Supplementary Material

Varant

Varant is an open source genetic variant annotation tool (written in Python, http://compbio.berkeley.edu/proj/varant). Varant provides five categories of annotation based on 17 data sources: variant identity and frequency, experimentally-defined genomic features, predicted genomic features, variant/gene phenotypes, and prediction of mutation impact. It has been used in a number of clinical research studies (Patel et al. 2015; Chan et al. 2016; Punwani et al. 2016).

QC Analysis Results

Supp. Fig. S2A shows that the Ts/Tv ratios for all the samples are clustered between 2.2 and 3.2. For the human genome based on 1000Genomes data, Wang et al. 2015 have shown that the Ts/Tv ratio is ~3 for SNVs inside exons and ~ 2 elsewhere. Since the capture regions cover more than just exons (1350 exonic and 39 intronic regions for 83 genes), the Ts/Tv ratio for SNVs is expected to lie between 2 and 3, consistent with the plot and the pattern is very similar for the 1000 Genomes samples. The Het/Hom ratios for all samples except one (P8) are clustered between 1.1 and 2.6. P8 is an outlier and carries more homozygous SNPs. It is established that on a genome scale, the Het/Hom ratio is close to 1.5 (McKernan et al. 2009; Schuster et al. 2010) but it also depends on whether a population incorporates recent admixture (skewing towards heterozygosity) or inbreeding (skewing towards homozygosity). The Het/Hom ratio range here is similar to that of the 1000Genomes sample. Thus, by these measures, the data, with the exception of P8, are of good quality.

Supp. Fig. S2B shows that for the capture v01 samples have 1200 to 2000 low quality and 200 to 940 no call sites. The captures v02 samples have 230 to 330 low quality and 90 to 180 no call sites. We expect that if any causative variant falls on one of these no call or low-quality sites it will be completely missed.

Supp. Fig. S2C shows that the number of common SNVs per sample cluster between 232 and 312 for the 96 samples sequenced using Capture v01 and between 132 and 175 for the 10 samples sequenced using Capture v02, consistent with Capture v02 covering 19 fewer genes than Capture v01. Non-African samples have a lower rare variant load (ranging from 8 to 33) than the African samples (ranging from 19 to 80). This pattern is also seen in the samples from 1000Genomes samples and has been previously reported in the literature (Durbin et al. 2010; Zawistowski et al. 2014). The novel (i.e. not found in 1000 Genomes and ExAC) SNV count is in the range of 0 to 8 with a median of 1 per sample and is similar to the count observed in 1000 Genomes dataset where the range is 0 to 12 with a median of 2 per sample. These novel variants be present in the patient's family or be de novo in the patient but it is not possible to distinguish these two situations given only the patient's variant data.

Supp. Fig. S2D shows that for the 96 Capture v01 samples, the common Indel count is between 5 and 16 where as for the 10 Capture v02 samples the count is 3 to 6 per sample. We observe that the distributions of the rare Indel is between 0 and 12 and novel Indel is between 0 and 4. Two African samples (P2 and P83) are identified as outliers carrying more rare Indels compared to rest of the Hopkins samples and 1000 Genomes dataset. However, the Indel counts are similar to 1000 Genomes dataset. By all these measures, the Hopkins data appears to be of high quality.

Supp. Fig. S3 shows the distribution of average read depth for 83 genes across all samples. Average read depth varies more substantially across the 106 samples (horizontal variation) than across the 83 genes (vertical). Genes not included (blue boxes) in the 10 capture v02 samples are evident. Though there is variation in the gene coverage across samples, from 107X to 983X, even the lowest coverage should be adequate for diagnostic analysis and confirmatory testing as shown by Strom et al. 2014.

We identified nine capture regions (occurring in eight genes) with anomalous read depth in more than 90 of the 106 samples (Supp. Table S1). One of these has very high coverage and the others have low coverage. Two of these regions lie in the major isoform of one gene, HYDIN, one high (Exon 53) and one low (Exon 60). Exon 53 has greater than 600X coverage in 70 samples (Supp. Fig. S4). Exon 60 has no coverage in 78 samples and less than 20X coverage in three more samples. The other anomalous regions are unlikely to affect downstream analysis because they are either deep intron, present in a minor isoform of the gene or the actual coverage of the region is at least 100X in most of the samples.

References

Chan AY, Punwani D, Kadlecek TA, Cowan MJ, Olson JL, Mathes EF, Sunderam U, Man Fu S, Srinivasan R, Kuriyan J, Brenner SE, Weiss A, et al. 2016. A novel human autoimmune syndrome caused by combined hypomorphic and activating mutations in ZAP-70. J Exp Med 213:155–65.

Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, La Vega FM De, Donnelly P, Egholm M, Flicek P, et al. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19:1527–1541.

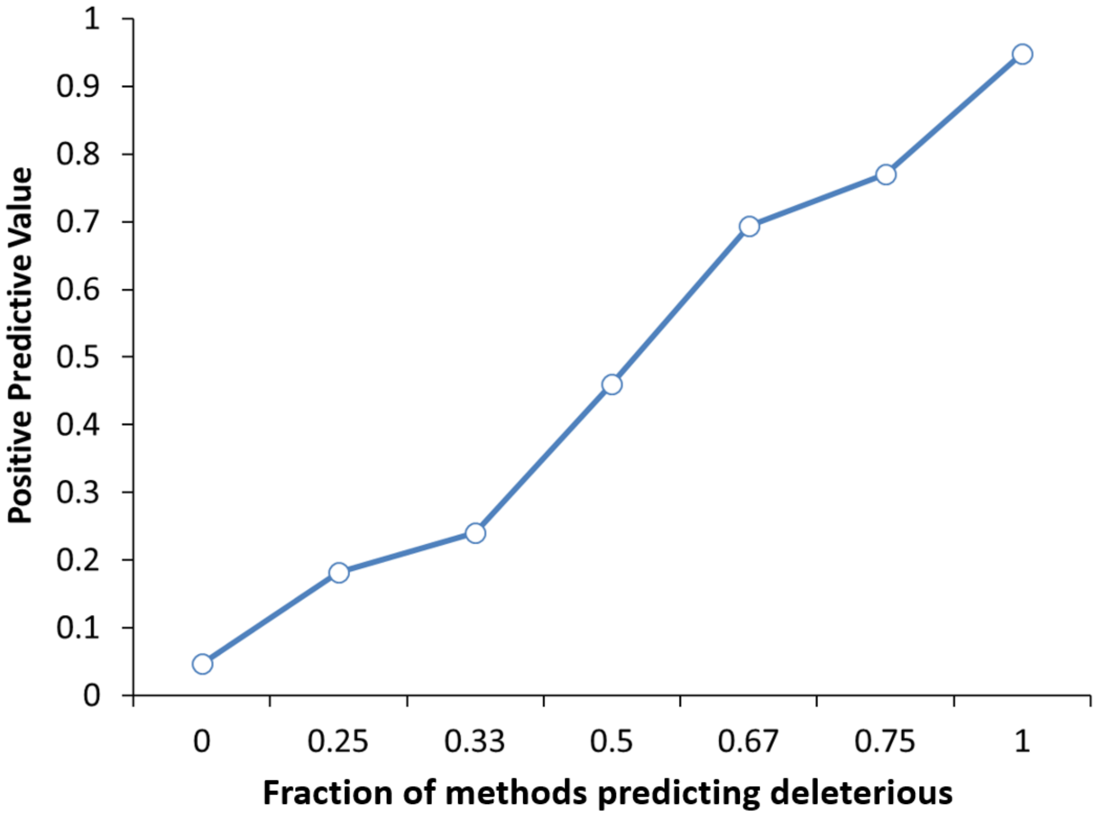
Patel JP, Puck JM, Srinivasan R, Brown C, Sunderam U, Kundu K, Brenner SE, Gatti RA, Church JA. 2015. Nijmegen Breakage Syndrome Detected by Newborn Screening for T Cell Receptor Excision Circles (TRECs). J Clin Immunol 35:227.

Punwani D, Zhang Y, Yu J, Cowan MJ, Rana S, Kwan A, Adhikari AN, Lizama CO, Mendelsohn BA, Fahl SP, Chellappan A, Srinivasan R, et al. 2016. Multisystem Anomalies in Severe Combined Immunodeficiency with Mutant BCL11B. N Engl J Med 375:2165–2176.

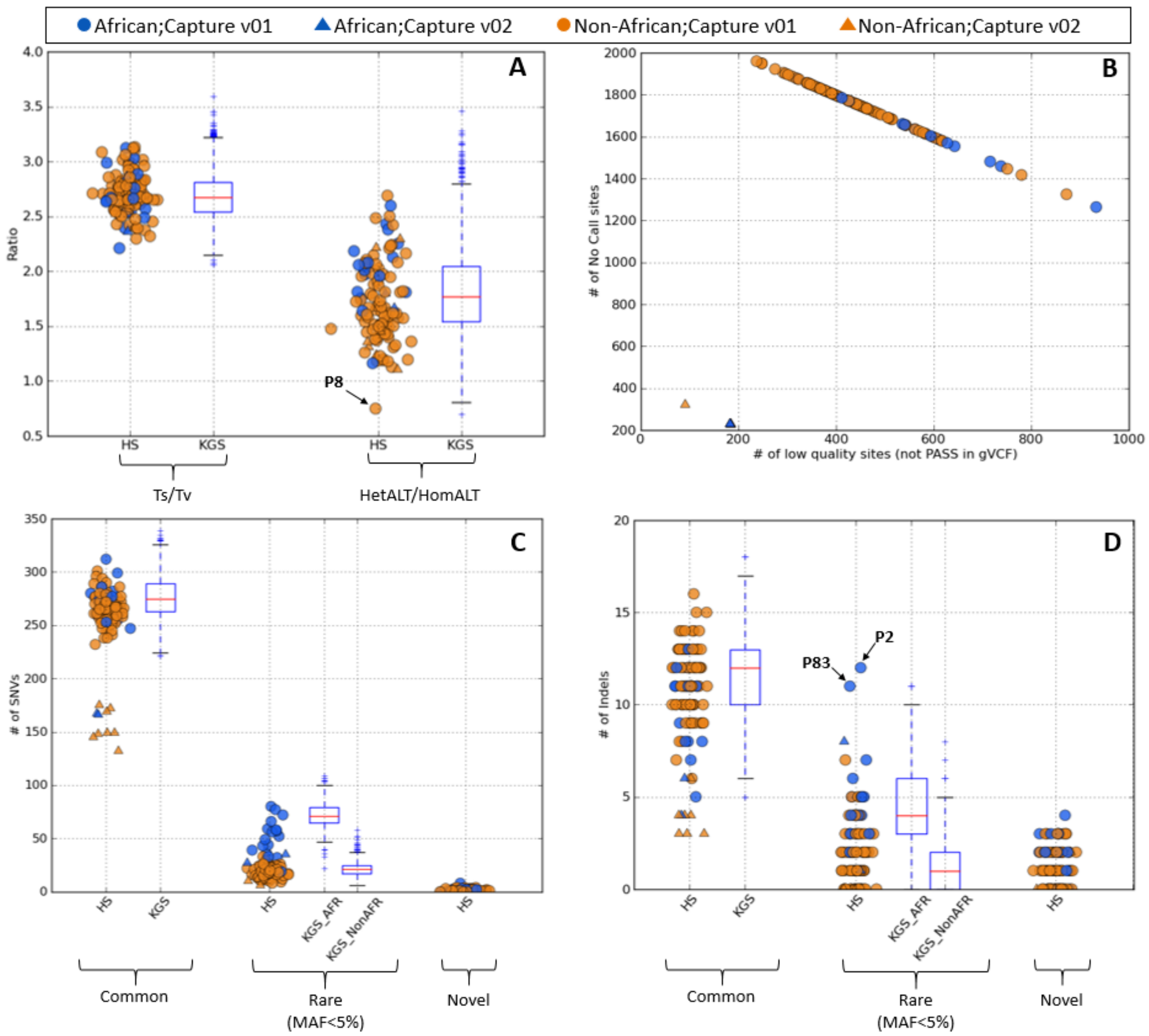
Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947.

Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, Deignan JL. 2014. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med 16:510–515.

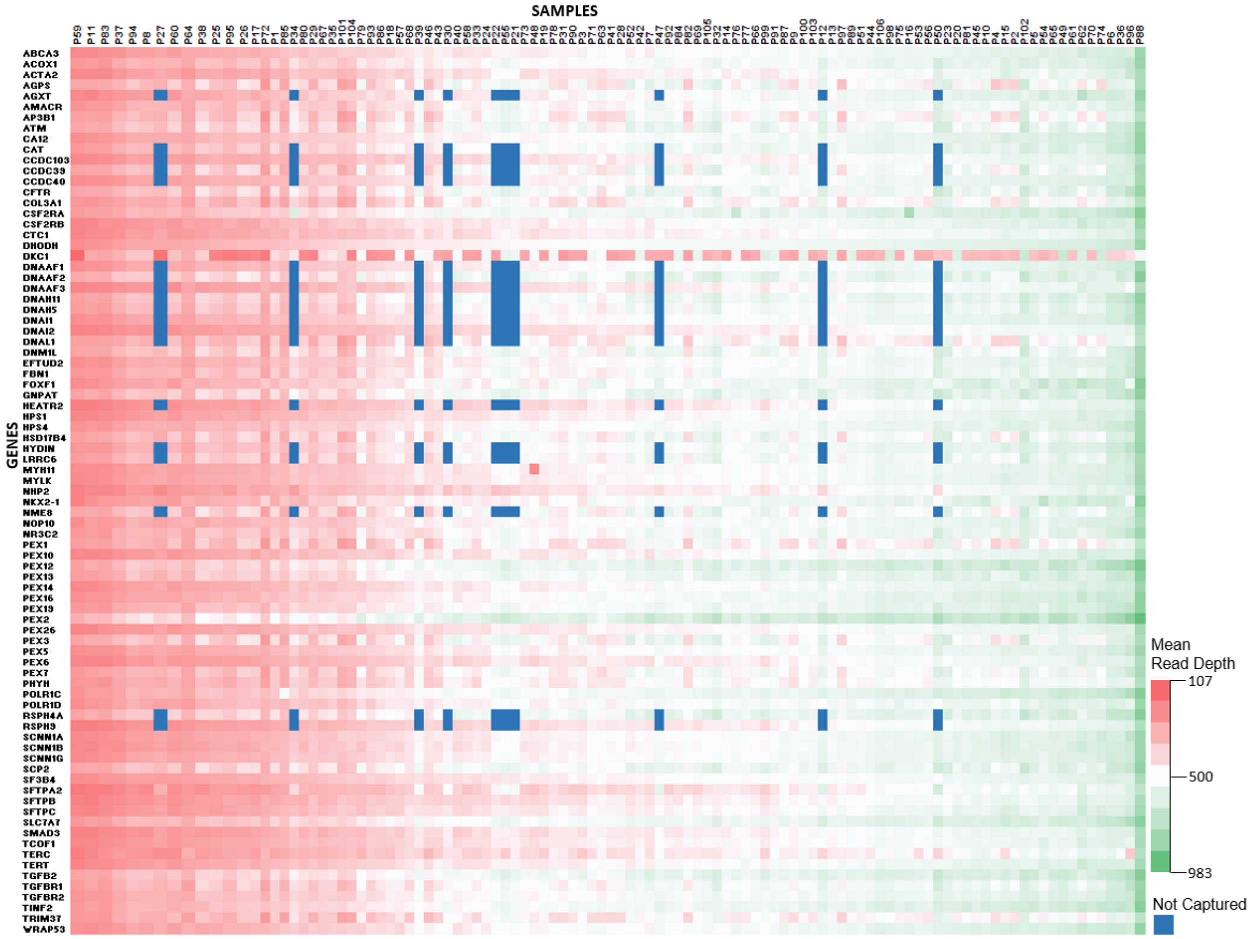
Zawistowski M, Reppell M, Wegmann D, St Jean PL, Ehm MG, Nelson MR, Novembre J, Zöllner S. 2014. Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. Eur J Hum Genet 22:1137–1144.



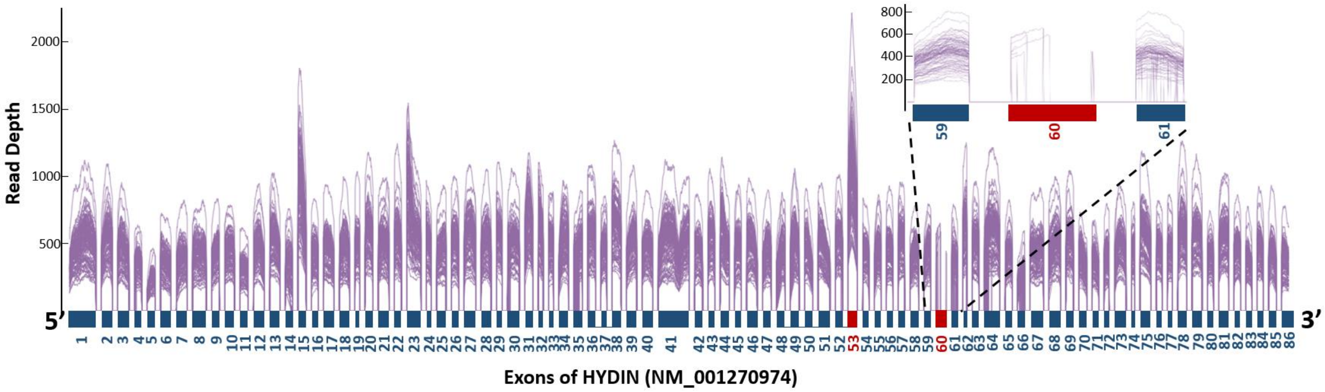
Supplementary Figure S1. Relationship between the fraction of methods that agree on a deleterious assignment for variants and the positive predictive value, PPV (fraction of predicted deleterious variants that are deleterious), for 10695 HGMD missense mutations and 10240 interspecies variants with available predictions for at least two out of the four methods (SNPs3D Profile, SIFT, Polyphen2 and CADD). By this measure, 77% of variants for which at least 3 of 4 methods predict deleterious are in fact deleterious.



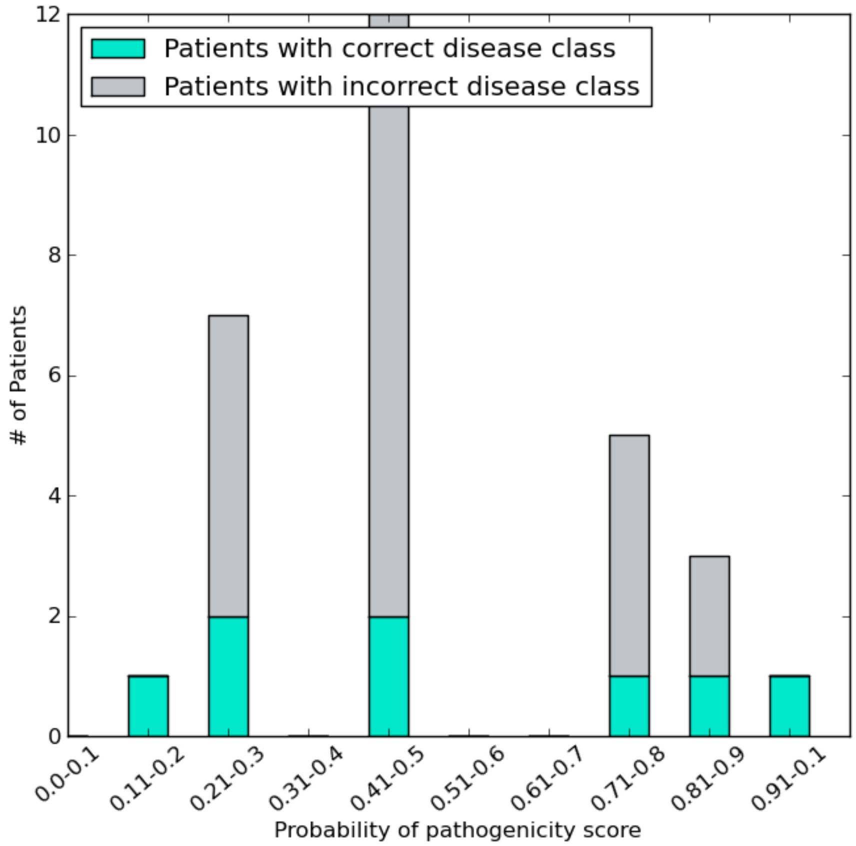
Supplementary Figure S2. Comparison of variant calling quality for 106 Hopkins samples versus 2,504 1000Genomes samples across the 83 genes in the panel. Only high-quality calls are included. HS: Hopkins Samples, KGS: 1000 Genomes samples, KGS_AFR: African samples in 1000Genomes, KGS_NonAFR: Non-African samples in 1000 Genomes. Circles represent HS sequenced using Capture v01 and triangles represent the HS sequenced using Capture v02. African samples are blue, Non-African are brown. Figure 1A shows the distribution of Transition vs. Transversion (Ts/Tv) and Heterozygous SNVs vs. Homozygous SNVs (HetALT/HomALT). By both measures, HS and KGS data are similar, except for the for HetALT/HomALT ratio of sample P8, an outlier with an excess of homozygous SNVs. Figure 1B shows the distribution of no call sites versus low-quality sites (not PASS in the gVCF file). Causative variants falling on any these sites will probably be missed. Figure 1C shows the distribution for common, rare and novel SNV types. For the common and rare SNV types, the HS and KGS distributions are similar. The lower counts of HS common SNVs for capture 2 reflects the fact that only 64 genes are included. There is a similar effect for rare variants, obscured by the crowding of points (medium 20 counts for capture v01 and 15 for v02). The rare variant distribution for both HS and KGS reflects the fact that Non-African samples have a lower rare variant load than the African samples. The novel variants load (between 0 and 8 variant with a median of 1 per sample) in HS is much lower than the rare variants. Figure 1D shows the distribution for common, rare and novel Indels. The distribution of common and rare Indels in HS is similar to the KGS distribution. Like SNVs, the lower counts of HS common indels for capture 2 is evident, in the plot. Two African samples (P2 and P83) are seen to carry more rare Indels and are outliers compared to the 1000Genomes dataset. For six sample there is a slightly higher number of rare than novel Indel in HS.



Supplementary Figure S3. Heat-map of average read depth for 83 genes across the 106 samples, color coded red (low depth: ~100) to green (high: ~950). Blue indicates the corresponding gene was not captured. Each column is for a different sample, and there are 83 rows, one for each gene. It is evident that coverage varies substantially across samples, from a low of 107X to a high of 983X.



Supplementary Figure S4. Exon-wise read depth for the HYDIN gene. Each purple line represents one sample and each rectangle represents one exon. The red rectangle indicates exons with anomalous coverage. The plot shows that Exon 53 has very high coverage and Exon-60 has very low coverage or no coverage for many samples compared to other exons in the gene. The inset shows a zoomed-in view of the read depth for Exon 59, Exon 60 and Exon 61.



Supplementary Figure S5. Distribution of correct and incorrect assignments of pathogenicity for patients based on missense mutations, as a function of the assigned probability of pathogenicity.

Supplementary Table S1. The nine regions with anomalous average read depth observed in more than 90 of the 106 samples. The eight "LOW COVERAGE" capture regions have low average read depth compared to other regions in the same gene. The one "HIGH COVERAGE" capture region has high read depth compared to other regions in the same gene. The "# of Samples" column is subdivided into coverage bins from no coverage to high coverage for "LOW COVERAGE" regions and from high to very high for "HIGH COVERAGE" regions. The HYDIN gene has two anomalous capture regions – 1. Low coverage, no reads in 78 samples and 2. A very high coverage of > 600X in 70 samples. The other anomalous regions are either deep intron, or present only in a minor isoform or actual coverage of the region is at least greater than or equal to 50X in all the samples.

	Gene	Capture Region	Genomic Region	Min. Mean Coverage in Samples	Max. Mean Coverage in Samples	# of Samples				
						No Coverage	1X – 20X	20X – 50X	50X – 100X	>100X
LOW COVERAGE	CCDC39	chr3:180369900-180370104	Exon (CDS)	60.27	381.93	-	-	-	8	88
	CSF2RA	chrX:1422103-1422305	Exon (CDS) in minor isoform	15.0	149.45	-	8	78	12	8
	DNM1L	chr12:32832247-32832449	First Exon of CDS	63.37	363.59	-	-	-	2	104
	CTC1	chr17:8151271-8151404	First Exon of CDS	68.18	314.30	-	-	-	7	99
	HYDIN	chr16:71021776-71022086	Exon (CDS)	0.0	262.83	78	3	5	6	4
	HSD17B4	chr5:118831376-118831558	Intron	0.0	172.08	9	1	3	64	29
	TERT	chr5:1294835-1295154	First Exon of CDS	55.67	270.49	-	-	-	8	98
	FBN1	chr15:48787269-48787507	Exon (CDS)	114.01	411.15	-	-	-	-	106

	Gene	Capture Region	Genomic Region	Min. Mean Coverage in *Samples	Max. Mean Coverage in *Samples	# of Samples				
						100X-300X	300X – 600X	600X – 900X	900X – 1300X	>1300X
HIGH COVERAGE	HYDIN	chr16:71007681-71007973	Exon (CDS)	357.29	1594.32	-	26	48	20	2

Supplementary Table S2. Five cases where two pairs of Indels in the CCDC40 gene were selected to satisfy a compound heterozygous model and leading to incorrect disease assignments. Each pair of Indels is very close to each other suggesting possible false variants arising from realignment errors or errors near repeat regions in the genome.

Patient ID	**Variant Position (chrom: position)**	**Reference Allele**	**Alternate Allele**	**Variant Type**
P31	chr17: 78064120	ACGCGCG	A	Deletion
	chr17: 78064128	AGGCACGTGCACGAACAAGGGACG	A	Deletion
P58	chr17: 78064144	A	ACC	Insertion
	chr17: 78064145	A	AC	Insertion
P61	chr17: 78064002	GC	G	Deletion
	chr17: 78064004	ACGTGCACGAAGAACACGGGACGCGCGCAGGCACGTGCACGAACAACACGGGACGCGCGCGGGC	A	Deletion
P91 & P66	chr17: 78063996	ACGCAGGCACGTGCACGAAGAACACGGGACGCG	A	Deletion
	chr17: 78064052:	CGGGACGCGCGCGGGCACGTGCACGAACAACACGGGACGCGCGCAGGCACGTGCACGAACAACACGGGACGCGCGCAGGCACGTGCACGAACAA	C	Deletion

Supplementary Table S3. Number of distinct variants that led to disease class prediction in 106 patients. 105 distinct potentially causative variants occurred only once in 78 patients. 14 potentially causative variants occurred twice or more in the remaining 28 patients. AD=Autosomal Dominant, HR=Homozygous Recessive and CH=Compound Heterozygous.

# of distinct variant that led to classification	**# of times seen in patients**	**Occurred as AD or HR**		**Occurred as part of CH pair**	
		# of patients with correct disease class	**# of patients with incorrect disease class**	**# of patients with correct disease class**	**# of patients with incorrect disease class**
105	1	17	29	14	18
11	2	2	12		6
1	3		3		
1	4	3		1	
1	6		4		2

Supplementary Table S4. Percentage of correct disease assignments in each of the three variant selection categories after removing HGMD from the method. Accuracy increases compared to the pipeline with HGMD. Overall trends remain the same - as expected, accuracy is highest in Category-1, then Category-2, then Category-3., and novel variant assignments are more accurate than for rare variants.

Category	**Variant Considered**	**Minor Allele Frequency**			**% Correct Assignment**
		Novel	<= 0.005	<=0.01	
Category-1	In ClinVar with Pathogenic or Likely pathogenic tag	1/1	4/7	0/1	5/9: 55%
Category-2	Missense (Predicted damaging either by SNPs3D, SIFT, PolyPhen2 or CADD) Frameshift / Non-Frameshift Indel NonSense Direct Splicing Any variant predicted damaging by dbscSNVs	11/16	13/39	3/7	27/62: 43%
Category-3	All other missense, UTR, and Intronic	5/18	2/13	1/2	8/33: 24%
		17/35: 49%	19/59: 32%	4/10: 40%	