Systematic Evaluation of Variants Linked to Hearing Loss using Minor Allele Frequency and Prediction Tools

Joonsuk Lee

Department of Medical Science

The Graduate School, Yonsei University

Systematic Evaluation of Variants Linked to Hearing Loss using Minor Allele Frequency and Prediction Tools

Joonsuk Lee

Department of Medical Science

The Graduate School, Yonsei University

Systematic Evaluation of Variants Linked to Hearing Loss using Minor Allele Frequency and Prediction Tools

Directed by Professor Heon Yung Gee

The Master's Thesis

submitted to the Department of Medical Science,

The Graduate School of Yonsei University

in partial fulfillment of the requirements for the degree of

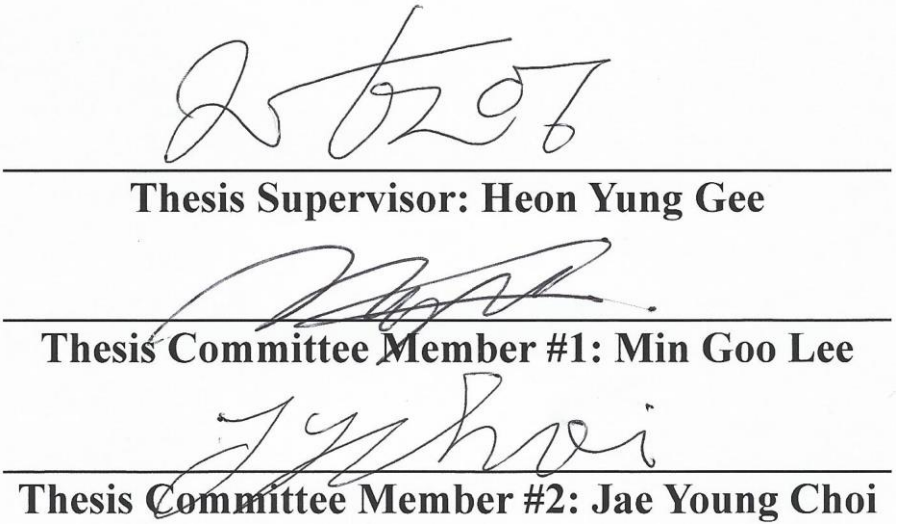
Master of Medical Science

Joonsuk Lee

December 2017

연세대학교
YONSEI UNIVERSITY

This certifies that The Master's Thesis of Joonsuk Lee is approved.



Thesis Supervisor: Heon Yung Gee

Thesis Committee Member #1: Min Goo Lee

Thesis Committee Member #2: Jae Young Choi

The Graduate School

Yonsei University

December 2017

ACKNOWLEDGEMENTS

석사 학위 졸업을 앞두며, 문득 지난 2 년 간 에비슨 2 층 휴게실 소파를 침대 삼아 잠을 청했던 날 들이 생각납니다. 학위를 시작할 때 '적어도 내가 하고 있는 것에 관해선 어딜 가더라도 부끄럽지 않은 사람이 되자' 는 각오로 학업에 임했고, 이를 위해서 그 동안 제가 할 수 있는 모든 노력을 다 했던 시간이었다고 생각합니다. 그 간의 정성을 논문으로 결실을 맺기 위해 정말 많은 분들께 큰 도움을 받았습니다. 무엇보다, 학위 과정 동안 저를 지도해주신 지헌영 교수님께 진심으로 감사 드립니다. 인식한 문제를 과학적으로 접근하고 해결해 나가는 방법과, 그 과정에서 요구되는 학생으로서의 마음가짐과 자세를 배울 수 있었습니다. 이는 앞으로 험난한 인생을 슬기롭게 헤쳐나가는 강력한 무기가 되어 줄 것이라 생각합니다. 그리고 저의 심사위원장이셨던 이민구 교수님께도 정말 감사 드립니다. 학위 시작 때부터 관심 있게 지켜 보아주시고, 더욱 발전할 수 있도록 아낌 없는 조언을 주신 덕분에 연구

과정에서 미처 생각하지 못 했던 점들을 보완할 수 있었습니다. 또한, 최재영 교수님, 정진세 교수님과 공동 연구를 하며 정말 즐겁게 연구할 기회를 가질 수 있었습니다. 그래서 두 교수님과 함께 연구 할 수 있었던 것이 저에게는 정말 큰 행운이었다고 생각합니다.

그리고 같은 연구실 일원으로서 가장 오랜 기간 함께 고생해 온 인정 많은 조경지 선생님, 책임감 있고 의리 있는 승부사 영익이, 매사 긍정적이고 배려심 많은 세영이, 같은 남자로서 배울 점도 많고, 부러운 점이 참 많은 듬직한 요준이, 유쾌한 웃음으로 주변 사람들의 기분을 밝게 만들어주는 에이스 정훈이, 그리고 좋은 이야기와 힘이 되어주는 말들로 마음의 위로가 되어준 김혜연 선생님께 고마움과 감사의 마음을 전합니다. 특히, 불안함과 초조함, 그리고 무언가에 쫓기는 듯한 느낌에 빠져있을 때, 커피 한 잔 건네며 기운을 북 돋아준 소중한 친구이자 한 해 연구실 Chief 로 고생해 온 지윤이를 비롯해서 같은 공간에서 서로를 도와주며 함께 동고동락해 온 김연정 선생님, 노신혜 선생님, 신동훈 선생님, 최성경 선생님, 강정민 선생님, 박학 선생님, 지훈이, 소원이,

그리고 이가형 선생님께 진심으로 감사의 인사를 전합니다. 그 밖에 지금은 같이 있지 않지만 늦은 밤 매일 같이 저녁을 먹으며 함께 연구로 밤을 지샜던 전익현 선생님과, 관심 있는 연구를 주제 삼아 시간 가는 줄 모르게 깊게 이야기를 나누었던 도현이, 부족한 나를 보고 끝까지 잘 따라와 준 이해심 많은 혜지, 안부를 건네며 먼 송도에서 가끔씩 얼굴을 보러 와준 요한이 모두에게 고마움을 전합니다. 끝으로 사랑하는 우리 아버지, 어머니, 큰 누나, 작은 누나, 형님들, 조카, 그리고 할머니에게 진심으로 감사 드립니다.

새벽 어두운 복도 끝 휴게실과 메케한 샤워실 냄새, 늦은 밤 군데 군데 환하게 켜진 에비슨의 연구실들, 나의 밥집 에비슨 식당, 지윤이와 함께 카페로 내려가는 그 짧은 시간의 기분과 소소한 행복, 그렇게 에비슨의 빛과 함께 했던 지난 2 년 간의 추억을 뒤로 하며, 미처 기억하지 못해 이 지면으로 마음을 표현하지 못한 모든 분들에게 다시 한 번 감사 드립니다. 자랑스러운 가족, 제자, 동료 그리고 친구가 될 수 있도록 앞으로 남은 인생도 후회 없이 최선을 다해 살아가겠습니다.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Systematic Evaluation of Variants Linked to Hearing Loss using Minor Allele Frequency and Prediction Tools

Joonsuk Lee

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Heon Yung Gee)

Non-syndromic hearing loss (NSHL) is extremely genetically heterogeneous, and to date, more than 96 genes have been linked to NSHL and explain about half of the clinical cases. Although high throughput DNA sequencing technology facilitates the identification of causative mutations in many human diseases, hundreds or thousands of variants identified by this method require interpretation to assess their likelihood of causing a disease. Here, we aim to systemically evaluate variants in 96 genes, which have been identified in NSHL patients, using minor allele frequency (MAF) and predictive tools. The MAF thresholds were determined considering allele frequency of the most common pathogenic variant of GJB2, and the prevalence of NSHL. For

the 96 NSHL known genes, 3,082 variants reported in HGMD and 1,210 reported as pathogenic or likely pathogenic in ClinVar were classified according to the MAF threshold and then according to the pLI scores of corresponding genes into three categories (pLI<0.1, 0.1<pLI<0.9, pLI>0.9). The number of missense variants reported in recessive (rec), dominant (dom) and dom/rec genes was 1,040, 244, and 668 respectively. The prediction scores of the missense variants were obtained using PolyPhen-2, SIFT, and Condel. As a result of analysis, the variants above the MAF threshold were 61, 23 and 14 in recessive, dominant and dom/rec genes, respectively. Using Korean control dataset, three variants that would be found more frequently in Koreans than in any other population were identified suggesting that several variants having MAF levels which are implausible for highly penetrance Mendelian disease could be found through other certain population control datasets. Additionally, there were statistical differences in prediction scores between the variants below and above the MAF threshold in recessive genes. Although prediction scores were not different between the variants below and above the MAF threshold for dominant genes, the scores were significantly different for dominant genes with > 0.9 pLI score. These data showed that prediction tools could be more useful for predicting variants in recessive genes and dominant genes with > 0.9 pLI score. Based on this study, we can prioritize novel candidate variants that have a causal relationship with the disease by using the MAF threshold and the prediction tool to evaluate variants in NSHL.

Key words: nonsyndromic hearing loss, minor allele frequency, prediction tool, pLI

Systematic Evaluation of Variants Linked to Hearing Loss using Minor Allele Frequency and Prediction Tools

Joonsuk Lee

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Heon Yung Gee)

I. INTRODUCTION

Hearing loss is a common sensory disorder that affects approximately one in every 500 newborns worldwide. At least 60% to 80% of hearing loss cases are hereditary, and over two-thirds are a nonsyndromic hearing loss (NSHL). The deafness phenotype includes more than 100 genes showing various patterns. Approximately 70% to 80% of them are estimated to cause NSHL in an autosomal recessive (AR) fashion.[1]

Although high-throughput DNA sequencing technology facilitates the

identification of causative mutations in many human diseases, the hundreds or thousands of variants identified with this method require interpretation to assess their likelihood of causing a disease. When researching genetic disorders, examining variant frequency is essential for finding candidates. The filtering efficiency is dependent on the number of control samples and racial diversity.[2,3] The amount of publicly annotated variant information has significantly increased with improved sequencing technology. At the same time, inaccurate information that has not been sufficiently verified has been also expanded. This means that variants reported as pathogenic in previous research may actually be benign.

A previous study used the prevalence of hearing loss to determine the maximum minor allele frequency (MAF) threshold as a classification criterion for pathogenic variants.[4] Although the established MAF threshold was fairly well-matched to "empirical" in classifying mutation pathogenicity, the theoretical rationale to establish the MAF threshold was unclear. This makes it difficult to directly apply this standard to the clinic. Moreover, the allele frequency estimates based on low allele counts were both upward-biased and imprecise because the sample size and population diversity of the control datasets were not sufficient to evaluate candidate variants using the allele frequencies.

Here, we systemically evaluated variants in 96 genes identified in NSHL patients using Exome Sequencing Project (ESP) and 1000 Genome Project (1000G) data, as well as Exome Aggregation Consortium (ExAC), which contains nearly as much

exome data as the control dataset. We also used single nucleotide variants (SNVs) genomic data of 397 Koreans to identify variants found more frequently in Korean subjects than other ethnicities and reclassified their pathogenicity. Finally, we confirmed the usefulness of various in-silico tools that classify mutations as deleterious or neutral by examining differences in prediction scores between variants classified by MAF.[5]

II. MATERIALS AND METHODS

1. Collecting pathogenic variant data for 96 NSHL genes

We selected 96 genes reported as causing NSHL on The Hereditary Hearing Loss Homepage (http://hereditaryhearingloss.org/) (**Table 1**). Next, we obtained variants and annotated information from The Human Gene Mutation Database (HGMD) and ClinVar database provided by the National Center for Biotechnology Information (NCBI). The ClinVar variants were filtered for mutations other than those reported as pathogenic or likely pathogenic.

2. Rationale for setting the MAF threshold of NSHL

There is no single mutation representing the majority of dominant NSHL genes in a given population.[4] Therefore, the maximum MAF threshold was obtained through the Hardy-Weinberg equilibrium based on the prevalence of hereditary NSHL.

In the case of recessive genes, we used the following proposed formula suggested by Whiffin et al.[6]

$$\textit{maximum credible population AF} = \sqrt{(\textit{prevalence})} \times \textit{maximum allelic contribution} \times \sqrt{(\textit{maximum genetic contribution})} \times 1/\sqrt{(\textit{penetrance})}$$

Based on the results of a large-scale genetic study of NSHL including up to 1,119 individuals,[7] we can assume that no newly identified variant will be more common. In the present cohort, 21.59% (95/440) of all NSHL patients with known genes had disease-causing variants in *GJB2*, and the c.35delG (p.Gly12Valfs*2) variant is estimated to account for 37.89% (72/190) of variant *GJB2* alleles. Finally, we assumed a penetrance of 1 as the phenotype of recessive genes has nearly 100% penetrance.

Table 1. 96 known NSHL genes examined in this study

#	Gene symbol	Gene Name	Accession #	MIM phenotype #	Mode	Repo-rted Variants
1	*ACTG1*	actin gamma 1	NM_001199954.1	102560	AD	29
2	*ADCY1*	adenylate cyclase 1	NM_021116.2	103072	AR	1
3	*BDP1*	B double prime 1, subunit of RNA polymerase III transcription initiation factor IIIB	NM_018429.2	607012	AR	1
4	*BSND*	barttin CLCNK type accessory beta subunit	NM_057176.2	606412	AR	18
5	*CABP2*	calcium binding protein 2	NM_016366.2	607314	AR	6
6	*CCDC50*	coiled-coil domain containing 50	NM_178335.2	611051	AD	3
7	*CD164*	CD164 molecule	NM_006016.4	603356	AD	1
8	*CDC14A*	cell division cycle 14A	NM_003672.3	603504	AR	2
9	*CDH23*	cadherin related 23	NM_022124.5	605516	AR	311
10	*CEACAM16*	carcinoembryonic antigen related cell adhesion molecule 16	NM_001039213.3	614591	AD	4
11	*CIB2*	calcium and integrin binding family member 2	NM_006383.3	605564	AR	12
12	*CLDN14*	claudin 14	NM_144492.2	605608	AR	11
13	*CLIC5*	chloride intracellular channel 5	NM_001114086.1	607293	AR	1
14	*COCH*	cochlin	NM_004086.2	603196	AD	30

15	*COL11A2*	collagen type XI alpha 2 chain	NM_080680.2	120290	AD/AR	18
16	*COL4A6*	collagen type IV alpha 6 chain	NM_001847.3	303631	XR	2
17	*CRYM*	crystallin mu	NM_001888.4	123740	AD	3
18	*DCDC2*	doublecortin domain-containing protein 2	NM_001195610.1	605755	AR	1
19	*DFNA5*	DFNA5, deafness associated tumor suppressor	NM_004403.2	608798	AD	6
20	*DFNB59*	deafness, autosomal recessive 59	NM_001042702.3	610219	AR	19
21	*DIABLO*	diablo IAP-binding mitochondrial protein	NM_019887.5	605219	AD	2
22	*DIAPH1*	diaphanous related formin 1	NM_005219.4	602121	AD	7
23	*DIAPH3*	diaphanous related formin 3	NM_001042517.1	614567	AD	5
24	*ELMOD3*	ELMO domain containing 3	NM_001135021.1	615427	AR	1
25	*EPS8*	epidermal growth factor receptor pathway substrate 8	NM_004447.5	600206	AR	4
26	*EPS8L2*	EPS8 like 2	NM_022772.3	614988	AR	1
27	*ESPN*	espin	NM_031475.2	606351	AD/AR	11
28	*ESRRB*	estrogen related receptor beta	NM_004452.3	602167	AR	20
29	*EYA4*	EYA transcriptional coactivator and phosphatase 4	NM_004100.4	603550	AD	23
30	*FAM65B*	family with sequence similarity 65 member B	NM_014722.3	611410	AR	1
31	*GIPC3*	GIPC PDZ domain containing family member 3	NM_133261.2	608792	AR	19
32	*GJB2*	gap junction protein beta 2	NM_004004.5	121011	AD/AR	387
33	*GJB3*	gap junction protein beta 3	NM_024009.2	603324	AD/AR	21

34	*GJB6*	gap junction protein beta 6	NM_001110219.2	604418	AD/AR	19
35	*GPSM2*	G-protein signaling modulator 2	NM_013296.4	609245	AR	3
36	*GRHL2*	grainyhead like transcription factor 2	NM_024915.3	608576	AD	3
37	*GRXCR1*	glutaredoxin and cysteine rich domain containing 1	NM_001080476.2	613283	AR	10
38	*GRXCR2*	glutaredoxin and cysteine rich domain containing 2	NM_001080516.1	615762	AR	1
39	*HGF*	hepatocyte growth factor	NM_000601.4	142409	AR	0
40	*HOMER2*	homer scaffolding protein 2	NM_199330.2	604799	AD	1
41	*ILDR1*	immunoglobulin like domain containing receptor 1	NM_001199799.1	609739	AR	21
42	*KARS*	lysyl-tRNA synthetase	NM_001130089.1	601421	AR	3
43	*KCNQ4*	potassium voltage-gated channel subfamily Q member 4	NM_004700.3	603537	AD	36
44	*KITLG*	KIT ligand	NM_000899.4	184745	AD	3
45	*LHFPL5*	lipoma HMGIC fusion partner-like 5	NM_182548.3	609427	AR	9
46	*LOXHD1*	lipoxygenase homology domains 1	NM_144612.6	613072	AR	31
47	*LRTOMT*	leucine rich transmembrane and O-methyltransferase domain containing	NM_001145309.3	612414	AR	17
48	*MARVELD2*	MARVEL domain containing 2	NM_001038603.2	610572	AR	14
49	*MCM2*	minichromosome maintenance complex component 2	NM_004526.3	116945	AD	1
50	*MET*	MET proto-oncogene, receptor tyrosine kinase	NM_001127500.2	164860	AR	1
51	*MIR96*	microRNA 96	NR_029512.1	611606	AD	8
52	*MSRB3*	methionine sulfoxide reductase B3	NM_001193460.1	613719	AR	4

53	*MYH14*	myosin heavy chain 14	NM_001145809.1	608568	AD	21
54	*MYH9*	myosin heavy chain 9	NM_002473.5	160775	AD	17
55	*MYO15A*	myosin XVA	NM_016239.3	602666	AR	223
56	*MYO3A*	myosin IIIA	NM_017433.4	606808	AR	19
57	*MYO6*	myosin VI	NM_004999.3	600970	AD/AR	53
58	*MYO7A*	myosin VIIA	NM_000260.3	276903	AD/AR	428
59	*NARS2*	asparaginyl-tRNA synthetase 2, mitochondrial	NM_024678.5	612803	AR	6
60	*OSBPL2*	oxysterol binding protein like 2	NM_144498.2	606731	AD	3
61	*OTOA*	otoancorin	NM_144672.3	607038	AR	22
62	*OTOF*	otoferlin	NM_194248.2	603681	AR	183
63	*OTOG*	otogelin	NM_001277269.1	604487	AR	7
64	*OTOGL*	otogelin like	NM_173591.3	614925	AR	20
65	*P2RX2*	purinergic receptor P2X 2	NM_174873.2	600844	AD	4
66	*PCDH15*	protocadherin related 15	NM_001142769.1	605514	AR	117
67	*PNPT1*	polyribonucleotide nucleotidyltransferase 1	NM_033109.4	610316	AR	13
68	*POU3F4*	POU class 3 homeobox 4	NM_000307.4	300039	XR	52
69	*POU4F3*	POU class 4 homeobox 3	NM_002700.2	602460	AD	15
70	*PRPS1*	phosphoribosyl pyrophosphate synthetase 1	NM_002764.3	311850	XL	17
71	*PTPRQ*	protein tyrosine phosphatase, receptor type Q	NM_001145026.1	603317	AR	1
72	*RDX*	radixin	NM_001260492.1	179410	AR	9

73	*S1PR2*	sphingosine-1-phosphate receptor 2	NM_004230.3	605111	AR	2
74	*SERPINB6*	serpin family B member 6	NM_001195291.2	173321	AR	3
75	*SIX1*	SIX homeobox 1	NM_005982.3	601205	AD	2
76	*SLC17A8*	solute carrier family 17 member 8	NM_139319.2	607557	AD	5
77	*SLC22A4*	solute carrier family 22 member 4	NM_003059.2	604190	AR	11
78	*SLC26A4*	solute carrier family 26 member 4	NM_000441.1	605646	AR	538
79	*SLC26A5*	solute carrier family 26 member 5	NM_198999.2	604943	AR	6
80	*SMPX*	small muscle protein, X-linked	NR_045617.1	300226	XD	8
81	*STRC*	stereocilin	NM_153700.2	606440	AR	55
82	*SYNE4*	spectrin repeat containing nuclear envelope family member 4	NM_001039876.2	615535	AR	3
83	*TBC1D24*	TBC1 domain family member 24	NM_001199107.1	613577	AD/AR	29
84	*TECTA*	tectorin alpha	NM_005422.2	602574	AD/AR	98
85	*TJP2*	tight junction protein 2	NM_004817.3	607709	AD	6
86	*TMC1*	transmembrane channel like 1	NM_138691.2	606706	AD/AR	91
87	*TMEM132E*	transmembrane protein 132E	NM_001304438.1	616178	AR	1
88	*TMIE*	transmembrane inner ear	NM_147196.2	607237	AR	15
89	*TMPRSS3*	transmembrane protease, serine 3	NM_024022.2	605511	AR	63
90	*TNC*	tenascin C	NM_002160.3	187380	AD	3
91	*TPRN*	taperin	NM_001128228.2	613354	AR	10

92	*TRIOBP*	TRIO and F-actin binding protein	NM_001039141.2	609761	AR	37
93	*TSPEAR*	thrombospondin type laminin G domain and EAR repeats	NM_001272037.1	612920	AR	3
94	*USH1C*	USH1 protein network component harmonin	NM_153676.3	605242	AR	46
95	*WFS1*	wolframin ER transmembrane glycoprotein	NM_001145853.1	606201	AD	93
96	*WHRN*	whirlin	NM_001083885.2	607928	AR	26

3. Control dataset

The following three control datasets were used: (1) The NHLBI Go Exome Sequencing Project (http://evs.gs.washington.edu/EVS/) (2) The 1000 Genomes Project (http://www.internationalgenome.org/), and (3) The ExAC, http://exac.broadinstitute.org/). The ESP database contains information of allele frequency in European Americans (EA) and African Americans (AA) from 4,300 and 2,203 individuals, respectively. The 1000 Genomes Phase 3 database consists of variants on 2,504 individuals with whole exome sequencing (WES) and whole genome sequencing (WGS) data from individuals with African ancestry (661), Americans (347), East Asian ancestry (504), European ancestry (503), and South Asian ancestry (489). ExAC has 60,706 WES datasets divided into seven groups of African/African American (AFR, 5,203), Latino (AMR, 5,789), East Asian (EAS, 4,327), Finnish (FIN, 3,307), Non-Finnish European (NFE, 33,370), South Asian (SAS, 8,256), and other (OTH, 454).

4. Predicting the functional impact of variants

Variant annotation was performed using Variant Effect Predictor (VEP) version 89 with GRCh37. VEP is used to determine gene symbols and NCBI Reference Sequences (RefSeq) for each functional consequence of the variant, and the PolyPhen-2 (PP2), Sorting Intolerant from Tolerant (SIFT), Consensus

deleteriousness of non-synonymous single nucleotide variants (Condel), and Combined Annotation Dependent Depletion (CADD) scores.[8-11] We separated 28 dominant genes among 96 NSHL genes into loss-of-function (LoF) intolerant (pLI ≥ 0.9) or LoF tolerant (pLI ≤ 0.1) categories using the probability of being LoF intolerant (pLI) reported previously.[3] Conversely, variants in recessive genes only cause disease when they are homozygous. Recessive genes were therefore classified using the probability of being intolerant of homozygous, but not heterozygous LoF variants (pRec) and the probability of being tolerant of both heterozygous and homozygous LoF variants (pNull) from intolerant (pLI, pRec > 0.9) to tolerant (pRec < 0.1, pNull > 0.9). The Functional Gene Constraint Scores for 96 NSHL genes were obtained from the ExAC download page (http://exac.broadinstitute.org/downloads).

5. Korean control dataset

To identify variants that are more frequently in Koreans than in other populations, we used 397 WGS sets from The National Biobank of Korea, Centers for Disease Control and Prevention (NBK) as a control Korean dataset. Derived MAF data for 19,368,798 SNVs of the NBK control data were obtained using VCFtools.[12]

III. RESULTS

1. Pathogenic variants of 96 NSHL genes reported in public databases

For the 96 NSHL genes, a total of 3,550 variants were reported in the HGMD and ClinVar databases. They reported 3,082 and 1,210 variants, respectively, with 742 variants in both databases. Missense variants were the most common (1,850). When classified by inheritance mode, there were 334 autosomal dominant, 1,982 autosomal recessive, 1,155 autosomal dom/rec, and 79 X-linked variants (**Table 2**). Among all the variants, 1,247 (35.13%) were reported in all three control datasets (**Figure 1**).

2. Calculating the MAF threshold for NSHL genes

No single mutation accounts for the majority of autosomal dominant NSHL in any given population.[4] Therefore, to set the MAF threshold of the 28 dominant genes, we assumed that the frequency of a single allele including a certain variant causing NSHL is not higher than the prevalence of total hearing loss. As a result, an MAF threshold of 0.1% was obtained through Hardy-Weinberg equilibrium.

For the 55 recessive genes, 95 of 440 individuals with known genes were identified as *GJB2* (95/440, 21.59%) according to a study by Soloan-Heggen et al. Among the 190 alleles of 95 individuals solved with *GJB2*, the most frequent variant

was in 72 alleles with c.35delG (72/190, 37.89%). Therefore, the maximum MAF threshold for this allele was $\sqrt{(0.002 \times 0.8 \times 0.7)} \times 0.3789 \times \sqrt{0.2159} \times 1 \leq 0.6\%$ as described in the Materials and Methods.

Although this threshold was based on the most prevalent known pathogenic variant c.35delG of *GJB2* from the study, the MAF threshold was lower than the Finnish and the European MAF of the ExAC control database for the variant. The reason for this is that while half of the 1,119 people with hearing loss recruited from the reference literature were Caucasian, the other half consisted of various races. Thus, the actual rate of pathogenic alleles most frequently found in a particular race may be higher than the proportion of pathogenic alleles that account for the largest portion of the hearing loss individuals referred to the literature.[7] For this reason, it is desirable to limit the application of this threshold to recessive genes where pathogenic alleles are frequently found in certain population, such as *GJB2* and *SLC26A4*. Therefore, the thresholds were applied to the 94 NSHL genes except for these two genes to filter variants.

Table 2. Consequence of Reported Deafness Variants according to inheritance mode

Consequence	All	AD	AR	AD/AR	X-linked
Missense_variant	1,969	244	1,015	668	42
Frameshift_variant	583	35	332	203	13
Stop_gained	448	17	278	141	12
Splice_region_variant	385	11	279	93	2
Inframe_deletion	67	12	22	31	2
Intron_variant	31	3	22	6	0
5_prime_UTR_variant	14	1	10	3	0
Non_coding_transcript_variant (ncRNA)	14	8	0	0	6
Inframe_insertion	10	0	6	4	0
Start_lost	10	0	8	2	0
Protein_altering_variant	6	2	1	3	0
Stop_lost	6	1	3	0	2
Upstream_gene_variant	6	0	5	1	0
3_prime_UTR_variant	1	0	1	0	0
SUM	3,550	334	1,982	1,155	79

Abbreviation: AD, autosomal dominant; AR, autosomal recessive; AD/AR, genes with both AD and AR inheritance; X-linked, x chromosome linked gene

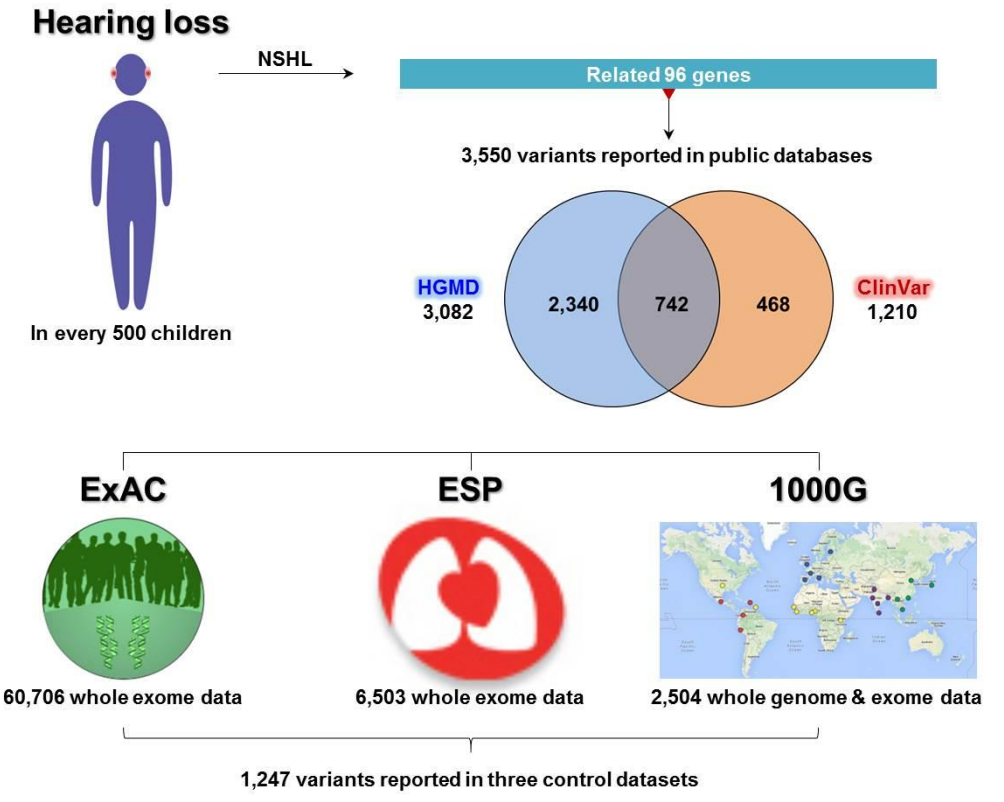


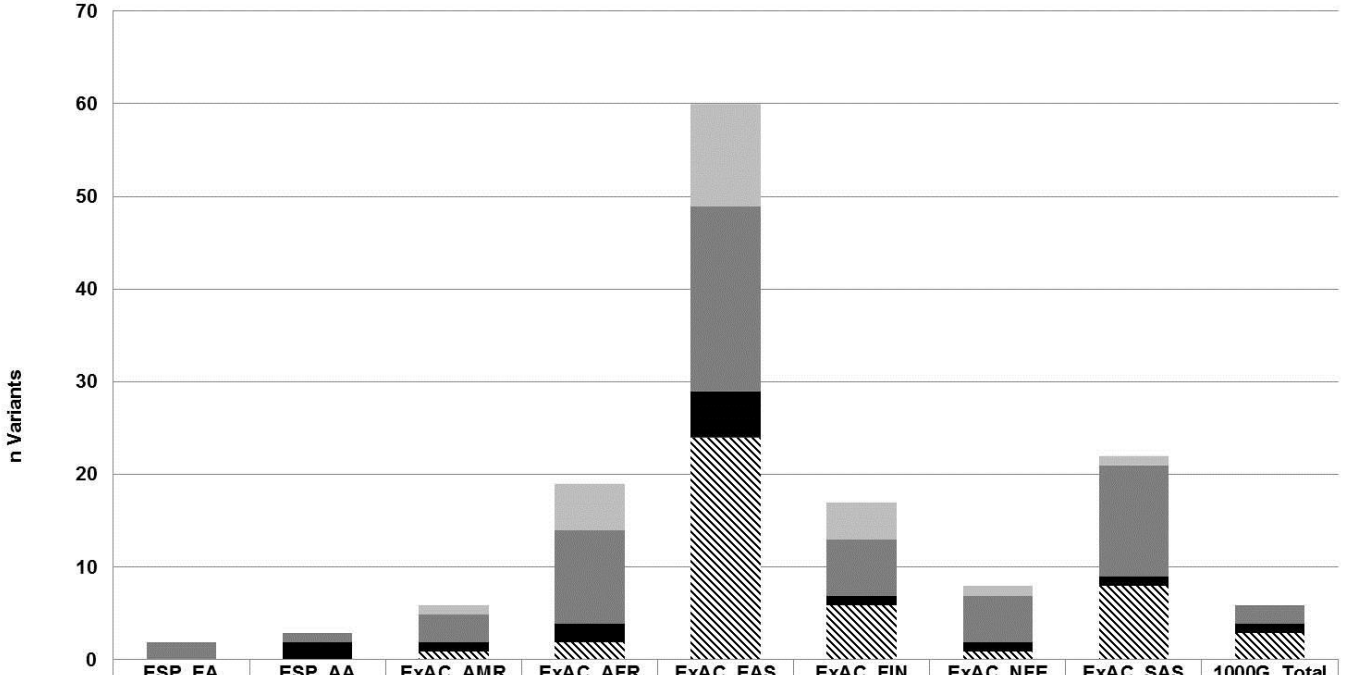
Figure 1. Deafness variants reported in public mutation databases. The variants reported in the Human Gene Mutation Database (HGMD) and ClinVar were 3,082, and 1,210, respectively. The remaining variants in ClinVar were those reported as likely pathogenic or pathogenic. Among the 3,550 variants, 1,247 were reported in EVS, ExAC, and 1000G control datasets.

3. Control dataset variants

Of the 1,247 variants reported in the control datasets, there were 104, 778, and 365 variants in the dominant, recessive, and dom/rec genes, respectively. We attempted to filter 1,247 variants except for the *GJB2* and *SLC26A4* genes according to the MAF threshold using the highest allele frequency in any one population (POPMAX_MAF) of each control dataset. When filtering with POPMAX_MAF, we noted that MAF data in each population of 1000G and 'OTHER' (OTH) population in ExAC exceeded the cut-off value in dominant genes even if the variants were found in only two alleles. In the case of recessive genes, the cut-off value of 0.006 was exceeded if at least five alleles with variants in the American population control data of 1000G were found. That is, when the total number of alleles used in the control dataset was low, the results of AF reported in the control dataset tended to be higher, resulting in an incorrect AF prediction value. Therefore, when using the MAF data for the OTH population in ExAC and the population-specific control data in 1000G for variant filtering, it would be more accurate to filter through the total MAF data. Also, the total adjusted AF data in ExAC (ExAC_Adj) that only include individuals with genotype quality (GQ) ≥ 20 and depth (DP) ≥ 10 was used in the variant that failed to pass filter criteria, that is VQSLOD (variant quality score log-odds) ≥ -2.632.

After classifying the variants according to cut-off values, there were 23 variants with an MAF > 0.1% in dominant genes and 61 variants with an MAF > 0.6% in

recessive genes. For dom/rec genes, variants with an MAF > 0.6% can be filtered regardless of inheritance mode, and variants with an MAF between 0.1% and 0.6% will not be causal variants unless they are recessive. Therefore, we regarded variants of dom/rec genes between MAF 0.1% and 0.6% as recessive variants. In the dom/rec case, there were 14 variants with an MAF > 0.6% (**Figure 2**). In particular, among the variants above the MAF threshold, the largest variants were filtered in East Asian subjects. This indicates that all variants associated with hearing loss are plausible, including those with lower causality, because many of the variants have only been studied in European and American population.



	ESP_EA (4,300)	ESP_AA (2,203)	ExAC_AMR (5,789)	ExAC_AFR (5,203)	ExAC_EAS (4,327)	ExAC_FIN (3,307)	ExAC_NFE (33,370)	ExAC_SAS (8,256)	1000G_Total (2,504)
Dom	0	0	1	5	11	4	1	1	0
Rec	2	1	3	10	20	6	5	12	2
Dom/Rec	0	2	1	2	5	1	1	1	1
Unique	0	0	1	2	24	6	1	8	3

Figure 2. Contribution of Three Different Control Datasets to Reclassification of 98 Reported Variants. The plot for each population is the number of variants found in the highest allele frequency in any one population above MAF threshold. Numbers of individuals per population are listed in parenthesis on the x axis. Light grey, number of variants above the MAF threshold in dominant genes (Dom); Dark grey, number of variants above the MAF threshold in recessive genes (Rec); Black, number of variants above the MAF threshold in Dom/Rec genes (Dom/Rec); Diagonal pattern, number of variants above the MAF threshold unique to that single population (Unique).

All three control datasets included variants that exceeded the MAF threshold in the ExAC and 1000G datasets, and all five variants filtered by ESP were also filtered by ExAC. On the other hand, the two variants filtered by 1000G failed to pass the above-described filter criteria in ExAC. These two variants could be considered as examples of WGS identifying a variant not found in WES, which is probably related to the hybridization/capture and polymerase chain reaction amplification steps required to prepare sequencing libraries with heterogeneous coverage.[13,14] In conclusion, when filtering newly discovered candidate variants using a population-specific control dataset, it is desirable to use a combination of exome and genome information as a control dataset, and it is sufficient to use ExAC as the control dataset for exome information. Based on this approach, the 98 variants above the MAF thresholds are most unlikely to be causal variants of NSHL. Information about these variants is listed in **Table 3**.

On the other hand, there were exceptions in 98 variants above the MAF thresholds which were expected to be benign. *MYO15A* (c.5925G>A; p.Trp1975Ter) and *OTOF* (c.5098G>C; p.Glu1700Gln) were classified as pathogenic by ACMG guideline, even though the POPMAX_MAFs of the variants in *MYO15A* and *OTOF* were 0.017067 in South Asian and 0.00743 in East Asian respectively. Thus, pathogenicity of variant cannot be determined by MAF thresholds only. But the only certain thing is that the group above the threshold is more likely to have a lower causality than the group below, which can be the key evidence for prioritizing the causality of the hearing loss

variants. Additionally, the variants of *MYO15A* and *OTOF* were reported from Iranian and Taiwanese hearing impaired patients respectively, which consistent with POPMAX in the ExAC control database.[15,16] This means that it is possible to investigate the causal variants frequently found in hearing-impaired patients by population. Therefore, it is possible to present a hearing loss gene responsible for major cause in a certain population by comparing the MAFs of disease-causing variant among races.

Table 3. 98 Reported NSHL variants present in controls above MAF thresholds

Gene Symbol	Nucleotide Change	Amino acid Change	Consequence	dbSNP 147	CNTL POP MAX	POP-MAX MAF	HG MD	Clin-Var	PP2 humvar	SIFT	Condel	CADD
DOMINANT GENES												
GJB3	c.94C>T	p.Arg32Trp	missense variant	rs1805063	ExAC FIN	0.0543	DM?	B, LB	Dam (1)	Del (0)	Del (0.945)	27.5
GJB3	c.529T>G	p.Tyr177Asp	missense variant	rs80297119	EVS AA	0.0189	DM?	B, LB	Dam (0.983)	Del (0)	Del (0.873)	24.5
GJB3	c.580G>A	p.Ala194Thr	missense variant	rs117385606	ExAC EAS	0.0072	DM?	LB, P	Benign (0.11)	Tol (1)	Neu (0.009)	15.62
KCNQ4	c.546C>G	p.Phe182Leu	missense variant	rs80358273	ExAC EAS	0.0045	DM	P	Benign (0.04)	Tol (1)	Neu (0.003)	19.93
KCNQ4	c.1365T>G	p.His455Gln	missense variant	rs34287852	ExAC FIN	0.3100	DP	B	Benign (0.005)	Tol (0.35)	Neu (0.021)	12.73
SLC17A8	c.1120G>T	p.Ala374Ser	missense variant	rs138307707	ExAC FIN	0.0089	DM	-	Dam (0.964)	Del (0)	Del (0.851)	29.6
MYH14	c.1150G>T	p.Gly384Cys	missense variant	rs119103280	ExAC NFE	0.0079	DM?	LB, P	Dam (0.977)	Del (0)	Del (0.863)	25.8
MYH14	c.1427G>A	p.Arg476His	missense variant	rs375694189	ExAC AFR	0.0021	DM	-	Dam (0.994)	Del (0)	Del (0.897)	34
MYH14	c.2921G>A	p.Arg974His	missense variant	rs113993956	ExAC AMR	0.0067	DM	P	Dam (0.892)	Del (0)	Del (0.796)	33
MYH14	c.4903G>A	p.Glu1635Lys	missense	rs140157424	ExAC	0.0030	DM	-	Dam (0.869)	Del (0.01)	Del (0.743)	34

			variant		EAS							
MYH9	c.5188C>T	p.Arg1730Cys	missense variant	rs201021615	ExAC EAS	0.0021	DM	VUS	Dam (0.953)	Del (0)	Del (0.841)	35
WFS1	c.353A>C	p.Asp118Ala	missense variant	rs71524349	ExAC AFR	0.0121	DM	LB	Dam (0.977)	Del (0)	Del (0.863)	24.5
WFS1	c.449C>T	p.Ala150Val	missense variant	rs113651985	ExAC EAS	0.0050	DM?	VUS	Dam (0.506)	Tol (0.17)	Neu (0.229)	26.3
WFS1	c.482G>A	p.Arg161Gln	missense variant	rs115346085	ExAC AFR	0.0159	DM?	-	Benign (0)	Tol (0.75)	Neu (0.002)	3.159
WFS1	c.577A>C	p.Lys193Gln	missense variant	rs41264699	ExAC SAS	0.0131	DM?	B, LB	Dam (0.46)	Tol (0.11)	Neu (0.422)	17.83
WFS1	c.1235T>C	p.Val412Ala	missense variant	rs144951440	ExAC EAS	0.0118	DM	LB	Dam (0.591)	Tol (0.05)	Del (0.549)	21.1
WFS1	c.2020G>A	p.Gly674Arg	missense variant	rs200672755	ExAC FIN	0.0014	DM	P	Dam (0.995)	Del (0)	Del (0.902)	24.4
WFS1	c.2195G>A	p.Arg732His	missense variant	rs149013740	ExAC AFR	0.0053	-	VUS, LP	Dam (0.995)	Del (0)	Del (0.902)	32
WFS1	c.2209G>A	p.Glu737Lys	missense variant	rs147834269	ExAC EAS	0.0129	DM	LB	Benign (0.099)	Del (0)	Neu (0.452)	25.5
WFS1	c.2335G>A	p.Val779Met	missense variant	rs141328044	ExAC AFR	0.0236	DM?	B, LB	Dam (0.685)	Tol (0.15)	Del (0.485)	24
WFS1	c.2611G>A	p.Val871Met	missense variant	rs71532874	ExAC FIN	0.0126	DM?	B, LB	Benign (0.019)	Tol (0.08)	Neu (0.301)	21.5
DIAPH1	c.2099T>A	p.Ile700Asn	missense variant	rs199830182	ExAC EAS	0.0011	DM?	-	Benign (0.247)	Tol (0.4)	Neu (0.041)	18.74
DIAPH1	c.2032C>T	p.Pro678Ser	missense variant	rs186370335	ExAC EAS	0.0119	DM?	LB	Dam (0.763)	Tol (0.14)	Del (0.532)	22.3

TJP2	c.334G>A	p.Ala112Thr	missense variant	rs144396411	ExAC EAS	0.0030	DM	-	Dam (1)	Del (0)	Del (0.945)	34
TJP2	c.2081G>A	p.Gly694Glu	missense variant	rs201366118	ExAC EAS	0.0027	DM	-	Dam (0.894)	Del (0)	Del (0.796)	25.7
TJP2	c.3562A>G	p.Thr1188Ala	missense variant	rs192802385	ExAC EAS	0.0020	DM	-	Dam (0.872)	Del_LC (0)	Del (0.784)	25.1
RECESSIVE GENES												
BSND	c.127G>A	p.Val43Ile	missense variant	rs34561376	ExAC EAS	0.1974	FP	B, LB	Benign (0)	Tol (0.44)	Neu (0.013)	0.045
SLC22A4			upstream_gene variant	rs3761661	1000G Total	0.1094	FP	No	No	No	No	5.459
SLC22A4	c.1046+5G>A		splice_region variant	rs2304081	ExAC EAS	0.2590	FP	No	No	No	No	12.04
MYO3A	c.4462A>G	p.Lys1488Glu	missense variant	rs34204285	ExAC EAS	0.0468	DM	VUS, B	Benign (0.039)	Tol_LC (0.05)	Neu (0.347)	22.7
PCDH15	c.4409+3011_4409+3013delAAC		intron_variant	rs113363047	EVS_AA	0.1097	DM?	B, P	No	No	No	8.44
PCDH15	c.4409+2222G>T		intron variant	rs148718874	ExAC EAS	0.0275	DM	VUS, B	No	No	No	9.546
PCDH15	c.4060C>A	p.Gln1354Lys	missense variant	rs61731387	ExAC AFR	0.0280	DM?	-	Dam (0.493)	Del (0)	Del (0.609)	27.5
PCDH15	c.3487G>A	p.Gly1163Arg	missense variant	rs149478475	ExAC EAS	0.0167	DM?	LB	Dam (0.994)	Del (0)	Del (0.897)	34
PCDH15	c.2920C>T	p.Arg974Cys	missense variant	rs201816080	ExAC EAS	0.0117	DM	VUS, B	Dam (0.84)	Tol (0.11)	Del (0.593)	26.6
PCDH15	c.1319A>C	p.Asp440Ala	missense variant splice_region	rs4935502	ExAC EAS	0.8434	DM	B, LB	Dam (0.492)	Del (0.01)	Del (0.570)	24.2

			variant								
CDH23	c.429+4G>A		splice_region variant intron variant	rs397517328	ExAC SAS	0.0184	DM?	-	No	No	No	16.35
CDH23	c.1096G>A	p.Ala366Thr	missense variant	rs143282422	ExAC NFE	0.0110	DM?	-	No	No	No	25.7
CDH23	c.1282G>A	p.Asp428Asn	missense variant	rs188376296	ExAC EAS	0.0066	DM?	-	No	No	No	25.1
CDH23	c.1423G>A	p.Val475Met	missense variant	rs62622410	ExAC AFR	0.0530	DM?	-	No	No	No	24.8
CDH23	c.2263C>T	p.His755Tyr	missense variant	rs181255269	ExAC SAS	0.0084	DM	-	No	No	No	23.3
CDH23	c.2568C>G	p.Ile856Met	missense variant	rs188498736	ExAC FIN	0.0621	DM?	VUS	No	No	No	23.5
CDH23	c.3074G>A	p.Gly1025Asp	missense variant	rs143179070	ExAC AFR	0.0149	DM	B	No	No	No	31
CDH23	c.3625A>G	p.Thr1209Ala	missense variant	rs41281314	ExAC AFR	0.1740	DM?	VUS, Not_provided. B, LB, P	No	No	No	23.5
CDH23	c.4858G>A	p.Val1620Met	missense variant	rs41281330	ExAC SAS	0.0598	DM?	B	No	No	No	29.1
CDH23	c.5418C>G	p.Asp1806Glu	missense variant	rs74145660	ExAC EAS	0.0962	DM?	B, LB	No	No	No	22.4

CDH23	c.5660C>T	p.Thr1887Ile	missense variant	rs397517340	ExAC SAS	0.0261	DM?	VUS, LB	No	No	No	20.2
CDH23	c.5753G>A	p.Arg1918Gln	missense variant	rs115113440	ExAC AFR	0.0096	DM	VUS	No	No	No	21.4
CDH23	c.6596T>A	p.Ile2199Asn	missense variant	rs111033494	ExAC AFR	0.0189	DM	B	No	No	No	15.74
CDH23	c.6847G>A	p.Val2283Ile	missense variant	rs41281334	ExAC AMR	0.1912	R	Not_provided, B, LB	No	No	No	4.949
CDH23	c.8120C>T	p.Pro2707Leu	missense variant	rs373230009	ExAC SAS	0.0123	DM	-	No	No	No	27.6
USH1C	c.388G>A	p.Val130Ile	missense variant splice_region variant	rs55843567	ExAC AFR	0.0462	DM?	VUS, B	Benign (0.022)	Tol (0.19)	Neu (0.051)	18.68
USH1C	c.307C>T	p.Arg103Cys	missense variant	rs397517880	ExAC SAS	0.0073	DM	VUS	Dam (1)	Del (0)	Del (0.945)	32
ESRRB	c.16A>G	p.Arg6Gly	missense variant	rs143477571	ExAC EAS	0.0435	DM?	B	Dam (0.997)	Del_LC (0)	Del (0.911)	22
ESRRB	c.1144C>T	p.Arg382Cys	missense variant	rs373131497	ExAC EAS	0.0097	DM	-	Dam (0.93)	Tol (0.07)	Del (0.696)	25.8
STRC	c.179T>C	p.Phe60Ser	missense variant	rs2729509	ExAC AFR	0.7655	R	Not_provided, B	Benign (0)	Tol (1)	Neu (0.000)	0.003
MYO15A	c.1783G>A	p.Ala595Thr	missense variant	rs2955365	ExAC SAS	0.7597	R	B, LB	Benign (0.186)	Tol_LC (0.06)	Neu (0.340)	23.1
MYO15A	c.2152T>G	p.Trp718Gly	missense variant	rs2955367	ExAC SAS	0.7607	R	B, LB	Benign (0.161)	Del_LC	Neu (0.398)	11.38

										(0.02)		
MYO15A	c.3026C>A	p.Pro1009His	missense variant	rs117612144	ExAC EAS	0.0460	DM	B	Benign (0.436)	Del_LC (0.01)	Del (0.547)	17.54
MYO15A	c.5287C>T	p.Arg1763Trp	missense variant	rs200146361	ExAC FIN	0.0064	DM?	VUS, B	Dam (1)	Del (0.01)	Del (0.905)	26.4
MYO15A	c.5925G>A	p.Trp1975Ter	stop_gained	rs375290498	ExAC SAS	0.0245	DM	VUS, P	No	No	No	37
MYO15A	c.6614C>T	p.Thr2205Ile	missense variant	rs121908970	ExAC FIN	0.0214	DM	B, P	Dam (0.94)	Del (0.02)	Del (0.768)	23.9
MYO15A	c.6796G>A	p.Val2266Met	missense variant	rs114274755	ExAC AFR	0.0308	DM?	B	Dam (0.981)	Del (0.01)	Del (0.831)	24.2
MYO15A	c.9478C>T	p.Leu3160Phe	missense variant	rs140029076	EVS EA	0.0099	DM?	B	Benign (0.277)	Del (0.04)	Neu (0.378)	22
MYO15A	c.10573A>G	p.Ser3525Gly	missense variant	rs182332665	ExAC SAS	0.0481	DM?	B	Dam (0.994)	Del (0)	Del (0.897)	23.7
LOXHD1	c.4526G>A	p.Gly1509Glu	missense variant	rs187587197	ExAC FIN	0.0496	DM	B	Dam (1)	Del (0)	Del (0.945)	33
LOXHD1	c.2825_2827delAGA	p.Lys942del	inframe_deletion	rs142960762	ExAC SAS	0.0413	DM	B	No	No	No	19.63
DFNB59	c.874G>A	p.Gly292Arg	missense variant	rs79399438	ExAC EAS	0.1528	DM?	B, LB	Benign (0.201)	Del_LC (0.01)	Neu (0.425)	23
OTOF	c.5098G>C	p.Glu1700Gln	missense variant	rs199766465	ExAC EAS	0.0079	DM	LP, P	Dam (0.856)	Del (0.02)	Del (0.714)	32
OTOF	c.5026C>T	p.Arg1676Cys	missense variant	rs139767460	ExAC EAS	0.0170	DM?	B	Dam (0.708)	Del (0.02)	Del (0.639)	28.7

OTOF	c.4023+1G>A		splice_donor variant	rs186810296	ExAC EAS	0.0110	DM?	VUS	No	No	No	26.3
OTOF	c.3751T>G	p.Cys1251Gly	missense variant	rs41288773	ExAC NFE	0.0254	DM?	B	Benign (0)	Tol (0.58)	Neu (0.005)	0.003
OTOF	c.3470G>A	p.Arg1157Gln	missense variant	rs56054534	EVS EA	0.0139	DM?	B	Dam (0.998)	Tol (0.91)	Del (0.474)	26.7
OTOF	c.2464C>T	p.Arg822Trp	missense variant	rs80356570	ExAC FIN	0.0294	DM?	B	Dam (0.619)	Del (0)	Del (0.660)	32
OTOF	c.367G>A	p.Gly123Ser	missense variant	rs116314622	ExAC SAS	0.0072	DM	-	Benign (0.044)	Tol (0.06)	Neu (0.325)	21.5
OTOF	c.158C>T	p.Ala53Val	missense variant	rs1879761	ExAC AMR	0.3016	DM?	B, LB	Benign (0.196)	Del (0.04)	Neu (0.367)	23.1
OTOF	c.157G>A	p.Ala53Thr	missense variant	rs144915302	ExAC EAS	0.0149	DM?	LB	Benign (0.178)	Del (0.05)	Neu (0.362)	23.3
OTOF	c.145C>T	p.Arg49Trp	missense variant	rs61746568	ExAC AMR	0.0384	DM	B	Dam (0.998)	Del (0)	Del (0.919)	34
CLDN14	c.11C>T	p.Thr4Met	missense variant	rs113831133	ExAC AFR	0.1299	DM?	B, LB	Benign (0.049)	Tol (1)	Neu (0.003)	6.269
TMPRSS3	c.617-3_617-2dupTA		splice_acceptor_variant	rs56283966	ExAC_EAS	0.3145	DM?	B, LB	No	No	No	23.5
TMPRSS3	c.268G>A	p.Ala90Thr	missense variant	rs45598239	ExAC NFE	0.0512	DM?	B	Dam (0.816)	Tol (0.22)	Neu (0.350)	22.9
TMPRSS3	c.212T>C	p.Phe71Ser	missense variant	rs185332310	ExAC EAS	0.0105	DM?	B	Dam (0.477)	Tol (0.23)	Neu (0.199)	24.4
TRIOBP	c.1193_1195delAAC	p.Gln398del	inframe_deletion	COSM5713391	ExAC EAS	0.5987	R	-	No	No	No	8.833
TRIOBP	c.3232C>T	p.Arg1078Cys	missense variant	rs200359708	ExAC FIN	0.0130	DM	B	Dam (0.996)	Del_LC (0)	Del (0.906)	29.9

TRIOBP	c.6736G>A	p.Glu2246Lys	missense variant splice_region variant	rs138139146	ExAC NFE	0.0091	DM	-	Dam (0.955)	Del (0.01)	Del (0.804)	34
TPRN	c.559G>T	p.Ala187Ser	missense variant	rs9411313	1000G Total	0.3472	DM	B	Benign (0.003)	Tol (0.28)	Neu (0.032)	0.002
TPRN	c.199G>C	p.Glu67Gln	missense variant	rs753739683	ExAC NFE	0.0116	DM	-	Dam (0.992)	Del (0)	Del (0.892)	20.7
DOMINANAT / RECESSIVE GENES												
TECTA	c.4315C>A	p.Leu1439Ile	missense variant	rs202199158	ExAC EAS	0.0095	DM	LB	Dam (0.998)	Tol (0.06)	Del (0.797)	22.6
MYO7A	c.2236G>A	p.Asp746Asn	missense variant	rs36090425	ExAC AFR	0.0272	DM	B	Benign (0.017)	Tol (0.35)	Neu (0.021)	22.9
MYO7A	c.2476G>A	p.Ala826Thr	missense variant	rs368341987	ExAC SAS	0.0092	DM	VUS, B, P	Benign (0.039)	Del (0)	Neu (0.447)	24.2
MYO7A	c.4697C>T	p.Thr1566Met	missense variant	rs41298747	ExAC NFE	0.0101	DM?	LB	Benign (0)	Tol (0.19)	Neu (0.050)	16.53
MYO7A	c.4805G>A	p.Arg1602Gln	missense variant	rs139889944	ExAC EAS	0.0585	DM?	LB, P	Dam (0.927)	Del (0.01)	Del (0.780)	34
MYO7A	c.5156A>G	p.Tyr1719Cys	missense variant	rs77625410	ExAC AFR	0.1412	DM?	B, LB	Dam (0.997)	Del (0.04)	Del (0.815)	25.2
MYO7A	c.6614_6634 dupTGAGC AAACAGC GGGGCTC CA	p.Met2205_Se r2211dup	inframe_insert ion	rs563508617	1000G Total	0.0242	-	-	No	No	No	17.61

COL11A2	c.4265C>T	p.Pro1422Leu	missense variant	rs555936333	ExAC EAS	0.0147	DM?	LB	Benign (0.015)	Del (0)	Neu (0.445)	25.6
COL11A2	c.2336C>T	p.Pro779Leu	missense variant	rs150877886	ExAC AMR	0.0092	DM	LB	Benign (0.334)	Tol (0.29)	Neu (0.134)	24.6
COL11A2	c.688G>T	p.Gly230Trp	missense variant	rs141430703	ExAC EAS	0.0090	DM	B, LB	Dam (0.814)	Tol (0.06)	Del (0.630)	22.9
TMC1	c.247_249delGAA	p.Glu83del	inframe_deletion splice_region variant	rs376040866	EVS AA	0.0577	DM	B, LB	No	No	No	20.5

Abbreviation: B, benign; CNTL, control dataset; Dam, damaging; Del, deleterious; DM, disease-causing mutations; DM?, disease-causing mutations?; DP, disease-associated polymorphism; FP, in vitro/laboratory or in vivo functional polymorphism; LB, likely_benign; LC, low_confidence (meaning that the protein alignment does not have enough sequence diversity); LP, likely_pathogenic; Neu, neutral; NP, not pathogenic; P, pathogenic; PD, probably_damaging; PNP, probably not pathogenic; PP, probably pathogenic; R, retired entry; Tol, tolerated; VUS, uncertain_significance

4. Comparing previous study

There were 55 genes overlapping between the 66 genes used in the previous study and the 96 genes used in this study.[4] Among the variants reported to the 55 genes, 66 variants were classified as benign through their control datasets and their MAF thresholds which were 0.005 and 0.0005 in AR and AD respectively. Of these 66 variants, *MYO15A* (c.4652C>A) was a variant not found in our control datasets and the other two variants were SNPs (LOXHD1, c.1381C>A; MYO7A, c.93C>T). Filtering the 63 variants using their MAF thresholds with our control dataset showed that 22 of the 63 variants did not exceed their MAF thresholds. According to the ACMG guideline, the 22 variants were not only classified as VUS but also including one pathogenic variant **(Table 4)**. This result showed that it provided a high filtering resolution with the control dataset composed of various population as well as larger sample size to prioritize the causality of variants through MAF threshold.

Next, we looked at the variants reported in the 55 genes with our control databases to examine the variants between their MAF thresholds and our MAF thresholds. There were 10 variants between MAF 0.5% and 0.6% in recessive genes and 13 variants between MAF 0.05% and 0.1% in dominant genes. As a result of classifying their pathogenicity by InterVar, which is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline, VUS and Likely benign variants were 17 and 6, respectively. VUS can be reclassified as

pathogenic or benign category any time when a new article is supported. Therefore, additional functional studies are needed to verify their pathogenicity.

Table 4. Variant which categorized as benign using control dataset in the previous study reclassified as pathogenic variant in this study.

Gene symbol	cDNA position	Amino acid substitution	dbSNP147[a]	CNTL (POPMAX)	POPMAX MAF (AC/AN)	NBK MAF (AC/AN)	PP2 Hum-var	SIFT	Con-del	CA-DD	DVD[b]	HG MD	Clin-Var	ACMG (Evidence)
PCDH15	c.748C>T	p.Arg250Ter	rs111033260, CM030933 With Pathogenic allele A=0.0002/27 (ExAC) A=0.0004/5 (GO-ESP) A=0.00007/2 (TOPMED)	ESP (EA)	0.000465 (4/8600)	0 (0/794)	No	No	No	35	Pathog enic	DM?	Patho genic	Pathogenic (PVS1, PP3, PP5)

Abbreviation: AC, allele count; AN, allele number; CADD, PHRED-like scaled CADD score; CNTL, control dataset; DM?, uncertain disease-causing mutations; DVD, Deafness Variation Database; HGMD, Human Gene Mutation Database; No, no data; PP2 Humvar, PolyPhen-2 humvar prediction score; SNP, single-nucleotide polymorphism.

[a]dbSNP database (http://www.ncbi.nlm.nih.gov/SNP).

[b]DVD (http://deafnessvariationdatabase.org/).

5. Prediction analysis of the reported variants in public databases using in-silico tools

We obtained the prediction scores of missense variants using not only PolyPhen-2 and SIFT, which are widely used as in-silico tools to predict the functional impact of the variants proposed in the American College of Medical Genetics and Genomics (ACMG) guideline, but also Condel, which provides consensus deleteriousness scores for an amino acid substitution based on PolyPhen-2, SIFT, and the other three prediction tools.

In the 28 dominant genes, the predicted scores of the 244 missense variants were not significantly statistically different between the variants below and above the MAF threshold (using Mann-Whitney tests) (**Figure 3A**). On the other hand, the prediction scores of 1,040 and 668 missense variants for 57 recessive and 10 Dom/Rec genes were statistically different between the variants below and above the MAF threshold (**Figure 3B, C**). Therefore, we concluded that the in-silico tool prediction of missense variants only according to the MAF in the dominant genes does not help classify variant pathogenicity.

Dominant genes are mostly dominant-negative or gain-of-function phenotypes except for haploinsufficient cases. Conversely, phenotypes due to recessive genes are mostly unmasked by LoF of the corresponding genes. We therefore hypothesized that prediction tools would better predict the functional impact of LoF mutations. If this is true, haploinsufficient dominant genes will impact the prediction score. To confirm

this, we classified the dominant genes according to the degree of intolerance to LoF through the pLI scores and identified the prediction scores of the variants in the dominant genes belonging to each pLI category. As a result of statistical analysis using Mann-Whitney test, Polyphen-2 and Condel showed statistical differences in the variants of dominant genes with > 0.9 pLI score, and SIFT showed statistical significance between prediction scores according to MAF in the second pLI category (0.1< pLI <0.9) (**Figure 3D**). Also, The prediction scores of missense variants were higher in the dominant genes with higher pLI scores (i.e., in haploinsufficient genes). SIFT showed a difference in prediction scores in the second pLI category. This difference may be due to the fact that SIFT does not take into account the structural and electrical changes of amino acids in missense unlike in the other two programs. These results suggest that prediction tools could be more useful for evaluating variants in haploinsufficient genes.

For this reason, the variants reported in dominant genes with a pLI score > 0.9 were considered less likely to be damaging if identified as benign by the prediction tools, even though the MAF was below the threshold. To figure out this, the two variants included in the above criteria were checked by the ACMG guideline. Consequently, they were judged as variants of unknown significance (VUS) (**Table 5**).

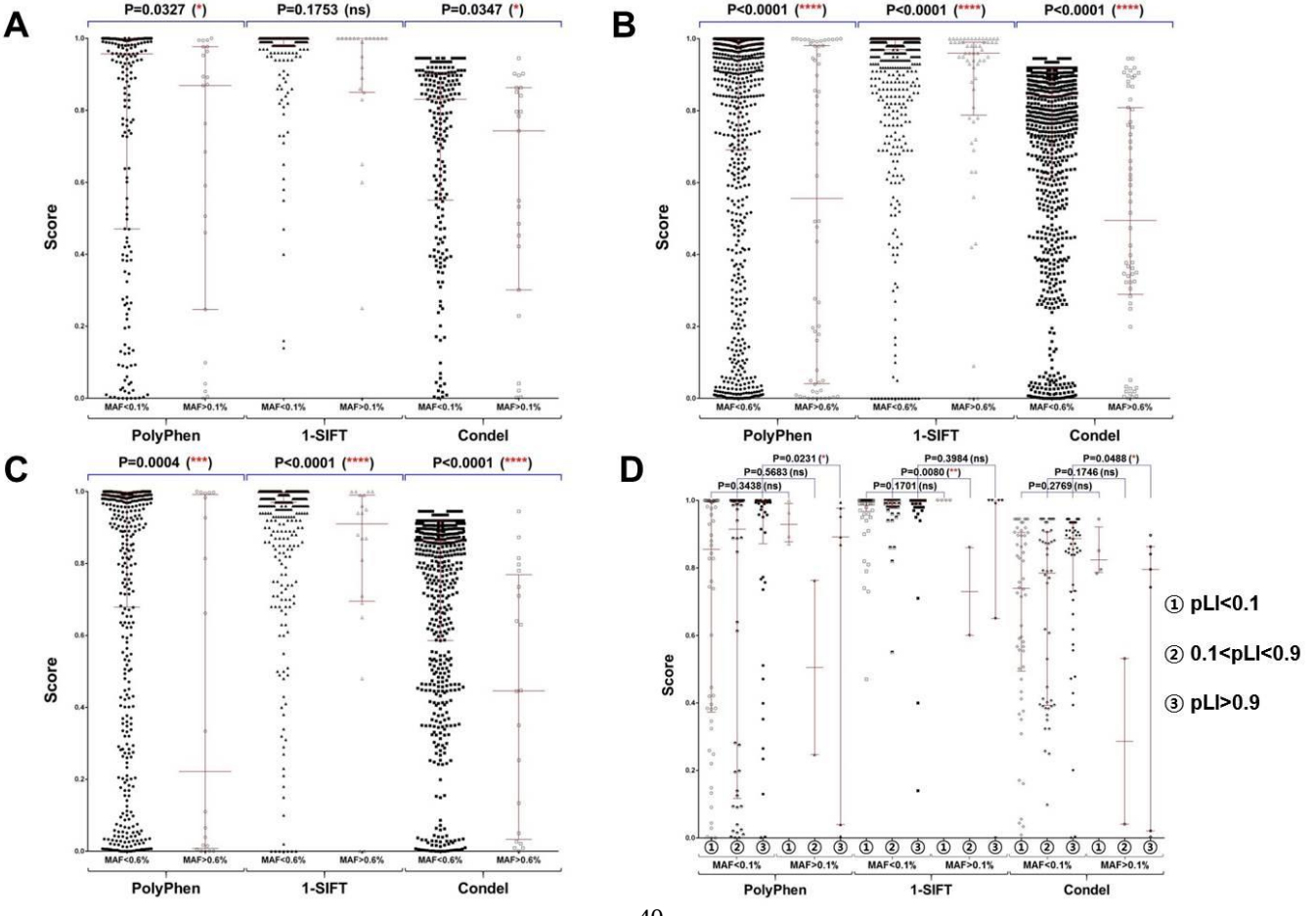


Figure 3. Prediction score of missense variants according to MAF threshold using PolyPhen-2, SIFT, and Condel. (A) In the 28 dominant genes, the predicted scores of the 244 missense variants were not significantly different when comparing variants below and above the MAF threshold. (B) The prediction scores of 1,040 missense variants for 57 recessive genes were statistically different between the variants below and above the MAF threshold. (C) The prediction scores of 668 missense variants for 10 dominant/recessive genes were significantly different between the variants below and above the MAF threshold. All comparisons were made with Mann-Whitney tests. (D) The prediction scores of missense variants showed statistical differences in the dominant genes with higher pLI scores. For all statistical analysis, the Mann-Whitney test was used for the mean comparison of the prediction scores between the variants below and above the MAF threshold. MAF, minor allele frequency; pLI, probability a gene is intolerant to a loss-of-function mutation.

Table 5. The variants predicted as benign in the prediction tools in dominant genes with a pLI score > 0.9 under the MAF threshold

Gene symbol	cDNA position	Amino acid substitution	dbSNP147[a]	CNTL (POPMAX)	POPMAX MAF (AC/AN)	NBK MAF (AC/AN)	PP2 Humvar	SIFT	Condel	CADD	DVD[b]	HGMD	Clin-Var	ACMG
MYH9	c.5137A>G	p.Ser1713Gly	rs764139009 C=0.000008/1 (ExAC)	ExAC (SAS)	0.000061 (1/16498)	No	Bn (0.002)	Tol (0.6)	Neu (0.004)	10.66	Unknown significance	No	Pathogenic	VUS
MYH9	c.3909C>A	p.Phe1303Leu	No	No	No	No	Bn (0.004)	Tol (0.86)	Neu (0.001)	15.66	No	DM	No	VUS

Abbreviation: AC, allele count; AN, allele number; Bn, benign; CADD, PHRED-like scaled CADD score; CNTL, control dataset; DM, disease-causing mutations; DVD, Deafness Variation Database; HGMD, Human Gene Mutation Database; Neu, neutral; No, no data; pLI, probability a gene is intolerant to a loss-of-function mutation; PP2 Humvar, PolyPhen-2 humvar prediction score; SNP, single-nucleotide polymorphism; Tol, tolerated; VUS, uncertain significance.

[a]dbSNP database (http://www.ncbi.nlm.nih.gov/SNP).

[b]DVD (http://deafnessvariationdatabase.org/).

6. Common SNVs associated with reported hearing loss in Korean

It was necessary to select only the SNVs from the reported variants in the 96 NSHL genes because all NBK variants were SNVs. Overall, there were 2,829 SNVs out of the 3,550 variants in the remaining 96 NSHL genes. Of these SNVs, 105 were found in the NBK database, and when these variants were sorted according to the inheritance pattern and mutation type, missense variants were the most frequent as expected (**Table 6**).

One must be wary of extrapolating to or from less well-characterized populations that could harbor founder mutations. Also, allele frequency estimates based on low allele counts are both upward-biased and imprecise. To overcome this limitation, we calculated how many times a variant with the MAF thresholds of 0.001 and 0.006 could be found in the NBK dataset, which has a low allele number compared to other control datasets. To facilitate these calculations, we used an online calculator (http://cardiodb.org/alleleFrequencyApp) that computes both the maximum credible population allele frequency and maximum sample allele count (AC) for a user-specified genetic architecture.[6] At a 5% error rate, this yields a maximum tolerated AC of 2 and 9 in dominant and recessive genes, respectively.

All except one variant (c.147C>G allele of *SLC26A4*; AC/AN in NBK = 1/794) out of the 105 found in the NBK dataset were found in the other three control datasets. As a result of filtering the 104 variants using the MAF thresholds according to

inheritance mode, three variants were only filtered by the NBK (**Table 7**). In the other control datasets, these variants showed lower MAFs than 0.001 and 0.006 in dominant and recessive genes, respectively. However, it is difficult to classify as benign for the three variants because the number of alleles exceeding maximum tolerated AC is not large enough. The difference between observed AC and maximum tolerated AC is likely to be caused by sequencing error. Therefore, a larger sample size is required to identify variants with MAF levels implausible for highly penetrance Mendelian disease using the Korean population control datasets.

Table 6. Summary of reported SNVs in NBK

Consequence	All	AD	AR	AD/AR	X-linked
Missense_variant	92	16	57	19	0
Stop_gained	5	1	4	0	0
Intron_variant	3	0	3	0	0
3_prime_UTR_variant	1	0	1	0	0
5_prime_UTR_variant	1	0	1	0	0
Splice_region_variant	2	0	2	0	0
Upstream_gene_variant	1	0	1	0	0
SUM	105	17	69	19	0

Abbreviation: AD, autosomal dominant; AR, autosomal recessive; AD/AR, genes with both AD and AR inheritance; X-linked, x chromosome linked gene

Table 7. The variants above the MAF threshold in the Korean datasets

Gene symbol	cDNA position	Amino acid change	dbSNP147[a]	CNTL (POPMAX)	POPMAX MAF (AC/AN)	NBK MAF (AC/AN)	PP2 Humvar	SIFT	Condel	CA-DD	DVD[b]	HG-MD	Clin-Var	Inte-rVar (Evid-ence)
P2RX2	c.817G>T	p.Asp273Tyr	rs767470753 T=0.00007/8 (ExAC)	ExAC (EAS)	0.000924 (8/8654)	0.006297 (5/794)	Dam (0.991)	Del (0)	Del (0.889)	27.1	Unknown significanc e	DM	No	VUS (PM1, PP2)
WFS1	c.1846G>T	p.Ala616Ser	rs553336498 With Uncertain significance allele T=0.000008/1 (ExAC) T=0.0002/1 (1000G)	1000G (Total)	0.0002 (1/5008)	0.003778 (3/794)	Bn (0.108)	Tol (0.84)	Neu (0.009)	0.151	Pathogenic	DM	No	VUS (PM2, PP5)
TECTA	c.3511G>A	p.Val1171Met	rs186780639 With Uncertain significance allele A=0.0003/33 (ExAC) A=0.0008/4 (1000G)	ExAC (EAS)	0.003843 (33/8586)	0.013854 (11/794)	Dam (0.79)	Tol (0.29)	Neu (0.326)	24.1	Benign	DM?	Uncer tain signifi cance	VUS (PM2, BS2)

Abbreviation: AC, allele count; AN, allele number; Bn, benign; CADD, PHRED-like scaled CADD score; CNTL, control dataset; DM, disease-causing mutations; DM?, disease-causing mutations?; DVD, Deafness Variation Database; HGMD, Human Gene Mutation Database; Neu, neutral; No, no data; PP2_Humvar, PolyPhen-2 humvar prediction score; SNP, single-nucleotide polymorphism; Tol, tolerated; VUS, uncertain significance.

[a]dbSNP database (http://www.ncbi.nlm.nih.gov/SNP).

[b]DVD (http://deafnessvariationdatabase.org/).

IV. DISCUSSION

We systematically evaluated the 96 NSHL genes with MAFs and prediction tools. The MAF threshold of NSHL was set based on hearing loss prevalence and the disease-specific genetic and allelic architecture. We also suggested an alternative for the variance problem in datasets with low allele numbers and/or low base quality regions among three different population-specific control datasets. In this way, the MAF of each variant reported in the control datasets was more precisely confirmed. We found that at least 98 of the 3,550 variants reported in the public mutation database were less likely to be causal variants of NSHL. In addition, variants in recessive genes, in which the phenotype is revealed by LoF, had large differences in prediction scores according to the MAF threshold. This result shows that the prediction score increases to a greater degree in a group with variants expected to act as LoF, which allowed us to prioritize haploinsufficient dominant gene variants through the pLI score. Furthermore, using the WGS data of 397 Koreans, 3 variants that could not be prioritized by the MAF threshold in the control datasets could be categorized as variants with weak causality.

There are some limitations to this evaluation approach. First, even if this method excluded 103 variants from the high priority group that are likely to be causal variants, the ACMG guideline should be used for clinical diagnoses. Second, since most widely used prediction tools are applied to missense variants, we evaluated only the 1,969

missense variants out of the reported 3,550 variants. However, if more useful prediction tools such as a CADD program capable of quantifying the degree of deleteriousness in multiple variant types are developed, a systematic evaluation of variants in the other types besides missense variants would be possible. Third, recessive genes are intolerant genes when the LoF variant is homozygous rather than heterozygous. Therefore, recessive genes were classified as pRec <0.1 + pNull> 0.9, 0.1 < pRec <0.9, and pRec > 0.9 + pLI <0.1. However, the variants below the MAF threshold showing the highest prediction score at 0.1 < pRec <0.9 showed no correlation between pRec categories and prediction score (**data not shown**). Fourth, we could not perform further analysis for seven genes (*LOXHD1*, *MIR96*, *MYO15A*, *OTOG*, *TPRN*, *WFS1*, *WHRN*) without pLI scores and for *PRPS1* (X-linked) with an unknown inheritance mode. Lastly, it is difficult to determine the pathogenicity of a variant using this MAF threshold only because of the presence of some genes that are frequently found pathogenic alleles such as *GJB2* in European and *SLC26A4* in East Asian. However, as we know, pathogenic alleles frequently found in certain ethnic groups are more likely to be shared among the individuals with hearing impairment, therefore, have been much more likely to be discovered and many studied. It is therefore unlikely that any single newly identified variant will explain a similarly large proportion of NSHL as the most common causal variant, at least in well-studied populations. Also, other causal variants found except common founder mutations will not show an MAF above the set threshold.

This assessment approach to prioritize variant causality can be applicable to a variety of Mendelian diseases for which the prevalence and genetic/allelic architecture are known. It can be used as part of a systematic approach evaluating variant causality.

V. CONCLUSION

1. Variants above the MAF thresholds are most unlikely to be causal variants of NSHL but there is an exception.

2. It is desirable to use exome and genome data together as a control dataset, and it is sufficient to use ExAC as the control dataset for exome.

3. All reported variants as hearing loss have a possibility that include those with lower causality because many of the variants so far have been studied in European and American populations.

4. Composing of various population as well as larger sample size to prioritize the causality of variants through MAF threshold provides a high filtering resolution.

5. The prediction score shows a statistical difference between variants below and above the MAF threshold in recessive and haploinsufficient genes.

REFERENCE

1. Hilgert N, Smith RJ, Camp GV. Function and expression pattern of nonsyndromic deafness genes. *Current molecular medicine.* 2009;9(5):546-564.
2. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature.* 2014;508(7497):469-476.
3. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291.
4. Shearer AE, Eppsteiner RW, Booth KT, Ephraim SS, Gurrola J, 2nd, Simpson A, et al. Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. *Am J Hum Genet.* 2014;95(4):445-453.
5. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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(5):405-424.
6. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med.* 2017.
7. Sloan-Heggen CM, Bierer AO, Shearer AE, Kolbe DL, Nishimura CJ, Frees KL, et al. Comprehensive genetic testing in the clinical evaluation of 1119 patients with hearing loss. *Hum Genet.* 2016;135(4):441-450.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7(4):248-249.
9. 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(7):1073-1081.
10. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of

nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet.* 2011;88(4):440-449.

11. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-315.
12. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156-2158.
13. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing datasets. *bioRxiv.* 2015.
14. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A.* 2015;112(17):5473-5478.
15. Fattahi Z, Shearer AE, Babanejad M, Bazazzadegan N, Almadani SN, Nikzat N, et al. Screening for MYO15A gene mutations in autosomal recessive nonsyndromic, GJB2 negative Iranian deaf population. *Am J Med Genet A.* 2012;158A(8):1857-1864.
16. Chiu YH, Wu CC, Lu YC, Chen PJ, Lee WY, Liu AYZ, et al. Mutations in the OTOF Gene in Taiwanese Patients with Auditory Neuropathy. *Audiology and Neurotology.* 2010;15(6):364-374.

Abstract (in Korean)

부 대립 유전자 빈도 와 예측 도구를 이용한 난청 관련 변이들의 체계적인 평가 방법

<지도교수 지헌영>

연세대학교 대학원 의과학과

이준석

비 증후군성 난청은 유전적으로 매우 이질적이고, 현재까지 96 개 이상의 유전자가 관련되어 있으며, 전체 난청 환자들의 절반을 차지하고 있다. 이러한 질병의 원인이 되는 돌연변이는 높은 처리량의 DNA 염기 서열 분석 기술을 통해서 더욱 용이하게 밝혀낼 수 있게 되었지만, 이 방법으로 확인된 수백 또는 수천 개의 변이들은 질병을 일으킬 가능성을 평가하기 위해서 추가적인 해석이 필요하다. 우리의 연구 목표는 비 증후군성 난청 환자에서 확인 된 96 개 유전자의 변이들을 부 대립 유전자 빈도 (MAF, Minor Allele Frequency) 및 예측 도구를 이용하여 체계적으로 평가하는 방법을 제시하는 것이다. MAF의 임계 값은 *GJB2*의 가장 흔한 병원성 변이 형의 대립 유전자 빈도와 비 증후군성 난청의 유병율을 고려하여 설정했다. 우리가 설정한 MAF 임계 값에 따라서 HGMD에 보고 된 3,082 개의 변이와

ClinVar에 보고 된 1,210개 변이들을 분류 했다. 그리고 pLI 값에 따라서 96개 유전자들을 다음과 같은 3가지 범주로 분류 했다 (pLI <0.1, 0.1 <pLI <0.9, pLI> 0.9). 열성, 우성, 그리고 우성/열성 유전자에 보고 된 missense 변이 수는 각각 1,040, 244 그리고 668개 였고, 각 missense 변이 형의 예측 점수는 PolyPhen-2, SIFT 그리고 Condel을 이용했다. 분석 결과, 열성, 우성 그리고 우성/열성 유전자에서 MAF 임계 값 이상의 변이들은 61개, 23개 그리고 14개였다. 또한, 한국인 대조군 데이터를 통해서 다른 다른 인종 보다 특히 한국인에서 더 자주 발견되는 세 개의 변이를 확인했다. 이는 다른 특정 인종의 대조군 데이터를 통해서 유전병을 일으키기에는 믿기 어려운 MAF 가진 변이들을 추가로 발견해 낼 수 있음을 시사한다. 추가로, 열성 유전자의 missense 변이들은 MAF 임계 값 보다 높거나 낮은 변이들 간의 예측 점수에서 통계적인 차이를 보였다. 그리고 우성 유전자는 MAF 임계 값 보다 높거나 낮은 변이들 간에 예측 점수의 통계적인 차이를 볼 수 없었던 반면, 0.9 보다 높은 pLI 값을 갖는 우성 유전자들의 변이들은 예측 점수에 유의한 차이를 보였다. 이를 통해서 예측 도구는 열성 유전자의 변이들과 pLI 값이 0.9 보다 높은 우성 유전자의 변이를 예측하는 데 더 유용하다고 볼 수 있다. 이러한 결과를 바탕으로, 비 증후군 성 난청의 병원성 변이를 평가하는 데 MAF 임계 값 및 예측 도구를 이용하여 해당 질병과 인과 관계가 높은 변이들을 우선 순위화 할 수 있다.

핵심되는 말: 비 증후군성 난청, 부 대립 유전자 빈도, 예측 도구, pLI

PUBLICATION LIST

1. Jung J*, **Lee JS***, Cho KJ, et al. Genetic Predisposition to Sporadic Congenital Hearing Loss in a Pediatric Population. *Sci Rep. 2017;7:45973. (* equal contribution)*

2. Oh CM, Chun S, Lee JE, **Lee JS** et al. A novel missense mutation in NR0B1 causes delayed-onset primary adrenal insufficiency in adults. *Clin Genet. 2017;92(3):344-346.*

3. Jun I, **Lee JS**, Lee JH, et al. Adult-Onset Vitelliform Macular Dystrophy caused by BEST1 p.Ile38Ser Mutation is a Mild Form of Best Vitelliform Macular Dystrophy. *Sci Rep. 2017;7(1):9146.*

4. Choi HJ*, **Lee JS***, Yu S, Cha DH, Gee HY, Choi JY, et al. Whole-exome sequencing identified a missense mutation in WFS1 causing low-frequency hearing loss: a case report. *BMC Med Genet. 2017;18(1):151. (* equal contribution)*