Identification of cancer and cell-cycle genes with protein interactions and literature mining

Loic Royer, Conrad Plake and Michael Schroeder
{loic.royer,conrad.plake,michael.schroeder}@biotec.tu-dresden.de
Biotec, TU Dresden, Dresden, Germany

Abstract: Gene prioritization based on background knowledge mined from literature has become an important method for the analysis of results from high-throughput experimental assays such as gene expression microarrays, RNAi screens and genome-wide association studies. We apply our gene mention identifier, which achieved the best result of over 80% in the BioCreative II text-mining challenge [HPR+08], and show how text-mined associations can be complemented using guilt-by-association on high confidence protein interaction networks.

First, we predict hand-curated gene-disease relationships in the OMIM database, Entrez Gene summaries and GeneRIFs with 37% success rate. Second, we confirm 24% of novel cell-cycle genes identified in a recent RNAi screen [KPH+07] by using text-mining and high confidence protein interactions. Moreover, we show how 71% of GOA cell-cycle annotations can be automatically recovered. Third, we devise a method to rank genes based on novelty, increasing interest, impact, and popularity.

1 Introduction

With high-throughput methods such as gene expression analyses, high-throughput RNA interference, and genome-wide association studies, gene prioritization becomes an important problem. Gene prioritization orders lists of genes according to their likelihood to be associated to a process, phenotype, or disease. In particular, genetic linkage analyses identify chromosomal regions, which are linked to a disease and which can contain hundred candidate genes linked to a disease. The problem becomes one of establishing indirect links from the candidate genes to the disease. These links can be of very different nature such as protein interactions [GLF+06, LKS+07], similarity of annotations from controlled vocabularies (phenotype in MeSH [LKS+07]), GeneOntology [AAE+05, PIBA02, TCS03], anatomy [TKP+05], sequence similarity [TCS03, GLF+06, AAE+05, LBO04, PIBA02], phylogeny [LBO04], or co-expression of genes [TKP+05, TCS03, vDCK+03, vDCK+05]. As a result, these approaches manage to significantly reduce the number of candidate genes [TAT+06] or even directly identify the disease gene such as [LKS+07], who predict for 298 out of 669 linkage intervals the correct disease gene.

While the above studies prioritize some hundred genes regarding their link to a disease, other efforts aim to establish large scale links between all genes of a genome and disease. [BK06] mine meta-information of all data sets in the Gene Expression Omnibus by ex-

tracting UMLS concepts from descriptions. This way they can identify novel genes linked to aging. Similarly, [GCV+07] link human protein interaction with expression and disease data. They conclude that disease genes are less likely to be essential interaction hubs and are functionally on the periphery.

In general, protein interaction data can be beneficial to infer indirect relationships by applying the principle of guilt-by-association. Both [GLF+06] and [LKS+07] make use of interactions in their analysis of linkage intervals and [GUT+08] find that half of their correct gene-disease associations are indirectly inferred via protein interactions.

Summarizing the above work, there are three interesting aspects:

- First, only few approaches (e.g. [TKP+05, LKS+07]), apply text-mining to associate genes and diseases and none of them apply large-scale identification of genes and diseases in the whole of the medical literature.
- Second, only few (e.g. [GLF+06, LKS+07, GUT+08]) use protein interactions and the principle of guilt-by-association.
- Third, experimental validation of prediction of novel disease genes is scarce, since such links are inherently difficult to verify due to the complexity of diseases.

In this paper, we address these three points. We show how state-of-the-art entity recognition of cancer genes, cell-cycle terminology, diseases, and their co-occurrences combined with the principle of guilt-by-association can predict hot cancer genes and novel cell-cycle genes. Hot cancer genes i) are novel, ii) have been published in high impact journals, iii) their popularity has not yet peaked, iv) and they attract a large group of researchers. This meta information, which is the result of the comprehensive mining of literature, lends itself to identify genes worthy of exploitation since there is a direct or indirect link and they are truly novel candidates. As argued above, the validation of gene-disease predictions is difficult, since there are no straight forward experiments. We address this problem by validating our approach on novel cell-cycle genes, which have been identified in a genome-wide RNAi screen [KPH+07]. The screen provides a gold standard of over 850 novel cell cycle genes, which have not been discussed in literature before.

The cornerstone of our approach is our entity recognition algorithm, which achieved the best results (81% success rate) in the recent BioCreative text-mining task of human gene name identification [HPR+08]. Since then we have further improved it to 86% success rate. We applied the algorithm, which is online accessible via the BioCreative meta server [LKRP+08], to over 17,000,000 abstracts from PubMed. Additionally, we considered for each abstract any annotated disease terms from the Medical Subject Heading, MeSH, and all stemmed cell-cycle terminology from the Gene Ontology literally appearing in the abstracts. Overall, our method identifies 2.74 million abstracts mentioning a gene, 1.71 million abstracts mentioning a gene and a disease, and 210.000 a gene and cell-cycle term. This resource is now available as GoGene [1] [PRW+09].

With this large data source, we define a simple co-occurrence model and set out to solve the following problems: First, how well can our model predict hand-curated gene-disease

[1] http://gopubmed.org/gogene

relationships in the OMIM database, and in Entrez Gene GeneRIFs. Second, how many of the over 850 novel cell-cycle genes identified in the RNAi screen [KPH+07] can be predicted with our method and which role does text-mining play and which inference through protein interactions? Third, we devise a method to rank genes based on novelty, increasing interest, impact, and popularity. With this ranking, we discuss 50 hot cancer genes and 20 hot cell-cycle genes in detail.

2 Methods

Identification of human genes in PubMed. We identify human gene mentions in PubMed by parsing each abstract with a dictionary of gene names, synonyms, and spelling variants. First we find as many hits as possible. In a second step, a context sensitive filter is applied to remove false positive matches by looking at tokens in the neighborhood of each name/hit and by resolving abbreviations to long forms. Finally, polysemous names, i.e. names referring to more than one gene, are disambiguated by comparing the text at hand against each candidate profile. A profile contains all known information of a gene, e.g. GO annotations, diseases, background texts etc. taken from the high quality databases Entrez Gene and SwissProt. The profile that best fits the text is taken as sense for the ambiguous gene name. For a detailed explanation of our gene identification method see [HPR+08].

Gene ranking. For researchers trying to obtain insights from large screening data, not all genes are equally important. Some genes have often been discussed in the literature. For some genes, the research interest of the community has reached saturation and has since then declined. These declining genes have well-known and stable functions. It is less probable that new insights into their function can be discovered . Yet, they do provide a rich source of information that can shed light on the experiment. In contrast, other genes lie at the forefront of research and have just recently received names and are only found in recent publications. The probability to discover new insights for these genes is higher due to their novelty. Moreover, a novel gene with many recent mentions in high impact journals constitutes an even better candidate. By compiling the publication dates of all human gene mentions in MEDLINE we can decide for each human gene whether research interest has peaked and is dwindling, if the gene belongs to some hot topic of research, or if the gene is discussed in a large body of high-impact literature.

Bibliometric features. We chose four features to measure how 'interesting' a gene is. First the category peaked/not-peaked: $peaked_g$, second the number of publications weighted by impact factor: $volume_g$, third how recent is the interest in the gene independently of the total number of papers: $novelty_g$, and finally the total number of distinct authors that contributed to the publications for that gene: $community_g$. We defined a gene has having peaked if the highest count of papers is at least 3 years old and if since then there was a consistent decrease in the number of papers. We compute novelty using a

simple exponential decrease of the relevance of old mentions divided by the total sum of all impact factor points: $novelty_g = \sum_{y=1950}^{2007} \alpha^{y-2007} c_{g,y}$

In this formula $c_{g,y}$ represents the cumulative impact factor for gene g for the year y. We chose to use a yearly decrease of 50% toward the past ($\alpha = 0.5$). The $community_g$ measure has a minor impact on the ranking as it is strongly correlated to the $volume_g$, but it does contribute additional information about the size of the research community for a gene. To combine these measures we use a Pareto ranking approach[MA94]. The advantage of this scheme is that it ranks genes according to all four features in a balanced manner. For example, shown in Table 4 are the top five genes ranked for breast cancer. Among these is SIRT7, a novel gene with many high-impact publications in recent years, but also BRCA1, a well known and important gene for breast cancer. Among the 36509 human genes, 16078 are found mentioned in MEDLINE and among these 31% have peaked (as of January 2008).

Association of Genes to MeSH and GO terms. To annotate genes with terms from MeSH and GO, we basically count the number of co-occurrences in the literature. For each gene-term pair we compute an association-score as follows: $\text{score}_{g,t} = log_2 \frac{N \times n_{g,t}}{n_g \times n_t}$

where N is the number of articles mentioning any gene and any term from the branch (e.g. a disease), $n_{g,t}$ is the number of articles mentioning the gene and the term, n_g is the number of articles mentioning the gene and any term from the branch, and n_t is the number of articles mentioning the term and any gene. The higher the association score the more likely this pair will be mentioned together in the literature. An association score of zero means that gene and term occur independently of one another. A negative score signals an underrepresentation of a pair in the literature.

Guilt-by-association. Guilt-by-association is the principle by which qualities can be transfered between associated items. In our case, we transfered the information *is cell-cycle related* to all direct interaction partners of a gene if it co-occurs significantly with cell-cycle terms in the literature. We experimented with several decision functions but observed that simply transferring to all direct neighbors of a gene performed best. In particular, adding more distant neighbors increases recall but leads to a significant decrease in precision (data not shown).

3 Results

Prediction of Gene-Disease Associations. We compare text mined gene-disease associations to Entrez Gene Summaries and GeneRIFs texts[2] and to the OMIM database (Tables 3 and 3). The achieved precision rates in Table 3 are underestimations because of the many incomplete GeneRIFs and Summaries (GRS) [BCF+07]. While the precision is low because of many incomplete GRS, the recall suffers from false positives in the benchmark

[2]Downloaded December 2007

Min. Score	Sig. level (%)	Precision (%)	Recall (%)	F1 (%)
12	0.04	**59.2**	1.7	3.3
10	0.22	44.0	6.1	10.6
8	0.97	21.8	11.8	**15.4**
6	3.5	9.3	19.6	12.6
4	10.8	4.7	33.0	8.2
2	28.6	3.0	53.7	5.6
0.6	50.0	2.5	**73.5**	4.8

Table 1: Results for text mined gene-disease associations. Comparison of gene-disease associations for different score thresholds. Predicted diseases are compared to automatically annotated Entrez Gene Summaries and GeneRIFs texts.

Min. Score	Sig. level (%)	Precision (%)	Recall (%)	F1 (%)
13	0.01	**88.5**	6.4	12.0
12	0.04	76.3	10.2	18.1
10	0.22	50.0	29.0	**36.7**
8	0.97	22.4	44.5	29.8
6	3.5	8.1	59.1	14.3
4	10.8	3.1	73.8	6.0
2	28.6	1.4	88.3	2.7
0.6	50.0	0.08	**93.8**	1.6

Table 2: Results for text mined gene-disease associations in OMIM. Comparison of gene-disease associations for different score thresholds. Predicted associations are compared to the OMIM gene-disease catalog.

data set. Annotation of GRS was done automatically by simple, context insensitive matching to have as many training examples as possible of all kinds of diseases at the expense of including false positives due to ambiguous disease terms. Since GRS are manually added to a gene's record, they might also contain information, which is not present in the abstracts of publications. Thus, only by looking at abstracts and not at full texts we are likely to miss gene-disease associations. Our predictions of human genes associated with genetic disorders are evaluated by comparing to the OMIM gene-disease catalog Table 3. We successfully mapped 358 disease concepts in MeSH to their respective counterparts in OMIM and used them as a benchmark data set. As expected, the achieved results in terms of recall are much better than for our GRS benchmark, since all annotations in OMIM can be regarded as more reliable as GRS entries.

Prediction of Cell-Cycle Genes. A recent genome-wide high-throughput RNAi screen identified 1351 genes important for cell division in HeLa cells [KPH$^+$07]. Among these 1351 genes, only 243 were previously associated with cell-cycle progression, and 252 previously uncharacterized genes were assigned this function. The remaining 882 genes were known to be implicated in other functions than cell-cycle. Another study characterized the genome-wide program of gene expression during the cell division cycle in HeLa cells us-

ing cDNA micro-arrays [WSS+02]. They identified genes periodically expressed during cell-cycle progression. These genes are not necessarily important to cell-cycle itself, but are nevertheless downstream of the cell-cycle machinery. As shown in Fig. 1D only 89 genes are both important for cell-cycle and periodically expressed, and 53 among these were previously known to be important for cell-cycle.

We compare these experimental knowledge to gene–cell-cycle associations mined from MEDLINE abstracts Table 3. To improve recall, we used protein sequence homology as well as protein interactions to transfer associations using the principle of guilt-by-association. First we evaluated how many genes associated to cell-cycle in GOA and KEGG can be recovered using text-mining. We achieve a recall of 40.6% and a precision of 43.3% using co-occurrence of genes with cell-cycle GO terms. Using the principle of guilt-by-association, the recall can be improved to 45.3% when using protein sequence homology, and to 67.9% when using protein interactions from the HPRD database [MSK+06]. Thus, high-quality protein interactions are a valuable resource for improving recall. Moreover, predicting all genes known to be related to cell-cycle either in GOA, KEGG, [KPH+07] or [WSS+02] leads to a decrease in recall of 19.3% except when using protein interactions. In this case, the recall more than doubles to 44.3% with the best F1-measure of 34.5%.

Next, we tried to predict the 882 new cell-cycle related genes identified by [KPH+07] using text-mining, protein sequence homology and protein interactions. Using solely gene–cell-cycle co-occurrences, our approach achieved a maximum p-value of 0.51 indicating the difficulty of predicting previously unknown cell-cycle genes using text-mining alone. Adding protein sequence homology leads to a slightly better p-value of 0.24 – still insignificant. However, using protein interactions improves the statistical significance of the results with a p-value of 0.016 (HPRD). In this case we are able to confirm 24.3% of the new cell-cycle genes identified by [KPH+07].

4 Discussion

The method we proposed can find novel links between genes and diseases not yet contained in databases such as Entrez Gene and OMIM. Especially genes that link to neoplasms are of high importance because of the high mortality of cancer patients. Among those genes, we can highlight interesting ones using a ranking that integrates different measures for interest such as *novelty*, *volume*, *community*, and *peaked*. We picked 10 different cancers categories (8 by site and 2 by type) and ranked the associated genes for each category (Tab. 4). Interestingly, none of our top ranked genes for pancreatic cancers is listed in OMIM. A manual inspection of those genes showed that each is indeed linked to pancreatic cancers. A possible explanation is that OMIM only includes genes shown to follow Mendelian inheritance patterns. For example, the top-ranked gene SIRT4 was first reported in 1999 in a publication about the characterization of five human yeast SIR2 homologs. Next publications followed in 2002, 2003, and 2005 revealing the regulation of SIRTs by histone deacetylase inhibitors. Then in 2006, two papers in the high-impact journal Cell were published reporting that SIRTs turn out to be critical regulators of metabolism

Prediction	Rec.(%)	Prec.(%)	F1(%)	p-value(<)
gene-cell cycle GO terms co-occurrence:				
GOA	55.7	32.2	40.4	10^{-186}
GOA+KEGG	40.6	43.3	41.4	10^{-204}
GOA+KEGG+Kittler	21.3	47.3	29.0	10^{-123}
GOA+KEGG+Kittler+Whitfield	19.3	52.2	27.8	10^{-125}
gene-cell cycle GO terms co-occurrence + sequence homology:				
GOA	61	28.0	38.3	10^{-191}
GOA+KEGG	45.3	38.4	41.2	10^{-209}
GOA+KEGG+Kittler	24.4	43.0	30.8	10^{-125}
GOA+KEGG+Kittler+Whitfield	19.3	52.2	27.8	10^{-125}
gene-cell cycle GO terms co-occurrence + PPI:				
GOA	**71.7**	13.5	21.9	10^{-136}
GOA+KEGG	**67.9**	17.1	27	10^{-154}
GOA+KEGG+Kittler	**45.3**	23.8	30.4	10^{-87}
GOA+KEGG+Kittler+Whitfield	**44.3**	28.3	**34.2**	10^{-100}
Predicting new cell cycle genes found in Kittler et al. [KPH$^+$07]:				
Cell cycle term co-occurence	3.4	7.0	4.2	0.51
+ protein sequence homology	6.7	7.7	6.4	0.24
+ protein interactions (HPRD)	**24.3**	8.0	12	**0.016**

Table 3: Predicting cell cycle related genes using GO term co-occurrence, protein sequence homology and protein interactions. Predicting known cell cycle related genes from GOA can be done at a maximal recall of 71%. Protein sequence homology only improves recall at the cost of a loss in precision. Using protein interactions improves recall, and when predicting all known cell cycle related genes (GOA+KEGG+Kittler+Whitfield) it achieves a higher F1-measure than pure co-occurrence. Predicting the new cell-cycle genes of [KPH$^+$07] does not work using pure text-mining ($P = 0.51$), is only marginally improved using protein sequence homology ($P = 0.24$), but becomes significant when using protein interactions from HPRD ($P = 0.016$).

Bone	Brain	Breast	Eye	Leukemia	Liver	Lymphoma	Pancreas	Prostate	Skin
FXYD6	CRB3	SIRT7	E2F5	CLLU1	LIN28B *	NPC1L1	SIRT4	OR51E1 *	KRT1
ADAM8	GHRHR	BRIP1 *	E2F1 *	ARL11 *	HNF1A *	ULBP2 *	G6PC2	OR51E2 *	MSH2 *
C9orf46	SCGN	BRCA1 *	CDK4 *	FCRL3	TCP10L *	TBX21	SOX2	TMPRSS2 *	MSH6 *
FGF23 *	HDAC-3	SERPINB5 *	KIF14 *	CCDC28A	UGT2B7	FHL2	FFAR1	PCA3 *	MLH1 *
TRIB2	SOX4	AP1S2	RBBP8 *	GATA1 *	ZNF689	CCR7	CDX2	PI16	KRT15

Table 4: Top five genes for 10 cancer. The top five genes according to the Pareto ranking for 10 different neoplasms. Most of the listed genes have seen an increase in research interest in recent years and have a high volume of high impact publications. Genes with a star are mentioned in OMIM to be related to the corresponding disease, genes without a star are not. Note that for brain and pancreas cancer none the top 5 genes identified are listed in OMIM.

and that SIRT4 acts in the mitochondria of pancreatic cells. The loss of SIRT4 in insulinoma cells up-regulates amino-acid-stimulated insulin secretion, which links SIRT4 to pancreatic cancers [HMH$^+$06]. Another member of the histone deacetylase gene family is SIRT7, which we found associated to breast cancers. Together with SIRT4, this gene was

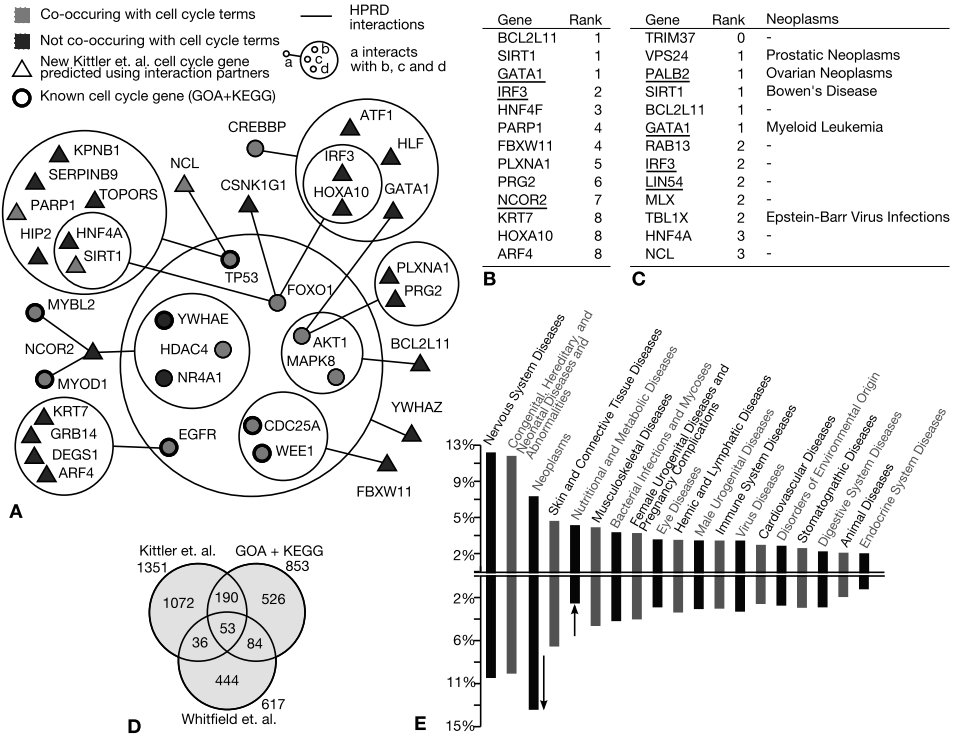
first reported in 1999. A recent publication in the British Journal of Cancer reports that levels of SIRT7 expression were significantly increased in breast cancers.

As expected, we found more links to cancers among genes known to be involved in cell-cycle progression, since defects in the cell-cycle are causative for cancers development (Fig. 1E). A recent RNAi screen identified more than 850 new genes with impact on cell-cycle progression [KPH$^+$07]. Out of those, 24% can be further confirmed by literature mining combined with high confidence protein interaction networks. A ranking of these genes highlights the interesting candidates for further research and confirmation studies (Fig. 1B and C). Figure 1A shows a sub-graph of the HPRD network with genes predicted by our method. For example, let's examine the gene IRF3 – an interferon regulatory factor. It forms a complex with CREBBP and thus interacts in the network with CREBBP [YLM$^+$02]. Moreover, CREBBP co-occurs with cell-cycle terms such as 'DNA replication checkpoint', 'centriole replication', 're-entry into mitotic cell-cycle'. These co-occurrences together with the interaction between IRF3 and CREBBP is our evidence for a link between IRF3 and cell-cycle. The importance of IRF3 for the cell-cycle can be further confirmed in that the target genes of IRF-3 are themselves involved in cell-cycle as shown in a recent Nature publication [AVC$^+$07].

PALB2 is a breast cancer susceptibility gene that interacts with BRCA2 to enable its recombinational repair and checkpoint functions [XSN$^+$06]. When mutated, it more than doubles the risk of breast and ovarian cancers [WK07]. PALB2 is among the genes identified by [KPH$^+$07] and predicted using both gene-cell-cycle co-occurrences and protein interactions (Fig. 1C). In HPRD, PALB2 is reported to interact with BRCA2, a gene which we find co-occurring significantly with cell-cycle terms in MEDLINE. Yet, PALB2 itself is not associated to cell-cycle in GOA or KEGG nor does it significantly co-occur with cell-cycle terms in the literature. This example shows how protein interactions can help to recover such candidates. As seen in Fig. 1C, our method correctly associates PALB2 as top ranked for ovarian cancers. PALB2 is definitely a 'hot' gene first researched in 2006 and discussed in 2007 in four Nature publications.

5 Conclusion

We showed the feasibility of a simple statistical co-occurrence model to find links between genes and diseases as well as between genes and cell-cycle processes by automatically searching the literature and using high confidence protein interactions. The achieved results for finding those associations are comparable to recent approaches to relationship extraction from texts, such as protein-protein interactions [KLV07]. The main contributions of our work are: i) the application of a state-of-the-art gene name identifier to all articles indexed in MEDLINE, ii) the ranking of all genes discussed in the literature by different measures of interest, iii) the potential to find novel links between genes, cancers, and cell-cycle processes not yet annotated in public databases, and iv) the support of high-throughput experiments by filtering results using knowledge from literature and known interaction networks to select the most promising gene candidates.



B

Gene	Rank
BCL2L11	1
SIRT1	1
GATA1	1
IRF3	2
HNF4F	3
PARP1	4
FBXW11	4
PLXNA1	5
PRG2	6
NCOR2	7
KRT7	8
HOXA10	8
ARF4	8

C

Gene	Rank	Neoplasms
TRIM37	0	-
VPS24	1	Prostatic Neoplasms
PALB2	1	Ovarian Neoplasms
SIRT1	1	Bowen's Disease
BCL2L11	1	-
GATA1	1	Myeloid Leukemia
RAB13	2	-
IRF3	2	-
LIN54	2	-
MLX	2	-
TBL1X	2	Epstein-Barr Virus Infections
HNF4A	3	-
NCL	3	-

Figure 1: Predicting Cell Cycle genes. (A) Example HPRD Sub-network of protein interactions for new cell cycle genes [KPH$^+$07] predicted using our method, and visualized using power graphs [RRAS08]. Genes like IRF3 and NCOR2 can be predicted using both gene–cell cycle term co-occurrence and high quality protein interactions from HPRD. (B) Top 12 hottest genes among genes shown in the example sub-network. Genes BCL2L11, SIRT1, GATA1 and IRF3 are at top. (C) Top 12 hottest genes among all new cell cycle genes from [KPH$^+$07] predicted by our method together with significantly co-occurring neoplasms. (D) Overlap between [KPH$^+$07] cell cycle genes, [WSS$^+$02] cell cycle periodic genes, and known cell cycle genes annotated in GOA or KEGG. Among the 53 genes in all three sets we find genes involved in the structural aspects of cell cycle such as histones, centromere proteins, tubulins and kinesins, that are both important and periodically expressed. Only 36 genes are both found by [KPH$^+$07] and [WSS$^+$02] but are not annotated in GOA or KEGG such as MELK and CBX3. (E) Disease associations mined from literature for all human genes (top) compared to disease associations for cell cycle genes. As expected, cell cycle genes are enriched in neoplasms and depleted in nutritional and metabolic diseases.

References

[AAE+05] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, 2005. disease.

[AVC+07] J. Andersen, S. Vanscoy, T-F. Cheng, D. Gomez, and N. C. Reich. IRF-3-dependent and augmented target genes during viral infection. *Genes Immun*, Dec 2007.

[BCF+07] WA Baumgartner, KB Cohen, LM Fox, G Acquaah-Mensah, and L Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48, Jul 2007.

[BK06] Atul J Butte and Isaac S Kohane. Creation and implications of a phenome-genome network. *Nat Biotechnol*, 24(1):55–62, Jan 2006. disease.

[GCV+07] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proc Natl Acad Sci U S A*, 104(21):8685–8690, May 2007. disease.

[GLF+06] Richard A George, Jason Y Liu, Lina L Feng, Robert J Bryson-Richardson, Diane Fatkin, and Merridee A Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*, 34(19):e130, 2006. disease.

[GUT+08] Graciela Gonzalez, Juan C Uribe, Luis Tari, Colleen Brophy, and Chitta Baral. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. *Pac Symp Biocomput*, pages 28–39, 2008.

[HMH+06] MC Haigis, R Mostoslavsky, KM Haigis, K Fahie, DC Christodoulou, AJ Murphy, DM Valenzuela, GD Yancopoulos, M Karow, G Blander, C Wolberger, TA Prolla, R Weindruch, FW Alt, and L Guarente. SIRT4 inhibits glutamate dehydrogenase and opposes the effects of calorie restriction in pancreatic beta cells. *Cell*, 126(5):941–54, Sep 2006.

[HPR+08] J. Hakenberg, C. Plake, L. Royer, H. Strobelt, U. Leser, and M. Schroeder. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*, 2008. to appear.

[KLV07] Martin Krallinger, Florian Leitner, and Alfonso Valencia. Assessment of the Second BioCreative PPI task: Automatic Extraction of Protein-Protein Interactions. In *Proceeding of the Second BioCreative Challenge Evaluation Workshop*, pages 41–54, 2007.

[KPH+07] Ralf Kittler, Laurence Pelletier, Anne-Kristine Heninger, Mikolaj Slabicki, Mirko Theis, Lukasz Miroslaw, Ina Poser, Steffen Lawo, Hannes Grabner, Karol Kozak, Jan Wagner, Vineeth Surendranath, Constance Richter, Wayne Bowen, Aimee L Jackson, Bianca Habermann, Anthony A Hyman, and Frank Buchholz. Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nat Cell Biol*, 9(12):1401–1412, Dec 2007. cell cycle.

[LBO04] Núria López-Bigas and Christos A Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32(10):3108–3114, 2004. disease.

[LKRP+08] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C. Kuo, C. Hsu, R. Tsai, H. Hung, W. Lau, C. Johnson, R. Sæ tre, K. Yoshida, Y. Chen, S. Kim, S. Shin, B. Zhang, W. Baumgartner, L. Hunter, B. Haddow, M. Matthews, X. Wang, P. Ruch, F. Ehrler, A. Ozgur, G. Erkan, D. Radev, M. Krauthammer, T. Luong, R. Hoffmann, C. Sander, and A. Valencia. Introducing Meta-Services for Biomedical Information Extraction. *Genome Biology*, 2008. accepted.

[LKS+07] Kasper Lage, E. Olof Karlberg, Zenia M Storling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Soren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–316, Mar 2007. disease.

[MA94] Osborne M.J. and Rubenstein A. *A Course in Game Theory.* MIT Press, 1994.

[MSK+06] Gopa R Mishra, M. Suresh, K. Kumaran, N. Kannabiran, Shubha Suresh, P. Bala, K. Shivakumar, N. Anuradha, Raghunath Reddy, T. Madhan Raghavan, Shalini Menon, G. Hanumanthu, Malvika Gupta, Sapna Upendran, Shweta Gupta, M. Mahesh, Bincy Jacob, Pinky Mathew, Pritam Chatterjee, K. S. Arun, Salil Sharma, K. N. Chandrika, Nandan Deshpande, Kshitish Palvankar, R. Raghavnath, R. Krishnakanth, Hiren Karathia, B. Rekha, Rashmi Nayak, G. Vishnupriya, H. G Mohan Kumar, M. Nagini, G. S Sameer Kumar, Rojan Jose, P. Deepthi, S. Sujatha Mohan, T. K B Gandhi, H. C. Harsha, Krishna S Deshpande, Malabika Sarker, T. S Keshava Prasad, and Akhilesh Pandey. Human protein reference database–2006 update. *Nucleic Acids Res*, 34(Database issue):D411–D414, Jan 2006.

[PIBA02] Carolina Perez-Iratxeta, Peer Bork, and Miguel A Andrade. Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 31(3):316–319, Jul 2002. disease.

[PRW+09] Conrad Plake, Loic Royer, Rainer Winnenburg, Jörg Hakenberg, and Michael Schroeder. GoGene: gene annotation in the fast lane. *Nucleic Acids Res*, 37(Web Server issue):W300–W304, Jul 2009.

[RRAS08] Loïc Royer, Matthias Reimann, Bill Andreopoulos, and Michael Schroeder. Unraveling protein networks with power graph analysis. *PLoS Comput Biol*, 4(7):e1000108, 2008.

[TAT+06] Nicki Tiffin, Euan Adie, Frances Turner, Han G Brunner, Marc A van Driel, Martin Oti, Nuria Lopez-Bigas, Christos Ouzounis, Carolina Perez-Iratxeta, Miguel A Andrade-Navarro, Adebowale Adeyemo, Mary Elizabeth Patti, Colin A M Semple, and Winston Hide. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res*, 34(10):3067–3081, 2006.

[TCS03] Frances S Turner, Daniel R Clutterbuck, and Colin A M Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4(11):R75, 2003. disease.

[TKP+05] Nicki Tiffin, Janet F Kelso, Alan R Powell, Hong Pan, Vladimir B Bajic, and Winston A Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*, 33(5):1544–1552, 2005. text mining disease.

[vDCK+03] Marc A van Driel, Koen Cuelenaere, Patrick P C W Kemmeren, Jack A M Leunissen, and Han G Brunner. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet*, 11(1):57–63, Jan 2003. disease.

[vDCK+05] M. A. van Driel, K. Cuelenaere, P. P C W Kemmeren, J. A M Leunissen, H. G. Brunner, and Gert Vriend. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res*, 33(Web Server issue):W758–W761, Jul 2005. disease.

[WK07] Tom Walsh and Mary-Claire King. Ten genes for inherited breast cancer. *Cancer Cell*, 11(2):103–105, Feb 2007.

[WSS+02] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O

Brown, and David Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*, 13(6):1977–2000, Jun 2002.

[XSN+06] Bing Xia, Qing Sheng, Koji Nakanishi, Akihiro Ohashi, Jianmin Wu, Nicole Christ, Xinggang Liu, Maria Jasin, Fergus J Couch, and David M Livingston. Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol Cell*, 22(6):719–729, Jun 2006.

[YLM+02] Hongmei Yang, Charles H Lin, Gang Ma, Melissa Orr, Michael O Baffi, and Marc G Wathelet. Transcriptional activity of interferon regulatory factor (IRF)-3 depends on multiple protein-protein interactions. *Eur J Biochem*, 269(24):6142–6151, Dec 2002.