DOI:
10.1002/humu.23413

Document Version
Peer reviewed version

Text-mined phenotype annotation and vector-based similarity to improve identification of similar phenotypes and causative genes in monogenic disease patients

Jake R. Saklatvala[1], Nick Dand[1], Michael A. Simpson[1]

[1] King's College London, Medical & Molecular Genetics, London, SE1 9RT, UK

Correspondence to: Jake R. Saklatvala, King's College London, Medical and Molecular Genetics, London, SE1 9RT, UK. Email: jake.saklatvala@kcl.ac.uk

Contract grant sponsors: The Generation Trust; The Peter Stebbings Memorial Charity; UK National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.

Abstract

The genetic diagnosis of rare monogenic diseases using exome/genome sequencing requires the true causal variant(s) to be identified from tens of thousands of observed variants. Typically a virtual gene panel approach is taken whereby only variants in genes known to cause phenotypes resembling the patient under investigation are considered. With the number of known monogenic gene-disease pairs exceeding 5000, manual curation of personalised virtual panels using exhaustive knowledge of the genetic basis of the human monogenic phenotypic spectrum is challenging.

We present improved probabilistic methods for estimating phenotypic similarity based on Human Phenotype Ontology annotation. A limitation of existing methods for evaluating a disease's similarity to a reference set is that reference diseases are typically represented as a series of binary (present/absent) observations of phenotypic terms. We evaluate a quantified disease reference set, using term frequency in phenotypic text descriptions to approximate term relevance.

We demonstrate an improved ability to identify related diseases through the use of a quantified reference set, and that vector space similarity measures perform better than established information content-based measures. These improvements enable the generation of bespoke virtual gene panels, facilitating more accurate and efficient interpretation of genomic variant profiles from individuals with rare Mendelian disorders. These methods are available online at https://atlas.genetics.kcl.ac.uk/~jake/cgi-bin/patient_sim.py

Keywords: Phenotype similarity, HPO, Genetic diagnosis, Monogenic, Rare disease, Mendelian, Variant prioritisation, whole exome sequencing, whole genome sequencing

Introduction

The use of phenotype ontologies to capture patient phenotypes in a consistent and comparable manner is becoming standard practice to assist diagnostics and novel gene discovery in rare disease (Beaulieu et al., 2014; The Deciphering Developmental Disorders Study, 2014; Thompson et al., 2014). The phenotypic components of an individual's disease are commonly collected using Human Phenotype Ontology (HPO) terms. The HPO (P. N. Robinson & Mundlos, 2010) is a specialised ontology designed to encompass phenotypic abnormalities that appear in human disease. Software has been developed to facilitate HPO term collection and storage (Girdea et al., 2013) and semantic similarity algorithms have been developed that calculate phenotypic similarity between patients to aid new gene discovery (Akawi et al., 2015; Westbury et al., 2015) or similarity between a patient's phenotype(s) and a reference set of diseases to aid clinical and molecular diagnosis (Amberger, Bocchini, Schiettecatte, Scott, & Hamosh, 2014; Köhler et al., 2009).

The increasing use of whole genome and whole exome sequencing necessitates the identification of a single disease-causing variant (or pair of variants for compound heterozygous disorders) from ~25,000 identified exonic sequence variants per individual (Hoischen et al., 2010; Musunuru et al., 2010; Ng et al., 2009). Often, standard variant filtering approaches are combined with virtual gene panels: prespecified lists of known causative genes that are prepared for specific phenotypic areas; only variants in the panel are considered as potentially causal. Virtual panels can be large – the DDG2P list (Wright et al., 2015) encompasses over 1000 genes – and such gene lists don't fully capitalise on the phenotypic similarity between the patient and known diseases, assuming a uniform distribution of probabilities across the panel that each gene could be causal. Selective augmentation of gene panels to create a more bespoke interpretation is often carried out, but with the number of known monogenic gene-disease pairs now exceeding 5,000 (Amberger et

al., 2014) it is becoming less feasible to manually curate personalised virtual panels for all phenotypes.

Systematically collected patient phenotype data provides an opportunity to implement automated methodology that can utilise exhaustive knowledge of the human phenotypic spectrum in order to identify candidate genes. Methods that generate candidate gene lists tailored to the patient offer improvements over standard virtual gene panel approaches. Firstly, automated searching across the entire human phenotypic spectrum mitigates the aforementioned issue of the growing number of known monogenic gene-disease pairs. Secondly, within a virtual panel all genes are considered equally likely to cause the disease, whereas searching across the entire human phenotypic spectrum confers the ability to score each gene based on its relevance to the patient's disease, enabling the construction of more concise and relevant gene panels.

Approaches have been developed that query patient phenotype terms against a reference set to assist clinical and molecular diagnosis (Köhler et al., 2009; Peter N. Robinson et al., 2014; Singleton et al., 2014). These methods incorporate ontology-based similarity measures and the reference set consists of curated phenotype annotations to known Mendelian disorders. Whilst these methods have been shown to be effective in a handful of cases or simulated patients/exomes, the underpinning phenotype similarity metrics have limitations. Clinical features are annotated to phenotypes as binary present/absent observations which are unable to describe the relevance of each phenotypic feature to the overall disease. For example, primary microcephaly-1 (MIM #251200) is characterised by 16 'Phenotypic abnormality' terms in the curated HPO annotation set, including the core feature 'Microcephaly' (HP: 0000252) as well as other features of lower penetrance (such as 'Renal hypoplasia/aplasia' – HP: 0008678). Binary annotation is unable to reflect the relative importance of terms in similarity calculations, in this case weighting the cardinal 'Microcephaly' feature equally to

non-obligate features such as ‘Renal hypoplasia’ and ‘Hyperreflexia’. Although penetrance data is recorded for a proportion of HPO phenotype annotations, termwise similarity measures such as the Resnik algorithm (Resnik, 1999) do not utilise this information, although BOQA (Bauer, Köhler, Schulz, & Robinson, 2012) is a Bayesian query tool that has been built to utilise the limited existing penetrance information.

In light of these limitations, we first investigated the use of a quantitatively annotated reference disease set, where HPO terms were weighted by relevance to their diseases. We used simple text mining of free-text disease descriptions to generate phenotype annotations, and used term frequency to approximate relevance. The utility of text-mined reference annotations was established by comparison to the curated annotation set of the same diseases employed by the Phenomizer, a differential diagnostic tool for human Mendelian disorders (Köhler et al., 2009). We also compared the quantified text-mined phenotype annotations against an unquantified version of the same annotation set. Secondly, we used a vector space model (VSM) to evaluate similarity between HPO-annotated diseases, comparing this to the Resnik similarity algorithm implemented in the Phenomizer (Resnik, 1999), as well as the BOQA algorithm (Bauer et al., 2012). Quantification of phenotype terms and the use of vector space similarity has been used outside of a clinical context in human phenome analysis (Lage et al., 2007; van Driel, Bruggeman, Vriend, Brunner, & Leunissen, 2006) and we hypothesised that the use of term quantification to represent relevance, combined with a suitable similarity measure would enhance our ability to identify similar diseases.

To test this hypothesis we developed evaluation metrics using OMIM Phenotypic Series (PS) (Amberger et al., 2014) as a set of known similar diseases. We evaluated methods based on their ability to group these diseases close to each other. Finally, we benchmarked different disease reference annotations and similarity metrics using an external diagnosed patient dataset (The Deciphering Developmental Disorders Study, 2014) in which patient HPO terms

were recorded prior to their genetic diagnosis (Wright et al., 2015). We evaluated the different reference sets (and query methods) based on their ability to return the correct gene for each patient after querying their HPO terms.

Materials and methods

Phenotype annotation

Manually curated OMIM phenotype annotations are made publically available for download by the HPO group (build #1233, Jan 13 2016, downloaded 09/02/16). Only 'phenotypic abnormality' annotations were used to ensure equivalence with annotations used by Phenomizer.

In order to text-mine phenotype annotations, the descriptive text was first extracted from OMIM (date: 05/02/16). The text entries were then submitted to Annotator (Shah et al., 2009) (date: 08/02/16), a free-to-use resource made available by the National Centre of Biomedical Ontology (Musen et al., 2012) which infers ontology annotations from text. Annotator utilises 'an exact string comparison (a "direct" match) between the text and ontology term names, synonyms, and IDs' (Shah et al., 2009), making it a reasonably simple text-mining tool. HPO terms were filtered to include only 'Phenotypic abnormality' terms (HP: 0000118). Text-mined phenotype annotation term counts were converted to penetrance statistics (to enable compatibility with BOQA) by dividing the count of each annotated HPO term by the highest HPO count within the phenotype.

Phenotype similarity

Once annotated, similarity between phenotypes was calculated. The performance of the Resnik algorithm implemented in the Phenomizer (Köhler et al., 2009; Resnik, 1999) and BOQA (Bauer et al., 2012) were compared to vector space, which is used in various information retrieval methods. Diseases are represented as vectors, with each dimension indicating the presence (or absence) of a particular HPO term. Similarity between two disease vectors is measured using the cosine of the angle between them (Equation 1). In this

application of cosine similarity, the score ranges from 1 (corresponding to an angle of 0^0, indicating identical vectors) to 0 (corresponding to 90^0, indicating orthogonality).

Equation 1: Cosine similarity between feature vectors *Q* and *D*

$$sim(Q, D) = \cos\theta = \frac{Q \cdot D}{\|Q\| \cdot \|D\|}$$

In its simplest form a vector space model requires vector components to be set to 1 (indicating the HPO term is present) or 0 (indicating the HPO term is absent). It can be advanced from this simple binary setting using a variety of techniques (our implementation incorporates all of the following):

Use of the semantic inheritance structure of terms; recalling that annotation of a particular ontology term implicitly annotates every ancestor of that term on the DAG, ancestral terms can also be included in the disease vectors.

Using term frequency; pertinent when using text-mined annotations, the number of times a disease is annotated with a particular HPO term is used.

Use of term weights; identical to the Resnik term-Information Content calculation, vector components can be modified by multiplying them by the inverse document frequency (IDF; Equation 2) of their terms. This up-weights vector features that correspond to specific terms.

Equation 2: IDF adjustment for term *t*

$$IDF(t) = \log\left(\frac{total\ phenotypes}{phenotypes\ annotated\ with\ t}\right)$$

Simulated queries (interpolated-precision recall)

When benchmarking information retrieval (IR) methods, approaches generally involve defining a set of entities (diseases) within the corpus that are 'known' to be similar (by consulting literature/experts) and then observing how well a particular method performs in

classifying such entities as similar. The approach developed here made use of the OMIM phenotypic series (PS) as the set of known similar diseases. The OMIM phenotypic series are defined as a set of 'phenotype entries [that] overlap significantly in their clinical manifestations', and that they are classified according to clinical judgement, not computed similarity (Amberger et al., 2014). Phenotypic series have variable size, grouping from 2 to 78 diseases (mean = 8.03). There are 353 phenotypic series, covering 2785 diseases in the OMIM catalogue (PS information downloaded: 15/02/16).

For each disease in a phenotypic series, the disease was removed and its terms used as a query to the remaining OMIM reference set. The diseases in the remaining set were ranked by similarity to the query, evaluating based on the ranks of the diseases within the same phenotypic series. Methods were evaluated by conducting this test on all diseases in phenotypic series and aggregating the results. Only diseases that were annotated by all methods (Table 1) were included in the analysis and diseases in multiple phenotypic series were not used (n=47), leaving 2317 phenotype queries (319 PS).

Precision-recall was used to evaluate the OMIM phenotypic series queries because these metrics are more applicable than ROC metrics when there is a highly skewed class distribution (a maximum of 2% of the query results are classed "positive"; the rest will be classed "negative"). Starting from rank 1 and iterating through further ranks, the precision and recall were calculated (defined in Equation 3 and Equation 4) for each query.

Equation 3: Precision (*P*) for a query at rank *r*

$$P_r = \frac{similar\ diseases\ within\ rank\ r}{r}$$

Equation 4: Recall (*R*) for a query at rank *r*

$$R_r = \frac{similar\ diseases\ within\ rank\ r}{total\ similar\ diseases}$$

To overcome difficulties in averaging performance over queries with variable numbers of positive results (due to the variable size of phenotypic series), an interpolation step was included. This involved defining 11 standard recall points (0 … 0.1 … 0.2 … … 1) and using the maximum precision found above each point as the 11-point interpolated precision recall (Equation 5).

Equation 5: Interpolated precision (*I*) for a query at standard recall level *R*

$$I_R = \max_{R' \geq R} P(R')$$

To evaluate a method where *n* queries were tested, the *n* interpolated precision points at each standard recall point were averaged, showing the decline of precision as recall increases. To facilitate statistical comparison between methods, the mean average precision (MAP) was also calculated across the queries (Equation 6). MAP is highly correlated with the area under a precision-recall curve and the single value metrics (rather than a curve) enable simple hypothesis testing using a Student's paired t test (Smucker, Allan, & Carterette, 2007).

Equation 6: Average precision (AP) for a query where *N* is the total number of results, *d* is the number of relevant results, *P(k)* is the precision at k results and $\Delta r(k)$ is the change in recall from cut-off *k-1* and *k* (thus only permits precision at 'relevant' ranks to be averaged).

$$AP = \frac{1}{d} \sum_{k=1}^{N} P(k) \Delta r(k)$$

DDD patient queries

To observe method performance in real cases of rare disease, patient diagnoses and phenotype information from the DDD consortium (The Deciphering Developmental Disorders Study, 2014) were used for additional benchmarking. A cohort of 1,133 patients with undiagnosed developmental disorders underwent exome sequencing and aCGH to identify SNVs, indels and CNVs. Following an analysis pipeline of filtering and evaluating variants that appear in the curated DDG2P developmental gene list (Wright et al., 2015) 351

patients were returned a (probable) diagnosis and their causative variants were released along with HPO terms describing their disease. Of these 351 diagnoses, 283 were due to pathogenic variants in a single gene (rather than digenic diagnoses, CNVs, UPDs or mosaicisms). 258 of these monogenic diagnoses were in a gene that mapped to an OMIM phenotype in all 3 reference sets (Table 1). The HPO terms for each patient were queried to the different disease reference sets with the different similarity methods. Only diseases in all reference sets were queried (n=6518). The 6518 phenotypic similarity scores for each query were converted to 1268 developmental gene scores by taking the top disease score for each OMIM-mapped gene in the DDG2P list (downloaded: 05/01/16) (Firth et al., 2009). Methods were evaluated based on their ability to prioritise the causative gene of each patient using only the phenotypic information provided. Analysis was initially conducted based on the ranks that each method assigned the correct causative gene in each patient.

Score-probability normalisation

The methods were tested further by undertaking analysis that quantified method performance ahead of choosing a developmental gene at random. Phenotypic similarity scores were converted to probabilities, which were then converted to normalised gene probabilities which could be plot against a baseline of the probability of selecting a random gene from the DDG2P list (n=1268). Rather than assuming a linear relationship between the phenotypic similarity score and the probability of similarity, we used similarity scores between diseases in OMIM phenotypic series to characterise this relationship. For each annotation/similarity method combination the following was done – for each disease *X* in a phenotypic series, we queried the HPO terms of *X* to the remaining annotation set. Similarity scores between *X* and diseases in the same series were recorded as examples of scores between truly similar diseases, while scores between *X* and all other phenotypes were recorded as examples of scores between non-similar diseases. The spectrum of recorded similarity scores was split

into bins at regular (100) intervals and the proportion of similarity scores representing true similarity *T* could be calculated for each bin (Figure 1, Equation 7). A generalised logistic function (Equation 8) was chosen to fit the data.

Equation 7: Calculation of *T* for each bin – TP is the number of scores between "true matches" within the bin; ALL is the total number of scores within the bin.

$$T_{bin} = \frac{TP_{bin}}{ALL_{bin}}$$

Equation 8: Generalised logistic function fit to the *T* data using midpoint *x* of each bin. Data was fit to using non-linear least squares, optimising variables *K*, *Q*, *B*, *M* and *v*.

$$T = \frac{K}{(1 + Qe^{-B(x-M)})^{1/v}}$$

The generalised logistic function (with appropriate calculated variables delineated in (Table 2)) was used to rescale the phenotypic similarity scores. The rescaled phenotypic similarity scores were then converted to gene scores using the highest disease score that mapped to each gene, and normalised so all gene scores summed to one.

Results

Phenotype annotation

We initially extracted HPO terms (and their frequencies) from OMIM disease descriptions using simple text mining (Materials and Methods). We compared the numbers of diseases that could be annotated and HPO terms extracted by text mining to the annotations curated by the HPO team (Köhler et al., 2014) (Table 1). Compared to the curated annotations, unquantified text mining assigned more phenotype terms to each disease on average, but detected a far narrower range of different terms overall. Text mining detected a lower proportion of the full range of HPO terms due to only using the OMIM text description as an input, whereas the curated annotations utilise additional data sources, such as published clinical studies and individual clinical experience. However, text mining resulted in an increased number of annotations because terms were encountered in a greater number of different disease descriptions on average. These terms were more general, hence having a lower average distance to the root HPO term (HP:0000001 – 'All'). When the text-mined terms were quantified, it resulted in over double the total annotation count.

Simulated queries

We developed a series of evaluation metrics to benchmark the performance of different combinations of phenotype annotation and similarity methods. We used the OMIM phenotypic series (PS), which comprise groups of clinically linked diseases, as a set of known similar diseases to generate queries. Only diseases that were annotated by all methods (Table 1) were included in the analysis and diseases in multiple phenotypic series were not queried (n=47), leaving 2317 disease queries (from 319 PS). We used precision-recall, a measure commonly utilised when evaluating information retrieval (IR) systems, to evaluate method performance on the phenotypic series queries (Equation 3, Equation 4). To overcome issues

of averaging over queries where the phenotypic series size was variable, precision-recall was converted to 11-point interpolated precision-recall (Equation 5). For each method 2317 queries were tested, so at each standard recall point the 2317 interpolated precision points were averaged to show the decline of precision as recall increases. Mean average precision (Equation 6) was also calculated to facilitate statistical comparison between methods using a paired Student's *t* test. *P*-values have been corrected for multiple testing under dependency (Benjamini & Yekutieli, 2001) unless stated otherwise.

Here we evaluated two different similarity measures (Vector space and Resnik). When HPO annotations were not quantified (curated (c) and unquantified text mining (u)) the performance of vector space similarity had a modest but significant advantage over Resnik ($P_c = 4.62 \times 10^{-13}$, $P_u = 4.47 \times 10^{-14}$), but vector space was far superior when the HPO terms were quantified ($P_q = 2.83 \times 10^{-95}$) (Figure 2A). When annotations were not quantified the similarity in performance between the two algorithms was expected due to their similar premise.

When phenotype annotations are quantified, vector space similarity performs better due to its ability to down-weight the vector features of more general terms (as they will have a lower IDF/IC as defined in Equation 2 (Materials and Methods)) and noise terms (which are likely to be found at a lower frequency to 'genuine' terms) that text mining is more prone to detecting. Using vector space similarity, we assessed the performance of each annotation method. We found that quantified text mining is superior to curated annotation ($P = 9.02 \times 10^{-58}$), although unquantified text mining is inferior to curated and quantified methods ($P = 2.38 \times 10^{-10}$ and $P = 1.02 \times 10^{-151}$, respectively). Having observed that querying a disease consisting of quantified phenotype terms against a quantified reference set was the optimal setting, we tested whether a quantified reference set was superior to the others even when the query was not quantified.

The quantified reference disease set compares favourably to both the curated and unquantified text-mined reference sets when using unquantified text-mined queries ($P = 2.55 \times 10^{-64}$ and $P = 1.38 \times 10^{-63}$, respectively) (Figure 2B). However, when querying with curated phenotype annotations, a quantified reference set provides no clear benefit in comparison to the curated annotation set ($P = 0.658$[a], Figure 2C). Curated queries were more effective with the curated reference set while text-mined queries were more effective with the text-mined reference sets.

DDD patient queries

Rank-based analysis

For further benchmarking we used an independent patient dataset released by the DDD consortium (The Deciphering Developmental Disorders Study, 2014) which contained the genetic diagnosis for 351 patients with developmental disorders as well as the HPO terms used to describe their disease prior to diagnosis. We used different combinations of reference disease sets and phenotype similarity calculation methods to rank genes in the DDG2P list (used as a virtual panel in the original study (Wright et al., 2015)) to determine which methods were most effective in prioritising genes within this large list. We ranked genes by querying the HPO terms for each of the 258 patients with monogenic disease diagnoses to the different OMIM disease reference sets. These disease similarity scores were converted to gene scores using the OMIM gene-disease mappings (taking the top disease score for genes that cause multiple OMIM diseases). The ranks of the correct gene for each of the 258 patients were compared across methods using the Wilcoxon test followed by adjustment for multiple testing under dependence.

[a] Without multiple testing correction

Contrary to our findings based on OMIM PS, we found that queries made using vector space similarity showed no significant improvement over Resnik with the unquantified text-mined reference set ($P = 0.481^{a}$), but the use of vector space resulted in an improvement for the curated ($P = 0.0210$) and quantified text-mined ($P = 1.57 \times 10^{-4}$) reference sets (Figure 3). Comparing reference sets under vector space similarity, we found that the unquantified text-mined annotation set showed no significant improvement over curated annotations ($P = 0.36^{a}$), but the quantified set showed a significant improvement in correct phenotype ranks ($P = 1.39 \times 10^{-4}$). The quantified text-mined reference set also showed a significant improvement in comparison to its unquantified counterpart ($P = 1.03 \times 10^{-7}$), supporting our previous observations in the OMIM PS benchmark dataset. Using quantified text mining in combination with vector space showed dramatic improvements over using BOQA with both the same quantified text-mined reference set, as well as the curated set for which BOQA could utilise penetrance statistics ($P = 1.85 \times 10^{-6}$ and $P = 8.62 \times 10^{-6}$ respectively). Interestingly, for each reference set, using both BOQA and Resnik similarity methods predicted more correct genes at rank 1 than using vector space, although this trend is reversed at rank 10, where vector space becomes more sensitive than other similarity methods.

Score-based analysis

In further analysis we attempted to re-scale the similarity scores to reflect the probability of causality for each of the genes in the DDG2P list based on the phenotypic similarity score between each DDD patient and the reference diseases. After querying the patient HPO terms using the different methods, the generalised logistic function (Equation 8, with appropriate variables delineated in Table 2) was used to rescale the phenotypic similarity scores. BOQA outputs a probability for each disease and therefore rescaling isn't required. The rescaled similarity scores were then converted to DDG2P gene scores using the OMIM gene-disease mappings (taking the top phenotype score for genes that cause multiple OMIM diseases) and

normalised to sum to one to give an estimate of the probability for each gene. These gene scores were plotted against a baseline of the probability of selecting a random gene from the DDG2P list (Figure 4).

For each reference set, vector space similarity outperformed Resnik similarity in assigning a higher average probability to the disease-causing gene (Table 3), although this advantage is significant only for the curated ($P = 2.03 \times 10^{-3}$) and the quantified text-mined reference sets ($P = 1.91 \times 10^{-3}$). Compared to vector space and Resnik, using BOQA to measure patient similarity to the respective reference sets resulted in a much higher mean probability assigned to the correct gene, although this was due to a handful of outlying patients having a very high probability assigned to the causative gene (Each BOQA method had 15-25 patients with $\Delta>0.2$ for the correct gene). However, both BOQA and Resnik approaches resulted in a low median Δ and performed poorly for the majority of patients (in the best case, 61 of 258 patients had a positive Δ). This was also the case when querying the curated reference set using vector space, which achieved a high average probability but was also offset by poor median performance, with over half of the correct genes having a lower probability than selecting the gene at random from the DDG2P list. The quantified reference set combined with vector space achieved the highest median probability for the correct genes. The quantified reference sets also achieved the highest number of patients with a positive Δ (probability subtracted by prior) for the correct gene. Additionally, the Resnik similarity measure has a slight advantage over vector space in this respect.

Discussion

Here we present a method for the identification of similar diseases through annotation with basic text mining, including term quantification to make optimal use of the most relevant phenotypic features. We have shown that this approach offers a clear advantage over current tools that use binary annotation settings to indicate only the presence or absence of clinical terms. Ideally, clinical terms could be quantified in the representation of a genetic disease by using data on their prevalence among individuals having that disease. However, such a comprehensive dataset does not yet exist (41% of current HPO curated annotations have quantification information with 48% of diseases containing at least one quantified phenotype term) and it is unclear how some currently employed similarity methods can integrate knowledge of feature prevalence. Whilst frequency extracted from text mining phenotypic descriptions is not expected to fully capture the penetrance of terms across affected cases, we find a weak positive correlation where term penetrance data is available (Supp. Figure S1) and it serves as an effective weighting scheme that encompasses all phenotypes. It seems likely that the performance of this text-mining approach could further improve through the use of additional relevant input phenotype descriptions. For example, text input from OMIM literature references and further literature on MEDLINE could be suitable. Despite limitations of this method of annotation, we find that this relatively simple method of term quantification aids the retrieval of similar diseases as well as the downstream identification of genes. Additionally, we investigated a vector space similarity measure as an alternative to currently used phenotype similarity measures, as its construction is less sensitive to noise associated with quantitative annotation.

In our initial validation stage we tested the ability of different combinations of annotation and similarity methods to group known similar OMIM diseases closely to each other. We found that quantification of phenotypic information enhanced our ability to identify similar

diseases, compared to both the curated annotation set and an unquantified version of the same reference set. Vector space similarity performed roughly as well as the Resnik measure of similarity when unquantified reference sets were used, but greatly outperformed it when quantifying reference phenotype terms. The use of OMIM phenotypic series as the known similar diseases enabled the methods to be tested across a great number of phenotypes (~35% of annotated OMIM records), representing a very diverse range of diseases. Although these groupings are curated based on clinical judgment rather than computed similarity (and thus not biased by annotations) the phenotypic series information may have biased the various annotation sets through the recycling of descriptive text. However, we did observe concordance between the DDD patient benchmarking and these OMIM PS tests.

The DDD patient diagnosis dataset used in subsequent testing should contain no such bias because each patient's HPO terms were recorded by their UK NHS or Irish Regional Genetics Service prior to diagnosis (Wright et al., 2015). When patient HPO terms were queried to the different disease reference sets, the quantified text-mined annotation set provided a significant benefit ahead of both unquantified reference sets in identifying relevant diseases for each patient and therefore ranking the true causative genes highly. Interestingly, the use of BOQA resulted in the identification of the diagnostic gene with very high confidence in a small number of patients (<10%) and predicted more causal genes at rank 1, although our methods were generally able to assign a higher rank and more probability to the causal gene in the majority of patient population. Despite the advances shown by our method, sensitivity remains relatively low for the challenging task of identifying disease-causing gene. Each combination of methods tested herein failed to assign a high rank to the causative developmental gene for a large subset of DDD patients. The top-performing method only found the causative gene within rank 10 in 23% of cases and rank 100 in 56% of cases (Figure 3). However, this is likely to reflect the dataset being enriched for patients that had a

developmental disorder that was initially difficult to diagnose. More straightforward cases would therefore be depleted in this dataset (The Deciphering Developmental Disorders Study, 2014), and we would expect this methodology to perform better at solving such cases - a valid alternative application. In addition, the DDD study also identified novel gene-phenotype links, such as that of the *MED13L* gene. *MED13L* variants have previously been described in patients with congenital heart defects such as dextro-looped transposition of the great arteries (d-TGA; MIM #608808) (Muncke et al., 2003), but this study identified 8 patients with variants in *MED13L* gene that had intellectual disability but lacked congenital heart malformations (Adegbola et al., 2015).

Text mining for phenotype annotations can also have utility in a clinical context; its quick and systematic nature could make it highly valuable in large clinical genetics services. Manual assignment of clinical ontology terms has only recently become widely undertaken and is performed with variable degrees of diligence. We propose that as an alternative, text mining of patient clinic letters for phenotypic terms would enable rapid and systematic definition of patient phenotypes. Term quantification would enable scoring of terms based on the belief that they are truly descriptive of the patient, although future work could incorporate more sophisticated text mining features such as detection of term negation and modifiers for severity. Text mining patient records over a long period would also enable longitudinal phenotype data to be collected. This is particularly pertinent in the context of syndromes where different clinical features appear at different ages. Patients could then be compared based on their clinical presentation at defined timepoints.

To summarise, we have shown that quantifying clinical terms can be an effective method of refining phenotype descriptions, beyond a simple representation as binary observations. When calculating similarity, term frequency becomes an additional feature by which terms can be weighted rather than only their IC/IDF. We have also utilised a suitable method for

calculating similarity between quantified phenotypic definitions, vector space models, which is able to take into account term frequency and specificity. Our methods show improvement compared to currently employed methods in classifying related OMIM diseases as similar. Querying patient phenotypes from a genuine developmental disorder study to a quantified disease reference set displayed a significant improvement in aiding the identification of the correct gene. These methods are available for use online at:

https://atlas.genetics.kcl.ac.uk/~jake/cgi-bin/patient_sim.py

Acknowledgements

This work was supported by the UK National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London, as well as a PhD studentship sponsored by The Generation Trust and funded by The Peter Stebbings Memorial Charity.

References

Adegbola, A., Musante, L., Callewaert, B., Maciel, P., Hu, H., Isidor, B., … Kalscheuer, V. M. (2015). Redefining the MED13L syndrome. *European Journal of Human Genetics*, (January), 1–10. doi:10.1038/ejhg.2015.26

Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A. F., … Hurles, M. E. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics*, *47*(11), 1363–1369. doi:10.1038/ng.3410

Amberger, J. S., Bocchini, C. a, Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(November 2014), 789–798. doi:10.1093/nar/gku1205

Bauer, S., Köhler, S., Schulz, M. H., & Robinson, P. N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics*, *28*(19), 2502–2508. doi:10.1093/bioinformatics/bts471

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., … Boycott, K. M. (2014). FORGE Canada consortium: Outcomes of a 2-year national rare-disease gene-discovery project. *American Journal of Human Genetics*, *94*(6), 809–817. doi:10.1016/j.ajhg.2014.05.003

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188. doi:10.1214/aos/1013699998

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., … Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, *84*(4), 524–533. doi:10.1016/j.ajhg.2009.03.010

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., … Brudno, M. (2013). PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*, *34*(8), 1057–1065. doi:10.1002/humu.22347

Hoischen, A., van Bon, B. W. M., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., … Veltman, J. A. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genetics*, *42*(6), 483–485. doi:10.1038/ng.581

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, *42*(Database issue), D966–74. doi:10.1093/nar/gkt1026

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., … Robinson, P. N. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *American Journal of Human Genetics*, *85*, 457–464. doi:10.1016/j.ajhg.2009.09.003

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., … Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, *25*(3), 309–316. doi:10.1038/nbt1295

Muncke, N., Jung, C., Rüdiger, H., Ulmer, H., Roeth, R., Hubert, A., … Rappold, G. (2003).

Missense Mutations and Gene Interruption in PROSIT240, a Novel TRAP240-Like Gene, in Patients with Congenital Heart Defect (Transposition of the Great Arteries). *Circulation*, *108*(23), 2843–2850. doi:10.1161/01.CIR.0000103684.77636.CD

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M.-A., & Smith, B. (2012). The National Center for Biomedical Ontology. *Journal of the American Medical Informatics Association*, *19*(2), 190–195. doi:10.1136/amiajnl-2011-000523

Musunuru, K., Pirruccello, J. P., Do, R., Peloso, G. M., Guiducci, C., Sougnez, C., … Kathiresan, S. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *The New England Journal of Medicine*, *363*(23), 2220–2227. doi:10.1056/NEJMoa1002926

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., … Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*(7261), 272–276. doi:10.1038/nature08250

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, *11*, 95–130. doi:10.1613/jair.514

Robinson, P. N., Köhler, S., Oellrich, A., Genetics, S. M., Wang, K., Mungall, C. J., … Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, *24*(2), 340–348. doi:10.1101/gr.160325.113

Robinson, P. N., & Mundlos, S. (2010). The Human Phenotype Ontology. *Clinical Genetics*, *77*(6), 525–534. doi:10.1111/j.1399-0004.2010.01436.x

Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., & Musen, M. A. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*, *10 Suppl 9*, S14. doi:10.1186/1471-2105-10-S9-S14

Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., … Yandell, M. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics*, *94*(4), 599–610. doi:10.1016/j.ajhg.2014.03.010

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management CIKM 07*, *308*(1-2), 623. doi:10.1145/1321440.1321528

The Deciphering Developmental Disorders Study. (2014). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, *10*(Chr X), 223–228. doi:10.1038/nature14135

Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I. G., … Lochmüller, H. (2014). RD-connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *Journal of General Internal Medicine*, *29*(SUPPL. 3), 780–787. doi:10.1007/s11606-014-2908-8

van Driel, M. a, Bruggeman, J., Vriend, G., Brunner, H. G., & Leunissen, J. a M. (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics : EJHG*, *14*(5), 535–542. doi:10.1038/sj.ejhg.5201585

Westbury, S. K., Turro, E., Greene, D., Lentaigne, C., Kelly, A. M., Bariana, T. K., … Robinson, P. N. (2015). Human phenotype ontology annotation and cluster analysis to

unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Medicine*, *7*, 36. doi:10.1186/s13073-015-0151-5

Wright, C. F., Fitzgerald, T. W., Jones, W. D., Clayton, S., McRae, J. F., Van Kogelenberg, M., … Firth, H. V. (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. *The Lancet*, *385*(9975), 1305–1314. doi:10.1016/S0140-6736(14)61705-0

Figure Legends

Figure 1: T, the fraction of true positive scores within each bin when OMIM PS phenotypes were queried using different annotation and similarity methods. Additionally the sigmoid curve fit is displayed for each setting with the solid lines.

Figure 2: Average 11-point interpolated precision-recall for 2317 queries using different combinations of phenotype annotation and similarity methods. Similarity measure is denoted by linestyle (solid = vector space; dashed = Resnik). Reference set annotation and query annotation is denoted by line and dot colour respectively (red = HPO curated annotation; green = unquantified text mining; blue = quantified text mining). Area under the precision-recall curve and mean average precision (MAP) are indicated in the legend. A: all combinations of annotation and similarity measure were tested, keeping the annotation setting of the query the same as the annotation of the reference set. B: queries were made using the unquantified text-mined set only. C: queries were made using the curated set only. B&C: only the vector space similarity measure was used as it was superior to Resnik in all cases.

Figure 3: Ranks of the 'correct' gene for 258 queries from the diagnosed DDD patient dataset, for the different combinations of reference annotations and query methods. Only phenotypes in all reference sets were queried (n=6518) and phenotype ranks were converted to DDG2P gene ranks (n=1268). Boxplot limits represent the 5th and 95th percentiles; black diamond indicates the mean rank.

Figure 4: Probability (after logistic function rescaling) assigned to the correct gene for 258 DDD patient queries to different reference sets using different query methods. Probability was plot against a baseline of selecting a DDG2P gene at random (1/1268). Number of patients for which probability is higher than randomly selecting a DDG2P gene is indicated on the plot.

Figures

Figure 1

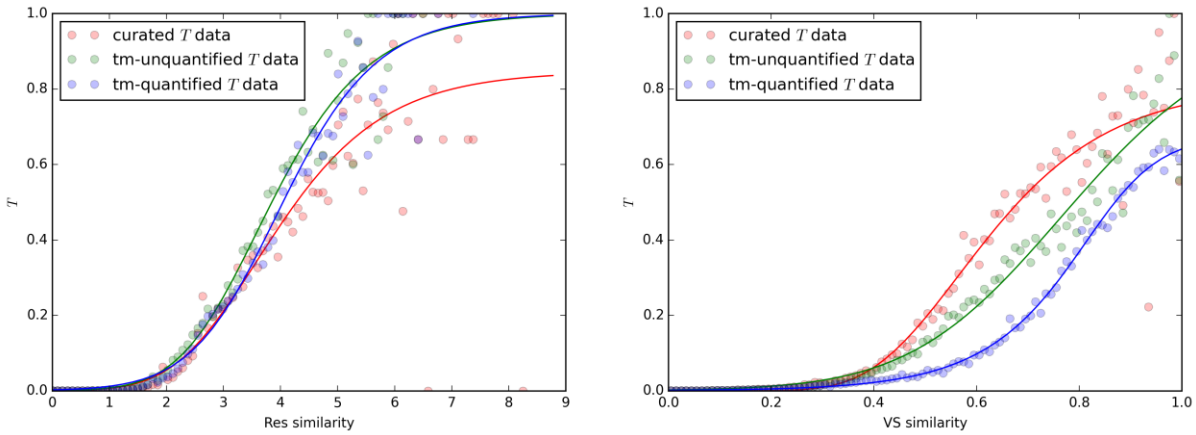


Figure 2

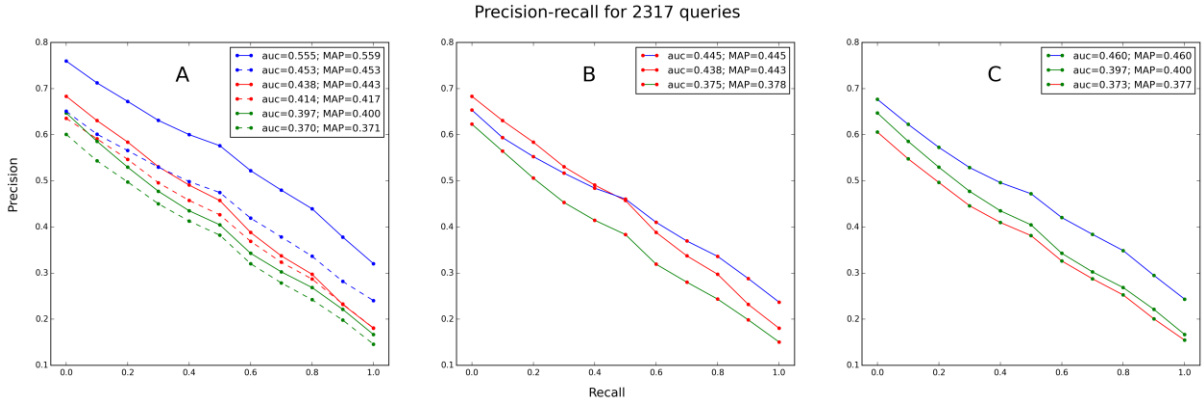
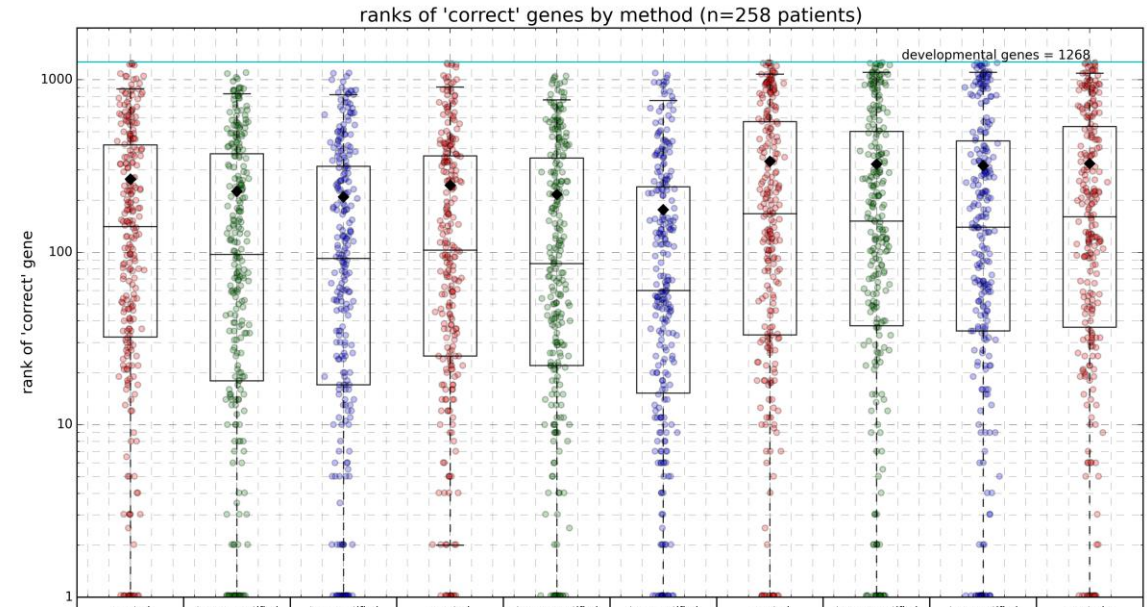
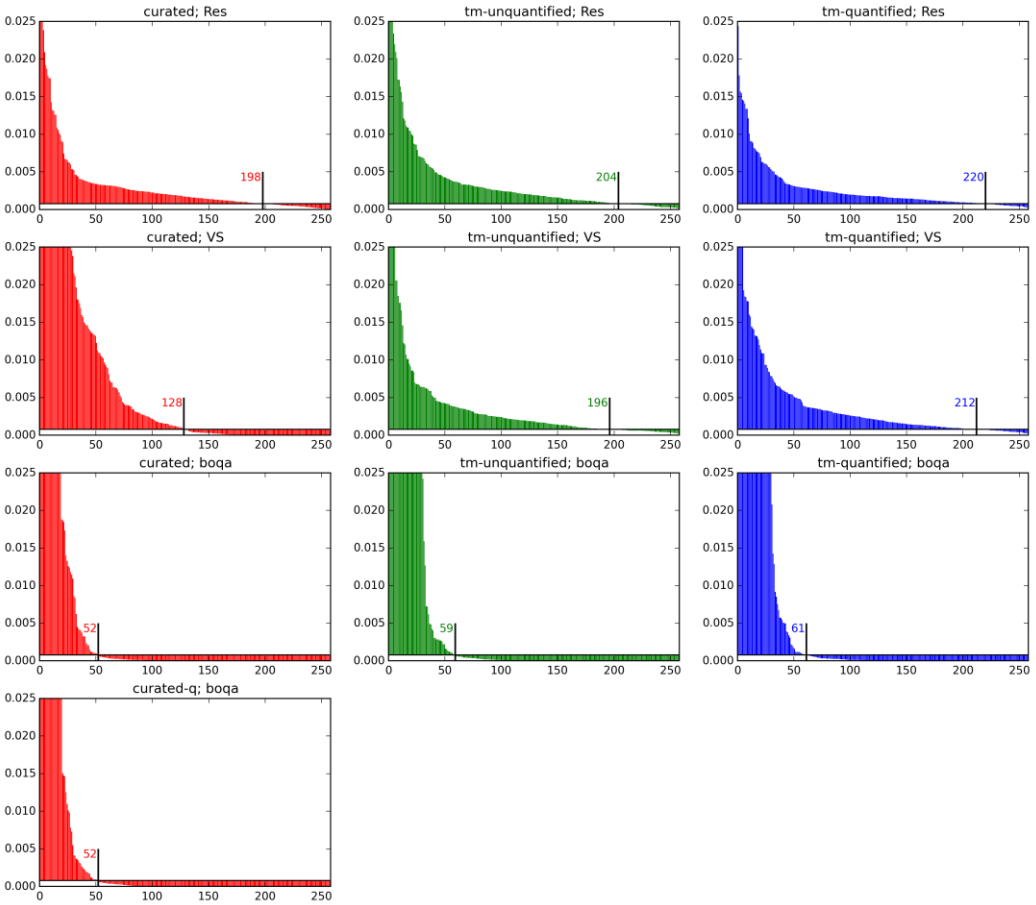


Figure 3



	curated Res	tm-unquantified Res	tm-quantified Res	curated VS	tm-unquantified VS	tm-quantified VS	curated boqa	tm-unquantified boqa	tm-quantified boqa	curated-q boqa
median	141.0	97.0	92.0	103.0	86.0	60.0	167.2	151.8	140.0	160.5
mean	265.6	225.1	209.1	244.2	216.6	176.6	335.9	323.9	317.4	326.7
solved@1	16	25	25	12	19	18	15	22	23	15
solved@10	32	48	52	35	48	59	31	39	39	30
solved@100	115	131	135	128	135	145	99	104	115	101

Figure 4



Tables

Annotation method	Phenotypes	Total annotations	Terms used	Average phenotypes per term	Average distance to root
Curated	6902	90236	6825	13.2	6.50
Text mining, unquantified	7600	105644	4719	22.4	6.43
Text mining, quantified	7600	230274	4719	22.4	6.43
Curated $_N$	6518	88533	6765	13.1	6.50
Text-mined, unquantified $_N$	6518	99126	4679	21.2	6.43
Text-mined, quantified $_N$	6518	215895	4679	21.2	6.43
Curated $_X$	384	1703	918	1.86	6.11
Text-mined, unquantified $_X$	1082	6518	1598	4.08	6.19
Text-mined, quantified $_X$	1082	14379	1598	4.08	6.19

Table 1: Metrics for different methods of annotating the OMIM phenotype catalogue with HPO terms. The subscript *N* denotes the group of phenotypes captured by all annotation methods. The subscript *X* denotes those phenotypes exclusively captured by each annotation method (curated vs. text-mined).

Annotation method	Similarity method	K	Q	B	M	v
Curated		0.847	6.92	0.832	-0.626	0.21
Text mining, unquantified	Resnik	1	3.5	0.965	1.15	0.334
Text mining, quantified		1	1.54	1.09	3.1	0.63
Curated		0.81	0.515	6.15	0.207	0.0569
Text mining, unquantified	Vector space	1	0.628	5.46	0.753	0.598
Text mining, quantified		0.678	2.17	15.1	0.807	2.02

Table 2: Optimised generalised logistic function variables from Equation 8 for each combination of annotation and similarity method. Functions are plot in Figure 1.

Annotation method	Similarity method	Average Δ	Average fold change	Median Δ	Median fold change	n(Δ > 0)
Curated	Resnik	2.31E-03	2.927	9.26E-04	1.175	198
Text mining, unquantified		2.69E-03	3.414	1.12E-03	1.422	204
Text mining, quantified		1.88E-03	2.388	9.01E-04	1.143	220
Curated	Vector space	2.27E-02	28.731	-5.86E-05	-0.074	128
Text mining, unquantified		2.79E-03	3.532	1.05E-03	1.327	196
Text mining, quantified		3.43E-03	4.345	1.17E-03	1.479	212
Curated	BOQA	5.33E-02	67.529	-7.79E-04	-0.987	52
Text mining, unquantified		8.31E-02	105.426	-7.69E-04	-0.975	59
Text mining, quantified		7.65E-02	97.020	-7.54E-04	-0.956	61
Curated, quantified		5.21E-02	66.053	-7.69E-04	-0.975	52

Table 3: Metrics for score-based analysis (after logistic function rescaling) of different phenotype annotation and similarity methods used to identify the causative genes for diagnosed DDD patients (from Figure 4). Δ = probability – prior; Fold change = Δ/prior.