Phen-Gen: Combining Phenotype and Genotype to Analyze Rare Disorders

Asif Javed[1], Saloni Agrawal[1], Pauline C. Ng[1]

[1]Computational & Systems Biology, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore

Supplementary material

VAAST, PHEVOR and eXtasy simulations

The performance of Phen-Gen was compared against VAAST[1], PHEVOR[2], and eXtasy[3]. PHEVOR is a phenotype-based add-on tool which relies on genomic predictions from methods like VAAST. Both VAAST and eXtasy predict only on the damaging impact of amino acid changing SNVs and do not score indels, splice or noncoding variants. Hence for this comparison only nonsynonymous mutations in Human Gene Mutation Database (HGMD) were used. VAAST uses a unique file format (Genome Variant Format, or GVF) and their analysis package comes with conversion software to convert from VCFs to this format. Unfortunately, we faced compatibility issues with VCFs extracted for majority of the samples from African American population (ASW), within the 1000 Genomes data. To keep things simple, all ASW samples were removed from these. For computational efficiency, a random set of 1000 individuals were set aside as controls and 26 individuals were used to generate the simulated patients. For each causal variant, one individual was randomly selected and eXtasy, VAAST and Phen-Gen were run independently. The comparison was conducted using HGMD reported disease causal variants for which the disease symptoms are defined in OMIM.

For VAAST the 1000 individuals served as controls to estimate the composite likelihood under the healthy model to be compared against the disease model. Phen-Gen and eXtasy provide a continuum of rankings whereas VAAST (on average) identifies 14 genes at the top of the list. For a fair comparison, the true causal gene for VAAST ranking was assigned best, average, and worst rank among similarly ranked genes. For example, if VAAST assigns the true gene along with 4 other genes the same top rank. The causal gene will be ranked as 1, 3 and 5 for best, average and worst respectively. This is reflected in the three components of the bar in VAAST results (Fig. 1). VAAST was able correctly narrow down the true causal gene to within 14 genes

(on average) in 62% of dominant and 78% of recessive cases. In comparison, Phen-Gen was able to identify the correct etiological gene in 82% cases for dominant and in 97% cases for recessive OMIM reported disorders.

PHEVOR was recently published by the authors of VAAST. It combines information from multiple disease and functional ontology databases to re-prioritize genomic predictions in light of patient-specific symptoms. PHEVOR is only available as an online tool; making a comprehensive simulation comparison not possible. In its publication, PHEVOR was evaluated in 100 simulations. We chose the same number for our comparison. 50 dominant and 50 recessive cases were randomly chosen from the VAAST simulations. As a sanity check, VAAST predictions of this subset were compared against VAAST results for all simulations to ensure that the subsample reflects the general trend. PHEVOR web server only allows up to 5 disease symptoms. If the phenotype description exceeds this number, 5 symptoms were randomly chosen. The combination of VAAST and PHEVOR was able to assign the causal gene top rank in 66% of dominant and 68% of recessive simulations. In comparison for the same cases, Phen-Gen was able to correctly identify the causal gene in 90% of dominant and 88% of recessive simulations (Fig. 1).

The better performance of Phen-Gen can be attributed to methodological differences as well as difference in data sources integrated in the prediction. PHEVOR's 'ontology propagation' does not take in account the information content of its various data sources. Phen-Gen estimates this in the gene interaction network construction by relying on pathway databases. Pathway databases are often the best curated gene functional information available[4]. PHEVOR currently does not use this resource, although its authors concede integrating pathway information to be an active area of PHEVOR's development.

eXtasy relies on locus and gene specific information and does not use a control population. Hence the control set was not used in its prediction. Under default settings it evaluates all variants both common and rare. In its manuscript it was evaluated with rare variants (MAF<1%) as well. eXtasy's performance is thus evaluated under both scenarios. In the first, all variants in the individual exome were provided as an input. In the second, only rare variants were used and variants with MAF>1% in 1000 Genomes, dbSNP or National Heart, Lung, and Blood Institute's Exome Sequencing Project (ESP) were discarded. eXtasy is able to narrow down the causal variant within top 10 variants in 69% of the cases for dominant and 76% of the cases for recessive disorders. In comparison Phen-Gen is able to narrow down the causal variant within 10 variants in 91% of the cases for dominant and 98% of the cases for recessive disorders. The relative advantage of Phen-Gen can potentially stem from multiple factors. eXtasy does not incorporate the disease inheritance pattern in its prediction. It also ranks variants based on

individual phenotypes. Although the information is combined in rank aggregation, the interplay of different combination of phenotypes may not be well represented.

For the comparison Phen-Gen used the identical control set of 1000 individuals from VAAST to define the null distribution of genes. Phen-Gen outperforms eXtasy, VAAST and VAAST+PHEVOR by 19-58% (Fig. 1). eXtasy currently only allows a subset of disease symptoms defined in Human Phenotype Ontology. To investigate Phen-Gen's relative advantage due to a more comprehensive symptom list, Phen-Gen simulations were repeated restricting it to the symptoms accepted by eXtasy; Phen-Gen's ability to identify the true disease causal gene dropped by 1% in these simulations (results not shown).

For further comparison with VAAST, 44 phenotypic heterogeneous dataset was employed. The results show the relative performance and highlights Phen-Gen's advantage even when disease symptoms are not completely specific (Supplementary Fig. 8).

Performance using real dataset

Phen-Gen was evaluated using a recently published real dataset comprising of one hundred parents-child trio families with the children exhibiting intellectual disability symptoms[5]. This data contains variant calls for each family, as well as a rich resource of medical history detailing each patient's unique symptoms. The information was initially electronically parsed, and then manually evaluated by two of the authors independently to absolve the translation to Human Phenotype Ontology terms from any ambiguities. Further advice from a clinician was sought to remove any errors of interpretation.

On the genotypic side, a key challenge in handling the real dataset was the noise level in the variant calls. In the original study itself, the focus was on de novo mutations and the provided 'high quality calls' cast an initial wide net to encompass large number of candidate mutations. In the original publication, these calls were pruned by further bioinformatic processing and finally bench validation. Replicating this effort required access to the patient samples themselves, which was not possible. The validation rate of de novo mutations reported even after further bioinformatic processing was 11.4%[3]. To sidestep validation issues due to noise level in the data and focus on the downstream analysis that Phen-Gen aims to provide, we initially focused on eighteen families with variants implicated in recessive or X-mode inheritance. Seven families were further removed from consideration as the variants reported in the publication were not observed in the correct inheritance pattern in the provided data

and had likely been corrected in the bench validation[3]. Hence eleven families were used in this current study.

The performance of Phen-Gen is evaluated for variants reported in the original publication (Supplementary Table 7). Phen-Gen allows for variants with MAF below 1% in public databases. The first column of Phen-Gen results depicts the performance in this scenario. The original study however discarded all known variants (already existing in dbSNP, or observed in their in-house database). After adopting the same filtering criteria as the original study, 16/21 genes agreed with the original study, and only five are potentially false positives. Of these five variants, two were homozygous in the patient and observed in heterozygous state in two different families and the remaining three are only observed in the respective family. Of the three novel variants two are indels which were not evaluated in the original study. These five variants cannot be ruled out as disease causal candidates based purely on the bioinformatics analysis of the variant calls and further bench validation and functional analysis would be needed to implicate or exonerate them.

Intellectual disability is a genetically diverse disorder and it has been estimated that more than a thousand different genes may be playing a role[6]. Thus despite the rich and detailed patient history, it is challenging to list it down to a few genes based purely on symptomatic knowledge. We analyzed the symptoms of 58 patients with reported recessively inherited deleterious mutations, de novo mutations, or X-linked mutations in males (Supplementary Tables 3 and 9 in the original publication[5]). The analysis relied only on just the phenotypic information to rank the genes for each patient. To quantify the performance of our approach, we computed the sum of the ranks of all reported genes in the respective patient's phenotypic match. The aim of this analysis is to highlight that the genes reported in the original study based on genetic evidence tend to have a significantly lower rank in our lists than observed just by chance. Each individual patient's rank should have a uniform distribution under the null hypothesis and hence their sum - an Irwin-Hall distribution- is approximated by a Normal distribution. The results indicate that the rank sum of reported genes is significantly lower than expected by chance ($P \leq 0.0036$).

The authors split the gene set into known, unknown and candidate genes. The 'known genes' had been implicated for intellectual disability in prior literature. 'Candidate genes' had not been directly reported for intellectual disability, but are linked to brain and embryonic development and there is further evidence of their involvement. The remaining genes are deemed 'unknown'. Even when we focused on individual gene subclasses, the rank sums in all three categories were significantly lower: known genes ($P \leq 2.44\times10^{-8}$), candidate genes ($P \leq 0.0051$), and unknown genes ($P \leq 0.0192$). In the original study, only known genes carrying deleterious

variants were considered confirmed as diagnosed; or if the same candidate gene harboring damaging mutations was observed in two different patients with similar symptoms. Known genes implicated in these patients are expected to be ranked lower. Our results indicate that although the unknown genes set could harbor a lot of incidental genes, they are likely to contain some true positives as well; and the diagnostic yield of this dataset can potentially be improved by further corroboration.

Comparison with PHIVE

PHIVE[7] was evaluated in a simulation framework similar to the one presented in this paper. The authors used a somewhat different HGMD version (869 diseases versus 765 for Phen-Gen) and this could potentially contribute to the difference in results. Since PHIVE is only available as an online server and not a downloadable package, it is not possible to do a comprehensive evaluation using the same dataset. The authors report 66% power to identify the causal gene in dominant and 83% in recessive disorders. Comparison of these reported results reveals Phen-Gen has 13-16% higher efficacy in identifying the true causal mutation in both dominant and recessive disorders (Table 1). This advantage may be attributed to different factors including the underlying prediction framework and better representation of the disease symptoms in human databases.

PHIVE uses phenotypes from mouse gene knockout experiments to establish phenotype to gene links. Using the mouse model to represent human phenotypes has a few limitations. Mouse knockout experiments have only been conducted for approximately a third of the human genes. There has been increasing focus in the recent past to generate a more comprehensive resource[8], and this would improve PHIVE predictions over time. In comparison human disease-gene association span only about 10% of the genes. Both these numbers support the usage of gene (or protein) interaction networks to extrapolate predictions to the remaining set of genes. PHIVE assigns a uniform phenotypic score to the 2/3rd gene set and the authors concede that integrating protein interaction information is a potential future direction to improve results. A second issue which would be more difficult to address is that human and mice do not share disease morphology for all disorders[9]. In particular it would not be representative of any primate specific, or even more constrained human-specific, traits. Ward and Kellis recently showed that a large number of human regulatory regions do not show evolutionary conservation in primate evolution[10], so the mouse model may not recapitulate recently acquired regulatory evolving traits. Despite these limitations, animal models provide an invaluable resource and integrating this information along with human disease matches is one avenue to improve Phen-Gen's predictions.

On the genotypic side, PHIVE assigns arbitrary pathogenicity scores for the different classes of coding mutations (all except missense). These scores were chosen to give optimal predictions within their simulation framework. Their performance declined by almost half in simulations spiking the causal variant in in-house exomes (Fig. 4a of the publication[7]). The authors attributed this decline to their reliance on allele frequency information which would be unavailable for novel variants observed in the in-house data. This issue will be faced during the analysis of any new dataset. Another factor potentially playing a role is that the 1000 Genomes data is highly curated and fine-tuning the predictions based on this dataset may have led to overfitting. For example, the indel calls in the public data have been improved in extensive bench validations[11], whereas in a practical scenario these calls tend to be more error prone.

Comparison with FunSeq

FunSeq[12] evaluates the regulatory role of a noncoding mutation using a combination of functionality categories similar to Phen-Gen. It identifies the top 0.4% of the genome as 'sensitive' to regulatory disrupting mutations. The selective constraint in each combination of annotations is estimated using enrichment of rare variants. FunSeq annotations are highly predictive of functionality but may not necessarily be all-encompassing. The method follows multiple screening steps to reduce the number of candidate mutations. Analysis of HGMD reported disease causal regulatory variants reveals that their initial filter (representing coding and sensitive regions) only captures 18% of these damaging variants and would annotate the rest as benign. Phen-Gen on the other hand assigns a continuum of probabilities to 80% of the variants. These variants are assigned a lower probability based on genotypic information (Supplementary Fig. 7). These variants would likely be poorly predicted by any genotype only approach and would require phenotypic support for improved predictions. Simulation results support this hypothesis; Phen-Gen is able to specify the true causal variant regulatory mutation in 49% of the cases with phenotype information (Supplementary Table 6). This predictive power stems from phenotypic corroboration with a 30 percentage points advantage in predicting the true causal gene compared to a purely genotype only approach; thus highlighting Phen-Gen's advantage of integrating this viable resource. In 69% of the cases, the disease gene appeared in Phen-Gen's top 10 list of candidate.

Prior probability for genomic predictor

It is estimated that about 10-15% of the genome is functionally active[13]. Assuming the average of these estimates and that all functional loci reside within our regions of interest span 19% of

the genome, we can compute the probability of a random annotated locus being functional to be 12.5/19 = 0.66. Next to compute the probability of a mutation at a functional region being deleterious, we used the PhyloP conserved bases as surrogates of the functional genome and computed the reduction in common mutations in dbSNPs (MAF>30%) at these sites in comparison to the rest of the genome. These numbers were combined to yield the probability of an annotated variant being damaging 0.0688 which is used as prior in the genomic predictor.

Incorporating pedigree information

For a dominant disease inheritance pattern genes harboring one or more damaging mutations are evaluated. Predicted damaging variants with one or more copies of the alternate allele in all cases and no copies of alternate allele in controls are considered. Only genes with a predicted probability higher (or equal) in cases than controls are incorporated in the downstream analysis.

For a recessive or compound heterozygous disease inheritance pattern, genes harboring two or more damaging mutations are evaluated. Predicted damaging variants with one or more copies of the alternate allele in all cases and one or zero copies of the alternate allele in controls are considered. Only genes with a predicted probability higher in cases than controls are incorporated in the downstream analysis. This allows for compound heterozygosity while reducing the false positives. If both parents are included in the analysis, there is a further filter to require at least one variant from each parent.

Phen-Gen allows the user to restrict the analysis to variants consistent with the pedigree structure of the family. This is the recommended settings as pedigree inconsistent variants harboring potential de novo mutations also tend to be enriched in sequencing errors. The software allows the user with leniency in this criterion by allowing pedigree-inconsistent variants if de novo mutations are the likely cause of the condition. In practice for real datasets however it is highly recommended to use a pedigree based variant caller (such as GATK 2 PhaseByTransmission walker) to prune out false de novo calls.

Evaluating null distribution of genes

To evaluate Phen-Gen's null distribution of genes we compared our predicted deleterious variant harboring genes with loss of function genes reported in ref. 11 and Residual Variation Intolerance Score (RVIS) reported in ref 14. Ref. 11 used a subset of 1000 Genomes individuals for their analysis and the predicted damaging variants were subsequently curated after further bench validations. Ref. 14 used NHLBI ESP allele frequencies and corrected for gene size as

larger genes are more likely to harbor incidental so-called damaging variants due to size. A cutoff of ninety fifth percentile was employed to their set of damaging variant harboring genes. Phen-Gen employs a one percentile cutoff for the null distribution. Despite methodological and dataset differences, genes which exceed this cutoff showed high enrichment in the respective datasets ($P \leq 8\times10^{-3}$ for McArthur et al and 10^{-2} for Petrovski et al using Fisher's exact test).

Supplementary Table 1
Additional statistics for results reported in Table 1

Performance in simulated patients with OMIM-listed disease symptoms and HGMD reported variants for coding predictor. The table is an extension of Table 1 with performance for genotype and phenotype-only predictors added for both known and unknown diseases in each category (see highlighted rows). The number of variants in each category is also included at the top of each table. The performance of both genotypic and phenotypic predictors in each category is presented. The percentage of known (gene disease association in the local Phenomizer database) and unknown variants in each category is also reported.

		Dominant				Recessive		
		Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**		**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Missense & Nonsense		9,194 variants, 1092 individuals				11,028 variants, 1092 individuals		
Phen-Gen	All	82	89	91	All	97	98	98
Genotype only		0	3	17		75	98	98
Phenotype only		30	66	72		23	49	67
Phen-Gen	Known (74%)	93	97	98	Known (91%)	97	98	98
Genotype only		0	3	19		75	98	98
Phenotype only		37	80	84		25	51	68
Phen-Gen	Unknown (26%)	51	67	72	Unknown (9%)	92	97	97
Genotype only		0	3	13		72	96	97
Phenotype only		12	26	40		13	29	54
Splice site		1,581 variants, 1092 individuals				1,899 variants, 1092 individuals		
Phen-Gen	All	80	85	85	All	87	87	87
Genotype only		0	4	29		72	87	87
Phenotype only		33	75	77		26	50	65

		Dominant				Recessive		
		Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**		**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Phen-Gen	Known (81%)	87	89	89	Known (94%)	87	87	87
Genotype only		0	4	30		72	87	87
Phenotype only		39	87	89		28	52	67
Phen-Gen	Unknown (19%)	49	67	70	Unknown (6%)	81	82	82
Genotype only		0	4	25		72	82	82
Phenotype only		5	20	26		3	16	43

Indel		5,972 variants, 1092 individuals				4,220 variants, 1092 individuals		
Phen-Gen	All	80	88	94	All	97	98	98
Genotype only		0	1	2		64	98	98
Phenotype only		31	72	74		24	56	72
Phen-Gen	Known (80%)	94	98	98	Known (91%)	97	98	98
Genotype only		0	1	1		64	98	98
Phenotype only		38	87	89		26	59	73
Phen-Gen	Unknown (20%)	43	61	73	Unkown (9%)	92	96	96
Genotype only		0	2	11		70	95	96
Phenotype only		8	21	32		11	27	56

Combined		16747 variants, 1092 individuals				17147 variants, 1092 indviduals		
Phen-Gen	All	81	88	91	All	96	97	97
Genotype only		0	3	13		72	97	97

		Dominant				Recessive		
		Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**		**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Phenotype only		31	69	74		24	51	68
Phen-Gen	Known (77%)	92	97	97	Known (91%)	96	97	97
Genotype only		0	3	14		72	97	97
Phenotype only		38	83	86		26	53	69
Phen-Gen	Unknown (23%)	43	61	73	Unknown (9%)	92	96	96
Genotype only		0	2	11		70	95	96
Phenotype only		8	21	32		11	27	56

Supplementary Table 2
Performance in novel disease gene discovery

The simulations are conducted with OMIM-listed disease symptoms and HGMD reported variants. In each category, Phen-Gen is evaluated with the knowledge of the respective known disease gene association masked from the simulation. The results are highlighted. Across all masked and unknown simulations Phen-Gen assigns the causal variant top rank in 71% of cases. For a comparison, Phen-Gen's performance in the known and unknown categories is also included from Supplementary Table 1. The results indicate a drop in performance for novel gene discovery in comparison to known associations, and highlight comparable performance to the true unknown cases. Since prior disease knowledge impacts only the phenotypic part of the prediction, Phen-Gen's prediction performance based solely on the phenotype is also included.

		Dominant			Recessive		
		Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**	**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Missense & Nonsense							
Known	Phen-Gen	93	97	98	97	98	98
	Phenotype only	30	66	72	25	51	68
Masked	Phen-Gen	58	78	84	90	97	98
	Phenotype only	13	23	26	12	17	18
Unknown	Phen-Gen	51	67	72	92	97	97
	Phenotype only	12	26	40	13	29	54
Splice site							
Known	Phen-Gen	87	89	89	87	87	87
	Phenotype only	39	87	89	26	50	65
Masked	Phen-Gen	47	61	64	83	87	87
	Phenotype only	5	18	20	14	22	23
Unknown	Phen-Gen	49	67	70	81	82	82
	Phenotype only	5	20	26	3	16	43

		Dominant			Recessive		
		Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**	**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Indel							
Known	Phen-Gen	94	98	98	97	98	98
	Phenotype only	38	87	89	26	59	73
Masked	Phen-Gen	55	69	74	88	97	97
	Phenotype only	12	23	25	11	21	23
Unknown	Phen-Gen	25	47	75	96	97	97
	Phenotype only	2	12	16	7	25	65
Combined							
Known	Phen-Gen	93	97	97	96	97	97
	Phenotype only	34	76	80	25	53	69
Masked	Phen-Gen	56	73	78	89	96	97
	Phenotype only	12	23	25	12	19	20
Unknown	Phen-Gen	43	61	73	92	96	96
	Phenotype only	8	21	32	11	27	56

Supplementary Table 3

Performance when only ESP is used as MAF filter

Performance in simulated patients with OMIM-listed disease symptoms and HGMD reported variants using only ESP common variants (minor allele frequency >1%) for filtration. The results when all three databases (ESP, dbSNP, 1000 Genomes) are used to define common variants are copied from Table 1 to indicate a drop in performance. The data shows that using all three databases to define common variants, confers a 13-21% advantage.

		Dominant			Recessive		
	MAF filtration database	**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**	**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Missense + Nonsense	ESP	69	81	84	76	92	94
	All 3	82	89	91	97	98	98
Splice site	ESP	69	80	82	73	86	86
	All 3	80	85	85	87	87	87
Indels	ESP	66	83	83	73	94	95
	All 3	80	88	94	97	98	98
Combined	ESP	68	82	84	75	92	94
	All 3	81	88	91	96	97	97

Supplementary Table 4

Performance in diseases with compound heterozygous inheritance pattern

Performance for compound heterozygous mutation pairs in coding and genomic regions is depicted. The table entries reflect the percentage of simulations in each category with the indicated result. The results are similar to the 96% observed for recessive simulations in Table 1 for coding variants.

	Coding			Genomic		
	Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**	**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Phen-Gen	97	98	98	30	59	62
Genotype only	25	98	98	8	16	35

Supplementary Table 5
Robustness to symptomatic heterogeneity

Performance in 44 disorders with variable symptoms is depicted. There were no indels reported for recessive disorders and only one genomic variant. Hence simulations for these variants were omitted. The table entries reflect the percentage of simulations in each category with the indicated result. This is a 5-7% drop off in performance in comparison to known diseases in Table 1 (92% for dominant and 96% for recessive). These results highlight Phen-Gen's robustness to symptomatic heterogeneity.

		Dominant			Recessive		
		Top gene (%)	**Top 5 genes (%)**	**Top 10 genes (%)**	**Top gene (%)**	**Top 5 genes (%)**	**Top 10 genes (%)**
Phen-Gen	Missense & Nonsense	88	95	97	90	95	96
	Splice site	86	89	89	86	88	89
	Indels	87	91	94			
	Combined	87	92	95	89	94	95

Supplementary Table 6

Performance in simulated patients for regulatory predictor

The performance in simulated patients with OMIM listed disease symptoms and HGMD regulatory disease causal variants is shown. Phen-Gen's genomic predictor was used for these simulations. The table entries reflect the percentage of simulations in each category with the indicated result.

	Dominant			Recessive			Combined (dominant + recessive)		
	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)	Top gene (%)	Top 5 genes (%)	Top 10 genes (%)
Regulatory									
Phen-Gen	58	64	65	40	61	72	49	62	69
Genotype only	1	2	2	34	51	62	19	29	34
Phenotype only	31	43	54	5	11	26	17	26	39

Supplementary Table 7

Performance in real patients

The table reflects performance of Phen-Gen for recessive and X-linked implicated variants in real familial data[5] . The trio IDs correspond to the family identifiers in ref 2. Similarly the gene classification reflects prior knowledge of the gene's involvement in intellectual disability as defined in the original publication. Phen-Gen employs a 1% MAF cutoff whereas the original publication removed all variants reported to dbSNP or observed in their in-house database. For the latter screen all variants common amongst the families was employed. The performance using both these cutoff is depicted.

			Phen-Gen's Rank	
Trio ID	**Gene**	**Classification**	**Inclusive of dbSNP MAF<1%**	**Filtration criteria from ref. 2**
4	*FANCB*	Unknown	18	3
4	*PDHA1*	Known	5	2
4	*GUCY2F*	Unknown	4	1
16	*ENOX2*	Unknown	2	1
18	*ARHGEF9*	Known	8	1
25	*GPM6B*	Unknown	7	1
41	*ARHGEF9*	Known	1	1
42	*DDX26B*	Unknown	13	3
72	*PDZD11*	Unknown	3	3
93	*TRPC5*	Candidate	4	1
12	*SYCP2L*	Unknown	2	1
12	*VPS13B*	Known	14	4
12	*C8orf59*	Unknown	4	2
12	*PRUNE2*	Unknown	7	3
24	*PCNT*	Known	7	2
70	*IQGAP2*	Unknown	6	1

References

1. Yandell, M. *et al. Genome Res.* **21,** 1529–1542 (2011).
2. Singleton, M.V *et al. Am. J. Hum. Genet.* **94,** 599–610 (2014).
3. Sifrim, A. *et al. Nat Methods* **10,** 1083–1084 (2013).
4. Wu, G., Feng, X. & Stein, L. *Genome Biol.* **11,** R53 (2010).
5. De Ligt, J. *et al. N. Engl. J. Med.* **367,** 1921–9 (2012).
6. Van Bokhoven, H. *Annu. Rev. Genet.* **45,** 81–104 (2011).
7. Robinson, P.N. *et al. Genome Res.* **24,** 340–8 (2014).
8. Brown, S.D. M. & Moore, M.W. *Mamm. Genome* **23,** 632–640 (2012).
9. Seok, J. *et al. Proc. Natl. Acad. Sci. U. S. A.* **110,** 3507–3512 (2013).
10. Ward, L. D. & Kellis, M. *Science* **337,** 1675–8 (2012).
11. MacArthur, D.G. *et al. Sci.* **335 ,** 823–828 (2012).
12. Khurana, E. *et al. Science* **342,** 1235587 (2013).
13. Hardison, R.C. & Ponting, C.P. *Genome Res.* **21,** 1769–1776 (2011).
14. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. *PLoS Genet.* **9,** (2013).