1 HPO Venn Diagram

Supplementary Figure 1: HPO Venn diagram highlighting DRAM2 at the intersection of “Autosomal Recessive” and “Visual Impairement”.

A Venn diagram visualisation of the HPO to gene mapping is rendered with Javascript D3 (Supplementary Figure 1). Only genes for which there are filtered variants in the patient are displayed by default. It is also possible to view all genes by selecting “Toggle All”. Since the computation of the Venn diagram scales exponentially with the number of HPO terms, we currently limit to the five most specific terms.

2 PubmedScore

In the variants table on the patient page, genes can be prioritised by PubmedScore, estimated from literature evidence. Pubmed publications are searched by gene name and phenotype keywords and the score is simply calculated by summing the appearance of each user specified keyword in every returned Pubmed papers title and abstract. If we take DRAM2 as an example, Pubmed search with keywords, “blindness”, “macula”, “macular”, “pigmentosa”, “retina”, “retinal”, “retinitis” and “stargardt”, returns three publications (pubmed IDs: 27518550, 26720460, 25983245) (Supplementary Figure 2). The keywords appear 16 times in the titles and abstracts of the three publications, so $\text{pubmedscore}(DRAM2)$ is set to 16.
Supplementary Figure 2: PubmedScore search for DRA M2 and the keywords “blindness”, “macula”, “macular”, “pigmentosa”, “retina”, “retinal”, “retinitis” and “stargardt”.

In order to define what constitutes a significant score, we obtained PubmedScores, using the same set of keywords, for retinal dystrophy genes from RETNET (https://sph.uth.edu/Retnet/disease.htm) and 1,000 randomly chosen non-retinal dystrophy genes.

As can be seen in Supplementary Figure 3, the kernel density plots of the PubmedScore distributions for RETNET and non-RETNET genes crosses over at 10.443. Based on this, a PubmedScore of 11 would be a good cutoff for best precision / recall outcome (precision=0.95, recall=0.62). We acknowledge there is still some limitations to this approach due to ambiguous gene names which can be mistaken for common acronyms such as for example, NHS (National Health Service). This issue could be addressed by adopting natural language processing to help contextualise genes and phenotypes.

3 Exomiser

Exomiser can help find candidate dysfunctional genes by performing cross-species phenotype comparisons of the observed patient phenotypes, as encoded by the HPO. It uses an implementation of the PHenotypic Interpretation of Variants in Exomes (PHIVE) (Robinson et al, 2014) algorithm to provide a similarity score to model organisms or known diseases. While Exomiser usually scores and prioritises variants, here we used only the gene-level phenotype-scoring in a trimmed-down version of the software provided by Julius Jacobsen (available on request). This version uses a modified version of PHIVE, known as HiPHIVE (Bone et al, 2016) which compares patient phenotype data with known human disease-gene phenotypes (OMIM (Amberger et al, 2011), Orphanet (Maiella et al, 2013), Decipher (Bragin et al, 2014)), mouse (MGI (Eppig et al, 2005), IMPC (Bult et al, 2013), MPO (Smith et al, 2005)), and zebrafish (Zfin (Van Slyke et al, 2014)) model data. PHIVE ranks models using pre-computed phenotype comparisons calculated by the PhenoDigm tool (Smedley
et al, 2013). This approach works particularly well for embryonic development genes such as \textit{PAX6} (Robinson et al, 2013; Smedley et al, 2015), whose function are conserved across species.

4 Phenogenon

Unrelated patients with observed HPO terms are included in the analysis, the number of which is denoted as $Pat_a$. The number of patients affected by a given HPO term \( h \) is denoted as $Pat_h$. The number of patients who have a specific genotype in a gene \( g \) is denoted as $Pat_g$ (this may be a single variant or more than one for compound hets). The variants are filtered by ExAC allele frequency and CADD phred score (Kircher et al, 2014) according to user specified thresholds so that only rare, predicted, damaging variants are included. The number of patients having both HPO term \( h \) and filtered variants in \( g \) is denoted as $Pat_{gh}$. Therefore one can construct a $2 \times 2$ table for gene \( g \) and HPO term \( h \) combination:

| Num of patients without variant in \( g \) | $Pat_a - Pat_h - Pat_g + Pat_{gh}$ | $Pat_h - Pat_{gh}$ |
| Num of patients with variant in \( g \) | $Pat_g - Pat_{gh}$ | $Pat_{gh}$ |

Fisher’s Exact test is used to test the non-independence between \( g \) and \( h \), and the $\phi$ correlation coefficient is used to quantify the correlation between \( g \) and \( h \). Fisher’s Exact test produces two p-values: right-tail and left-tail p-values. A small right-tail p value indicates a positive correlation between \( g \) and \( h \). The Phenogenon score is defined as the $-\log_{10}(p\text{value})$. The sign and magnitude of $\phi$ also determines the type of correlation. A strong positive $\phi$ indicates that rare damaging variants in the gene appear more often than expected by chance in patients with that HPO term. This suggests the disrupted gene may cause the phenotype. A negative $\phi$ would indicate a gene has less rare damaging variants than expected by chance, which could suggest a phenotype is driven by a functionally enhanced gene. The method is applied to a gene with two possible inheritance modes, dominant where the patient has to have at least one qualified variant, and recessive where the patient has to have at least two qualified variants or a homozygote variant. The significance of the top p-values can also help infer whether a gene is more likely to be dominant or recessive.

One drawback to this approach is its susceptibility to sequencing bias. For instance, if all individuals which share a particular HPO term have improved coverage at a given gene then there will be an enrichment of rare variants in that gene. An approach we have used to correct for this, is to filter out genes which have very low number of variants to number of patient ratios (VP). We used a threshold of 0.8 for recessive gens and 1.0 for dominant.

In order to get a handle on an appropriate p-value threshold to select in order to limit the number of false positives, we applied increasing thresholds $10^{-6}$, $10^{-5}$ and $10^{-4}$ to all Phenogenon results (Supplementary Table 1). As expected given our cohort HPO term distribution (Supplementary Figure 8), we have the most power to detect gene to ophthalmic phenotypes associations. However, as we relax the threshold and tolerate some false positives, we start to detect associations with other phenotypes such as “premature loss of teeth”, “dermatological manifestations” and “skeletal dysplasia” from \textit{TERT} and \textit{ERCC6L2}. Interestingly, some of the false positives detected, \textit{GRHL2}, \textit{USP25}, \textit{TEMN2}, \textit{NOTCH2}, \textit{ATRNL1} and \textit{PTCH1}, are conserved genes with a high probability of being loss intolerant according to the ExAC database (Samocha et al, 2014). These genes are linked to a range of developmental phenotypes and so may well be worthy of further investigation. Other false positives phenotypic associations, such as “Neurodevelopmental abnormality” in \textit{ERCC6L2}, may be linked to known phenotype associations such as “Microcephaly”. The Phenogenon HPO graph view, available on the gene page, is particularly suited to explore these phenotype spectrums. For example, as shown in Supplementary Figure 4, \textit{USH2A}, which causes Usher 2 syndrome, shows both enrichment of “rod-cone dystrophy” and “hearing impairment”.
<table>
<thead>
<tr>
<th>Phenogenon p-value threshold</th>
<th>True Positive Gene-Phenotype Associations</th>
<th>False Positive Gene-Phenotype Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-06</td>
<td><em>GUCY2D</em>: Nystagmus <em>GUCY2D</em>: Undetectable electroretinogram <em>GUCY2D</em>: Visual loss</td>
<td><em>KRTAP4-3</em>: Cataract <em>KRTAP4-3</em>: Constriction of peripheral visual field</td>
</tr>
<tr>
<td>1e-05</td>
<td><em>CNGB1</em>: Constriction of peripheral visual field <em>GUCY2D</em>: Nystagmus <em>GUCY2D</em>: Undetectable electroretinogram <em>GUCY2D</em>: Visual loss <em>USH2A</em>: Rod-cone dystrophy</td>
<td><em>PTCH1</em>: Dermatological manifestations of systemic disorders <em>USP25</em>: Ichthyosis</td>
</tr>
<tr>
<td>1e-04</td>
<td><em>ABCA4</em>: Juvenile onset <em>CNGB1</em>: Cataract <em>CNGB1</em>: Constriction of peripheral visual field <em>CNGB1</em>: Nyctalopia <em>CNGB3</em>: Achromatopsia <em>ERCC6L2</em>: Microcephaly <em>GUCY2D</em>: Nystagmus <em>GUCY2D</em>: Visual loss <em>TERT</em>: Premature loss of teeth <em>USH2A</em>: Autosomal recessive inheritance <em>USH2A</em>: Rod-cone dystrophy <em>USH2A</em>: Visual impairment</td>
<td><em>ATRNL1</em>: Abnormality of the genital system <em>C12orf50</em>: Aplastic anemia <em>ERCC6L2</em>: Neurodevelopmental abnormality <em>GRHL2</em>: Abnormality of the urinary system <em>KRTAP4-3</em>: Cataract <em>KRTAP4-3</em>: Constriction of peripheral visual field <em>KRTAP4-3</em>: Nyctalopia</td>
</tr>
</tbody>
</table>

Supplementary Table 1: True positive and false positive gene-phenotype associations reported by Phenogenon for different p-value thresholds in Phenopolis.
5 SimReg

SimReg implements an inference procedure under the "similarity regression" model described in (Greene et al, 2016). It is used here to estimate a probability of association between HPO encoded phenotypes and a binary genotype vector indicating whether a sequenced individual harbours a rare variant in a particular gene. This is done by comparing the evidence for a random model with one in which the log odds of observing a rare variant is linked to the similarity between HPO profiles and an estimated characteristic HPO profile. The method was applied to all genes under dominant and recessive mode of inheritance assumptions, i.e. setting the binary genotype to 1 if individuals carried at least 1 and 2 rare alleles respectively, and 0 otherwise. One application was to the gene GUCY2D, which was estimated to have a probability of association of 0.98 (Supplementary Figure 5). The estimated characteristic phenotype, shown below, corresponds well to the literature phenotype (Gregory-Evans et al, 2000). SimReg is available for download on CRAN https://cran.r-project.org/web/packages/SimReg/index.html.

6 Use cases from the UK Inherited Retinal Dystrophy Consortium (UKIRDC)

An interactive tutorial is available on the individual page which explains how to use the system. Here is an illustration of how the Phenopolis gene prioritisation has been used in the UKIRDC to identify the causal genetic mutations for two patients affected by retinal dystrophy:

- A known dominant retinal dystrophy gene: PAX6 (Supplementary Figure 6).
- A novel recessive retinal dystrophy gene: GNB3 (Arno et al, 2016) (Supplementary Figure 7).
Supplementary Figure 3: Distribution of PubmedScore for 245 RETNET genes and 1,000 random non-RETNET genes. The keywords included in the search were: “retina, retinal, retinitis, blindness, macula, macular, stargardt, pigmentosa, amaurosis”. The green line shows the threshold of 11.
Supplementary Figure 4: Phenogenon HPO diagram for recessive $USH2A$ mutations detects “hearing impairment” and “rod-cone dystrophy” as enriched HPO terms. HPO terms which are already known to be linked to the gene are highlighted in red.
Supplementary Figure 5: SimReg HPO diagram for GUCY2D.

Supplementary Figure 6: Missense mutation in PAX6 known dominant retinal dystrophy gene. PAX6 is ranked third on Exomiser score.
Supplementary Figure 7: Homozygous stop gained in GNB3, a novel recessive retinal dystrophy gene. GNB3 is ranked first on Phenogenon score and third on Exomiser score.
Supplementary Figure 8: Distribution of HPO terms in Phenopolis.
References


