Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation

Sami S. Amr, PhD[1,2], Saeed H. Al Turki, PhD[1], Matthew Lebo, PhD[1,2], Mahdi Sarmady, PhD[3], Heidi L. Rehm, PhD[1,2] and Ahmad N. Abou Tayoun, PhD[3]

Purpose: Classification of novel variants is a major challenge facing the widespread adoption of comprehensive clinical genomic sequencing and the field of personalized medicine in general. This is largely because most novel variants do not have functional, genetic, or population data to support their clinical classification.

Methods: To improve variant interpretation, we leveraged the Exome Aggregation Consortium (ExAC) data set (N = ~60,000) as well as 7,000 clinically curated variants in 132 genes identified in more than 11,000 probands clinically tested for cardiomyopathies, rasopathies, hearing loss, or connective tissue disorders to perform a systematic evaluation of domain level disease associations.

Results: We statistically identify regions that are most sensitive to functional variation in the general population and also most commonly impacted in symptomatic individuals. Our data show that a significant number of exons and domains in genes strongly associated with disease can be defined as disease-sensitive or disease-tolerant, leading to potential reclassification of at least 26% (450 out of 1,742) of variants of uncertain clinical significance in the 132 genes.

Conclusion: This approach leverages domain functional annotation and associated disease in each gene to prioritize candidate disease variants, increasing the sensitivity and specificity of novel variant assessment within these genes.

Genet Med advance online publication 22 September 2016

Key Words: disease burden; domains; intragenic; intolerance; variant interpretation

INTRODUCTION

Genomic sequencing in the form of large gene panels, whole-genome sequencing or whole-exome sequencing has entered into genetics clinics with thousands or even tens of thousands of such tests performed to date, significantly increasing the diagnostic yields in several clinical scenarios.[1–4] The promise of genomic medicine has long been acknowledged to be at the center of precision medicine by the scientific community and has just recently gained national recognition.[5] With such comprehensive testing, however, come several challenges, the key one of which is the interpretation of the large number of variants—estimated to be 3–5 million per human genome—generated in the process, with thousands of which having uncertain clinical significance. Even more alarming is the fact that the majority of variants identified are very rare, having been seen in only one family.[6] Characterizing the functional impact of all variants identified requires tremendous resources that are beyond the capacity of any single molecular diagnostic laboratory.

Recognizing this challenge, the National Institutes of Health has recently funded three different groups under the Clinical Genome Resource (ClinGen) Program, whose main goal is to build and maintain a publicly accessible genomic knowledge base to promote variant data sharing among clinical laboratories, researchers, and clinicians.[6] Another major goal of this program is dissemination of standards and guidelines for variant interpretation and gene–disease association.[7,8] In a similar approach for gene–disease associations, we have recently shown that a systematic evaluation of genes associated with hearing loss can largely eliminate unnecessary interpretation of variants in genes with weak disease associations.[9] Still, a large number of novel variants of uncertain clinical significance are routinely identified in genes with strong disease associations, necessitating novel approaches to prioritize candidate disease-causing variants.

The use of population sequencing data sets such as the Exome Aggregation Consortium (ExAC) has been useful in genome-wide assessment of the tolerance of genes to variation and has shown a greater degree of intolerance for genes that are associated with Mendelian disorders.[10–13] A similar strategy at a gene level can potentially identify tolerant or sensitive intragenic regions for a particular class of variants for any given disease gene.[14] The tolerance predictions of these regions to different classes of variation can be further enhanced through the use of clinically interpreted variant data sets from diagnostic laboratories. Here, we used variant data sets from population databases and from a clinical laboratory to statistically define domain-level disease associations and to assess exon-level tolerance of loss-of-function variants across 132 genes included in gene

[1]Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, Massachusetts, USA; [2]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA; [3]Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA. Correspondence: Heidi L. Rehm (hrehm@partners.org) or Ahmad N. Abou Tayoun (aboutayoua@chop.edu)

Submitted 13 May 2016; accepted 28 July 2016; advance online publication 22 September 2016. doi:10.1038/gim.2016.134

panels offered at our laboratory. We evaluated the utility of this approach by examining the impact of regions defined through our analysis of classification of variants previously determined to be of uncertain clinical significance. In addition, we show a significant role for our approach in refining the variant interpretation process.

MATERIALS AND METHODS

Clinically curated variants

A total of 186,009 variant observations representing 6,978 unique variants (**Supplementary Table S1** online) in 132 genes (**Supplementary Table S2** online) were identified in a cohort of 11,219 probands tested at the Laboratory for Molecular Medicine between 2005 and 2015. This cohort consisted of individuals affected with cardiomyopathies ($n = 5,466$), rasopathies ($n = 3,022$), hearing loss ($n = 1,990$), and connective tissue disorders ($n = 741$).

All 6,978 unique variants were manually assessed for clinical significance and each was classified into one of five categories based on its potential impact and role in causing the patient's disease: pathogenic (P), likely pathogenic (LP), variant of uncertain significance (VUS), likely benign (LB), and benign (B). Details of the variant classification workflow at our laboratory have been described.[15] The third category, VUS, is further refined into three subcategories (VUS–favor benign (FB), VUS, and VUS–favor pathogenic (FP)) based on whether evidence favors a benign or pathogenic classification but does not meet thresholds or criteria for a likely pathogenic or likely benign classification. VUS-FP leans toward LP and VUS-FB leans toward LB, whereas the VUS subcategory denotes a lack of evidence to favor a causative or neutral role or equally conflicting evidence.

Protein domain burden analysis

The numbers of probands positive or negative for each of the 6,978 clinically curated variants were determined using our internal database. For controls, we queried the Exome Aggregation Consortium (release 0.3)[12] database (60,706 whole exome samples) for all high-quality (PASS) variants and their allele frequencies in these 132 genes. Then, the numbers of individuals positive or negative for each of the rare variants were determined. To minimize any potential annotation discrepancies, we reannotated all variants in cases and controls using SnpEff v4.0e[16] with a gene model from RefSeq.[17] Annotations include exon, transcript, gene, amino acid change, and the variant's impact on the protein.

Based on their functional impact, reannotated case and control variants were then grouped into "functional" variants (mostly missense and in-frame insertions or deletions), "loss of function" variants, or both (lof_functional) (**Supplementary Table S3** online, http://snpeff.sourceforge.net/SnpEff_manual.html). Domain burden was analyzed by first calculating, within each variant group, the total number of rare variants—defined as variants with allele frequency <1% in cases and controls—at each protein domain ($N = 645$) based on boundaries from the Pfam database[18] (domain boundaries were predicted using hidden Markov models, downloaded from UCSC tables, last updated 1 July 2013). Coding regions that do not overlap with any Pfam domain were labeled as "outside domains" for each transcript. We then generated a 2×2 table at each protein domain/"outside domain" region for the number of positives and negatives in cases and controls and assessed significance using Fisher's exact test. All P values were adjusted by Bonferroni correction based on the total number of domains (the significance level was $<1.52 \times 10^{-5}$). All significant, tolerant, and intolerant regions are listed in **Table 1** and **Supplementary Table S4** online.

As a quality check, significantly "intolerant" regions were confirmed to have adequate coverage in ExAC (**Supplementary Table S4** online), thus ruling out any potential false-positive results due to the lack of coverage in the general population. Furthermore, all identified regions were validated for enrichment of clinically significant variants in our internal database, HGMD, and ClinVar.

Loss-of-function variant analysis per RefSeq transcripts

After reannotating variants from the Exome Aggregation Consortium, we investigated all RefSeq transcripts in the 132 genes (361 transcripts and 2,825 unique exons) to identify exons with high-allele frequency (MAF ≥0.1%) loss-of-function (LoF) variant(s) and/or with multiple such variants. The loss-of-function effect of each variant was predicted by SnpEff v4.0e (see details under "Protein Domain Burden Analysis") and included mainly stop-gains, splice sites (± 1 or 2), frameshifts, and start-losses.

Exons with high-allele frequency LoF variants were interpreted in the context of disease prevalence, mode of inheritance, potential annotation errors, and/or gene structure (see "Results").

RESULTS

Intragenic disease burden

A significant number of variants in *known* genes are continually identified that have uncertain clinical significance due to absence of supporting functional and/or statistical genetic evidence. For example, the clinical significance of 35% of the variants (2,413/6,978) identified in 132 genes analyzed in this study could not be determined (see below) despite the fact that most of those genes have well-established associations with cardiomyopathies, rasopathies, connective tissue disorders, or hearing loss. In such genes, quantifying *intragenic* intolerance to variation can potentially support variant prioritization and interpretation.

We made the assumption that intragenic regions depleted of functional variation in the general population but enriched for variants in individuals with a disease are most likely to be clinically relevant relative to other regions. To quantify such intragenic region–disease associations, variant counts (<1% MAF) in cases versus controls were calculated at each of the annotated Pfam domains in the 132 genes ("Materials and Methods"). The number of cases, genes, domains, and clinically curated

Table 1 Domains or intragenic regions with significant enrichment of variants in cases or controls

Gene	Inheritance	Domain	Effect	Disease	$P_{\text{Intolerance}}$[a]	$P_{\text{Tolerance}}$[b]
PTPN11	AD	Y_Phosphatase	Functional	Rasopathies	6.05E-162	
		SH2_4147			5.70E-175	
		SH2_1434			1.03E-12	
		Outside domains			6.64E-11	
BRAF	AD	C1_1			5.2E-55	
		Pkinase_Tyr			9.3E-39	
KRAS	AD	Ras			9.8E-17	
SOS1	AD	Outside domains				2.06E-07
HRAS	AD	Ras_955			3.44E-08	
FBN1	AD	Outside domains	LoF	Marfan syndrome	3.22E-30	
		EGF	Functional		1.17E-09	
MYH7	AD	Myosin_head	Functional	Cardiomyopathy	9.36E-100	
		IQ			1.09E-10	
		Outside domains			1.26E-34	
		Myosin_tail_1				1.66E-03
MYBPC3	AD	I-set_896	LoF		6.11E-40	
		fn3_371			2.83E-25	
		Outside domains			1.71E-20	
		I-set_620			5.4E-13	
		fn3_478			4.2E-11	
		fn3_928			7.67E-09	
		ig_1553			9.5E-09	
		I-set_304	Functional		2.78E-28	
		I-set_896			1.01E-09	
		Outside domains			5.02E-15	
RYR2	AD	SPRY_6760	Functional			1.84E-10
		SPRY_23214				7.2E-06
		RYDR_ITPR_11082				9.34E-06
		Outside domains				1.67E-07
DSP	AD	Outside domains	LoF		3.44E-16	
			Functional			3.43E-14
PKP2	AD	Arm_18076	LoF		7.62E-09	
		Outside domains			2.6E-20	
LMNA	AD	Filament_5591	Functional		5.4E-12	
		Filament_21209	LoF		2.02E-08	
NEBL	AD	LIM	Functional			3.73E-12
MYH6	AD	Myosin_tail_1	Functional			5.00E-06
TPM1	AD	Tropomoysin_20090	Functional		8.97-E07	
TNNT2	AD	Troponin_3028	Functional			7.04E-42
		Outside domains				2.46E-19
LAMA4	AD	Laminin_G_1_5762	Functional			1.25E-11
		Laminin_I_22365				3.64E-11
TMPO	AD	Outside domains	Functional			1.21E-23
TTR	AD	Transthyretin	Functional		2.82E-07	
SGCD	AD	Sarcoglycan_1_413169	Functional			3.6E-12
DES	AD	Filament_4997	Functional			1.32E-20
LDB3	AD	LIM_8042	Functional			5.05E-12
CASQ2	AD	Outside domains	LoF_Functional			8.25E-06

Additional information can be found in **Supplementary Table S4** online.

[a]$P_{\text{Intolerance}}$ = *P* value for enrichment of variants in cases relative to controls. [b]$P_{\text{Tolerance}}$ = *P* value for enrichment of variants in controls relative to cases.

Table 1 Continued on next page

Table 1 Continued

Gene	Inheritance	Domain	Effect	Disease	$P_{\text{Intolerance}}$[a]	$P_{\text{Tolerance}}$[b]
MYO7A	AD/AR	Outside domains	LoF	Hearing loss	5.64E-08	
		MyTH4_6052	Functional		6.72E-07	
CDH23	AR	Cadherin_36780	Functional			1.07E-09
		Outside domains				2.02E-05
DFNB31	AR	Outside domains	Functional			4.19E-06
GPR98	AR	Outside domains	Functional			3.54E-37
OTOF	AR	Outside domains	LoF			2.24E-05
SLC26A4	AR	Sulfate_transp_21084	Functional		1.97E-12	
PCDH15	AR	Cadherin_37570	LoF_Functional			3.11E-06
		Outside domains				3.9E-12
MYO15A	AR	Outside domains	Functional	Hearing loss		2.29E-10
MYO3A	AR	Outside domains	Functional			5.61E-07
TRIOBP	AR	Outside domains	Functional			1.47E-11

Additional information can be found in **Supplementary Table S4** online.

[a]$P_{\text{Intolerance}}$ = *P* value for enrichment of variants in cases relative to controls. [b]$P_{\text{Tolerance}}$ = *P* value for enrichment of variants in controls relative to cases.

variants per disease are shown in **Supplementary Table S1** online. The ExAC database (*N* = 60,706 individuals) was used as a control data set. We statistically determined whether any domain(s) had a significant enrichment of variants in cases versus controls. Within a total of 34 genes, we identified 33 regions with a significant disease burden in addition to 25 regions that were relatively tolerant to variation in the general population ($P < 1.52\times10^{-5}$, **Table 1** and **Supplementary Table S4** online). The remaining 98 genes did not exhibit any significant enrichment, most likely owing to lack of sufficient, especially clinical, variant data (see "Discussion").

Validation of the intragenic burden analysis

Because all manually curated variants with MAF <1% were included regardless of their clinical classification, we sought to determine whether constrained or intolerant regions were enriched for pathogenic or likely pathogenic variants and vice versa. A total of 6,978 unique variants with clinical significance ranging from benign to pathogenic were used in the analysis (**Supplementary Figure S1A** online; see also "Materials and Methods" for more information on variant classification). The distribution of all classified variants across the four diseases and the 132 genes is shown in **Supplementary Figure S1B** and **Supplementary Table S2** online, respectively. As expected, there was a significant enrichment of P and LP variants relative to those classified as VUS-FB in all disease-constrained intragenic regions ($P < 0.001$, **Supplementary Table S5** online). However, no such enrichment was observed in most (26/29) tolerant regions (**Supplementary Table S6** online). Instead, the latter regions contained, on average, threefold more VUS-FB variants ($P < 0.006$) but 13-fold fewer P and LP variants relative to the disease-intolerant regions ($P < 0.001$, **Figure 1a**).

To further validate our findings, we determined the distribution of disease-causing variants from ClinVar and HGMD across the regions identified (**Supplementary Information** and **Supplementary Table S4** online). We detected a fivefold enrichment of ClinVar P/LP to ClinVar B/LB variants in the intolerant regions ($P = 2\times10^{-4}$, **Figure 1b**). Again, no such enrichment was observed in most tolerant regions. Rather, similar to our laboratory-curated variants, the latter regions contained, on average, 2.6-fold more ClinVar B/LB variants ($P = 0.0145$) and 3- and 2.3-fold fewer ClinVar P/LP ($P = 0.011$) and HGMD DM ($P = 0.017$) variants, respectively, relative to the disease-constrained regions (**Figure 1b**). These findings highlight the utility of our approach and show that the identified intragenic regions can predict the distribution of disease-causing and benign variants across the relevant genes.

Examples of constrained domains

As might be expected, most of the constrained regions were within the 65 genes that are primarily associated with autosomal dominant disorders due to gain-of-function pathogenic variants, including cardiomyopathies, rasopathies, and connective tissue disorders. A total of 25 of these 65 (~40%) genes had constrained regions, with some genes shown to be generally intolerant to variation (*PTPN11*, *BRAF*, *SOS1*, *HRAS*, *FBN1*, *MYH7*, *MYBPC3*, *PKP2*, and *LMNA*) and others having certain intragenic regions that were significantly more or less constrained; this observation could be used to support variant prioritization in the relevant gene. One example is the *MYH7* gene, for which most pathogenic variation for cardiomyopathy is due to missense variants. Despite its overall intolerance to variation, the myosin head domain in this gene is most disease-burdened relative to other domains, such as the myosin tail, which appears to be slightly tolerant to variation (**Figure 2**). As expected, the myosin head domain contained 186 P/LP variants but only 9 VUS-FB variants. By contrast, there were only 3 and 14 VUS-FB and P/LP variants, respectively, in the tail domain (**Supplementary Tables S5 and S6** online). Another example is *PTPN11*, for which the pathogenic variation for rasopathies is due to missense variants. *PTPN11* is generally intolerant to variation, but, as shown in **Supplementary Figure S2** online, the SH2 and phosphatase domains in this gene are extremely constrained, such that novel variants identified in those domains are more likely to be clinically significant relative to others regions of the gene. In fact, 88% (376/428) of the P and LP variants identified in this gene were restricted to those two domains (**Supplementary Table S5** online).

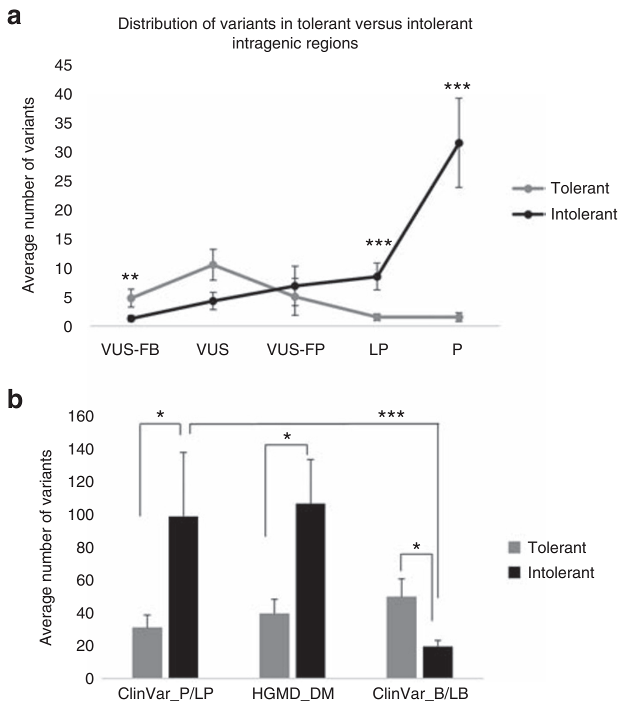


Figure 1 Intolerant intragenic regions show enrichment of clinically significant variants relative to tolerant regions that are depleted for such variants but enriched for benign or likely benign variants. (**a**) The distribution of classified variants in disease-constrained (black) versus disease-tolerant (grey) regions. Average numbers ± standard error are shown. ****P* < 0.001; ***P* < 0.006 (Mann-Whitney nonparametric two-sample test). (**b**) The distribution of ClinVar P/LP, B/LB, and HGMD DM variants in disease-constrained (red) versus disease-tolerant (green) regions. Average numbers ± standard error are shown. ****P* < 0.001; **P* = 0.01 (Mann-Whitney nonparametric two-sample test).

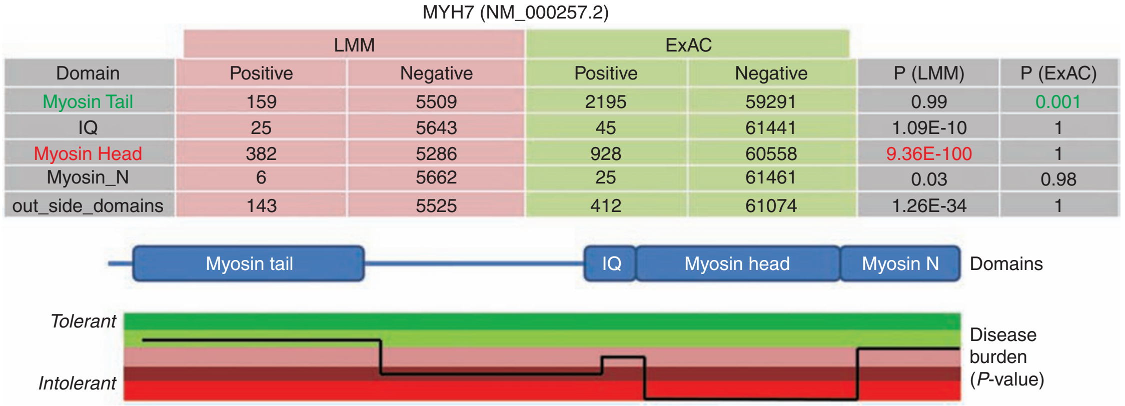
Unlike autosomal dominant disorders, the 67 autosomal recessive hearing-loss genes rarely had regions that were constrained. Only two genes had such regions, although one of them can also cause disease in an autosomal dominant fashion (*MYO7A*). By contrast, there were nine genes with regions that were highly tolerant to variation in the general population (**Table 1**). One interesting example is *GPR98*, which has been shown to be an extremely tolerant gene.[10,12] Our analysis, however, shows that this tolerance is significant only in the regions that are "outside domains" (**Table 1**), whereas other regions of the gene did not appear to be significantly tolerant (data not shown).

Impact of burden analysis

To determine the impact of this analysis on variant classification, we investigated the extent to which the identified statistically significant regions can support classification of variants that lack any other evidence, so-called VUS variants in our classification system (see "Materials and Methods"). Identification of such variants in the tolerant or disease-constrained regions is very likely to support reclassifying them at least to VUS-favor benign or VUS-favor pathogenic, respectively. Of the total 1,742 VUS variants identified in the 132 genes, 450 (or 26%) resided in these regions and can thus be considered for reclassification (**Supplementary Figure S3** and **Supplementary Tables S5 and S6** online). Based on this, we anticipate that our intragenic constraint bins will enable variant prioritization in the relevant genes.

Exons with high burden of loss-of-function variants

Although most clinical laboratories tend to include the longest transcripts in their assays because of limited available information on the critical exons required for gene function, detailed analysis of LoF variants and transcript structure can be helpful to further refine the clinical relevance of intragenic disease regions. We used allele frequencies from the general population[12] to identify exons harboring high-allele frequency



MYH7 (NM_000257.2)

	LMM		ExAC			
Domain	Positive	Negative	Positive	Negative	P (LMM)	P (ExAC)
Myosin Tail	159	5509	2195	59291	0.99	0.001
IQ	25	5643	45	61441	1.09E-10	1
Myosin Head	382	5286	928	60558	9.36E-100	1
Myosin_N	6	5662	25	61461	0.03	0.98
out_side_domains	143	5525	412	61074	1.26E-34	1

Figure 2 A heat map showing the distribution of variation in cases versus controls across the *MYH7* gene. The table shows the raw data, i.e., the numbers of probands and controls that are positive or negative for variants in certain domains or regions within the gene. LMM refers to the laboratory for molecular medicine. P(LMM) refers to the *P*-value for statistical significance in cases relative to controls. P(ExAC) is the *P*-value for statistically significant enrichment in controls.

Table 2 Exons with high-allele frequency loss-of-function variants in the general population

Gene	Transcript	Ex	Variant (cDNA; protein change)	Al. count	MAF[a]	Eth.	Clin.Sig.Var	Comment
ABCC9	NM_020297.2	18	c.2238-1G>A	118	0.7%[b]	SA	None	
ATP6V1B1	NM_001692.3	1	c.2T>C; p.M1?	49581	41%[b]	All	None	AM
CABP2	NM_016366.2	6–7	c.637+1G>T	32	0.5%[c]	Fin.	NA	Last ex.
CCDC50	NM_178335.2	6	c.827C>G; p.S276*	19	0.1%[d]	SA	None	AS
COCH	NM_004086.2	2	c.-23-1G>T	77	0.5%[b]	SA	None	AS
DFNA5	NM_004403.2	2	c.120insA; p.R42Efs*112	137	0.2%[b]	NF	None	AS/WDM
		6	c.712C>T; p.R238*	27	0.4%[d]	Fin.	None	WDM
DSC2	NM_024422.3	16	c.2688_2689insTC; p.Ala897Lysfs∗4	816	1.2%[b]	NF	None	AS/Last ex.
EDN3	NM_000114.2	4	c.565G>GA; p.T189Nfs*10	36	0.5%[b]	Fin.	NA	AS
				29	0.3%	Afr.		
				164	0.2%	NF		
EDNRB	NM_001201397.1	1	c.169G>T; p.G57*	24	0.2%	SA	NA	AS
LAMA4	NM_001105209.2	2	c.326C>G; p.Ser109*	65	0.6%[d]	Afr.	None	AS
LOXHD1	NM_144612.6	1	c.2T>A; p.M1?	929	4.5%	All	None	AS/AM
LRTOMT	NM_001145307.4	6	c.438-2A>C	41	0.5%	NF	None	AS
MYO15A	NM_016239.3	2	c.3524dup; p.S1176Vfs*14	22	0.26%[d]	EA	None	AS
		26	c.5925G>A; p.W1975*	138	1.7%[b]	SA	None	INF (18aa)
OTOF	NM_194248.2	32	c.4023+1G>A	95	1.0%	EA	None	INF (43aa)
PAX3	NM_000438.5	4	c.638C>A; p.S213*	90	0.6%[b]	SA	NA	AS
	NM_181460.3	8	c.1204+1G>C	22	0.1%	SA	NA	AS
PCDH15	NM_033056.3	33	c.4714_4715insAACA; p.T1572Kfs∗11	99	1.0%	Afr.	None	AS
P2RX2	NM_170682.2	4	c.426C>A; p.C142*	41	0.9%	Fin.	NA	AS
		11	c.1325_1335del; p.Ser442*	349	3.4%	Afr.	NA	AS/Last ex.
SLC26A4	NM_000441.1	8	c.919-2A>G	38	0.4%	EA	None	
TMC1	NM_138691.2	17	c.1534C>T; p.R512*	12	0.2%	Fin.	None	INF (54aa)
TRIOBP	NM_007032.5	1	c.124C>T; p.R42*	46	2.5%[b]	SA	None	AS
USH1C	NM_153676.3	19	c.2014-1G>A	61	0.6%	Afr.	None	AS
		20	c.2167C>T; p.Q723*	26	0.3%	Afr.	None	AS

Exons with reduced or absent expression in GTEx are underscored in the "Comment" column.

Afr., African; Al. count, allele count; AM, alternative "methionine"; AS, alternatively spliced; Clin.Sig.Var., clinically significant variant in that exon; EA, East Asian; Eth., ethnicity; Ex, exon; INF, in-frame exon (included in parentheses is the number of amino acids, aa, encoded by this exon); Fin., Finnish; Last ex., last exon; MAF, minor allele frequency; NA, not applicable, no clinical data available; NF, non-Finnish; SA, South Asian; WDM, wrong disease mechanism.

[a]Frequencies are from the highest subpopulation, which is noted in the ethnicity column. [b]≥1 homozygote(s) in ExAC. [c]General MAF (all populations) is 0.1%. [d]Several other loss-of-function variants in this exon.

LoF variants in the 132 genes studied. Such exons could cause disease but with reduced penetrance, or they could be clinically benign, potentially due to alternative splicing leaving out exons not critical to gene function. For routine interpretation, the laboratory uses the relevant disease prevalence, estimated penetrance, and genetic heterogeneity to calculate the minor allele frequency above which variants in the associated genes can be considered likely benign, with higher frequencies used for a benign classification. For hearing loss, 0.1% and 0.3% likely benign cutoffs were used for autosomal dominant and autosomal recessive hearing loss, respectively. For the remaining diseases, 0.3% allele frequency was conservatively used.

Our first-tier analysis identified a large number of LoF variants in the genes of interest. However, we performed manual curation to exclude very-low-quality variants, especially frameshifts, and "apparent" nonsense variants, whereby adjacent single-nucleotide variants, likely to be in *cis*, changed their interpretation; these are the so-called multinucleotide polymorphisms (MNPs).[12] After this filtration, we identified 26 exons from 21 genes, each harboring at least one LoF variant exceeding the aforementioned allele frequency cutoffs (**Table 2**). Some of those exons also had several rare LoF variants in ExAC. Most exons (18 or 70%) appeared to be alternatively spliced and therefore not expressed by all transcripts, probably explaining why they are found at a frequency too high for the disorder. Three exons were small in-frame exons; basal exon skipping[19] or nonsense-induced alternative splicing[20] might rescue the expected protein loss of function due to such variants. Two exons—one of which was also alternatively spliced—had high-allele frequency start-loss variants with nearby secondary methionine suggesting either annotation errors or start reinitiation. Three "LoF" variants were in the last exons of *CABP2*, *DSC2*, and *P2RX2* and thus are not expected to be true LoF variants due to escape from nonsense-mediated decay (NMD) and limited protein impact. The last exons in *DSC2* and *P2RX2* were also alternatively spliced (**Table 2**).

Finally, three exons did not fit any of these categories. One was in *DFNA5*, wherein only variants leading to exon 8 skipping have been shown to cause autosomal dominant hearing loss through a potential gain of function mechanism.[21,22] In fact, a frameshift variant in exon 5 of this gene failed to segregate with disease in an Iranian family with hearing loss,[23] confirming that loss of function is not a disease mechanism. The remaining two LoF variants are splice acceptor variants in *ABCC9* and *SLC26A4* found in 0.7% South Asian alleles (with one homozygous individual) and 0.4% East Asian alleles, respectively (**Table 2**). Whether these two variants are true, potentially founder, LoF mutations in the Asian subpopulation has yet to be determined. Nevertheless, none of the 26 exons contained pathogenic variants in our patient cohorts, further suggesting they are not required for gene function.

Examples of exons with high-allele frequency LoF variants

One example is the *PCDH15* gene, wherein pathogenic variants have been strongly associated with Usher syndrome type 1 (refs. 24,25). One LoF variant with 1% allele frequency is found in exon 33 of the NM_033056 transcript (**Table 2**). Additionally, 37 other LoF variants in ExAC are also found within this exon, which is alternatively spliced from other transcripts, suggesting that it is unlikely to harbor any disease-causing variants. In fact, this finding has also been confirmed in a recent study.[26]

Another example is the *USH1C* gene, which has also been associated with Usher syndrome type 1.[27] This gene contains 28 exons and produces several transcripts that share exons 1–14 and 16–21 encoding for three PDZ domains and one coiled-coil domain. A longer isoform with an additional seven exons (NM_ 153676.3: Exons 15A-20A, 26A) encoding a second coiled-coil domain and a PST (proline, serine, threonine-rich) domain has also been described.[28] Two LoF variants, a splice site (0.6%) in exon 19A and a nonsense variant (0.3%) in exon 20A, were identified by ExAC in the African subpopulation (**Figure 3a**). This allele frequency is relatively high given the Usher type I disease prevalence, suggesting that those exons are very unlikely to be involved in this disease. To examine this, we assessed the expression of *USH1C* exons 19A and 20A in different human tissues using the recently published Genotype-Tissue Expression (GTEx) Project in which multitissue RNA

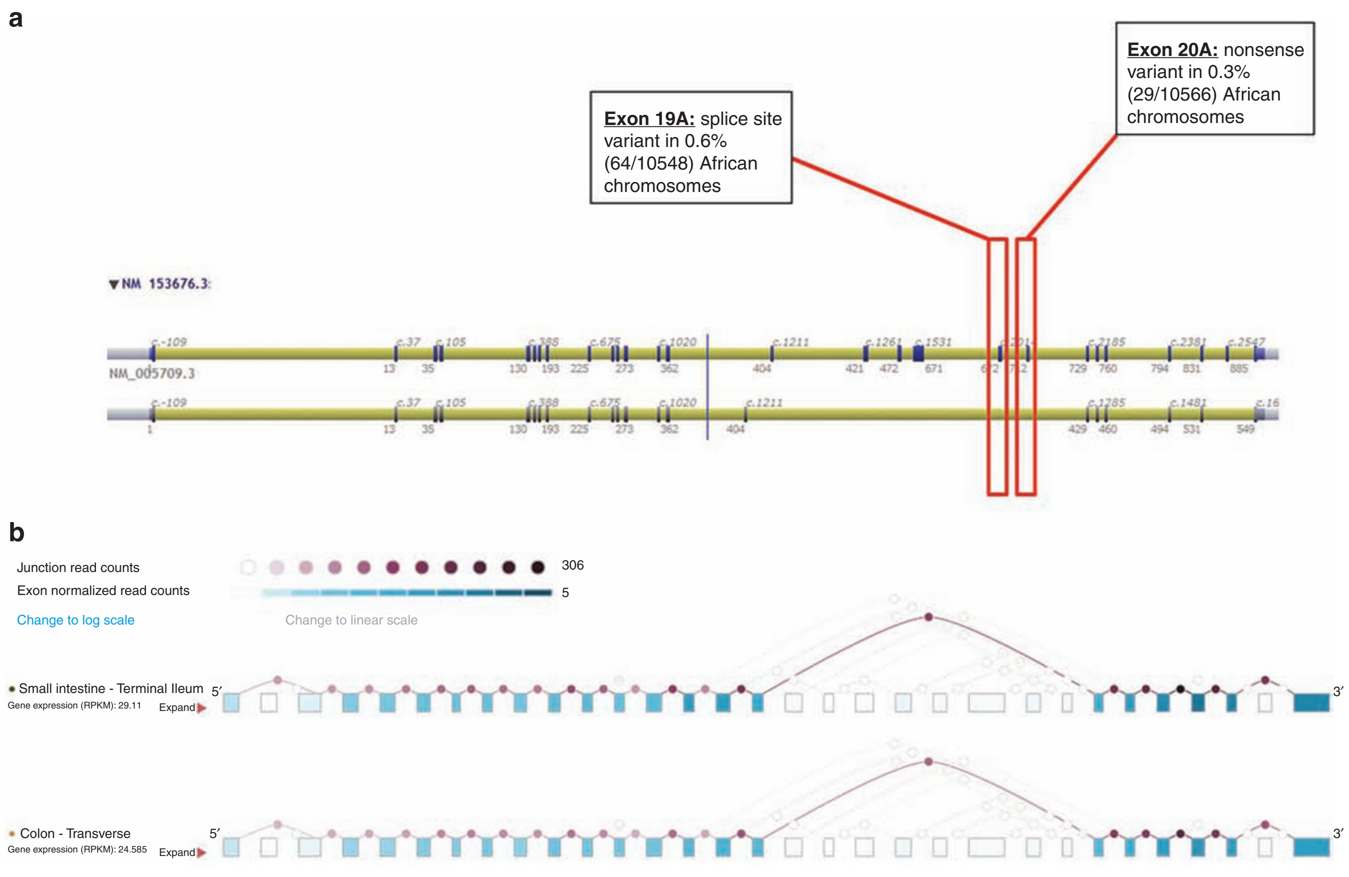


Figure 3 High-allele LoF variants in two alternatively spliced exons (19A and 20A) in the *USH1C* gene. (**a**) A schematic adapted from Alamut Visual Software showing two *USH1C* transcripts and highlighting exons 19A and 20A and the LoF variants in those two exons. (**b**) This schematic is adapted from the GTEx Portal website showing the exon-level expression of *USH1C* in two different tissues (only the two tissues with the highest expression are shown). These data were generated using RNA sequencing and clearly show the lack of expression of several exons, including 19A and 20A, as signified by exon skipping. Note that this trend holds for all other tissues where *USH1C* is present at lower levels.

sequencing was performed for 173 individuals.[29] Gene- and exon-level expression patterns were made publicly accessible through the GTEx portal (http://www.gtexportal.org/home/documentationPage). Interestingly, exons 19A and 20A and a few others showed very low expression, if at all, in the tissues where *USH1C* was found to be expressed. It is thus strongly predicted that these exons are spliced out of the primary isoform of the protein in most, if not all, tissues (**Figure 3b**). Because cochlear and retinal tissues are not represented in the GTEx database, we cannot exclude the possibility that this isoform is expressed only in these tissues involved in Usher syndrome; however, given the high-frequency LoF variants identified in the ExAC database but not found in our disease population, the most likely conclusion is that these exons are also not required in retina and cochlea.

We also found a similar low expression pattern in GTEX for several exons in **Table 2**, including *CCDC50* exon 6, *COCH* exon 2, *EDN3* exon 4, *LAMA4* exon 2, *LOXHD1* exon 1, *MYO15A* exons 2 and 26, and *PAX3* exon 4. In addition, 5 of the 26 exons (*MYO15A* exons 2 and 26, *OTOF* exon 32, *PCDH15* exon 33, and *TRIOBP* exon 1) overlapped with the tolerant regions identified through the burden analysis described (see also **Supplementary Table S6** online). We did not expect high concordance because the latter approach is extremely dependent on clinically curated variants that are lacking for several genes and/or regions (see "Discussion") and uses domain boundaries that often consist of multiple exons whose tolerance is averaged into an overall domain tolerance.

In summary, although tolerance to 23 of the 26 high-allele frequency LoF variants is probably explained by gene structure (alternative splicing, in-frame exons, start reinitiation, NMD escape) and one due to a distinct disease mechanism (*DFNA5*), we cannot exclude for any of these variants genuine loss-of-function effect but with reduced penetrance and/or variable expressivity. Nevertheless, such information is extremely important to use during clinical variant interpretation. Extra caution should be exercised in interpreting variants in these exons given that all variations may be benign, as supported by the fact that all 26 exons were devoid of pathogenic variants in our patient population.

DISCUSSION

Because it is impossible to have sufficient evidence to support the classification of all potential variants in disease genes, new approaches are needed to better inform the variant interpretation process. In this study, we leveraged information about the frequency of variants in cases and controls to statistically bin intragenic regions of disease genes into those with higher or lower tolerance to variation. We showed that such regions exhibit the expected enrichment of pathogenic or benign variants based on regional overall tolerance. We also showed that this approach provides additional information to support the classification of variants with limited or no evidence and to prioritize variants in the relevant genes. Additionally, we highlighted exons that harbor LoF variants at frequencies in the general population that exceed what would be expected based on disease prevalence, suggesting that these exons might not be expressed and/or disease-relevant. In fact, most such exons seem to be alternatively spliced and were devoid of pathogenic variants, supporting the premise that LoF variants with high frequencies in the general population, if supported with strong analytical data, can guide transcript annotation to aid in interpretation.

It should be noted that additional expert interpreted variants, along with deeper allele frequency information from variants in large general populations, are always needed to gain more resolution across intragenic regions. This is especially important for large domains where unequal tolerance to variation can average significant signals. Furthermore, because most diseases can be subdivided into different forms based on phenotypic variation, more clinical data will be needed to stratify based on phenotype. In addition, our approach does not take into consideration information about missense change types (for example, conservative, nonconservative, or cysteine changes) that can affect this analysis. Although some domains might tolerate conservative or nonconservative changes, they might be extremely intolerant to disulfide bond disruptions (cysteine changes), for example. Finally, with increased variant allele frequency in patients and in the general population, this analysis can be performed based on ethnicity to uncover regions that have subpopulation-specific tolerance or intolerance to variation. All these reasons, in addition to lack of variant data for many genes as described, might explain why only 34 genes had significantly tolerant or intolerant regions in our analysis while 98 genes did not.

An illustrative example to demonstrate some of these complexities is the troponin domain, encoded for by exons 10–14 of the *TNNT2* gene, which is associated with both hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM). This domain was found to be tolerant to variation using our burden analysis (**Supplementary Table S6** online), although missense P/LP variants have been reported in this domain. Interestingly, most of those variants are localized to exon 10, disrupt highly conserved basic arginine residues, and are carried by patients with DCM only and not HCM, consistent with previous findings.[30,31] This example clearly highlights the issues of amino acid properties and conservation, phenotype stratification, and variable tolerance within any domain.

It is also important to note that sufficient variant data were not available for all genes. Well-curated variant databases would be needed to take this approach for other genes. Furthermore, our approach relies on sufficient understanding of disease prevalence, penetrance, and extent of genetic heterogeneity. Because such information is not always available, conservative approaches must be taken in assuming variation of high-allele frequency is likely benign. Systematic collection of these data for all diseases will be a useful resource that can inform data-analysis approaches such as those presented here.

Despite the limitations, our study clearly demonstrates the importance of using the relative distribution of variants in controls versus affected individuals to support variant interpretation and/or prioritization within the relevant genes. Incorporating our intragenic disease burden statistics into the existing phylogenetic-based in silico algorithms, such as SIFT[32] and PolyPhen-2,[33] is likely to enhance their prediction of variant clinical significance. Finally, our study highlights the need for appropriately capturing and sharing disease sequencing data to enable such approaches, which are likely to reduce the interpretation challenge facing clinical genomics as well as guide high-impact functional research in disease genes.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

We thank the many members of the Partners Healthcare Laboratory of Molecular Medicine (LMM) who carefully interpreted variant evidence to form the database used in this analysis. We also thank bioinformatics members at the LMM and CHOP for providing and organizing the input data for this analysis and for performing part of the statistical analysis. Some of the data in this paper were presented at the American Society of Human Genetics (ASHG) annual meeting in Baltimore, Maryland, in 2015.

DISCLOSURE

All authors work for fee-for-service laboratories to perform clinical genetic testing. The other authors declare no conflict of interest.

REFERENCES

1. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* 2013;14:295–300.
2. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;369:1502–1511.
3. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 2014;312:1870–1879.
4. Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA* 2014;311:1035–1045.
5. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793–795.
6. Rehm HL, Berg JS, Brooks LD, et al.; ClinGen. ClinGen–the clinical genome resource. *N Engl J Med* 2015;372:2235–2242.
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. Richards S, Aziz N, Bale S, et al.; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–424.
9. Abou Tayoun AN, Al Turki SH, Oza AM, et al. Improving hearing loss gene testing: a systematic review of gene evidence toward more efficient next-generation sequencing-based diagnostic testing and interpretation. *Genet Med* 2016;18:545–553.
10. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;9:e1003709.
11. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 2014;46:944–950.
12. Exome Aggregation Consortium; Karczewski K, Minikel E, Samocha K, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv preprint, 10 May 2016. http://dx.doi.org/10.1101/030338.
13. Piton A, Redin C, Mandel JL. XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *Am J Hum Genet* 2013;93:368–383.
14. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol* 2016;17:9.
15. Duzkale H, Shen J, McLaughlin H, et al. A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet* 2013;84:453–463.
16. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;6:80–92.
17. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005;33(database issue):D501–D504.
18. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. *Nucleic Acids Res* 2014;42(database issue):D222–D230.
19. Drivas TG, Wojno AP, Tucker BA, Stone EM, Bennett J. Basal exon skipping and genetic pleiotropy: A predictive model of disease pathogenesis. *Sci Transl Med* 2015;7:291ra97.
20. Wang J, Chang YF, Hamilton JI, Wilkinson MF. Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol Cell* 2002;10:951–957.
21. Cheng J, Han DY, Dai P, et al. A novel DFNA5 mutation, IVS8+4 A>G, in the splice donor site of intron 8 causes late-onset non-syndromic hearing loss in a Chinese family. *Clin Genet* 2007;72:471–477.
22. Yu C, Meng X, Zhang S, Zhao G, Hu L, Kong X. A 3-nucleotide deletion in the polypyrimidine tract of intron 7 of the DFNA5 gene causes nonsyndromic hearing impairment in a Chinese family. *Genomics* 2003;82:575–579.
23. Van Laer L, Meyer NC, Malekpour M, et al. A novel DFNA5 mutation does not cause hearing loss in an Iranian family. *J Hum Genet* 2007;52:549–552.
24. Ahmed ZM, Riazuddin S, Aye S, et al. Gene structure and mutant alleles of PCDH15: nonsyndromic deafness DFNB23 and type 1 Usher syndrome. *Hum Genet* 2008;124:215–223.
25. Ahmed ZM, Riazuddin S, Bernstein SL, et al. Mutations of the protocadherin gene PCDH15 cause Usher syndrome type 1F. *Am J Hum Genet* 2001;69: 25–34.
26. Perreault-Micale C, Frieden A, Kennedy CJ, et al. Truncating variants in the majority of the cytoplasmic domain of PCDH15 are unlikely to cause Usher syndrome 1F. *J Mol Diagn* 2014;16:673–678.
27. Verpy E, Leibovici M, Zwaenepoel I, et al. A defect in harmonin, a PDZ domain-containing protein expressed in the inner ear sensory hair cells, underlies Usher syndrome type 1C. *Nat Genet* 2000;26:51–55.
28. Ouyang XM, Xia XJ, Verpy E, et al. Mutations in the alternatively spliced exons of USH1C cause non-syndromic recessive deafness. *Hum Genet* 2002;111: 26–30.
29. Rivas MA, Pirinen M, Conrad DF, et al.; GTEx Consortium; Geuvadis Consortium. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 2015;348:666–669.
30. Li D, Czernuszewicz GZ, Gonzalez O, et al. Novel cardiac troponin T mutation as a cause of familial dilated cardiomyopathy. *Circulation* 2001;104:2188–2193.
31. Rani DS, Dhandapany PS, Nallari P, Narasimhan C, Thangaraj K. A novel arginine to tryptophan (R144W) mutation in troponin T (cTnT) gene in an indian multigenerational family with dilated cardiomyopathy (FDCM). *PLoS One* 2014;9:e101451.
32. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4: 1073–1081.
33. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.