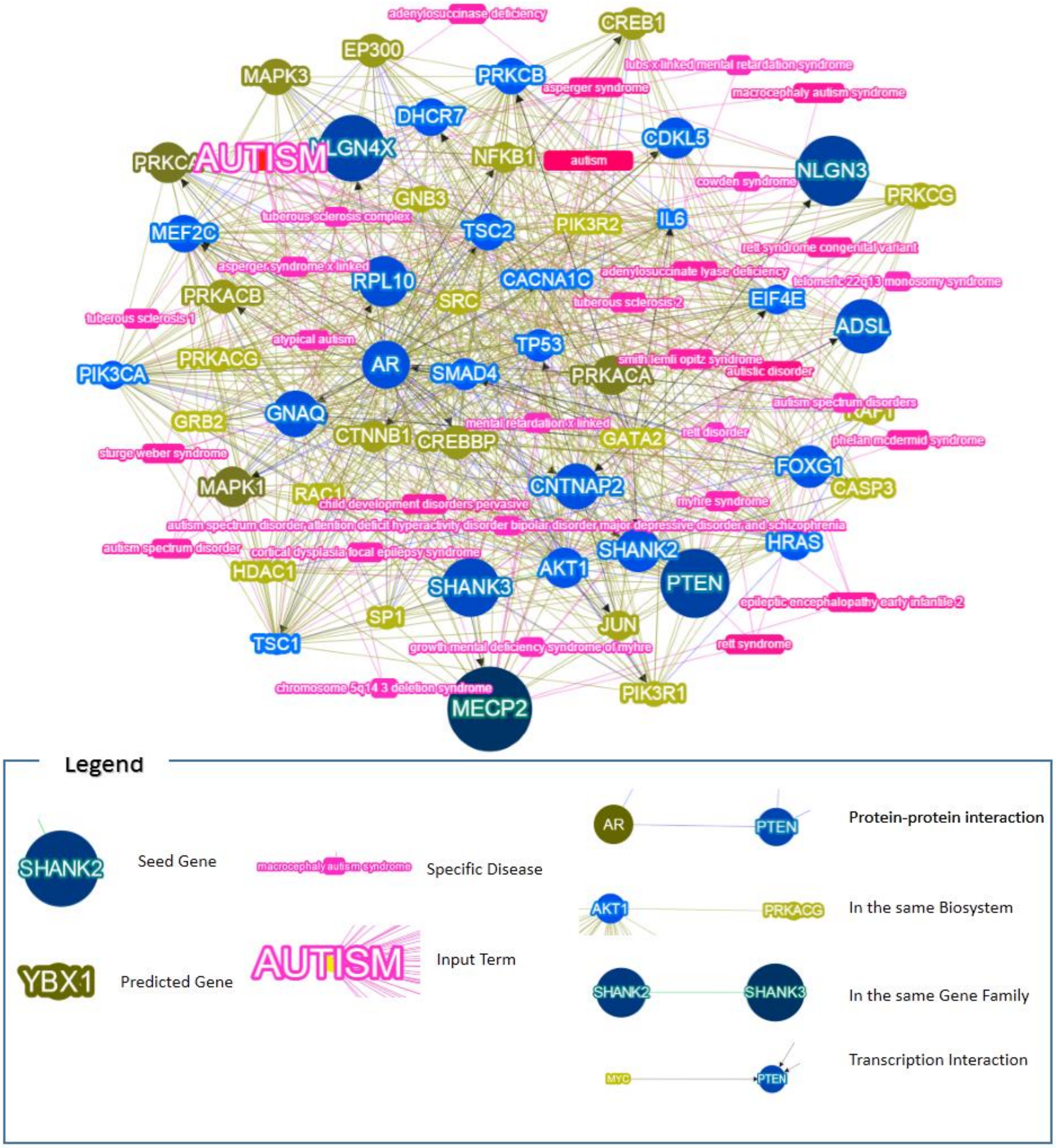
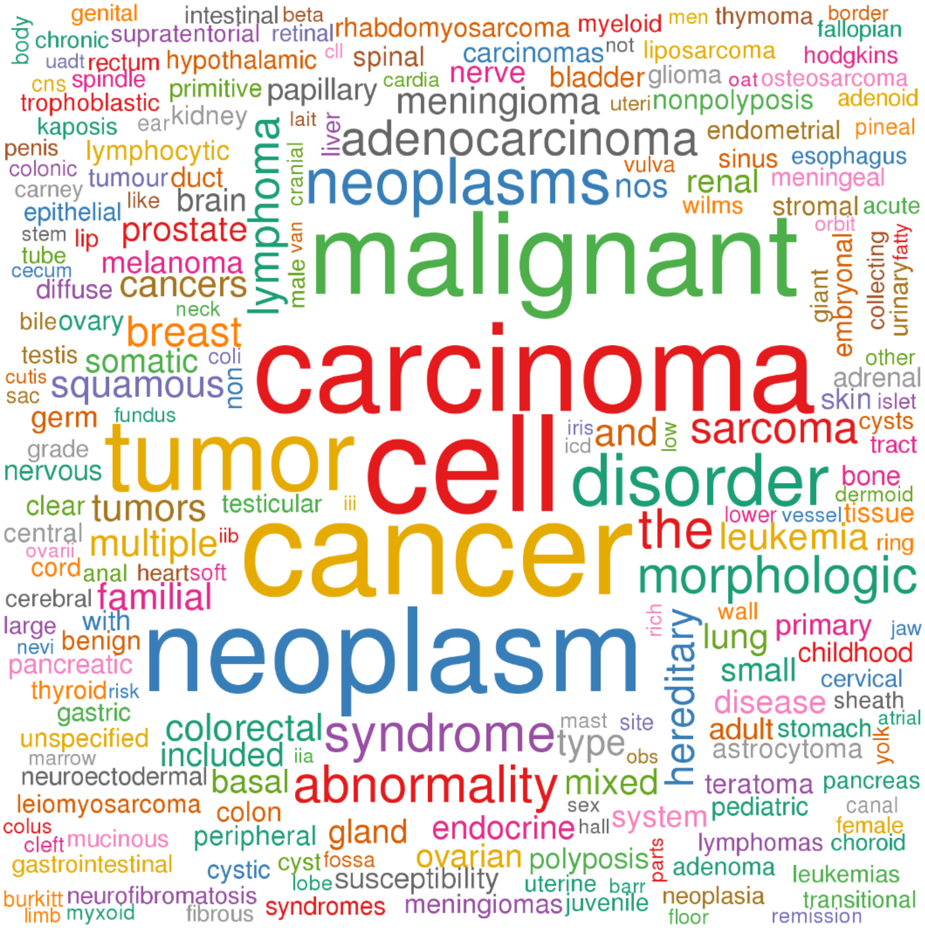

adenylosuccinase deficiency
CREB1
EP300
MAPK3
PRKCB
lubs x linked mental retardation syndrome
asperger syndrome
macrocephaly autism syndrome
DHCR7
PRKCA
AUTISM
NLGN4X
NFKB1
autism
CDKL5
NLGN3
cowden syndrome
GNB3
PRKCG
tuberous sclerosis complex
MEF2C
TSC2
PIK3R2
IL6
rett syndrome congenital variant
asperger syndrome x linked
RPL10
CACNA1C
adenylosuccinate lyase deficiency
telomeric 22q13 monosomy syndrome
PRKACB
SRC
tuberous sclerosis 2
EIF4E
ADSL
tuberous sclerosis 1
atypical autism
TP53
PIK3CA
PRKACG
AR
SMAD4
smith lemli opitz syndrome
PRKACA
autistic disorder
autism spectrum disorders
RAP1
GRB2
GNAQ
CTNNB1
CREBBP
mental retardation x linked
GATA2
rett disorder
phelan mcdermid syndrome
sturge weber syndrome
FOXG1
MAPK1
RAC1
CNTNAP2
CASP3
child development disorders pervasive
myhre syndrome
autism spectrum disorder attention deficit hyperactivity disorder bipolar disorder major depressive disorder and schizophrenia
autism spectrum disorder
cortical dysplasia focal epilepsy syndrome
SHANK2
HRAS
HDAC1
AKT1
PTEN
SHANK3
epileptic encephalopathy early infantile 2
SP1
JUN
TSC1
growth mental deficiency syndrome of myhre
rett syndrome
chromosome 5q14 3 deletion syndrome
MECP2
PIK3R1
Legend
SHANK2
Seed Gene
macrocephaly autism syndrome
Specific Disease
YBX1
Predicted Gene
AUTISM
Input Term
AR
PTEN
Protein-protein interaction
AKT1
PRKACG
In the same Biosystem
SHANK2
SHANK3
In the same Gene Family
MYC
PTEN
Transcription Interaction

Supplementary Figure 1

An example of the gene-term-disease network automatically generated by Phenolyzer web server for 'autism'.

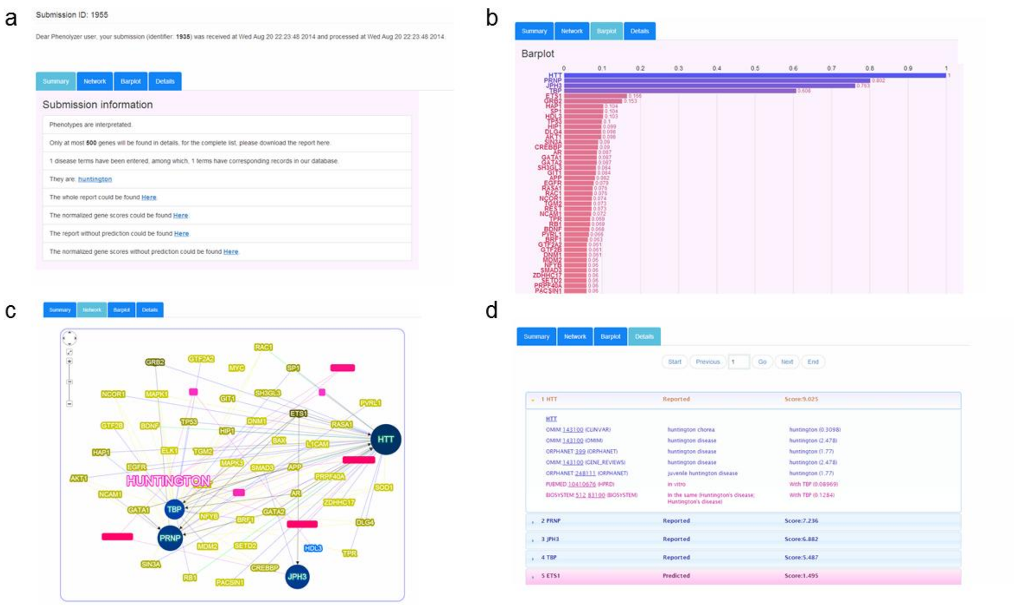
The largest word represents the user's input term, 'Autism'. The pink round rectangles represent disease names corresponding to each term. The balls represent genes. The darker the color, the more the node contributes. The longer the round rectangles, the more a disease name contributes. The larger the ball, the more significantly a gene is related with the term. A reported gene is blue and a predicted gene is yellow. Four types of edges correspond to four different types of gene-gene relationships as illustrated in the legend. The figure can be zoomed in by mouse operations in the web server to facilitate closer examination of the edge types.



Supplementary Figure 2

The wordcloud for all the interpreted names related to 'Cancer'.

This is the wordcloud generated by Phenolyzer corresponding to the input term 'cancer'. Some of the most frequently occurring words include 'cancer', 'neoplasm', 'cell', 'carcinoma', 'tumor' and 'malignant'.



Submission ID: 1955

Dear Phenolyzer user, your submission (identifier: **1935**) was received at Wed Aug 20 22:23:48 2014 and processed at Wed Aug 20 22:23:48 2014

Submission information

Phenotypes are interpretated.

Only at most **500** genes will be found in details, for the complete list, please download the report here.

1 disease terms have been entered, among which, 1 terms have corresponding records in our database.

They are: huntington

The whole report could be found Here

The normalized gene scores could be found Here

The report without prediction could be found Here

The normalized gene scores without prediction could be found Here

1 HTT Reported Score:9.025

HTT

OMIM:143100 (CLINVAR)	huntington chorea	huntington (0.3098)
OMIM:143100 (OMIM)	huntington disease	huntington (2.478)
ORPHANET:399 (ORPHANET)	huntington disease	huntington (1.77)
OMIM:143100 (GENE_REVIEWS)	huntington disease	huntington (2.478)
ORPHANET:248111 (ORPHANET)	juvenile huntington disease	huntington (1.77)
PUBMED:10410676 (HPRD)	in vitro	With TBP (0.08969)
BIOSYSTEM:512 83100 (BIOSYSTEM)	In the same (Huntington's disease; Huntington's disease)	With TBP (0.1284)

2 PRNP Reported Score:7.236

3 JPH3 Reported Score:6.882

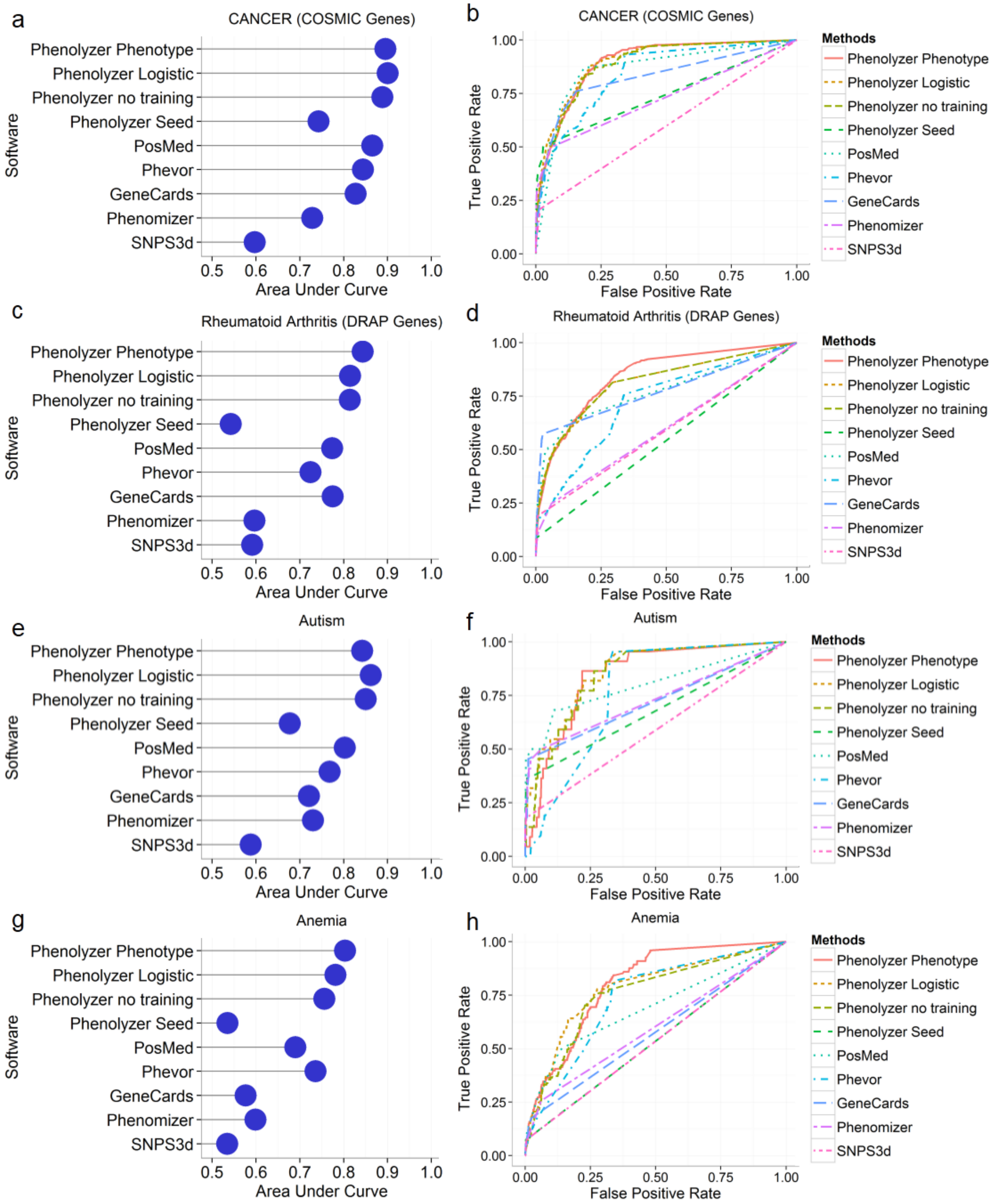
4 TBP Reported Score:5.487

5 ETS1 Predicted Score:1.495

Supplementary Figure 3

A snapshot of the output from the Phenolyzer web server.

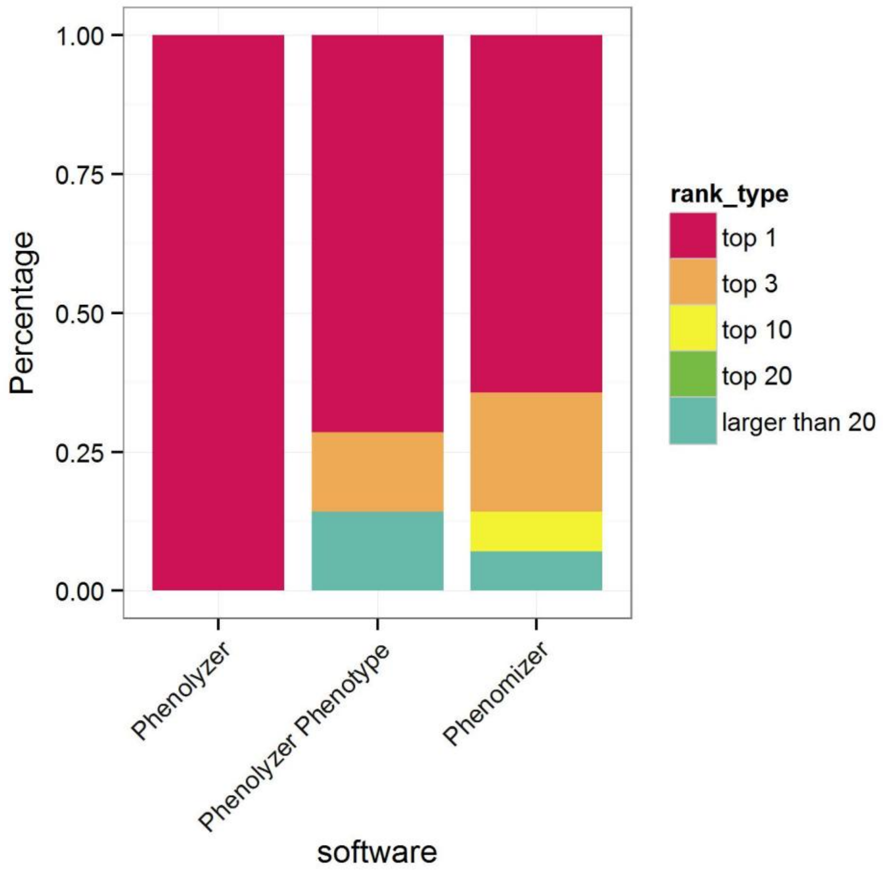
(a) ‘Summary’ includes **the** link to wordcloud, as well as the input settings and **the** links to output files for download. (b) ‘Barplot’ is a bar plot on at most 500 highest-ranked genes with normalized scores. (c) ‘Network’ is the interactive gene-disease-term network, with controlling buttons below it. (d) ‘Details’ shows how the score for each gene is calculated, including the links to each publication or database website.



Supplementary Figure 4

Comparison between Phenolyzer and other tools to find disease genes for cancer, rheumatoid arthritis, autism and anemia.

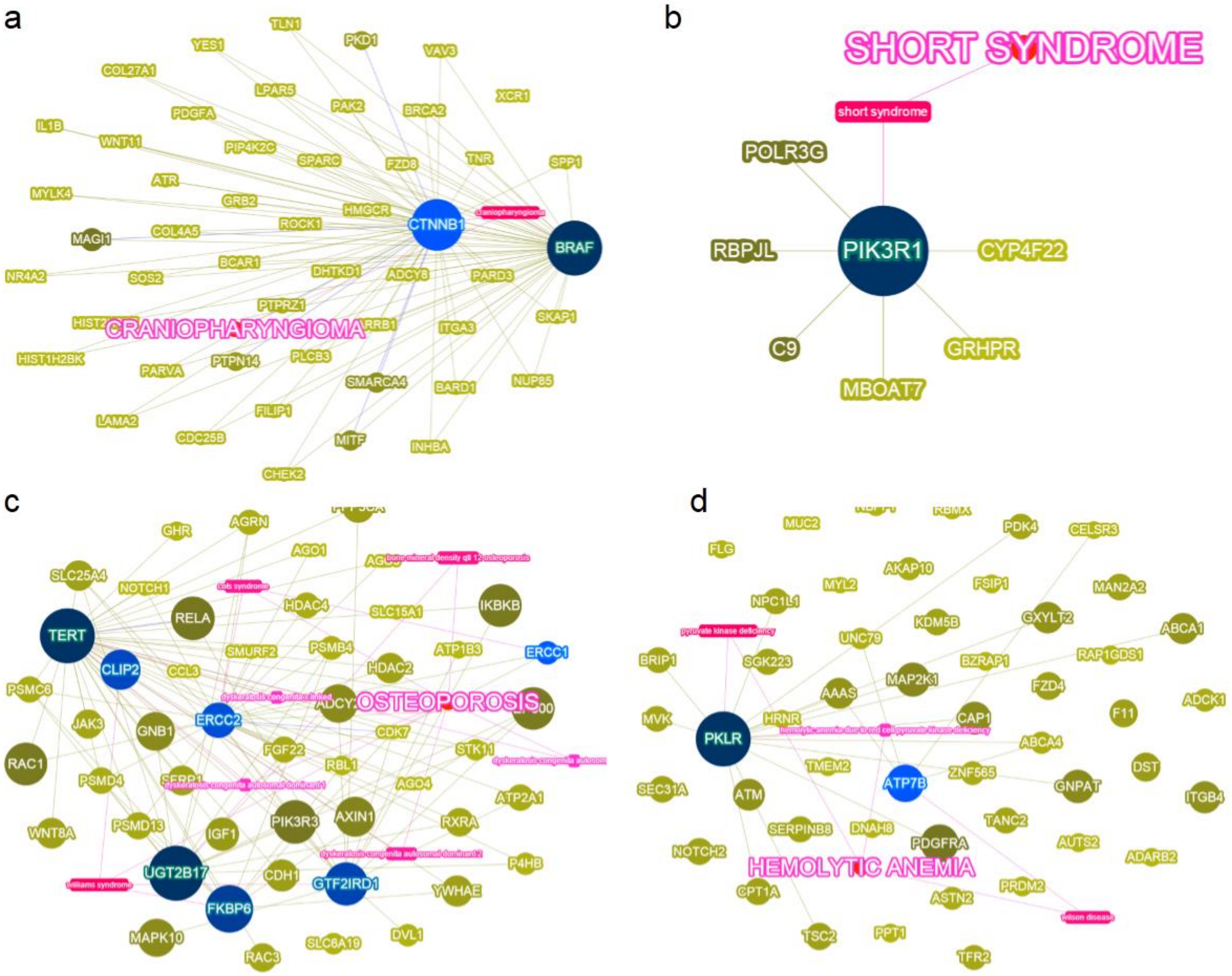
The AUC and ROC curve plot showing the performance comparison between Phenolyzer and other tools, on four gene sets of different complex diseases. (a, c, e and g) For each software, the AUC is calculated as the area under the ROC curve. (b, d, f and h) The ROC curve is plotted as True Positive Rate versus False Positive Rate. ‘Phenolyzer Phenotype’ is the Phenolyzer results with phenotype terms as input (the same input as Phenomizer). ‘Phenolyzer Logistic’ is Phenolyzer with weights trained with Logistic Regression model, compared with ‘Phenolyzer no training’. ‘Phenolyzer Seed’ is Phenolyzer’s seed gene result without the seed gene growth step, thus only representing the genes found in Phenolyzer’s disease-gene mapping knowledgebase.



Supplementary Figure 5

Evaluation of Phenolyzer by phenotype terms as input.

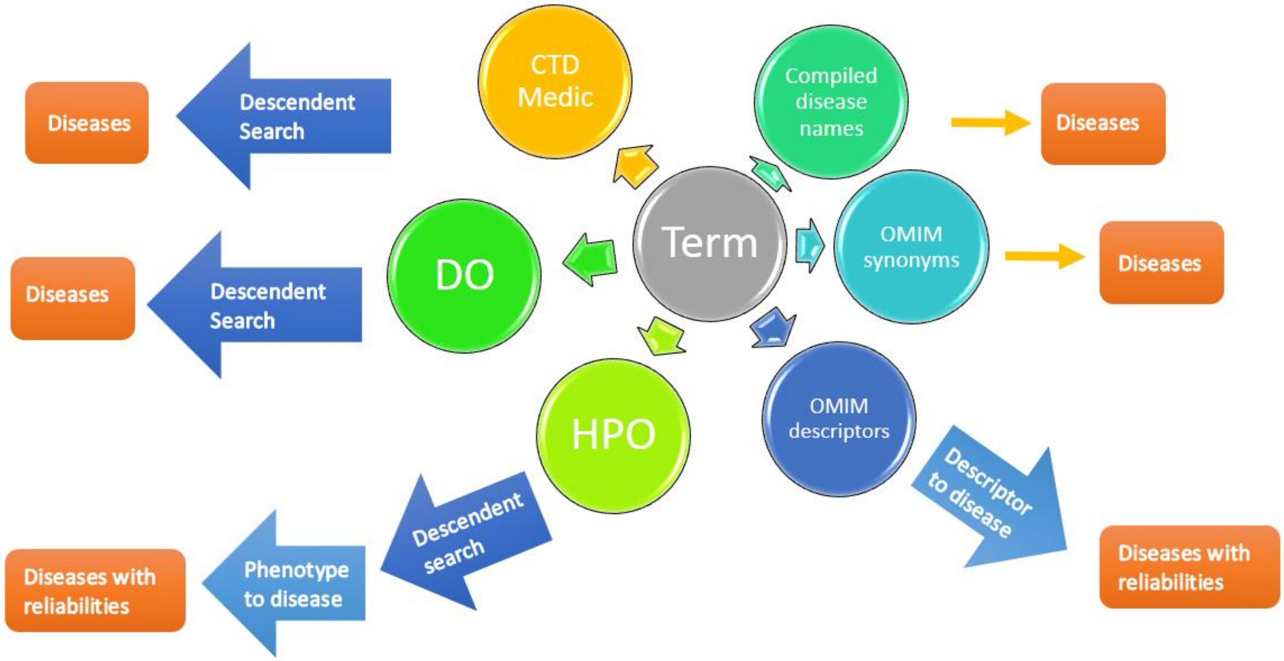
Phenolyzer with phenotype terms (rather than disease names) as input is able to prioritize most genes as 'Top 1' for the 14 monogenic diseases, which is similar as Phenomizer.



Supplementary Figure 6

Phenolyzer's results on four case studies.

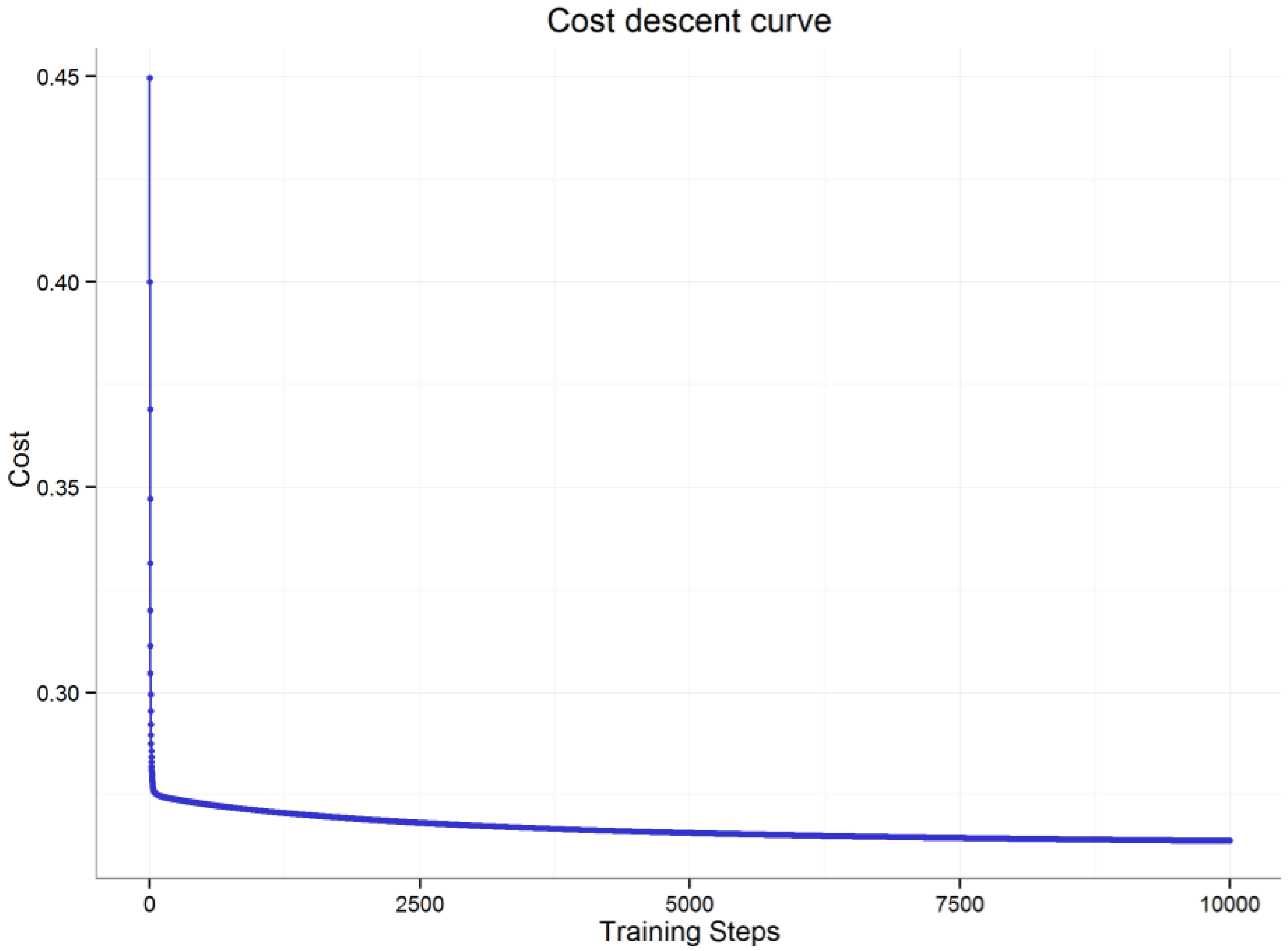
(a and b) The candidate gene lists generated from two studies on 'Craniopharyngiomas' and 'SHORT syndrome' were used as input into Phenolyer. The network plot shows that *BRAF* and *PIK3R1* are the genes with the highest scores corresponding to each disease separately. (c) For the CNV study of 'Osteoporosis', the generated significant CNV regions were used as input, and the Phenolyzer network successfully identified the correct gene, *UGT2B17*. (d) Combined with wANNOVAR, we first filtered the variants into a small list, then included all the genes in the variant list as the input into Phenolyzer. The correct gene *PKLR* was identified as the top gene for 'hemolytic anemia'.



Supplementary Figure 7

Illustration of the disease or phenotype term interpretation process.

The term is first processed through a word match to several different data source, DO (Disease Ontology), CTD Medic disease ontology vocabulary, HPO (Human Phenotype Ontology), OMIM synonym, OMIM descriptors, and Phenolyzer's compiled disease vocabulary. After the first match, the disease names are directly returned for Phenolyzer's compiled disease names and OMIM synonyms. For DO and CTD, an ontology search will retrieve all the descendent disease names and synonyms. For OMIM descriptors, they are mapped into OMIM diseases with a conditional probability as reliability. For HPO, an ontology search first finds all the descendent phenotypes, then the phenotypes are mapped into diseases with reliabilities.



Supplementary Figure 8

Learning curve through the training with Gradient Descent Algorithm.

The gradient descent algorithm iteratively reduces the cost and changes weight vector *w* into the direction opposite to the gradient of the cost function. After 10,000 iterations, with learning rate at 1, the slope of the curve is close to 0 and demonstrates that the number of learning steps is sufficient.

Supplementary Note

The impact of input tokenization on Phenolyzer

Based on the implementation of Phenolyzer, if a long disease name is tokenized into multiple short terms, the possibility of a match for each short term is larger than the original long term; as a result, the number of genes in the result should be increased and the recall rate may be improved. However, the cost associated with the improved recall rate is that the precision may be affected. To demonstrate this, we thoroughly tokenized 13 disease names from the 14 Mendelian disease names used in the manuscript; for example, 'Spermatogenic failure, nonobstructive, Y-linked' is tokenized into 'Spermatogenic;failure;nonobstructive;Y;linked', with each single word being treated as a single term. Our result shows that thorough tokenization increased the number of returned genes from 2,809 to 22,222 (median) and from 3,287 to 21,286 (mean), at a cost of incorrectly prioritizing 3 genes out of 15 genes (**Supplementary Dataset 1**).

Case study on exome sequencing and copy number variation data

To better illustrate the power of Phenolyzer, below we present three case studies on how to use Phenolyzer to analyze exome-sequencing data and copy number variation (CNV) data and identify disease causal genes.

One group identified that a *BRAF* gene mutation drives growth of papillary craniopharyngiomas and targeted genotyping identified this mutation in all papillary craniopharyngiomas[1]. They first analyzed the whole-exome sequencing data from a cohort of adamantinomatous (n=12) and papillary craniopharyngiomas (n=3) and identified a small number of non-synonymous somatic mutations in both subtypes, including 490 variants within 444 genes. This gene list was supplied into Phenolyzer, together with the disease term ‘craniopharyngiomas’ (**Supplementary Fig. 6**). *BRAF* shows up as the top gene. We also evaluated other competing tools, with the same term as input, and restricted results by the same input gene list. PosMed ranked *BRAF* as 6th. Genecards and SNPs3d returned a result list without the *BRAF* gene. As Phevor does not include the record for disease term ‘craniopharyngiomas’, we used ‘craniopharyngioma’ and the *BRAF* gene is ranked as 321st (see details in **Supplementary Table 6**).

Another group identified a *PI3KR1* mutation underlying SHORT syndrome and verified its functional impact in fibroblasts[2]. Whole exome-sequencing was conducted in 6 individuals from two families affected by SHORT syndrome. After variant calling and filtering, 22 variants within 22 genes were left in the candidate list. This gene list was then provided as input into Phenolyzer, together with the term ‘SHORT syndrome’. In Phenolyzer’s network, it is easily seen that *PI3KR1* is the most significant gene (**Supplementary Fig. 6**). After gene selection, PosMed can also rank *PIK3R1* as top 1. Genecards ranked *PIK3R1* as 2nd. SNPs3d does not have *PIK3R1* in the output. Phevor had no record for term ‘SHORT syndrome’ at all. In addition, we also tried ‘partial lipodystrophy’, ‘low body mass index’, ‘short stature’, ‘progeroid face’, and ‘Rieger anomaly’ (the phenotype descriptions in the original paper for SHORT syndrome) as input, and *PIK3R1* is still ranked as the ‘top1’ gene by Phenolyzer. We used ‘Lipodystrophy ‘,’ Weight loss ‘, ‘Progeroid facial appearance ‘,‘Rieger anomaly’,’ Short stature’ as HPO terms for Phevor and *PIK3R1* is also ranked as top 1 (See details in **Supplementary Dataset 7**).

In the third example, a CNV study identified a CNV affecting the gene *UGT2B17* as a contributing factor for osteoporosis[3]. Case-control genome-wide CNV studies were conducted using Affymetrix Human Mapping 500K Arrays, with the cohort of 350 Han Chinese individuals with a history of hip osteoporosis and 350 healthy matched controls. 727 significant CNVs were called from this study. These regions were used as input into Phenolyzer, together with the disease term ‘osteoporosis’. From the Phenolyzer’s network plot, *UGT2B17* shows up as the top gene (**Supplementary Fig. 6**). We also evaluated other competing tools, by first compiling a gene list from these CNVs (1,510 genes in total). For PosMed, *UGT2B17* ranked at 4th. For Genecards, *UGT2B17* ranked at 3rd. For SNPs3d and Phevor, the final gene list did not include *UGT2B17* (See details in **Supplementary Dataset 8**).

Pipeline between wANNOVAR and Phenolyzer

To facilitate users with whole genome or exome sequencing data rather than just a list of genes, we also implemented a wANNOVAR-Phenolyzer pipeline for analysis on VCF files at http://wannovar.usc.edu[4]. Disease or phenotype terms are accepted as optional input fields here, then Phenolyzer is called to prioritize candidate genes directly from wANNOVAR output. In the wANNOVAR result page, the gene list prioritized by Phenolyzer can be directly retrieved. Additionally, the link to the Phenolyzer result page is also available as the ‘Network Visualization’ link.

For example, we previously reported an exome sequencing study identifying a mutation in *PKLR* as "unrelated finding" in a patient with hemolytic anemia, through a study originally designed to uncover the genetic basis of attention deficit and hyperactivity disorder (ADHD)[5]. The VCF file is used as the input into wANNOVAR, with ‘rare recessive Mendelian disease' selected as disease model. In total, 87 variants were left after the filtration, whose corresponding genes are then submitted automatically as input into Phenolyzer together with the term ‘anemia’ or ‘hemolytic anemia’, by wANNOVAR. From the result network, the *PKLR* gene is ranked top with the term ‘hemolytic anemia’ (**Supplementary Fig. 6**). However, with the term ‘anemia’, *PKLR* ranked third. Thus, integrating wANNOVAR and Phenolyzer in the same pipeline can facilitate and expedite identifying disease causing variants, yet more refined disease terms can further improve the accuracy of disease gene finding, when the disease can be caused by multiple genes.

Additional Discussions

Phenolyzer follows a strictly defined procedure including four steps - term interpretation, seed gene generation, seed gene growth and data integration. For term interpretation, at least a word match is needed to process the term, and the possibility to match the terms depends on the vocabularies of several different disease ontology systems, the synonyms provided by OMIM, and phenotype descriptors provided by the Human Phenotype Ontology. Some improvements can be made in the future to better explore all possible terms, including integrating several medical dictionaries and standardized vocabularies, such as Webster's medical dictionary, ICD9 and ICD10 vocabulary.

For the seed gene generation step, other databases will be incorporated in the future, such as DECIPHER[6]. We note that several heuristics are used in this step (Gene Score System), to translate association information and publication count into scores. The underlying logic is that the translated score is normalized proportionally to the rank percentage. For the seed gene growth step, similar heuristics are also used (Gene Prediction Score System). Despite the use of heuristics, our results demonstrated the effectiveness of the approach. To facilitate integration of gene-disease and gene-gene relationship databases, we now made an add-on system for the Phenolyzer command-line tool. As long as the user compile their own databases into a pre-defined format, it is straightforward to add databases into the Phenolyzer system. To demonstrate this, we integrated Mentha protein interaction database[7], the Genetic Association Database[8], DisGeNet[9] and the Genecards database into our web server as an option, with an *ad hoc* weight.

The Logistic Regression with Gradient Descent method is designed to optimize weights for integrating the scores of seed genes and different relationship types. We examined how the performance will change if we instead assign *ad hoc* weights to each feature (**Supplementary Fig. 4**). Although the logistic regression model is trained from four diseases (type 1 and 2 diabetes, coronary artery disease and Crohn’s disease) that are totally different from the test set, the improvement over "no training" is still significant. Therefore, we used the optimal weights rather than arbitrarily selected weights in our software. In addition, we also provide the ‘Weight Adjust’ option for users to apply their own arbitrary weights, or even turn off a database by setting the weight to zero.

For the tool comparison, we noticed that inconsistent results appeared between monogenic diseases and complex diseases. We think the reason is that the monogenic disease and complex disease comparisons are two different types of comparisons. In monogenic disease comparisons, as long as the tool has the knowledge of the gene-disease mapping in a database, and can retrieve the record and prioritize it as top one, then it has a perfect performance. However, in complex disease comparison, we chose four disease gene sets – two from public databases and two from scientific literature. Each disease has a large number of genes to be positive genes. Thus the computational tool needs both high precision and recall rates to achieve good performance: precision means that the disease genes can be ranked higher than others, and recall means that a large candidate gene list can be found that includes the true disease genes. Therefore, it is not surprising that SNPs3d performs differently between these two types of tasks, as SNPs3d has high precision but has very limited recall.

Although Phenolyzer works well to find candidate genes for many diseases, its ability to identify the correct genes is to a large extent limited by the available biological knowledge. For example, if a phenotype or disease has never been reported to be associated with any gene before, Phenolyzer is much less likely to find the candidate genes, unless a similar or related phenotype is available. Nevertheless, as Phenolyzer gives a continuous normalized score from 0 to 1 for each candidate gene, the integration between Phenolyzer score and other kinds of variant prediction scores (for example, SIFT scores, PolyPhen scores) may further increase power for finding candidate gene and variants, even if the genes are previously uncharacterized. Finally, we believe that an improved algorithm that integrates both previous gene-disease and gene-gene relationship knowledge and an improved score for variant deleteriousness may offer the greatest power to prioritize variants from whole genome and exome sequencing data.

Reference

1. Brastianos, P.K. et al. *Nature genetics* **46**, 161-165 (2014).
2. Chudasama, K.K. et al. *The American Journal of Human Genetics* **93**, 150-157 (2013).
3. Yang, T.-L. et al. *The American Journal of Human Genetics* **83**, 663-674 (2008).
4. Chang, X. & Wang, K. *Journal of medical genetics* **49**, 433-436 (2012).
5. Lyon, G.J. et al. *Discov Med* **12**, 41-55 (2011).
6. Bragin, E. et al. *Nucleic acids research*, gkt937 (2013).
7. Calderone, A., Castagnoli, L. & Cesareni, G. *Nature methods* **10**, 690-691 (2013).
8. Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. *Nature genetics* **36**, 431-432 (2004).
9. Bauer-Mehren, A., Rautschka, M., Sanz, F. & Furlong, L.I. *Bioinformatics* **26**, 2924-2926 (2010).

Supplementary Table 1. Descriptive comparison of functionality between Phenolyzer and other similar tools.

Tool	Gene disease phenotype mapping	Phenotype and disease interpretation	Input format for disease and phenotype	Algorithm	Platform	Data visualization	Gene prioritization (without variant)	Variant prioritization(disease or phenotype specific)	CNV prioritization
Phenolyzer	OMIM, ClinVar,Orphanet, GWAS, GeneReviews	DO, HPO, CTD Medic, OMIM descriptors and synonyms	Any disease or phenotype related terms, OMIM ID	Ontology search; gene-disease and gene-gene score system; logistic regression	Web, command line	Interactive network; interactive bar plot, wordcloud	Yes	Yes	Yes
PosMed	Literature	None	Keywords	Neural network; literature mining	Web	None	Yes	No	No
Genecards	About 10 databases	None	Keywords	Database mining	Web	None	Yes	No	No
SNPs3d	OMIM, literature	None	Keywords	Literature and database mining	Web	Network by java plugin	Yes	No	No
Phevor	Ontology annotation(DO , HPO, etc)	DO, HPO, or depends on Phenomizer	Specific names, Phenomizer output	Ontology propagation	Web	Variant plot	No	Yes	No
Phenomizer	HPO, OMIM, Orphanet, Decipher	HPO	HPO terms	Ontology similarity search with p values	Web	Ontology visualization	Yes	No	No

Supplementary Table 2. Phenolyzer's performance on discovering the 590 known disease genes from a newborn sequencing study and on predicting the 55 newly published disease genes from four human genetics journals.

590 Known disease genes	Top1	Top5	Top10	Top50	In the list
Phenolyzer	81.2%	89.3%	90.7%	92.9%	93.4%
55 Newly published disease genes	**Top 5% ratio**	**Top 10% ratio**	**Top 20% ratio**	**Top 50% ratio**	**In the list**
Phenolyzer	43.6%	47.3%	56.4%	70.9%	96.4%
PosMed	36.4%	47.3%	50.9%	60.0%	70.9%
Phvor	20.0%	29.1%	41.8%	50.9%	85.5%
Phenomizer	30.9%	38.2%	41.8%	50.9%	60.0%
Genecards	5.5%	14.5%	21.8%	41.8%	58.2%