

Predicting disease genes using protein–protein interactions

M Oti, B Snel, M A Huynen, H G Brunner



J Med Genet 2006;43:691–698. doi: 10.1136/jmg.2006.041376

Background: The responsible genes have not yet been identified for many genetically mapped disease loci. Physically interacting proteins tend to be involved in the same cellular process, and mutations in their genes may lead to similar disease phenotypes.

Objective: To investigate whether protein–protein interactions can predict genes for genetically heterogeneous diseases.

Methods: 72 940 protein–protein interactions between 10 894 human proteins were used to search 432 loci for candidate disease genes representing 383 genetically heterogeneous hereditary diseases. For each disease, the protein interaction partners of its known causative genes were compared with the disease associated loci lacking identified causative genes. Interaction partners located within such loci were considered candidate disease gene predictions. Prediction accuracy was tested using a benchmark set of known disease genes.

Results: Almost 300 candidate disease gene predictions were made. Some of these have since been confirmed. On average, 10% or more are expected to be genuine disease genes, representing a 10-fold enrichment compared with positional information only. Examples of interesting candidates are AKAP6 for arrhythmogenic right ventricular dysplasia 3 and SYN3 for familial partial epilepsy with variable foci.

Conclusions: Exploiting protein–protein interactions can greatly increase the likelihood of finding positional candidate disease genes. When applied on a large scale they can lead to novel candidate gene predictions.

Many human genetic diseases can be caused by multiple genes. Since they lead to the same or similar disease phenotypes, the underlying genes are likely to be functionally related. Such functional relatedness can be exploited to aid in the finding of novel disease genes.¹ Direct protein–protein interactions are one of the strongest manifestations of a functional relation between genes, so interacting proteins may lead to the same disease phenotype when mutated. Indeed, several genetically heterogeneous hereditary diseases are known to be caused by mutations in different interacting proteins, such as Hermansky-Pudlak syndrome and Fanconi anaemia.^{2,3} Also, a recent study showed that interacting proteins tend to lead to similar disease phenotypes when mutated.⁴ Therefore protein–protein interactions might in principle be used to identify potentially interesting disease gene candidates.

Many human protein–protein interactions have been reported.⁵ These literature based interactions are reliable, but are naturally biased toward better studied proteins and have already been exploited by the community for disease

gene prediction. Protein–protein interactions from high throughput experiments do not have this bias, though they are also generally less reliable than literature based interactions.⁶ These high throughput sets are especially interesting for novel disease gene prediction as they can contain previously undescribed protein–protein interactions. There are two human high throughput protein–protein interaction sets available,^{7,8} but more are available from other species. These first have to be mapped to interactions between human proteins before they can be applied to disease gene prediction.

We investigated how successful protein–protein interactions are in predicting candidate disease genes for genetically heterogeneous hereditary diseases using a systematic large scale bioinformatics approach. To be as comprehensive and unbiased as possible we used both literature-based and high throughput human protein–protein interactions, and human mapped high throughput interactions from three other species—*Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (nematode), and *Saccharomyces cerevisiae* (baker's yeast). To identify potential new candidate disease genes, we examined whether disease proteins had interaction partners which were located within other loci associated with that same disease; such interaction partners were considered to be candidate disease genes. Several of these predictions have since been confirmed.

METHODS

Genetic disease data

Disease data were obtained from the Online Mendelian Inheritance in Man (OMIM, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>) “Morbid Map” list of diseases. This list contains disease loci and known disease genes from the OMIM database.⁹ We selected genetically heterogeneous hereditary diseases with at least one known disease causing gene and at least one disease locus lacking an identified causative gene. Disease subtypes were pooled into single diseases. A disease can have several loci, some of which may overlap each other. For instance, there are several X linked mental retardation subtypes, many of which have been mapped to overlapping loci. Therefore, a single protein–protein interaction could result in multiple candidate gene predictions for different disease subtypes.

The loci vary in length and gene count, with a median of 88 genes per locus, and a mean of 123.6; the loci lengths are not normally distributed. Whole chromosome loci were excluded from the analyses. In total, there were 383 diseases in the dataset. Together these diseases have 1195 disease loci with identified disease genes and 432 disease loci lacking identified causative genes.

The dataset used in the benchmark tests contained all the genetically heterogeneous hereditary diseases from Morbid

Abbreviations: HPRD, Human Protein Reference Database; OMIM, Online Mendelian Inheritance in Man; Y2H, yeast two-hybrid protein–protein interaction assay

Table 1 Protein–protein interaction sets used in the study

Interaction set	Source of interaction data*	Number of proteins in human mapped interactions†	Number of human mapped interactions‡	References	Comments
HPRD set	Published reports	6005	19 728	5	Downloaded October 21 2005
Human Y2H set	High throughput experiments (Y2H)	2686	5211	7, 8	Interactions from both experiments pooled
Fly set	High throughput experiment (Y2H)	4706	16 313	10	All interactions were used, regardless of confidence level
Worm set	High throughput experiment (Y2H)	1933	5771	11	Downloaded from DIP database ¹²
Yeast set	High throughput experiments (Y2H, PCP)	2455	27 098	Y2H: 13, 14 PCP: 15, 16	Interactions from all four experiments pooled
Combined high throughput set	All high throughput sets	8162	54 048	See above	Excluding HPRD
Total combined set	All interaction sets	10 894	72 940	See above	Including HPRD

*Y2H, yeast two-hybrid protein–protein interaction assay; PCP, protein complex purification experiment (based on mass spectrometry).

†These are the proteins that could be automatically mapped to human Ensembl gene IDs.

‡These are the interactions from the original interaction sets that could be automatically mapped to human Ensembl gene IDs.

Map with at least two known disease genes (289 diseases, 1114 disease loci with known disease genes, 1003 distinct genes). Only loci with known disease genes were used in the benchmark tests.

Protein–protein interaction sets

Five protein–protein interaction datasets were used in this study, from four different species—human, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (table 1). One of the human interaction sets contained manually curated protein–protein interactions culled from the literature, while all the other protein interaction data were from high throughput protein–protein interaction experiments.

The non-human protein–protein interactions were mapped to human proteins using orthology relations. Orthology between the other species' and human proteins was determined using the Inparanoid program,¹⁷ with default settings, on the whole (protein) genomes of the species. Genomes were acquired from Ensembl¹⁸ for the metazoan species and from the *Saccharomyces* genome database¹⁹ for yeast. Where one non-human protein was orthologous to several human proteins, the interaction was assumed to be valid for all these proteins.

Candidate gene prediction

For all the diseases in the heterogenic diseases dataset, the following actions were undertaken:

- The protein interaction partners were determined for each known disease protein.
- The chromosomal locations of the genes coding for these interacting proteins were determined using gene location data from the Ensembl database. The cytogenetic loci were mapped to chromosomal base pair ranges using the Ensembl database, which specifies cytogenetic band boundaries as exact base pair positions, rounded off to the nearest 100 kb.
- These chromosomal locations were checked to see if they fell within one or more disease loci (of the same disease) that lacked a known disease gene.
- Each interacting protein gene that was located within one of these loci was considered a candidate gene prediction.
- If a candidate gene lay within multiple (overlapping) loci of the same disease, each of them was counted as a separate prediction.

This procedure was carried out using a custom written C++ program (available on request), as were the benchmark and randomisation tests described below.

Benchmark tests

The benchmark test is introduced to examine how well a protein–protein interaction set performs in recovering known disease genes from different loci known to be involved in the same disease. These tests were therefore carried out analogously to the candidate gene prediction tests, with the exception that the protein interactor positions were examined against the disease loci with known disease genes (as opposed to the loci with unidentified disease genes). As with the prediction tests, these genes and their associated disease loci were taken from OMIM Morbid Map.

If an interactor lay within a disease locus it was considered a candidate gene prediction (a positive). If this interactor was indeed the known disease gene in that locus, it was considered a correct prediction (true positive). If it was not the known disease gene for that locus, it was considered a wrong prediction (false positive).

Randomisation tests

Owing to the complex nature of the data—potentially overlapping loci with different gene counts and networked protein–protein interactions—protein interactor randomisation tests were used to estimate the significance of the candidate gene prediction and benchmark results. In each case the genome was randomly shuffled and each protein in the interaction set was replaced with its counterpart from the shuffled genome. This approach retains the original structure of the interaction network; it only randomises the protein identities.

One thousand randomisation tests were carried out for each of the 10 analyses: five protein–protein interaction sets, each of which was used for both novel disease gene prediction and for benchmarking. In addition, separate randomisation tests were carried out for the two combined datasets, the combined high throughput set, and the total combined set.

Two other types of randomisation tests were carried out for each dataset—namely, the randomisation of the gene positions on the genome, and the shuffling of the protein interactors in the interactor sets. These led to similar results as the interactor identity randomisation (data not shown) and were left out of the results for brevity.

RESULTS

Benchmark tests perform well above random expectation

In order to examine how well protein–protein interaction sets predict disease genes in other loci of the same disease, we first undertook benchmark tests that attempted to predict known disease genes. These tests were carried out using only

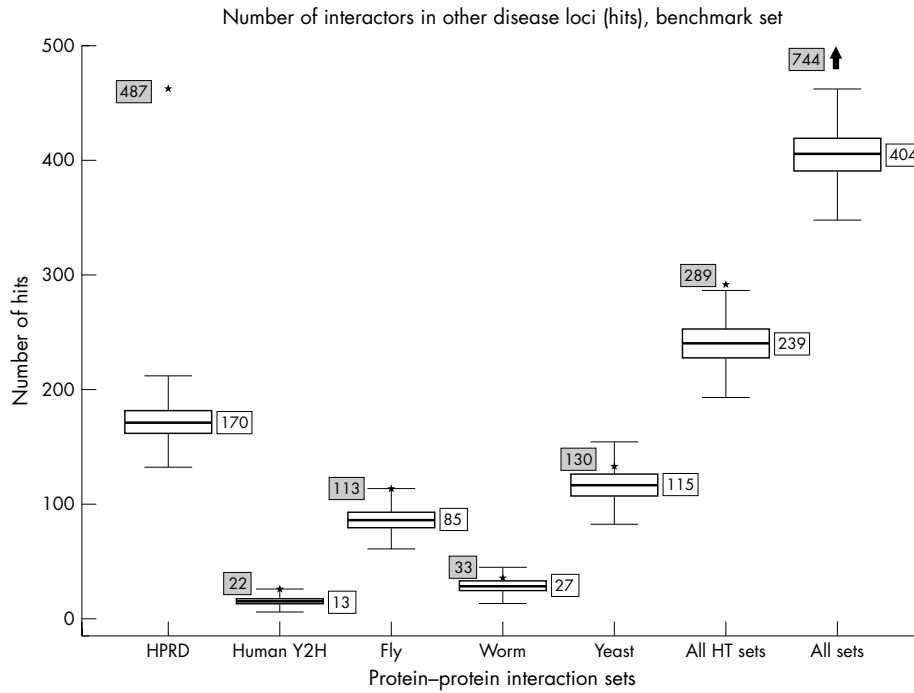


Figure 1 Overrepresentation of physically interaction proteins from loci associated with the same disease. The different protein-protein interaction sets used are on the X axis, while the Y axis contains the number of disease protein interactors falling within another locus associated with the same disease (hits) in the benchmark locus set. Black stars and associated shaded boxes indicate the values based on the real interaction datasets, while the box plots indicate the values resulting from the 1000 randomised interactor controls per set (numbers in clear boxes are medians). The value for the combined interaction dataset (indicated by the black arrow) is not included in the plot, to keep the Y axis scale manageable. The HPRD and total combined sets score much higher than all their randomised controls; the difference is smaller for the high throughput sets. HPRD, Human Protein Reference Database; HT, high throughput.

diseases with multiple known causative genes and their disease associated loci from OMIM Morbid Map, allowing the candidate gene predictions to be evaluated for accuracy and overrepresentation. With regard to the number of disease gene interaction partners that are located in another locus of the same disease, the HPRD interaction stood out as having over twice as many as would be expected by chance (fig 1). The high throughput sets all scored higher than the vast majority of their randomised controls, though the magnitude of this differed from the fly set, scoring higher than all but two controls, to the worm set, scoring higher than 823 of the

1000 randomised controls. We thus show here an overrepresentation of disease gene interaction partners in other disease associated loci, suggesting that disease genes encode proteins that tend to interact with each other.

The tendency for proteins associated with the same disease to interact with each other can be tested more directly by examining what percentage of these correctly localised interaction partners are indeed the known disease causing genes in those loci (fig 2). Once again, the HPRD protein-protein interactions performed very well. Almost 60% of these interacting proteins corresponded to the known disease

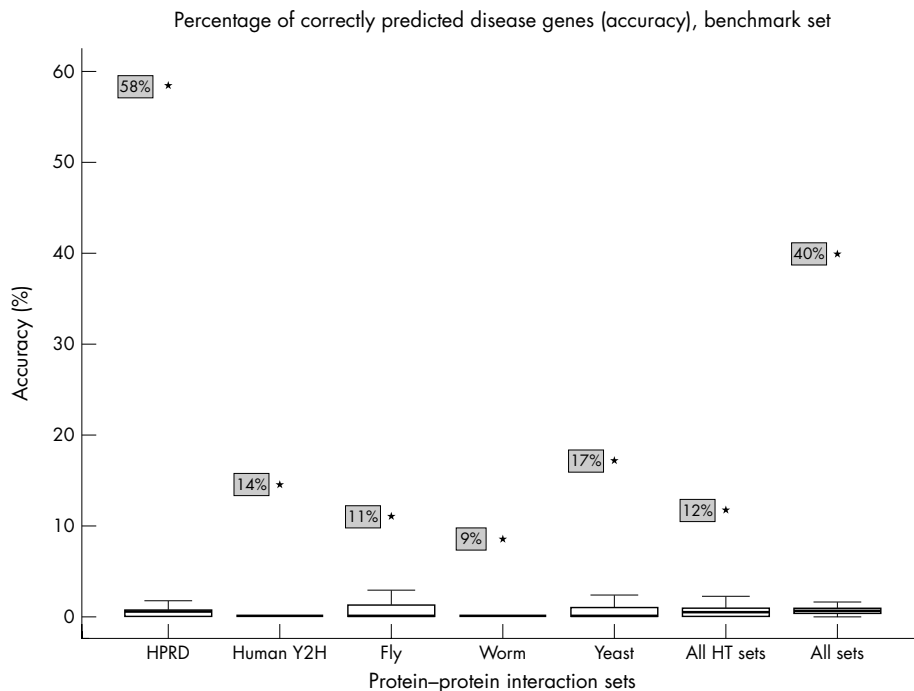


Figure 2 Relatively high likelihood of finding a disease gene from protein interaction data in a given locus. The different protein-protein interaction sets used are on the X axis, while the Y axis contains the percentage of disease protein interactors falling within other disease loci that correspond to the known disease genes in those loci. Black stars (and associated shaded boxes) indicate the values based on the real interaction datasets, while the box plots indicate the values resulting from the 1000 randomised interactor controls per set. Randomised controls have median accuracies of less than 1%. All interaction sets substantially outperform all their controls, with the HPRD scoring exceptionally high. HPRD, Human Protein Reference Database; HT, high throughput.

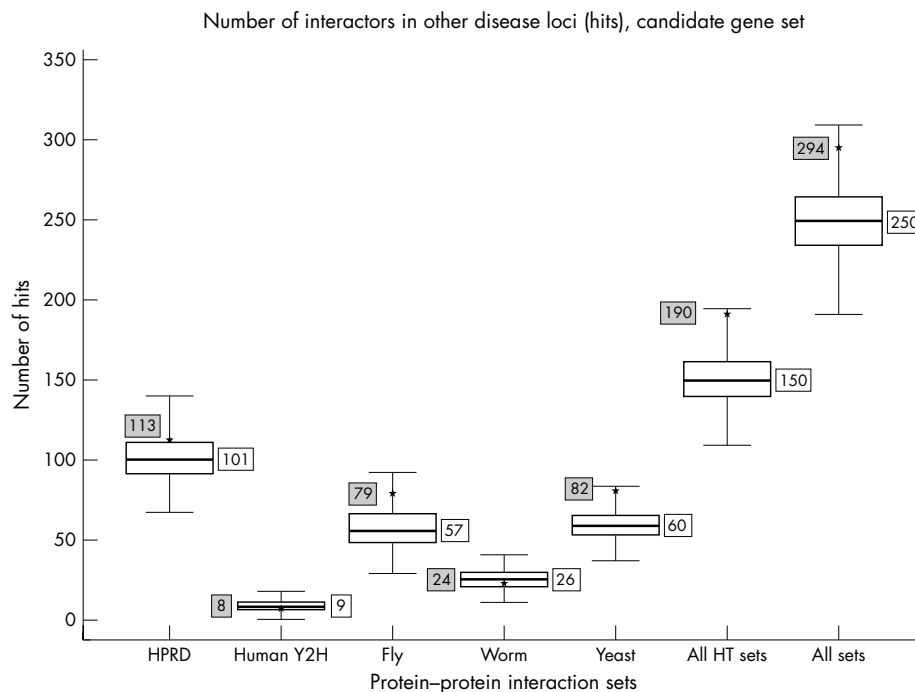


Figure 3 Candidate disease gene prediction hit counts. The different protein-protein interaction sets used are on the X axis, while the Y axis contains the number of disease protein interactors falling within another locus associated with the same disease (hits) in the candidate gene prediction locus set. Black stars (and associated shaded boxes) indicate the values based on the real interaction datasets, while the box plots indicate the values resulting from the 1000 randomised interactor controls per set (numbers in clear boxes are medians). The fly and worm sets score higher than the majority of their randomised controls, but the two smallest sets do not perform above random expectation. The HPRD scores relatively lower than the fly and worm sets, but still above the majority of its controls. HPRD, Human Protein Reference Database; HT, high throughput.

causing genes in these loci. The high throughput protein interaction sets had a lower performance (9–17%, with an average of 12%), but they all did much better than their randomised controls. Except for the two smallest sets, all interaction sets substantially outperformed every single run of their corresponding 1000 randomised controls. This implies that disease protein interaction partners, when located within other loci of the same disease, are at least 10-fold more likely to be involved in that disease than the other genes in these loci, given that randomly chosen locus genes have on average a 1% (1/88) chance of being correct.

An interesting result is that the yeast interaction set is more accurate in predicting candidate genes than the other high throughput sets, including those using “native” human proteins. Most of the yeast interactions are based on protein complex purification rather than yeast two-hybrid assays. A likely explanation for the relatively good performances of the yeast interactions is the previously observed higher quality of protein-protein interaction data from protein complex purification experiments relative to yeast two-hybrid assays.⁶

Candidate disease gene prediction results follow similar patterns

Having established the validity of the approach for known disease genes we compared the likelihood of mapping an interactor to a disease locus to chance expectation for disease loci for which we do not know the disease gene. As shown in fig 3, the three largest datasets and the combined data all showed an increase relative to the medians of the randomisation experiments. This result is consistent with that of the benchmark experiments for the high throughput data. It indicates an appreciable enrichment for true disease genes in the prediction results. Nevertheless the majority of the candidates are still probably false positives.

In contrast to the high throughput results from fly and yeast, the HPRD results do show a large discrepancy between the benchmark results and the novel gene prediction results. In the latter the enrichment of interactors in disease loci is much smaller than in the former. This suggests that the majority of disease genes which could be found using HPRD

protein-protein interactions have already been detected by the community, which is unsurprising given that all these interactions have previously been reported. Some of the interactions in the HPRD are even based on research on disease genes in the first place.^{20–23} Thus the HPRD benchmark results are not representative for its novel gene prediction results.

The two smallest high throughput interaction sets, human and worm, show no signal for the prediction of novel disease genes. It should be noted that the candidate gene prediction disease dataset contains fewer target loci (432) than the benchmark dataset (1114). One explanation for this might be that the combination of a small number of interactions and a small number of target loci prevents the two smallest sets from showing any signal here.

The full list of predicted candidate disease genes and their corresponding interactions are given in supplementary table 1 (the supplementary table can be viewed on the journal website (<http://www.jmedgenet.com/supplemental>)).

Confirmed and refuted predictions

A few candidate disease gene predictions could be confirmed or refuted because the disease causing genes are known but absent from the list of known disease genes used in the study (table 2). These disease loci were treated as having unidentified causative genes during the study, but manual inspection of the results led to their identification as disease loci with known causative genes.

It is encouraging to see that so many of these predictions were confirmed, though these are generally from the HPRD interaction set (of which 10 were confirmed and five refuted). From the high throughput sets there were two confirmed and 13 refuted predictions, which is consistent with their benchmark results (fig 2). This list may not be exhaustive owing to the complexities of thoroughly investigating all these predictions by hand, but we do not expect these proportions to change substantially. In any case the benchmark results remain valid, as the known gene disease loci are not susceptible to this misannotation problem.

Table 2 Confirmed and refuted candidate disease gene predictions

Disease subtype	Candidate genes	Based on interaction with*	Original disease subtype	Status	Comments
Branchiootic syndrome 3 [608389]	SIX1 (sine oculis homeobox homolog 1) [601205]†	EYA1 (eyes absent homolog 1) [601653]	Branchiootic syndrome 1 [602588]	Confirmed	Confirmed, ²⁴ but not in Morbid Map version used; interaction occurs in fly, worm and HPRD sets
SCID, autosomal recessive, T-negative/B-positive type [600802]	JAK3 (Janus kinase 3)	LCK (lymphocyte specific protein-tyrosine kinase) [153390] PTPRC (protein-tyrosine phosphatase, receptor type, C) [151460] IL2RG (interleukin 2 receptor, γ) [308380]	SCID caused by LCK deficiency [153390] SCID due to LCK deficiency [151460] SCID, X linked [300400]	Confirmed	Relevant Ensembl gene ID erroneously mapped to INSL3 symbol instead of JAK3, thus JAK3 not mapped to an Ensembl gene ID. All interactions from HPRD
Nephronophthisis 4 [606966] Senior-Loken syndrome 4 [606996]	NPHP4 (nephrocystin 4) [607215]	NPHP1 (nephrocystin 1) [607100]	Nephronophthisis, juvenile [256100] Senior-Loken syndrome 1 [266900]	Confirmed	NPHP4 gene symbol not mapped to corresponding Ensembl gene ID in Ensembl database version used. HPRD interaction
Charcot-Marie-Tooth disease, type 2L [608673]	HSPB8 (heat shock 22kDa protein 8) [608014]	HSPB1 (heat shock 22 kDa protein 1) [602195]	Charcot-Marie-Tooth disease, axonal, type 2F [606595]	Confirmed	HSPB8 identified as disease gene in OMIM database, but not in Morbid Map; HPRD interaction
Polycystic kidney disease, infantile severe, with tuberous sclerosis [600273] Pachyonychia congenita, Jadassohn-Lewandowsky type [167200]	PKD1 (polycystin 1) [601313] KRT6A (keratin 6A) [148041]	PKD2 (polycystin 2) [173910] KRT17 (keratin 17) [148069]	Polycystic kidney disease, adult, type II [173910] Pachyonychia congenita, Jackson-Lawler type [167210]	Confirmed Confirmed	Disease caused by chromosomal deletion which affects two genes, PKD1 and TSC2 ²⁵ ; mentioned in OMIM, but not in Morbid Map; HPRD interaction Ensembl ID maps to KRT6E, KRT6D, KRT6C and KRT6A; OMIM uses KRT6A name, whereas Ensembl uses KRT6E as primary name, thus mapping to corresponding Ensembl gene ID failed; from human high throughput set
Charcot-Marie-Tooth disease, type 2L [608673]	DNCL1 (dynein light chain, LC8-type 1) [601562] RNF10 (ring finger protein 10) [not in OMIM] MAPKAPK5 (mitogen activated protein kinase-activated protein kinase 5) [606723]	DNM2 (dynamin 2) [602378] GARS (glycyl-tRNA synthetase) [600287] HSPB1 (heat shock 22kDa protein 1) [602195]	Charcot-Marie-Tooth disease, dominant intermediate B [606482] Charcot-Marie-Tooth disease, axonal, type 2D [601472] Charcot-Marie-Tooth disease, axonal, type 2F [606595]	Refuted	HSPB8 is causative (see above). HSPB8 identified as disease gene in OMIM database, but not in Morbid Map; two HPRD interactions, one yeast
Marfan-like connective tissue disorder [154705]	FBN2 (fibulin 2) [135821]	FBN1 (fibrillin 1) [134797]	Marfan syndrome [154700]	Refuted	Causative gene is TGFBR2 (TGF β receptor II) [190182]. FBN2 was suspected but refuted ²⁶ ; mentioned in OMIM, but not in Morbid Map; HPRD interaction
Retinitis pigmentosa 26 [608380]	ENSG00000163510 [not in OMIM] HECW2 (HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2) [not in OMIM]	PRPF3 (Pre-mRNA processing factor 3 homolog) [607301] PRPF8 (Pre-mRNA processing factor 8 homolog) [607300] PRPF31 (Pre-mRNA processing factor 31 homolog) [606419] PRPF3 (Pre-mRNA processing factor 3 homolog) [607301]	Retinitis pigmentosa 18 [601414] Retinitis pigmentosa 13 [600059] Retinitis pigmentosa 11 [600138] Retinitis pigmentosa 18 [601414]	Refuted	Causative gene is CRKL (ceramide kinase-like) [608381]; gene name not mapped to Ensembl gene ID in Ensembl; three interactions from yeast set, one from fly set
Retinitis pigmentosa 10 [180105]	LUC7L2 (LUC7-like 2) [not in OMIM] METTL2 (methyltransferase like 2A) [607846]	PRPF3 (Pre-mRNA processing factor 3 homolog) [607301] PRPF31 (Pre-mRNA processing factor 31 homolog) [606419] CRX (cone-rod homeobox) [602225]	Retinitis pigmentosa 18 [601414] Retinitis pigmentosa 11 [600138] Retinitis pigmentosa, late onset dominant [268000]	Refuted	Causative gene is IMPDH1 (IMP dehydrogenase 1) [146690]; Morbid Map version used contains two entries for this subtype, one with and one without associated gene; two from yeast set, one from fly set
Spastic paraplegia 17 [270685]	SF3B2 (splicing factor 3b, subunit 2) [605591] KLC2 (kinesin light chain 2) [601334]	HSPD1 (heat shock 60kDa protein 1) [118190] KIF5A (kinesin family member 5A) [602821]	Spastic paraplegia 13 [605280] Spastic paraplegia 10 [604187]	Refuted	Known gene is BSCL2 (seipin) [606158]; mentioned in OMIM, but not listed in Morbid Map; both from HPRD set, KLC2 also from fly and worm sets

Table 2 Continued

Disease subtype	Candidate genes	Based on interaction with*	Original disease subtype	Status	Comments
Dyskeratosis congenita, autosomal dominant [127550]	EIF4G1 (eukaryotic translation initiation factor 4- γ 1) [600495] EIF4A2 (eukaryotic translation initiation factor 4A, isoform 2) [601102] KPNA1 (karyopherin α 1) [600686] CPA3 (carboxypeptidase A3, mast cell) [114851]	DKC1 (dyskerin 1) [300126]	Dyskeratosis congenital 1 [305000]	Refuted	Causative gene is TERC (telomerase RNA component) [602322]; gene symbol not mapped to Ensembl gene ID in Ensembl database version used; three from fly set (EIF4G1, EIF4A2, KPNA1) one from yeast (CPA3)

*These are the known disease genes, associated with other disease subtypes, which have physical protein–protein interactions with the candidate disease genes.

†Square brackets contain OMIM numbers for both diseases and genes.

A prediction is considered confirmed if it is known in published reports to be causative for the relevant disease, and considered refuted if a different gene in the same locus is known to be causative for that disease. It is important to note that a “refuted” candidate gene may not have been screened and excluded, and may thus still be a valid candidate.

Promising leads

In addition to these confirmed prediction the protein interaction sets also led to several plausible but unconfirmed candidate gene predictions. For instance, the AKAP6 (A-kinase (PRKA) anchor protein 6) gene is predicted as a candidate gene for arrhythmogenic right ventricular dysplasia 3 (OMIM 602086), based on an HPRD interaction with RYR2 (ryanodine receptor 2). A-kinase anchor proteins are involved in cardiac myocyte contractility and possibly also in heart failure.^{27–28} SYN3 (synapsin III) was predicted as a candidate gene for familial partial epilepsy with variable foci (OMIM 604364) based on an HPRD interaction with SYN1 (synapsin I), which is causative for epilepsy, X linked, with variable learning disabilities and behaviour disorders (OMIM 300491). There are several other interesting examples, which can be viewed in supplementary table 1 (<http://www.jmedgenet.com/supplemental>).

The HPRD interaction set is biased toward disease proteins

As it is based on interactions described in published reports, the HPRD interaction set is expected to be biased toward known disease genes. These proteins would be better studied, and known interactions with candidate disease proteins would have been investigated by the community already. This bias can already be seen in the HPRD benchmark results, but it can be quantified more directly by the proportion of the interacting proteins that are known disease proteins. Here once again the bias is clear—the proportion of interacting proteins that occurs in the heterogeneous disease protein set is twice as high for the HPRD set as for the high throughput sets, and is well over twice the proportion of Ensembl proteins in the disease protein set (table 3).

Interestingly, even the high throughput sets are enriched for disease genes. This suggests that genetically heterogeneous disease proteins are more likely to have protein–protein interactions, or that they have more easily detectable interactions.

Protein–protein interactions add as much information as localization

Hereditary diseases do not always have genetic loci associated with them. It is therefore interesting to see how much protein–protein interactions can at all predict the candidate disease gene pool in the absence of any genetic localisation data. When disregarding localisation information entirely in the benchmark disease gene set, the combined high throughput protein–protein interaction set still has a prediction accuracy of 0.7%. This is two orders of magnitude higher than the chance of randomly picking the disease protein from the entire genome and is of the same order of enrichment as genetic localisation, which generally reduces the candidate gene pool from ~20 000 to ~100. Combining these two information sources enriches the candidate gene pool a further order of magnitude to a one in 10 chance (12%) of picking the right disease gene (fig 2), which corresponds to a 1000-fold enrichment relative to the entire genome.

Needless to say, the HPRD interaction set performs much better than the combined high throughput set, resulting in a prediction accuracy of 6.6% when localisation data are disregarded. This corresponds to a 1000-fold enrichment even before localisation is taken into account. Once again, combining this with localisation information leads to a further 10-fold enrichment resulting in the 58% accuracy found in this study.

Table 3 Overrepresentation of heterogeneous disease genes in HPRD protein interaction set (χ^2 test).

	Number of proteins in interaction set	Subset also in disease protein set	Subset also in disease protein set (percentage)	χ^2 Test	p Value
HPRD set (literature based)	6005	678	11.29%	550.2098	<2.2e-16
Human Y2H set (high throughput)	2686	146	5.44%	4.845	0.03
Fly set (high throughput)	4706	276	5.86%	18.109	2.1e-5
Worm set (high throughput)	1933	101	5.23%	2.086	0.15
Yeast set (high throughput)	2455	141	5.74%	7.838	0.005
Reference set – all human protein coding genes in Ensembl					
	Total	In disease set	Percentage		
Ensembl known genes	22242	1003	4.51%		

The disease gene enrichment in HPRD is highly significantly higher than in the high throughput sets ($p < 1e-13$ after Bonferroni correction for every case).

DISCUSSION

The hypothesis being investigated here is that interacting proteins would often lead to similar disease phenotypes when mutated, enabling the usage of protein–protein interactions to suggest candidate disease genes. Our results suggest that this is indeed the case. Given the average locus size of close to 100 genes and high throughput interaction benchmark accuracies of 9–17%, positional candidate genes that interact with known disease genes have a more than 10-fold higher likelihood of being disease causing genes than random locus genes.

There are several practical limitations to the degree to which protein–protein interactions can predict disease gene candidates. To begin with, high throughput protein–protein interaction sets—especially yeast two-hybrid sets—are inherently noisy and contain a lot of interactions with no biological relevance.^{10 11 13 14} Therefore we might be predicting a disease gene based on an interaction that does not occur *in vivo*, but which did erroneously appear in a yeast two-hybrid assay. Indeed, only 5.8% of the human, fly, and worm Y2H interactions were confirmed by the HPRD, even among proteins common to both sets. However, given the Y2H set prediction accuracies of over 10% and the fact that the HPRD is not exhaustive, the proportion of Y2H interactions that are genuine is probably substantially higher than this figure suggests. Nevertheless, these high noise levels could reduce the accuracy of the Y2H based predictions relative to other techniques, as evidenced by the higher performance of the mainly protein complex purification based yeast interaction set.

Another practical limitation is the mapping of the high throughput interactions from other species to human proteins. In this study, when a protein in the other species had multiple human orthologues, the interaction was transferred to all of them. However, this need not be the case in reality. Encouragingly, we have previously shown that interactions between proteins are quite conserved across species and that conserved interactions tend to involve functionally related proteins.²⁹ Also, the yeast set outperforms the other sets despite its evolutionary distance to humans—though this may reflect the fact that most yeast interactions were from more reliable protein complex purification experiments rather than yeast two-hybrid assays.

Apart from the protein interactions, the designation of the candidate disease loci can also be a source of noise. Some of the candidate disease loci were designated based on incorrect reasoning, or faulty linkage assignment. For instance, we have recently shown that a family with EEC syndrome linked to chromosome 19 (EEC2, OMIM 602077)³⁰ actually has a mutation in the P63 gene denoted EEC3 (OMIM 604292) which is localised on human chromosome 3q27.³¹ However, the EEC2 locus remains in OMIM as a separate EEC locus with unidentified causative gene.

Furthermore, the use of cytogenetic bands to designate disease loci in OMIM Morbid Map can lead to problems in locating the genes in the Ensembl database.

Though they do not have sharp boundaries in reality, the Ensembl database uses specific base pair positions (rounded off to the nearest 100 kb) as band boundaries. Thus genes lying in the vicinity of a band boundary could easily be assigned to separate bands in published reports and in the Ensembl database. Indeed over 20% of the known disease genes in OMIM Morbid Map are associated with loci that differ from their Ensembl annotation. Most of these genes lie between 1 Mb and 10 Mb of their Morbid Map annotated loci on the same chromosome. The use of markers instead of cytogenetic bands could improve this; however, OMIM Morbid Map does not include marker information.

Finally, phenotypically similar diseases can be functionally related, even though they are classified as different diseases. As this study used pre-existing disease classifications rather than systematic phenotypic similarity analysis, potential links between disease genes causing similar but differently classified disease phenotypes would be overlooked. This would reduce the number of predictions made, without affecting the accuracy of those predictions that have been made.

All these practical limitations reduce the accuracy of the predictions, meaning that the true degree to which proteins involved in the same genetic disease interact is likely to be much higher. With higher quality protein interaction sets, more precise locus demarcation, and more systematic disease phenotype descriptions the value of this approach to disease gene prediction should increase even further.

Apart from the practical limitations, there are fundamental limits to the prediction capacity of protein–protein interactions. Two interacting proteins need not lead to similar disease phenotypes when mutated—for instance, they may have different but overlapping functions or one may be more dispensable than the other. Also, disease proteins may lie at different points in a molecular pathway and need not interact with each other directly. Disease mutations need not even involve proteins, as is the case with TERC (telomerase RNA component) in congenital autosomal dominant dyskeratosis (see table 2). Protein–protein interactions will thus not be capable of detecting every novel disease protein. Despite these fundamental limitations, the high proportion of disease proteins among correctly localised HPRD interaction partners is promising, although this interaction set is biased. And despite their practical limitations, even the high throughput datasets have prediction accuracies of up to 17%. Thus, in the absence of practical limitations, these fundamental limitations should result in a prediction accuracy that lies between these two values.

Outlook

This study provides evidence that the systematic use of protein–protein interaction data may lead to an approximately 10-fold improvement in positional candidate gene prediction. At the same time, the quality and quantity of the data available can be much improved. Though around 73 000 interactions between almost 11 000 proteins were used in this study, the actual number of interactions between these proteins should be much greater as all interaction assaying techniques miss large numbers of interactions.^{6 7} In addition, a more systematic phenotypic classification of diseases, such as our recently developed text mining approach,³² may lead to more interactions between related disease genes being identified. With increasing quantity and quality of interaction and phenotypic data and more dense interaction networks, the performance and utility of this approach to disease gene prediction should improve even further.

ACKNOWLEDGEMENTS

We thank Bas Dutilh for doing the orthology determination, and Gert Friend, Marc van Driel, René van der Heijden, Vera van Noort, and Toni Gabaldon for discussions and suggestions. This work is part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).



A supplementary table containing a full list of predicted candidate disease genes and their corresponding interactions is available on the journal website (<http://www.jmedgenet.com/supplemental>).

Authors' affiliations

M Oti, B Snel, M A Huynen, Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands
H G Brunner, Department of Human Genetics, University Medical Centre Nijmegen – St Radboud, Nijmegen, Netherlands

Conflicts of interest: none declared.

Correspondence to: Dr Han G Brunner, Department of Human Genetics, University Medical Centre Nijmegen – St Radboud, Geert Grooteplein 10, 6525 GA Nijmegen, Netherlands; H.Brunner@antrg.umcn.nl

Received 31 January 2006

Revised version received 10 March 2006

Accepted for publication 14 March 2006

Published Online First 12 April 2006

REFERENCES

- Brunner HG, van Driel MA. From syndrome families to functional genomics. *Nat Rev Genet* 2004;5:545–51.
- Di Pietro SM, Dell'Angelica EC. The cell biology of Hermansky-Pudlak syndrome: recent advances. *Traffic* 2005;6:525–33.
- Mace G, Bogliolo M, Guervilly JH, Dugas du Villard JA, Rosselli F. 3R Coordination by Fanconi anemia proteins. *Biochimie* 2005;87:647–58.
- Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* 2006;38:285–93.
- Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, 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 L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobel GC, Dang CV, Garcia JG, 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 Res* 2003;13:2363–71.
- von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 2002;417:399–403.
- Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgaard N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albaladejo JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 2005;437:1173–8.
- Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koepfen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droegge A, Krobatsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 2005;122:957–68.
- Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;33(Database issue):D514–17.
- Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaup E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM. A protein interaction map of Drosophila melanogaster. *Science* 2003;302:1727–36.
- Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, PIANO F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. A map of the interactome network of the metazoan *C. elegans*. *Science* 2004;303:540–3.
- Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 2004;32(Database issue):D449–51.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001;98:4569–74.
- 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 M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;403:623–7.
- 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, Lewtunew 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;415:141–7.
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreau 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 CW, Figeys D, Tyers M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002;415:180–3.
- Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001;314:1041–52.
- Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehtvaslahti H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Poccock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. The Ensembl genome database project. *Nucleic Acids Res* 2002;30:38–41.
- Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res* 1998;26:73–9.
- Huber PA, Medhurst AL, Youssoufian H, Mathew CG. Investigation of Fanconi anemia protein interactions by yeast two-hybrid analysis. *Biochem Biophys Res Commun* 2000;268:73–7.
- Schenck A, Bardoni B, Moro A, Bagni C, Mandel JL. A highly conserved protein family interacting with the fragile X mental retardation protein (FMRP) and displaying selective interactions with FMRP-related proteins FXR1P and FXR2P. *Proc Natl Acad Sci U S A* 2001;98:8844–9.
- Singaraja RR, Hadano S, Metzler M, Givan S, Wellington CL, Warby S, Yanai A, Gutekunst CA, Leavitt BR, Yi H, Fichter K, Gan L, McCutcheon K, Chopra V, Michel J, Hersch SM, Ikeda JE, Hayden MR. HIP14, a novel ankyrin domain-containing protein, links huntingtin to intracellular trafficking and endocytosis. *Hum Mol Genet* 2002;11:2815–28.
- Tsuchiya H, Iseda T, Hino O. Identification of a novel protein (VBP-1) binding to the von Hippel-Lindau (VHL) tumor suppressor gene product. *Cancer Res* 1996;56:2881–5.
- SIX1 mutations cause branchio-oto-renal syndrome by disruption of EYA1-SIX1-DNA complexes. *Proc Natl Acad Sci USA* 2004;101:8090–5.
- Brook-Carter PT, Peral B, Ward CJ, Thompson P, Hughes J, Maheshwar MM, Nellist M, Gamble V, Harris PC, Sampson JR. Deletion of the TSC2 and PKD1 genes associated with severe infantile polycystic kidney disease – a contiguous gene syndrome. *Nat Genet* 1994;8:328–32.
- Collod G, Chu ML, Sasaki T, Coulon M, Timpl R, Renkart L, Weissenbach J, Jondeau G, Bourdarias JP, Junien C, Boileau C. Fibulin-2: genetic mapping and exclusion as a candidate gene in Marfan syndrome type 2. *Eur J Hum Genet* 1996;4:292–5.
- Fink MA, Zakhary DR, Mackey JA, Desnoyer RW, Apperson-Hansen C, Damron DS, Bond M. AKAP-mediated targeting of protein kinase A regulates contractility in cardiac myocytes. *Circ Res* 2001;88:291–7.
- Zakhary DR, Moravec CS, Bond M. Regulation of PKA binding to AKAPs in the heart: alterations in human heart failure. *Circulation* 2000;101:1459–64.
- Huynen MA, Snel B, van Noort V. Comparative genomics for reliable protein–protein prediction from genomic data. *Trends Genet* 2004;20:340–4.
- O'Quinn JR, Hennekam RC, Jorde LB, Bamshad M. Syndromic ectrodactyly with severe limb, ectodermal, urogenital, and palatal defects maps to chromosome 19. *Am J Hum Genet* 1998;62:130–5.
- Celli J, Duijff P, Hamel BC, Bamshad M, Kramer B, Smits AP, Newbury-Ecob R, Hennekam RC, Van Buggenhout G, van Haeringen A, Woods CG, van Essen AJ, de Waal R, Vriend G, Haber DA, Yang A, McKeon F, Brunner HG, van Bokhoven H. Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome. *Cell* 1999;99:143–53.
- van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 2006, Feb 22 (epub ahead of print).