

Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies

AUTHOR(S)

E Teber, J Liu, S Ballouz, D Fatkin, Merridee Wouters

PUBLICATION DATE

01-01-2009

HANDLE

10536/DRO/DU:30038987

Downloaded from Deakin University's Figshare repository

Deakin University CRICOS Provider Code: 00113B

Deakin Research Online

This is the published version:

Teber, Erdahl T., Liu, Jason Y., Ballouz, Sara, Fatkin, Diane and Wouters, Merridee A. 2009, Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies, *BMC bioinformatics*, vol. 10, no. Supplement 1:S69, pp. 1-10

Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30038987

This is an open access article distributed under the terms of the attached BioMed Central License. See license for details.

Copyright : 2009, Teber et al

BMC Bioinformatics

BioMed Central

Research

Open Access

Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies

Erdahl T Teber[1], Jason Y Liu[1], Sara Ballouz[1], Diane Fatkin[1,2] and Merridee A Wouters*[1,2]

Address: [1]Victor Chang Cardiac Research Institute, 384 Victoria St, Darlinghurst, 2010, NSW, Australia and [2]School of Medical Sciences, University of New South Wales, Sydney, Australia

Email: Erdahl T Teber - e.teber@victorchang.edu.au; Jason Y Liu - j.liu@victorchang.edu.au; Sara Ballouz - s.ballouz@victorchang.edu.au; Diane Fatkin - d.fatkin@victorchang.edu.au; Merridee A Wouters* - m.wouters@victorchang.edu.au

* Corresponding author

from The Seventh Asia Pacific Bioinformatics Conference (APBC 2009)
Beijing, China. 13–16 January 2009

Published: 30 January 2009

BMC Bioinformatics 2009, **10**(Suppl 1):S69 doi:10.1186/1471-2105-10-S1-S69

This article is available from: http://www.biomedcentral.com/1471-2105/10/S1/S69

© 2009 Teber et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Automated candidate gene prediction systems allow geneticists to hone in on disease genes more rapidly by identifying the most probable candidate genes linked to the disease phenotypes under investigation. Here we assessed the ability of eight different candidate gene prediction systems to predict disease genes in intervals previously associated with type 2 diabetes by benchmarking their performance against genes implicated by recent genome-wide association studies.

Results: Using a search space of 9556 genes, all but one of the systems pruned the genome in favour of genes associated with moderate to highly significant SNPs. Of the 11 genes associated with highly significant SNPs identified by the genome-wide association studies, eight were flagged as likely candidates by at least one of the prediction systems. A list of candidates produced by a previous consensus approach did not match any of the genes implicated by 706 moderate to highly significant SNPs flagged by the genome-wide association studies. We prioritized genes associated with medium significance SNPs.

Conclusion: The study appraises the relative success of several candidate gene prediction systems against independent genetic data. Even when confronted with challengingly large intervals, the candidate gene prediction systems can successfully select likely disease genes. Furthermore, they can be used to filter statistically less-well-supported genetic data to select more likely candidates. We suggest consensus approaches fail because they penalize novel predictions made from independent underlying databases. To realize their full potential further work needs to be done on prioritization and annotation of genes.

Background

The process of linking genes to disease phenotypes is rapidly gaining momentum since the first disease-causing gene was identified 25 years ago [1]. Alternative approaches adopted in the past to identify disease genes are the candidate gene approach, where likely suspects are prioritised and screened on a genome-wide basis; and linkage analysis where specific loci are determined systematically using family studies. The two approaches have been synthesized into a pipeline by completion of the Human Genome Project; and further enabled by the increased availability of high-throughput experimental data and the development of sophisticated bioinformatics tools. In addition there have been efforts in the bioinformatics community to systematize and automate candidate gene prediction. Automated prediction systems provide geneticists with a reduced list of genes estimated to have a high probability of involvement in the disease phenotype by sifting through hundreds to thousands of genes. Ultimately, these tools aim to give the researcher the best possible guidance in honing in on the gene culprits for further biological confirmation. Since their introduction in the early 2000s, the predictive powers of automated candidate gene prediction systems have improved, largely due to increases in biological systems knowledge and more effective algorithms.

Candidate gene prediction systems vary in their approach and the data sources they draw on in generating predictions. These are summarised in Figure 1 and Table 1. Comparing the performance of these systems can be difficult because of the use of custom benchmark test sets by individual groups. Typically, benchmarking data is derived from genotype-phenotype information from the Online Mendelian Inheritance in Man (OMIM) database [2], but groups have used varying subsets of diseases. Several groups have tried to use standard benchmark sets [3-5], but these efforts have been limited. In addition, it is difficult to predict whether benchmarks which predominately contain data on well characterised diseases with Mendelian transmission patterns (i.e. dominant, recessive, X-linked) resulting from mutations in single genes [6] will be effective in predicting genes involved in less well characterised diseases, or in complex diseases.

A recent effort by Tiffin and colleagues [7] to identify candidate disease genes for the complex disease type II diabetes (T2D) and the related obesity trait predicted 12 genes in previously implicated chromosomal regions. The study also allowed a limited comparison of seven candidate gene prediction systems. Since that time two genome-wide association studies (GWAs) on T2D undertaken by the Wellcome Trust Case Control Consortium (WTCCC) and the Genetics Replication and Meta-analysis Consortium (DIAGRAM) have been published [8,9]. GWAs are a powerful tool for identifying genetic variants linked to complex diseases because they are more sensitive than linkage studies to small to moderate effect size contributions from polygenic and oligogenic diseases. The data from these GWAs allow the assessment of the predictions made by Tiffin *et al.*, as well as evaluation of the effectiveness of predictions made by the individual automated candidate gene prediction systems used in their study and our system, *Gentrepid* [4]. We assessed the candidate gene predictions systems' ability to select robustly supported genes from the GWAs and used them to filter noisy data from statistically less well supported genes to select favoured candidates.

Results

Predictions

All methods were given the starting set of 9556 genes mapped to chromosomal intervals implicated in T2D as assessed by Tiffin *et al.*, except for *POCUS* which was run against a search space of 562 genes. The *POCUS* method was confined to the smaller search space because "poorly defined susceptibility regions or regions with questionable association with the disease are obscured by background noise" [7]. The number of candidate gene predictions made by the eight methods varied from two to 3093. *POCUS* generated the smallest number of candidates but neither of the two predictions matched genes in either the highly significant (HS) or medium-to-highly significant (MHWD) data sets. Other candidate gene prediction methods made considerably more predictions. The largest numbers of predictions were made by *G2D* (3,093 candidate gene predictions) and *eVOC* (2,496 predictions). These comprise almost one third and one quarter of the search space respectively. Thus neither of these methods prune the search space particularly well. Excluding *POCUS*, the least number of predictions was made by *Gentrepid* comprising 502 genes in known-disease-gene mode.

Accuracy of predictions

To assess the accuracy of the predictions, all eight systems were compared with genes found in previously-implicated intervals strongly linked to T2D by the GWAs. Figure 2 shows the comparative performance of seven of these methods in selecting the 11 genes in the HS GWA data set. Several metrics were calculated to assess accuracy. No metrics were calculated for *POCUS* as neither of its two predictions matched genes in either the HS or MHWD data sets.

The Enrichment Ratio is a general measure of the system's ability to accurately prune the search space. Enrichment Ratios ranged from 1 to 5 for the seven remaining prediction systems. The highest Enrichments Ratios were obtained by *Gentrepid* and *GeneSeeker*. These results were robust when the upper and lower (not shown) 95%

Table 1: Automated Candidate Gene Prediction Systems

Semi-Automated Systems
GeneSeeker is a semi-automated web-server tool which selects positional candidates based on expression and phenotypic data from human and mouse. Queries must be formulated by the end-user using Boolean expressions [13,33]. ♠ ◈
Systems Biology Techniques
Prioritizer uses pathway and interaction data from KEGG [17,34], Reactome [35], and HPRD [36]. Interactions are also predicted using a Bayesian technique based on GO keywords [23] and other databases [5].
In *Gentrepid* Common Pathway Scanning (CPS), pathways are associated with phenotypes using either known disease genes, or by searching for enrichment of pathways across multiple disease intervals associated with the phenotype [4]. ♠◈
Oti et al use protein-protein interaction data from HPRD [36], Y2H [37,38], and PCP [39,40] giving coverage of 10 894 human genes [24].
Genotype-Phenotype Mapping Methods
G2D [32] uses biomedical literature to associate pathological conditions with GO terms [23]. Candidate genes are identified by homology to GO-annotated disease-associated genes. ♠◈
Gentrepid Common Module Profiling (CMP) searches for enrichment of particular domains in gene clusters associated with a particular phenotype. Domains are extracted either from known disease genes or by comparison of multiple disease intervals [4]. ♠◈
POCUS searches for over-representation of functional annotation among multiple loci associated with the same disease. Functional annotation is based on keywords from InterPro domains [22] and GO [23]. No *a priori* knowledge of the phenotype is required [3]. ♠
Techniques based on a bipartite distribution of "disease" and "non-disease" genes
The *eVOC* system uses text mining of biomedical literature to associate a phenotype with anatomy terms and links these with human expression data to produce a ranked list of disease genes. The classifier is a machine-learning technique, based on a bipartite training set of 17 known "disease genes" and 306 "non-disease genes" [30]. ♠
DGP (Disease Gene Prediction) is a web tool which selects genes based on protein sequence properties. The properties analysed by DGP include protein length, degree of sequence conservation, the extent of phylogenetic relationship and paralogy patterns [31,41]. ♠
PROSPECTR (PRiOrization by Sequence and Phylogenetic Extent of CandidaTe Regions) uses an alternating decision tree to discriminate "disease genes" from "non-disease genes" using a classifier based on sequence features such as gene length, protein length, and similarity of homologs in other species [12]. ♠
Hybrid techniques
SUSPECTS combines a genotype-phenotype mapping method based on disease-gene-associated keywords from InterPro and GO, and expression libraries, with the *PROSPECTR* Boolean classifier. Disease genes are prioritized [21]. ♠ ◈

♠ Assessed here, ◈ Webserver.

confidence interval limits were taken into account. The lowest Enrichment Ratios were associated with the Machine Learning methods. This is not surprising, as the classifiers are trained to distinguish "disease genes" from "non-disease genes" and are ignorant of any concept of phenotype. The Specificity of a system measures its ability to reject genes not associated with the phenotype. Specificity scores among all seven methods ranged from 0.68 to 0.99, with a median of 0.92. As a group, the Machine Learning methods were poorer at rejection. *G2D* also performed poorly on this metric, but this result is slightly misleading because it does not take into account *G2D*'s prioritization method which will be discussed later.

The Sensitivity is a measure of a system's ability to find the disease genes in the search space. A caveat here is not all of the GWA predictions are currently confirmed. *G2D* is by far the standout performer in Sensitivity, with *eVOC* ranked second. However, as can be seen from the other metrics, this result is obtained at the expense of Specificity for both systems. *Gentrepid's* Sensitivity is on par with most of the Machine Learning methods but with higher Specificity. The high Specificity reflects the high quality of the data in the underlying databases. The lower Sensitivity is due to incompleteness of these databases with respect to all human genes.

Figure 2 shows the comparative performance of methods when assessed against the 61 T2D associated genes with moderate to strong SNP signals (MHWD) in the Tiffin chromosomal intervals. The MHWD data set is not as statistically well supported as the HS set, and would be expected to contain some genes associated with T2D and others that are false positives. Perhaps the most interesting metric to look at here is the Sensitivity which should fall compared to the values for the HS set because of the lower signal to noise ratio in the MHWD set. All the systems except one, *SUSPECTS*, passed this negative test. More importantly, application of the systems to this noisy genetic data allows selection of a subset of candidates on the basis of molecular data (see below).

The results shown for *Gentrepid* in Figure 2 are for the known-disease-gene mode. In *ab initio* mode, *Gentrepid*'s CPS method identified 506 pathways containing a total of 1980 candidate gene predictions. This resulted in Enrichment Ratios of 3.3 and 2.1 when the HS and MHWD full gene sets were considered (Table 2).

In *ab initio* mode, the CMP method generated 527 predictions by limiting the selection to the top 10% most probable genes. This resulted in correct prediction of one gene from the HS set and five from the MHWD set, yielding a

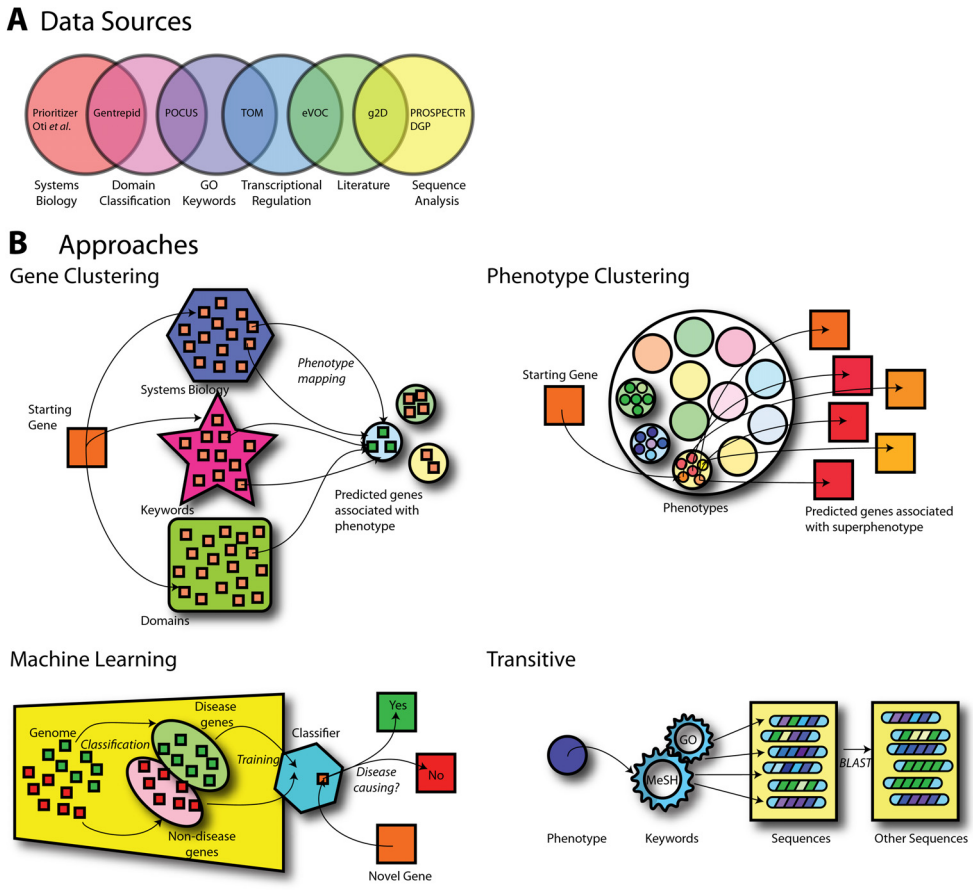


Figure 1

Data sources and approaches used in automated candidate gene prediction methods. **(A):** Most systems draw on at least two types of data. *SUSPECTS* [21] (not shown) uses keywords from InterPro [22] and GO [23], co-expression data, and also incorporates the *PROSPECTR* module [12] (shown on right). **(B)**: *Upper left* Gene clustering approaches associate a gene cluster with a phenotype via a group member. For example, Systems Biology approaches [4,5,24] group genes whose protein products interact; and link them to a phenotype using a group-member gene associated with the phenotype. Systems Biology methods assume oligogenic diseases are associated with disruption in proteins that participate in a common complex or pathway [25]. Other gene clustering systems look for enrichment of keywords or domains associated with particular phenotypes and suggest candidate genes with similar properties. These systems are based on the principle that candidate genes have similar functions to disease genes already determined [26-28]. *Upper right* Phenotype clustering approaches such as that of Freudenberg & Propping [29] group related phenotypes into superphenotypes. *Lower left* Most of the Machine Learning approaches do not use phenotype information and are based on the concept that the genome consists of a bipartite distribution of genes: those which cause diseases, and those that do not. By analysing these two gene sets with respect to discriminating variables, a profile for "non-disease genes" and "disease genes" is produced which enables training of a classifier. A novel gene submitted to the classifier is flagged as either "disease-causing" or "non-disease causing". Systems include *eVOC* [30], *PROSPECTR* [12], *SUSPECTS* [21] and *DGP* [31]. Finally *G2D*, *lower right*, is a transitive method that maps phenotypes to genes [32] by interfacing literature- and keyword-based ontologies.

Enrichment Ratio of 2.2 when applied to the HS and 2.0 for the MHWD gene data sets.

It is also interesting to note the effect of lack of annotation on these results. Only five of 11 genes in the HS dataset, and 19 of 61 genes in the MHWD set contained KEGG or BioCarta pathway annotations.When we included only genes containing pathway information from the gene datasets (designated 'annotated' in Table 2) we observed Enrichment Ratios of 7.2 against the HS and 6.8 against the MHWD pathway-annotated sets. Sensitivities also improved by a factor of 2 for the HS dataset. By extrapolation, if all genes were pathway annotated, we could expect approximately two- to three-fold improvement in Enrichment and Sensitivity scores.

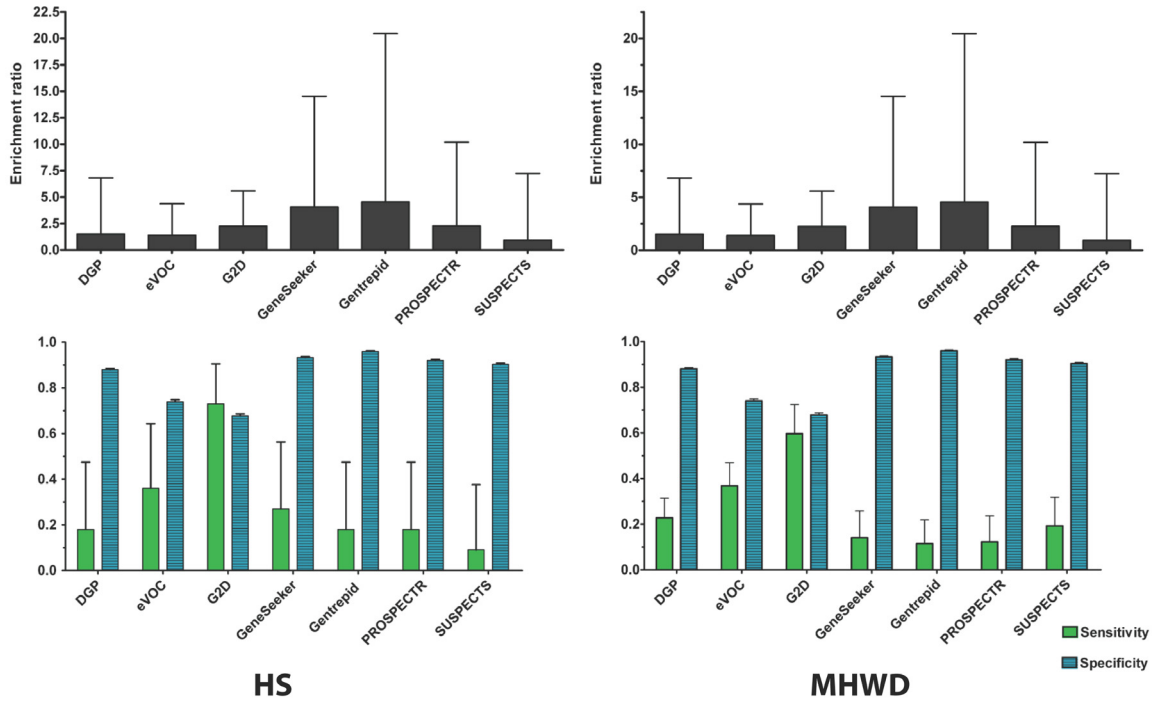


Figure 2
Comparison of methods against the HS (left) and MHWD (right) T2D gene data sets. Top: Relative Enrichment Ratios. Bottom: Comparisons based on Sensitivity and Specificity.

Prioritization

Although the metrics discussed provide useful measures of a candidate gene prediction system's performance, another criterion of importance to geneticists is the system's ability to prioritize predictions. Although several methods claim the ability to prioritize (Table 1), only *G2D* provided prioritized predictions in Tiffin *et al*. [7]. Hence only *G2D* and *Gentrepid* will be discussed here. Because *Gentrepid* only made 502 predictions *in toto*, we took the top 502 predictions made by *G2D* and recalculated the Enrichment Ratio, Sensitivity and Specificity for this restricted set of favoured predictions. ER and Specificity significantly improved to 3.1 and 0.95 such that *G2D* surpassed *Gentrepid*'s gross ER and almost equalled *Gentrepid* in Specificity. The improvement in these two metrics came at the expense of Sensitivity which was reduced to 0.16, but the *G2D* system still managed to maintain its lead on this metric.

In *G2D*'s prioritization system, a GO-metric is calculated for each gene in the search space based on how well its GO profile fits the GO profile of the disease genes inferred from MeSH terms. An R-score is calculated for each gene by normalizing against the number of genes in the genome with better GO-metrics for the phenotype. Genes with R-scores closer to zero are better fits to the phenotype.

Gentrepid CPS ranks genes by the number of loci in the search space involved in a particular pathway. In *ab initio* mode, of the 53 intervals searched, the top pathway, focal adhesion, was represented in 35 of these. All five of the HS dataset genes were represented by pathways found in at least eight intervals. Pathways implicated in at least eight intervals constituted the top 40% of the 506 pathways containing 1749 candidates. In these pathways, *Gentrepid* identified all five pathway annotated genes from the HS dataset, and 18 of the 19 pathway annotated genes from the MHWD gene data set. Other figures for CMP are given in Table 2.

CMP *ab initio* looks for protein domains enriched in the search space compared to the genome by taking a census of domains in the search space and the genome. Two expectation values are calculated to estimate the frequency of occurrence of genes with domains of interest based on a random combination of these domains e_a and the rarest domain e_b [4]. Figure 3 shows the data ranked on a χ^2 test based on e_a was most effective in prioritizing the HS data. This reflects our experience of phenotypes with genotypes

Table 2: Gentrepid *ab initio* results

Predictions	Reference list	ER	L95%	U95%	S	L95%	U95%
CPS rank 8+ pathways	HS	3.3	1.1	9.4	0.45	0.21	0.72
CPS rank 8+ pathways	HS – annotated	7.2	2.1	25	1.00	0.57	1.00
CPS rank 8+ pathways	MHWD	2.1	1.3	3.6	0.30	0.20	0.42
CPS rank 8+ pathways	MHWD – annotated	6.8	3.6	13	0.95	0.75	0.99
CPS interactions top 50%	HS	4.1	1.2	15	0.27	0.10	0.57
CPS interactions top 50%	HS – annotated	9.0	2.2	37	0.60	0.23	0.88
CPS interactions top 50%	MHWD	1.7	0.79	3.8	0.11	0.06	0.22
CPS interactions top 50%	MHWD – annotated	8.1	3.2	20	0.54	0.29	0.77
CMP top 10%	HS	2.2	0.3	17	0.1	0.02	0.38
CMP top 10%	MHWD	2.0	0.8	4.8	0.1	0.08	0.18

Abbreviations in Table: ER – Enrichment Ratio, L95% – Lower 95% confidence limit, U95% – Upper 95% confidence limit, S – Sensitivity

encoded by multi-domain proteins, as would be expected for diseases associated with signaling. For metabolic diseases associated with single-domain proteins, e_b may be a better measure.

Although the *G2D* prioritization system appears more sensitive than the coarse-grained prioritization of *Gentrepid*, the performance of both systems was roughly equivalent against the HS set. Both systems were moderately successful in prioritizing the HS data. For example, of the seven genes in the HS dataset predicted by *G2D*, four were ranked in the top 15% by *G2D*'s prioritization method (bold in Figure 3). Significant work needs to be done to improve the prioritization schemes of both *G2D* and *Gentrepid*.

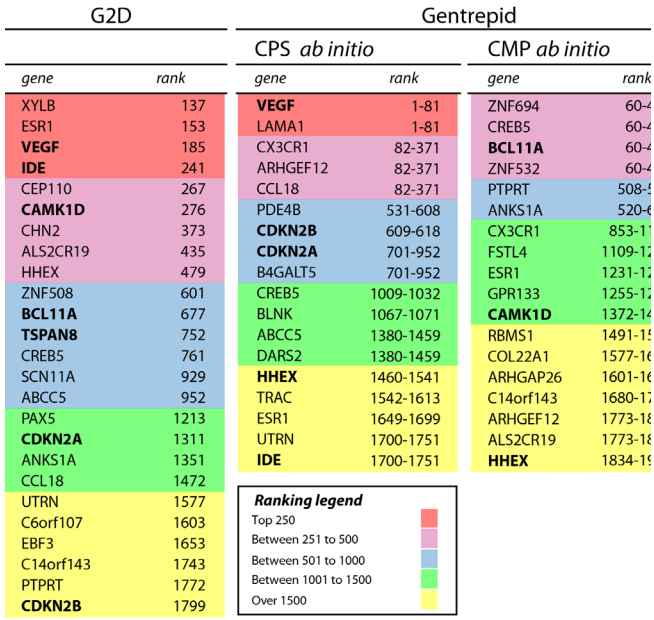
Finally, we used the candidate gene predictions systems to filter the less statistically-well-supported MHWD data set (MHWD – HS): effectively adding more power to the GWA study. Prioritized predictions are the unbolded genes in Figure 3. Additional unprioritized predictions made for the MHWD dataset using the other candidate predictions systems are given as supplementary data in Additional file 1.

Discussion

Candidate disease gene prediction is a rapidly moving area of bioinformatics research with the potential to deliver great benefits to human health. By assisting geneticists to use existing biological information to investigate disease loci obtained by linkage analysis and association studies, disease genes can be identified more rapidly. The need for good applications in the area of candidate gene prediction is becoming increasing important as the proliferation of SNP-based association studies produces valuable genetic information in need of analysis.

The biggest problem facing candidate gene prediction today is the accuracy and completeness of the underlying databases. Failure to make a prediction is mostly due to incomplete data coverage. For example, 65% of human proteins have GO terms but only 25% of these are manually annotated. Systems drawing on GO terms like *G2D* are potentially able to make predictions for 65% of genes but only around one third of these are likely to be accurate. Systems Biology methods like *Gentrepid* CPS are reliant on pathway and protein-protein interaction data. One of the databases CPS draws on is OPHID [10], one of the most complete protein-protein interaction datasets, containing over 48 000 interactions. However these 48,000 interactions are estimated to be only 13% of the complete human interactome [11]. Completeness of the underlying data clearly impacts the Sensitivity of the *Gentrepid* CPS method. As time goes on this constraint will ease as these databases are further populated. In the meantime, we have shown that the use of independent biological data to make complementary candidate gene predictions is one way to ameliorate the problem of incomplete data coverage (see Figure 3) [4].

In addition to the predictions made by the individual candidate gene prediction systems in Tiffin *et al.*, a set of nine "winners" were chosen using a consensus approach [7]. These nine candidate genes were independently predicted by six of the seven prediction systems studied. A larger consensus set, chosen by five of the seven methods, contained 94 genes [7]. None of the genes in either of these consensus lists matched any of the genes in the HS and MHWD gene sets. Even if we compile a third tier of consensus genes from any four of the seven methods (269 genes) only one gene (VEGF) fell within the HS data set and only three genes (CHN2, B4GALT5, VEGF) matched the MHWD data set. Clearly the consensus approach is not working and it is easy to see why when the underlying databases are considered (Figure 1A). Candidate gene prediction systems that use an independent data set, not drawn upon by most of the other methods, will be penalized. Possibly the only benefit of a consensus approach is



G2D		Gentrepid			
		CPS *ab initio*		CMP *ab initio*	
gene	*rank*	*gene*	*rank*	*gene*	*rank*
XYLB	137	**VEGF**	1-81	ZNF694	60-4
ESR1	153	LAMA1	1-81	CREB5	60-4
VEGF	185	CX3CR1	82-371	**BCL11A**	60-4
IDE	241	ARHGEF12	82-371	ZNF532	60-4
CEP110	267	CCL18	82-371	PTPRT	508-5
CAMK1D	276	PDE4B	531-608	ANKS1A	520-6
CHN2	373	**CDKN2B**	609-618	CX3CR1	853-1
ALS2CR19	435	**CDKN2A**	701-952	FSTL4	1109-12
HHEX	479	B4GALT5	701-952	ESR1	1231-12
ZNF508	601	CREB5	1009-1032	GPR133	1255-12
BCL11A	677	BLNK	1067-1071	**CAMK1D**	1372-14
TSPAN8	752	ABCC5	1380-1459	RBMS1	1491-15
CREB5	761	DARS2	1380-1459	COL22A1	1577-16
SCN11A	929	**HHEX**	1460-1541	ARHGAP26	1601-16
ABCC5	952	TRAC	1542-1613	C14orf143	1680-17
PAX5	1213	ESR1	1649-1699	ARHGEF12	1773-18
CDKN2A	1311	UTRN	1700-1751	ALS2CR19	1773-18
ANKS1A	1351	**IDE**	1700-1751	**HHEX**	1834-19
CCL18	1472				
UTRN	1577				
C6orf107	1603				
EBF3	1653				
C14orf143	1743				
PTPRT	1772				
CDKN2B	1799				

Figure 3
MHWD dataset filtered against prioritized automatic candidate gene predictions. Genes in bold are robustly supported genes from the GWA studies (HS set).

to give the user a false sense of accuracy when confronted with noisy data.

Clearly much work still remains to improve the sensitivity and specificity of candidate gene prediction methods but some general conclusions are possible. Machine Learning methods were not as effective as other methods. Most of the Machine Learning approaches do not use phenotype information and are based on the concept that the genome consists of a bipartite distribution of genes: those which cause diseases, and those that do not. The evidence supporting this assumption is limited [12]. We believe the concept that there is a difference between "disease genes" and "nondisease genes" is intrinsically flawed and no such Boolean classification exists. We hypothesize that the ability of these methods to predict disease genes in test sets is based on selection effects in the data: possibly rare, highly penetrant monogenic diseases, such as those involved in metabolic syndromes, are over-represented among known disease genes because they have been easier targets to identify. Although these systems were not as effective as the other candidate gene prediction systems, their performance was not greatly different. However, we believe that unlike systems which attempt to map genotype to phenotype, Machine Learning systems based on the disease gene/non-disease gene concept will not improve as more biological data becomes available.

Conclusion

Candidate gene prediction systems have typically been benchmarked on well characterized oligogenic phenotypes. GeneSeeker [13] produced a 10-fold enrichment using a data set consisting of eight diseases. *Gentrepid*'s combined methods [4] produced an Enrichment Ratio of 13 when 29 diseases with a total of 170 known disease genes were used. For 29 diseases with 163 genes, POCUS [3] reported Enrichment Ratios between 12 to 42-fold, depending on the size of the intervals in the search space. The PRIORITIZER [5] method yielded a 2.8-fold enrichment using a data set consisting of 96 heritable disorders. In summary, Enrichment Ratios of 3 to 13 have been reported in benchmarks, but a substantial part of the data used for these studies has been limited to oligogenic phenotypes, where several different genes may cause the disease, but a single mutation in each case or family has a large effect.

Some doubts have been raised about the ability of systems to predict candidates for complex polygenic diseases such as T2D where multiple genes interact to create a permissive gene pool for disease genesis. The candidate gene prediction systems did prune the genome in favour of moderately to highly significant SNPs identified by the GWAs under semi-blind testing on a complex polygenic disease. Enrichment Ratios calculated in this study suggest that most of the oligogenic benchmarks have been reasonably good predictors of system performance.

Methods

Benchmark datasets

Eight candidate gene prediction systems were assessed on their ability to predict genes involved in T2D by comparison against genes implicated by recent GWAs. Two data sets of T2D-implicated genes were used as the benchmark: a **H**ighly **S**ignificant gene set (HS) of 21 genes and a **M**oderate to **H**ighly significant gene set derived from the **WTCCC** and **DIAGRAM** studies (MHWD) of 172 genes [8,9]. The HS gene set contained 11 genes which mapped to the chromosomal regions investigated by Tiffin *et al.* (hereafter Tiffin intervals) [7]. Genes associated with 706 moderately significant SNPs with a frequentist additive p-value of <0.001, good clustering and intact NCBI build 36 reference ids were taken from WTCCC T2D data [8]. SNPs positioned between the 5' UTR and 3' UTR of a known gene structure, 1000 bases upstream of a 5' UTR or 1000 bases downstream of 3' UTR of a known gene were considered to implicate the gene in T2D disease susceptibility. This moderately significant list was combined with the genes from the HS data set to generate the MHWD dataset, yielding 172 genes genome wide of which 61 genes mapped to the Tiffin intervals [7].

Data sources for predictions

The search space available to all eight automated candidate gene prediction systems consisted of 9556 genes in 53 chromosomal loci assessed by Tiffin *et al*. to be involved in T2D by various linkage and association studies. We matched 96.5% of all Ensembl gene entries [14] provided to NCBI Entrez ids. All remaining genes were unable to be matched due either to the Ensembl entry having an unknown gene symbol label or because the entry was ambiguous or associated with a redundant gene symbol name entry. Ensembl entries and NCBI id matching was carried out at four levels, in order: approved symbol name, previous symbol names, Uniprot/SwissProt Accession and RefSeq Ids. Data conversion keys for matching between databases were acquired from BioMart [15].

Predictions made by seven candidate gene prediction methods were also obtained from Tiffin *et al*. [7]. Nine disease-implicated genes were available to the systems as seeds (PPARG, GYS1, IRS1, INS, KCNJ11, ABCC8, SLC2A1, PPARGC1, CAPN10). Two of the genes – PPARG and KCNJ11, are implicated by the highly significant SNPs detected by GWAs but are not in the Tiffin intervals and are thus not included in the search space or benchmark set.

Candidate gene predictions

The candidate gene predictions for seven of the systems are detailed elsewhere [7]. Briefly, *GeneSeeker* selected genes from the search space using a Boolean expression based on 14 keywords selected by an expert user [7]. *PROSPECTR*, *DGP* and *eVOC* are Boolean classifiers which require only the search space as input. *G2D* and *Gentrepid* in *ab initio* mode, also only require the search space. *POCUS* potentially only needs the search space as input, but this was restricted to the seven best supported intervals of the 53 available, as judged by the *POCUS* team. *SUSPECTS* and *Gentrepid*, in known-disease-gene mode, used the nine known disease genes associated with the phenotype as seeds. *SUSPECTS* additionally draws on predictions from the *PROSPECTR* Boolean classifier.

Gentrepid predictions are discussed in detail here for the first time. *Gentrepid* implements two different modules to derive predictions: CPS – a systems biology method; and CMP – a method that associates phenotypes with particular domains. CPS and CMP can be used in two input modes: using known disease genes as a seed or using only the search space (*ab initio* mode).

In known-disease-gene input mode, CPS searches all pathway and interaction data in BioCarta [16], KEGG [17] and I2D (formerly OPHID) [10] to extract all genes associated with the disease gene, and then filters this list against implicated loci. Genes are ranked based on the total number of genes implicated in the pathway. For example, if two known disease gene seeds and three genes in the loci being investigated are found in the same pathway, the pathway is given a rank of five against the phenotype. CMP parses the protein sequences of the known disease genes associated with the phenotype into domains using the Pfam library of Hidden Markov Models (HMMs) [18] and then retrieves any other genes with related domain content from the genome. A score between 0 and 1 is generated reflecting the candidate gene's similarity to a known disease gene [4]. The same nine disease genes and 53 chromosomal cytogenetic bands were used by *Gentrepid* as per Tiffin *et al*..

In *ab initio* mode, *Gentrepid* can make predictions in the absence of known disease genes if two or more loci are provided as input. *Gentrepid*'s CPS *ab initio* method is based on the premise that pathways whose genes are more prevalent within disease-implicated loci (chromosomal regions) compared to the entire genome have a higher probability of involvement in the pathoetiology of the disease phenotype of interest. Analogous to the known disease gene mode, pathways are ranked by the number of loci involved. The CMP *ab initio* method searches for enrichment of domains in the loci with respect to the genome and ranks genes based on the statistical significance of the domain enrichment (equations 2 and 4 in [4] where *mn* is replaced by Σ – the total number of genes in the intervals examined).

For each input mode, a final list of predictions is made by consolidating all predictions from both the CMP and CPS modules.

Metrics for comparisons

Systems were compared using three metrics: Enrichment Ratio, Sensitivity and Specificity. The Enrichment Ratio calculations were calculated as below [4]:

$$EnrichmentRatio = \frac{TP/(TP+FP)}{\sum genes_{implicated} / \sum genes_{all}} \quad (1)$$

The denominator was obtained by dividing the number of T2D implicated genes by the total number of genes within all surveyed chromosomal regions.

Sensitivity and Specificity were calculated as below:

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (2)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (3)$$

TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives. Sensitivity is the proportion of true positives among all disease genes in the chromosomal regions. Specificity is the proportion of true negatives among genes not associated with the disease in chromosomal regions. Confidence intervals were estimated using the method of Newcombe [19] implemented using the CIcalculator software [20].

List of abbreviations used

T2D: Type II diabetes; WTCCC: Welcome Trust Case Control Consortium; DIAGRAM: Genetics Replication and Meta-analysis Consortium; SNP: Single nucleotide polymorphism; GWA: Genome wide association studies; HPRD: Human Protein Reference Database; BIND: Biomolecular Interaction Network Database; DGP: Disease Gene Prediction; PROSPECTR: PRiOrization by Sequence and Phylogenetic Extent of CandidaTe Regions; OMIM: Online Mendelian Inheritance in Man; OPHID: Online Predicted Human Interaction Database; KEGG: Kyoto Encyclopedia of Genes and Genomes; GO: Gene Ontology; MeSH: Medical Subject Headings; HS: Highly Significant gene set; MHWD: Moderate to Highly gene set derived from WTCCC and DIAGRAM studies; ER: Enrichment Ratio; S: Sensitivity; TP: true positives; FP: false positives; TN: true negatives; FN: false negatives.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MW design, concept and oversight of study, and manuscript author. ET manuscript author, and data generation for study. JL implementation, construction and maintenance of database. SB additional data generation, figures and manuscript preparation. DF Genetics consultant. All authors read and approved the final manuscript.

Additional material

Additional file 1

Additional unprioritized MHWD matches. *Additional unprioritized MHWD matches data.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S1-S69-S1.txt]

Acknowledgements

The authors wish to acknowledge funding from the Ronald Geoffrey Arnott Foundation.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 1, 2009: Proceedings of The Seventh Asia Pacific Bioinformatics Conference (APBC) 2009. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S1

References

1. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB: **A Polymorphic DNA Marker Genetically Linked to Huntingtons-Disease.** *Nature* 1983, **306(5940):**234-238.
2. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders.** *Nucleic Acids Research* 2002, **30:**52-55.
3. Turner FS, Clutterbuck DR, Semple CAM: **POCUS: mining genomic sequence annotation to predict disease genes.** *Genome Biology* 2003, **4(11):**R75.
4. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction.** *Nucleic Acids Research* 2006, **34(19):**e130.
5. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *American Journal of Human Genetics* 2006, **78(6):**1011-1025.
6. Motulsky AG: **Genetics of complex diseases.** *J Zhejiang Univ Sci B* 2006, **7(2):**167-8.
7. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CAM, Hide W: **Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes.** *Nucleic Acids Research* 2006, **34(10):**3067-3081.
8. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447(7145**661-678 [http://dx.doi.org/10.1038/nature05911].
9. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PIW, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney ASF, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CNA, Payne F, Perry JRB, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.** *Nat Genet* 2008, **40(5):**638-645.
10. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21(9):**2076-2082.
11. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM: **Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.** *Genome Biology* 2005, **6(5):**R40.
12. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005, **6:**55.
13. van Driel MA, Cuelenaere K, Kemmeren PPCW, Leunissen JAM, Brunner HG, Vriend G: **GeneSeeker: extraction and integra-**

tion of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 2005, **33(Web Server issue):**W758-W761.
14. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJP: **Ensembl 2006.** *Nucleic Acids Research* 2006, **34:**D556-D561.
15. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **A generic system for fast and flexible access to biological data.** *Genome Research* 2004, **14:**160-169.
16. **BioCarta** [http://www.biocarta.com]
17. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Research* 2004, **32:**D277-D280.
18. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam Protein Families Database.** *Nucleic Acids Research* 2002, **30:**276-280.
19. Newcombe RG: **Improved confidence intervals for the difference between binomial proportions based on paired data.** *Statistics in Medicine* 1998, **17(22):**2635-2650.
20. **CIcalculator software.** . http://www.pedro.fhs.usyd.edu.au/calculator.html
21. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates.** *Bioinformatics* 2006, **22(6):**773-774.
22. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJA, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C: **New developments in the InterPro database.** *Nucleic Acids Research* 2007, **35(Database issue):**D224-D228.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25:**25-29.
24. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *Journal of Medical Genetics* 2006, **43(8):**691-8.
25. Badano JL, Katsanis N: **Beyond Mendel: An evolving view of human genetic disease transmission.** *Nature Reviews Genetics* 2002, **3(10):**779-789.
26. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes.** *Nature* 2001, **409(6822):**853-855.
27. Dudley AM, Janse DM, Tanay A, Shamir R, Church GM: **A global view of pleiotropy and phenotypically derived gene function in yeast.** *Molecular Systems Biology* 2005:2005.0001.
28. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, Suzuki G, Watanabe M, Hirata A, Ohtani M, Sawai H, Fraysse N, Latge JP, Francois JM, Aebi M, Tanaka S, Muramatsu S, Araki H, Sonoike K, Nogami S, Morishita S: **High-dimensional and large-scale phenotyping of yeast mutants.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(52):**19015-19020.
29. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics* 2002, **18:**S110-S115.
30. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates.** *Nucleic Acids Research* 2005, **33(5):**1544-1552.
31. Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease.** *Nucleic Acids Research* 2004, **32(10):**3108-3114.
32. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining.** *Nature Genetics* 2002, **31(3):**316-319.
33. **GeneSeeker web tool** [http://www.cmbi.ru.nl/geneseeker]
34. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BFF, Hogue CWV: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Research* 2005, **33:**D418-D424.
35. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L: **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Research* 2005, **33:**D428-D432.
36. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao ZX, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang LL, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Research* 2003, **13(10):**2363-2371.
37. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98(8):**4569-4574.
38. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang MJ, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403(6770):**623-627.
39. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415(6868):**141-147.
40. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang LY, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CWV, Figeys D, Tyers M: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415(6868):**180-183.
41. **DGP web tool** [http://cgg.ebi.ac.uk/services/dgp]