Improving the diagnostic yield of exome-sequencing, by predicting gene-phenotype associations using large-scale gene expression analysis

Patrick Deelen,[1,2,4] Sipko van Dam,[1,4] Johanna C. Herkert,[1,5] Juha M. Karjalainen,[1,5] Harm Brugge,[1,5] Kristin M. Abbott,[1] Cleo C. van Diemen,[1] Paul A. van der Zwaag,[1] Erica H. Gerkes,[1] Pytrik Folkertsma,[1] Tessa Gillett,[1] K. Joeri van der Velde,[1,2] Roan Kanninga,[1,2] Peter C. van den Akker,[1] Sabrina Z. Jan,[1] Edgar T. Hoorntje,[1,3] Wouter P. te Rijdt,[1,3] Yvonne J. Vos,[1] Jan D.H. Jongbloed,[1] Conny M.A. van Ravenswaaij-Arts,[1] Richard Sinke,[1] Birgit Sikkema-Raddatz,[1] Wilhelmina S. Kerstjens-Frederikse,[1] Morris A. Swertz,[1,2] Lude Franke[1]

[1] University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, 9700 VB, the Netherlands

[2] University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, 9700 VB, the Netherlands

[3] Netherlands Heart Institute, Utrecht, the Netherlands

[4] These authors contributed equally to this work

[5] These authors contributed equally to this work

Corresponding author:
Lude Franke
E-mail: Lude@ludesign.nl

Keywords: Diagnostic sequencing; Variant prioritization; Human Phenotype Ontology; RNA sequencing; Network analysis; Gene co-regulation

Abstract

Clinical interpretation of exome and genome sequencing data remains challenging and time consuming, with many variants with unknown effects found in genes with unknown functions. Automated prioritization of these variants can improve the speed of current diagnostics and identify previously unknown disease genes. Here, we used 31,499 RNA-seq samples to predict the phenotypic consequences of variants in genes. We developed GeneNetwork Assisted Diagnostic Optimization (GADO), a tool that uses these predictions in combination with a patient's phenotype, denoted using HPO terms, to prioritize identified variants and ease interpretation. GADO is unique because it does not rely on existing knowledge of a gene and can therefore prioritize variants missed by tools that rely on existing annotations or pathway membership. In a validation trial on patients with a known genetic diagnosis, GADO prioritized the causative gene within the top 3 for 41% of the cases. Applying GADO to a cohort of 38 patients without genetic diagnosis, yielded new candidate genes for seven cases. Our results highlight the added value of GADO (www.genenetwork.nl) for increasing diagnostic yield and for implicating previously unknown disease-causing genes.

Introduction

With the increasing use of whole-exome sequencing (WES) and whole-genome sequencing (WGS) to diagnose patients with a suspected genetic disorder, diagnostic yield is steadily increasing [1]. Although our knowledge of the genetic basis of Mendelian diseases has improved considerably, the underlying cause remains elusive for a substantial proportion of cases. The diagnostic yield of genome sequencing varies from 8% to 70% depending on the patient's phenotype and the extent of genetic testing [2]. Sequencing all ~20,000 protein-coding genes by WES and entire genomes by WGS usually increases sensitivity but decreases specificity: it results in off-target noise and reveals many variants of uncertain clinical significance. In a study by Yang *et al.*, proband-only WES identified approximately 875 variants in each patient, even after removing low quality variants [3].

One strategy to manage the list of genetic variants is to perform trio analysis of samples from the proband and both of his or her biological parents to ascertain, for instance, whether a variant has *de novo* status [4]. Another strategy is to limit the analyses to a gene panel of Online Mendelian Inheritance in Men (OMIM) disease-annotated genes [5] or genes known to be directly related to the patient's phenotype. However, determining the actual disease-causing variant requires further variant filtering based on information about its predicted functional consequence, population frequency data, conservation, disease-specific databases (such as the Human Gene Mutation Database [6]), literature, and segregation analysis [7].

Several tools have been developed that aid in variant filtering and prioritization [8,9]. Annotation tools, such as VEP [10] and GAVIN [9], offer additional functionality that allows variants to be filtered according to their population frequency and variant class. Other tools use phenotype descriptions to rank potential candidates genes [11]. The phenotypes are typically described in a structured manner, e.g. using Human Phenotype Ontology (HPO) terms [12]. AMELIE (Automatic Mendelian Literature Evaluation), for example, prioritizes candidate genes by their likelihood of causing the patient's phenotype based on automated literature analysis [13]. However, this focus on what is known may inadvertently filter out variants in potential novel disease genes. Alternatively, the causative gene defect could be missed if a patient's phenotype differs from the features previously reported to be associated to a disease gene. Tools like Exomiser can identify novel human disease genes, as it prioritizes variants based on semantic phenotypic similarity between a patient's phenotype described by HPO terms and HPO-annotated diseases, Mammalian Phenotype Ontology (MPO)-annotated mouse and Zebrafish Phenotype Ontology (ZPO)-annotated fish models associated with each exomic candidate and/or its neighbors in an interaction network [14]. However, most available algorithms are based on existing knowledge on human disease genes, their orthologues in animal models, or well-described biological pathways (for a detailed review see [11]).

To overcome this, we hypothesized that co-regulation of expression data could be used to prioritize variants, including those in less well studied genes. We assumed that if a gene or a gene set is known to cause a specific disease or disease symptom, these genes will often have similar molecular functions or be involved in the same biological process or pathway. We reasoned that variants in genes with yet unknown function that are involved in the same biological pathway or co-regulated with known disease genes likely result in the same phenotype. In order to identify groups of genes with a related biological function, we used an expansive compendium of 31,499 RNA-sequencing (RNA-seq) gene expression samples to predict functions for genes with high accuracy.

We then developed a user-friendly tool that can prioritize variants in known *and* unknown genes based on our functional predictions, which we designated GeneNetwork Assisted Diagnostic Optimization (GADO). GADO ranks variants based on gene co-regulation in publicly available expression data of a wide range of tissues and cell types using HPO terms to describe a patient’s phenotype. To validate our prioritization method, we tested how well our method predicts disease-causing genes based on features described for each of the genes in the OMIM database. We then used exome sequencing data of patients with a known genetic diagnosis to benchmark GADO. Finally, we applied our methodology to previously inconclusive WES data and identified several genes that contain variants that likely explain the phenotype of the respective patients. Thus, we show that our methodology is successful in identifying variants in novel, potentially relevant genes explaining the patient’s phenotype.

Results

Gene prioritization using GADO

We have developed GADO to perform gene prioritizations using the phenotypes observed in patients denoted as HPO terms [15]. In combination with a list of candidate genes (i.e. genes harboring rare and possibly damaging variants), this results in a ranked list of genes with the most likely candidate genes on top (**Figure 1**a). The gene prioritizations are based on the predicted involvement of the candidate genes for the specified set of HPO terms. These predictions are made by analyzing public RNA-seq data from 31,499 samples (**Figure 1**b), resulting in a gene prediction score for each HPO term. These predictions are solely based on co-regulation of genes annotated to a certain HPO term with other genes. This makes it possible to also prioritize genes that currently lack any biological annotation.

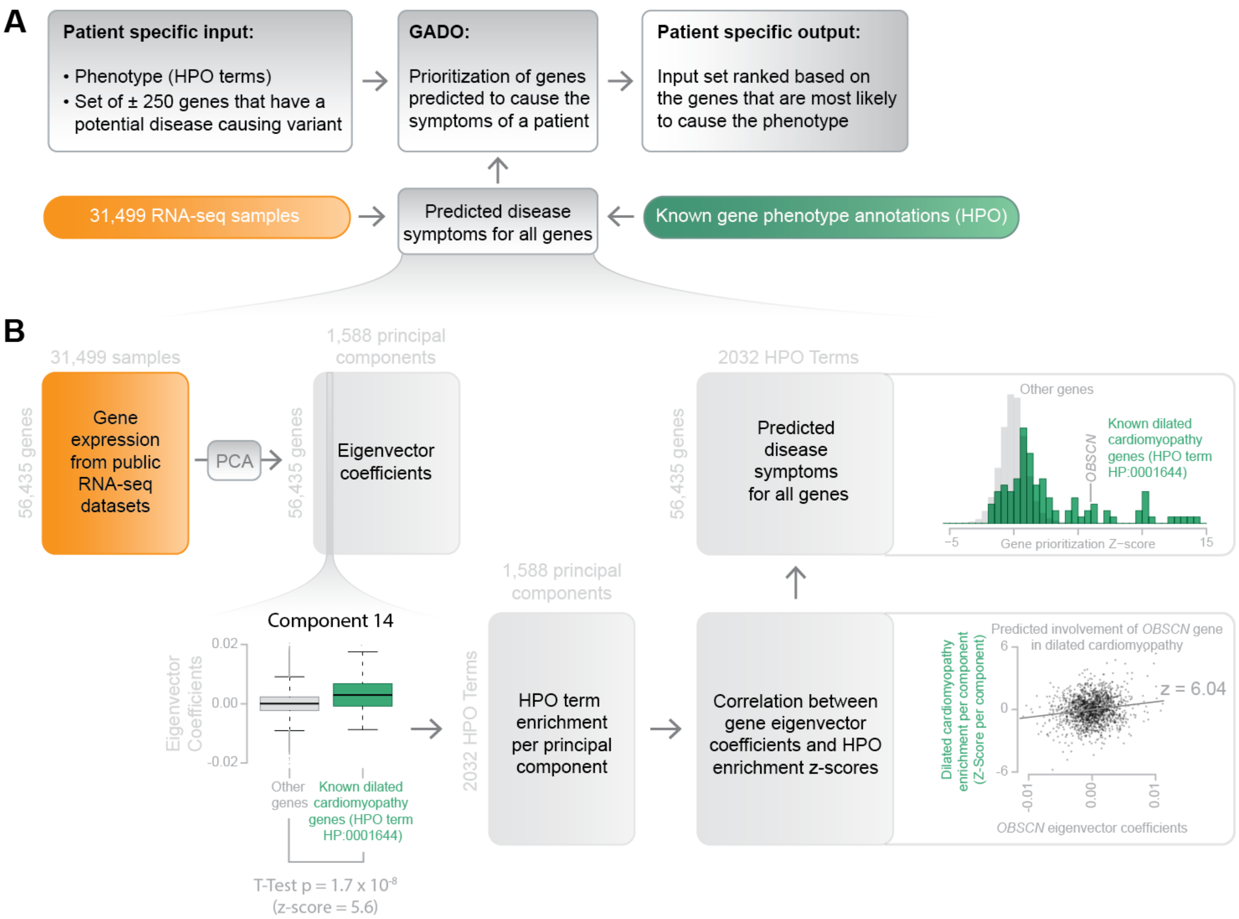

A
Patient specific input:
• Phenotype (HPO terms)
• Set of ± 250 genes that have a potential disease causing variant
GADO:
Prioritization of genes predicted to cause the symptoms of a patient
Patient specific output:
Input set ranked based on the genes that are most likely to cause the phenotype
31,499 RNA-seq samples
Predicted disease symptoms for all genes
Known gene phenotype annotations (HPO)
B
31,499 samples
56,435 genes
Gene expression from public RNA-seq datasets
PCA
1,588 principal components
56,435 genes
Eigenvector coefficients
2032 HPO Terms
56,435 genes
Predicted disease symptoms for all genes
Other genes
Known dilated cardiomyopathy genes (HPO term HP:0001644)
OBSCN
−5
Gene prioritization Z−score
15
Component 14
0.02
0.00
-0.02
Eigenvector Coefficients
Other genes
Known dilated cardiomyopathy genes (HPO term HP:0001644)
T-Test p = 1.7 x 10-8
(z-score = 5.6)
1,588 principal components
2032 HPO Terms
HPO term enrichment per principal component
Correlation between gene eigenvector coefficients and HPO enrichment z-scores
Predicted involvement of OBSCN gene in dilated cardiomyopathy
6
0
-6
Dilated cardiomyopathy enrichment per component (Z-Score per component)
z = 6.04
-0.01
0.00
0.01
OBSCN eigenvector coefficients

Figure 1: Schematic overview of GADO. *(a) Per patient, GADO requires a set of phenotypic features and a list of candidate genes (i.e. genes harboring rare alleles that are predicted to be pathogenic) as input. It then ascertains whether genes have been predicted to cause these features, and which ones are present in the set of candidate genes that has been provided as input. The predicted HPO phenotypes are based on the co-regulation of genes with sets of genes that are already known to be associated with that phenotype. (b) Overview of how disease symptoms are predicted using gene expression data from 31,499 human RNA-seq samples. A principal component analysis on the co-expression matrix results in the identification of 1,588 significant principal components. For each HPO term we investigate every component: per component we test whether there is a significant difference between eigenvector coefficients of genes known to cause a specific phenotype and a background set of genes. This results in a matrix that indicates which principal components are informative for every HPO term. By correlating this matrix to the eigenvector coefficients of every individual gene, it is possible to infer the likely HPO disease phenotype term that would be the result of a pathogenic variant in that gene.*

Public RNA-seq data acquisition and quality control

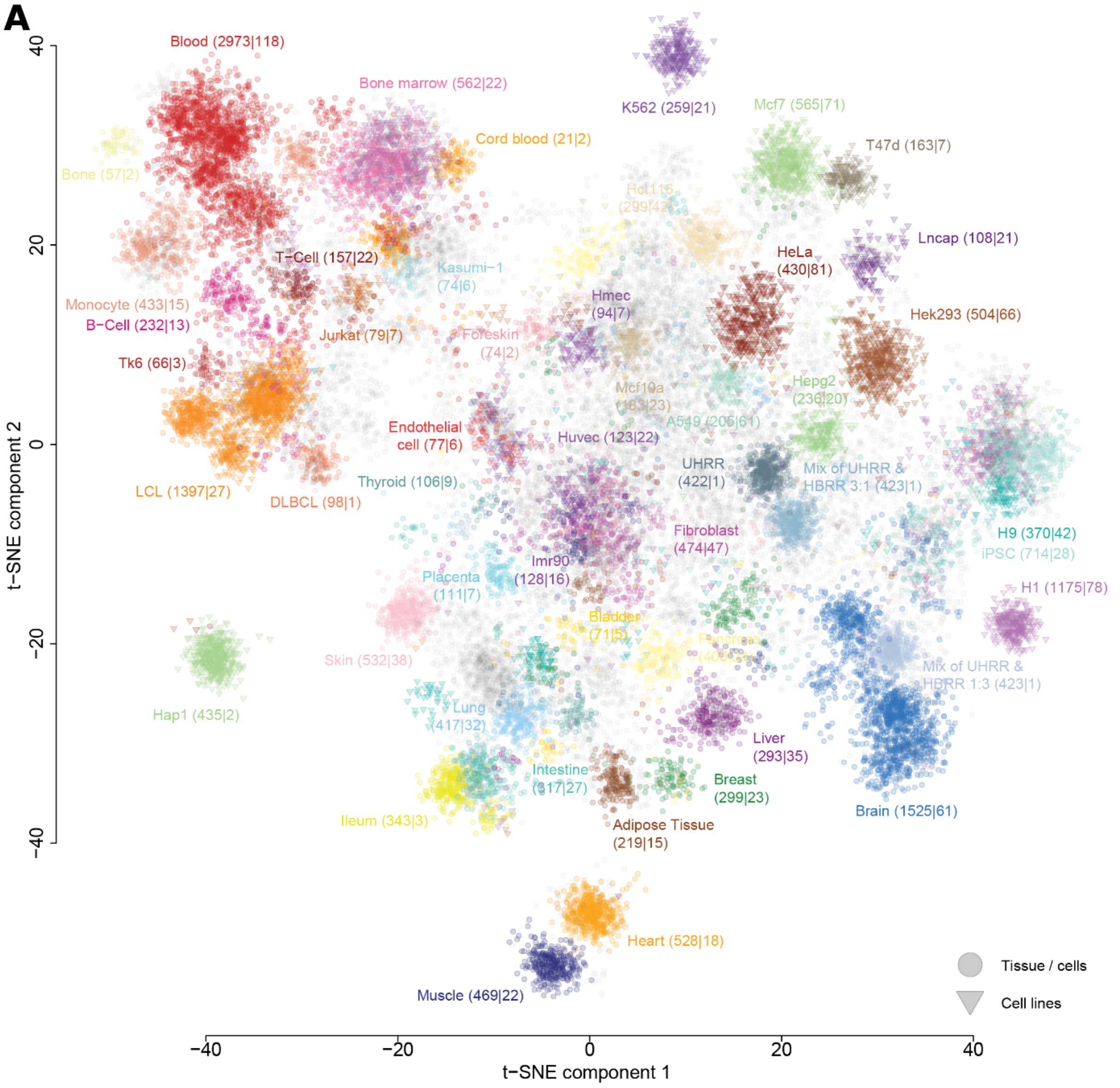
To predict functions of genes and HPO term associations, we downloaded all human RNA-seq samples publicly available in the European Nucleotide Archive (accessed June 30, 2016) (supplementary table 1) [16]. We quantified gene-expression using Kallisto [17] and removed samples for which a limited number of reads are mapped. We used a principal component analysis (PCA) on the correlation matrix to remove low quality samples and samples that were annotated as RNA-seq but turned out to be DNA-seq. In the end, we included 31,499 samples and quantified gene expression levels for 56,435 genes (of which 22,375 are protein-coding).

Although these samples are generated in many different laboratories, we previously observed that, after having corrected for technical biases, it is possible to integrate these samples into a single expression dataset [18]. We validated that this is also true for our new dataset by visualizing the data using t-Distributed Stochastic Neighbor Embedding (t-SNE). We labeled the samples based on cell-type or tissue and we observed that samples cluster together based on cell-type or tissue origin (**Figure 2**a). Technical biases, such as whether single-end or paired-end sequencing had been used, did not lead to erroneous clusters, which suggests that this heterogeneous dataset can be used to ascertain co-regulation between genes and can thus serve as the basis for predicting the functions of genes.

Prediction of gene HPO associations and gene functions

To predict HPO term associations and putative gene functions using co-regulation (**Figure 1**b), we used a method that we had previously developed and applied to public expression microarrays [19]. Since these microarrays only cover a subset of the protein-coding genes (n = 14,510), we decided to use public RNA-seq data instead. This allows for more accurate quantification of lower expressed genes and the expression quantification of many more genes, including a large number of non-protein-coding genes. [20].

We applied this prediction methodology [19] to the HPO gene sets and also to Reactome [21], KEGG pathways [22], Gene Ontology (GO) molecular function, GO biological process and GO cellular component [23] gene sets. For 5,088 of the 8,657 gene sets (59%) with at least 10 genes annotated, the gene function predictions had significant predictive power (see materials and methods). For the 8,657 gene sets with at least 10 genes annotated, the median predictive power, denoted as Area Under the Curve (AUC), ranged between 0.73 (HPO) to 0.87 (Reactome) (**Figure 2**b).



B

Database	Number of gene sets	Gene sets ≥ 10 genes	Gene sets with significant predictive power	Median AUC
Reactome	2,143	1,388	1,150	0.87
GO molecular function	4,070	726	398	0.82
GO biological process	11,753	2,576	1,115	0.82
GO cellular component	1,609	500	370	0.84
KEGG	186	186	168	0.84
HPO	7,920	3,281	1,887	0.73

Figure 2: A compendium of gene expression profiles that can be used for gene function prediction *(a) 31,499 RNA-seq samples derived from many different studies show coherent clustering after correcting for technical biases. Generally, samples originating from the same tissue, cell-type or cell-line cluster together. The two axes denote the first t-SNE components. (b) Gene co-expression information of 31,499 samples is used to predict gene functions. We show the prediction accuracy for gene sets from different databases. AUC, Area Under the Curve, GO, Gene Ontology, HPO, Human Phenotype Ontology.*

Prioritization of known disease genes using the annotated HPO terms

Once we had calculated the prediction scores of HPO disease phenotypes, we leveraged these scores to prioritize genes found by sequencing the DNA of a patient. For each individual HPO term–gene combination, we calculated a prediction z-score that can be used to rank genes. In practice, however, patients often present with not one feature but a combination of multiple features. Therefore, we combined the z-scores for each HPO term [24] to generate an overall z-score that explains the full spectrum of features in a patient. GADO uses these combined z-scores to prioritize the candidate genes: the higher the combined z-score for a gene, the more likely it explains the patient's phenotype.

Because many HPO terms have fewer than 10 genes annotated, and since we were unable to make significant predictions for some HPO terms, certain HPO terms are not suitable to use for gene prioritization. We solved this problem by taking advantage of the way HPO terms are structured. Each term has at least one parent HPO term that describes a more generic phenotype and thus has also more genes assigned to it. Therefore, if an HPO term cannot be used, GADO will make suggestions for suitable parental terms (supplementary figure 1).

To benchmark our prioritization method, we used the OMIM database [5]. We tested how well our method was able to retrospectively rank disease-causing genes listed in OMIM based on the annotated symptoms of these diseases. We took each OMIM disease gene (n = 3,382) and used the associated disease features (15 per gene on average) as input for GADO. What we found was that for 49% of the diseases GADO ranks the causative gene in the top 5% (**Figure 3**a, b). Moreover, we observed a statistically significant difference between the performance of GADO on true gene-phenotype combinations and its performance using a random permutation of gene-phenotype combinations (p-value = 2.16 × 10^{-532}).

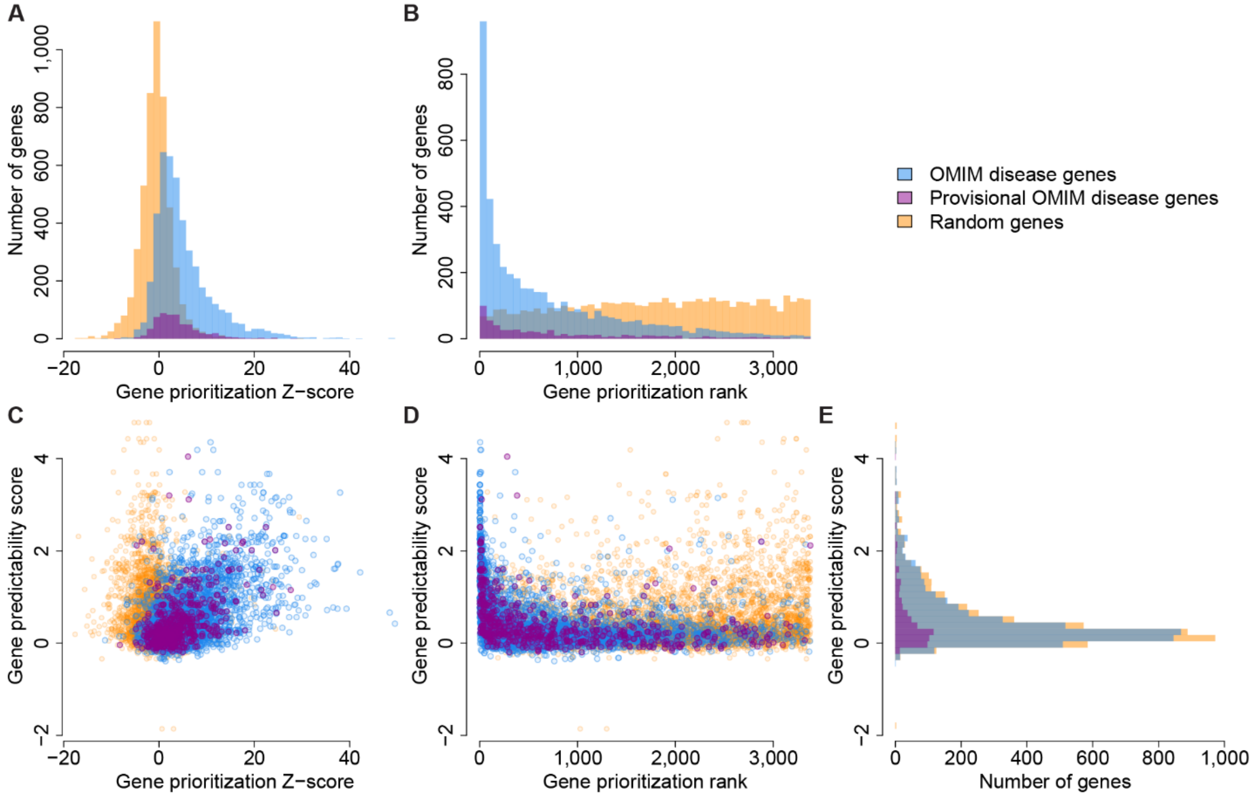


Figure 3: Performance of disease gene prioritization compared to random permutation. *(a) OMIM disease genes and provisional disease genes have significantly stronger z-scores compared to permuted disease genes (T-test p-values: 2.16×10^{-532} & 5.38×10^{-80}, respectively). We also observe that the predictions of the provisional OMIM genes are, on average, weaker than the other OMIM disease genes (T-test p-value: 1.89×10^{-7}). (b) Ranking the disease based on z-scores shows GADO's ability to prioritize the causative gene for a disease among all OMIM genes. For 49% of the disorders the causative gene is ranked in the top 5%. (c) We observe a clear relation between the prioritization z-scores and the gene predictability scores (Pearson r = 0.54). We don't observe this relation in the permuted results. (d) GeneNetwork performs best for genes with high predictability scores. (e) The different groups have similar distributions of gene predictability scores.*

Gene predictability scores explains performance differences between genes

For some combinations of genes and HPO terms listed in OMIM, GADO could not establish the gene-phenotype combination (**Figure 3**). For example, variants in *SLC6A3* are known to cause infantile Parkinsonism-dystonia (MIM 613135) [25–27], but GADO was unable predict the annotated HPO terms related to the Parkinsonism-dystonia for this gene. This may, however, be due to very low expression levels of *SLC6A3* in most tissues except specific brain regions [28].

To better understand why we can't predict HPO terms for all genes, we used the Reactome, GO and KEGG prediction scores. Jointly these databases comprise thousands of gene sets. Since these databases describe such a wide range of biology, we assumed that if a gene does not show any prediction signal for any gene set in these databases, gene co-expression is probably not informative for this gene. To quantify this, we calculated, per gene, the average skewness of the z-score distribution of the Reactome, GO and KEGG gene sets. From this we were able to derive a 'gene predictability score' for every gene that is independent of whether this gene is already known to play a role in any a disease or pathway (**Figure 3**c, d, e). We then ascertained whether these 'gene predictability scores' are correlated with the prediction z-score of the OMIM diseases, and found a strong correlation (Pearson r = 0.54, p-value = 1.14×10^{-332}) between the gene predictability scores and GADO's ability to identify a known disease gene (**Figure 3**c).

To investigate why some genes have a high 'gene predictability score' but low prediction performance, we scored a set of genes known to cause cardiomyopathy (CM) for the amount of literature evidence that these genes cause CM. We found several genes for which the prediction score for the CM phenotype is lower than expected based on the gene predictability scores (supplementary figure 2a). Pathogenic variants in the *TTR* gene implicated in hereditary amyloidosis (MIM 105210) [29], for instance, cause accumulation of the transthyretin protein in different organ systems, including the heart, resulting in CM. However, this gene is primarily expressed in the liver. Therefore, its disease mechanism is different from other mechanisms resulting in CM, as many inherited CMs are caused by deleterious variants in genes highly expressed in the heart and directly affecting the function of the cardiac sarcomere. Therefore, the phenotypic function prediction for this gene may be worse than we would expect based on the predictability score. We performed a similar analysis using the HPO term ‘dilated cardiomyopathy’ and observed a low prediction performance for the *TMPO* gene, despite a high gene predictability score (supplementary figure 2b). Previously, this gene was reported to be related to dilated cardiomyopathy (DCM) and listed as such by OMIM. However, recent reclassification of the reported variants using the ExAC data revealed that the reported variant was far too common to be causative for DCM [30].

Benchmarking GADO using solved cases with realistic phenotyping

Although *in silico* benchmarking demonstrated the potential of GADO, it used all annotated HPO terms for a disease. In practice, however, patients may only present with a limited number of the annotated features. To perform a validation that was a more realistic reflection of clinical practice, we used exome sequencing data of 83 patients with a known genetic diagnosis. We used their phenotypic features as listed in their medical records prior to the genetic diagnosis (supplementary table 2). On average, per patient, GADO yielded 56 possible disease-causing genes with variants that are rare and predicted to be deleterious. In 41% of the patients the actual causative gene was ranked in the top 3 and in 50% of the cases it was in the top 5 (mean rank 10) (**Figure 4**a).

Clustering of HPO terms

In addition to ranking potentially causative genes based on a patient’s phenotype, we observed that GADO can be used to cluster HPO terms based on the genes that are predicted to be associated to these HPO terms. This can help identify pairs of symptoms that often occur together, as well as symptoms that rarely co-occur, and we actually observed this for a patient suspected of having two different diseases. This patient is diagnosed with a glycogen storage disease, GSD type Ib, caused by compound heterozygous variants in *SLC37A4* (MIM 602671) and DCM that is probably caused by a truncating variant in *TTN* (MIM 188840). Clustering of the assigned HPO terms placed the phenotypic features related to GSD type Ib (‘leukopenia’ (HP:0001882) and ‘inflammation of the large intestine’ (HP:0002037)) together, while

Cardiomyopathy (HP:0001638) was only weakly correlated to these specific features (**Figure 4**b).

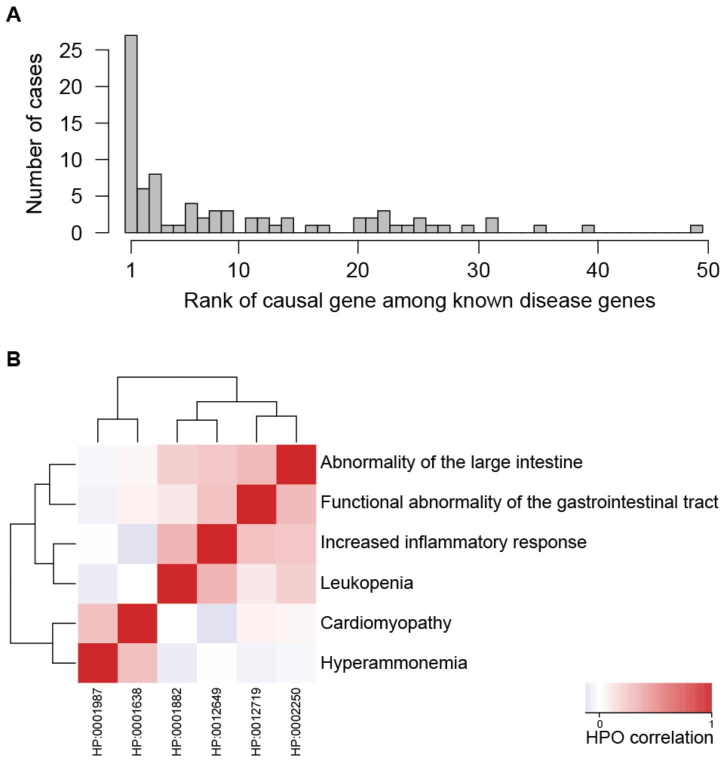


Figure 4: Performance of GeneNetwork on solved cases *(a) Rank of the known causative gene among the candidate disease causing variants. (b) Our cohort contained a case with two distinct conditions, and clustering showed the HPO terms of the same disease are closest to each other. Note, the HPO term "Inflammation of the large intestine" did not yield a significant prediction profile and therefore the parent terms "Abnormality of the large intestine", "Increased inflammatory response" and "Functional abnormality of the gastrointestinal tract" where used for this case.*

Reanalysis of previously unsolved cases

To assess GADO's ability to discover new disease genes, we applied it to data from 38 patients who are suspected to have a Mendelian disease but who have not had a genetic diagnosis. All patients had undergone prior genetic testing (WES with analysis of a gene panel according to their phenotype, supplementary table 3). On average three genes had a z-score ≥ 5 (which we used as an arbitrary cut-off and that correspond to a p-value of 5.7 X 10^{-7}) and were further assessed. In seven cases, we identified variants in genes not associated to a disease in OMIM or other databases, but for which we could find literature or for which we gained functional evidence implicating their disease relevance (**Table 1**). For example, we identified two cases with DCM with rare compound heterozygous variants in the *OBSCN* gene (MIM 608616) that are predicted to be damaging. In literature, inherited variant(s) in *OBSCN*, encoding obscurin, are associated with hypertrophic CM [31] and DCM [32]. Furthermore, obscurin is a known interaction partner of titin (TTN), a well-known DCM-related protein [31]. Another example came from a patient with ichthyotic peeling skin syndrome, which is caused by a damaging variant in *FLG2 (*MIM 616284). We recently published this case where we prioritized this gene using an alpha version of GADO [33].

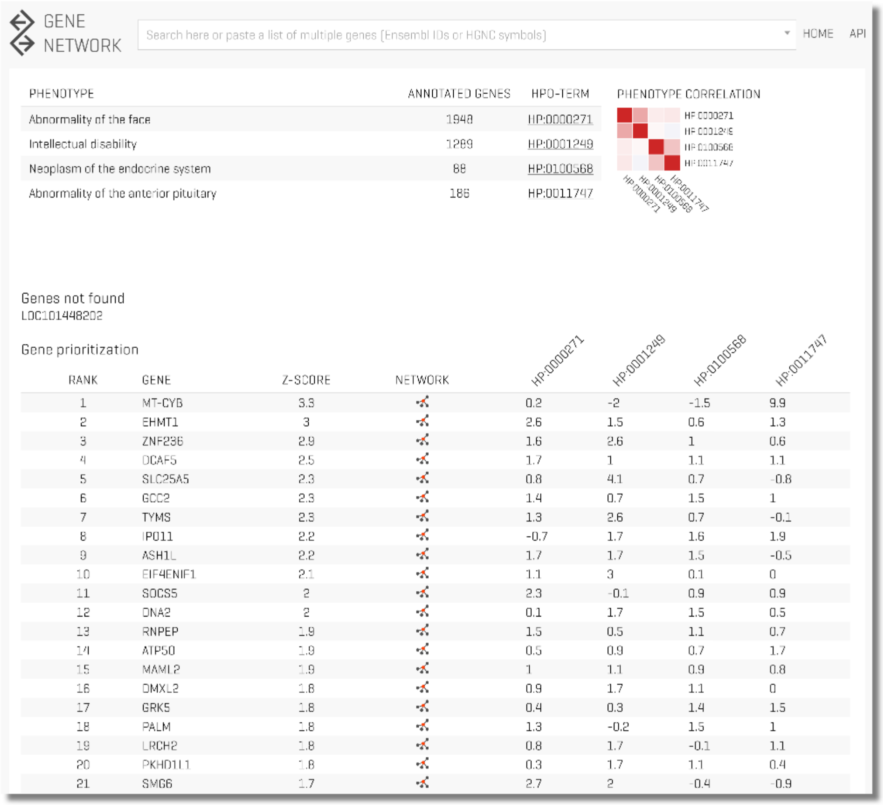
HPO terms used	Number of genes with candidate variant	Number of genes with z ≥ 5	Candidate gene	Variants	CADD scores	GnomAD minor allele frequency	Supporting papers	Expression in relevant tissue
HP:0001644	247	5	*OBSCN*	NM_001098623.2: c.[15037C>T]; [20963delC]	24.8 25.2	8.0 x 10^{-5} 1.7 x 10^{-3}	[31, 32]	Yes
HP:0001644	226	3	*OBSCN*	NM_001098623.2: c.[5545C>T]; [22384+3_22384+21del]	14.7 7.8	3.2 x 10^{-4} 0	[31, 32]	Yes
HP:0008066 HP:0008064	359	3	*FLG2*	NM_001014342.2: c.[632C>G]; [632C>G]	35.0 35.0	1.1 x 10^{-5} 1.1 x 10^{-5}	[34]	Yes
HP:0001263 HP:0001249 HP:0000717 HP:0000708 HP:0002167 HP:0002360 HP:0000664	206	12	*INO80*	NM_017553.2: c.[898C>T]	34	0	[35, 36]	Yes
HP:0001644	346*	2	MB	NM_00203377.1: c.[214G>A]	22.4	3.6 x 10^{-5}	[37]	Yes
HP:0001644	126*	1	*SYNPO2L***	NM_001114133.2: c.[473G>A]	24.1	5.4 x 10^{-4}	[38]	Yes
HP:0001638	336	4	*NRAP***	NM_001261463.1: c.[4648C>T]	20.4	8.7 x 10^{-4}	[39]	Yes

Table 1: unsolved cases with new candidate genes. *Out of the 38 unsolved patients investigated, we identified candidate genes in seven patients. For these genes we have found literature that indicates these genes fit the phenotype of these patients or for which we gained functional evidence implicating their disease relevance. *These variants where pre-filtered for family segregation. **The variants in these genes do not fully explain the phenotype but are likely contributing to the phenotype.*

www.genenetwork.nl

All analyses described in this paper can be performed using our online toolbox at www.genenetwork.nl. Users can perform gene prioritizations using GADO by providing a set of HPO terms and a list of candidate genes (**Figure 5**a). Per gene, it is also possible to download all prediction scores for the HPO terms and pathways. Our co-regulation scores between genes can be used for clustering. Furthermore, the predicted pathway and HPO annotations of genes can be used to perform function enrichment analysis (**Figure 5**b). We also support automated queries to our database.

A



GENE NETWORK

Search here or paste a list of multiple genes (Ensembl IDs or HGNC symbols)

HOME API

PHENOTYPE	ANNOTATED GENES	HPO-TERM
Abnormality of the face	1948	HP:0000271
Intellectual disability	1289	HP:0001249
Neoplasm of the endocrine system	88	HP:0100568
Abnormality of the anterior pituitary	186	HP:0011747

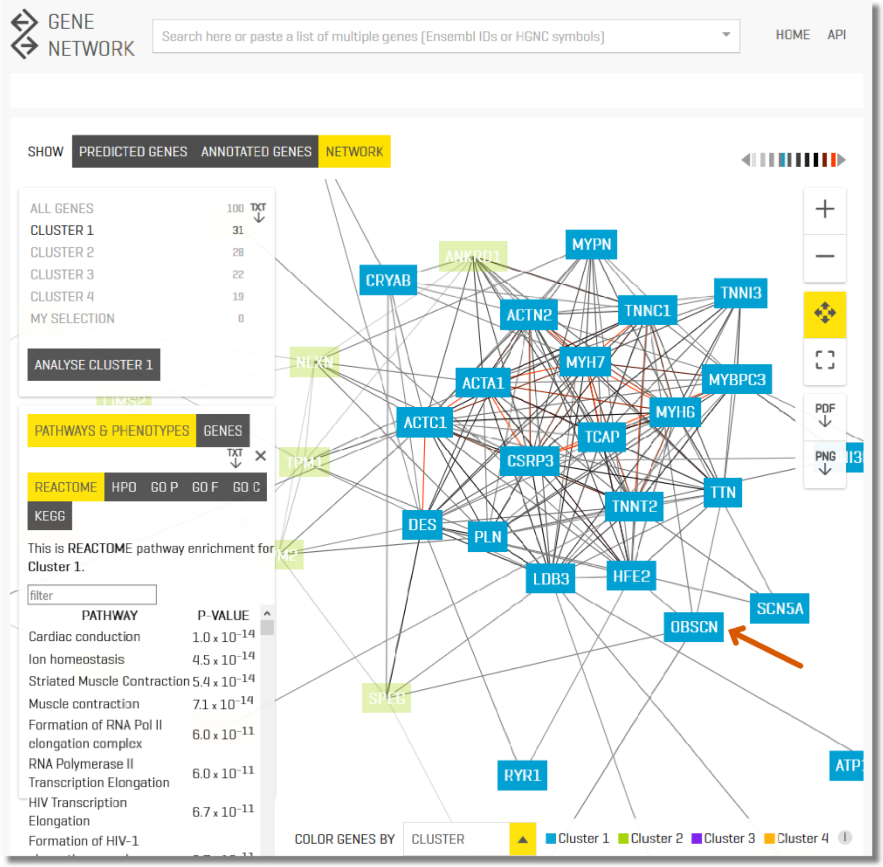
PHENOTYPE CORRELATION

Genes not found
LOC101448202

Gene prioritization

RANK	GENE	Z-SCORE	NETWORK	HP:0000271	HP:0001249	HP:0100568	HP:0011747
1	MT-CYB	3.3		0.2	-2	-1.5	9.9
2	EHMT1	3		2.6	1.5	0.6	1.3
3	ZNF236	2.9		1.6	2.6	1	0.6
4	DCAF5	2.5		1.7	1	1.1	1.1
5	SLC25A5	2.3		0.8	4.1	0.7	-0.8
6	GCC2	2.3		1.4	0.7	1.5	1
7	TYMS	2.3		1.3	2.6	0.7	-0.1
8	IPO11	2.2		-0.7	1.7	1.6	1.9
9	ASH1L	2.2		1.7	1.7	1.5	-0.5
10	EIF4ENIF1	2.1		1.1	3	0.1	0
11	SOCS5	2		2.3	-0.1	0.9	0.9
12	DNA2	2		0.1	1.7	1.5	0.5
13	RNPEP	1.9		1.5	0.5	1.1	0.7
14	ATP5O	1.9		0.5	0.9	0.7	1.7
15	MAML2	1.9		1	1.1	0.9	0.8
16	DMXL2	1.8		0.9	1.7	1.1	0
17	GRK5	1.8		0.4	0.3	1.4	1.5
18	PALM	1.8		1.3	-0.2	1.5	1
19	LRCH2	1.8		0.8	1.7	-0.1	1.1
20	PKHD1L1	1.8		0.3	1.7	1.1	0.4
21	SMG6	1.7		2.7	2	-0.4	-0.9

B



GENE NETWORK

Search here or paste a list of multiple genes (Ensembl IDs or HGNC symbols)

HOME API

SHOW PREDICTED GENES ANNOTATED GENES NETWORK

ALL GENES	100	TXT
CLUSTER 1	31	
CLUSTER 2	28	
CLUSTER 3	22	
CLUSTER 4	19	
MY SELECTION	0	

ANALYSE CLUSTER 1

PATHWAYS & PHENOTYPES GENES

TXT ×

REACTOME HPO GO P GO F GO C KEGG

This is REACTOME pathway enrichment for Cluster 1.

filter

PATHWAY	P-VALUE
Cardiac conduction	1.0×10^{-14}
Ion homeostasis	4.5×10^{-14}
Striated Muscle Contraction	5.4×10^{-14}
Muscle contraction	7.1×10^{-14}
Formation of RNA Pol II elongation complex	6.0×10^{-11}
RNA Polymerase II Transcription Elongation	6.0×10^{-11}
HIV Transcription Elongation	6.7×10^{-11}
Formation of HIV-1	

COLOR GENES BY CLUSTER

🟦 Cluster 1 🟩 Cluster 2 🟪 Cluster 3 🟧 Cluster 4

***Figure 5: www.genenetwork.nl** (a) Prioritization results of one of our previously solved cases. This patient was diagnosed with Kleefstra syndrome. The patient only showed a few of the phenotypic features associated with Kleefstra syndrome and additionally had a neoplasm of the pituitary (which is not associated with Kleefstra syndrome). Despite this limited overlap in phenotypic features, GADO was able to rank the causative gene (EHMT1) second. Here, we also show the value of the HPO clustering heatmap, the two terms related to the neoplasm cluster separately from the intellectual disability and the facial abnormalities that are associated to Kleefstra syndrome. (b) Clustering of a set of genes allowing function / HPO enrichment of all genes or specific enrichment of automatically defined sub clusters. Here we loaded all known DCM genes and OBSCN, and we focus on a sub-cluster of genes containing OBSCN (highlighted by the arrow). We see that it is strongly co-regulated with many of the known DCM genes. Pathway enrichment of this sub-cluster reveals that these genes are most strongly enriched for the muscle contraction Reactome pathway. DCM, Dilated Cardiomyopathy.*

Discussion

Prioritizing genes from WES or WGS data remains challenging. To meet this challenge, we developed GADO, a novel tool to prioritize genes based on the phenotypic features of a patient. Since the classification of variants is labor-intensive, prioritization of the most likely candidate variants saves time in the diagnostic process.

Importantly, GADO can also aid in the discovery of currently unknown disease genes. The main advantage of our methodology is that it does not rely on any prior knowledge about disease-gene annotations. Instead, we used predicted gene functions based on co-expression networks extracted from a large compendium of publicly available RNA-seq samples. RNA-seq has previously shown to be very helpful to accurately quantify expression levels of lowly expressed genes and non-coding genes [18]. To evaluate our diagnostic algorithm, we developed a testing scenario based on simulated patients presenting with all clinical features listed in OMIM for a certain disease or syndrome. This validation test showed that for 49% of the diseases the causative gene ranks in the top 5%. We also investigated the OMIM “provisional” category of genes for which there is limited evidence. Both the OMIM disease-gene annotation and the provisional annotations perform significantly better than a random permutation. While we do find a small but significant difference in prediction performance between the provisionally annotated genes and the more established disease associated genes, we conclude, based on our findings, that these provisional OMIM annotations are generally of similar reliability to the other OMIM disease annotations.

Benchmarking on sequence data of patients with a known genetic diagnosis revealed that GADO returned the real causative variant within the top 3 results for 41% of the samples, indicating the potential power of GADO for a large number of diseases. Finally, in seven patients, GADO was able to identify potential novel disease genes that are strong candidates based on literature or functional evidence. For other cases we have identified genes with a strong prediction score harboring variants that might explain the phenotype. However, since very little is known about these genes it is not yet possible to draw firm conclusions. Hopefully this will become possible in the near future through initiatives like Genematcher [40].

Potential to discover novel human disease genes

Over the last decade, several computational tools have been developed to prioritize variants in genes. Some, such as GAVIN, focus on variant filtering and prioritization based on deleteriousness scores, allele frequency and inheritance model [9]. Other methods measure the similarity between the clinical manifestations observed in a patient and those representing each of the diseases in a database or literature. Exomiser is closely related to GADO as it prioritizes genes based on specified HPO terms and also infers HPO annotation for unknown genes [14]. The gene prioritization by Exomiser is based on the effects of orthologs in model organisms and applies a guilt-by-association method using protein-protein associations provided by STRING [41]. Exomiser performs better than GADO in ranking known disease-causing genes (supplementary figure 3, supplementary table 4) and is also able to identify potential new genes in human disease. However, Exomiser has a limitation in that only a subset of the protein-coding genes has orthologous genes in other species for which a knockout model also exists. Additionally, the used STRING interactions are biased towards well studied genes and rely heavily on existing annotations to biological pathways (supplementary figure 4). There are however, still 3,922 protein-coding genes that are not currently annotated in any of the databases we used, and there are even more non-coding genes for which the biological function or role in disease is unknown. Since GADO does not rely on prior knowledge, it can be used to prioritize variants in both coding *and* non-coding genes (for which no or limited information is available). GADO thus enables the discovery of novel human disease genes and can complement existing tools in analyzing the genomic data of patients who have a broad spectrum of phenotypic abnormalities.

Limitations

The gene predictability score indicates for which genes we can reliably predict phenotypic associations and for which genes we cannot based on gene co-regulation. This score gives insight into which genes are expected to perform poorly in our prioritization. We found strong correlation between these gene predictability scores and the gene prioritization z-scores. Thus, genes with a high predictability score have more accurate HPO term predictions. However, since our predictions primarily rely on co-activation patterns that we identified from RNA-seq data, our method does not perform well for genes where gene-expression patterns are not informative of their function. This could, for instance, be the case for proteins relying heavily on post-translation modifications for regulation or genes for which different transcripts have distinct functions. This last limitation can potentially be overcome by predicting HPO-isoform associations by using transcript-based expression quantification.

Insufficient statistical power to obtain accurate predictions may be another explanation for the low predictability scores of certain genes. This may be true for genes that are poorly expressed or expressed in only a few of the available RNA-seq samples. The latter issue we expect to overcome in the near future as the availability of RNA-seq data in public repositories is rapidly increasing. Initiatives such as Recount enable easy analysis on these samples [42], allowing us to update our predictions in the future, thereby increasing our prediction accuracy.

For some genes we are unable to predict annotated disease associations despite having a high gene predictability scores. Some genes, such as *TTR,* simply act in a manner unique to a specific phenotype. Other genes, such as *TMPO,* turned out to be false positive disease associations. These examples show that our gene predictability score has the potential to flag genes acting in a unique manner as well as genes that might be incorrectly assigned to a certain disease or phenotype.

We noted that the median prediction performance of HPO terms is lower compared to the other gene sets databases used in our study, such as Reactome. This may be due to the fact that phenotypes can arise by disrupting multiple distinct biological pathways. For instance, DCMs can be caused by variants in sarcomeric protein genes, but also by variants in calcium/sodium handling genes or by transcription factor genes [43]. As our methodology makes guilt-by-association predictions based on whether genes are showing similar expression levels, the fact that multiple separately working processes are related to the same phenotype can reduce the accuracy of the predictions (although it is often still possible to use these predictions as the DCM HPO phenotype prediction performance AUC = 0.76).

Complexity

Given that nearly 5% of patients with a Mendelian disease have another genetic disease [44], it is important to consider that multiple genes might each contribute to specific phenotypic effects. Clinically, it can be difficult to assess if a patient suffers from two inherited conditions, which may hinder variant interpretation based on HPO terms. We showed that GADO can disentangle the phenotypic features of two different diseases manifesting in one patient by correlating and subsequently clustering the profiles of HPO terms describing the patient's phenotype. If the HPO terms observed for a patient do not correlate, it is more likely that they are caused by two different diseases. An early indication that this might be the case for a specific patient can simplify subsequent analysis because the geneticist or laboratory specialist performing the variant interpretation can take this in consideration. GADO also facilitates separate prioritizations on subsets of the phenotypic features.

Conclusion

Connecting variants to disease is a complex multistep process. The early steps are usually highly automated, but the final most critical interpretations still rely on expert review and human interpretation. GADO is a novel approach that can aid users in prioritizing genes using patient-specific HPO terms, thereby speeding-up the diagnostic process. It prioritizes variants in coding *and* non-coding genes, including genes for which there is no current knowledge about their function and those that have not been annotated in any ontology database. This gene prioritization is based on co-regulation of genes identified by analyzing 31,499 publicly available RNA-seq samples. Therefore, in contrast to many other existing prioritization tools, GADO has the capacity to identify novel genes involved in human disease. By providing a statistical measure of the significance of the ranked candidate variants, GADO can provide an indication for which genes its predictions are reliable. GADO can also detect phenotypes that do not cluster together, which can alert users to the possible presence of a second genetic disorder and facilitate the diagnostic process in patients with multiple non-specific phenotypic features. GADO can easily be combined with any filtering tool to prioritize variants within WES or WGS data and can also be used in gene panels such as PanelApp [45]. GADO is freely available at www.genenetwork.nl to help guide the differential diagnostic process in medical genetics.

Materials and Methods

Gene co-regulation and function predictions

We used publicly available RNA-seq samples from the European Nucleotide Archive (ENA) database [46] to predict gene functions and gene-HPO term associations. After processing and quality control we included 31,499 sample for which we have expression quantification on 56,435 genes (supplementary methods 1). We performed a PCA on the gene correlation matrix and selected 1,588 reliable principal components (PCs) (Cronbach's Alpha ≥ 0.7).

We used the eigenvectors of these 1,588 PCs to predict gene functions and to predict HPO term associations [19]. We applied this methodology to the gene sets described by terms in the following databases: Reactome and KEGG pathways, Gene Ontology (GO) molecular function, GO biological process and GO cellular component terms and finally to HPO terms. We excluded terms for which fewer than 10 genes are annotated because predictions for smaller groups of genes are less accurate and might be misleading. Predictions were made for 8,657 gene sets in total.

The following steps were taken to obtain the gene prediction scores per gene set (**Figure 1**). First, for each PC, a student's T-test was conducted between the eigencoefficients of the genes annotated to a particular gene set and a group of genes serving as a background. This background consisted of the genes annotated to any term in a specific database, excluding those annotated to the current term. Second, the resulting p-values of the T-test were transformed into a z-score, which indicate to which extend each PC represents a part of the biology underlying a gene set. This is done for each PC, resulting in a profile how important each PC is for a gene set. Finally, to predict which genes can be associated to a particular gene set, we correlated the 1,588 T-test z-scores for that gene set (as calculated above) with the 1,588 eigenvector coefficients of a gene. The p-value of this correlation indicates the fit between a gene and a pathway / HPO term, these p-values were transformed to predictions z-scores. When a gene was already explicitly annotated to a gene-set and we wanted to predict whether that gene is involved in that gene set, then there is a small circular bias as the predictions profile of this set was partly calculated based on this gene. To remove this bias, the 1,588 z-scores for a gene set were first re-calculated while assuming this gene is not involved in that gene set, after which the gene prediction was made.

To determine the accuracy of our predictions we assessed our ability to predict back known gene set annotations. For each gene-set, we calculated an Area Under the Curve (AUC), using a Mann-Whitney U test, on the predictions z-scores of the genes that are part of a set versus those that are not part of a set. These AUCs indicate how accurate the predictions were, with an AUC of 1 indicating perfect predictions and an AUC of 0.5 indicating no predictive power. The average AUC for each category was calculated based on all gene sets with at least 10 annotated genes and with a p-value ≤ 0.05 (Bonferroni corrected for the number of pathways in a database).

Gene predictability scores

To explain why for some genes we cannot predict known HPO annotation, we have established a gene predictability score. We have calculated this gene predictability using the prioritization z-scores based on Reactome, GO and KEGG. For each gene and for each database we calculated the skewness in the distribution of the prioritization z-scores of the gene sets. We used the average skewness as the gene predictability score.

GADO predictions

To identify potential causative variants in patients, we used HPO terms to describe a patient’s features. We only used the HPO terms which have significant predictive power (based on the p-value of U test to calculate the AUC). If the predictions for a patient's HPO term were not significant, the parent/umbrella HPO terms were used (supplementary figure 1). The online GADO tool suggests the parent terms from which the user can then select which terms should be used in the analysis. The gene prediction z-scores for an HPO term were used to rank the genes. If a patient's phenotype was described by more than one HPO term, a meta-analysis was conducted. In these cases a weighted z-score was calculated by adding the z-scores for each of the patient's HPO terms and then dividing by the square root of the number of HPO terms [24]. The genes with the highest combined z-scores are predicted to most likely candidate causative genes for a patient. This analysis can be conducted at: https://www.genenetwork.nl.

Validation of disease-gene predictions

To benchmark our method we used the OMIM morbid map [5] downloaded on March 26, 2018, containing all disease-gene-phenotype entries. From this list, we extracted the disease-gene associations, excluding non-disease and susceptibility entries. We extracted the provisional disease-gene associations separately. For each disease in OMIM, we used GADO to determine the rank of the causative gene among all genes in the OMIM morbid map. For this we used all phenotypes annotated to the OMIM disease. If any of the HPO terms did not have significant predictive power, the parent terms were used.

To determine if these distributions were significantly different from what we expect by chance, we permuted the data. We replaced the existing gene-OMIM annotation but assigned every gene to a new disease (keeping the phenotypic features for a disease together), assuring that the randomly selected gene was not already annotated to any of the phenotypes of the original gene.

Cohort of previously solved cases

To test if GADO could help prioritize genes that contain the causative variant, we used 83 samples of patients who were previously genetically diagnosed through whole exome analysis or gene panel analysis. These samples encompass a wide variety of different Mendelian disorders (supplementary table 2). To assess which genes harbor potentially causative variants, we first called and annotated the variants from the exome sequencing files (Supplementary methods 3). For 11 of the previously solved cases, GAVIN did not flag the causative variant as a candidate. To be able to include these samples in our GADO benchmark, we added the causative genes for these cases manually to the candidate list.

The phenotypic features of a patient were translated into HPO terms, which were used as input to GADO. Here we only used features reported in the medical records prior to the molecular diagnosis. If any of the HPO terms did not have significant predictive power, the parent terms were used. From the resulting list of ranked genes, the known disease genes harboring a potentially causative variant were selected. Next, we determined the rank of the gene with the known causative variant among the selected genes. If a patient harbored multiple causative variants in different genes, in case of di-genic inheritance or two inherited conditions, the median rank of these genes was reported (supplementary table 2).

Unsolved cases cohorts

In addition to the patients with a known genetic diagnosis, we tested 38 unsolved cases (supplementary table 3). These are patients with mainly cardiomyopathies or developmental delay. All patients were previously investigated using exome sequencing, by analyzing a gene panel appropriate for their phenotype. To allow discovery of potential novel disease genes, we used GADO to rank genes with candidate variants (Supplementary methods 3). For genes with a prediction z-score ≥ 5, a literature search for supporting evidence was performed to assess whether these genes are likely candidate genes.

Website

To make our method and data available we have developed a website available at www.genenetwork.nl that can be used to run GADO, lookup gene functions predictions, visualize networks using co-regulations scores and perform function enrichments of sets of genes (Supplementary methods 4).

Description of Supplemental Data

Supplementary methods 1. Processing and quality control of public RNA-seq data

Supplementary methods 2. Benchmark comparison with Exomiser

Supplementary methods 3. Variant calling and processing of benchmark samples

Supplementary methods 4. GeneNetwork website

Supplementary figure 1. Selection of parent HPO term if GADO does not have significant predictive power for query term

Supplementary figure 2. Comparison of GADO performance with the level of evidence for each cardiomyopathy-related gene

Supplementary figure 3. Comparison between GADO and Exomiser rankings

Supplementary figure 4. Correcting for biases in co-expression networks

Supplementary figure 5. Histogram of the gene types included in our analyses

Supplementary figure 6. PCA plot of 36,761 samples

Supplementary figure 7. Investigation of principal components capturing technical biases

Supplementary figure 8. Variance explained by first 1588 PCs

Supplementary figure 9. Visualization of PC1 to PC 10 of PCA over gene correlation matrix

Supplementary figure 10. Outlier genes in PC 8 and PC 9 of PCA over gene correlation matrix

Supplementary figure 11. PC sample scores to distinguish different tissues

Supplementary figure 12. Outlier samples in PC sample scores of PC 8 and PC 9

Supplementary table 1. A list of samples annotated in the European Nucleotide Archive June 30, 2016

Supplementary table 2. A list of 83 diagnosed patients with Mendelian disorders and corresponding predictions with GADO

Supplementary table 3. A list of 38 undiagnosed patients with suspected Mendelian disorders

Supplementary table 4. A comparison between GADO and Exomiser predictions using a list of 83 diagnosed patients with Mendelian disorders

Acknowledgments

We are grateful for the participation of the patients and their parents in this study. We thank Kate Mc Intyre for editing the manuscript and Marieke Bijlsma, Gerben van der Vries, Sido Haakma and Pieter Neerincx for support with the computational analyses. This work was carried out on the Groningen Center for Information Technology (Strikwerda, W. Albers, R. Teeninga, H. Gankema and H. Wind) and Target storage (E. Valentyn and R. Williams). Target is supported by Samenwerkingsverband Noord Nederland, the European Fund for Regional Development, the Dutch Ministry of Economic Affairs, Pieken in de Delta and the provinces of Groningen and Drenthe. Wouter P. te Rijdt is supported by Young Talent Program (CVON PREDICT) grant 2017T001 from the Dutch Heart Foundation. Netherlands Heart Institute, Utrecht, the Netherlands. This work is supported by a grant from the European Research Counsil (ERC Starting Grant agreement number 637640 ImmRisk) to Lude Franke and a VIDI grant (917.14.374) from the Netherlands Organisation for Scientific Research (NWO) to Lude Franke. This work was supported by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007).

References

1. Brown TL, Meloche TM. Exome sequencing a review of new strategies for rare genomic disease research. Genomics. Academic Press; 2016. p. 109–14.

2. Wright CF, FitzPatrick DR, Firth H V. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet [Internet]. Nature Publishing Group; 2018 [cited 2018 Jul 11];19:253–68. Available from: http://www.nature.com/doifinder/10.1038/nrg.2017.116

3. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA [Internet]. NIH Public Access; 2014 [cited 2018 Jul 11];312:1870–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25326635

4. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet (London, England) [Internet]. Elsevier; 2015 [cited 2018 Jul 11];385:1305–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25529582

5. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. Online Mendelian Inheritance in Man, OMIM [Internet]. Available from: https://omim.org/

6. Stenson PD, Mort M, Ball E V., Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet [Internet]. Springer Berlin Heidelberg; 2017 [cited 2018 Jul 11];136:665–77. Available from: http://link.springer.com/10.1007/s00439-017-1779-6

7. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature [Internet]. Nature Research;

2016 [cited 2017 Jun 23];536:285–91. Available from: http://www.nature.com/doifinder/10.1038/nature19057

8. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet [Internet]. Nature Publishing Group; 2017 [cited 2018 Jun 17];18:599–612. Available from: http://www.nature.com/doifinder/10.1038/nrg.2017.52

9. van der Velde KJ, de Boer EN, van Diemen CC, Sikkema-Raddatz B, Abbott KM, Knopperts A, et al. GAVIN: Gene-Aware Variant INterpretation for medical sequencing. Genome Biol [Internet]. 2017 [cited 2017 Jun 16];18:6. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1141-7

10. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. BioMed Central; 2016 [cited 2018 Jul 11];17:122. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

11. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Med [Internet]. 2015 [cited 2018 Jul 11];7:81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26229552

12. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. Am J Hum Genet [Internet]. 2008 [cited 2018 Jul 11];83:610–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18950739

13. Birgmeier J, Haeussler M, Deisseroth CA, Jagadeesh KA, Ratner AJ, Guturu H, et al. AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2017 [cited 2018 Jun 17];171322. Available from: https://www.biorxiv.org/content/early/2017/08/02/171322

14. Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. Genet Med [Internet]. 2016 [cited 2018 Jun 27];18:608–17. Available from: http://www.nature.com/articles/gim2015137

15. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aym? S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res [Internet]. 2017 [cited 2017 Jun 16];45:D865–76. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1039

16. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Res [Internet]. Oxford University Press; 2011 [cited 2017 Jun 16];39:D28-31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20972220

17. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol [Internet]. 2016 [cited 2017 Jul 5];34:525–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27043002

18. Deelen P, Zhernakova D V, de Haan M, van der Sijde M, Bonder MJ, Karjalainen J, et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. Genome Med [Internet]. 2015 [cited 2015 Apr 7];7:30. Available from: http://genomemedicine.com/content/7/1/30

19. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra H, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet [Internet]. 2015;47:115–25. Available from: http://www.nature.com/doifinder/10.1038/ng.3173

20. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. Zhang S-D, editor. PLoS One

[Internet]. Public Library of Science; 2014 [cited 2018 Jun 27];9:e78644. Available from: http://dx.plos.org/10.1371/journal.pone.0078644

21. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46:D649–55.

22. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res [Internet]. 2017 [cited 2018 Jul 12];45:D353–61. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1092

23. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res [Internet]. 2017 [cited 2018 Jul 12];45:D331–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27899567

24. Zaykin D V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol [Internet]. NIH Public Access; 2011 [cited 2018 May 29];24:1836–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21605215

25. Kurian MA, Zhen J, Cheng S-Y, Li Y, Mordekar SR, Jardine P, et al. Homozygous loss-of-function mutations in the gene encoding the dopamine transporter are associated with infantile parkinsonism-dystonia. J Clin Invest [Internet]. 2009 [cited 2018 Jun 28];119:1595–603. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19478460

26. Puffenberger EG, Jinks RN, Sougnez C, Cibulskis K, Willert RA, Achilly NP, et al. Genetic Mapping and Exome Sequencing Identify Variants Associated with Five Novel Diseases. Janecke AR, editor. PLoS One [Internet]. 2012 [cited 2018 Jun 28];7:e28936. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22279524

27. Kurian MA, Li Y, Zhen J, Meyer E, Hai N, Christen H-J, et al. Clinical and molecular characterisation of hereditary dopamine transporter deficiency syndrome: an observational

cohort and experimental study. Lancet Neurol [Internet]. 2011 [cited 2018 Jun 28];10:54–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21112253

28. The Gtex Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet [Internet]. 2013 [cited 2014 Jan 20];45:580–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23715323

29. Benson MD. Inherited amyloidosis. J Med Genet [Internet]. BMJ Publishing Group; 1991 [cited 2018 Jul 10];28:73–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1848299

30. Nouhravesh N, Ahlberg G, Ghouse J, Andreasen C, Svendsen JH, Haunsø S, et al. Analyses of more than 60,000 exomes questions the role of numerous genes previously associated with dilated cardiomyopathy. Mol Genet genomic Med [Internet]. Wiley-Blackwell; 2016 [cited 2018 Jul 10];4:617–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27896284

31. Arimura T, Matsumoto Y, Okazaki O, Hayashi T, Takahashi M, Inagaki N, et al. Structural analysis of obscurin gene in hypertrophic cardiomyopathy. Biochem Biophys Res Commun [Internet]. Academic Press; 2007 [cited 2018 Jun 28];362:281–7. Available from: https://www.sciencedirect.com/science/article/pii/S0006291X07015963?via%3Dihub

32. Marston S, Montgiraud C, Munster AB, Copeland O, Choi O, dos Remedios C, et al. OBSCN Mutations Associated with Dilated Cardiomyopathy and Haploinsufficiency. Thangaraj K, editor. PLoS One [Internet]. Public Library of Science; 2015 [cited 2018 Jun 28];10:e0138568. Available from: http://dx.plos.org/10.1371/journal.pone.0138568

33. Bolling MC, Jan SZ, Pasmooij AMG, Lemmink HH, Franke LH, Yenamandra VK, et al. Generalized Ichthyotic Peeling Skin Syndrome due to FLG2 Mutations. J Invest Dermatol [Internet]. 2018 [cited 2018 Jul 10]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/29505760

34. Alfares A, Al-Khenaizan S, Al Mutairi F. Peeling skin syndrome associated with novel variant in *FLG2* gene. Am J Med Genet Part A [Internet]. Wiley-Blackwell; 2017 [cited 2018 Jul 11];173:3201–4. Available from: http://doi.wiley.com/10.1002/ajmg.a.38468

35. Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, et al. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. Cell Rep [Internet]. Elsevier; 2015 [cited 2018 Jul 24];10:148–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25558065

36. Runge JS, Raab JR, Magnuson T. Identification of Two Distinct Classes of the Human INO80 Complex Genome-Wide. G3 (Bethesda) [Internet]. G3: Genes, Genomes, Genetics; 2018 [cited 2018 Jul 24];8:1095–102. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29432129

37. Meeson AP, Radford N, Shelton JM, Mammen PP, DiMaio JM, Hutcheson K, et al. Adaptive mechanisms that preserve cardiac function in mice without myoglobin. Circ Res [Internet]. 2001 [cited 2018 Jul 20];88:713–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11304494

38. van der Harst P, van Setten J, Verweij N, Vogler G, Franke L, Maurano MT, et al. 52 Genetic Loci Influencing Myocardial Mass. J Am Coll Cardiol [Internet]. Elsevier; 2016 [cited 2018 Jul 20];68:1435–48. Available from: https://www.sciencedirect.com/science/article/pii/S0735109716346642?via%3Dihub

39. Truszkowska GT, Bilińska ZT, Muchowicz A, Pollak A, Biernacka A, Kozar-Kamińska K, et al. Homozygous truncating mutation in NRAP gene identified by whole exome sequencing in a patient with dilated cardiomyopathy. Sci Rep [Internet]. Nature Publishing Group; 2017 [cited 2018 Jul 20];7:3362. Available from: http://www.nature.com/articles/s41598-017-03189-8

40. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. Hum Mutat [Internet]. 2015 [cited 2018 Jul 12];36:928–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26220891

41. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res [Internet]. 2017 [cited 2018 Jul 12];45:D362–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27924014

42. Collado-Torres L, Nellore A, Jaffe AE. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. F1000Research [Internet]. 2017 [cited 2018 Jul 12];6:1558. Available from: https://f1000research.com/articles/6-1558/v1

43. Posafalvi A, Herkert JC, Sinke RJ, van den Berg MP, Mogensen J, Jongbloed JDH, et al. Clinical utility gene card for: dilated cardiomyopathy (CMD). Eur J Hum Genet [Internet]. Nature Publishing Group; 2013 [cited 2018 Jun 22];21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23249954

44. Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, et al. Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N Engl J Med [Internet]. NIH Public Access; 2017 [cited 2018 Jun 27];376:21–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27959697

45. Genomics England. PanelApp [Internet]. Available from: https://panelapp.genomicsengland.co.uk

46. Silvester N, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Gibson R, et al. Content discovery and retrieval services at the European Nucleotide Archive. Nucleic Acids Res. 2015;43:D23–9.