White Paper

Variant Analysis: Phenotype-Driven Ranking filter

Sample to Insight

Introduction

Whole genome and exome sequencing is widely used to identify disease-causing variants in patients with multiple congenital abnormalities and rare, undiagnosed genetic disorders. However, a key challenge in using this approach is finding the true causal variant among the hundreds of rare, functional (coding and/or regulatory) variants. It can take hours to evaluate the relationship between variants in a patient's sequence data and his phenotype or disease, in order to identify the disease-causing mutation (1). In addition, the disease-causing variant is successfully identified in only 25-30% cases (2,3). Therefore, we propose using phenotype and genotype data in conjunction to prioritize variants for further evaluation and increase overall solve rate. This approach draws from a network of phenotype-phenotype, phenotype-disease, and disease-gene relationships established from the QIAGEN Knowledge Base, and looks for plausible diseases that can explain both the phenotypes observed as well as the genetic variations detected (Figure 1 - A). For each disease, we can compute a score that represents the similarity between phenotype profile and disease, and this score is in turn used to rank variants that reside in disease-implicated genes.

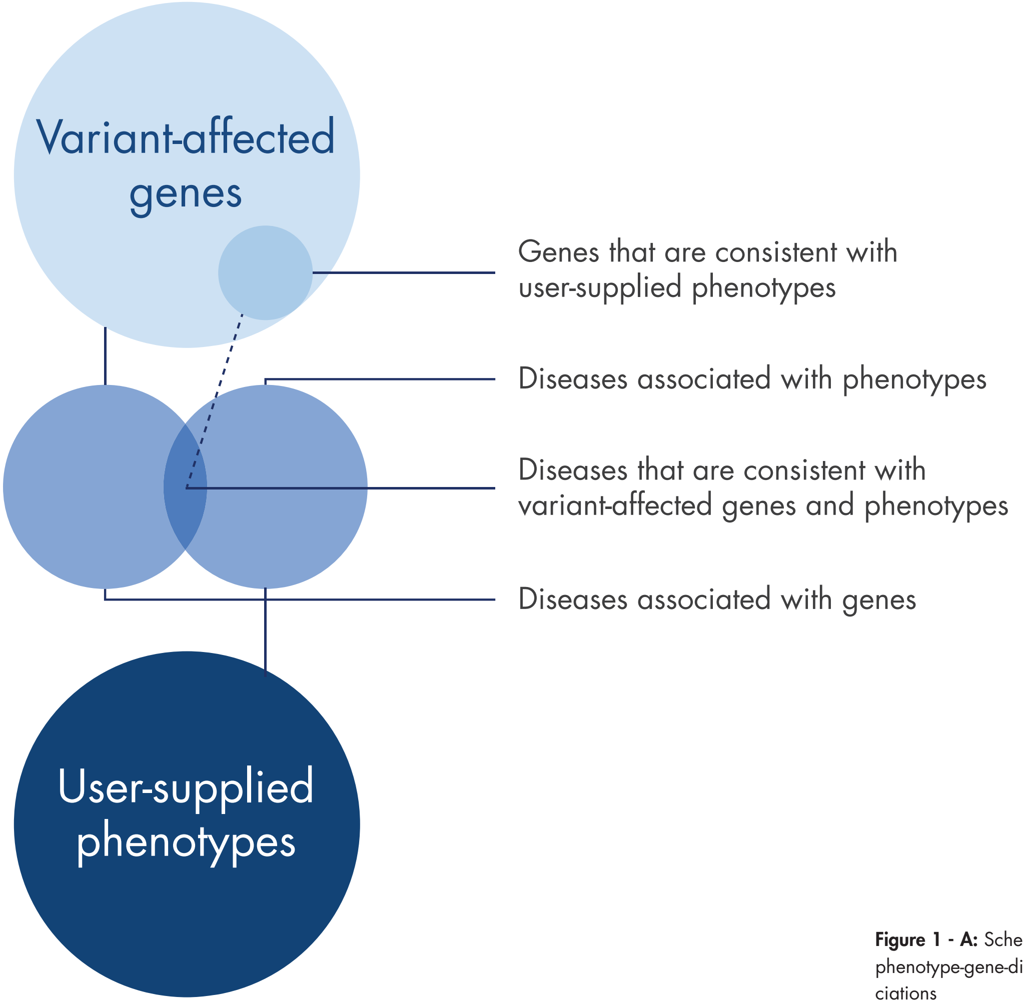

Variant-affected genes
Genes that are consistent with user-supplied phenotypes
Diseases associated with phenotypes
Diseases that are consistent with variant-affected genes and phenotypes
Diseases associated with genes
User-supplied phenotypes

Figure 1 - A: Schematic of phenotype-gene-disease associations

Mapping of phenotypes

Users may enter a phenotype term as free-text or provide an HPO identifier in standard format (e.g. HP:0000213). As a term is entered, the QIAGEN Knowledge Base supplies phenotypes matching the text as an autocompleted entry or as alternatives for selection. Spacing, capitalization, and hyphenation are normalized during fuzzy-matching of entered terms. Supported phenotype terms include all names and synonyms for any disease, abnormality, or biological process computationally associated with findings in the QIAGEN Knowledge Base. More than 60,000 phenotypes are available, including 44,000 phenotypes associated with variants in our Knowledge Base.

We currently support HPO phenotypes cited in 92% of the phenotype annotations described by HPO or Orphanet for OMIM or Orphanet diseases.

For supported HPO phenotypes, both primary and alternate identifiers, as well as primary term and all synonyms, are available for mapping. Inclusion of HPO terms has been prioritized based on frequency of their use in phenotype annotations, and improvements in coverage are ongoing.

Scoring algorithm

The scoring algorithm is based on a heuristic that uses evidence from the QIAGEN Knowledge Base to connect genes and associated diseases with user-provided disease phenotypes (Figure 2 - A). A directed network is built from gene/disease relationships and disease/phenotype relationships, as well as the process hierarchy (ontology) that relates more specific terms of diseases and phenotypes to more general terms in a hierarchical manner. For each gene/disease combination, a score is calculated indicating its relevance in the context of the user-provided, observed phenotypes. By and large, this score counts how many disease phenotypes can be explained by the disease; however, it also takes into account phenotype prevalence among all diseases represented in the Knowledge Base (measured as "specificity" below), as well as the confidence of connecting a given phenotype to a given disease when traversing the process hierarchy (expressed as "path weight" below). The specificity of a phenotype is given by

$$specificity = \frac{1}{1 + \log(\max(1, N_d))}$$

where N_d is the number of diseases that the phenotype is connected to in the network. The path weight of a (shortest) path from a phenotype to a disease in the network is

$$path\ weight = 0.75^N$$

where N is the number of links traversed through the process hierarchy. In the special case where a phenotype is a gene-associated disease itself, the path weight score is set to 1. The maximum path length when traversing the process hierarchy is 4. The total score for a given gene/disease combination is then computed as the sum over all phenotypes connected to the disease through at least one path:

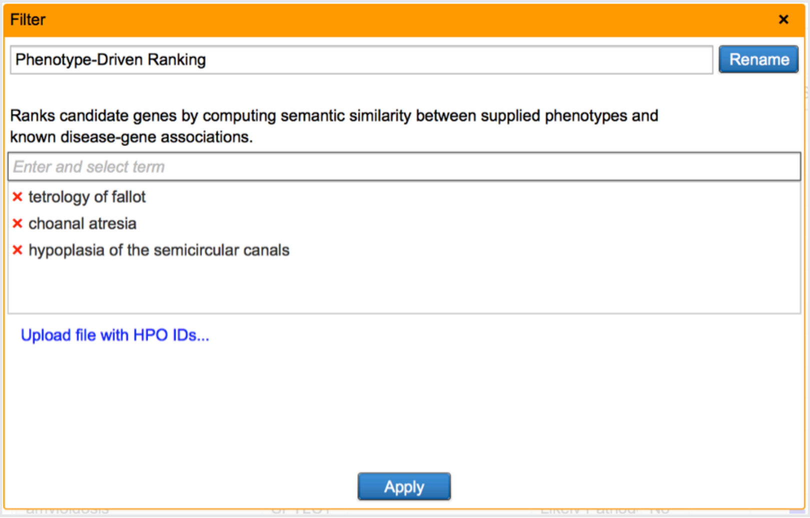
$$score = \sum_{phenotypes} specificity \times path\ weight$$

Note that the score only depends on the connected disease and will be the same for all genes that are associated with it.

The algorithm consists of the following steps:

1. Given genes coming through the filter cascade, determine set D of diseases correlated or caused by it.

2. From any given phenotype, determine shortest path(s) to a disease in D under the condition the path may not contain other diseases in D unless it corresponds to the given phenotype, and the last link in the path is a phenotype-disease relationship.



Filter ×

Phenotype-Driven Ranking

Ranks candidate genes by computing semantic similarity between supplied phenotypes and known disease-gene associations.

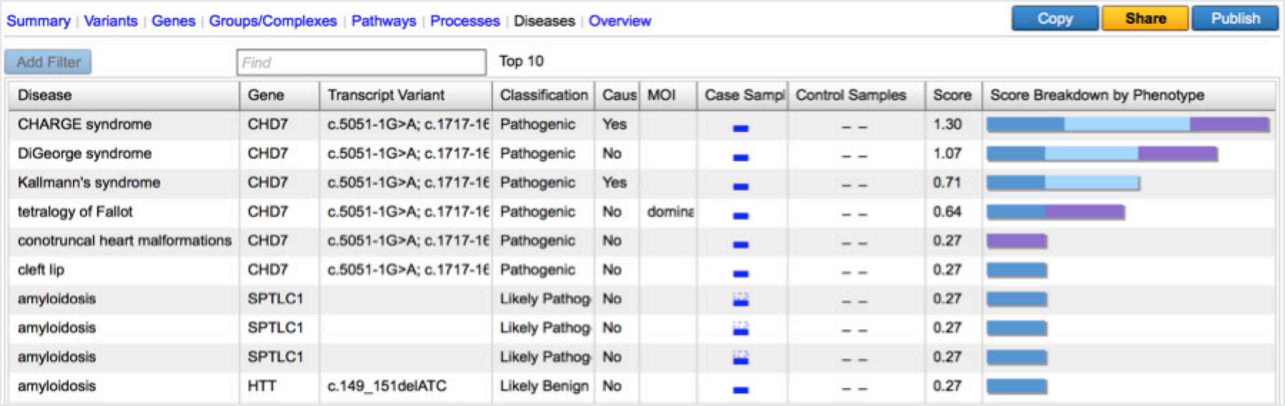
Enter and select term

× tetrology of fallot

× choanal atresia

× hypoplasia of the semicircular canals

Figure 2 - A: Phenotype-Driven Ranking filter showing user entered phenotypes



Summary | Variants | Genes | Groups/Complexes | Pathways | Processes | Diseases | Overview
Copy
Share
Publish
Add Filter
Find

Top 10

Disease	Gene	Transcript Variant	Classification	Caus	MOI	Case Sampl	Control Samples	Score	Score Breakdown by Phenotype
CHARGE syndrome	CHD7	c.5051-1G>A; c.1717-1€	Pathogenic	Yes		▬	– –	1.30	
DiGeorge syndrome	CHD7	c.5051-1G>A; c.1717-1€	Pathogenic	No		▬	– –	1.07	
Kallmann's syndrome	CHD7	c.5051-1G>A; c.1717-1€	Pathogenic	Yes		▬	– –	0.71	
tetralogy of Fallot	CHD7	c.5051-1G>A; c.1717-1€	Pathogenic	No	domina	▬	– –	0.64	
conotruncal heart malformations	CHD7	c.5051-1G>A; c.1717-1€	Pathogenic	No		▬	– –	0.27	
cleft lip	CHD7	c.5051-1G>A; c.1717-1€	Pathogenic	No		▬	– –	0.27	
amyloidosis	SPTLC1		Likely Pathog	No		▬	– –	0.27	
amyloidosis	SPTLC1		Likely Pathog	No		▬	– –	0.27	
amyloidosis	SPTLC1		Likely Pathog	No		▬	– –	0.27	
amyloidosis	HTT	c.149_151delATC	Likely Benign	No		▬	– –	0.27	

Figure 2 - B: Disease-gene pair ranked by the score generated using phenotype-disease relationship

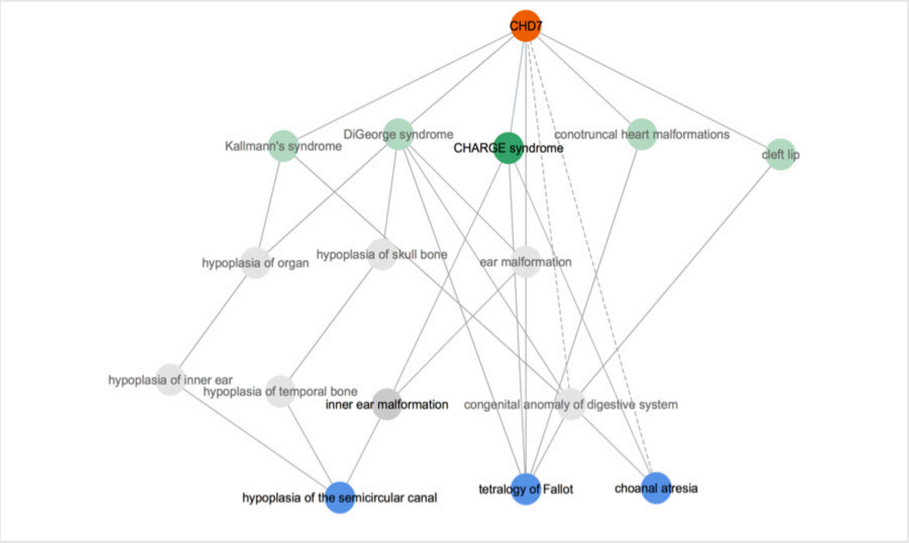


Figure 2 - C: Network diagram showing gene-disease-phenotype relationships

3. For a given disease in D, collect all paths connecting that disease to a phenotype, compute the total score above, and combine all paths into a network that is displayed.

Gene/disease pairs are then annotated with the type of relationship, i.e., whether they are causal (using OMIM) or represent an observed correlation, as well as by mechanism of inheritance if known. Gene/disease pairs are listed in a table and rank-ordered by their score (Figure 2 - B). For context and exploration, displayed networks also show paths connecting the same gene to other inferred diseases as well as to other genes connected to the displayed disease nodes (Figure 2 - C).

Benchmarking

In a benchmarking study, we used 29 cases with rare, congenital abnormalities from Inova Translational Medicine Institute (Fairfax, VA). The disease-causing variant for all these cases was previously identified. We used Ingenuity Variant Analysis to annotate and filter variants to a short list of rare, deleterious variants based on best practice guidelines using the Common Variants and Predicted Deleterious filters. The variants in this list were further filtered and ranked by likelihood of inferred diseases that are characterized by input phenotypes, using the Phenotype-Driven Ranking filter. This filter sorts genes and variants using phenotype-disease-gene relationships, as explained above. Furthermore, genes and variants are secondarily sorted using variant classifications and the table also lists the mode of inheritance when available, so that a reviewer can pick the most likely causal variant between two variants with the same score. In 22 out of the 29 cases (76%), the disease that was previously diagnosed and reported by the clinicians, along with the gene/variant linked to it, was correctly identified using the Phenotype-Driven Ranking filter. In 20 out of the 22 solved the causal variant ranked among the top 5 variants on the list.

Variant Rank	Percentage of solved cases
1	60
Top 5	90
Top 10	96

The Phenotype-Driven Ranking filter in Ingenuity Variant Analysis uses phenotypes to infer and rank matching diseases and enables prioritization of disease-causing variants from whole genome and exome sequence data for individuals with genetics disorders. This enables fast and accurate disease prediction based on clinical signs and symptoms observed alongside genotype information.

Reference

1. Frederick E. Dewey, et. al. Clinical Interpretation and Implications of Whole-Genome Sequencing. JAMA. 2014 Mar 12; 311(10): 1035–1045.

2. Clinical application of whole-exome sequencing across clinical indication. Kyle Retterer MS, et al. Genetics in Medicine (2015). doi: 10.1038/gim.2015.148

3. Clinical Impact and Cost-Effectiveness of Whole Exome Sequencing as a Diagnostic Tool: A Pediatric Center's Experience. Valencia CA, et al. Front Pediatr. (2015). 3: 67. doi: 10.3389/fped.2015.00067

QIAGEN Bioinformatics

EMEA
Silkeborgvej 2 · Prismet
8000 Aarhus C
Denmark
Phone: +45 7022 5509

Americas
1700 Seaport Boulevard #3
Redwood City · CA 94063
USA
Phone: +1 (617) 945 0178