text{dis} = i)$ is the probability of obtaining the observed data if candidate number i was in fact the correct one. This likelihood is computed from the interaction data and any associated phenotype descriptions, and using the estimated parameters, in a straightforward manner (**Supplementary Methods** online).

Case studies. Case studies were made by downloading complex data available for all putative disease complexes (http://www.cbs.dtu.dk/suppl/dgf/) and creating an interactive graph in the free software cytoscape (http://www.cytoscape.org/). Data in these files combined with literature studies were used to generate the hypotheses. More data on the case studies can be found in **Supplementary Data** online. Proteins are named by using the corresponding gene name according to HUGO gene nomenclature (http://www.gene.ucl.ac.uk/nomenclature/).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors wish to thank Ulrik de Lichtenberg and Thomas Skøt Jensen for critical reading of the manuscript, editing and help in developing the protein interaction score. We also thank Christopher Workman and Zoltan Szallasi for valuable discussions and help with the manuscript. Y.M. is supported by K.U. Leuven GOA AMBioRICS, CoE EF/05/007 SymBioSys, BELSPO IUAP P6/25 BioMaGNet, EU-FP6-NoE Biopattern and EU-FP6-MC-EST Bioptrain. Z.M.S. is supported by an EU Biosapiens (NoE), FP6 grant. The Center for Biological Sequence Analysis and the Wilhelm Johannsen Center for Functional Genome Research are supported by the Danish National Research Foundation.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Brunner, H.G. & van Driel, M.A. From syndrome families to functional genomics. *Nat. Rev. Genet.* **5**, 545–551 (2004).
2. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. & Pickard, B.S. Speeding disease gene discovery by sequence-based candidate prioritization. *BMC Bioinformatics* **6**, 55 (2005).
3. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006).
4. Franke, L. *et al.* TEAM: a tool for the integration of expression, and linkage and association maps. *Eur. J. Hum. Genet.* **12**, 633–638 (2004).
5. Turner, F.S., Clutterbuck, D.R. & Semple, C.A. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.* **4**, R75 (2003).
6. Perez-Iratxeta, C., Bork, P. & Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316–319 (2002).
7. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M.A. G2D: a tool for mining genes associated with disease. *BMC Genet.* **6**, 45 (2005).
8. Masseroli, M., Galati, O. & Pinciroli, F. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res.* **33**, W717–W723 (2005).
9. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A. & Brunner, H.G. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. J. Hum. Genet.* **11**, 57–63 (2003).
10. van Driel, M.A. *et al.* GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.* **33**, W758–761 (2005).
11. Hristovski, D., Peterlin, B., Mitchell, J.A. & Humphrey, S.M. Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.* **74**, 289–298 (2005).
12. Freudenberg, J. & Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18** Suppl 2, S110–S115 (2002).
13. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
14. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. & Leunissen, J.A. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **14**, 535–542 (2006).
15. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* (2006).
16. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster. Science* **302**, 1727–1736 (2003).
17. Walhout, A.J. *et al.* Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12**, 1952–1958 (2002).
18. Boulton, S.J. *et al.* Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**, 127–131 (2002).
19. Gandhi, T.K. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **38**, 285–293 (2006).
20. Lim, J. *et al.* A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125**, 801–814 (2006).
21. Oti, M., Snel, B., Huynen, M.A. & Brunner, H.G. Predicting disease genes using protein-protein interactions. *J. Med. Genet.* (2006).
22. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat. Biotechnol.* **24**, 537–544 (2006).
23. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
24. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).

25. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
26. Korbel, J.O. *et al.* Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* **3**, e134 (2005).
27. Schijvenaars, B.J. *et al.* Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics* **6**, 149 (2005).
28. Butte, A.J. & Kohane, I.S. Creation and implications of a phenome-genome network. *Nat. Biotechnol.* **24**, 55–62 (2006).
29. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33** (Database Issue), D514–D517 (2005).
30. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* 17–21 (2001).
31. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
32. Gerard Salton, M.J.M. *Introduction to Modern Information Retrieval* (Neal-Schuman Publishers, New York, 1983).
33. Divita, G., Tse, T. & Roth, L. Failure analysis of MetaMap Transfer (MMTx). *Medinfo* **11**, 763–767 (2004).
34. Gu, S., Kumaramanickavel, G., Srikumari, C.R., Denton, M.J. & Gal, A. Autosomal recessive retinitis pigmentosa locus RP28 maps between D2S1337 and D2S286 on chromosome 2p11–p15 in an Indian family. *J. Med. Genet.* **36**, 705–707 (1999).
35. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
36. Sohocki, M.M. *et al.* A range of clinical phenotypes associated with mutations in *CRX*, a photoreceptor transcription-factor gene. *Am. J. Hum. Genet.* **63**, 1307–1315 (1998).
37. Sekine, M. *et al.* Localization of a novel susceptibility gene for familial ovarian cancer to chromosome 3p22–p25. *Hum. Mol. Genet.* **10**, 1421–1429 (2001).
38. Demuth, I. *et al.* An inducible null mutant murine model of Nijmegen breakage syndrome proves the essential function of NBS1 in chromosomal stability and cell viability. *Hum. Mol. Genet.* **13**, 2385–2397 (2004).
39. Matsuura, S. *et al.* Positional cloning of the gene for Nijmegen breakage syndrome. *Nat. Genet.* **19**, 179–181 (1998).
40. Castilla, L.H. *et al.* Mutations in the *BRCA1* gene in families with early-onset breast and ovarian cancer. *Nat. Genet.* **8**, 387–391 (1994).
41. Lancaster, J.M. *et al.* *BRCA2* mutations in primary breast and ovarian cancers. *Nat. Genet.* **13**, 238–240 (1996).
42. Taniguchi, T. *et al.* Disruption of the Fanconi anemia–BRCA pathway in cisplatin-sensitive ovarian tumors. *Nat. Med.* **9**, 568–574 (2003).
43. Thompson, L.H. Unraveling the Fanconi anemia–DNA repair connection. *Nat. Genet.* **37**, 921–922 (2005).
44. Dechairo, B. *et al.* Replication and extension studies of inflammatory bowel disease susceptibility regions confirm linkage to chromosome 6p (IBD3). *Eur. J. Hum. Genet.* **9**, 627–633 (2001).
45. Hampe, J. *et al.* Linkage of inflammatory bowel disease to human chromosome 6p. *Am. J. Hum. Genet.* **65**, 1647–1655 (1999).
46. Hosler, B.A. *et al.* Linkage of familial amyotrophic lateral sclerosis with frontotemporal dementia to chromosome 9q21–q22. *J.A.M.A.* **284**, 1664–1669 (2000).
47. Koyama, S. *et al.* Alteration of familial ALS-linked mutant SOD1 solubility with disease progression: its modulation by the proteasome and Hsp70. *Biochem. Biophys. Res. Commun.* **343**, 719–730 (2006).
48. Polavarapu, N. *et al.* Investigation into biomedical literature classification using support vector machines. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 366–374 (8-11 August 2005).
49. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
50. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
51. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32** (Database Issue), D452–D455 (2004).
52. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–D357 (2006).
53. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
54. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
55. Lehner, B. & Fraser, A.G. A first-draft human protein-interaction map. *Genome Biol.* **5**, R63 (2004).
56. O'Brien, K.P., Remm, M. & Sonnhammer, E.L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33** (Database Issue), D476–D480 (2005).