bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Curating clinically relevant transcripts for the interpretation of sequence variants

Marina T. DiStefano[1], Sarah E. Hemphill[1], Brandon J. Cushman[1], Mark J. Bowser[1], Elizabeth Hynes[1], Andrew R. Grant[1], Rebecca K. Siegert[1], Andrea M. Oza[1], Michael A. Gonzalez[2], Sami S. Amr[1,3], Heidi L. Rehm[1,4,5], and Ahmad N. Abou Tayoun[2]

[1]Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA, USA.

[2]Division of Genomic Diagnostics, The Children's Hospital of Philadelphia, The University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

[3]Department of Pathology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA USA

[4]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston MA USA

[5]The Broad Institute of MIT and Harvard, Cambridge, MA, USA

Disclosures: None of the authors have a direct financial or personal interest in the subject matter or materials discussed in the article

Number of text pages, tables and figures: 14 text pages (excluding references and figure legends), 1 table, 5 figures, 4 supplementary tables

Running head (40 characters or less): Transcript Curation for Sequence Variants Interpretation

Grant numbers and sources of support (including departmental or institutional funding when no extramural funding was received); Research reported in this publication was supported by the National Human Genome Research Institute (NHGRI), in conjunction with additional funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) under award number: U41HG006834. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Corresponding author: Ahmad N. Abou Tayoun, The Children's Hospital of Philadelphia

3615 Civic Center Boulevard, RM 716 ARC Bldg. Philadelphia, PA 19104

Office: 215-590-3856, Fax: 215-590-2156, Email: aboutayoua@email.chop.edu

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Variant interpretation depends on accurate annotations using biologically relevant transcripts. We have developed a systematic strategy for designating primary transcripts, and applied it to 109 hearing loss-associated genes that were divided into 3 categories. Category 1 genes (n=38) had a single transcript, Category 2 genes (n=32) had multiple transcripts, but a single transcript was sufficient to represent all exons, and Category 3 genes (n=38) had multiple transcripts with unique exons. Transcripts were curated with respect to gene expression reported in the literature and the Genotype-Tissue Expression Project. In addition, high frequency loss of function variants in the Genome Aggregation Database, and disease-causing variants in ClinVar and the Human Gene Mutation Database across the 109 genes were queried. These data were used to classify exons as "clinically relevant", "uncertain significance", or "clinically insignificant". Interestingly, 7% of all exons, containing >124 "clinically significant" variants, were of “uncertain significance”. Finally, we used exon-level next generation sequencing quality metrics generated at two clinical labs, and identified a total of 43 technically challenging exons in 20 different genes that had inadequate coverage and/or homology issues which might lead to false variant calls. We have demonstrated that transcript analysis plays a critical role in accurate clinical variant interpretation.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

With the rapid growth of genomic testing and the dropping cost of sequencing, proper analysis of genetic variants is critical for patient care. The American College of Medical Genetics and Genomics (ACMG) has set forth guidelines for the interpretation of sequence variants [1]. However, use of the guidelines requires an understanding of the transcriptional architecture of each gene. There can be several mRNA transcripts for each gene, and each laboratory individually determines which transcript to use when annotating, interpreting, and reporting variants in any gene. Human transcripts are currently designated and annotated by multiple groups. The most commonly used sets of coding transcripts that are currently available come from GENCODE (Ensembl and HAVANA, CCDS, LRG), UCSC, RefSeq (LRG), and AceView [2-8]. Each group annotates transcripts with a combination of computational and manual literature curation. Although current HGVS standards mention that describing variants in the context of exons and introns is optional [9], in many genes there are exon-specific factors that influence interpretation and tracking this data on an exon level is important.

In addition to the above annotation challenges, technical limitations of NGS can also lead to inaccurate variant calls. Several genes contain coding sequences that can pose several technical problems, including sequences with high homology to other genomic regions, high GC content, and repetitive sequences. If a gene has significant sequence overlap with another gene or a pseudogene, it can be difficult to align the short NGS reads to the right genomic location, leading to false negative and/or false positive variant calls. DNA with high GC content is not easily amplified, and highly repetitive DNA is prone to sequencing and/or alignment errors. All such regions should be systematically investigated in the targeted genes of interest to address test limitations and design necessary ancillary assays [10].

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Upon passing sequencing quality metrics and filtration cutoffs, a rare variant would then be evaluated based on the most biologically relevant transcript for the disease of interest. It is common to choose the longest transcript for sequencing pipelines. However, variants are often evaluated in the context of this transcript, which does not necessarily encompass all essential exons and can also contain non-biologically functional exons. Thus, choosing a medically relevant transcript is essential for variant interpretation, and for understanding the molecular consequence of a variant on the gene's function.

Here we provide a framework for transcript curation and selection using a combination of tissue expression and genomic datasets, protein functional domains, and published work from animal and human studies. We apply this framework to hearing loss, a relatively common condition that affects 1 in 300 infants, half of which have a genetic etiology [11]. Due to the complexity of the auditory system, it is also highly heterogeneous with over 100 genes causative for nonsyndromic hearing loss alone [12]. We also use clinical NGS datasets generated at two different diagnostic laboratories to systematically highlight technically challenging regions across the hearing loss genes. We demonstrate the utility of our framework and its impact on variant annotation and interpretation. While our analysis was limited to hearing loss, we recommend that this guidance be used for all genes that are definitively associated with fully penetrant diseases.

Methods

Transcript Curation process

109 hearing loss-associated genes largely from the OtoGenome™ Test (GTR000509148.8) at the Laboratory for Molecular Medicine (LMM) were included for transcript curation. All known (NM) RefSeq transcripts in these genes were curated with respect to function, tissue specificity

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and temporal expression from published literature (**Figure 1**). Exon-specific expression data were extracted from the Genotype-Tissue Expression Project (GTEx) on 01/15/18. GTEx was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. In addition, high allele frequency (>0.3%) predicted loss of function (LoF) variants (nonsense, frameshift and +/-1,2 splice site) were queried from the Genome Aggregation Database (gnomAD) [13], and exons containing such LoFs were flagged. An allele frequency of 0.3% was chosen since variants in hearing loss genes that are above this frequency can be considered likely benign as defined by Duzkale et al [14].

Exon Numbering and Classification

All coding and noncoding exons in the primary transcript were numbered sequentially. Coding and noncoding exons in minor transcripts were also numbered separately and sequentially per transcript. Transcripts were aligned and viewed using Alamut (Version 2.6.1 Interactive Biosoftware). Each minor transcript was given a different letter that was added after the exon number. For example, if there were two transcripts with unique alternate exon 1, they were numbered as 1A and 1B. To define a minimal curated transcript list, unique exons were listed in the minimal number of minor transcripts and the designated longest transcript contained the most coding bases **(Table 1, Supplemental Tables 1-3)**.

Exons were classified as "clinically significant", "uncertain significance", or "clinically insignificant" **(Figure 1C)**. Exons were classified as "clinically significant" if there was no evidence they were alternatively spliced, they did not contain high frequency exonic LoF variants, or they were supported by tissue-specific inner ear expression in the literature. "Uncertain significance" exons were spliced out of major transcripts, had no expression data and, for some, contained one high frequency LoF variant. Finally, "clinically insignificant" exons

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were noncoding, had non-supporting human or animal tissue expression data, or had multiple high frequency LoF variants.

Variant Counts

Pathogenic (P), likely pathogenic (LP), benign (B), likely benign (LB) variants, and variants of uncertain clinical significance (VUS) in the ClinVar database [15] in addition to Disease Mutations (DMs) in the Human Gene Mutation Database (HGMD) [16] were counted across all uncertain and insignificant exons. Each variant was evaluated based on transcript location and predicted molecular consequence to the gene.

Technically challenging regions

Exon-level next generation sequencing (NGS) quality metrics, including average mapping quality (MQ) and average minimum depth of coverage (DP), were calculated across all 109 genes using exome sequencing data. Exome targets were captured using the Agilent Clinical Research Exome (CRE) V5 kit, after which 2x100bp paired-end sequencing was performed on the Illumina HiSeq platform.

Exome sequencing using the above conditions was performed at the Children's Hospital of Philadelphia (CHOP) [17] and the Laboratory for Molecular Medicine (LMM) with an overall average coverage of ~100x and 180x, respectively. Poor quality regions as defined in **Results** section were compared between the two sites. Additionally, 87 of the 109 genes were targeted at the LMM using a custom-based capture kit (Agilent), followed by 2x150bp paired-end sequencing to an average coverage of ~600x. Exon-level NGS quality metrics from this capture-based sequencing approach were calculated and compared to those from exome sequencing.

Results

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Transcript curation

A total of 109 genes on LMM's hearing loss panel were curated for clinically relevant transcripts as outlined in **Figure 1**. These genes had between 1 and 17 NCBI reference sequence RefSeq transcripts and between 1 and 72 unique exons, for a total of 340 unique transcripts and 2161 unique exons across all genes (**Figure 2A**). Genes were divided into 3 categories using RefSeq transcripts [5]. All genes with only one RefSeq transcript were classified as category 1 (C1) genes. Genes with multiple transcripts were considered category 2 (C2) genes if the longest transcript included all annotated exons and those with multiple transcripts and mutually exclusive exons were grouped under category 3 (C3) (**Figure 1**).

Category 1 genes. Of the 109 genes evaluated, 38 had a single RefSeq transcript each with an average of 19 exons (total of 38 transcripts and 725 exons across 38 genes, **Figure 2A** and **Supplementary Table 1**). Because each gene had only one known RefSeq transcript, C1 transcripts had minimal curation and almost all exons, except for noncoding ones (n=24), were considered to be critical for inclusion in diagnostic testing and variant interpretation. Interestingly, however, based on exon-level counts of high allele frequency loss of function (LoF) variants in the general population (see **Methods**), there was one insignificant exon (*MYO15A*, NM_016239.3, Exon 26) and another of uncertain significance (*ATP6V1B1*, NM_001692.3, Exon 1) in this category.

The *MYO15A* exon 26 contained a nonsense variant (c.5925G>A; p.Trp1975*) that was present in 1.3% (316/23,712) South Asian alleles including 3 homozygotes in the Genome Aggregation Database (gnomAD). Interestingly, RNA sequencing data from the Genotype-Tissue Expression (GTEx) study showed that this exon is not expressed across all tested tissues. We therefore considered this exon to be clinically insignificant and that variants identified therein are more

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

likely to be benign. The *ATP6V1B1* exon 1 contained a start loss variant (c.2T>C; p.M1?) that was present in 40% total alleles in gnomAD, including 23,280 homozygotes. It is possible that the exon start is erroneously annotated or that re-initiation might occur elsewhere, including at any of the two downstream methionines in this exon. However, there is currently no functional data to support either possibility or to rule out potential re-initiation at downstream exons. We therefore classified this exon as uncertain clinical significance wherein sequence variants should be carefully interpreted.

Category 2 genes. This category included 33 genes each with an average of 4 transcripts and 18 unique exons (**Figure 2A** and **Supplementary Table 2**). A total of 118 transcripts, including 582 unique exons, were curated in this category. Although the longest transcript represented all annotated RefSeq exons and was presumed to be the major transcript for C2 genes, we curated the shorter (minor) transcripts for potentially identifying non-biologically or non-clinically relevant exons in the longer transcript. Apart from the 23 noncoding exons in these genes, there were 26 coding exons not contained in minor transcripts, thus questioning their clinical relevance (**Figure 2B**). Of those, 7 exons were not expressed in any tissue in the GTEx database including 4 exons (*CCDC50* exon 6, *DFNA5* exon 2, *EDN3* exon 4, and *ILDR1* exon 6) harboring high allele frequency LoF variants in gnomAD (**Supplementary Table 2**). Of the 15 exons that showed expression in the GTEx database, 5 exons (*CEP78* exons 1, 2, and 16, *DFNA5* exon 6, and *KARS* exon 15) also contained high frequency LoF variants in gnomAD (**Supplementary Table 2**), while the remaining 10 exons did not, although more information is needed to clarify their biological or clinical relevance.

An illustrative example in this category is the *EDN3* gene known to cause Waardenburg syndrome (WS) type 4 [18, 19]. This gene has 5 RefSeq transcripts sharing coding exons 1-3 and 5

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

but differing in the inclusion of coding exon 4. Specifically, the NM_001302456.1 and NM_207033.2 transcripts do not contain the fourth coding exon shared by the three other transcripts (**Figure 3**). Interestingly, a frameshift variant (NM_207034.2: c.559_560insA; p.Thr189Asnfs) in this exon was present in 0.6% (157/25790) Finnish European alleles in gnomAD, with a high quality variant score, including 2 homozygotes (**Figure 3**), supporting the clinical insignificance of this exon.

Category 3 genes. There were 38 genes in this category, each with an average of 5 transcripts and 22 unique exons (**Figure 2A** and **Supplementary Table 3**). In total, C3 genes had 184 RefSeq transcripts and 854 unique exons. Given the multiple transcripts with mutually exclusive exons in this category, a thorough curation was carried out to select the most clinically relevant transcript for each C3 gene (**Supplementary Table 3**). Published human and/or animal tissue expression studies supported transcript selection for 30 C3 genes; the longest transcript was supported in 25 genes while a shorter isoform was most relevant in 5 genes. There were no expression data to guide selection of the most biologically relevant transcript for 8 C3 genes for which we defaulted to the longest transcript. Overall, 104 coding exons in the C3 genes met our criteria for “uncertain significance”, while 7 coding exons were classified as "clinically insignificant" (see **Methods** and **Figure 2B**).

A C3 example is the *PAX3* gene which is a common cause of WS, type 1 [20-22]. This gene has 8 RefSeq transcripts with varied tissue and temporal expression [23-25], and with significant alternative splicing; a transcript can include 4, 5, 8, 9, or 10 exons. Certain exons use alternate splice junctions which can also change reading frame for the terminal exon (**Figure 4A**). Interestingly, one putative LoF variant (c.638C>A; p.S213*) in exon 4d of the NM_000438.5 transcript was found in 0.54% (166/30,592) South Asian alleles including one homozygous

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

individual. This allele frequency is inconsistent with the estimated disease prevalence for autosomal dominant WS of approximately 1/40,000[22, 26]. Exon 4d is only 20 amino acids longer than exon 4 on all other transcripts. Human expression data in the GTEx database strongly supports usage of the exon 4 (and not 4d) splice donor site, suggesting that exon 4d is not biologically relevant (**Figure 4B**).

Similarly, exon 8c in the NM_181460.3 transcript of the *PAX3* gene is unlikely to be biologically relevant as supported by lack of its expression (GTEx database) and the presence of a putative LoF splicing variant (c.1024+1G>C) impacting this exon in 0.15% (47/30,764) South Asian alleles in gnomAD. This same variant has a missense effect (p.Arg402Pro) on the NM_181461.3 transcript, further highlighting the importance of appropriate transcript selection for variant annotation and interpretation **(Figure 4C)**.

Impact on clinical testing and interpretation

Transcript selection can significantly alter variant interpretation because variants can have differing molecular consequences on each transcript. For example, pathogenic missense and nonsense variants in the *MITF* gene, encoding a transcription factor critical for melanocyte development, are known to cause WS, type 2 [22, 27]. This gene has 13 curated RefSeq transcripts. On 4 out of 13 transcripts, a particular pathogenic variant (which segregated with disease in >10 members of a family with WS [27]) is annotated as a variant in a +1 canonical splice donor site (c.33+1G>A). However, this nucleotide change is a deep intronic variant in the other 9 transcripts (e.g. NM_001184967.1:c.199-1066G>A), and could easily be misclassified if only the deep intronic consequence were interpreted.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Tissue-specific transcript expression can also significantly alter variant interpretation. Variants in *TBC1D24*, a GTPase-activating protein, are associated with either nonsyndromic hearing loss, DOORS syndrome, or a spectrum of epilepsy conditions [28-33]. This gene has 2 curated RefSeq transcripts: NM_001199107.1, the longest which contains 8 exons and is most abundant in mouse neurons, and NM_020705.2, which is missing exon 3, contains only 7 exons, and is expressed in mouse cochlea and non-neuronal tissues [29]. NM_001199107.1:c.969_970delGT (p.Ser324Thrfs), a frameshift variant in exon 3, is not present in the shorter transcript and was identified in the homozygous state in five members of a consanguineous family with severe lethal epileptic encephalopathy but no hearing loss, thereby supporting the tissue-specific expression of the longest transcript [31].

Based on our comprehensive curation of all transcripts, there were 139 coding exons with no or uncertain clinical significance, constituting 7% of all 2089 coding exons across all 109 genes (**Figure 2B**). Because of the limited evidence supporting those exons' clinical relevance, variants therein should be carefully interpreted as they can be a source of false positive diagnoses. Interestingly, there are 124 variants that are labeled as disease causing (DM or P/LP) in disease databases (HGMD and ClinVar, respectively), in addition to 224 VUSs and 151 B/LB variants across those exons (**Figure 2C**). These variants interpreted as clinically significant will all require further assessment to ensure sufficient evidence is present to implicate them in hearing loss. This highlights the importance of our transcript curation approach and the impact it could have on functional and clinical annotations.

Technical Assessment

Due to its genetic heterogeneity, most clinical genetics laboratories use targeted or exome-based panels to sequence a comprehensive set of genes known to cause hearing loss. Although very

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

robust, such approaches have limitations inherent to the next generation sequencing (NGS) technology including the inability to reliably capture and sequence low complexity and/or high homology genomic regions.

We sought to identify regions in a set of 109 hearing loss genes that are technically challenging to sequence using current short read (100-150bp) NGS in clinical laboratories. We used clinical exome sequencing data generated in two different sites (CHOP and LMM) to calculate exon-level quality metrics across all exons in the 109 hearing loss genes. We have recently shown that an average mapping quality (MQ) and/or an average minimum depth of coverage (DP) cutoffs of 20 and/or 15, respectively, are strong indicators of poor quality regions [17]. We identified 43 well-baited (~90% baited bases) exons in 20 genes with the above cutoffs despite being exome sequenced to an overall average coverage of up to 180x at the two clinical laboratories (**Supplementary Table 4** and **methods**). Of those, 31 exons were sequenced to an overall coverage of ~600x using a different targeted capture (average % baited exons: 95%) and longer reads (150bp), but still had low quality metrics (**Supplementary Table 4**).

The 43 regions included exons with high homology to other genomic sequences (n=21 exons in *STRC* and *OTOA*) or exons that have GC-rich or repeat sequences (n=22, e.g. *KCNQ1*, *MYO15A*, and *TPRN*) (**Figure 5**). It is unlikely that sequence variants in all 43 exons will be reliably detected using available NGS chemistries, and therefore false positive and/or false negative variant calls in those exons should be highly expected.

Discussion

Transcript selection is critical for determining DNA variants' potential effects on RNA and/or protein expression, function and stability. This annotation, in turn, significantly impacts variant

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

interpretation. Although most clinical laboratories use one set of coding transcripts (commonly RefSeq) for variant annotation, any set might contain multiple transcripts for each gene; some of which are true clinically relevant isoforms, while others can be false annotations. Even for those genes with multiple isoforms, deciphering the relevant isoform for a given disease – often based on tissue-specific expression data – is necessary for interpretation. In the absence of uniform guidelines for transcript selection, each lab applies different internal rules for identifying the most appropriate transcript(s) for interpretation and reporting.

Here we provide a comprehensive evidence-based framework for transcript curation and selection. We apply this framework to 109 hearing loss genes, and illustrate its utility in transcript selection and variant annotation and interpretation. We also use a new exon classification system, and show that 7% of all coding (RefSeq) exons in these genes have no or questionable clinical validity rendering them a potential source of false variant calls irrespective of their predicted protein effect (missense, loss of function, etc.).

A challenge with our approach is that it requires significant manual curation, though such curation is essential for accurate interpretation, and is arguably more effective if performed ahead of testing, and not retrospectively to minimize analysis and wet bench burdens. Another challenge is that it is highly dependent on availability of human and/or animal expression data in the relevant disease tissue – the inner ear in this current work. However, leveraging existing large human genomic population (gnomAD) and transcriptome (GTEx) sequencing data as well as high quality variant databases (ClinVar) can support the selection of clinically relevant transcripts in our genes.

Finally, we use exon-level NGS quality metrics to highlight regions that are inaccessible to sequencing and/or accurate variant calling, especially with short read (100-150bp) chemistries

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

304 that are mostly used in clinical and research labs. It is possible that some of those regions can be
305 recovered with longer reads and improved bioinformatics pipelines. Until then, however, it is
306 highly important that different ancillary assays, such as Sanger sequencing, be validated to
307 accurately capture sequence variants in those regions.

308 In summary, we recommend that our transcript selection framework and exon classification
309 system be used in other disease areas for more efficient and accurate variant interpretation, and
310 to avoid erroneous annotations and, potentially, misdiagnoses.

311 **Acknowledgements**

312 Research reported in this publication was supported by the National Human Genome Research
313 Institute (NHGRI), in conjunction with additional funding from the Eunice Kennedy Shriver
314 National Institute of Child Health and Human Development (NICHD) under award number:
315 U41HG006834. The content is solely the responsibility of the authors and does not necessarily
316 represent the official views of the National Institutes of Health.

317

318

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL: 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, 17:405-424.

2. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res, 22:1760-1774.

3. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, Hinrichs AS, Haeussler M, Guruvadoo L, Navarro Gonzalez J, Gibson D, Fiddes IT, Eisenhart C, Diekhans M, Clawson H, Barber GP, Armstrong J, Haussler D, Kuhn RM, Kent WJ: The UCSC Genome Browser database: 2018 update. Nucleic Acids Res, 46:D762-D769.

4. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ: ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res, 41:D56-63.

5. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, 44:D733-745.

6. Thierry-Mieg D, Thierry-Mieg J: AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 2006, 7 Suppl 1:S12 11-14.

7. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D: The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 2009, 19:1316-1323.

8. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Juettemann T, Keenan S, Laird MR, Lavidas I, Maurel T, McLaren W, Moore B, Murphy DN, Nag R, Newman V, Nuhn M, Ong CK, Parker A, Patricio M, Riat HS, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Wilder SP, Zadissa A, Kostadima M, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Cunningham F, Yates A, Zerbino DR, Flicek P: Ensembl 2017. Nucleic Acids Res, 45:D635-D642.

9. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE: HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat, 37:564-569.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

10. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB: Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn, 20:4-27.

11. Alford RL, Arnos KS, Fox M, Lin JW, Palmer CG, Pandya A, Rehm HL, Robin NH, Scott DA, Yoshinaga-Itano C: American College of Medical Genetics and Genomics guideline for the clinical evaluation and etiologic diagnosis of hearing loss. Genet Med, 16:347-355.

12. Abou Tayoun AN, Al Turki SH, Oza AM, Bowser MJ, Hernandez AL, Funke BH, Rehm HL, Amr SS: Improving hearing loss gene testing: a systematic review of gene evidence toward more efficient next-generation sequencing-based diagnostic testing and interpretation. Genet Med, 18:545-553.

13. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG: Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536:285-291.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

14. Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, Funke BH, Rehm HL, Lebo MS: A systematic approach to assessing the clinical significance of genetic variants. Clin Genet, 84:453-463.

15. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR: ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res, 44:D862-868.

16. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet, 133:1-9.

17. Niazi R, Gonzalez MA, Balciuniene J, Evans P, Sarmady M, Abou Tayoun AN: ExomeSlicer: a resource for the development and validation of exome-based clinical panels. bioRxiv.

18. Edery P, Attie T, Amiel J, Pelet A, Eng C, Hofstra RM, Martelli H, Bidaud C, Munnich A, Lyonnet S: Mutation of the endothelin-3 gene in the Waardenburg-Hirschsprung disease (Shah-Waardenburg syndrome). Nat Genet 1996, 12:442-444.

19. Hofstra RM, Osinga J, Tan-Sindhunata G, Wu Y, Kamsteeg EJ, Stulp RP, van Ravenswaaij-Arts C, Majoor-Krakauer D, Angrist M, Chakravarti A, Meijers C, Buys CH: A homozygous mutation in the endothelin-3 gene associated with a combined Waardenburg type 2 and Hirschsprung phenotype (Shah-Waardenburg syndrome). Nat Genet 1996, 12:445-447.

20. Baldwin CT, Hoth CF, Amos JA, da-Silva EO, Milunsky A: An exonic mutation in the HuP2 paired domain gene causes Waardenburg's syndrome. Nature 1992, 355:637-638.

21. Tassabehji M, Read AP, Newton VE, Harris R, Balling R, Gruss P, Strachan T: Waardenburg's syndrome patients have mutations in the human homologue of the Pax-3 paired box gene. Nature 1992, 355:635-636.

22. Song J, Feng Y, Acke FR, Coucke P, Vleminckx K, Dhooge IJ: Hearing loss in Waardenburg syndrome: a systematic review. Clin Genet.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

23. Blake JA, Ziman MR: Pax3 transcripts in melanoblast development. Dev Growth Differ 2005, 47:627-635.

24. Monsoro-Burq AH: PAX transcription factors in neural crest development. Semin Cell Dev Biol, 44:87-96.

25. Tsukamoto K, Nakamura Y, Niikawa N: Isolation of two isoforms of the PAX3 gene transcripts and their tissue-specific alternative expression in human adult tissues. Hum Genet 1994, 93:270-274.

26. Shi Y, Li X, Ju D, Li Y, Zhang X, Zhang Y: A novel mutation of the MITF gene in a family with Waardenburg syndrome type 2: A case report. Exp Ther Med, 11:1516-1518.

27. Tassabehji M, Newton VE, Read AP: Waardenburg syndrome type 2 caused by mutations in the human microphthalmia (MITF) gene. Nat Genet 1994, 8:251-255.

28. Azaiez H, Booth KT, Bu F, Huygen P, Shibata SB, Shearer AE, Kolbe D, Meyer N, Black-Ziegelbein EA, Smith RJ: TBC1D24 mutation causes autosomal-dominant nonsyndromic hearing loss. Hum Mutat, 35:819-823.

29. Rehman AU, Santos-Cortez RL, Morell RJ, Drummond MC, Ito T, Lee K, Khan AA, Basra MA, Wasif N, Ayub M, Ali RA, Raza SI, Nickerson DA, Shendure J, Bamshad M, Riazuddin S, Billington N, Khan SN, Friedman PL, Griffith AJ, Ahmad W, Leal SM, Friedman TB: Mutations in TBC1D24, a gene associated with epilepsy, also cause nonsyndromic deafness DFNB86. Am J Hum Genet, 94:144-152.

30. Falace A, Filipello F, La Padula V, Vanni N, Madia F, De Pietri Tonelli D, de Falco FA, Striano P, Dagna Bricarelli F, Minetti C, Benfenati F, Fassio A, Zara F: TBC1D24, an ARF6-interacting protein, is mutated in familial infantile myoclonic epilepsy. Am J Hum Genet, 87:365-370.

31. Guven A, Tolun A: TBC1D24 truncating mutation resulting in severe neurodegeneration. J Med Genet, 50:199-202.

32. Campeau PM, Kasperaviciute D, Lu JT, Burrage LC, Kim C, Hori M, Powell BR, Stewart F, Felix TM, van den Ende J, Wisniewska M, Kayserili H, Rump P, Nampoothiri S, Aftimos S, Mey

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A, Nair LD, Begleiter ML, De Bie I, Meenakshi G, Murray ML, Repetto GM, Golabi M, Blair E, Male A, Giuliano F, Kariminejad A, Newman WG, Bhaskar SS, Dickerson JE, Kerr B, Banka S, Giltay JC, Wieczorek D, Tostevin A, Wiszniewska J, Cheung SW, Hennekam RC, Gibbs RA, Lee BH, Sisodiya SM: The genetic basis of DOORS syndrome: an exome-sequencing study. Lancet Neurol, 13:44-58.

33. Zhang L, Hu L, Chai Y, Pang X, Yang T, Wu H: A dominant mutation in the stereocilia-expressing gene TBC1D24 is a probable cause for nonsyndromic hearing impairment. Hum Mutat, 35:814-818.

34. Modamio-Hoybjor S, Mencia A, Goodyear R, del Castillo I, Richardson G, Moreno F, Moreno-Pelayo MA: A mutation in CCDC50, a gene encoding an effector of epidermal growth factor-mediated cell signaling, causes progressive hearing loss. Am J Hum Genet 2007, 80:1076-1089.

35. Namburi P, Ratnapriya R, Khateb S, Lazar CH, Kinarty Y, Obolensky A, Erdinest I, Marks-Ohana D, Pras E, Ben-Yosef T, Newman H, Gross M, Swaroop A, Banin E, Sharon D: Bi-allelic Truncating Mutations in CEP78, Encoding Centrosomal Protein 78, Cause Cone-Rod Degeneration with Sensorineural Hearing Loss. Am J Hum Genet 2016, 99:777-784.

36. Robertson NG, Resendes BL, Lin JS, Lee C, Aster JC, Adams JC, Morton CC: Inner ear localization of mRNA and protein products of COCH, mutated in the sensorineural deafness and vestibular disorder, DFNA9. Hum Mol Genet 2001, 10:2493-2500.

37. Robertson NG, Skvorak AB, Yin Y, Weremowicz S, Johnson KR, Kovatch KA, Battey JF, Bieber FR, Morton CC: Mapping and characterization of a novel cochlear gene in human and in mouse: a positional candidate gene for a deafness disorder, DFNA9. Genomics 1997, 46:345-354.

38. Van Laer L, Huizing EH, Verstreken M, van Zuijlen D, Wauters JG, Bossuyt PJ, Van de Heyning P, McGuirt WT, Smith RJ, Willems PJ, Legan PK, Richardson GP, Van Camp G: Nonsyndromic hearing impairment is associated with a mutation in DFNA5. Nat Genet 1998, 20:194-197.

39. Delmaghani S, del Castillo FJ, Michel V, Leibovici M, Aghaie A, Ron U, Van Laer L, Ben-Tal N, Van Camp G, Weil D, Langa F, Lathrop M, Avan P, Petit C: Mutations in the gene encoding

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

pejvakin, a newly identified protein of the afferent auditory pathway, cause DFNB59 auditory neuropathy. Nat Genet 2006, 38:770-778.

40. Cheng J, Zhu Y, He S, Lu Y, Chen J, Han B, Petrillo M, Wrzeszczynski KO, Yang S, Dai P, Zhai S, Han D, Zhang MQ, Li W, Liu X, Li H, Chen ZY, Yuan H: Functional mutation of SMAC/DIABLO, encoding a mitochondrial proapoptotic protein, causes human progressive hearing loss DFNA64. Am J Hum Genet 2011, 89:56-66.

41. Grifa A, Wagner CA, D'Ambrosio L, Melchionda S, Bernardi F, Lopez-Bigas N, Rabionet R, Arbones M, Monica MD, Estivill X, Zelante L, Lang F, Gasparini P: Mutations in GJB6 cause nonsyndromic autosomal dominant deafness at DFNA3 locus. Nat Genet 1999, 23:16-18.

42. Borck G, Ur Rehman A, Lee K, Pogoda HM, Kakar N, von Ameln S, Grillet N, Hildebrand MS, Ahmed ZM, Nurnberg G, Ansar M, Basit S, Javed Q, Morell RJ, Nasreen N, Shearer AE, Ahmad A, Kahrizi K, Shaikh RS, Ali RA, Khan SN, Goebel I, Meyer NC, Kimberling WJ, Webster JA, Stephan DA, Schiller MR, Bahlo M, Najmabadi H, Gillespie PG, Nurnberg P, Wollnik B, Riazuddin S, Smith RJ, Ahmad W, Muller U, Hammerschmidt M, Friedman TB, Riazuddin S, Leal SM, Ahmad J, Kubisch C: Loss-of-function mutations of ILDR1 cause autosomal-recessive hearing impairment DFNB42. Am J Hum Genet 2011, 88:127-137.

43. Santos-Cortez RL, Lee K, Azeem Z, Antonellis PJ, Pollock LM, Khan S, Irfanullah, Andrade-Elizondo PB, Chiu I, Adams MD, Basit S, Smith JD, University of Washington Center for Mendelian G, Nickerson DA, McDermott BM, Jr., Ahmad W, Leal SM: Mutations in KARS, encoding lysyl-tRNA synthetase, cause autosomal-recessive nonsyndromic hearing impairment DFNB89. Am J Hum Genet 2013, 93:132-140.

44. Beisel KW, Rocha-Sanchez SM, Morris KA, Nie L, Feng F, Kachar B, Yamoah EN, Fritzsch B: Differential expression of KCNQ4 in inner hair cells and sensory neurons is the basis of progressive high-frequency hearing loss. J Neurosci 2005, 25:9285-9293.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

45. Xu T, Nie L, Zhang Y, Mo J, Feng W, Wei D, Petrov E, Calisto LE, Kachar B, Beisel KW, Vazquez AE, Yamoah EN: Roles of alternative splicing in the functional properties of inner ear-specific KCNQ4 channels. J Biol Chem 2007, 282:23899-23909.

46. Zazo Seco C, Serrao de Castro L, van Nierop JW, Morin M, Jhangiani S, Verver EJ, Schraders M, Maiwald N, Wesdorp M, Venselaar H, Spruijt L, Oostrik J, Schoots J, Baylor-Hopkins Center for Mendelian G, van Reeuwijk J, Lelieveld SH, Huygen PL, Insenser M, Admiraal RJ, Pennings RJ, Hoefsloot LH, Arias-Vasquez A, de Ligt J, Yntema HG, Jansen JH, Muzny DM, Huls G, van Rossum MM, Lupski JR, Moreno-Pelayo MA, Kunst HP, Kremer H: Allelic Mutations of KITLG, Encoding KIT Ligand, Cause Asymmetric and Unilateral Hearing Loss and Waardenburg Syndrome Type 2. Am J Hum Genet 2015, 97:647-660.

47. Riazuddin S, Ahmed ZM, Fanning AS, Lagziel A, Kitajiri S, Ramzan K, Khan SN, Chattaraj P, Friedman PL, Anderson JM, Belyantseva IA, Forge A, Riazuddin S, Friedman TB: Tricellulin is a tight-junction protein necessary for hearing. Am J Hum Genet 2006, 79:1040-1051.

48. Ahituv N, Sobe T, Robertson NG, Morton CC, Taggart RT, Avraham KB: Genomic structure of the human unconventional myosin VI gene. Gene 2000, 261:269-275.

49. Rehman AU, Santos-Cortez RL, Morell RJ, Drummond MC, Ito T, Lee K, Khan AA, Basra MA, Wasif N, Ayub M, Ali RA, Raza SI, University of Washington Center for Mendelian G, Nickerson DA, Shendure J, Bamshad M, Riazuddin S, Billington N, Khan SN, Friedman PL, Griffith AJ, Ahmad W, Riazuddin S, Leal SM, Friedman TB: Mutations in TBC1D24, a gene associated with epilepsy, also cause nonsyndromic deafness DFNB86. Am J Hum Genet 2014, 94:144-152.

50. Adato A, Lefevre G, Delprat B, Michel V, Michalski N, Chardenoux S, Weil D, El-Amraoui A, Petit C: Usherin, the defective protein in Usher syndrome type IIA, is likely to be a component of interstereocilia ankle links in the inner ear sensory cells. Hum Mol Genet 2005, 14:3921-3932.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

519 51. Mathur PD, Zou J, Zheng T, Almishaal A, Wang Y, Chen Q, Wang L, Vashist D, Brown S, Park
520 A, Yang J: Distinct expression and function of whirlin isoforms in the inner ear and retina: an
521 insight into pathogenesis of USH2D and DFNB31. Hum Mol Genet 2015, 24:6213-6228.
522 52. Yang J, Liu X, Zhao Y, Adamian M, Pawlyk B, Sun X, McMillan DR, Liberman MC, Li T:
523 Ablation of whirlin long isoform disrupts the USH2 protein complex and causes vision and
524 hearing loss. PLoS Genet 2010, 6:e1000955.
525 53. Ebrahim S, Ingham NJ, Lewis MA, Rogers MJC, Cui R, Kachar B, Pass JC, Steel KP:
526 Alternative Splice Forms Influence Functions of Whirlin in Mechanosensory Hair Cell
527 Stereocilia. Cell Rep 2016, 15:935-943.
528 54. Wayne S, Robertson NG, DeClau F, Chen N, Verhoeven K, Prasad S, Tranebjarg L, Morton CC,
529 Ryan AF, Van Camp G, Smith RJ: Mutations in the transcriptional activator EYA4 cause late-
530 onset deafness at the DFNA10 locus. Hum Mol Genet 2001, 10:195-200.
531 55. Wang Y, Liu Y, Nie H, Ma X, Xu Z: Alternative splicing of inner-ear-expressed genes. Front
532 Med 2016, 10:250-257.
533 56. Lagziel A, Overlack N, Bernstein SL, Morell RJ, Wolfrum U, Friedman TB: Expression of
534 cadherin 23 isoforms is not conserved: implications for a mouse model of Usher syndrome type
535 1D. Mol Vis 2009, 15:1843-1857.
536 57. Riazuddin S, Belyantseva IA, Giese AP, Lee K, Indzhykulian AA, Nandamuri SP, Yousaf R,
537 Sinha GP, Lee S, Terrell D, Hegde RS, Ali RA, Anwar S, Andrade-Elizondo PB, Sirmaci A,
538 Parise LV, Basit S, Wali A, Ayub M, Ansar M, Ahmad W, Khan SN, Akram J, Tekin M,
539 Riazuddin S, Cook T, Buschbeck EK, Frolenkov GI, Leal SM, Friedman TB, Ahmed ZM:
540 Alterations of the CIB2 calcium- and integrin-binding protein cause Usher syndrome type 1J and
541 nonsyndromic deafness DFNB48. Nat Genet 2012, 44:1265-1271.
542 58. Wilcox ER, Burton QL, Naz S, Riazuddin S, Smith TN, Ploplis B, Belyantseva I, Ben-Yosef T,
543 Liburd NA, Morell RJ, Kachar B, Wu DK, Griffith AJ, Riazuddin S, Friedman TB: Mutations in

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the gene encoding tight junction claudin-14 cause autosomal recessive deafness DFNB29. Cell 2001, 104:165-172.

59. Adato A, Vreugde S, Joensuu T, Avidan N, Hamalainen R, Belenkiy O, Olender T, Bonne-Tamir B, Ben-Asher E, Espinos C, Millan JM, Lehesjoki AE, Flannery JG, Avraham KB, Pietrokovski S, Sankila EM, Beckmann JS, Lancet D: USH3A transcripts encode clarin-1, a four-transmembrane-domain protein with a possible role in sensory synapses. Eur J Hum Genet 2002, 10:339-350.

60. Vastinsalo H, Jalkanen R, Dinculescu A, Isosomppi J, Geller S, Flannery JG, Hauswirth WW, Sankila EM: Alternative splice variants of the USH3A gene Clarin 1 (CLRN1). Eur J Hum Genet 2011, 19:30-35.

61. Lynch ED, Lee MK, Morrow JE, Welcsh PL, Leon PE, King MC: Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the Drosophila gene diaphanous. Science 1997, 278:1315-1318.

62. Sanchez-Mejias A, Fernandez RM, Lopez-Alonso M, Antinolo G, Borrego S: New roles of EDNRB and EDN3 in the pathogenesis of Hirschsprung disease. Genet Med 2010, 12:39-43.

63. Cui L, Wong EH, Cheng G, Firmato de Almeida M, So MT, Sham PC, Cherny SS, Tam PK, Garcia-Barcelo MM: Genetic Analyses of a Three Generation Family Segregating Hirschsprung Disease and Iris Heterochromia. PLoS One 2013, 8:e66631.

64. Grillet N, Schwander M, Hildebrand MS, Sczaniecka A, Kolatkar A, Velasco J, Webster JA, Kahrizi K, Najmabadi H, Kimberling WJ, Stephan D, Bahlo M, Wiltshire T, Tarantino LM, Kuhn P, Smith RJ, Muller U: Mutations in LOXHD1, an evolutionarily conserved stereociliary protein, disrupt hair cell function in mice and cause progressive hearing loss in humans. Am J Hum Genet 2009, 85:328-337.

65. Ahmed ZM, Masmoudi S, Kalay E, Belyantseva IA, Mosrati MA, Collin RW, Riazuddin S, Hmani-Aifa M, Venselaar H, Kawar MN, Tlili A, van der Zwaag B, Khan SY, Ayadi L, Riazuddin SA, Morell RJ, Griffith AJ, Charfedine I, Caylan R, Oostrik J, Karaguzel A, Ghorbel

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A, Riazuddin S, Friedman TB, Ayadi H, Kremer H: Mutations of LRTOMT, a fusion gene with alternative reading frames, cause nonsyndromic deafness in humans. Nat Genet 2008, 40:1335-1340.

66. Hershey CL, Fisher DE: Genomic analysis of the Microphthalmia locus and identification of the MITF-J/Mitf-J isoform. Gene 2005, 347:73-82.

67. Chen L, Guo W, Ren L, Yang M, Zhao Y, Guo Z, Yi H, Li M, Hu Y, Long X, Sun B, Li J, Zhai S, Zhang T, Tian S, Meng Q, Yu N, Zhu D, Tang G, Tang Q, Ren L, Liu K, Zhang S, Che T, Yu Z, Wu N, Jing L, Zhang R, Cong T, Chen S, Zhao Y, Zhang Y, Bai X, Guo Y, Zhao L, Zhang F, Zhao H, Zhang L, Hou Z, Zhao J, Li J, Zhang L, Sun W, Zou X, Wang T, Ge L, Liu Z, Hu X, Wang J, Yang S, Li N: A de novo silencer causes elimination of MITF-M expression and profound hearing loss in pigs. BMC Biol 2016, 14:52.

68. Ahmed ZM, Yousaf R, Lee BC, Khan SN, Lee S, Lee K, Husnain T, Rehman AU, Bonneux S, Ansar M, Ahmad W, Leal SM, Gladyshev VN, Belyantseva IA, Van Camp G, Riazuddin S, Friedman TB, Riazuddin S: Functional null mutations of MSRB3 encoding methionine sulfoxide reductase are associated with human deafness DFNB74. Am J Hum Genet 2011, 88:19-29.

69. Kelley PM, Weston MD, Chen ZY, Orten DJ, Hasson T, Overbeck LD, Pinnt J, Talmadge CB, Ing P, Mooseker MS, Corey D, Sumegi J, Kimberling WJ: The genomic structure of the gene defective in Usher syndrome type Ib (MYO7A). Genomics 1997, 40:73-79.

70. Haraksingh RR, Jahanbani F, Rodriguez-Paris J, Gelernter J, Nadeau KC, Oghalai JS, Schrijver I, Snyder MP: Exome sequencing and genome-wide copy number variant mapping reveal novel associations with sensorineural hereditary hearing loss. BMC Genomics 2014, 15:1155.

71. Zwaenepoel I, Mustapha M, Leibovici M, Verpy E, Goodyear R, Liu XZ, Nouaille S, Nance WE, Kanaan M, Avraham KB, Tekaia F, Loiselet J, Lathrop M, Richardson G, Petit C: Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying acellular gels, is defective in autosomal recessive deafness DFNB22. Proc Natl Acad Sci U S A 2002, 99:6240-6245.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

596 72. Choi BY, Ahmed ZM, Riazuddin S, Bhinder MA, Shahzad M, Husnain T, Riazuddin S, Griffith
597 AJ, Friedman TB: Identities and frequencies of mutations of the otoferlin gene (OTOF) causing
598 DFNB9 deafness in Pakistan. Clin Genet 2009, 75:237-243.
599 73. Cohen-Salmon M, El-Amraoui A, Leibovici M, Petit C: Otogelin: a glycoprotein specific to the
600 acellular membranes of the inner ear. Proc Natl Acad Sci U S A 1997, 94:14450-14455.
601 74. Schraders M, Ruiz-Palmero L, Kalay E, Oostrik J, del Castillo FJ, Sezgin O, Beynon AJ, Strom
602 TM, Pennings RJ, Zazo Seco C, Oonk AM, Kunst HP, Dominguez-Ruiz M, Garcia-Arumi AM,
603 del Campo M, Villamar M, Hoefsloot LH, Moreno F, Admiraal RJ, del Castillo I, Kremer H:
604 Mutations of the gene encoding otogelin are a cause of autosomal-recessive nonsyndromic
605 moderate hearing impairment. Am J Hum Genet 2012, 91:883-889.
606 75. Housley GD, Kanjhan R, Raybould NP, Greenwood D, Salih SG, Jarlebark L, Burton LD, Setz
607 VC, Cannell MB, Soeller C, Christie DL, Usami S, Matsubara A, Yoshie H, Ryan AF, Thorne
608 PR: Expression of the P2X(2) receptor subunit of the ATP-gated ion channel in the cochlea:
609 implications for sound transduction and auditory neurotransmission. J Neurosci 1999, 19:8377-
610 8388.
611 76. Lynch KJ, Touma E, Niforatos W, Kage KL, Burgard EC, van Biesen T, Kowaluk EA, Jarvis
612 MF: Molecular and functional characterization of human P2X(2) receptors. Mol Pharmacol 1999,
613 56:1171-1181.
614 77. Monsoro-Burq AH: PAX transcription factors in neural crest development. Semin Cell Dev Biol
615 2015, 44:87-96.
616 78. Alagramam KN, Miller ND, Adappa ND, Pitts DR, Heaphy JC, Yuan H, Smith RJ: Promoter,
617 alternative splice forms, and genomic structure of protocadherin 15. Genomics 2007, 90:482-492.
618 79. Alagramam KN, Yuan H, Kuehn MH, Murcia CL, Wayne S, Srisailpathy CR, Lowry RB, Knaus
619 R, Van Laer L, Bernier FP, Schwartz S, Lee C, Morton CC, Mullins RF, Ramesh A, Van Camp
620 G, Hageman GS, Woychik RP, Smith RJ: Mutations in the novel protocadherin PCDH15 cause
621 Usher syndrome type 1F. Hum Mol Genet 2001, 10:1709-1718.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

80. Ahmed ZM, Riazuddin S, Ahmad J, Bernstein SL, Guo Y, Sabar MF, Sieving P, Riazuddin S, Griffith AJ, Friedman TB, Belyantseva IA, Wilcox ER: PCDH15 is expressed in the neurosensory epithelium of the eye and ear and mutant alleles are responsible for both USH1F and DFNB23. Hum Mol Genet 2003, 12:3215-3223.

81. Khan SY, Ahmed ZM, Shabbir MI, Kitajiri S, Kalsoom S, Tasneem S, Shayiq S, Ramesh A, Srisailpathy S, Khan SN, Smith RJ, Riazuddin S, Friedman TB, Riazuddin S: Mutations of the RDX gene cause nonsyndromic hearing loss at the DFNB24 locus. Hum Mutat 2007, 28:417-423.

82. Jin H, May M, Tranebjaerg L, Kendall E, Fontan G, Jackson J, Subramony SH, Arena F, Lubs H, Smith S, Stevenson R, Schwartz C, Vetrie D: A novel X-linked gene, DDP, shows mutations in families with deafness (DFN-1), dystonia, mental deficiency and blindness. Nat Genet 1996, 14:177-180.

83. Nakane T, Inada Y, Ito F, Itoh N, Tazawa S, Chiba S: Cloning and expression of mouse deafness dystonia peptide 1 cDNA. Biochem Biophys Res Commun 2000, 273:759-764.

84. Scott HS, Kudoh J, Wattenhofer M, Shibuya K, Berry A, Chrast R, Guipponi M, Wang J, Kawasaki K, Asakawa S, Minoshima S, Younus F, Mehdi SQ, Radhakrishna U, Papasavvas MP, Gehrig C, Rossier C, Korostishevsky M, Gal A, Shimizu N, Bonne-Tamir B, Antonarakis SE: Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. Nat Genet 2001, 27:59-63.

85. Verpy E, Leibovici M, Zwaenepoel I, Liu XZ, Gal A, Salem N, Mansour A, Blanchard S, Kobayashi I, Keats BJ, Slim R, Petit C: 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.

86. Khateb S, Zelinger L, Ben-Yosef T, Merin S, Crystal-Shalit O, Gross M, Banin E, Sharon D: Exome sequencing identifies a founder frameshift mutation in an alternative exon of USH1C as the cause of autosomal recessive retinitis pigmentosa with late-onset hearing loss. PLoS One 2012, 7:e51566.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

87. Bahloul A, Michel V, Hardelin JP, Nouaille S, Hoos S, Houdusse A, England P, Petit C: Cadherin-23, myosin VIIa and harmonin, encoded by Usher syndrome type I genes, form a ternary complex and interact with membrane phospholipids. Hum Mol Genet 2010, 19:3557-3565.

88. Santos-Cortez RL, Lee K, Giese AP, Ansar M, Amin-Ud-Din M, Rehn K, Wang X, Aziz A, Chiu I, Hussain Ali R, Smith JD, University of Washington Center for Mendelian G, Shendure J, Bamshad M, Nickerson DA, Ahmed ZM, Ahmad W, Riazuddin S, Leal SM: Adenylate cyclase 1 (ADCY1) mutations cause recessive hearing impairment in humans and defects in hair cell function and hearing in zebrafish. Hum Mol Genet 2014, 23:3289-3298.

89. Nyegaard M, Rendtorff ND, Nielsen MS, Corydon TJ, Demontis D, Starnawska A, Hedemand A, Buniello A, Niola F, Overgaard MT, Leal SM, Ahmad W, Wikman FP, Petersen KB, Cruger DG, Oostrik J, Kremer H, Tommerup N, Frodin M, Steel KP, Tranebjaerg L, Borglum AD: A Novel Locus Harbouring a Functional CD164 Nonsense Mutation Identified in a Large Danish Family with Nonsyndromic Hearing Impairment. PLoS Genet 2015, 11:e1005386.

90. Gagnon LH, Longo-Guess CM, Berryman M, Shin JB, Saylor KW, Yu H, Gillespie PG, Johnson KR: The chloride intracellular channel protein CLIC5 is expressed at high levels in hair cell stereocilia and is essential for normal inner ear function. J Neurosci 2006, 26:10188-10198.

91. Diaz-Horta O, Subasioglu-Uzak A, Grati M, DeSmidt A, Foster J, 2nd, Cao L, Bademci G, Tokgoz-Yilmaz S, Duman D, Cengiz FB, Abad C, Mittal R, Blanton S, Liu XZ, Farooq A, Walz K, Lu Z, Tekin M: FAM65B is a membrane-associated protein of hair cell stereocilia required for hearing. Proc Natl Acad Sci U S A 2014, 111:9864-9868.

92. Zhao B, Wu Z, Muller U: Murine Fam65b forms ring-like structures at the base of stereocilia critical for mechanosensory hair cell function. Elife 2016, 5.

93. Thoenes M, Zimmermann U, Ebermann I, Ptok M, Lewis MA, Thiele H, Morlot S, Hess MM, Gal A, Eisenberger T, Bergmann C, Nurnberg G, Nurnberg P, Steel KP, Knipper M, Bolz HJ:

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

673 OSBPL2 encodes a protein of inner and outer hair cell stereocilia and is mutated in autosomal
674 dominant hearing loss (DFNA67). Orphanet J Rare Dis 2015, 10:15.
675 94. Grati M, Shin JB, Weston MD, Green J, Bhat MA, Gillespie PG, Kachar B: Localization of
676 PDZD7 to the stereocilia ankle-link associates this scaffolding protein with the Usher syndrome
677 protein network. J Neurosci 2012, 32:14288-14293.
678 95. Schneider E, Marker T, Daser A, Frey-Mahn G, Beyer V, Farcas R, Schneider-Ratzke B,
679 Kohlschmidt N, Grossmann B, Bauss K, Napiontek U, Keilmann A, Bartsch O, Zechner U,
680 Wolfrum U, Haaf T: Homozygous disruption of PDZD7 by reciprocal translocation in a
681 consanguineous family: a new member of the Usher syndrome protein interactome causing
682 congenital hearing impairment. Hum Mol Genet 2009, 18:655-666.
683 96. Baig SM, Koschak A, Lieb A, Gebhart M, Dafinger C, Nurnberg G, Ali A, Ahmad I, Sinnegger-
684 Brauns MJ, Brandt N, Engel J, Mangoni ME, Farooq M, Khan HU, Nurnberg P, Striessnig J,
685 Bolz HJ: Loss of Ca(v)1.3 (CACNA1D) function in a human channelopathy with bradycardia and
686 congenital deafness. Nat Neurosci 2011, 14:77-84.
687 97. Schultz JM, Khan SN, Ahmed ZM, Riazuddin S, Waryah AM, Chhatre D, Starost MF, Ploplis B,
688 Buckley S, Velasquez D, Kabra M, Lee K, Hassan MJ, Ali G, Ansar M, Ghosh M, Wilcox ER,
689 Ahmad W, Merlino G, Leal SM, Riazuddin S, Friedman TB, Morell RJ: Noncoding mutations of
690 HGF are associated with nonsyndromic hearing loss, DFNB39. Am J Hum Genet 2009, 85:25-39.

691

692

693

694

695

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

696

697 **Figure Legends:**

698 **Figure 1: A) Transcript Curation Workflow:** 109 hearing loss-associated genes,
699 predominantly from the OtoGenome™ Test (GTR000509148.7), were categorized. Genes were
700 divided into 3 categories using NCBI reference sequence (RefSeq) transcripts. Category 1 (C1)
701 contained genes that had a single transcript, genes in category 2 (C2) had multiple transcripts,
702 but the longest transcript encompassed all exons, and Category 3 (C3) genes had multiple
703 transcripts with unique exons. **B) Category 2 and 3 Curation Process:** Category 2 and 3 genes
704 were manually curated. Exon-specific expression data was pulled from GTEx and gEAR.
705 Literature searches were performed for information about functional domains, additional
706 expression data, such as tissue-specific transcript expression, and temporal expression. To
707 evaluate population variation, loss of function variants were pulled from gnomAD. To evaluate
708 interpreted variation, LP/P variants were pulled from our internal database (also in ClinVar) and
709 ClinVar and DM variants were pulled from HGMD. **C) Exon Curation Process:** Exons were
710 categorized as “Clinically relevant,” “Uncertain significance,” or “Clinically Insignificant” based
711 on the pieces of evidence listed in the table.

712 **Figure 2: A) Gene, exon, and transcript Counts:** Genes were categorized and transcripts and
713 exons were counted. **B) Exon Counts Across Categories:** For each of the three categories,
714 exons were classified as Noncoding, Uncertain, or Insignificant as per the definitions in Figure
715 1C. **C) Impact on Interpretation:** Variant counts in uncertain and insignificant exons for each
716 category were collected from HGMD and ClinVar. DM, Disease causing mutation; LP, Likely
717 Pathogenic; P, Pathogenic; VUS, Variant of Uncertain Significance.

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

718 **Figure 3: Visualization of category 2 example, *EDN3*. A)** Transcript view of the *EDN3* gene.
719 The high frequency loss of function variant is pulled from gnomAD and is located in the exon
720 boxed in red.

721 **Figure 4: Visualization of category 3 example, *PAX3*. A)** Transcript view of the *PAX3* gene.
722 **B)** A close-up of the high frequency nonsense variant in exon 4d. **C)** A close-up of the two
723 uncertain exons in PAX3, and 9a (NM_181459.3) and 8c (NM_181460.3).

724 **Figure 5. Visualization of three genes with known technically challenging regions.** Exons in
725 Otoancorin, *OTOA* (*top*) or Stereoclin, *STRC* (*bottom*) with high homology to other genomic
726 sequences; and GC-rich first exon in the potassium channel, *KCNQ1* (*middle*). MQ and coverage
727 plots are displayed for *OTOA*, *KCNQ1* and *STRC*. Green bars indicate exons with both average
728 MQ $\geq$ 20 and min DP $\geq$ 15. Orange bar represents either average MQ < 20 or min DP < 15. Red
729 bars indicate poor exons where both average MQ < 20 and min DP < 15. Each bar in MQ and
730 coverage plots shows minimum and maximum range for each exon (top and bottom of the bar),
731 average is shown by a tick mark in the middle of each bar.

732

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

733 **Table 1: Curated transcripts for category 2 and 3 genes.** The minimal curated transcript set
734 with all unique exons for categories 2 and 3 are listed. C2 genes in which the curated transcript is
735 not the longest one are listed in bold.

Gene	Category	Transcript Set
ACTG1	2	NM_001199954.1†
BCS1L	2	NM_004328.4†
CATSPER2	2	NM_001282310.1†
CCDC50	2	NM_178335.2†
CEP78	2	NM_001098802.1†
CHD7	2	NM_017780.3†
COCH	2	NM_001347720.1†
COL11A2	2	NM_080680.2†
COL4A5	2	NM_033380.2†
DFNA5	2	NM_004403.2†
DFNB59 (PJVK)	2	NM_001042702.3†
DIABLO	2	NM_019887.5†
EDN3	2	NM_207034.2†
GJB6	2	NM_001110219.2†
GPSM2	**2**	**NM_013296.4**
HARS	2	NM_002109.5†
HARS2	2	NM_012208.2†
ILDR1	2	NM_001199799.1†
KARS	2	NM_001130089.1†
KCNE1	2	NM_000219.5†
KCNQ4	2	NM_004700.3†
KITLG	2	NM_000899.4†
MARVELD2	2	NM_001038603.2†
MYH14	2	NM_001145809.1†
MYO6	**2**	**NM_001300899.1**
NLRP3	2	NM_004895.4†
PRPS1	2	NM_002764.3†
SLC52A2	2	NM_024531.4†
SYNE4	2	NM_001039876.2
TBC1D24	**2**	**NM_020705.2**
USH1G	2	NM_173477.4†
USH2A	2	NM_206933.2†
	2	NM_007123.5

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

WFS1	2	NM_006005.3†
ADCY1	3	NM_021116.2*†
	3	NM_001281768.1
CABP2	3	NM_016366.2*
	3	NM_001318496.1†
CACNA1D	3	NM_000720.3*†
	3	NM_001128839.2
CDC14A	3	NM_033312.2*†
	3	NM_003672.3
	3	NM_033313.2
	3	NM_001319210.1
CDH23	3	NM_022124.5*†
	3	NM_001171935.1
	3	NM_001171932.1
	3	NM_052836.3
	3	NM_001171931.1
	3	NM_001171930.1
	3	NM_001171936.1
CD164	3	NM_006016.5*†
	3	NM_001142404.2
CIB2	3	NM_006383*†
	3	NM_001301224
CLIC5	3	NM_001114086.1*†
	3	NM_016929.4
	3	NM_001256023.1
CLDN14	3	NM_144492*†
CLRN1	3	NM_174878.2*
	3	NM_052995.2
	3	NM_001195794.1†
	3	NM_001256819.1
DFNB31 (WHRN)	3	NM_015404.3*†
	3	NM_001346890.1
DIAPH1	3	NM_005219.4*†
	3	NM_001314007.1
EDNRB	3	NM_000115.4*†
	3	NM_001201397.1
	3	NM_003991.3
EYA1	3	NM_000503.5*†
	3	NM_172060.3

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

EYA4	3	NM_004100.4*
	3	NM_172105.3
	3	NM_001301013.1†
HGF	3	NM_000601.5*†
	3	NM_001010931.2
	3	NM_001010934.2
HSD17B4	3	NM_001199291.1*†
	3	NM_000414.3
	3	NM_0012929027.1
	3	NM_001292028.1
KCNQ1	3	NM_000218.2*†
	3	NM_181798.1
LOXHD1	3	NM_144612.6*†
	3	NM_001145472.2
LRTOMT	3	NM_00145309.5*†
	3	NM_001205138.3
	3	NM_001145307.4
MITF	3	NM_000248.3*
	3	NM_198159.2†
MSRB3	3	NM_001193460.1*
	3	NM_198080.3†
MYO7A	3	NM_000260.3*†
	3	NM_001127179.2
OSBPL2	3	NM_144498.2*†
	3	NM_001278649.1
OTOA	3	NM_144672.3*†
	3	NM_001161683.1
	3	NM_170664.2
OTOF	3	NM_001287489.1*
	3	NM_194248.2†
	3	NM_194322.2
OTOG	3	NM_001277269.1*†
	3	NM_001292063.1
PAX3	3	NM_001127366.2*
	3	NM_181459.3†
	3	NM_181457.3
	3	NM_181461.3
	3	NM_000438.5
	3	NM_013942.4

bioRxiv preprint doi: https://doi.org/10.1101/276287; this version posted March 5, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

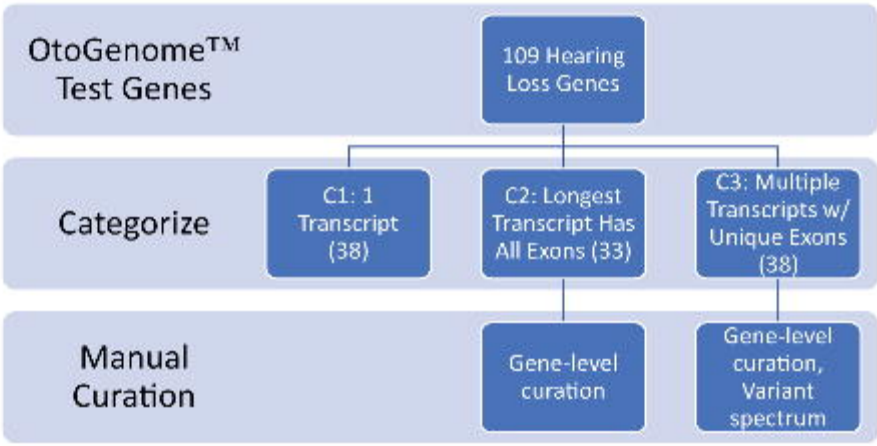
PCDH15	3	NM_001142763.1*†
	3	NM_001142769.1
	3	NM_001142771.1
	3	NM_001354430.1
	3	NM_001142770.1
PDZD7	3	NM_001195263.1*†
	3	NM_001351044.1
	3	NM_024895.4
P2RX2	3	NM_174873.1*
	3	NM_170683.2†
	3	NM_001282164.1
	3	NM_001282165.1
RIPOR2	3	NM_014722.3*†
	3	NM_001286445.1
	3	NM_001286446.1
	3	NM_015864.3
RDX	3	NM_002906.3*†
	3	NM_001260494.1
SERPINB6	3	NM_004568.5*
	3	NM_001271822.1
	3	NM_001271823.1
TIMM8A	3	NM_004085.3*†
	3	NM_001145951.1
TMPRSS3	3	NM_024022.2*†
	3	NM_032405.1
TRIOBP	3	NM_001039141.2*†
	3	NM_138632.2
USH1C	3	NM_153676.3*†
	3	NM_001297764.1

736 *Primary transcript, †Longest transcript

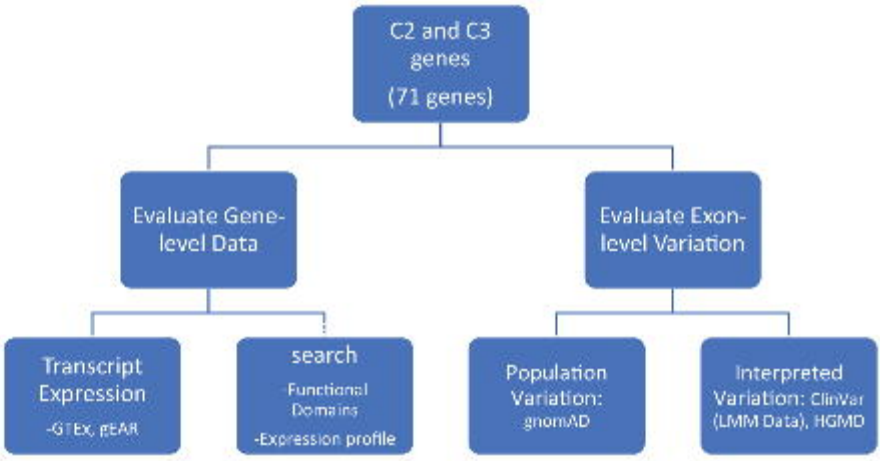
737

FIGURE 1

A.



B.

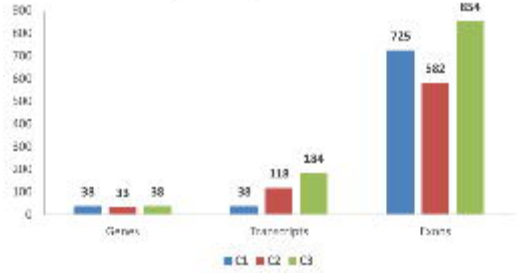


C.

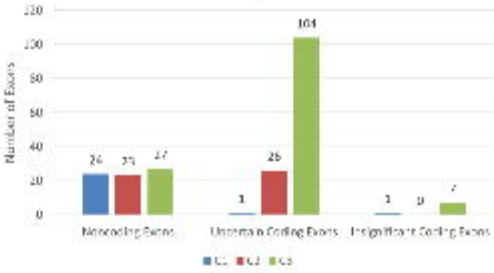
Exon Designation (in genes with a clinical validity of at least Moderate)	Evidence Used for Designation
Clinically Significant Exon	• Evidence that exon is required for biological function (e.g. pathogenic variants) AND/OR • No alternative splicing
Exon of Uncertain Significance	• No expression data in the literature AND • Alternative splicing AND/OR • 1 high frequency LoF variant in gnomAD (>0.3%)
Clinically Insignificant Exon	• Literature confirms exon is not expressed in tissue of clinical relevance AND/OR • >1 high quality, high frequency LoF variant in gnomAD (>0.3%)

FIGURE 2

A.


Gene, transcript and exon counts
38
38
38
38
113
184
725
582
854
Genes
Transcripts
Exons
C1
C2
C3

B.


Uncertain or Insignificant Exons
Number of Exons
24
23
27
1
26
104
1
0
7
Noncoding Exons
Uncertain Coding Exons
Insignificant Coding Exons
C1
C2
C3

C.

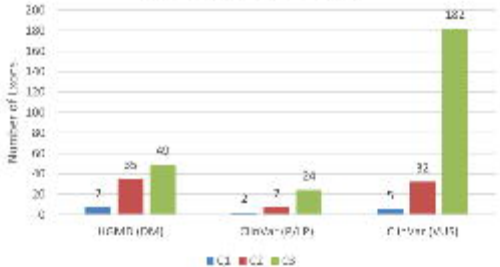

Impact on Interpretation
Number of Exons
7
35
49
2
7
24
5
32
182
C1
C2
C3

FIGURE 3

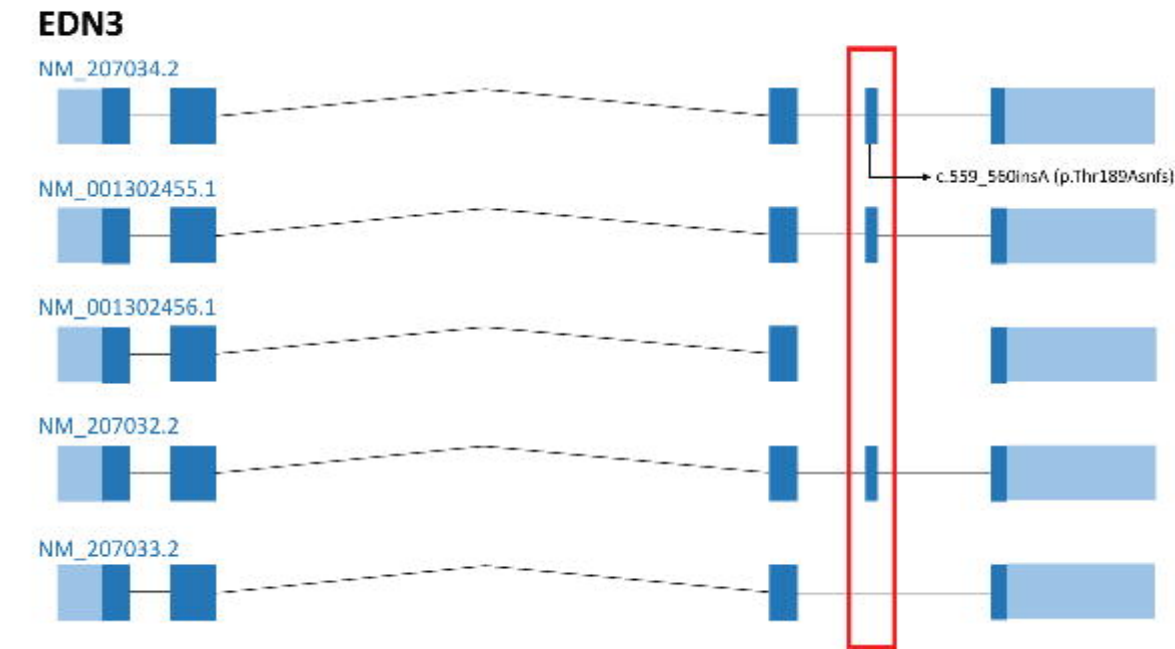
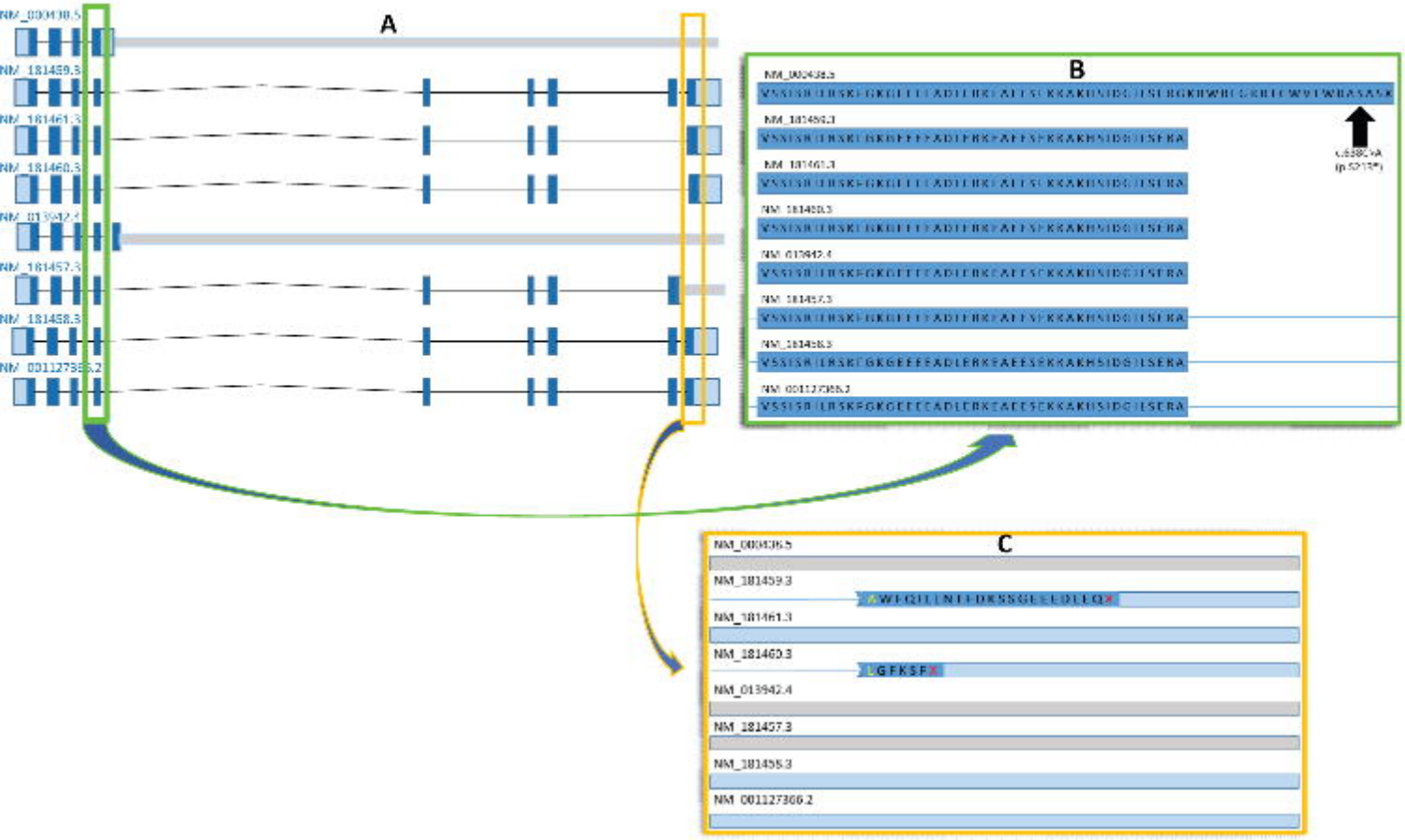

EDN3
NM_207034.2
NM_001302455.1
NM_001302456.1
NM_207032.2
NM_207033.2
c.559_560insA (p.Thr189Asnfs)

FIGURE 4


A
B
C
NM_000438.5
NM_181459.3
NM_181461.3
NM_181460.3
NM_013942.4
NM_181457.3
NM_181458.3
NM_001127366.2


Avgerage MQ ≥ 20 AND minimum DP ≥ 15
Average MQ < 20 OR minimum DP < 15
Average MQ < 20 AND min DP < 15
OTOA
Mapping Quality
Depth
Exons
KCNQ1
Mapping Quality
Depth
Exons
STRC
Mapping Quality
Depth
Exons