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

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

66 candidate genes by their likelihood of causing the patient's phenotype based on automated
67 literature analysis [13]. However, this focus on what is known may inadvertently filter out
68 variants in potential novel disease genes. Alternatively, the causative gene defect could be
69 missed if a patient's phenotype differs from the features previously reported to be
70 associated to a disease gene. Tools like Exomiser can identify novel human disease genes,
71 as it prioritizes variants based on semantic phenotypic similarity between a patient's
72 phenotype described by HPO terms and HPO-annotated diseases, Mammalian Phenotype
73 Ontology (MPO)-annotated mouse and Zebrafish Phenotype Ontology (ZPO)-annotated fish
74 models associated with each exomic candidate and/or its neighbors in an interaction
75 network [14]. However, most available algorithms are based on existing knowledge on human
76 disease genes, their orthologues in animal models, or well-described biological pathways
77 (for a detailed review see [11]).

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

87 We then developed a user-friendly tool that can prioritize variants in known *and* unknown
88 genes based on our functional predictions, which we designated GeneNetwork Assisted
89 Diagnostic Optimization (GADO). GADO ranks variants based on gene co-regulation in
90 publicly available expression data of a wide range of tissues and cell types using HPO terms
91 to describe a patient's phenotype. To validate our prioritization method, we tested how well
92 our method predicts disease-causing genes based on features described for each of the
93 genes in the OMIM database. We then used exome sequencing data of patients with a
94 known genetic diagnosis to benchmark GADO. Finally, we applied our methodology to
95 previously inconclusive WES data and identified several genes that contain variants that
96 likely explain the phenotype of the respective patients. Thus, we show that our methodology
97 is successful in identifying variants in novel, potentially relevant genes explaining the
98 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.

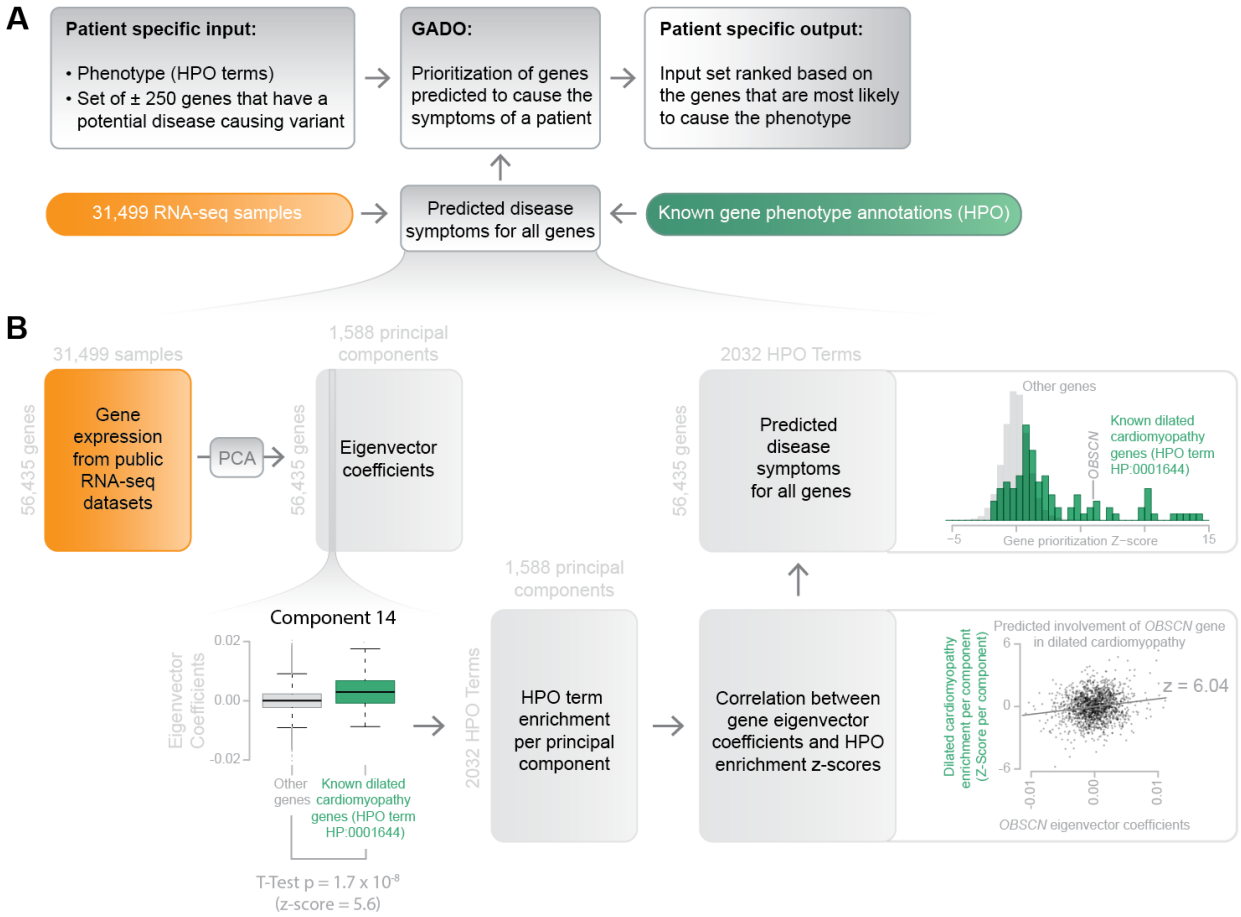


Figure 1: Schematic overview of GADO. *(a) Per patient, GADO requires a set of phenotypic features and a list of candidate genes (i.e. genes harboring rare alleles that are predicted to be pathogenic) as input. It then ascertains whether genes have been predicted to cause these features, and which ones 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

132 samples and quantified gene expression levels for 56,435 genes (of which 22,375 are
133 protein-coding).

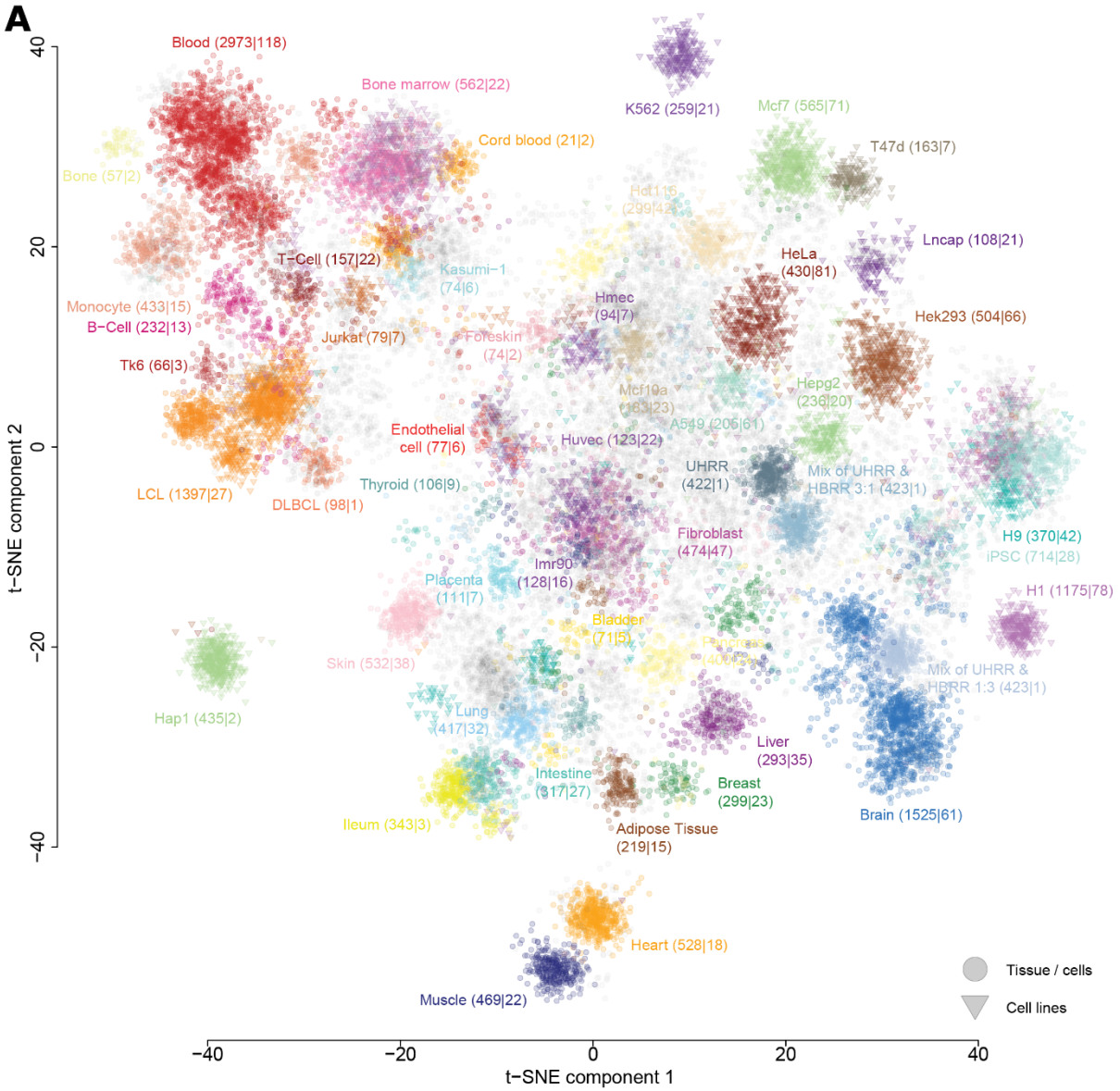
134 Although these samples are generated in many different laboratories, we previously
135 observed that, after having corrected for technical biases, it is possible to integrate these
136 samples into a single expression dataset [18]. We validated that this is also true for our new
137 dataset by visualizing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

143 **Prediction of gene HPO associations and gene functions**

144 To predict HPO term associations and putative gene functions using co-regulation (**Figure**
145 **1**b), we used a method that we had previously developed and applied to public expression
146 microarrays [19]. Since these microarrays only cover a subset of the protein-coding genes (n
147 = 14,510), we decided to use public RNA-seq data instead. This allows for more accurate
148 quantification of lower expressed genes and the expression quantification of many more
149 genes, including a large number of non-protein-coding genes. [20].

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

A



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}).

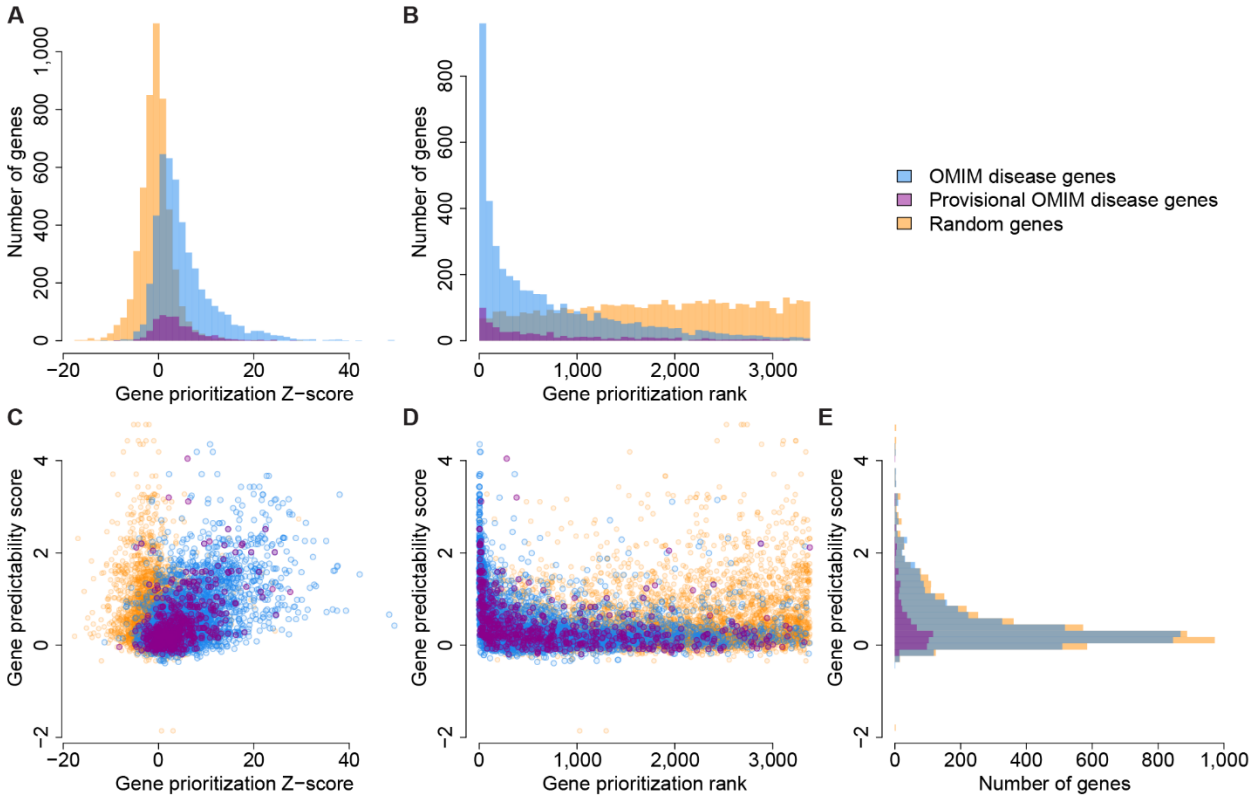


Figure 3: Performance of disease gene prioritization compared to random permutation. *(a) OMIM disease genes and provisional disease genes have significantly stronger z-scores compared to permuted disease genes (T-test p-values: 2.16×10^{-532} & 5.38×10^{-80}, respectively). We also observe that the 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-

211 expression is probably not informative for this gene. To quantify this, we calculated, per
212 gene, the average skewness of the z-score distribution of the Reactome, GO and KEGG gene
213 sets. From this we were able to derive a 'gene predictability score' for every gene that is
214 independent of whether this gene is already known to play a role in any a disease or
215 pathway (**Figure 3**c, d, e). We then ascertained whether these 'gene predictability scores'
216 are correlated with the prediction z-score of the OMIM diseases, and found a strong
217 correlation (Pearson r = 0.54, p-value = 1.14×10^{-332}) between the gene predictability
218 scores and GADO's ability to identify a known disease gene (**Figure 3**c).

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

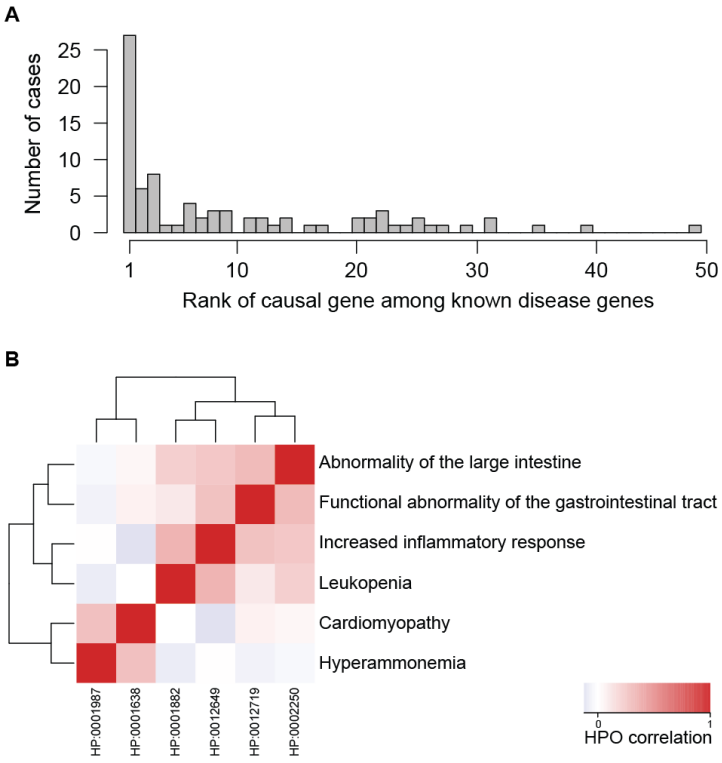
237 **Benchmarking GADO using solved cases with realistic phenotyping**

238 Although *in silico* benchmarking demonstrated the potential of GADO, it used all annotated
239 HPO terms for a disease. In practice, however, patients may only present with a limited
240 number of the annotated features. To perform a validation that was a more realistic
241 reflection of clinical practice, we used exome sequencing data of 83 patients with a known
242 genetic diagnosis. We used their phenotypic features as listed in their medical records prior
243 to the genetic diagnosis (supplementary table 2). On average, per patient, GADO yielded 56
244 possible disease-causing genes with variants that are rare and predicted to be deleterious.

245 In 41% of the patients the actual causative gene was ranked in the top 3 and in 50% of the
246 cases it was in the top 5 (mean rank 10) (**Figure 4**a).

247 **Clustering of HPO terms**

248 In addition to ranking potentially causative genes based on a patient's phenotype, we
249 observed that GADO can be used to cluster HPO terms based on the genes that are predicted
250 to be associated to these HPO terms. This can help identify pairs of symptoms that often occur
251 together, as well as symptoms that rarely co-occur, and we actually observed this for a patient
252 suspected of having two different diseases. This patient is diagnosed with a glycogen storage
253 disease, GSD type Ib, caused by compound heterozygous variants in *SLC37A4* (MIM 602671)
254 and DCM that is probably caused by a truncating variant in *TTN* (MIM 188840). Clustering of
255 the assigned HPO terms placed the phenotypic features related to GSD type Ib ('leukopenia'
256 (HP:0001882) and 'inflammation of the large intestine' (HP:0002037)) together, while
257 Cardiomyopathy (HP:0001638) was only weakly correlated to these specific features (**Figure**
258 **4**b).



259

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

266 **Reanalysis of previously unsolved cases**

267 To assess GADO's ability to discover new disease genes, we applied it to data from 38
268 patients who are suspected to have a Mendelian disease but who have not had a genetic
269 diagnosis. All patients had undergone prior genetic testing (WES with analysis of a gene
270 panel according to their phenotype, supplementary table 3). On average three genes had a
271 z-score ≥ 5 (which we used as an arbitrary cut-off and that correspond to a p-value of 5.7 X
272 10^{-7}) and were further assessed. In seven cases, we identified variants in genes not
273 associated to a disease in OMIM or other databases, but for which we could find literature or
274 for which we gained functional evidence implicating their disease relevance (**Table 1**). For
275 example, we identified two cases with DCM with rare compound heterozygous variants in
276 the *OBSCN* gene (MIM 608616) that are predicted to be damaging. In literature, inherited
277 variant(s) in *OBSCN*, encoding obscurin, are associated with hypertrophic CM [31] and DCM [32].
278 Furthermore, obscurin is a known interaction partner of titin (TTN), a well-known DCM-
279 related protein [31]. Another example came from a patient with ichthyotic peeling skin
280 syndrome, which is caused by a damaging variant in *FLG2 (*MIM 616284). We recently
281 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	8.0×10^{-5} 1.7×10^{-3}	[31, 32]	Yes
HP:0001644	226	3	*OBSCN*	NM_001098623.2: c.[5545C>T]; [22384+3_22384+21del]	14.7 7.8	3.2×10^{-4} 0	[31, 32]	Yes
HP:0008066 HP:0008064	359	3	*FLG2*	NM_001014342.2: c.[632C>G]; [632C>G]	35.0 35.0	1.1×10^{-5} 1.1×10^{-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	3.6×10^{-5}	[37]	Yes
HP:0001644	126*	1	*SYNPO2L***	NM_001114133.2: c.[473G>A]	24.1	5.4×10^{-4}	[38]	Yes
HP:0001638	336	4	*NRAP***	NM_001261463.1: c.[4648C>T]	20.4	8.7×10^{-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

293 annotations of genes can be used to perform function enrichment analysis (**Figure 5**b). We
294 also support automated queries to our database.

A

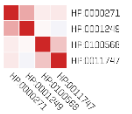
GENE NETWORK

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	OCAF5	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	ATP50	1.9		0.5	0.9	0.7	1.7
15	MAML2	1.9		1	1.1	0.9	0.8
16	OMXL2	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
CLUSTER 1	31
CLUSTER 2	28
CLUSTER 3	22
CLUSTER 4	19
MY SELECTION	0

ANALYSE CLUSTER 1

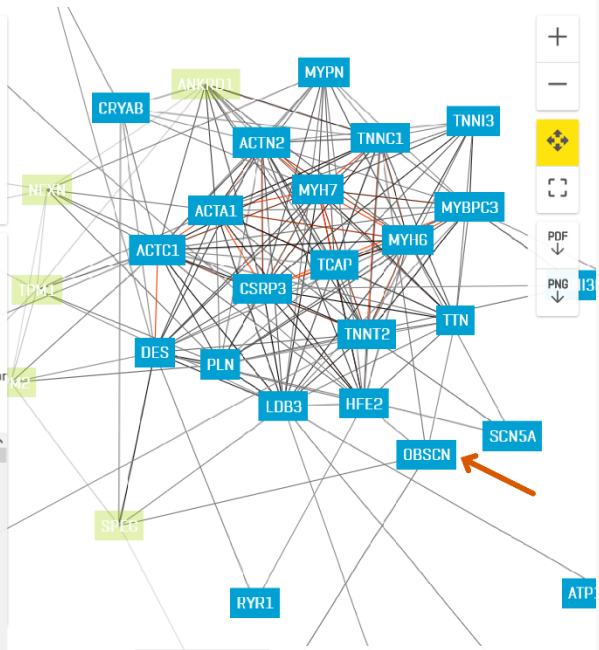
PATHWAYS & PHENOTYPES GENES

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

295

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

308 # Discussion

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

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

329 Benchmarking on sequence data of patients with a known genetic diagnosis revealed that
330 GADO returned the real causative variant within the top 3 results for 41% of the samples,
331 indicating the potential power of GADO for a large number of diseases. Finally, in seven
332 patients, GADO was able to identify potential novel disease genes that are strong candidates
333 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 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

369 case for proteins relying heavily on post-translation modifications for regulation or genes for
370 which different transcripts have distinct functions. This last limitation can potentially be
371 overcome by predicting HPO-isoform associations by using transcript-based expression
372 quantification.

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

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

386 We noted that the median prediction performance of HPO terms is lower compared to the
387 other gene sets databases used in our study, such as Reactome. This may be due to the
388 fact that phenotypes can arise by disrupting multiple distinct biological pathways. For
389 instance, DCMs can be caused by variants in sarcomeric protein genes, but also by variants
390 in calcium/sodium handling genes or by transcription factor genes [43]. As our methodology
391 makes guilt-by-association predictions based on whether genes are showing similar
392 expression levels, the fact that multiple separately working processes are related to the
393 same phenotype can reduce the accuracy of the predictions (although it is often still
394 possible to use these predictions as the DCM HPO phenotype prediction performance AUC =
395 0.76).

396 **Complexity**

397 Given that nearly 5% of patients with a Mendelian disease have another genetic disease [44],
398 it is important to consider that multiple genes might each contribute to specific phenotypic
399 effects. Clinically, it can be difficult to assess if a patient suffers from two inherited
400 conditions, which may hinder variant interpretation based on HPO terms. We showed that
401 GADO can disentangle the phenotypic features of two different diseases manifesting in one
402 patient by correlating and subsequently clustering the profiles of HPO terms describing the

403 patient's phenotype. If the HPO terms observed for a patient do not correlate, it is more
404 likely that they are caused by two different diseases. An early indication that this might be
405 the case for a specific patient can simplify subsequent analysis because the geneticist or
406 laboratory specialist performing the variant interpretation can take this in consideration.
407 GADO also facilitates separate prioritizations on subsets of the phenotypic features.

408 **Conclusion**

409 Connecting variants to disease is a complex multistep process. The early steps are usually
410 highly automated, but the final most critical interpretations still rely on expert review and
411 human interpretation. GADO is a novel approach that can aid users in prioritizing genes
412 using patient-specific HPO terms, thereby speeding-up the diagnostic process. It prioritizes
413 variants in coding *and* non-coding genes, including genes for which there is no current
414 knowledge about their function and those that have not been annotated in any ontology
415 database. This gene prioritization is based on co-regulation of genes identified by analyzing
416 31,499 publicly available RNA-seq samples. Therefore, in contrast to many other existing
417 prioritization tools, GADO has the capacity to identify novel genes involved in human
418 disease. By providing a statistical measure of the significance of the ranked candidate
419 variants, GADO can provide an indication for which genes its predictions are reliable. GADO
420 can also detect phenotypes that do not cluster together, which can alert users to the
421 possible presence of a second genetic disorder and facilitate the diagnostic process in
422 patients with multiple non-specific phenotypic features. GADO can easily be combined with
423 any filtering tool to prioritize variants within WES or WGS data and can also be used in gene
424 panels such as PanelApp [45]. GADO is freely available at www.genenetwork.nl to help guide
425 the differential diagnostic process in medical genetics.

426 Materials and Methods

427 **Sample acquisition**

428 All RNA-seq data used in this project was acquired from the European Nucleotide Archive
429 (ENA) database [46]. Of the 67,090 human RNA-seq samples, with at least 500,000 reads,
430 registered in the ENA on June 30, 2016 (supplementary table 1), 67,019 were successfully
431 downloaded. For 71 of the registered samples, the files were missing. Sample annotations
432 were acquired from [18,47] and through manual curation based on study meta-information in
433 the ENA database (supplementary table 1).

434 **Gene expression quantification**

435 The 67,019 downloaded samples were mapped to transcript annotations from Ensemble
436 release 83 which uses build GRCh38.p5 of the human genome [48] using Kallisto [17] version
437 0.42.4, and the number of reads assessed. The number of reads mapped per sample was
438 obtained from the Kallisto summary file. The following genome files were used:
439 ftp://ftp.ensembl.org/pub/release-
440 83/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz
441 ftp://ftp.ensembl.org/pub/release-
442 83/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz
443 These files were merged and used to build the Kallisto reference index file. The following
444 setting, in addition to all default settings, was used: –k 31.

445 The following Kallisto settings were used mapping all 67,019 samples using default settings
446 for paired-end data mapping. For single-end data mapping we used the following settings in
447 addition to the defaults: –l 200 and –s 20 –bias.

448 After obtaining the transcript counts per sample, these transcript-level counts were summed
449 to gene-level counts for each sample.
450

451 **Gene quality control**

452 We quantified 66,233 genes, which were filtered on the criteria described below, after which
453 56,435 genes remained. Twenty-nine gene names were duplicates/identical. After these
454 were removed, 66,203 genes remained. Of these, 3,628 genes are not expressed (0 reads
455 detected among 31,499 samples) and were removed, leaving 62,575 genes. Next, we
456 detected a number of duplicate genes (100% sequence similarity). Since these genes with
457 perfect sequence similarity have exactly the same number of reads mapping, we were
458 concerned they would appear as perfectly co-expressed genes in our analysis. Most of these
459 genes are either incorrectly mapped genes in the genome build or duplicates of their
460 biological counterpart. Due to their high sequence similarity they are indistinguishable to the
461 mapping tool (potentially introducing false correlations). To avoid potential biases resulting
462 in deceptively high co-expression values, we decided to remove this bias prior to our
463 analysis. 5,471 of these were not located on chromosomes (but on scaffolds), and were
464 removed, leaving 57,104 genes. Another 665 genes had identical transcripts: different IDs,
465 but 100% identical sequences (e.g. ENST00000442165 and ENST00000446969).
466 An additional four genes had no expression in any of the remaining samples after removing
467 outlier/poor-quality samples, as described below, and were also removed prior to the PCA

468 analysis. The 56,435 genes that remained were used for our analyses (supplementary figure
469 5).

470 **RNA-seq sample quality control**

471 We excluded all samples in which less than 70% of the reads successfully mapped to the
472 genome, as reported by Kallisto, resulting in 36,761 samples.

473 *Principal component analysis to identify outlier samples*

474 To identify outlier samples, we conducted a principal component analysis (PCA) along the
475 following steps. First, all estimated counts were log2 transformed. Second, the data was
476 quantile normalized. Third, the covariance over the samples was calculated. Fourth, genes
477 without variance were removed from the dataset. Fifth, a PCA was conducted on the
478 covariance matrix. An arbitrary cut-off on PC 1 was selected at 0.0049 (supplementary
479 figure 6), leaving us with 32,142 samples.

480 *Removal of non-Illumina samples*

481 Since only a small number of samples that passed quality control (147 samples, <0.5% of
482 the total number of samples) were not sequenced on Illumina machines, we removed these
483 to avoid potential biases as a result of these different sequencing tools. This left 31,995
484 samples in our dataset.

485 *Removing duplicate samples*

486 A number of samples had identical values for all genes. Upon inspection, some of these
487 samples appeared to be have been used by multiple studies and uploaded to the ENA
488 database multiple times. To remove duplicate samples, we identified all samples with a
489 correlation >0.9999, randomly selected one of them to include and removed the other.
490 After this step, 31,499 samples remained.

491 *Removal of technical biases*

492 To identify potential technical biases in our data, we calculated the correlation between the
493 PC-scores for each PC and the following potential confounders: read length, paired/single
494 end, total reads in the dataset and percentage mapping reads (supplementary figure 7). We
495 found that all these factors significantly correlated to our sample PC scores for multiple PCs
496 (p-value < 0.01), indicating that these technical factors would affect the co-expression
497 detected in the dataset, if not removed. We decided not to correct for GC content per gene
498 as this may also have biological meaning [49]. For a manual of the covariate removal pipeline
499 we refer to: https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-
500 pipeline. To remove covariates, we used the "adjustcovariates" option.

501 *PCA*

502 After correcting our dataset for technical biases, we conducted the following steps on the
503 matrix. First, we calculated the correlation over the genes. Second, we conducted a PCA
504 over the correlation matrix over the genes. Third, we calculated PC scores for each sample
505 for all PCs.

506 *Inspection of gene PC eigencoefficients*

507 To investigate if any technical biases were present for the different gene types (coding,
508 miRNA, pseudogene, etc.), we plotted the gene eigencoefficients for the first 10 PCs and
509 colored the genes by biotype (supplementary figure 8) and detected an outlier cluster on
510 PC8 and PC9, which were further investigated (supplementary figure 9).

511 *Inspection of sample PC scores*

512 To better understand the origin of the outlier genes in eigenvector coefficients of PC 8 and
513 PC 9, we investigated the PC scores of the samples for these PCs. Additionally, we created a
514 plot for each of the sample PC scores of the first 10 PCs (supplementary figure 10). We
515 observed that there is a clear biological explanation for these outliers, and therefore we
516 decided to retain these signals in the data (supplementary figure 11).

517 **Gene co-regulation analysis**

518 After the quality control steps described above, we conducted a co-regulation analysis using
519 the 31,499 sample by 56,435 gene matrix. The co-regulation analysis was performed using
520 the PC eigencoefficients of the genes for each of the reliable PCs obtained from our gene-co-
521 expression matrix. To determine which PCs are reliable, Cronbach's alpha [50] was calculated
522 for each PC (based on PCA of the gene-correlation matrix). Those PCs with a Cronbach's
523 Alpha ≥ 0.7 were considered reliable, and is a commonly used cutoff [51]. In total, 1,588 PCs
524 have a Cronbach's Alpha ≥ 0.7. Additionally, we calculated the variance explained by each
525 of these PCs and found the first 1588 PCs explain 66 percent of the variance
526 (supplementary figure 12). By including signals from only these PCs, we aimed to remove
527 signals that are not reliable from our analysis. This method was previously shown to
528 perform better than using the correlation matrix directly [19]. The co-regulation scores were
529 calculated by calculating the correlation between the eigencoefficients of each gene pair.
530 Prior to this step the eigencoefficients were standard normalized per gene, after which the
531 eigencoefficients per PC were standard normalized. The logic to this step is to let the signal
532 a gene has for each PC weigh equally when determining the correlation between 2 genes.
533 Here we presumed each PC represents some biological process and those genes that are co-
534 expressed in multiple processes should be reported as strongly co-expressed. This is

535 illustrated and further explained in [19]. The p-values of co-regulated genes can be queried via
536 the website.

537 **Data visualization of sample PC scores using a t-SNE plot**

538 To identify clusters for each cell type and tissue type, we used the sample PC scores, which
539 indicate how strong the signal of each sample is for each PC in the data. Here, each PC is a
540 gene expression signature for the complete set of genes. To visualize how the samples
541 cluster in a two dimensional figure, we constructed a t-SNE plot [52] based on these sample
542 PC-scores using the Rtsne library [53] (version 0.13). The t-SNE was run with a perplexity of
543 50, and we ran 10,000 iterations on our sample PC score matrix. We found that single
544 clusters were visible for many cell- and tissue-types (**Figure 2**a). Most of these clusters
545 contain samples from different studies, which suggests that these clusters are not merely a
546 representation of study-specific biases. The fact that studies with multiple cell/tissue types
547 show multiple clusters further supports the suggestion that the clusters are not driven by
548 non-biological inter-study differences.

549 **Gene function and HPO association predictions**

550 Next, we used the PC eigenvector coefficients calculated in the previous steps to predict
551 functions for genes and to predict which phenotypes they are most likely to play a role in
552 (also described in [19]). For each of the 1,588 reliable PCs, we determined the extent to which
553 each PC captures the activity of a biological module (defined as a group of genes annotated
554 to a term, e.g. a GO function term or HPO phenotype).

555 To do this, the following steps were taken. First, for each PC, a student's T-test was
556 conducted between the eigencoefficients of the genes annotated to a particular term and a
557 group of genes serving as a background. This background consisted of all genes annotated
558 to any term in a specific database, except for those annotated to the term for which the T-
559 test was conducted. Genes that were not annotated to any term in a database were
560 excluded from this background, as these genes have not yet been annotated to any
561 biological functions/terms (because they have not been studied yet). Second, the resulting
562 p-values were transformed into a z-score, which are indicating to which extend each PC
563 represents a biological function/term. This was repeated for each of the 1,588 significant
564 PCs, resulting in a z-score for each PC-term combination. Higher absolute z-values between
565 a term and a PC indicate that the signal for that PC is more strongly related to that term.
566 We applied this methodology to the gene sets described by terms in the following
567 databases: Reactome and KEGG pathways, Gene Ontology (GO) molecular function, GO
568 biological process and GO cellular component terms and finally to HPO terms. We excluded

569 terms for which fewer than 10 genes are annotated because predictions for smaller groups
570 of genes are less accurate and might be misleading. Predictions were made for 8,657 gene
571 sets in total. For each term, we calculated how well each PC captured the signal of the
572 genes that are annotated to that term. Third and last, to predict which genes are correlated
573 to a particular HPO term, we correlated the 1,588 z-scores for that term (as calculated
574 above) with the 1,588 eigenvector coefficients of a gene. These correlations were
575 transformed into z-scores, which we refer to as prioritization scores. This can be done for
576 any gene-to-HPO term combination. However, when a gene is already explicitly annotated
577 to the term and we wish to predict whether that gene is predicted to be involved in that
578 term, there is a small circular bias as the z-scores for this term were partly calculated based
579 on this gene. To remove this bias in these circumstances, the 1,588 z-scores for a gene set
580 were first re-calculated while assuming these gene is not involved in that term, after which
581 the prediction for this gene was made.

582 **Validation of the GO, HPO and Reactome term predictions**

583 To determine the accuracy of our GO, HPO and Reactome term predictions, we calculated
584 how well we could predict genes that are part of a term. To do so, we used the prioritization
585 z-scores that the genes had for a particular term. For each term, we calculated an Area
586 Under the Curve (AUC), using a Mann-Whitney U test, on the prioritization scores of the
587 genes that are part of the term versus those that are not part of the term. These AUCs
588 indicate how accurate the predictions were, with an AUC of 1 indicating perfect predictions
589 and an AUC of 0.5 indicating no predictive power. The average AUC for each category was
590 calculated based on all terms with at least 10 genes annotated and for which the p-value
591 was less than 0.05 (Bonferroni corrected for the number of pathways for the category
592 tested) (**Figure 2**b).

593 **GADO predictions**

594 To identify potential causative variants in patients, we used HPO term annotations
595 describing the patient's features. The gene prediction z-scores for an HPO term were used
596 to rank the genes. If a patient's phenotype was described by more than one HPO term, a
597 meta-analysis was conducted. In this case a weighted z-score was calculated by adding the
598 HPO z-scores for all the patient's HPO terms and then dividing by the square root of the
599 number of HPO terms. In this calculation, we used only those HPO terms, which have
600 significant predictive power (based on whether genes annotated to this term have
601 significantly absolute higher z-scores than those not annotated to the term as calculated in
602 the section "Gene function and HPO association predictions"). If the predictions for a
603 patient's HPO term were not significant, the parent/umbrella HPO term(s) was used. (The

604 online GADO tool supplies the user with a list of parent terms from which the user can then
605 manually select which terms should be used in the analysis (supplementary figure 1)). If
606 this parent term also did not have significant predictive power, the parent's parent term was
607 used (thus moving up the HPO tree until a parent term is found which has significant
608 predictive power). If an HPO term has multiple parents, predictions were made using each
609 parent and the results are reported separately. The genes with the highest z-scores are
610 most relevant for the patient according to GADO's predictions. This analysis can be
611 conducted at: https://www.genenetwork.nl/gado.

612 **Validation of disease-gene predictions**

613 To benchmark our method we used the OMIM morbid map [5] downloaded on March 26,
614 2018, containing all disease-gene-phenotype entries. From this list, we extracted the
615 disease-gene associations, excluding non-disease and susceptibility entries. We extracted
616 the provisional disease-gene associations separately. For each disease in OMIM, we used
617 GADO to determine the rank of the causative gene among all genes in the OMIM morbid
618 map. For this we used all phenotypes annotated to the OMIM disease. If any of the HPO
619 terms did not have significant predictive power, the parent term(s) was used.

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

625 **Cohort of previously solved cases**

626 To test if GADO could help prioritize genes that contain the causative variant, we used 83
627 samples of patients who were previously genetically diagnosed through whole exome
628 analysis or gene panel analysis. These samples encompass a wide variety of different
629 Mendelian disorders (supplementary table 2). To assess which genes harbor potentially
630 causative variants, we first called and annotated the variants from the exome sequencing
631 files.

632 *Variant calling*

633 We used the available WES or WGS data from patients with and without genetic diagnosis.
634 These samples were genotyped using a relatively standard BWA and GATK pipeline. For a
635 detailed description of the genotype pipeline see: https://molgenis.gitbooks.io/ngs_dna/
636 (version 3.4.0). For the WGS samples, we confined our analysis to the exome.

637 *Variant annotation*

638 We used GAVIN to annotate our variants to obtain a list of candidate variants. GAVIN
639 prioritizes genes based on, among other factors, minor allele frequency and gene-
640 recalibrated CADD scores (for details see [9]). For 11 of the previously solved cases, GAVIN
641 did not flag the causative variant as a candidate. To be able to include these samples in our
642 GADO benchmark, we added the causative genes for these cases manually to the candidate
643 list.

644 *GADO ranking*

645 The phenotypic features of a patient were translated into HPO terms, which were used as
646 input to GADO for ranking all genes based on how likely they are to cause that set of
647 features. If any of the HPO terms did not have significant predictive power, the parent
648 term(s) was used. From the resulting list of ranked genes, the known disease genes
649 harboring a potentially causative variant were selected. Next, we determined the rank of the
650 gene with the known causative variant among the selected genes. If a patient harbored
651 multiple causative variants in different genes, in case of di-genic inheritance or two
652 inherited conditions, the median rank of these genes was reported (supplementary table 2).

653 **Benchmark comparison with Exomiser**

654 To evaluate GADO's performance, we compared GADO with Exomiser [54] (version 10.1.0,
655 with exomiser-phenotype-1802 and exomiser-genome-hg19-1805 files from
656 https://data.monarchinitiative.org/exomiser/data/). Both GADO and Exomiser were given
657 each patient candidate gene list along with their respective set of phenotypes as input.
658 Default settings were used. We used the gene rankings based on
659 "EXOMISER_GENE_COMBINED_SCORE" and identified the rank of the causative gene
660 (supplementary table 4). In case of a tie, the average rank of the ties was reported. If a
661 patient harbored multiple causative variants, the median rank of the genes harboring the
662 causative variants was reported. To ensure a fair comparison, we used GADO on the set of
663 genes reported by Exomiser (supplementary table 4).

664 **Unsolved cases cohorts**

665 In addition to the patients with a known genetic diagnosis, we tested 38 unsolved cases
666 (supplementary table 3). These are patients with mainly cardiomyopathies or developmental
667 delay. All patients were previously investigated using exome sequencing, by analyzing a
668 gene panel appropriate for their phenotype. To allow discovery of potential novel disease
669 genes, we used GADO to score all genes with candidate variants. For genes with a

670 prediction z-score ≥ 5, a literature search for supporting evidence was performed to assess
671 whether these genes are likely candidate genes.

672 **GADO web-tool**

673 To make the gene-co-regulation-based HPO predictions publicly available a website was
674 constructed: www.genenetwork.nl

675 On this website the user can conduct the following analyses:
676 *1. Predict putative functions of genes*. This can be achieved by querying a gene, for which
677 gene-network will then predict the function based on the functional enrichment of its co-
678 regulation partners. Enrichment for GO, Reactome, KEGG and HPO phenotypes can be
679 retrieved.

680 *2. Prioritize potential causative disease genes for patients*: Based on HPO terms or a group
681 of genes annotated to a patient, the GADO tool will rank all genes based on how likely they
682 are to be related to the patient's phenotype. These can be further filtered for genes of
683 interest, by providing a list of genes known to harbor likely causative variants.

684 **Gene network visualization**

685 Edges are drawn between two genes/nodes based on a z-score cutoff. The cutoff at which a
686 line/edge between two genes should be drawn can be manually altered with the bar in the
687 top right corner. The network is drawn based on a force directed layout and clusters are
688 assigned using affinity propagation [55]

689 **HPO, Reactome, KEGG and GO enrichment calculations**

690 On the network page it is possible to retrieve which HPO, Reactome, KEGG and GO
691 categories are enriched among the visualized genes. It is also possible to retrieve this for a
692 sub-selection of these genes. The enrichment is calculated based on the z-scores of each of
693 these genes for each category. For each category/term, a Mann-Whitney U test is conducted
694 between the z-scores of the genes in the network versus the z-scores of genes that are not
695 part of the visualized network. The pathways with the most significant p-values are then
696 ranked highest.

697 It is also possible to identify which other genes are strongly co-regulated with those
698 visualized in the network. This is done similarly to how the correlation between a gene and
699 a pathway is calculated, as described above in "Gene function and HPO association
700 predictions". First, the z-scores for each PC of the genes visualized in the network is
701 calculated. After the z-scores of this group of genes have been calculated for each4

702 pathway, the correlation of the PC coefficients for each gene not in the network with these
703 z-scores is calculated. The genes with the most significant correlation are ranked highest.

704 **Gene predictability scores**

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

710 Description of Supplemental Data

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

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

715 Supplementary figure 3. Comparison between GADO and Exomiser rankings

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

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

718 Supplementary figure 6. PCA plot of 36,761 samples

719 Supplementary figure 7. Investigation of principal components capturing technical biases

720 Supplementary figure 8. Visualization of PC1 to PC 10 of PCA over gene correlation matrix

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

722 Supplementary figure 10. PC sample scores to distinguish different tissues

723 Supplementary figure 11. Outlier samples in PC sample scores of PC 8 and PC 9

724 Supplementary figure 12. Variance explained by first 1588 PCs

725 Supplementary table 1. A list of samples annotated in the European Nucleotide Archive June
726 30, 2016

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

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

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

733 Declaration of Interests

734 The authors declare no competing interests.

735 Acknowledgments

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

750 Web Resources

751 Gene Network, www.genenetwork.nl
752 GADO, https://www.genenetwork.nl/gado
753 European Nucleotide Archive, https://www.ebi.ac.uk/ena
754 Ensembl, https://www.ensembl.org
755 OMIM, https://www.omim.org
756 Genotyping pipeline, https://molgenis.gitbooks.io/ngs_dna/
757 Covariate removal pipeline, https://github.com/molgenis/systemsgenetics/tree/master/eqtl-
758 mapping-pipeline

759 # References

760 1. Brown, T.L., and Meloche, T.M. (2016). Exome sequencing a review of new strategies for
761 rare genomic disease research. Genomics *108*, 109–114.

762 2. Wright, C.F., FitzPatrick, D.R., and Firth, H. V. (2018). Paediatric genomics: diagnosing
763 rare disease in children. Nat. Rev. Genet. *19*, 253–268.

764 3. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang,
765 M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-
766 exome sequencing. JAMA *312*, 1870–1879.

767 4. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg,
768 M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis
769 of developmental disorders in the DDD study: a scalable analysis of genome-wide research
770 data. Lancet (London, England) *385*, 1305–1314.

771 5. McKusick-Nathans Institute of Genetic Medicine, and Johns Hopkins University Online
772 Mendelian Inheritance in Man, OMIM , https://omim.org/.

773 6. Stenson, P.D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M.,
774 Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a
775 comprehensive repository of inherited mutation data for medical research, genetic diagnosis
776 and next-generation sequencing studies. Hum. Genet. *136*, 665–677.

777 7. Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-
778 Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding
779 genetic variation in 60,706 humans. Nature *536*, 285–291.

780 8. Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization
781 and Mendelian disease. Nat. Rev. Genet. *18*, 599–612.

782 9. van der Velde, K.J., de Boer, E.N., van Diemen, C.C., Sikkema-Raddatz, B., Abbott, K.M.,
783 Knopperts, A., Franke, L., Sijmons, R.H., de Koning, T.J., Wijmenga, C., et al. (2017).
784 GAVIN: Gene-Aware Variant INterpretation for medical sequencing. Genome Biol. *18*, 6.

785 10. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P.,
786 and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

787 11. Smedley, D., and Robinson, P.N. (2015). Phenotype-driven strategies for exome
788 prioritization of human Mendelian disease genes. Genome Med. *7*, 81.

789 12. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008).
790 The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary
791 Disease. Am. J. Hum. Genet. *83*, 610–615.

792 13. Birgmeier, J., Haeussler, M., Deisseroth, C.A., Jagadeesh, K.A., Ratner, A.J., Guturu, H.,
793 Wenger, A.M., Stenson, P.D., Cooper, D.N., Re, C., et al. (2017). AMELIE accelerates
794 Mendelian patient diagnosis directly from the primary literature. BioRxiv 171322.

795 14. Bone, W.P., Washington, N.L., Buske, O.J., Adams, D.R., Davis, J., Draper, D., Flynn,
796 E.D., Girdea, M., Godfrey, R., Golas, G., et al. (2016). Computational evaluation of exome
797 sequence data using human and model organism phenotypes improves diagnostic efficiency.
798 Genet. Med. *18*, 608–617.

799 15. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aym?, S., Baynam,
800 G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology
801 in 2017. Nucleic Acids Res. *45*, D865–D876.

802 16. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland,
803 I., Faruque, N., Goodgame, N., Gibson, R., et al. (2011). The European Nucleotide Archive.
804 Nucleic Acids Res. *39*, D28-31.

805 17. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic
806 RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

807 18. Deelen, P., Zhernakova, D. V, de Haan, M., van der Sijde, M., Bonder, M.J., Karjalainen,
808 J., van der Velde, K.J., Abbott, K.M., Fu, J., Wijmenga, C., et al. (2015). Calling genotypes
809 from public RNA-sequencing data enables identification of genetic variants that affect gene-
810 expression levels. Genome Med. *7*, 30.

811 19. Fehrmann, R.S.N., Karjalainen, J.M., Krajewska, M., Westra, H., Maloney, D., Simeonov,
812 A., Pers, T.H., Hirschhorn, J.N., Jansen, R.C., Schultes, E.A., et al. (2015). Gene expression
813 analysis identifies global gene dosage sensitivity in cancer. Nat. Genet. *47*, 115–125.

814 20. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of
815 RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS One *9*,
816 e78644.

817 21. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw,
818 R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase.
819 Nucleic Acids Res. *46*, D649–D655.

820 22. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG:
821 new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–
822 D361.

823 23. The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase
824 and resources. Nucleic Acids Res. *45*, D331–D338.

825 24. Zaykin, D. V (2011). Optimally weighted Z-test is a powerful method for combining
826 probabilities in meta-analysis. J. Evol. Biol. *24*, 1836–1841.

827 25. Kurian, M.A., Zhen, J., Cheng, S.-Y., Li, Y., Mordekar, S.R., Jardine, P., Morgan, N. V.,
828 Meyer, E., Tee, L., Pasha, S., et al. (2009). Homozygous loss-of-function mutations in the
829 gene encoding the dopamine transporter are associated with infantile parkinsonism-
830 dystonia. J. Clin. Invest. *119*, 1595–1603.

831 26. Puffenberger, E.G., Jinks, R.N., Sougnez, C., Cibulskis, K., Willert, R.A., Achilly, N.P.,
832 Cassidy, R.P., Fiorentini, C.J., Heiken, K.F., Lawrence, J.J., et al. (2012). Genetic Mapping
833 and Exome Sequencing Identify Variants Associated with Five Novel Diseases. PLoS One *7*,
834 e28936.

835 27. Kurian, M.A., Li, Y., Zhen, J., Meyer, E., Hai, N., Christen, H.-J., Hoffmann, G.F.,
836 Jardine, P., von Moers, A., Mordekar, S.R., et al. (2011). Clinical and molecular
837 characterisation of hereditary dopamine transporter deficiency syndrome: an observational
838 cohort and experimental study. Lancet Neurol. *10*, 54–62.

839 28. The Gtex Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat.
840 Genet. *45*, 580–585.

841 29. Benson, M.D. (1991). Inherited amyloidosis. J. Med. Genet. *28*, 73–78.

842 30. Nouhravesh, N., Ahlberg, G., Ghouse, J., Andreasen, C., Svendsen, J.H., Haunsø, S.,
843 Bundgaard, H., Weeke, P.E., and Olesen, M.S. (2016). Analyses of more than 60,000
844 exomes questions the role of numerous genes previously associated with dilated
845 cardiomyopathy. Mol. Genet. Genomic Med. *4*, 617–623.

846 31. Arimura, T., Matsumoto, Y., Okazaki, O., Hayashi, T., Takahashi, M., Inagaki, N.,
847 Hinohara, K., Ashizawa, N., Yano, K., and Kimura, A. (2007). Structural analysis of obscurin
848 gene in hypertrophic cardiomyopathy. Biochem. Biophys. Res. Commun. *362*, 281–287.

849 32. Marston, S., Montgiraud, C., Munster, A.B., Copeland, O., Choi, O., dos Remedios, C.,
850 Messer, A.E., Ehler, E., and Knöll, R. (2015). OBSCN Mutations Associated with Dilated

851 Cardiomyopathy and Haploinsufficiency. PLoS One *10*, e0138568.

852 33. Bolling, M.C., Jan, S.Z., Pasmooij, A.M.G., Lemmink, H.H., Franke, L.H., Yenamandra,
853 V.K., Sinke, R.J., van den Akker, P.C., and Jonkman, M.F. (2018). Generalized Ichthyotic
854 Peeling Skin Syndrome due to FLG2 Mutations. J. Invest. Dermatol.

855 34. Alfares, A., Al-Khenaizan, S., and Al Mutairi, F. (2017). Peeling skin syndrome
856 associated with novel variant in *FLG2* gene. Am. J. Med. Genet. Part A *173*, 3201–3204.

857 35. Alazami, A.M., Patel, N., Shamseldin, H.E., Anazi, S., Al-Dosari, M.S., Alzahrani, F.,
858 Hijazi, H., Alshammari, M., Aldahmesh, M.A., Salih, M.A., et al. (2015). Accelerating novel
859 candidate gene discovery in neurogenetic disorders via whole-exome sequencing of
860 prescreened multiplex consanguineous families. Cell Rep. *10*, 148–161.

861 36. Runge, J.S., Raab, J.R., and Magnuson, T. (2018). Identification of Two Distinct Classes
862 of the Human INO80 Complex Genome-Wide. G3 (Bethesda). *8*, 1095–1102.

863 37. Meeson, A.P., Radford, N., Shelton, J.M., Mammen, P.P., DiMaio, J.M., Hutcheson, K.,
864 Kong, Y., Elterman, J., Williams, R.S., and Garry, D.J. (2001). Adaptive mechanisms that
865 preserve cardiac function in mice without myoglobin. Circ. Res. *88*, 713–720.

866 38. van der Harst, P., van Setten, J., Verweij, N., Vogler, G., Franke, L., Maurano, M.T.,
867 Wang, X., Mateo Leach, I., Eijgelsheim, M., Sotoodehnia, N., et al. (2016). 52 Genetic Loci
868 Influencing Myocardial Mass. J. Am. Coll. Cardiol. *68*, 1435–1448.

869 39. Truszkowska, G.T., Bilińska, Z.T., Muchowicz, A., Pollak, A., Biernacka, A., Kozar-
870 Kamińska, K., Stawiński, P., Gasperowicz, P., Kosińska, J., Zieliński, T., et al. (2017).
871 Homozygous truncating mutation in NRAP gene identified by whole exome sequencing in a
872 patient with dilated cardiomyopathy. Sci. Rep. *7*, 3362.

873 40. Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: A
874 Matching Tool for Connecting Investigators with an Interest in the Same Gene. Hum. Mutat.
875 *36*, 928–930.

876 41. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A.,
877 Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-
878 controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res.
879 *45*, D362–D368.

880 42. Collado-Torres, L., Nellore, A., and Jaffe, A.E. (2017). recount workflow: Accessing over
881 70,000 human RNA-seq samples with Bioconductor. F1000Research *6*, 1558.

882 43. Posafalvi, A., Herkert, J.C., Sinke, R.J., van den Berg, M.P., Mogensen, J., Jongbloed,
883 J.D.H., and van Tintelen, J.P. (2013). Clinical utility gene card for: dilated cardiomyopathy
884 (CMD). Eur. J. Hum. Genet. *21*,.

885 44. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H.,
886 Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes
887 Resulting from Multilocus Genomic Variation. N. Engl. J. Med. *376*, 21–31.

888 45. Genomics England PanelApp , https://panelapp.genomicsengland.co.uk.

889 46. Silvester, N., Alako, B., Amid, C., Cerdeño-Tárraga, A., Cleland, I., Gibson, R.,
890 Goodgame, N., ten Hoopen, P., Kay, S., Leinonen, R., et al. (2015). Content discovery and
891 retrieval services at the European Nucleotide Archive. Nucleic Acids Res. *43*, D23–D29.

892 47. Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein,
893 M.C., and Ma 'ayan, A. Massive mining of publicly available RNA-seq data from human and
894 mouse.

895 48. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva,
896 D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res.
897 *43*, D662–D669.

898 49. Vinogradov, A.E. (2003). DNA helix: the importance of being GC-rich. Nucleic Acids Res.
899 *31*, 1838–1844.

900 50. Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests.
901 Psychometrika *16*, 297–334.

902 51. Bresciani, M.J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., and
903 Hickmott, J. (2009). Examining Design and Inter-Rater Reliability of a Rubric Measuring
904 Research Quality across Multiple Disciplines - Practical Assessment, Research &
905 Evaluation. *14*,.

906 52. Van Der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach.
907 Learn. Res. *9*, 2579–2605.

908 53. Krijthe, J.H. (2015). T-Distributed Stochastic Neighbor Embedding using Barnes-Hut , .

909 54. Smedley, D., Jacobsen, J.O.B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M.,
910 Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation
911 diagnostics and disease-gene discovery with the Exomiser. Nat. Protoc. *10*, 2004–2015.

55. Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972–976.