Systems Biology

An unbiased evaluation of gene prioritization tools

Daniela Börnigen[1,2,*], Léon-Charles Tranchevent[1,*], Francisco Bonachela-Capdevila[3,*], Koenraad Devriendt[4], Bart de Moor[1], Patrick De Causmaecker[3], and Yves Moreau[1#]

[1]Department of Electrical Engineering, ESAT-SCD, IBBT-KULeuven Future Health Department, Katholieke Universiteit Leuven, Leuven, Belgium
[2]Biostatistics Department, Harvard School of Public Health, Harvard University, Boston, MA, USA
[3]CODeS Group, ITEC-IBBT-KULEUVEN, Katholieke Universiteit Leuven campus Kortrijk, Kortrijk, Belgium
[4]Center for Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium
*Contributed equally to this work

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Gene prioritization aims at identifying the most promising candidate genes among a large pool of candidates—so as to maximize the yield and biological relevance of further downstream validation experiments and functional studies. During the past few years, several gene prioritization tools have been defined and some of them have been implemented and made available through freely available web tools. In this study, we aim at comparing the predictive performance of eight publicly available prioritization tools on novel data. We have performed an analysis in which 42 recently reported disease gene associations from literature are used to benchmark these tools before the underlying databases are updated.

Results: Cross-validation on retrospective data provides performance estimate likely to be overoptimistic because some of the data sources are contaminated with knowledge from the disease-gene association. Our approach mimics a novel discovery more closely and thus provides more realistic performance estimates. There are however marked differences, and tools that rely on more advanced data integration schemes appear more powerful.

Contact: yves.moreau@esat.kuleuven.be

1 INTRODUCTION

A major challenge in human genetics is to discover novel disease causing genes, both for Mendelian and complex disorders. Identifying disease genes is a crucial first step in unraveling molecular networks underlying diseases, and thus understanding disease mechanisms, also towards the development of effective therapies. The discovery of a novel disease gene often starts with a cytogenetic study, a linkage analysis, a high-throughput omics experiment, or a genome-wide association study (GWAS). However, these studies do not always pinpoint the disease gene uniquely, but often result in large lists of candidate genes that are potentially relevant (Hardy and Singleton, 2009). Moreover, recent advances in next-generation sequencing offer promising opportunities to explore the genomic alterations of patients (Schuster, 2008). However, thousands of mutations in hundreds of genes are often detected, among which only a few are in fact linked to the genetic condition of interest (Lupski et al., 2010). The experimental validation of these candidate genes, for instance through resequencing, pathway or expression analysis, is still expensive and time consuming. An efficient way to reduce the validation cost is to narrow down the large list of candidate genes to a small and manageable set of highly promising genes, a process called gene prioritization. Prioritization was in the past achieved manually by geneticists and biologists and was mainly based on their own expertise. Nowadays, biologists and geneticists can use computational approaches that can handle and analyze the large amount of genomic data currently available.

In the past few years, many gene prioritization methods have been proposed, some of which have been implemented into publicly available tools that users can freely access and use (Moreau et al., 2012; Doncheva et al., 2012; Piro et al., 2012; Tiffin 2011, Oti 2011; Tranchevent et al., 2010). Information about these tools is summarized in our Gene Prioritization Portal (http://www.esat.kuleuven.be/gpp) that currently describes 33 prioritization tools. This web site has been designed to help researchers to carefully select the tools that best correspond to their needs. For instance, only few tools can prioritize the whole genome, which can be necessary when no positive regions can be identified beforehand, or when selecting candidates for a medium-throughput screen (instead of low-throughput validation). Another example is the study of a poorly characterized disorder, for which a prioritization tool not relying on a set of known disease genes might be more suited. Recently, several studies have demonstrated that gene prioritization tools can help geneticists to discover novel disease genes (Thienpont et al., 2010; Calvo et al., 2006). For instance, a KIF1A mutation was discovered in hereditary spastic paraparesis patients after KIF1A was predicted to be the best candidate gene from the locus using multiple prioritization tools (Erlich et al., 2011). Another study discovered homozygous mutations in the PTRF-CAVIN gene in patients with congenital generalized lipodystrophy with muscle rippling after PTRF-CAVIN was predicted as the most probable candidate gene for high expression

#To whom correspondence should be addressed.

© Oxford University Press 2005

in muscle and adipose tissue (Rajab et al., 2010). A third study identified the HHEX gene to be associated with type 2 diabetes in a Dutch cohort after investigating the T2D-susceptibility loci using candidate gene prioritization (Vliet-Ostaptchouk et al., 2008).

However, beyond these conceptual differences, one essential parameter to consider when selecting gene prioritization tools is their respective performance—that is, their ability to identify the true positive genes as promising candidate genes in order to maximize the yield of the follow-up experimental validation. A common standard in bioinformatics is to estimate the performance with a benchmark analysis. Several publications that introduce a novel prioritization approach also describe a comparative benchmark with several existing methods (Hutz et al., 2008; Köhler et al., 2008; Thornblad et al., 2007). However, these benchmarks are most of the time cross-validations of gold-standard disease data sets (e.g., known data). Therefore, the estimation of the performance is likely an overestimate of the real performance (i.e., on novel data). Because different types of data are dependent on each other (for example, GO annotation, KEGG pathway membership, and MEDLINE abstracts) it becomes impossible to remove all cross-talk effects between data sources (e.g., removing MEDLINE data does not remove all information from the biomedical literature since much of it is present in GO and KEGG) to prevent contamination of the prediction of the disease gene by actual retrospective knowledge of this association. This makes it challenging to create benchmarks on retrospective data that are indicative of the performance of the method in an actual research setting. Next to benchmarking, some studies use several prioritization methods to analyze disease associated loci, mostly for type 2 diabetes and obesity (Tiffin et al., 2006; Elbers et al., 2007; Teber et al., 2009). However, the results have not been experimentally validated, which means that it is not possible to identify which methods made better predictions. Also, a few studies combine computational and experimental analysis: in silico generated hypothesis are then validated in vivo. We have, for instance, performed a computationally-supported genetic screen in Drosophila that led to the identification of 12 novel atonal genetic interactors (Aerts et al., 2009). Although useful, such studies often rely on the use of a single tool and therefore cannot be used to compare different approaches. They also give no indication of the performance of the method in general, but only illustrate it on a single well-validated case.

In this study, we aim at comparing the performance of several freely accessible web-based gene prioritization tools on novel data, which, to our knowledge, has never been performed before. To this aim, we selected recently reported disease gene associations from literature and use several gene prioritization tools to make predictions immediately after publication (typically within two days). Our approach relies on the fact that, when the prioritization tools are used, the novel disease gene association of interest is not yet included in the databases that underlie these tools. As a consequence, our approach mimics a novel discovery, and therefore the estimation of the performance is more accurate. It has to be mentioned that we compare tools and not the underlying algorithms (we see a tool as an algorithm plus some data sources), because this is what is most relevant to geneticists.

2 METHODS

2.1 Gene Prioritization tools

We aim at comparing the gene prioritization tools that can easily be used, and therefore only select the tools for which a free web-based implementation is available. The main objective is to assess the ability of the gene prioritization tools to predict potential novel disease genes which can then be experimentally validated. We have therefore not selected the tools whose ranking strategies depend exclusively on text as they would most likely work only when the novel disease gene was already considered a good candidate gene prior to discovery. One exception is Candid that also uses other data sources beside MEDLINE (e.g., protein domains, interactions, and expression data). In total, we have selected eight tools: Suspects (Adie et al., 2006), ToppGene (Chen et al., 2007), GeneDistiller (Seelow et al., 2008), GeneWanderer (Köhler et al., 2008), Posmed (Yoshida et al., 2009), Candid (Hutz et al., 2008), Endeavour (Aerts et al., 2006), and Pinta (Nitsch et al., 2010). The tools are run with their settings recommended by the developers. When applicable, multiple configurations are defined to explore several possibilities (for instance, several ranking algorithms within one tool). Originally, Pinta was developed to use expression data as input data, but here, we replace the continuous data (coming from expression data) with binary data using training genes: a *1* is inputed for each training gene, and a *0* is associated to the other genes. For an overview of the tools, please see Supplementary Table S1. All tools except Candid are used to prioritize a set of candidate genes (from a chromosomal region), and Candid is used to prioritize the whole genome. Pinta and Endeavour support both genome-wide and candidate set based prioritizations, and are used for both in this study (Endeavour-GW and Pinta-GW for genome-wide prioritization, Endeavour-CS and Pinta-CS for the candidate set prioritization). In addition, GeneWanderer can be run with up to four different ranking strategies (random walk, diffusion kernel, shortest path and direct interaction). We present the results for the first two strategies (GeneWanderer-RW for random walk, GeneWanderer-DK for diffusion kernel) since they have been showed to outperform the other two, simpler, approaches (Köhler et al., 2008) and since they can be efficiently used with many training genes. The performance of Posmed shows a strong dependency on the set of keywords used as an input and we ran it twice with different inputs. In the first run, we use the complete keyword set (Posmed-KS), and in the second, we only use the name of the disease (Posmed-DN). GeneDistiller is trained with both genes and keywords. These keywords are then used to find additional genes through the mining of OMIM, which in our case has less influence since OMIM is already used to derive the training genes. We therefore consider that GeneDistiller is trained with genes only. Candid is the only tool that can also be trained with disease specific tissues, when available, tissues relevant to the disease under study are used. Notice that Suspects went offline during our study after the 27th association and is not supported anymore (Euan Adie, personal communication), therefore, Suspects results are based on 27 associations over 42.

2.2 Validation data set

The validation data set is built by mining the scientific literature to identify the recently discovered disease-gene associations. This is achieved manually to avoid false positive associations. We select 6 journals that frequently publish papers that describe such associations: Nature Genetics, American Journal of Medical Genetics (part A / part B), Human Genetics, Human Molecular Genetics, and Human Mutation. We select all the novel disease-gene associations regardless of the disease under study, of the methodology used, and of whether the findings are confirmed or not. Novelty is assessed by using OMIM (McKusick, 1998), the Genetic Association Database (Becker et al., 2004), GoPubmed (Doms and Schroeder, 2005), and GeneCards (Safran et al., 2010). More precisely, we assess novelty at the gene level, and therefore novel mutations within already known genes are not considered. This process was kept active for 6 months (May 15 - November 15, 2010) and led to a collection of 42 associations (see Table 1

and Supplementary Table S2). For each association, the tools are run as soon as the association is identified following the defined workflow (see below). By doing this, we simulate as much as possible the prediction of a novel disease gene since the underlying databases are still unaware of the association. Once an association is identified, the exact inputs for the different tools have to be defined. For instance, ToppGene, GeneDistiller, GeneWanderer, Pinta and Endeavour require training genes (genes already known to be associated to the disease under study) whereas Suspects, Posmed, GeneDistiller and Candid require keywords that describe the disease. Training genes and keywords are collected from the corresponding OMIM pages, GAD pages and from recently published reviews when possible. BioMart (Haider et al., 2009) is used to map between gene symbols and tool specific gene identifiers (e.g., EntrezGene or Ensembl identifiers). As mentioned above, most of the tools require in addition a set of candidate genes (from the whole genome). Several tools accept chromosomal coordinates whereas some prefer cytogenetics bands. For each association, we select the cytogenetics bands that cover approximately 10Mb around the novel disease gene and derive the chromosomal coordinates. We choose 10Mb to obtain on average at least 100 candidate genes. Once again, BioMart is used to retrieve specific gene identifiers. For an overview of the inputs for the 42 associations, please see Supplementary Table S3.

The resulting 42 novel disease gene associations do not represent a homogeneous set. Therefore, we have divided them into confirmed (for monogenic diseases, the mutation is found in at least 2 unrelated patients; for multifactorial diseases, a GWAS is replicated in a separate cohort), intermediate (a single study, but additional functional evidence is provided), and unconfirmed (a single study) associations.

2.3 Performance measures

For each tool, we then assess its ability to identify the novel disease genes as promising genes using several statistical measures. We first compute the median of the rank ratio over all associations. We preferably use rank ratio over rank because tools do not necessarily return the same number of candidate genes even when fed with the same inputs. In addition, we also draw the boxplots of these rank ratios to give a more comprehensive view of the tool performance. Another method to compare the tools is to build the Receiver Operating Characteristic (ROC) curves, and to compute the Area Under the Curve (AUC) as an estimate of the global performance. To compare the tools even further, we computed the true positive rates when setting the threshold for validation at the top 5% (TPR in top 5% of candidates), 10% (TPR in top 10%) and 30% (TPR in top 30%). This is motivated by the fact that in a real situation, the number of candidate genes to assay often needs to be limited because of financial and time constraints. We have selected three thresholds that represent reasonable biological hypotheses, as we previously illustrated in a genetic screen (Aerts et al., 2009). The corresponding TPR measures are used to estimate how efficient the tools are if only the top 5%, 10%, or 30% candidate genes would be assayed. Notice that these values correspond to the shape of the lower end of the ROC curve (the sharper the curve, the higher the TPR). There are cases for which some tools are not able to identify the novel disease gene at all, we therefore include a response rate. It is defined as the percentage of associations for which each tool does return a prioritization result for the novel disease gene (in some cases a tool will not return any result, for example because it could not correctly map the gene identifier or some candidates are otherwise filtered out). For example, if one of the 42 disease genes could not be ranked (i.e., gene is missing), the response rate drops down to ~98% (41/42).

Lastly, we also derive a heat map to detect any correlation between tools by computing the pairwise cosine similarity of the rankings presented in Tables 2 (see Supplementary Figure S1).

Table 1. The validation data set consisting of 42 recently discovered disease gene associations.

Gene	Disease/phenotype	Reference(s)
HCCS	Congenital Diaphragmatic Hernia	Qidwai et al. (2010)
BRCA2	Bipolar Disorder	Tesli et al. (2010)
TNFRSF19	Nasopharyngeal carcinoma	Bei et al. (2010)
MECOM	Nasopharyngeal carcinoma	Bei et al. (2010)
ATF7IP	Testicular germ cell tumor	Turnbull et al. (2010)
DMRT1	Testicular germ cell tumor	Turnbull et al. (2010)
FUT2	Crohn's disease	McGovern et al. (2010)
CSF1R	Asthma	Shin et al. (2010)
GLI3	Metopic craniosynostosis	McDonald-McGinn et al. (2010)
STOM	Nonsyndromic cleft lip/palate	Letra et al. (2010)
UTRN	Arthrogryposis	Tabet et al. (2010)
GABRR1	Bipolar schizoaffective disorder	Green et al. (2010)
UBE2L3	Crohn's disease	Fransen et al. (2010)
BCL3	Crohn's disease	Fransen et al. (2010)
EZH2	Myelodysplastic syndromes	Nikoloski et al. (2010)
TRAF6	Parkinson's disease	Zucchelli et al. (2010)
IL10	Behcet's disease	Remmers et al. (2010); Mizuki et al. (2010)
DAB2IP	Abdominal aortic aneurysm	Gretarsdottir et al. (2010)
SPIB	Primary biliary cirrhosis	Liu et al. (2010)
MMEL1	Primary biliary cirrhosis	Hirschfield et al. (2010)
TBX2	Complex heart defect	Radio et al. (2010)
RUNX2	Single-suture craniosynostosis	Mefford et al. (2010)
CRHR1	Multiple sclerosis	Briggs et al. (2010)
IFNG	Leprosy	Cardoso et al. (2010)
SH2B1	Congenital Anomalies of the Kidney and Urinary Tract	Sampson et al. (2010)
DISP1	Congenital Diaphragmatic Hernia	Kantarci et al. (2010)
G6PC3	Dursun syndrome	Banka et al. (2010)
PQBP1	Periventricular heterotopia	Sheen et al. (2010)
CD320	Methylmalonic aciduria	Quadros et al. (2010)
CHST14	Ehlers-Danlos syndrome	Miyake et al. (2010)
PLCE1	Esophageal squamous cell carcinoma	Wang et al. (2010); Abnet et al. (2010)
C20orf54	Esophageal squamous cell carcinoma	Wang et al. (2010)
SDCCAG8	Retinal-renal ciliopathy	Otto et al. (2010)
TP63	Lung adenocarcinoma	Miki et al. (2010)
UBE2E2	Type 2 diabetes	Yamauchi et al. (2010)
LPP	Tetralogy of Fallot	Arrington et al. (2010)
RANBP1	Smooth pursuit eye movement abnormality	Cheong et al. (2011)
HTR7	Alcohol dependence	Zlojutro et al. (2010)
SOX17	Congenital anomalies of the kidney and the urinary tract	Gimelli et al. (2010)
ACAD9	Mitochondrial complex I deficiency	Haack et al. (2010)
TRAF3IP2	Psoriasis	Ellinghaus et al. (2010); Hüffmeier et al. (2010)
WDR62	Autosomal recessive primary microcephaly	Yu et al. (2010); Nicholas et al. (2010)

2.4 Integration of predictions

In order to get an estimate of the usefulness of a meta-predictor, the results of the different tools are combined using the Order Statistics as within Endeavour. Integration happens separately for the genome-wide tools and for the candidate set based tools, and tools that return only few rankings (Suspects and Posmed) were not included. For each experiment, the gene identifiers of the different tools are mapped using Biomart. In order to avoid getting artificially favorable rankings, the size of the merged ranking is set to the maximum size of the underlying rankings.

3 RESULTS

The overall ranking results of all gene prioritization tools are summarized in Table 2, the complete results are presented in Supplementary Tables S9 and S10. These results have also been added to the Gene Prioritization Portal (http://www.esat.kuleuven.be/gpp).

Table 2. Results for the genome-wide and candidate set based prioritization tools. (*) Values computed only on the first 27 associations.

	Median	Response rate	TPR in top 5%	TPR in top 10%	TPR in top 30%
Genome-wide prioritization tools					
Candid	18.10	100%	21.4%	33.3%	64.3%
Endeavour-GW	15.49	100%	28.6%	38.1%	71.4%
Pinta-GW	19.03	100%	26.2%	31.0%	71.4%
Integration	12.45	100%	19.1%	38.1%	78.6%
Candidate set based prioritization tools					
Suspects	12.77*	88.9%*	33.3%*	33.3%*	63.0%*
ToppGene	16.80	97.6%	35.7%	42.9%	52.4%
GeneWanderer-RW	22.10	95.2%	16.7%	26.2%	61.9%
GeneWanderer-DK	22.97	88.1%	11.9%	21.4%	52.4%
Posmed-DN	45.45	50.0%	4.7%	11.9%	23.8%
Posmed-KS	31.44	47.6%	4.7%	7.1%	23.8%
GeneDistiller	11.11	97.6%	26.2%	47.6%	78.6%
Endeavour-CS	11.16	100%	26.2%	42.9%	90.5%
Pinta-CS	18.87	100%	28.6%	31.0%	71.4%
Integration	6.99	100%	40.5%	57.1%	83.3%

3.1 Performance measures

When considering the median of the rank ratios, GeneDistiller, Endeavour-CS, and Suspects are the tools that perform the best on this benchmark (respectively 11.11, 11.16, and 12.77). They are followed by Endeavour-GW (15.49), ToppGene (16.8), Candid (18.1), Pinta-CS (18.87), Pinta-GW (19.03), GeneWanderer-RW (22.11), GeneWanderer-DK (22.97), Posmed-KS (31.44), and Posmed-DN (45.45). The boxplots presented in Figure 1 illustrate that both, GeneDistiller and Endeavour-CS perform better than the other candidate set based prioritization tools (Figure 1-right). Among the genome-wide tools, Endeavour-GW performs slightly better than Pinta-GW and Candid (Figure 1-left).

When considering the response rate, Endeavour (both modes), Candid, and Pinta (both modes) performed the best study with 100% closely followed by ToppGene, GeneDistiller, and GeneWanderer-RW with more than 95% (meaning that only one or two associations are missing). At the other hand of the spectrum, Posmed-KS and Posmed-DN only work for about half of the experiments in our benchmark (respectively 47.6% and 50%).

When we compare the tools based on the global AUC (see Figure 2), we observe that GeneDistiller appears as the best performing tool overall with an AUC of 86%. It is followed by Endeavour-CS (82%), Endeavour-GW (79%), Pinta-GW (77%), Suspects (76%), Pinta-CS (75%), Candid (73%), GeneWanderer-RW (71%), GeneWanderer- DK (67%), ToppGene (66%), Posmed-KS (58%), and Posmed-DN (56%). However, the ROC curves are in general intertwined meaning that none of the approaches is clearly performing better than the other. However, we postulate that, in our case, the most important section of the ROC curve is the beginning and therefore use three other measures, the true positive rates at 5%, at 10%, and at 30%. These measures indicate how efficient the tools would be if only the top candidate genes would be assayed.

Considering the TPR in top 10% and 30%, we can observe a similar trend. Indeed, at 10%, GeneDistiller is first with a rate of 47.6% (20 associations found over 42), followed by both ToppGene and Endeavour-CS with 42.9% (18 associations). However, at 30%, the best tool is Endeavour-CS (90.5% - 38 associations), followed by GeneDistiller (78.6% - 33 associations). The other tools show smaller TPR at both levels: Pinta-CS (31%, 71.4%), Suspects (33.3%, 63%), GeneWanderer-RW (26.2%, 61.9%), GeneWanderer-DK (21.4%, 52.4%), Posmed-KS (7.1%, 23.8%), and Posmed-DN (11.9%, 23.8%). Among the genome-wide prioritization tools, Endeavour-GW shows highest TPR in top 10% and 30% (38.1%, 71.4%), followed by Candid (33.3%, 64.3%) and Pinta-GW (31%, 71.4%).

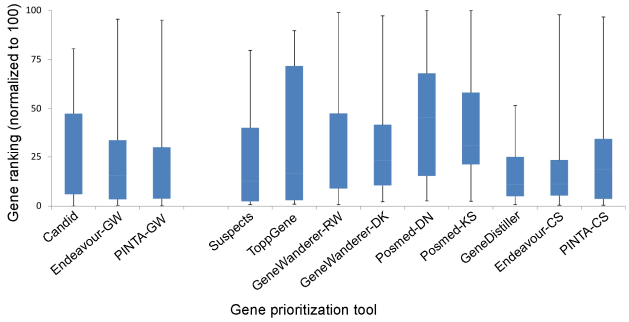


Figure 1: Boxplots of the 42 novel disease genes from the validation data set illustrated for the genome-wide (left) and candidate gene set based (right) prioritization tools.

3.2 Correlations

Supplementary Figure S1 shows the heat map of the novel disease gene ranking positions for all tools in this study. For the tools that have two modes (i.e., Posmed, GeneWanderer, Endeavour, Pinta), the two modes are highly correlated (> 0.89). There is also a significant correlation between Candid and GeneWanderer-DK (0.82). The other values are within 0.4 and 0.7, indicating that all tools are moderately correlated.

3.3 Integration of predictions

Our meta-analysis reveals that the best results are obtained when predictions are combined over the different tools (see Table 2 and Supplementary Table S11). For the genome-wide tools, all performance measures are improved by the integrative method (e.g.,

median of 12.45 for the meta-predictor versus 15.49 for Endeavour-GW). Similar results are obtained for the candidate set based tools (e.g., median of 6.99 for the meta-predictor versus 11.11 for GeneDistiller), although the TPR in the top 30% of the integrative method is still lower than for Endeavour-CS (83.3% versus 90.5%).

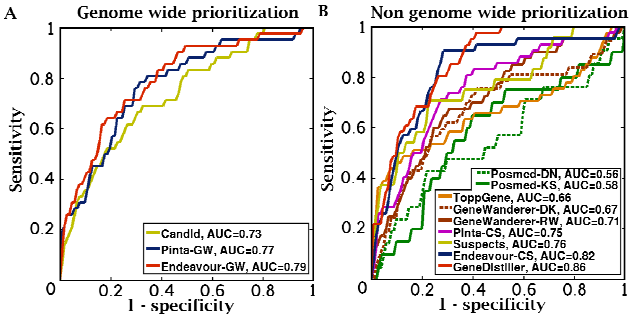


Figure 2: ROC curves of the genome-wide (A) and candidate gene set based (B) prioritization tools.

4 DISCUSSION

We aim at assessing the usefulness of eight gene prioritization tools that are freely available via web applications. We have built a validation based on 42 recently discovered disease-gene associations from literature containing novel genes for both monogenic conditions and complex disorders. We have selected novel disease-gene associations regardless of their strength, and of the underlying methodology. To mimic a real discovery, we have run the tools as soon as the article appeared online so that all databases used for gene prioritization are still not contaminated by the knowledge of the novel disease-gene association. This also means that we had to exclude tools that query MEDLINE online since their results would be biased.

We want to compare the performance of the tools even if the inputs are different (genes vs. keywords, genome-wide vs. candidate set). Among the eight gene prioritization tools that we have analyzed in this study, only Endeavour, Candid, and Pinta have been used for genome-wide prioritization. The input data for Endeavour and Pinta are training genes, whereas Candid requires keywords. The gene prioritization tools that we have used to prioritize candidate genes within a region of interest are Suspects, ToppGene, GeneWanderer, Posmed, GeneDistiller, and again Endeavour and Pinta. Suspects and Posmed are trained with keywords, the other tools require training genes. We have extensively searched through literature and dedicated databases to identify as many reliable training genes as possible for the disease of interest, as well as a set of appropriate keywords to derive fair and meaningful comparisons. However, different, and possibly better, results might be obtained by refining the inputs.

Our validation is too small to claim that the differences among the tools are significant. However, a trend can still be observed, GeneDistiller and Endeavour-CS consistently appear as the best tools when looking at all performance measures. It is interesting to notice that the best results are in general obtained with tools that use many data types in conjunction (up to eight for Endeavour, as compared to the three data sources used by Posmed), but there is no perfect correlation. This is in agreement with the conclusion of the recent review by Tiffin et al. (2009), who indicate that successful computational applications will be facilitated by improved data integration.

All tools except Posmed have a high response rate ranging from 88% to 100%, meaning that at least 37 of the 42 novel disease genes are prioritized (or 24 of 27 for Suspects). However, the response rates for Posmed-KS and Posmed-DN are respectively 47.6% and 50%, which can be explained by the fact that Posmed also acts as a filter on the candidate genes to obtain a reduced list of genes in the end. There are therefore cases for which the novel disease gene has been removed by the filter. This is different from the other tools for which missing genes basically correspond to genes that are not recognized by the tool (it happens most of the time with poorly characterized genes, such as C20orf54). Another special case is Suspects that went offline during the validation and therefore could only be validated with the first 27 associations. We therefore calculated the response rate only on the first 27 associations.

Two types of tools can be distinguished, the ones that are trained with already known genes and the ones that are trained with descriptive keywords. It appears that gene-based tools seem to work better than keyword-based tools (the average of medians is 17.2 for genes based tools and 27 for keyword based tools - similar results are obtained with the other measures, see Supplementary Table S8). This could be because we use in general more genes than keywords for training (18.8 genes on average for 6 keywords). This also indicates that more keywords might be needed to model a disease, a small text (such as an OMIM entry) might even be necessary (van Driel et al., 2006).

There is in general an agreement between the five performance measures we use throughout our study. One notable exception exists for ToppGene, whose AUC is 66%, and corresponds to rank 10th (out of the 12 prioritization tools). In contrast its associated TPR in top 10% is 42.9%, which corresponds to rank 2nd. This apparent contradiction can be explained by observing Figure 2, in which the ROC curve exhibits a non convex shape. This is because ToppGene either ranks the novel disease gene on top or at the bottom (i.e., the disease genes are rarely ranked in the middle). And therefore the TPR in top 10% will be high because it only takes into account the top of the list, while the AUC will be lower because it basically behaves like an average over all cases. Another important point is that our observations are in line with the 'no free lunch' theorem. Indeed each tool can perform better than all the others for some cases, or, in other words, none of the tools outperforms another on the complete data set (if we do not consider the special case of Posmed that also acts as a filter).

Posmed-KS has been trained with the complete keyword set, whereas Posmed-DN has been trained only with the disease name.

The median rank ratio is 31.44 when the complete keyword set is used and drops to 45.45 when only the disease name is inputted. If we only compare the results over the 19 associations for which both tools are able to prioritize the novel disease gene, the difference becomes even larger (29.6 and 50 respectively for Posmed-KS and Posmed-DN). Altogether, these results indicate that Posmed does not rely on the use of the single disease name and that the extra keywords are indeed important. It can be observed that the performance measures for Posmed are worse than for the other tools in our benchmark study. However, when looking at the individual ranks, it can be observed that Posmed returns far fewer genes than the other tools because it also acts as a filter. As a result, the rank ratios are on general larger and the performance

measures are therefore worse. As such, it becomes difficult to fairly compare Posmed to the other tools because our measures of performance naturally penalize the fact that Posmed returns prioritizations for a limited set of candidates. Changing our performance measures to counterbalance this effect would then give an unfair advantage to Posmed because it returns prioritizations only for the "safer bets".

GeneWanderer has also been run twice with different network algorithms: random walk (RW) and diffusion kernel (DK). The respective performance are very similar although the random walk approach is performing a little bit better than the diffusion kernel albeit non significant (22.11 to 22.97 for median rank ratio – similar differences are observed with the other measures). The heat map indicates a strong correlation (>0.9, see Supplementary Figure S1) between the two modes, which was expected since applying diffusion to a kernel can be interpreted as equivalent to applying a random walk on the underlying network. Altogether, this indicates that these two algorithms are similar.

Endeavour and Pinta are used to prioritize both the whole genome (Endeavour-GW and Pinta-GW) and the defined chromosomal region (Endeavour-CS and Pinta-CS) allowing us to identify the influence of the size of the gene list to prioritize. The median rank ratio is better for Endeavour-CS (11.16) than for Endeavour-GW (15.49) in our benchmark. The difference is smaller but remains when considering the AUC, and the TPR in top 10% and 30%.

The same training genes are used, and therefore the observed difference is only caused by extending the small candidate gene set to the whole genome. This confirms previous findings that prioritizing the whole genome is more difficult than prioritizing a rather small positive locus. The heat map indicates that the two Endeavour modes are strongly correlated as expected since the core algorithm is the same in both modes (>0.9, see Supplementary Figure S1). At contrary, the results for both Pinta modes are very similar (correlation of 0.99) and seem to indicate that the size of the candidate set does not influence this algorithm.

In this study, we consider the tools as off the shelf solutions, and use them as recommended by the developers without fine tuning of the parameters. However, an important feature that might influence the results is the date of the last data update. The latest genomic data (still prior to discoveries considered in this study) is likely to give the best results since it will model more accurately what is currently known, when compared to data that is two years old. In our setup, we have no control over the genomic data used and cannot identify if variation in performance among tools can be explained by this.

In addition, the quality of both the data sources and the integration methodologies are also influencing the outcome of the prioritization process. However, we aim at estimating the usefulness of some prioritization tools for geneticists. And therefore an in-depth comparison of the implementation of the tools is beyond the scope of this study.

It is important to notice that the 42 novel disease gene associations do not represent a very homogeneous set. Indeed, the median of the rank ratios over the tools show that some associations seem to be easier to predict than others. This also explains why all tools are moderately correlated on the heat map (> 0.4). A plausible explanation is the disparity in the available data between the novel disease genes. Since only little data can be gathered for poorly characterized genes, such as C20orf54, they are more difficult to prioritize. However, we also hypothesize that the nature of the underlying genetic disorder, as well as the quality of the reported association might influence the ability of the tools to predict correctly that association. We have therefore divided the associations between confirmed, intermediate, and unconfirmed. Among the 42 associations, 23 are confirmed, 8 are intermediate, and 11 are unconfirmed (see Supplementary Table S2). We hypothesize that this might influence our validation since some unconfirmed associations might in fact be spurious. We observe that Suspects and ToppGene perform better for the 23 confirmed associations than for the 19 unconfirmed ones (see Supplementary Tables S4 and S5). This trend is however not always shared as the situation is opposite for GeneDistiller and GeneWanderer. Although informative, these comparisons are not significant due to the small number of associations.

In our validation data set, there are 17 monogenic diseases and 25 multifactorial disorders (see Supplementary Tables S6 and S7). It has been shown that it is more difficult to make predictions for multifactorial diseases than for monogenic diseases (Linghu et al., 2009). Our results however seem to indicate that not all tools are influenced by the intrinsic complexity of multifactorial diseases. For instance, Endeavour and ToppGene seem to perform better for monogenic conditions while GeneWanderer and Suspects perform better for complex disorders. However, the size of our validation data set does not allow for a complete statistical analysis. Larger validation data sets and real predictive studies will be pursued to complement our preliminary study.

We are aware of the limited coverage of available literature in human genetics in our study that report novel disease-gene associations. However, we aimed at estimating the real performance of gene prioritization tools and therefore have decided to keep under strict control all the factors that could potentially bias the benchmark. We were further interested in finding novel disease gene associations for defining a proper benchmark, and there is no guarantee that these associations are uniformly distributed over the whole literature. We have used journals about genetic disorders in general and favor journals that report novel associations and have avoided specialized journals that focus on few diseases to avoid introducing bias towards one disease class. Our choice of the 6 selected journals may not be perfect, but they allowed us to cover most disease types and most situations.

Several studies have shown that combining predictions of several tools lead to even better predictions (Tiffin et al., 2006; Elbers et al., 2007). However, no performance criteria were used to select the tools to be combined. With this comparison of tools, we ease the selection of the most efficient tools, whose combination may lead to more accurate predictions. In addition, we report that the meta-predictors that integrate the predictions made by several tools perform better than the best individual tools as already reported (Thornblad et al., 2007).

Our results indicate that cross-validation based benchmarks tend to overestimate the real predictive performance. Indeed, all the tools for which such a benchmark exists have lower AUC than anticipated using our dataset (see Supplementary Table S12). We therefore believe that developers should take extra care when benchmarking their tools as to avoid these pitfalls. Also, some hard constraints have made this study small enough not to reach significance (e.g., only few tools have a programmatically queryable interface).

As already discussed in (Moreau et al., 2012), this field needs to consolidate through improved benchmarking efforts due to the lack of a ground truth for evaluating the performance of prioritization methods. Therefore we see a need for a large-scale community effort to compare multiple tools across common prospective benchmarks. We hope our work represents the first step towards a collaborative effort to tackle this problem at a larger scale.

ACKNOWLEDGEMENTS

We thank Peter Konings for his help regarding the statistics.

Funding:

Research Council KUL [CIF/07/02 DE CAUSMAE / DEFIS - SOCK, ProMeta, GOA Ambiorics, GOA MaNet, GOA 2006/12, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys, START 1, several PhD/postdoc and fellow grants]; Flemish Government [FWO: PhD/postdoc grants, projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR), IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3, FOD:Cancer plans, IBBT]; Belgian Federal Science Policy Office [IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011)]; EU-RTD [ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED].

REFERENCES

Abnet, C. C. et al., (2010). A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet, 42(9), 764–767.

Adie, E. A. et al., (2006). SUSPECTS: enabling fast and effective prioritization ofpositional candidates. Bioinformatics, 22(6), 773 –774.

Aerts, S. et al., (2006). Gene prioritization through genomic data fusion. Nat Biotech, 24(5), 537–544.

Aerts, S. et al., (2009). Integrating computational biology and forward genetics in drosophila. PLoS Genet, 5(1), e1000351.

Arrington, C. B. et al., (2010). Haploinsufficiency of the LIM domain containing preferred translocation partner in lipoma (LPP) gene in patients with tetralogy of fallot and VACTERL association. American Journal of Medical Genetics. Part A, 152A(11), 2919–2923. PMID: 20949626.

Banka, S. et al., (2010). Mutations in the G6PC3 gene cause dursun syndrome. American Journal of Medical Genetics. Part A, 152A(10), 2609–2611. PMID: 20799326.

Becker, K. G. et al., (2004). The genetic association database. Nat Genet, 36(5), 431–432.

Bei, J. et al., (2010). A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet, 42(7), 599–603.

Briggs, F. B. S. et al., (2010). Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12,566 individuals. Human Molecular Genetics, 19(21), 4286–4295. PMID: 20699326.

Calvo, S., et al. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet, 38(5):576-82.

Cardoso, C. C. et al., (2010). IFNG +874 TntextgreaterA single nucleotide polymorphism is associated with leprosy among brazilians. Human Genetics, 128(5), 481–490. PMID: 20714752.

Chen, J. et al., (2007). Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics, 8(1), 392.

Cheong, H. S. et al., (2011). Association of RANBP1 haplotype with smooth pursuit eye movement abnormality. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics, 156B(1), 67–71. PMID: 21184585.

Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Research, 33(Web Server issue), W783–786. PMID: 15980585.

Doncheva, N. T., et al., (2012). Recent approaches to the prioritization of candidate disease genes. WIREs Syst Biol Med. doi: 10.1002/wsbm.1177.

Elbers, C. C. et al., (2007). A strategy to search for common obesity and type 2 diabetes genes. Trends in Endocrinology and Metabolism: TEM, 18(1), 19–26. PMID: 17126559.

Ellinghaus, E. et al., (2010). Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat Genet, 42(11), 991–995.

Erlich, Y. et al., (2011). Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 21: 658-664.

Fransen, K. et al., (2010). Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for crohn's disease. Human Molecular Genetics, 19(17), 3482–3488. PMID: 20601676.

Gimelli, S. et al., (2010). Mutations in SOX17 are associated with congenital anomalies of the kidney and the urinary tract. Human Mutation, 31(12), 1352–1359. PMID: 20960469.

Green, E. K. et al., (2010). Variation at the GABAA receptor gene, rho 1 (GABRR1) associated with susceptibility to bipolar schizoaffective disorder. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics, 153B(7), 1347–1349. PMID: 20583128.

Gretarsdottir, S. et al., (2010). Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nature Genetics, 42(8), 692–697. PMID: 20622881.

Haack, T. B. et al., (2010). Exome sequencing identifies ACAD9 mutations as a cause of complex i deficiency. Nat Genet, 42(12), 1131–1134.

Haider, S. et al., (2009). BioMart central portal–unified access to biological data. Nucleic Acids Research, 37(Web Server), W23–W27.

Hardy, J. and Singleton, A. (2009). Genomewide association studies and human disease. The New England Journal of Medicine, 360(17), 1759–1768. PMID: 19369657.

Hirschfield, G. M. et al., (2010). Variants at IRF5-TNPO3, 17q12-21 and MMEL1 are associated with primary biliary cirrhosis. Nat Genet, 42(8), 655–657.

Hüffmeier, U. et al., (2010). Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. Nature Genetics, 42(11), 996–999. PMID: 20953186.

Hutz, J. E. et al., (2008). CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genetic Epidemiology, 32(8), 779–790. PMID: 18613097.

Kantarci, S. et al., (2010). Characterization of the chromosome 1q41q42.12 region, and the candidate gene DISP1, in patients with CDH. American Journal of Medical Genetics. Part A, 152A(10), 2493–2504. PMID: 20799323.

Köhler, S. et al., (2008). Walking the interactome for prioritization of candidate disease genes. American Journal of Human Genetics, 82(4), 949–958. PMID: 18371930.

Letra, A. et al., (2010). Follow-up association studies of chromosome region 9q and nonsyndromic cleft lip/palate. American Journal of Medical Genetics. Part A, 152A(7), 1701–1710. PMID: 20583170.

Linghu, B. et al., (2009). Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biology, 10(9), R91. PMID: 19728866.

Liu, X. et al., (2010). Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nature Genetics, 42(8), 658–660. PMID: 20639880.

Lupski, J. R. et al., (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. The New England Journal of Medicine, 362(13), 1181–1191. PMID: 20220177.

McDonald-McGinn, D. M. et al., (2010). Metopic craniosynostosis due to mutations in GLI3: a novel association. American Journal of Medical Genetics. Part A, 152A(7), 1654–1660. PMID: 20583172.

McGovern, D. P. B. et al., (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with crohn's disease. Human Molecular Genetics, 19(17), 3468–3476. PMID: 20570966.

McKusick, V. A. (1998). Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders. The Johns Hopkins University Press, 12th edition.

Mefford, H. C. et al., (2010). Copy number variation analysis in single-suture craniosynostosis: multiple rare variants including RUNX2 duplication in two cousins with metopic craniosynostosis. American Journal of Medical Genetics. Part A, 152A(9), 2203–2210. PMID: 20683987.

Miki, D. et al., (2010). Variation in TP63 is associated with lung adenocarcinoma susceptibility in japanese and korean populations. Nat Genet, 42(10), 893–896.

Miyake, N. et al., (2010). Loss-of-function mutations of CHST14 in a new type of Ehlers-Danlos syndrome. Human Mutation, 31(8), 966–974. PMID: 20533528.

Mizuki, N. et al., (2010). Genome-wide association studies identify IL23R-IL12RB2 and IL10 as behc¸et's disease susceptibility loci. Nature Genetics, 42(8), 703–706. PMID: 20622879.

Moreau Y., et al. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discover. Nat Rev Genet.13(8):523-36.

Nicholas, A. K. et al., (2010). WDR62 is associated with the spindle pole and is mutated in human microcephaly. Nat Genet, 42(11), 1010–1014.

Nikoloski, G. et al., (2010). Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. Nat Genet, 42(8), 665–667.

Nitsch, D. et al., (2010). Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC Bioinformatics, 11(1), 460.

Oti, M. (2011). Web tools for the prioritization of candidate disease genes. Methods Mol Biol.760:189-206.

Otto, E. A. et al., (2010). Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. Nat Genet, 42(10), 840–850.

Piro, R. M. et al., (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS Journal, 279: 678–696.

Qidwai, K. et al., (2010). Deletions of xp provide evidence for the role of holocytochrome c-type synthase (HCCS) in congenital diaphragmatic hernia. American Journal of Medical Genetics. Part A, 152A(6), 1588–1590. PMID: 20503342.

Quadros, E. V. et al., (2010). Positive newborn screen for methylmalonic aciduria identifies the first mutation in TCblR/CD320, the gene for cellular uptake of transcobalamin-bound vitamin b(12). Human Mutation, 31(8), 924–929. PMID:20524213.

Radio, F. C. et al., (2010). TBX2 gene duplication associated with complex heart defect and skeletal malformations. American Journal of Medical Genetics. Part A, 152A(8), 2061–2066. PMID: 20635360.

Rajab, A., et al. (2010). Fatal Cardiac Arrhythmia and Long-QT Syndrome in a New Form of Congenital Generalized Lipodystrophy with Muscle Rippling (CGL4) Due to *PTRF-CAVIN* Mutations. PLoS Genet 6(3): e1000874.

Remmers, E. F. et al., (2010). Genome-wide association study identifies variants in the MHC class i, IL10, and IL23R-IL12RB2 regions associated with behc¸et's disease. Nat Genet, 42(8), 698–702.

Safran, M. et al., (2010). GeneCards version 3: the human gene integrator. Database: The Journal of Biological Databases and Curation, 2010, baq020. PMID: 20689021.

Sampson, M. G. et al., (2010). Evidence for a recurrent microdeletion at chromosome 16p11.2 associated with congenital anomalies of the kidney and urinary tract (CAKUT) and hirschsprung disease. American Journal of Medical Genetics. Part A, 152A(10), 2618–2622. PMID: 20799338.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. Nat Meth, 5(1), 16–18.

Seelow, D. et al., (2008). GeneDistiller—Distilling candidate genes from linkage intervals. PLoS ONE, 3(12), e3874.

Sheen, V. L. et al., (2010). Mutation in PQBP1 is associated with periventricular heterotopia. American Journal of Medical Genetics. Part A, 152A(11), 2888–2890. PMID: 20886605.

Shin, E. K. et al., (2010). Association between colony-stimulating factor 1 receptor gene polymorphisms and asthma risk. Human Genetics, 128(3), 293–302. PMID: 20574656.

Tabet, A. et al., (2010). Molecular characterization of a de novo 6q24.2q25.3 duplication interrupting UTRN in a patient with arthrogryposis. American Journal of Medical Genetics Part A, 152A(7), 1781–1788.

Teber, E. T. et al., (2009). Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies. BMC Bioinformatics, 10 Suppl 1, S69. PMID: 19208173.

Tesli, M. et al., (2010). Association analysis of PALB2 and BRCA2 in bipolar disorder and schizophrenia in a scandinavian case-control sample. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics, 153B(7), 1276–1282. PMID: 20872766.

Thienpont B. et al., (2010). Haploinsufficiency of TAB2 causes congenital heart defects in humans. Am. J. Hum. Genet. 86(6): 839-849.

Thornblad, T. A. et al., (2007). Prioritization of positional candidate genes using multiple web-based software tools. Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies, 10(6), 861–870. PMID: 18179399.

Tiffin, N. et al., (2006). Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Research, 34(10), 3067–3081. PMID: 16757574

Tiffin, N. et al., (2009). Linking genes to diseases: it's all in the data. Genome Medicine, 1(8), 77. PMID: 19678910.

Tiffin, N. (2011). Conceptual thinking for in silico prioritization of candidate disease genes. Methods Mol Biol.760:175-87.

Tranchevent, L. et al., (2010). A guide to web tools to prioritize candidate genes. Briefings in Bioinformatics.

Turnbull, C. et al., (2010). Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. Nature Genetics, 42(7), 604–607.

van Driel, M. A. et al., (2006). A text-mining analysis of the human phenome. European Journal of Human Genetics: EJHG, 14(5), 535–542. PMID: 16493445.

Vliet-Ostaptchouk, J. V. et al., (2008). HHEX gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort, Eur J Hum Genet. 16, 652–656.

Wang, L. et al., (2010). Genome-wide association study of esophageal squamous cell carcinoma in chinese subjects identifies susceptibility loci at PLCE1 and c20orf54. Nat Genet, 42(9), 759–763.

Yamauchi, T. et al., (2010). A genome-wide association study in the japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4AC2CD4B. Nat Genet, 42(10), 864–868.

Yoshida, Y. et al., (2009). PosMed (Positional medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. Nucleic Acids Research, 37(Web Server issue), W147–152. PMID: 19468046.

Yu, T. W. et al., (2010). Mutations in WDR62, encoding a centrosome-associated protein, cause microcephaly with simplified gyri and abnormal cortical architecture. Nat Genet, 42(11), 1015–1020.

Zlojutro, M. et al., (2010). Genome-wide association study of theta band event-related oscillations identifies serotonin receptor gene HTR7 influencing risk of alcohol dependence. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics. PMID: 21046636.

Zucchelli, S. et al., (2010). TRAF6 promotes atypical ubiquitination of mutant DJ-1 and alpha-synuclein and is localized to lewy bodies in sporadic parkinson's disease brains. Human Molecular Genetics, 19(19), 3759–3770. PMID: 20634198.