Genetics inMedicine | ORIGINAL RESEARCH ARTICLE

Phenotype-driven gene target definition in clinical genome-wide sequencing data interpretation

Maver Ales, MD[1], Lovrecic Luca, MD, PhD[1], Volk Marija, MD[1], Rudolf Gorazd, MD, MSc[1], Writzl Karin, MD, PhD[1], Blatnik Ana, MD[1], Hodzic Alenka, PhD[1] and Borut Peterlin, MD, PhD[1]

Purpose: Genome-wide sequencing approaches are increasingly being used in place of disease gene panel sequencing approaches. Despite the well-recognized benefits of these approaches, they also carry with them an increased burden of analyzing overwhelmingly large gene targets and an increased possibility of detecting incidental findings.

Methods: We propose a novel approach for design of individualized phenotype gene panels using the set of signs and symptoms observed and selecting relevant genes on the basis of known phenotype–gene associations.

Results: We used results of diagnostic exome sequencing in 405 cases submitted to our institution to show retrospectively that using the phenotype gene panel increases the sensitivity of masked exome analysis (increase from 25.4 to 29.7% in overall diagnostic yield). We also show that such a strategy enables the possibility of masked analysis of genome-wide sequencing data in patients with poorly defined and multifaceted clinical presentations. Ultimately, we show that this approach enables control over the incidental findings rate (0.25% in phenotype gene panels). Finally, we provide a Web tool for customized phenotype panel creation (available at http://www.kimg.eu/generator).

Conclusion: In conclusion, we present a novel approach to a phenotype-driven diagnostic process of genome scale sequencing data that harnesses the sensitivity of these approaches while restricting the analysis to genes relevant to clinical presentation in patient.

Genet Med advance online publication 31 March 2016

Key Words: clinical genetics; exome sequencing; genome sequencing; phenotype ontology

INTRODUCTION

Recent technological advances in human genome sequencing have enabled the collection of a wealth of data at progressively diminishing costs and have significantly facilitated the identification of causative genetic variations in human disorders.[1] Two approaches are mainly used in clinical practice for identification of causative sequence variation: undirected whole-exome or mendeliome resequencing and directed approaches of targeted gene panel resequencing.[2]

Disease gene panel next-generation sequencing (NGS) is currently the most widely used approach for genetic testing of genetically heterogeneous disorders.[2] This approach has some benefits: it is a highly focused analysis of well-defined pertinent genes, it minimizes the possibility of identifying incidental findings, and it offers relatively low-cost diagnostics after the initial implementation stage.[2]However, it is restricted to surveying the genes that have been directly related to a genetic disorder, and the targeted list of genes varies notably between sequencing providers.[3] Additionally, such an approach may fail to address the heterogeneity in clinical presentations of disorders and may miss the cases in which only a partial clinical presentation of the underlying genetic variant is present in the patient, and thus the original diagnosis may not correctly reflect the underlying genetic condition. Consequently, it is not uncommon for the clinical target in panel sequencing approaches to be selected too restrictively, and this may result in a failure to detect the actual causative variation in patients.[4]

Selection of appropriate target genes for diagnostic sequencing is challenging for several reasons. First, human diseases, especially hereditary genetic disorders, usually affect multiple organ systems and present with a wide variety of clinical symptoms.[5] For this reason, it is often difficult to unequivocally attribute the clinical presentation of a patient to a specific disease or disease group based on the disease gene panel selected. Second, clinical presentation in any given patient commonly varies and often does not fully conform to established clinical diagnostic criteria, potentially misleading the selection of a gene target. Furthermore, clinical presentation may overlap significantly among disease groups, and the choice of gene target may not be straightforward. Finally, the selection of the disease gene panels is still mostly arbitrary, resulting in a large variation between assortments of genes offered by various sequencing providers aiming for diagnosis of the same clinical condition.

As the cost of clinical exome sequencing has fallen closer to that of focused gene panel sequencing, a single-exome sequencing test presents a viable alternative for the establishment of separate panels for numerous disorder groups. For this reason, several diagnostic institutions are already redirecting

[1]Clinical Institute of Medical Genetics, University Medical Centre Ljubljana, Ljubljana, Slovenia. Correspondence: Borut Peterlin (borut.peterlin@guest.arnes.si)

Submitted 7 September 2015; accepted 25 January 2016; advance online publication 31 March 2016. doi:10.1038/gim.2016.22

diagnostics from targeted gene panel sequencing to whole-exome sequencing.[5,6] This approach has the potential to outperform focused gene panel sequencing in diagnostic performance, but it also carries with it an increased burden of analyzing overwhelmingly large gene targets and an increased risk of identifying a number of findings not pertinent to patients' clinical presentations.[7,8] Current European Society of Human Genetics recommendations for whole-genome sequencing in clinical settings warrant initially focusing the genetic analysis on gene targets with known relationships to phenotypes observed and only secondarily expand the search, aiming to minimize the rate of incidental findings.[9] In adherence to these recommendations, a common approach to address this issue is to utilize a two-tiered strategy: initially masking the exome-level data with a panel of clinically relevant genes and if the patient agrees to be informed about incidental findings, proceeding to whole-exome analysis. A principal limitation of this approach is in narrowing first-tier analysis to arbitrary gene panels developed at each institution, which shares the limitation of the focused panel testing approach.

We hypothesized that generating a gene target based on the observed phenotypes collected in a standardized way may be a viable solution for both sensitive and controlled expansion of gene targets in cases submitted for sequencing. We have identified human phenotype ontology (HPO) as the most comprehensive framework as the basis for such an approach because it contains a wide representation of terms associated with phenotypes observed in human diseases in a well-defined and formalized form.[10] Because the HPO annotations are also associated with causative genes, we found it to be a plausible basis for phenotype-driven gene target generation. This approach obviates the need to limit the search of genes in terms of specific diagnosis hypothesis and presents a natural way of using the phenotype–gene relationships in designing a gene panel specific to each patient.

Based on these considerations, we developed a bioinformatic approach for phenotype-driven gene target generation. To establish the utility of this approach, we initially assessed whether phenotype-based annotations provide a framework that sufficiently reflects clinically observed relationships between genes and how these compare to the selection of genes within disease gene panels. Subsequently, we also demonstrated the benefits of phenotype-based gene target generation in a set of hypothetical clinical situations. Furthermore, we performed a validation study using a set of 405 cases submitted for sequencing at our institution to survey whether phenotype-driven gene target generation allows for efficient inclusion of pertinent gene targets and compared it to the disease gene panel–based diagnostic approach.

MATERIALS AND METHODS

We developed an approach for phenotype-based gene target generation in genome-wide sequencing approaches. To substantiate the utility of the HPO resource as the basis of the phenotype-driven gene target generation, we aimed to show that phenotype–gene annotations within the HPO database were sufficiently informative to reflect the landscape of associations between diseases and serve as a viable replacement for manually curated disease gene panels. Furthermore, we demonstrated the added value of phenotype-driven gene target generation in two clinically relevant cases: Parkinson disease and Usher syndrome. To formally quantify the added benefits of phenotype-based gene target generation, we performed a retrospective analysis of cases submitted for exome sequencing at our institution and also compared the causative gene inclusion rate using disease panels versus phenotype-based generated gene targets.

The data in this study were obtained from results gathered from results of routine clinical diagnostics and were not generated specifically for the purpose of this study.

HPO as the source of phenotype–gene associations

We used the latest stable build of the HPO resource (February 2015) as the source of information on phenotype–genotype associations.[10] Because HPO includes associations between diseases and phenotype elements in addition to associations between diseases and causative genes, it is possible to identify the implied associations between genes and phenotype traits, which has been done systematically in the HPO project.

The relationships of HPO ontology were parsed using the ontoCAT package for R.[11] The associations between genes and phenotypes were collected from standardized output results generated by the HPO pipeline. Based on these sources of information, we could associate unique phenotype identifiers characterizing cases of gene annotations.

Evaluation of gene–phenotype associations in HPO as an alternative to disease gene panels

In the present study, we aimed to demonstrate the possibility of generating gene panels based on phenotype–gene associations of genes with the clinical signs and symptoms observed in patients referred for diagnostic sequencing. To show that phenotype gene panels represent an alternative to disease gene panels, we performed two systematic analyses of the landscape of phenotype–gene associations and their relationship to disease gene panels.

First, we aimed to determine whether the phenotype–gene associations in the HPO resource can reflect the composition of genes in the disease gene panels and whether it may serve as a viable alternative to disease panels. We investigated whether genes co-occurring in the disease gene panels share an excess of overlapping phenotype terms than would be expected by random chance. To minimize the spurious nonspecific annotations and precipitate specific phenotype–gene associations, we established a measure of phenotype similarity score for each pair of genes (P_{G_i,G_j}). This score estimates the pairwise phenotype similarity for pairs of genes G_i and G_j and associated sets of HPO phenotype terms Ph_{G_i} and Ph_{G_j} by comparing the intersection of associated HPO terms with the union of associated HPO terms (equation 1). A higher phenotype similarity score

of a gene pair corresponded to greater phenotype compatibility of the genes in the pair.

$$P_{G_i,G_j} = \frac{Ph_{G_i} \cap Ph_{G_j}}{Ph_{G_i} \cup Ph_{G_j}} \quad (1)$$

We utilized this phenotype score to verify whether genes co-occurring in disease panels display increased compatibility in the profiles of associated HPO terms. We used Mann-Whitney *U* statistics to ascertain whether the distribution of pairwise phenotype scores of genes in distinct disease gene panels is increased in comparison to randomly paired genes.

Second, we utilized network analysis of pairwise gene phenotype similarity associations to determine whether clusters of genes with similar phenotype associations display overlap with the composition of disease gene panels. We computed pairwise comparisons of all pairs of genes with phenotype profiles captured within the HPO resource using Cytoscape 2.8 software[12] for analysis and visualization (http://www.cytoscape.org) and using force-directed layout to cluster the genes based on phenotype similarities.

Assessment of phenotype-driven panel generation and its efficiency in the clinical setting

To determine the performance properties of HPO-driven gene target generation, we performed a retrospective analysis of the cases submitted for diagnostic evaluation at our institution. The cases included in the study were referred for whole human exome or mendeliome sequencing in a variety of human genetic disorders from diverse groups of human genetic disorders (**Supplementary Table S1** online). We denote mendeliome as a representative set of genes containing clinically relevant variants associated with human disease; specifically, we used the Illumina TruSight One panel, which targets coding regions of 4,813 genes in the human genome (http://www.illumina.com/products/trusight-one-sequencing-panel.html). Considering that the mendeliome capture used in this study encompasses a large majority of the genes currently annotated by HPO phenotypes (84.4%), we have treated mendeliome sequencing and whole-exome sequencing equally in the subsequent sections and did not make a distinction between results of undirected sequencing undertaken by either capture approach.

Altogether, we included 405 cases with HPO annotations in the survey pool; in 159 of these patients, a definitely or likely disease-causing variant was identified. Phenotypic characterization was performed using a locally established instance of the Phenotips platform, in which patient phenotype collection is based on HPO terminology.[13]

To compare the relative yield of phenotype gene panel with that of the disease gene panel approach, we also prompted the clinical geneticists examining the patients to select the single most relevant classic NGS panel for each case submitted from the selection of NGS disease panels compiled in the EuroGenTest NGS panel database, which was obtained in July 2014.[3] The EuroGenTest panel database can be accessed at http://www.eurogentest.org/index.php?id=668, under heading titled NGS panel database.

After collecting this information, we generated a custom phenotype-based gene target and investigated whether the gene with a causative variant in the set of patients was captured in the disease gene panel, the phenotype gene panel, or both gene sets.

To compare the performance of phenotype gene panels versus disease gene panels, we utilized a measure of *causal gene inclusion rate* as the proportion of causative genes captured by either the phenotype or disease gene panel.

Web tool for phenotype panel generation

To enable the use of publicly available phenotype panels, we prepared a Web tool that enables phenotype gene panel generation based on clinical symptoms and signs of patients in clinical practice. The tool allows the user to find and select the phenotype traits observed in their patient and identify genes associated with relevant clinical symptoms and signs. The tool then creates a panel of genes associated with the observed phenotypes. Users can inspect and download the complete panel of genes associated with the set of signs and symptoms observed in a patient, making it possible to use the generated set in post hoc analysis for filtration of exome and genome sequencing data. We also implemented the labeling of genes that are included in the set of ACMG genes, for which reporting of potential incidental findings has been recommended.[14] The tool is available at http://kimg.eu/generator/.

RESULTS

Landscape of phenotype–gene associations in the HPO database

We initially assessed the phenotype similarities between pairs of genes co-occurring in disease gene panels. The analysis of phenotype–gene associations has shown that genes co-existing in disease gene panels share a significant excess of phenotype associations in contrast to pairs of genes originating from discordant gene panels (**Figure 1**). Notably, we could determine

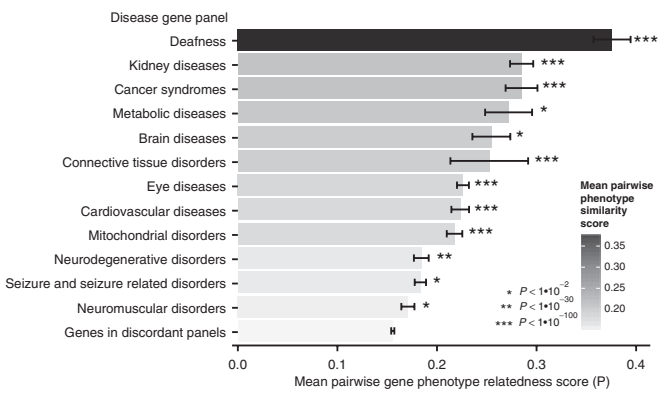

Figure 1 Summary of phenotype similarity scores for pairs of genes in classic groups of gene panels. Mean values for phenotype similarity of genes within disease gene panels are presented, along with 95% confidence intervals. *Mann-Whitney *U* test significance values when comparing pairwise gene phenotype relatedness scores in disease gene panels with pairwise scores of randomly selected gene pairs.

that genes belonging to a subset of panels share particularly extensive phenotype associations, most notably deafness, kidney disorders, and cancer syndrome genes, reflecting distinct and homogeneous clinical phenotype similarities among genes included in these panels. By contrast, subsets of genes assigned to panels related to neurological disorders (neurodegenerative disorders, epilepsy, and neuromuscular disorders) presented less extensive phenotypic conformance, suggesting greater heterogeneity in phenotype presentation for this subset of genes.

We also performed network analysis of overall associations based on the phenotype similarity of genes, which has accordingly revealed the spontaneous formation of gene clusters that, in general, mimic the assortment of genes in disease gene panels (**Figure 2**). Although most genes in disease panels related to eye and kidney disorders occur in tightly connected phenotype hubs, other genes were found to display more complex patterns of network allocation. Genes in neuromuscular panels naturally formed two clusters, one associated with phenotypes related to myopathies and the other related to clinical presentation of neuropathies. Furthermore, genes associated with muscle disorders closely associated with genes related to cardiomyopathies, reflecting the overlap between genes related to cardiomyopathies and various forms of muscular disorders observed in clinical presentation of these two disease groups. Metabolic disorders with multifaceted clinical presentations were interspersed among a variety of disease groups.

Although an overall phenotype similarity of genes recapitulated broad patterns of gene combinations in classic gene panels, we observed several cases of genes departing from such assortments. Examples of this occurrence include the *DPM3* gene, a cause of the congenital disorder glycosylation, which usually occurs in metabolic panels. Phenotype similarity–based clustering, however, placed the gene in close association with cardiomyopathy and myopathy genes, in accordance with predominant involvement of heart and muscle encountered in this disorder (**Figure 2b**). Another example is the positioning of the *TGFBR2* gene, which was, surprisingly, among the genes related to cancer-susceptibility syndromes (**Figure 2c**) instead of the expected association with cardiovascular disorders. Further examination of gene–phenotype associations supports this observation because mutations in the gene are also related to familial cancer syndromes in addition to cardiovascular diseases.

Specific cases of utility for phenotype-driven gene panels

We subsequently selected a subset of cases illustrating the utility of the approach using the generation of gene panels.

We generated a phenotype-based gene target for Parkinson disease, which has clearly defined hallmark symptoms of bradykinesia (HP:0002067), tremor (HP:0001337), and rigidity (HP:0002063).[15] Mutations in more than 15 genes have been associated with the inheritance of PD in families and are tracked in the OMIM database. We have compared familial PD gene panels offered by various diagnostic centers (as tracked by the EuroGenTest NGS panel database). Surprisingly, we identified considerable discrepancies in gene panels offered, in not only the size of panels, which ranged from 9 to 17 genes, but also their composition; only four genes were found to overlap across all three panels used for these comparisons (**Supplementary Figure S1a** online). We attempted panel generation based on hallmark HPO terms associated with Parkinson disease and generated a target of 240 genes related to at least one of the three phenotype traits associated with PD. This panel has captured a majority of genes related to PD and was sensitive enough to detect all the genes included across existing gene panels (with the exception of *UCHL1*, which has so far been associated only with Parkinson disease in association studies).

Additionally, several additional genes of interest were captured by our approach but were absent in currently offered panels. Notably, *JPH3*, *SLC30A10*, *EIF4G1*, and *DNAJC5* genes were found to be associated with phenotypes of rigidity, tremor, and Parkinsonism phenotypes but were absent from currently offered gene panels despite significant phenotypic overlap with PD.

We also evaluated the utility of phenotype-based panels in the design of gene panels for diagnosis of specific genetic syndromes. In the example of Usher syndrome, we were able to capture all genes related to the disorder after generating the panel based on two phenotypic traits: retinitis pigmentosa and hearing impairment (**Supplementary Figure S1b** online), whereas coverage of pertinent genes varied significantly among various sequencing providers.

Validation of phenotype-driven gene target generation in the clinical setting

Of the 405 cases included in the retrospective analysis, 159 cases were those with likely or definitely causative variants identified by exome sequencing (with overall diagnostic yield of the original clinical analysis estimated to be 39.3%). Original exome sequencing analysis established a diagnosis in 35 patients with previously unclassified disorders, and in 9 patients the diagnosis was reclassified.

In 25.4% (103 of 405) cases, the detected causative mutation was captured within the disease panel, but the phenotype gene panel rate was 29.7% (120 of 405). The causative gene inclusion rate was 64.8% for disease gene panels and 75.5% for phenotype gene panels (**Figure 3**). Phenotype gene panels evidently outperformed the disease gene panels in the cases with diagnosis establishment or reclassification. Disease gene panels captured the causative gene in only 45.7% (16 of 35) of diagnosis establishment cases and 22.2% (2 of 9) of reclassification cases, whereas the phenotype gene panel captured the causative gene in 62.9% (22 of 35) and 66.7% (6 of 9) of the diagnosis establishment and reclassification cases, respectively.

We further observed that the two types of panels were largely complementary. Utilizing the combined disease and phenotype gene panel, we attained the highest overall diagnostic yield of 35.8% of the cases (145 of 405), capturing a majority of diagnosis establishment cases (82.9%, 29 of 35) and a majority of reclassification cases (77.8%, 7 of 9).

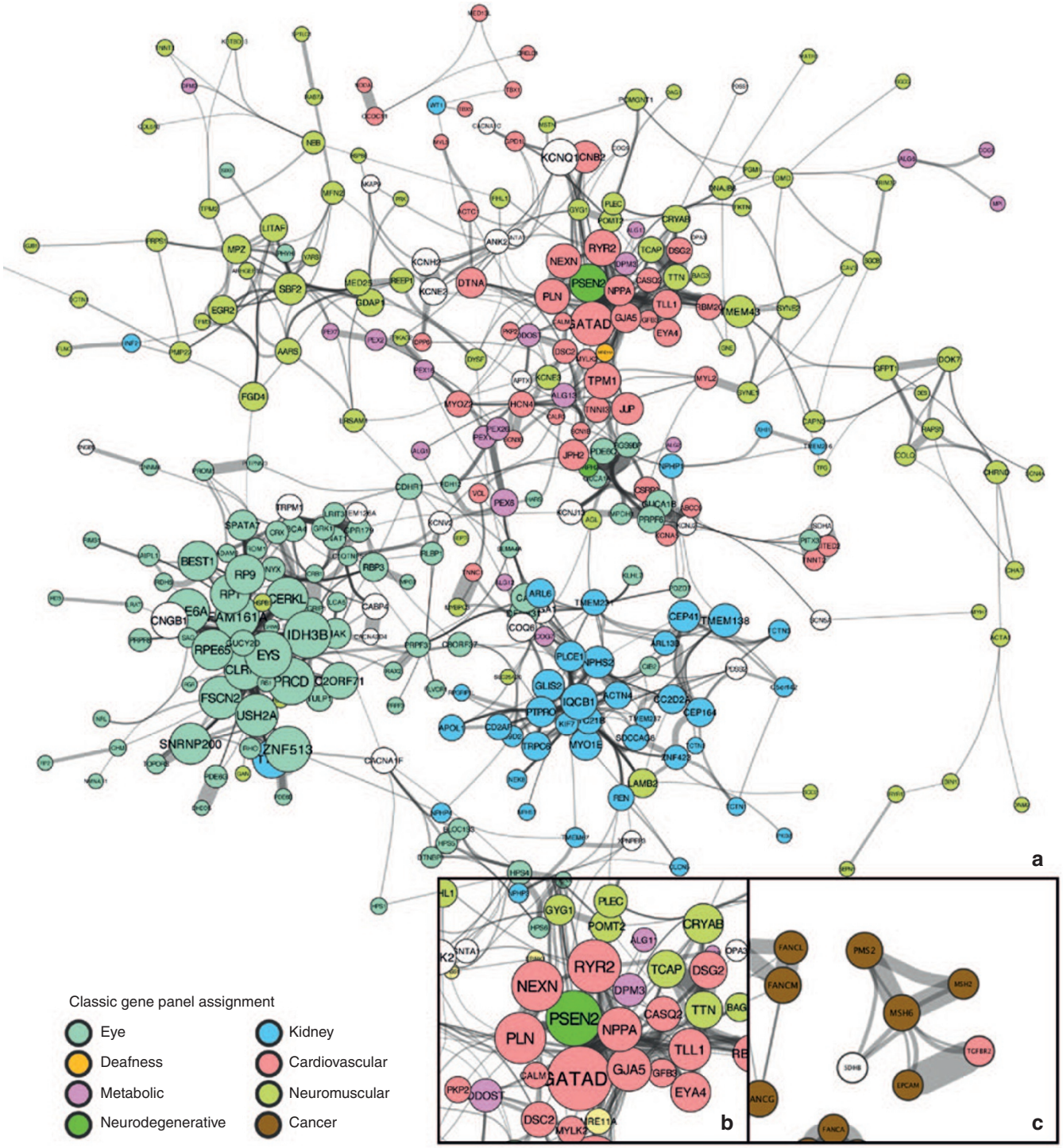

Figure 2 The network of genes originating from eight different gene panels clustered according to phenotype similarity score using the force-directed network layout weighted by the strength of phenotype similarity of presented gene profiles. The network (**a**) shows a clear trend toward co-clustering of genes originating from shared panels. (**b**) The unexpected positioning of the DPM3 gene in proximity to cardiomyopathy and myopathy genes, despite the fact that it originates from the panel of genes related to metabolic disorders. Similarly, (**c**) shows the association of TGFBR2 gene with genes related to cancer syndromes, reflecting its association with cancer development predisposition.

On average, phenotype gene panels contained 458 genes and were based on the average number of 4 HPO tags attributed to each case by the referring clinician. Disease gene panels contained an average number of 29 genes. As expected, the overlap was nonrandom, with the disease and phenotype gene panel sharing, on average, 12 genes, signifying convergence

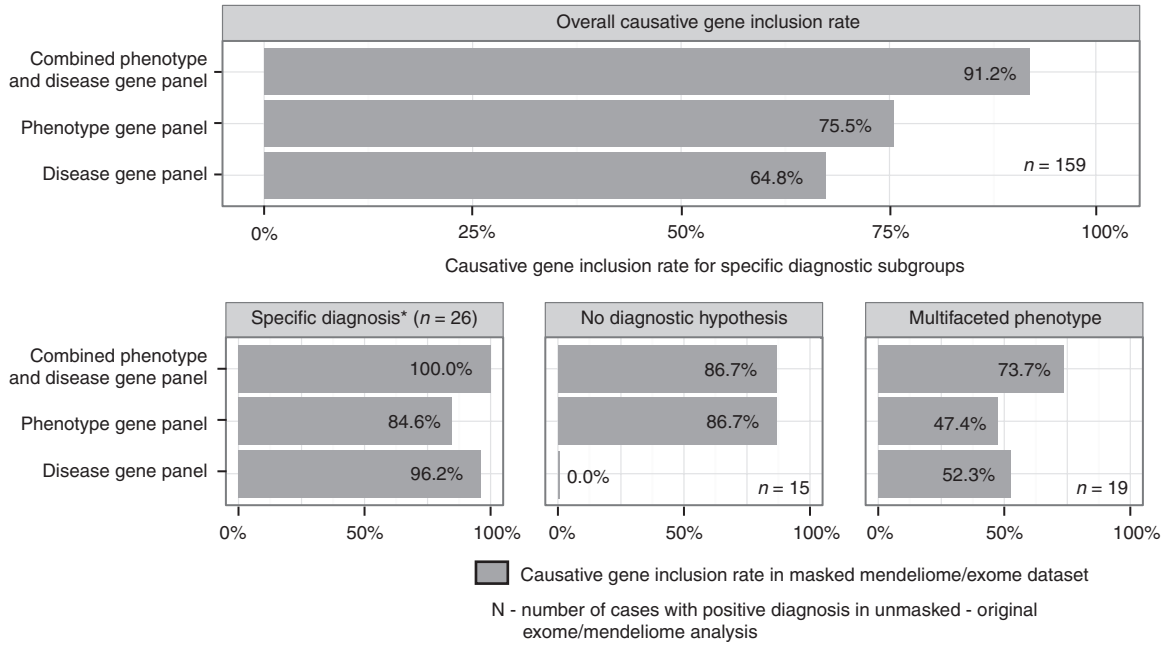

Figure 3 Comparative capture rate of genes with definitely or likely pathogenic variants when comparing disease versus phenotype gene panels overall and in specific diagnostic cases of specific diagnoses, cases without diagnostic hypothesis, and cases with multifaceted, complex phenotypes.

of gene content in both panel types. We also noted that overlap between disease gene panels in phenotype gene panels increased with the number of HPO terms used in case submissions, again signifying the convergence of gene sets represented by both approaches to masking exome data (data not shown).

Of the cases submitted, 15 were classified as having complex phenotypes by the clinician, so direction of diagnostic approach and disease panel selection were not possible. In these patients, phenotype-based gene target generation reached a high causative gene inclusion rate of 86.7% (13 of 15, **Figure 3**). In an additional 19 cases, referring clinicians noted the need for investigation of multiple disease gene panels. In this category, both strategies reached comparable performance rates, but they also reached a causative gene inclusion rate of 73.7% (14 of 19; **Figure 3**) when we used a combined set of genes from a patient's disease and phenotype gene panel.

In the subset of 26 cases, the suspected diagnosis suggested a narrow gene target with clinical presentation specific enough to limit the detection to a minimal set of genes (such as neurofibromatosis type 1 or polycystic kidney disorder). In these cases, use of a disease gene panel resulted in a causal gene inclusion rate of 96.2% (25 of 26), whereas phenotype gene panels included the causal gene in 84.6% (21 of 26), making this situation the sole exception in which phenotype gene panels were outperformed by disease gene panels.

We subsequently analyzed the patterns of HPO annotations and compared the dynamics of phenotype gene panel sensitivity in relation to comprehensiveness of patient phenotype annotation. We found a clear increase in the sensitivity of generated panels with an increasing number of HPO terms attributed

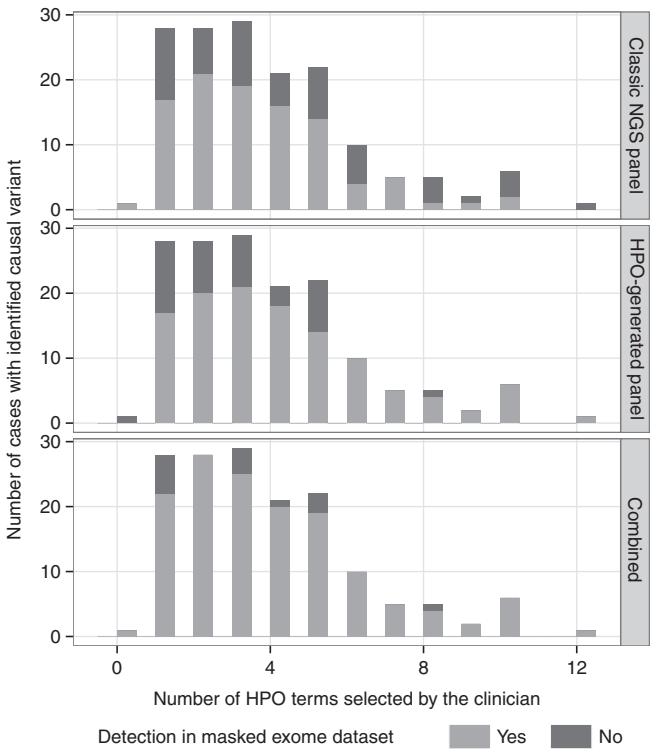

Figure 4 Causative gene inclusion rate for phenotype gene panels and its dependence on the number of human phenotype ontology (HPO) terms assigned to each diagnostic case. With the increasing number HPO identifiers, an improvement in the causative gene inclusion rate is observed for phenotype gene panels. In cases annotated with more than five HPO terms, phenotype gene panels included the identified causative gene in 96.6% (with a single identified causative gene missed from the panel).

to each case (**Figure 4**). This effect was most notable in cases where more than five annotations were attributed in case submissions. In comparison to disease gene panels, which captured the causative gene 44.9% of the time, the phenotype gene panels proved to be robust in clinically well-annotated cases with at least five selected HPO terms, reaching a causative gene inclusion rate of 96.6%.

We further surveyed cases in which there was a discrepancy in the capture of a causative gene by disease or phenotype gene panel, and **Table 1** provides a set of illustrative examples. In general, disease panels failed due to incompleteness of panel composition, in cases with etiological disease reclassification, cases with complex phenotypes without any diagnostic direction, and cases in which disease panels did not contain an updated set of phenotype-associated genes (**Table 1**, Section A). Failure of causative gene inclusion in the phenotype gene panel was attributed primarily to suboptimal HPO annotations of a causative gene and incomplete phenotype characterization of patients using HPO terms (**Table 1**, Section B). As expected, both approaches to masking the exome data failed to capture genes with novel or recently reported disease associations (**Table 1**, Section C).

Finally, we estimated the rate of incidental findings with analysis limited to phenotype gene panels. In the retrospectively analyzed population of 405 cases, we analyzed variants in 53 genes with potentially actionable findings compiled by the American College of Medical Genetics and Genomics. We focused on variants with clear pathogenic or likely pathogenic assertions collected in the ClinVar database. In only a single case was the presence of a reportable incidental finding

Table 1 Examples with different capture outcomes for classic (disease gene panels) versus phenotype-based gene targets

Case	Leading phenotype	Number of HPO terms	Classic panel, selected by the clinician	Disease panel provider	Gene with causative variant	Classic panel hit	HPO panel hit	Likely cause of disparity
Section A, detection by HPO-generated panel but not by classic panel								
P0100	Nonspecific syndrome with intellectual disability, autism and multiple malformations	5	No specific panel could be assigned by the clinician	NA	*FOXP1*	No	Yes	Complex phenotype, no predefined panels captured the presentation of the patient sufficiently, preference to select multiple gene panels was chosen for this case
P0124	Arrhythmogenic right ventricular dysplasia	2	Arrhythmogenic right ventricular dysplasia panel	Provider A	*MYH7*	No	Yes	Reclassification of cardiac phenotype, overlapping phenotype presentation
P0623	Dystonia	1	Dystonia panel	Provider B	*NPC1*	No	Yes	Specific, atypical phenotype presentation in the patient
P0312	Premature ovarian failure	2	Premature ovarian failure panel	Provider B	*STAG3*	No	Yes	Gene has been described only recently, the HPO has already been updated, and classic disease gene panels do not have the mutated gene in the defined gene content
P0464	Nonsyndromic deafness	2	Deafness, nonsyndromic sensorineural, AD	Provider C	*SLC26A4*	No	Yes	Gene panel selected by the clinician was too narrow and limited to autosomal dominant mode of inheritance (affected parent was also affected by deafness)
P0257	Early-onset dementia	9	Dementia, all causes	Provider A	*TYROBP*	No	Yes	The selected panel was incomplete and the causative gene was not included, clinician remarked a preference to select multiple gene panels in this case
Section B, detection by classic panel but not by HPO-generated panel								
P0952	Syndromic hearing loss	1	Syndromic hearing loss	Provider A	*USH2A*	Yes	No	Incomplete HPO annotation of *USH2A* gene, low number of HPO annotations provided by the clinician
P0092	Congenital cavernous malformation	1	Congenital cavernous malformation	Provider C	*CCM2*	Yes	No	Suboptimal gene–phenotype annotations of *CCM2* gene, low number of HPO annotations provided by the clinician
P0480	Early-onset blindness	3	Stargardt disease and macular dystrophies	Provider B	*ABCA4*	Yes	No	Suboptimal gene–phenotype annotations of *ABCA4* gene
P0904	Cardiomyopathy noncompactive	2	Pan cardiomyopathy panel	Provider E	*DTNA*	Yes	No	Suboptimal gene–phenotype annotations of *DTNA* gene, low number of HPO annotations provided by the clinician
Section C, detection by neither classic panel nor HPO-generated panel								
P0730	Epileptic encephalopathy	3	Epileptic encephalopathy	Provider A	*SOX5*	No	No	Recently identified candidate causative gene with a *de novo* mutation—neither disease nor phenotype panel has the capacity to capture novel gene–phenotype associations

HPO, human phenotype ontology.

observed, resulting in an overall incidental finding rate of 0.2% (1 of 405) for analysis limited to phenotype gene panels.

DISCUSSION

In the present study we developed a new approach to phenotype-driven gene panel generation that offers the potential to meaningfully narrow the search of genes in diagnostic (NGS) to those specifically related to disease phenotype traits observed in the patient. We provide evidence that phenotype-based associations between genes reflect the associations between genes as they are observed in a clinical setting while expanding the gene target to accommodate phenotypically related genes not yet included in disease gene panels. In a validation study performed in a realistic clinical setting, we show that this strategy has the potential to improve the sensitivity of masked genome sequencing data analysis by introducing phenotype-guided gene selection, and that it is an approach that naturally fits into genome-wide sequencing workflows.

Progressively increasing the complexity and richness of known phenotype–gene associations and decreasing sequencing costs have given genome-wide sequencing approaches a promising future direction toward unified diagnostics of human genetic disorders.[1] Shifting genetic diagnostics from targeted gene panel approaches to genome-wide sequencing approaches requires the establishment of strategies to focus on pertinent gene targets aiming to minimize the burden of interpreting excessively large gene targets and to allow comprehensive analysis while minimizing the rate of incidental findings.[2,16] In addition to the most commonly utilized approach that employs predefined gene sets in curated gene panels, we introduce an alternative strategy in which gene target generation is based on systematically collected phenotype information in each patient.

We identified the HPO resource as a plausible basis for capturing phenotype–gene associations in such gene target generation. We surveyed the properties of HPO phenotype–gene associations and how they reflect the composition of currently used classic gene panels. We showed that the phenotype-based associations spontaneously recapitulate the classic gene panels solely on the basis of the phenotype compatibility of genes. In addition to this, we observed that, when clustered according to phenotype resemblance, some genes show unexpected associations with those from disparate sets of genes, signifying that the currently used approaches using disease panels may be limiting in various clinical settings.

In general, our experience in clinical sequencing is that clinical disease presentation in patients is commonly atypical or only partially penetrant, which may lead to the selection of an incorrect gene panel. This difficulty in a priori gene panel definition may result in the need to perform further rounds of genetic testing or may ultimately lead to nondetection of a genetic cause in the case. Defining the clinical target in terms of the phenotypes observed in patients provides the flexibility to capture the genes when only minor overlap is present between the clinically observed presentation and the phenotypes associated with the gene. As an example, a case submitted for sequencing for a suspected specific form of cardiomyopathy (ARVD) may carry a likely pathogenic variant in a gene associated with hypertrophic cardiomyopathy, which is in line with recently recognized overlapping of phenotype–gene associations in the field of cardiomyopathies.[17]

The introduction of new tools for genome-wide analysis in patients with rare diseases is constantly advancing the number of novel genotype–phenotype associations, including identification of novel candidate genes and expanding phenotype implications for genes with known disease-causing effects.[18] Therefore, it is essential to constantly update the selection of candidate genes to investigate cases submitted for sequencing. We show this in the case of identification of STAG3 gene mutation in a patient with premature ovarian failure identified in mid-2014 that was, for this reason, absent from the disease panels utilized in this study, but it was seamlessly captured in the phenotype gene panel.

Furthermore, considering the complex clinical presentation in some patients, it may be impossible to define a plausible clinical target prior to genetic testing. This situation is most evident in patients with undiagnosed genetic syndromes with multiple affected organ systems and multimodal phenotype presentation in which a diagnostic process cannot be directed toward any specific disease. In such cases, phenotype-based panel generation naturally supplants this deficiency of disease panels, directing diagnostics into an expanded set of phenotypically compatible genes. Among such cases in the present study, phenotype-based gene targeting captured a causative gene in the majority of cases (86.7%).

Another challenge in gene panel sequencing as it is utilized presently is the wide disparity between compositions across sequencing providers. This presents an open need to find a consensus approach to define the core set of genes for inclusion in the panel across institutions, aiming to maximize the comprehensiveness of genetic tests offered across institutions. Although this is difficult to achieve in the widely disparate assortment of panels defined across different institutions, HPO as a unified, centralized, and constantly updated resource could serve as the resource for such consensus and updated panel generation.

Our validation study has shown that phenotype-based gene target definition supersedes the disease panels in their capacity to capture the causal gene in the masked genome data analysis. Interestingly, combining the selection of a single disease gene panel along with a phenotype gene panel resulted in significantly improved capture rates, detecting causative genes in more than 90% of cases in our masked first-tier exome analysis. This observation suggests that the disease gene panel selection along with phenotype-based panel selection act as two complementary approaches. This probably stems from the fact that the mapping of genes to phenotypes is still progressing, and it is our expectation that with continual improvement the HPO phenotype–gene associations will be comprehensive and contain

enough redundancy to account for variability in phenotype annotation among submitting clinicians. As in disease panels, our approach to masked analysis does not allow for identification of novel genes that have no known phenotype annotations in the HPO database.

Our results also show that the efficiency of phenotype-based gene targets is highly dependent on the number of HPO terms attributed to each case submitted for sequencing and that comprehensive phenotype characterization substantially improves the capture rate when using phenotype-based gene target generation.

In conclusion, we present a novel approach for phenotype-driven gene target generation to facilitate interpretation of genome-wide interpretation in the clinical setting. We showed that HPO phenotype language provides a useful resource for the generation of phenotype gene panels. We also demonstrated that this approach improves the diagnostic rate in comparison to currently used disease gene panel–oriented approaches while still offering specificity to focused interpretation of pertinent findings. The presented strategy may be used in most diverse clinical situations and outperforms currently used disease panel approaches in clinical cases in which disease diagnosis cannot be established a priori. Additionally, we demonstrated its robustness in situations with partially penetrant diseases and cases with misleading clinical presentation.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENT

This study was supported by grant P3-0326 from Slovenian Research Agency.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* 2013;14:681–691.
2. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* 2013;14:295–300.
3. Dierking A, Schmidtke J. The future of Clinical Utility Gene Cards in the context of next-generation sequencing diagnostic panels. *Eur J Hum Genet* 2014;22:1247.
4. Neveling K, Feenstra I, Gilissen C, et al. A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum Mutat* 2013;34:1721–1726.
5. Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 2014;312:1880–1887.
6. Korf BR, Rehm HL. New approaches to molecular diagnosis. *JAMA* 2013;309:1511–1521.
7. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;508:469–476.
8. Dorschner MO, Amendola LM, Turner EH, et al.; National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am J Hum Genet* 2013;93:631–640.
9. van El CG, Cornel MC, Borry P, et al.; ESHG Public and Professional Policy Committee. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur J Hum Genet* 2013;21(suppl 1):S1–S5.
10. Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet* 2010;77:525–534.
11. Kurbatova N, Adamusiak T, Kurnosov P, Swertz MA, Kapushesky M. ontoCAT: an R package for ontology traversal and search. *Bioinformatics* 2011;27:2468–2470.
12. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011;27:431–432.
13. Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat* 2013;34:1057–1065.
14. Green RC, Berg JS, Grody WW, et al.; American College of Medical Genetics and Genomics. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–574.
15. Klein C, Djarmati A. Parkinson disease: genetic testing in Parkinson disease-who should be assessed? *Nat Rev Neurol* 2011;7:7–9.
16. Ormond KE, Wheeler MT, Hudgins L, et al. Challenges in the clinical application of whole-genome sequencing. *Lancet* 2010;375:1749–1751.
17. Lopes LR, Zekavati A, Syrris P, et al.; Uk10k Consortium. Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *J Med Genet* 2013;50:228–239.
18. Iglesias A, Anyane-Yeboa K, Wynn J, et al. The usefulness of whole-exome sequencing in routine clinical practice. *Genet Med* 2014;16:922–931.