

Comprehensive genetic analysis of *STRC* variants in hereditary hearing impairment using long-read sequencing

Cheng-Yu Tsai^{1,2}, Yue-Sheng Lu¹, Yu-Ting Chiang², Ming-Yu Lo¹, Pei-Hsuan Lin^{1,3}, Shih-Feng Tsai^{4,5}, Chuan-Jen Hsu^{1,6}, Pei-Lung Chen^{2,3,7,8}*, Jacob Shu-Jui Hsu²*, and Chen-Chi Wu^{1,2,9,10}*

- 1) Department of Otolaryngology, National Taiwan University Hospital, 100225 Taipei, Taiwan.
- 2) Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, 100233 Taipei, Taiwan.
- 3) Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, 100229 Taipei, Taiwan.
- 4) Institute of Molecular and Genomic Medicine, National Health Research Institute, 350401 Miaoli, Taiwan.
- 5) Department of Life Sciences and Institute of Genome Sciences, National Yang-Ming University, 112304 Taipei, Taiwan
- 6) Department of Otolaryngology, Buddhist Tzuchi General Hospital, 427003 Taichung, Taiwan
- 7) Institute of Molecular Medicine, National Taiwan University College of Medicine, 100225 Taipei, Taiwan
- 8) Department of Medical Genetics, National Taiwan University Hospital, 100226 Taipei, Taiwan
- 9) Department of Medical Research, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu 302041, Taiwan
- 10) Department of Otolaryngology, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu 302041, Taiwan

* **Correspondence:** Pei-Lung Chen (P.-L.C.), Jacob Shu-Jui Hsu (J.-S.H), and Chen-Chi Wu (C.-C.W.)

- Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, No.1 Jen-Ai Road Sec.1, 100233 Taipei, Taiwan (P.-L.C.) & (J.-S.H); paylong@ntu.edu.tw (P.-L.C.) jacobhsu@ntu.edu.tw (J.-S.H)
- Department of Otolaryngology, National Taiwan University Hospital, No.7 Chung-Shan South Rd., 100225 Taipei, Taiwan (C.-C.W.); chenchiwu@ntuh.gov.tw (C.-C.W.)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Abstract

Background: Sensorineural hearing impairment (SNHI) is a common disorder with a significant genetic component. Genetic testing for SNHI often involves next-generation sequencing (NGS), but SNHI-related pathogenic *STRC* variants cannot be directly addressed by conventional NGS due to the complex genomic scenario derived from large genomic rearrangements and a highly homologous pseudogene. Long-read sequencing (LRS) offers an unprecedented resolution to these challenges.

Methods: We developed a comprehensive workflow that integrates the PacBio-based LRS approach with marker-mediated refinements to effectively address pseudogene contamination. This methodology was applied to analyze the *STRC* gene in a cohort of 100 unrelated Taiwanese patients diagnosed with SNHI of unknown genetic cause after first-tier NGS testing.

Results: We identified bi-allelic *STRC* variants in 11 patients (11% diagnostic yield), including homozygous deletions, compound heterozygous deletions and conversions, and compound heterozygous SNVs and CNVs. In total, we detected *STRC* variants in 27 patients, with 81.6% of these variants occurring in patients with mild to moderate SNHI.

Conclusions: This study represents the first large-scale clinical investigation utilizing LRS technology for the genetic diagnosis of SNHI. Our study highlights the diagnostic capabilities of LRS in detecting complex variants within the *STRC* and advancing our understanding of the genetic etiology of SNHI that remains unresolved by conventional NGS.

Key words: hereditary hearing impairment (HHI), stereocilia (*STRC*), long-reads sequencing (LRS), high-fidelity (HiFi) reads, copy number variants (CNVs), marker-mediated refinements

Introduction

Sensorineural hearing impairment (SNHI) is a common childhood disorder with diverse etiologies. Genetic causes account for more than 50% of childhood SNHI cases and are classified as hereditary hearing impairment (HHI). To date, more than 120 genes have been associated with non-syndromic HHI [1, 2]. Among these, the stereocilin (*STRC*) gene, inherited in a recessive pattern (DFNB16, OMIM #603720), is known to cause non-syndromic HHI. The stereocilin protein encoded by *STRC* is expressed in the outer hair cell (OHC) bundles of the cochlea and is responsible for connecting adjacent stereocilia and interacting with the tectorial membrane [3]. This gene has been associated with mild-to-moderate SNHI and deafness-infertility syndrome, especially when long-range deletions occur across the *STRC* and *CATSPER2* region (15q15.3) [4, 5].

Pathogenic variants in *STRC* are a major contributor to HHI in different populations, with prevalence ranging from 5.6-10.2% in Europeans [6-8], 6.1-11.2% in Americans [9, 10], and 2.8-10.3% in East Asians [11-13]; studies have also confirmed *STRC* variants in Hispanic and Middle Eastern populations, with an overall prevalence of 16.1% in multi-ethnicity studies [14]. In addition to single nucleotide variants (SNVs), copy number variants (CNVs), also called structural variants (SVs), are important genetic causes of the *STRC*-related HHI. CNVs include a variety of large-scale genomic rearrangements ranging from 50 bp to megabases in size, such as deletions, duplications, translocations, and conversions [15]. CNVs have been identified in at least 29 HHI-related genes, with *STRC*, *OTOA*, and *GJB2/GJB6* being the most frequently implicated [16]. A meta-analysis by Han et al. [17], which included data from 37 relevant articles, showed that 14.4% of mild to moderate

HHI patients have bi-allelic pathogenic *STRC* variants. Homozygous *STRC* deletions account for the largest proportion (~70%) of these variants, highlighting their significant etiological role.

Another diagnostic challenge for *STRC* variants is the presence of the homologous pseudo-*STRC* (*STRCPI*) located ~80kb upstream of the *STRC* region. *STRCPI* shares ultra-high sequence identity (97%) with *STRC* and has been implicated in gene-to-pseudogene conversion, complicating genetic testing [18]. While next generation sequencing (NGS) has demonstrated efficacy in identifying HHI-associated variants [2], its short-read nature (150-300 bp) limits its ability to comprehensively address both CNVs and pseudogenes simultaneously. This limitation hinders the accurate detection of SNVs and CNVs in the *STRC* "dark region", which extends from the five-prime untranslated region to exon 18 and shows 99.8% identity between *STRC* and *STRCPI* (with less than 30 bp nucleotide differences in an approximately 11 kb region). To overcome this challenge, several bioinformatic tools and fluorescence-based approaches have been developed to complement NGS for CNV detection. These include CNV calling algorithms [19, 20], array comparative genomic hybridization (aCGH) [21-23], allelic-specific droplet digital PCR [9, 10], or multiplex ligation-dependent probe amplification (MLPA) [24, 25]. In addition, long-range (LR) PCR/nested PCR [8, 9] has been used to reduce pseudogene contamination and improve the accuracy of conventional NGS approaches. However, these methods often require the integration of multiple approaches and complex hierarchical pipelines, requiring the preparation of numerous primers specific for each exon or intron for subsequent validations. A one-round genetic assay that can quickly, accurately, and efficiently address the challenging questions posed by the *STRC* gene would be of significant clinical and academic value.

Long-read sequencing (LRS) assays offer a promising solution to these challenges by enabling direct sequencing of large genomic regions, ranging from kilobases to megabases. Two commercial platforms, Pacific Biosciences (PacBio) [26] and Oxford Nanopore Technologies (ONT) [27], offer high-resolution LRS for the detection of SNVs, indels, and CNVs [28]. In particular, PacBio's high-fidelity (HiFi) sequencing via circular consensus sequencing (CCS) has demonstrated high accuracy rates for both SNVs (>99%) and CNVs (>95%) [29], which has been successfully used to implement direct LRS assays on 193 medically relevant genes, including the challenging *STRC*. However, the application of LRS-mediated CNV detection in large-scale clinical studies is still limited.

In this study, we use amplicon-based LRS technology to perform two-step target enrichment of *STRC* genomic segments, followed by generation of 10kb HiFi reads using the PacBio CCS pipeline to detect SNVs and CNVs hidden in the *STRC* gene. This study provides a potential solution to the challenges of *STRC* variant detection in clinical genetics.

Materials and Methods

Subjects

This study included 100 patients with SNHI of unknown genetic cause after first-tier NGS [30], including 73 with mild-to-moderate SNHI and 27 with severe-to-profound SNHI. Three National Institute of Standards and Technology (NIST) reference samples-HG001 (NA12878), HG002 (NA24385), and HG005 (NA24631)-were used for benchmarking [31], whose lymphoblastoid cell line DNAs were obtained from the Coriell Institute for Medical Research (Camden, NJ, USA).

Genomic DNAs of patients were extracted using the MagCore® genomic DNA extraction kit (Cat. No. MGB400-03). This study was approved by the National Taiwan University Hospital (approval number 202012083RIND), and informed consent was obtained from all participants or their legal guardians.

Long-read amplification of STRC

Prior to the implementation of LRS assays using single-molecule real-time (SMRT) technology, the recommended PacBio workflow (PN 101-921-300) for the construction of amplicon-based PacBio SMRTbell® libraries was followed. A two-step PCR procedure was performed for each sample prior to SMRTbell library construction. In the first step, four pairs of 42 bp primers were designed, each consisting of a 25 bp targeted primer (**Table 1**) and a 17 bp M13 sequence connected at the 5' prime end (5'-GTAAAACGACGGCCAGT for forward and 5'-CAGGAAACAGCTATGAC for reverse). These primers, supplied by MB MISSION BIOTECH Co., Ltd. (Taipei, Taiwan), were capped with the 5' amino modifier group C6 (5AmMC6) and used with the TaKaRa LA Taq® kit (code no. RR002A) to generate the initial PCR amplicons. In the second step, multiple pairs of asymmetric M13-barcoded forward/reverse primers (as specified in the PacBio workflow) and the KAPA HiFi HotStart ReadyMix (KK2602) were used to generate M13-barcoded amplicons .

These amplicons were then purified using AMPure® PB Beads and capillary electrophoresis, followed by SMRTbell library construction using the SMRTbell® Express Template Prep Kit 2.0 (PN 100-938-900). The SMRTbell products were then subjected to SMRT sequencing on the PacBio

Sequel IIe platform using the Sequel II Sequencing Kit 2.0 (PN 101-820-200).

Table 1. Targeted primers for amplifying *STRC* genomic regions.

Primer	Loci of targeted primer (hg19)	25-bp targeted primers
LRP1-F	chr15:43905486-43905510	CCCTCTGTAGCCCTGCTGTCTAGAT
LRP1-R	chr15:43911226-43911250	CCAGTTCTCCATCTCCATGTCTGCC
LRP2-F	chr15:43901216-43901240	GCCCAAGAGGACAAGTAGGTGACAG
LRP2-R	chr15:43906328-43906352	TG TTCAGAGGGCTTGCTTTAGTGGG
LRP3-F	chr15:43896367-43896391	AAGGAGAGTGGGGAGATCTGGACAG
LRP3-R	chr15:43901539-43901563	GGGAAGGGGCTCGTCTTTCTTTCTC
LRP4-F	chr15:43891467-43891491	ATTA ACTCCCCATCGCCAGCTGATG
LRP4-R	chr15:43896550-43896574	TTCCTTGATCCCCAGGGTGAGATGA

LRP: long-range PCR primer; F: forward; R: reverse.

Data analysis

Raw data from LRS assays were processed using SMRT[®] Link (ver. 10.2.1, PacBio), the recommended software for analysis of PacBio sequencing data. Reads were mapped to the hg19 reference genome using pbmm2 (ver. 1.9) [32]. Variant calling, normalization, and annotation were performed using DeepVariant (ver. 1.4.0) [33], BCFtools (ver. 1.13) [34], and ANNOVAR (ver. 2021Oct19) [35], respectively. The merging and conversion of mapped BAM files was implemented using SAMtools (ver. 1.15.1) [36], and the refinement of long reads based on specific markers was conducted using Bamq1 (ver. 1.6) [37].

Multiple predictive scores were applied for the pathogenicity prediction of variants, including SIFT [38] and PolyPhen-2 [39] retrieved from dbNSFP (ver4.1) [40], CADD (GRCh37-ver. v1.7)

[41], and SpliceAI (ver. 1.3.1) [42]. The allele frequencies of variants were retrieved from population databases gnomAD [43] (ver. 2.1.1, last accessed Oct 19, 2024) and Taiwan Biobank [44] (last accessed Oct 10, 2024). The pathogenicity assertions of variants were retrieved from disease databases ClinVar [45] (last accessed Oct 10, 2024) and Deafness Variant Database (DVD, ver. 9) [45].

Multiplex ligation-dependent probe amplification (MLPA) assays

DNA samples (> 500ng) were prepared for validation of CNVs by MLPA assays. MLPA assays were performed using a commercial kit (SALSA[®] MLPA[®] Probemix P461-B1, MRC Holland, Amsterdam, The Netherlands). This experimental kit consisted of 45 MLPA probes within the chromosome 15q15.3 and 16q12.2 regions, including *STRC*, *CATSPER2*, *STRCP1*, *OTOA*, and other nearby regions. Seven and four probes in this kit were designed for the range of exons 19 to 28 on *STRC* and *STRCP1*, respectively. The experimental pipeline was based on the recommended general MLPA[®] protocol (version-008).

Results

Target enrichment of the STRC region for generating 10kb amplicons

In this study, a total of 100 unrelated SNHI cases were recruited for the PacBio-mediated LRS assay, including 73 with mild-to-moderate SNHI and 27 with severe-to-profound SNHI. **Figure 1A** shows the amplicon-based target enrichment strategy for *STRC* using four pairs of PCR primers (see

Methods) to generate 5.1-5.9kb amplicons. These amplicons, covering a quarter of the *STRC* region (19.4kb) and containing hundreds of base pairs of overlapping regions, were successfully amplified in the control samples. The amplicon length was further extended to 10kb using interlaced primer pairs, generating PCR products Amp-01, Amp-02 and Amp-03 in the control samples (**Figure 1B**). This approach also amplified homologous segments within the pseudogene *STRCP1*. The expected regions of the *STRC/STRCP1* amplifications are listed in **Table S1**.

SMRTbell libraries were constructed using these 10kb amplicons as circular templates for the PacBio CCS approach, generating HiFi sequencing data with 98-99% accuracy (**Figure 1C**). All samples, including three NIST reference samples (HG001, HG002, HG005), were mapped to the hg19 reference genome using PacBio-recommended pbmm2 software, followed by variant calling and annotation.

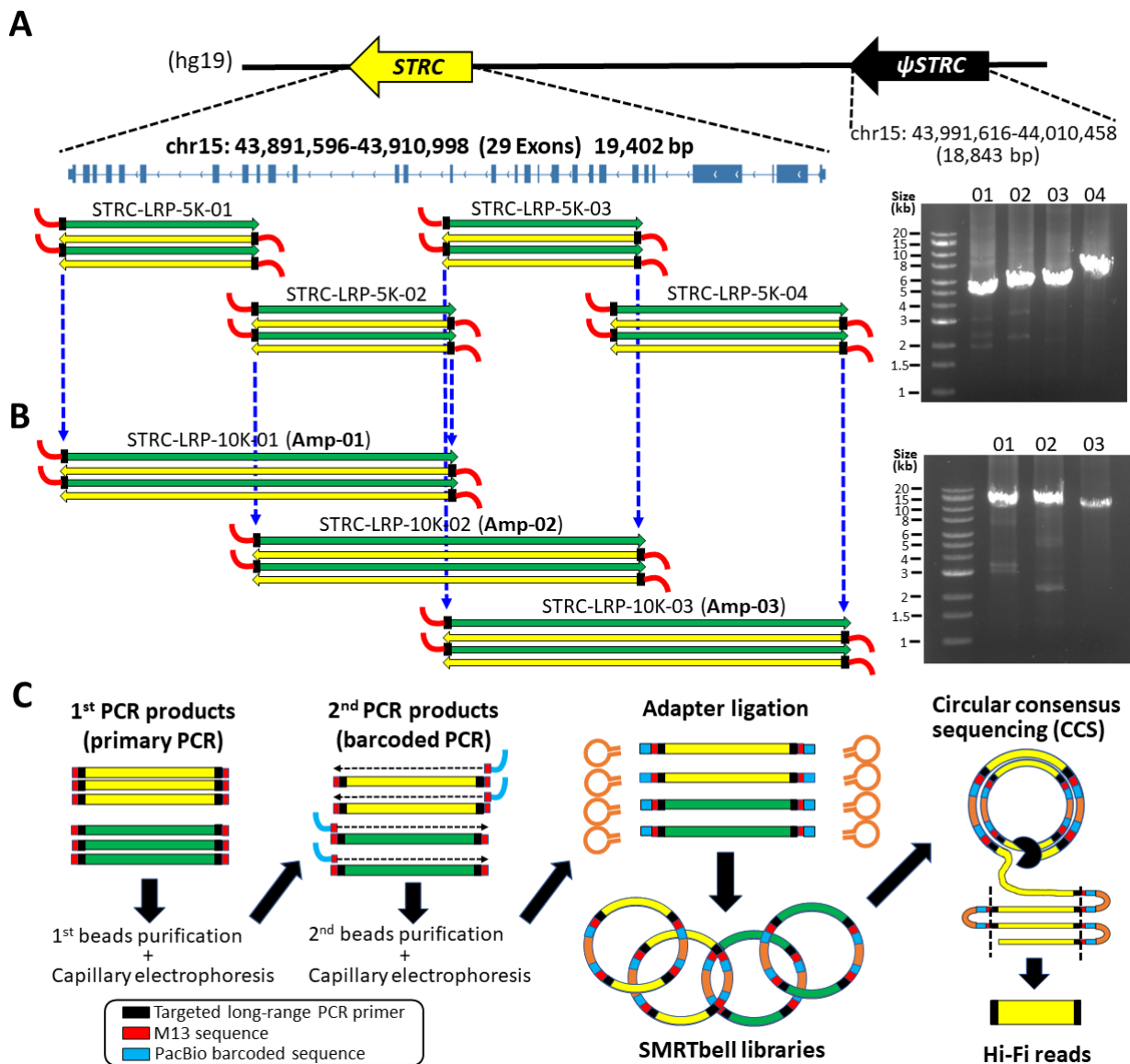


Figure 1. The workflow of *STRC* amplicon-based target enrichment, SMRTbell libraries, and circular consensus sequencing (CCS) for long-read sequencing (LRS). (A) The expected 5kb amplicons (STRC-LRP-5K-01 to -04) with their corresponding electrophoretic evidences. (B) The expected 10kb amplicons (STRC-LRP-10K-01 to -03, also referred to as Amp-01 to -03) with their corresponding electrophoretic evidences. (C) The two-step PCR of 10kb amplicons for SMRTbell library construction and CCS approach for Hi-Fidelity (Hi-Fi) reads.

Benchmarking and refinement of the STRC/STRCP1 copy number ratios using NIST reference samples

To establish benchmarks for our LRS assay, we first analyzed three NIST reference samples (HG001, HG002, and HG005). In addition to our LRS data, we incorporated external sequencing data generated using the PacBio CCS pipeline (11kb) curated in the open access GIAB (Genome in a

Bottle project) repository [46]. **Figure 2A** shows the mapping content of these reference samples on *STRC* and *STRCPI*. The distribution of mapped reads from the external source (upper panel) appeared symmetrical in HG001 and HG002, but asymmetrical in HG005. In contrast, our LRS assay showed an asymmetric distribution for HG001 and HG002 that was inconsistent with the external data (**Figure 2A**, bottom panel). This discrepancy suggested potential pseudogene contamination in our LRS assay.

To address this, we developed a refinement process using a set of divergent markers between *STRC* and *STRCPI*, spanning from intron 26 to intron 15 (**Table 2 & Figure S1**). These markers were grouped into 14 clusters in *STRC* (M1-M14) and *STRCPI* (pM1-pM14), respectively, based on their loci. To ensure accurate filtering of mis-mapped reads, these *STRC* markers, which represent the major differences between *STRC* and *STRCPI*, were selected based on their absence from the population genome (gnomAD database) with sufficient confidence (i.e., adequate coverage documented in the database). **Figure 2B** illustrates the filtering process, which produces four groups of filtered reads. Reads where all markers within a cluster at *STRC* and *STRCPI* contain the reference nucleotide (Ref) are categorized as group (1) and (3), respectively. Conversely, reads where all markers within a cluster at *STRC* and *STRCPI* match the alternative nucleotide (Alt) are classified as group (2) and (4), respectively. The refined copy number (rCN) ratios of *STRC* and *STRCPI* are then calculated as the sum of read counts in the combined group (1+4) divided by the sum of read counts in the combined group (1+2+3+4), and the sum of read counts in the combined group (2+3) divided by the sum of read counts in the combined group (1+2+3+4), respectively (**Figure 2B**, bottom panel).

Applying this rCN calculation (**Figure 2C**), HG001 and HG002 showed a similar trend of equal proportion per cluster between *STRC* and *STRCPI* in both the external source and our LRS data, supporting the feasibility of our refinement strategy. Interestingly, the *STRC* of HG005 showed a higher proportion in both datasets, suggesting a duplicated *STRC* in this sample. To confirm this finding, we performed MLPA analysis. As shown in **Figure 2D**, HG005 showed 1.5-fold MLPA signals at *STRC* and 0.5-fold signals at *STRCPI*, consistent with the rCN ratios and indicating a one-copy gain at *STRC* and one-copy deletion at *STRCPI* in HG005. Both HG001 and HG002 showed symmetrical *STRC/STRCPI* proportions in both rCN ratios (**Figure 2C**) and MLPA results (**Figure S2**), further validating the reliability of our refinement strategy.

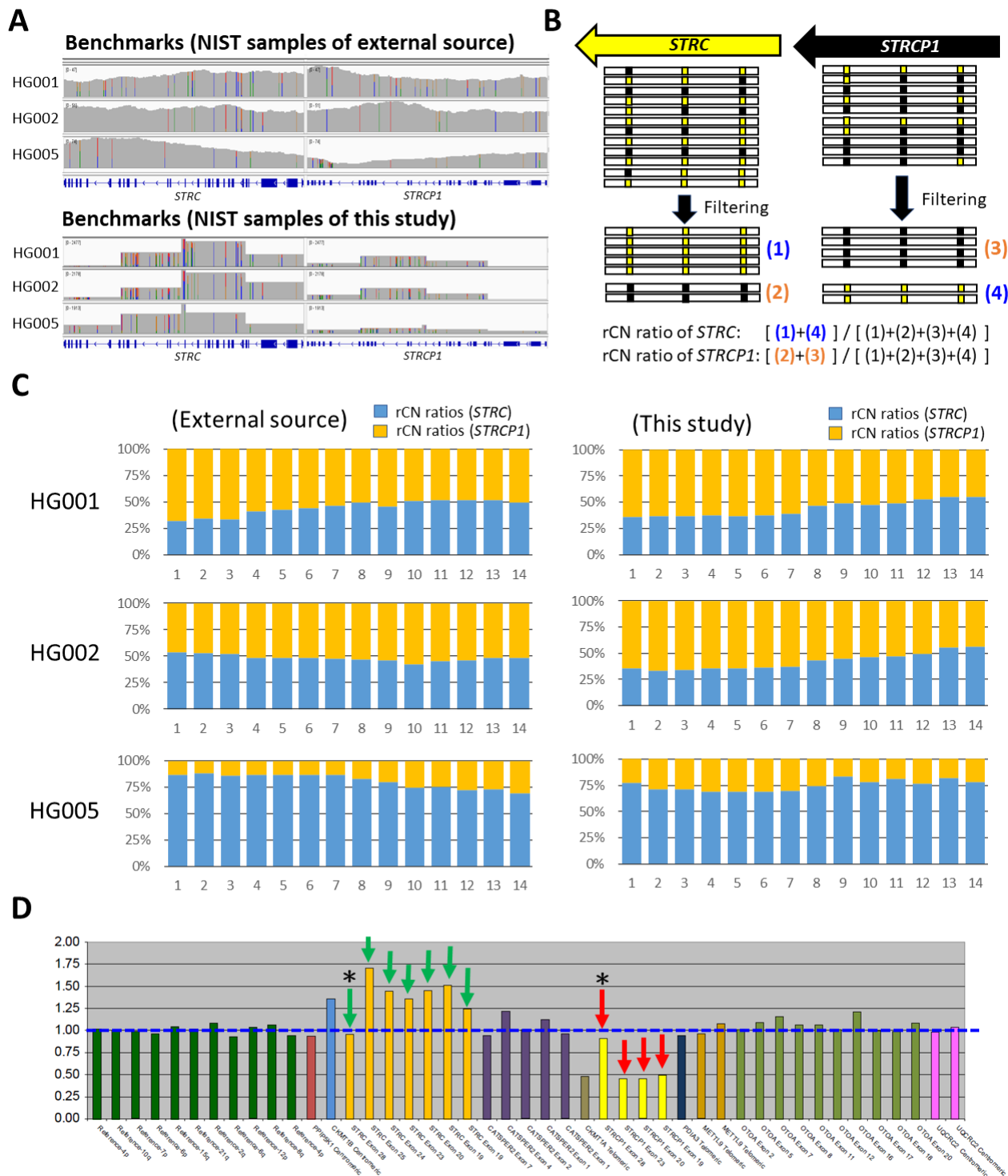


Figure 2. The refinement and validation of copy number ratios between *STRC* and *STRCP1* in benchmarks of NIST reference samples. (A) The overview of mapped LRS reads from external source and this study. (B) Illustrative processing of *STRC/STRCP1* refined copy number (rCN). (C) The rCN plots of reference samples (HG001, HG002, HG005) from external source and this study. (D) The MLPA plots of HG005. Green arrows: the MLPA signal peaks for *STRC*; red arrows: the MLPA signal peaks for *STRCP1*; y-axis: copy number compared to wild-type controls; blue dashed line: the baseline of normal copy number (copy number = 1.00). Asterisks indicate peaks directed at exon 28 of the *STRC/STRCP1* region. (Abbreviations) NIST: National Institute of Standards and Technology; MLPA: multiplex ligation-dependent probe amplification.

Table 2. The 14 *STRC/STRCP1* divergent marker clusters

Loci of marker (<i>STRC</i>)						Loci of marker (<i>STRCP1</i>)						Amplicon
	Ref	Alt	Exon / Intron	Allele frequency (Coverage)		Ref	Alt	Allele frequency (Coverage)				
M1-1	chr15-43892640	C	A	Intron 26	N/A (>30)	pM1-1	chr15-43992458	A	C	N/A (>30)	Amp 01	
M1-2	chr15-43892646	T	G	Intron 26	N/A (>30)	pM1-2	chr15-43992464	G	T	N/A (>30)	Amp 01	
M1-3	chr15-43893008	T	C	Intron 26	N/A (>30)	pM1-3	chr15-43992826	C	T	N/A (>30)	Amp 01	
M2-1	chr15-43893462	G	A	Intron 24	N/A (>30)	pM2-1	chr15-43993281	A	G	N/A (>30)	Amp 01	
M2-2	chr15-43893466	C	T	Intron 24	N/A (>30)	pM2-2	chr15-43993285	T	C	AC=2 (>30)	Amp 01	
M2-3	chr15-43893476	T	C	Intron 24	N/A (>30)	pM2-3	chr15-43993295	C	T	AC=1 (>30)	Amp 01	
M3-1	chr15-43893830	C	A	Intron 23	N/A (>30)	pM3-1	chr15-43993648	A	C	AC=3 (>30)	Amp 01	
M3-2	chr15-43893859	T	C	Intron 23	N/A (>30)	pM3-2	chr15-43993677	C	T	N/A (>30)	Amp 01	
M3-3	chr15-43894034	A	G	Intron 23	N/A (~30)	pM3-3	chr15-43993854	G	A	AC=1 (~30)	Amp 01	
M4-1	chr15-43895366	C	T	Intron 23	N/A (>30)	pM4-1	chr15-43994876	T	C	N/A (>30)	Amp 01	
M4-2	chr15-43895375	T	C	Intron 23	N/A (>30)	pM4-2	chr15-43994885	C	T	N/A (>30)	Amp 01	
M4-3	chr15-43895417	C	T	Intron 23	N/A (>100)	pM4-3	chr15-43994927	T	C	N/A (>30)	Amp 01	
M5-1	chr15-43895520	C	T	Exon 23	N/A (~120)	pM5-1	chr15-43995030	T	C	N/A (>30)	Amp 01	
M5-2	chr15-43895542	T	A	Exon 23	N/A (~120)	pM5-2	chr15-43995052	A	T	N/A (>30)	Amp 01	
M6-1	chr15-43895679	A	G	Intron 22	N/A (>30)	pM6-1	chr15-43995158	G	A	N/A (>30)	Amp 01	
M6-2	chr15-43895682	G	A	Intron 22	N/A (>30)	pM6-2	chr15-43995161	A	G	N/A (>30)	Amp 01	
M6-3	chr15-43895686	G	C	Intron 22	N/A (>30)	pM6-3	chr15-43995165	C	G	N/A (>30)	Amp 01	

Table 2. The 14 *STRC/STRCP1* divergent marker clusters (continued)

Loci of marker (<i>STRC</i>)	Ref	Alt	Exon / Intron	Allele frequency (Coverage)	Loci of marker (<i>STRCP1</i>)	Ref	Alt	Allele frequency (Coverage)	Amplicon		
M7-1	chr15-43896031	A	G	Intron 22	N/A (>30)	pM7-1	chr15-43995511	G	A	N/A (>30)	Amp 01
M7-2	chr15-43896042	C	T	Intron 22	N/A (>30)	pM7-2	chr15-43995522	T	C	N/A (>30)	Amp 01
M7-3	chr15-43896057	G	A	Intron 22	N/A (>30)	pM7-3	chr15-43995537	A	G	N/A (>30)	Amp 01
M8-1	chr15-43896750	G	A	Intron 20	N/A (>30)	pM8-1	chr15-43996230	A	G	N/A (~30)	Amp 01、02
M8-2	chr15-43896760	A	G	Intron 20	N/A (>30)	pM8-2	chr15-43996240	G	A	N/A (~30)	Amp 01、02
M9-1	chr15-43897383	G	A	Intron 19	N/A (>30)	pM9-1	chr15-43996861	A	G	N/A (~30)	Amp 01、02
M9-2	chr15-43897621	C	T	Intron 18	N/A (>30)	pM9-2	chr15-43997099	T	C	N/A (>25)	Amp 01、02
M9-3	chr15-43897623	G	T	Intron 18	N/A (>30)	pM9-3	chr15-43997101	T	G	N/A (>25)	Amp 01、02
M9-4	chr15-43897624	G	C	Intron 18	N/A (>30)	pM9-4	chr15-43997102	C	G	N/A (>25)	Amp 01、02
M10-1	chr15-43899232	G	A	Intron 18	N/A (~25)	pM10-1	chr15-43998706	A	G	N/A (>25)	Amp 01、02
M10-2	chr15-43899690	T	C	Intron 18	N/A (~30)	pM10-2	chr15-43999164	C	T	N/A (~30)	Amp 01、02
M11-1	chr15-43900188	T	A	Intron 17	N/A (>120)	pM11-1	chr15-43999662	A	T	N/A (~30)	Amp 01、02
M11-2	chr15-43900267	C	T	Intron 17	N/A (>120)	pM11-2	chr15-43999741	T	C	N/A (~30)	Amp 01、02
M12-1	chr15-43900975	G	A	Intron 16	N/A (~25)	pM12-1	chr15-44000447	A	G	N/A (~25)	Amp 01、02
M13-1	chr15-43901397	C	T	Intron 16	N/A (~30)	pM13-1	chr15-44000869	T	C	N/A (~30)	Amp 01、02、03
M13-2	chr15-43901476	C	T	Exon 16	N/A (>120)	pM13-2	chr15-44000948	T	C	N/A (>30)	Amp 01、02、03
M14-1	chr15-43901755	A	G	Intron 15	N/A (~30)	pM14-1	chr15-44001227	G	A	N/A (>25)	Amp 02、03

Marker-mediated refinement for SNV genotyping of STRC/STRCP1 in reference samples

In **Figure 2D**, the MLPA signal peaks on exon 28 of both *STRC* and *STRCP1* (arrows with asterisks) indicate an inconsistent copy number in contrast to the other signal peaks. This implies that there is an SNV present at this peak of *STRC* in HG005 that affects the trend of one-copy gain in the MLPA results. This SNV can be inferred as c.5125A>G (hg19:chr15:43892272-T-C, NM_153700.2:p.T1709A) because it is the only divergent nucleotide within exon 28 between *STRC* (chr15:43892272-T) and *STRCP1* (chr15:43992088-C) (**Figure S3**). This SNV is documented as “likely pathogenic” in ClinVar and is confirmed in the NIST external source data of HG005 but not of HG001 and HG002 (**Figure 3A**).

However, our initial LRS assay mapping results indicated a heterozygous genotype of c.5125A>G in all three reference samples (top panel of **Figure 3B**), which is inconsistent with the NIST external source data. This discrepancy is likely due to pseudogene contamination. To address this, we applied a refinement process based on the *STRC/STRCP1* divergent markers (**Figure 3C**) to filter out mismapped segments using cluster M1, which is closest to the c.5125A>G variant. This refinement improved the mapping results for HG001 and HG002, while the result for HG005 remained unchanged (lower panel of **Figure 3B**). These refined results are consistent with the NIST external source data, demonstrating the effectiveness of marker-mediated refinement for accurate SNV genotyping.

Furthermore, the allele ratio of c.5125A>G in HG005 (NIST external source data) was 35% (Ref:Alt=40:22) rather than 50% (**Figure 3A**), reflecting the one-copy gain of *STRC*. Based on these

results, we propose that reference sample HG005 harbors a one-copy gain of *STRC* with a heterozygous SNV c.5125A>G on *STRC* (**Figure 3D**). Therefore, we established benchmarks using NIST reference samples and applied marker-mediated refinement to our LRS results to ensure consistency with external source data.

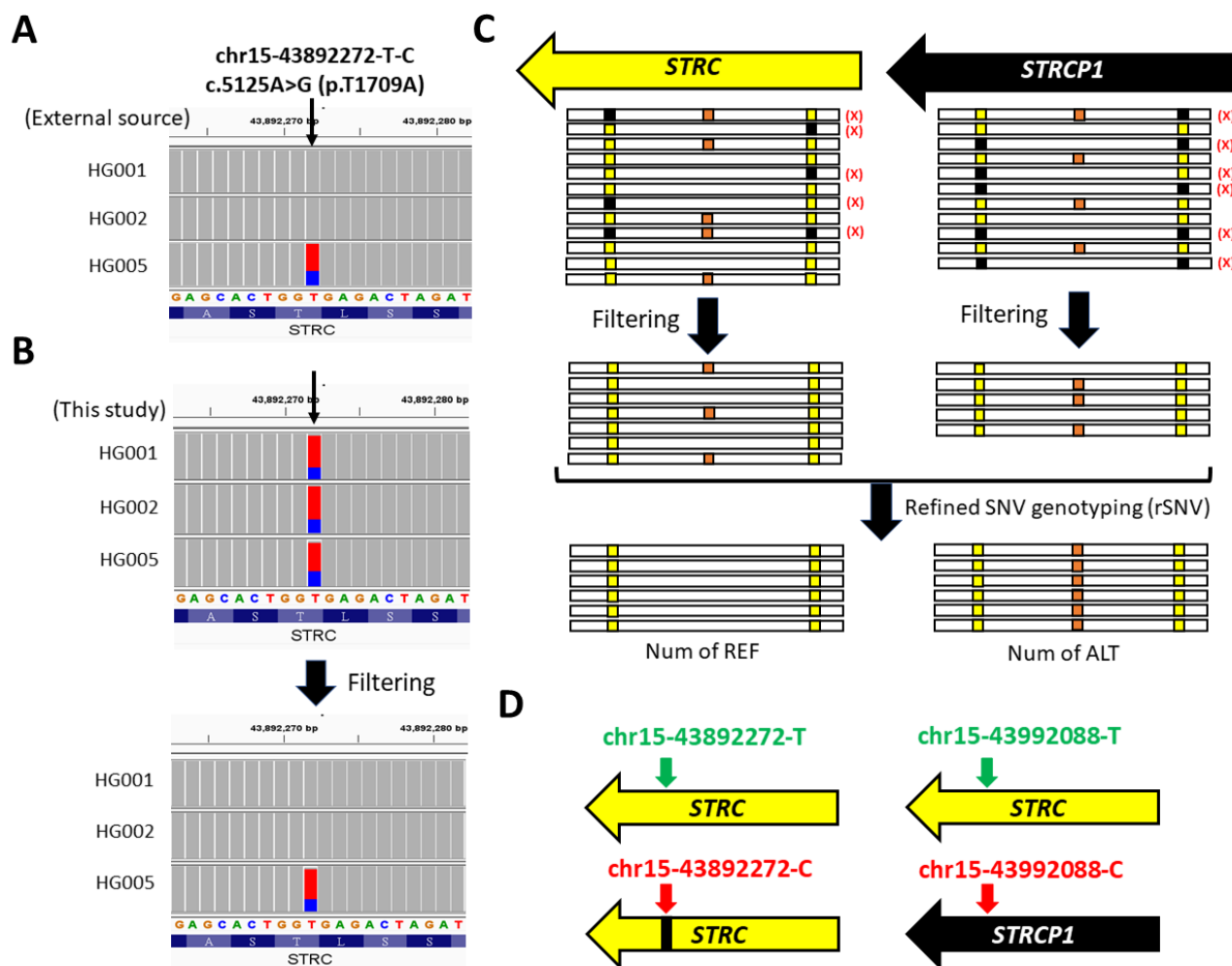


Figure 3. Marker-mediated refinement of single nucleotide variation (SNV) in reference samples. (A) Mapped reads from reference samples (NIST external source data). (B) Illustrative process of marker-mediated refinement for pre-filtered mapped reads resulting in refined SNV (rSNV) calls. (C) Illustrative process of marker-mediated refinement for accurate SNV genotyping of *STRC/STRCP1*. (D) Proposed *STRC/STRCP1* status of reference sample HG005.

STRC variants (CNVs or SNVs) identified in HHI patients

Based on the marker-mediated refinement from reference sample benchmarks, both rCN ratios and refined SNV (rSNV) ratios of *STRC* are calculated for the detection of CNVs and SNVs,

respectively, in each of the HHI patients included in this study. CNV analysis was performed using the following rCN ratio thresholds: two-copy loss (serial rCN ratios < 5% or mapping reads < 10 within *STRC* region), one-copy loss (serial rCN ratios < 30%), one-copy gain (serial rCN ratios > 70%). Samples with potential CNVs were further validated using MLPA assays.

For SNV detection, rSNV ratios were obtained by marker-mediated refinement using the closest marker cluster to each annotated variant. SNV types were defined as follows: homozygote (rSNV ratio > 80%), heterozygote (rSNV ratio 40-60%), and heterozygote with one-copy gain (rSNV ratio 30-35%).

Our analysis revealed bi-allelic *STRC* variants in 11 cases (**Figure 4A**), including homozygous deletion (n=7), compound heterozygosity of deletion and conversion (n=2), and compound heterozygosity of pathogenic SNV and deletion/conversion (n=2). In addition, 16 cases had mono-allelic *STRC* variants, including heterozygous deletion (n=2), conversion (n=2), and SNV (n=12). Analysis of allele frequencies (**Figure 4B**) showed that *STRC* deletions (n=19, 9.5% frequency in the entire cohort) were predominantly found in the mild-to-moderate HHI group (13%). Similarly, *STRC* conversions (n=5, 2.5% frequency) were also predominantly observed in the mild-to-moderate HHI group (2.7%). In contrast, *STRC* SNVs were detected with similar allele counts in both mild-to-moderate (n=8) and severe-to-profound (n=6) groups.

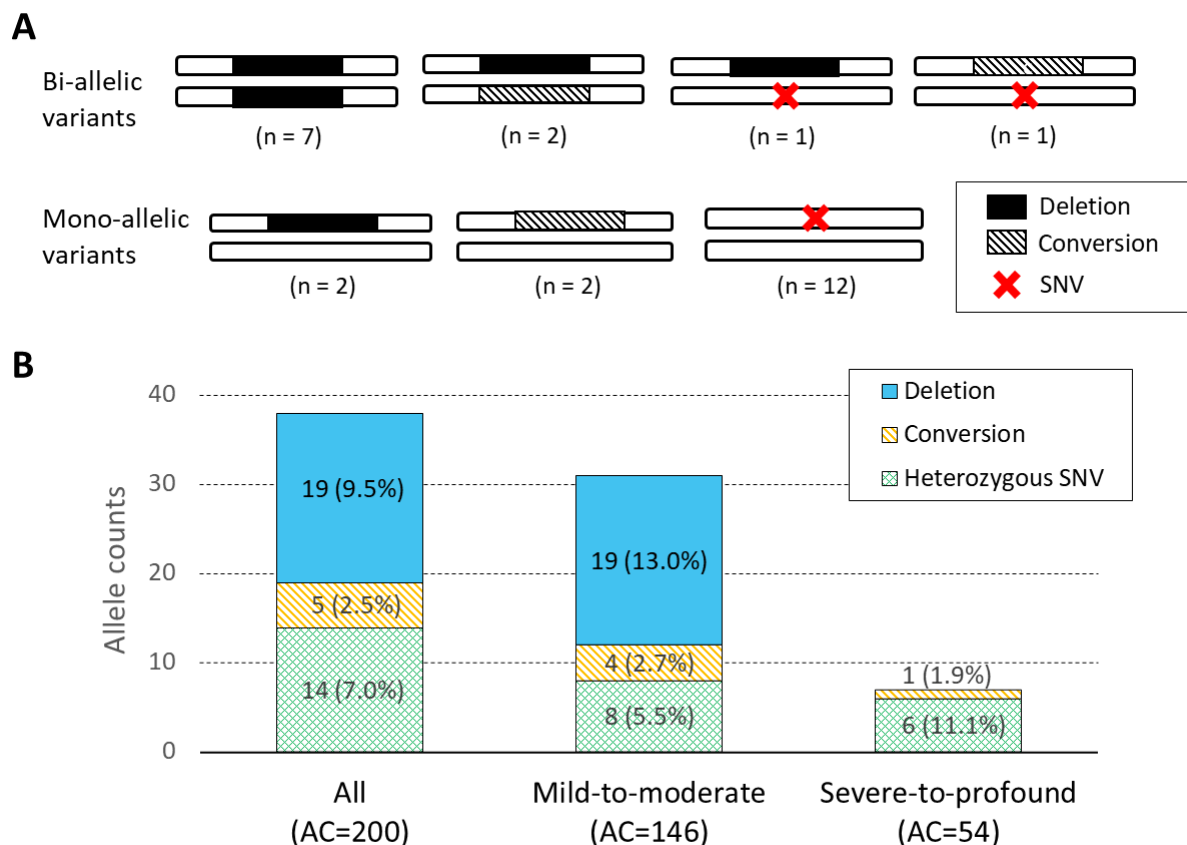


Figure 4. Statistical results of *STRC* variants found in the LRS assay. (A) Distribution of HHI cases with bi-allelic (n=11) and mono-allelic (n=16) variants. (B) Allele counts (AC) of identified CNVs and SNVs in all HHI cases and subgroups (mild-to-moderate, n=73; severe-to-profound, n=27).

Our LRS assays identified both 2-copy and 1-copy losses of *STRC*, as shown by the representative rCN ratio plots (**Figure 5**). These results, combined with MLPA validations, revealed several bi-allelic and mono-allelic CNVs, including deletions and gene conversions. Based on our LRS results, these *STRC* deletions and conversions spanned from 3' end to intron 18 (markers M1 to M9) or to intron 15 (markers M1 to M14).

In addition, we identified four pathogenic SNVs in *STRC* (**Table 3**). All SNVs were heterozygous with sufficient rSNV ratios for confident identification. These included c.5125A>G (n=9), c.4622G>A (n=3), c.4402C>T (n=1), and c.4143G>A (n=1). Of note, c.5125A>G (p.T1709A)

represents the only divergent nucleotide between *STRC* and its pseudogene *STRCP1* reference transcript in exon 28 (i.e., the Alt nucleotide in *STRC* being as Ref nucleotide in *STRCP1*; see **Figure S3**). This variant is likely pathogenic based on predictive scores and ClinVar assertions (rs1336307815). The c.4622G>A (p.R1541Q) also shows likely pathogenicity according to predictive scores. In addition, we identified a previously reported pathogenic variant, c.4402C>T (p.R1468X), and a novel pathogenic variant, c.4143G>A (p.W1381X), both predicted to result in protein truncation.

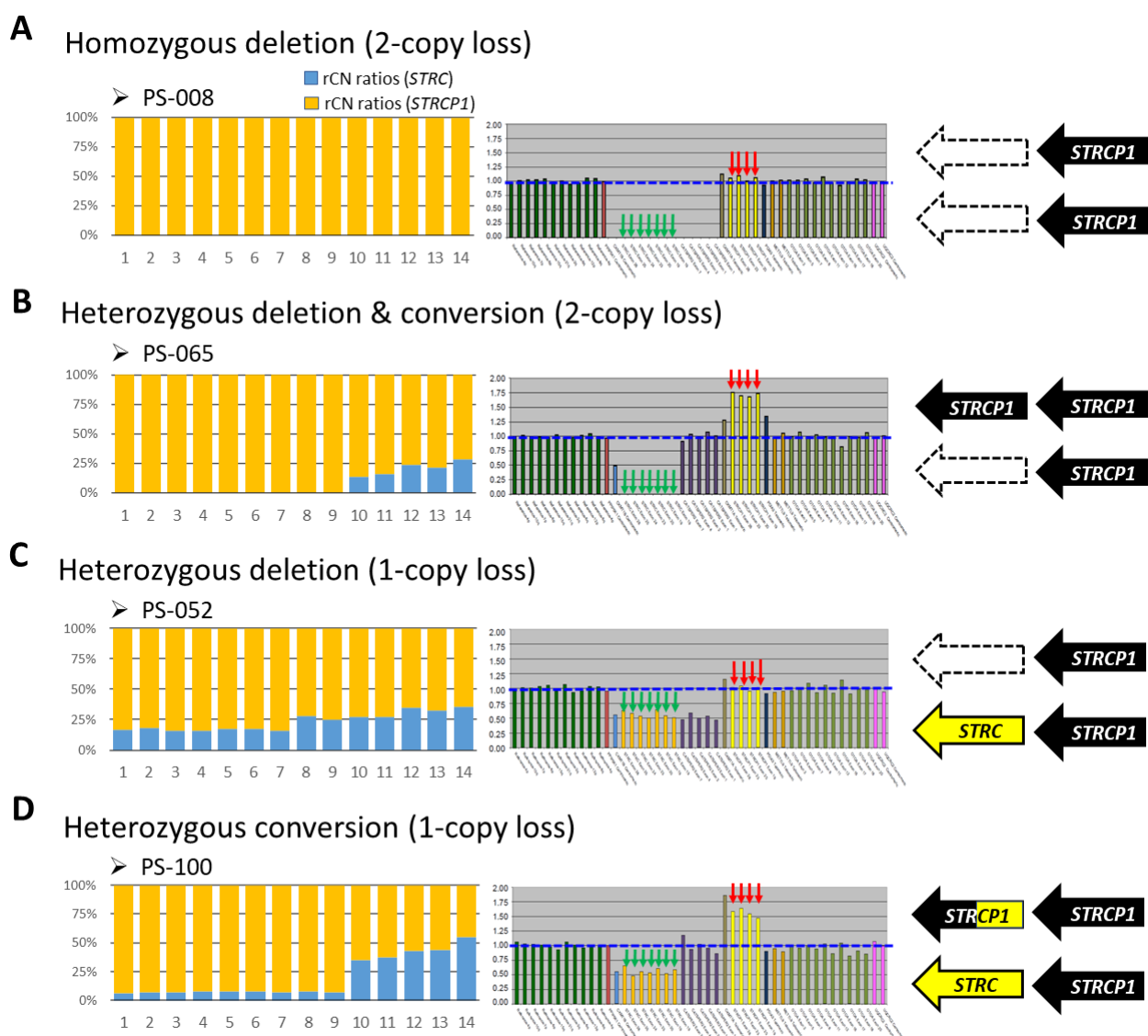


Figure 5. Four types of CNVs identified in this study, characterized by aberrant rCN ratios and MLPA peaks. (A) Homozygous deletions (PS-008 as an example); (B) Compound heterozygous deletion and conversion (PS-065 as an example); (C) Heterozygous deletion (PS-052 as an example); (C) Heterozygous conversion (PS-100 as an example).

Table 3. Pathogenic SNVs of *STRC* confirmed in this study.

Variant loci (GRCh37) HGVS (NM_153700.2) Reference SNP ID (dbSNP)	Exon	SIFT	PolyPhen-2	CADD #	SpliceAI †	ClinVar	DVD	Grpmax_AF § (gnomAD)	MAF* (TWB)	Marker cluster ¶	Cases	Ref
<i>I5-43892272-T-C</i>												
<i>c.5125A>G</i> (p.T1709A) (rs1336307815)	28/29	0.02 (D)	0.994 (D)	23.9	0.08 (DS-AL)	LP	LP	N/A	N/A	M1	9	[8]
<i>c.4622G>A</i> (p.R1541Q) <i>I5-43893673-C-T</i> (rs77493867)	24/29	0.17 (T)	0.876 (PD)	22.2	0.02 (DS-AL)	N/A	P	0.00251 (EAS)	0.0047	M1	3	[9]
<i>c.4143G>A</i> (p.W1381X) <i>I5-43896634-C-T</i> (No RSID)	21/29	N/A	N/A	41	0.06 (DS-AL)	N/A	N/A	N/A	N/A	M8	1	Novel
<i>c.4402C>T</i> (p.R1468X) <i>I5-43895583-G-A</i> (rs377480477)	23/29	N/A	N/A	39	0.06 (DS-AG)	P	P	0.000140 (EUP)	N/A	M8	1	[8]

CADD PHRED score of the targeted variant (CADD ver.1.7).

† SpliceAI maximal DS score of the targeted variant (parameter of maximum distance: 500bp). DS (delta score): probability that targeted variant causes altered splicing events at the predicted position (score scale: 0-1 with recommended threshold 0.5); AL (Acceptor Loss); AG (Acceptor Gain).

§ Grpmax_AF: The maximal sub-population allele frequency in the gnomAD database [43] (ver. 2.1.1, last accessed on Oct 10, 2024). EAS: East Asian; EUP: European.

* MAF: The minor allele frequency (database: whole genome sequence) in the Taiwan Biobank [44] (<https://taiwanview.twbiobank.org.tw/variant.php>, last accessed Oct 19, 2024).

¶ Marker cluster: The selected cluster of diverse markers documented in Table 2 for marker-mediated SNV refinements.

(Other abbreviations) HGVS: Human Genome Variation Society; (SIFT & PolyPhen-2) D: damaging; PD: possibly damaging; T: tolerant

Compound heterozygosity of SNVs and CNVs in HHI cases

Among these HHI patients with confirmed *STRC* variants, we identified two individuals, PS-071 and PS-032, harboring compound heterozygous variants consisting of a SNV and a CNV. In patient PS-071, we observed homozygosity of the protein truncating variant c.4402C>T (**Figure 6A**). In addition, LRS and MLPA analysis revealed gene conversion spanning marker clusters from M1 to M12 in the LRS rCN ratio plot (**Figure 6B**). These findings confirm compound heterozygosity (c.[4402C>T];[conversion]) in PS-071.

Patient PS-032 was recruited from a multiplex family with an affected sibling (PS-036), an unaffected father (PS-037), and an unaffected mother (PS-038). A novel protein-truncating variant, c.4143G>A, was identified with homozygosity in both affected siblings and heterozygosity in the father (**Figure 6C**). Subsequent LRS and MLPA analysis revealed a heterozygous deletion spanning marker clusters from M1 to M9 in the LRS rCN ratio plot. This deletion, present in both affected siblings, was inherited from the maternal allele. (**Figure 6D & Figure S4**). These results confirmed the compound heterozygosity (c.[4143G>A];[deletion]) in both PS-032 and PS-036, indicating the autosomal recessive inheritance in this multiplex family.

However, in unaffected mother (PS-038), LRS analysis revealed missing rCN ratios for *STRC* from M10 to M14 (intron 15 to 18), a region beyond the detection range of the MLPA validation (**Figure S1**). Given the mother's normal hearing phenotype, we conclude that she carries a benign partial gene conversion. Analysis of divergent nucleotides by *STRC/STRCP1* pairwise alignment (**Figure S5**) revealed only three exonic variants within the region spanning exons 1 to 18. All three

are synonymous variants where the *STRC* sequence is directly replaced by the homologous *STRCP1* segment (**Table S2**), suggesting that this partial pseudogene conversion is likely harmless.

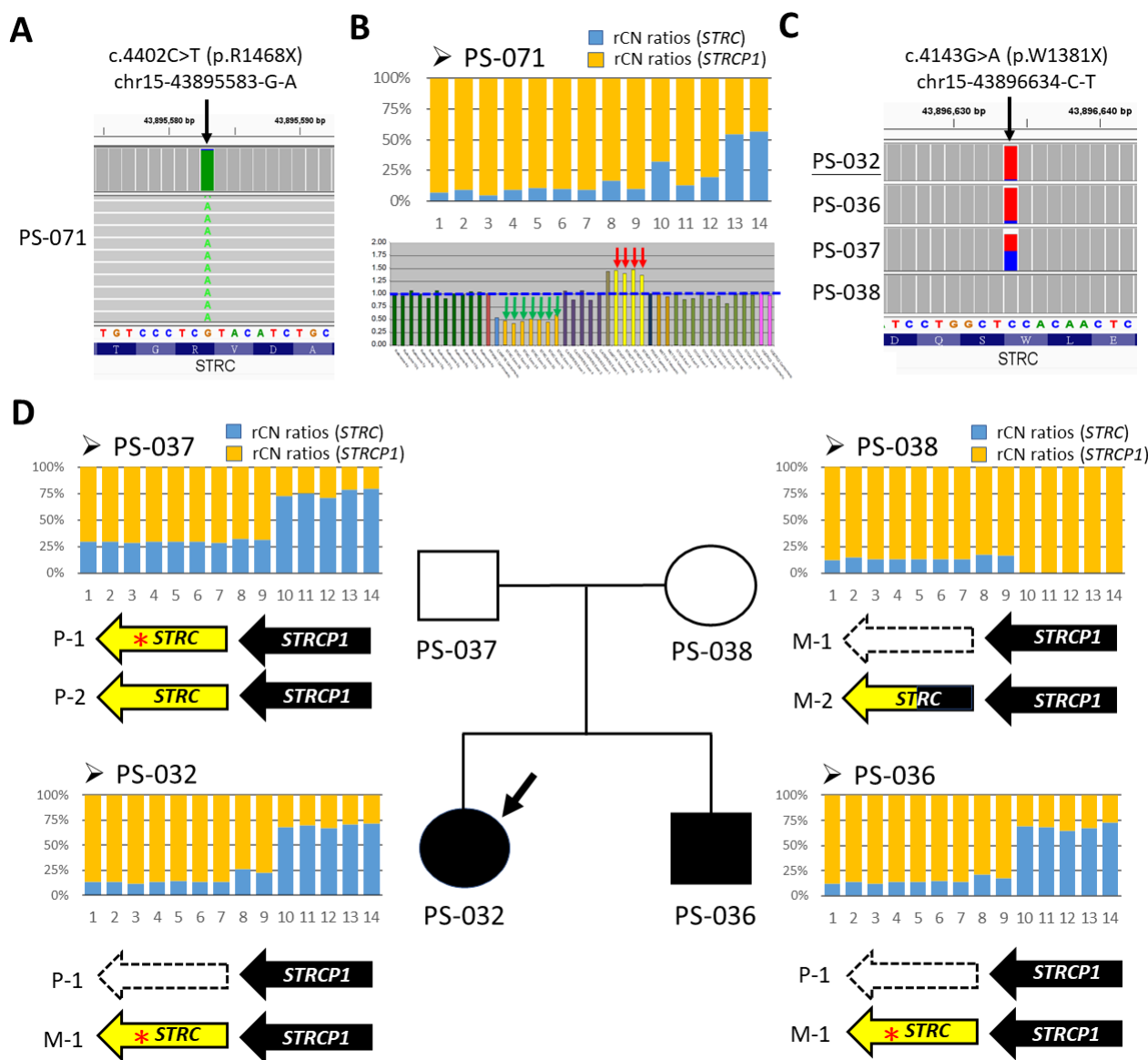


Figure 6. Compound heterozygous genotypes in two unrelated patients with truncating variants and either gene conversion or deletion. (A) c.4402C>T (p.R1468X) in PS-071 from a simplex family, forming a compound heterozygous genotype with (B) a single copy loss caused by gene conversion of *STRCP1*. (C) c.4143G>A (p.W1381X) in PS-032 and PS-036, inherited from the paternal allele, coupled with (D) a large deletion of *STRC*, inherited from the maternal allele, confirming compound heterozygosity in an autosomal recessive inheritance pattern.

Discussion

In this study, we used the amplicon-based LRS assays to investigate the genetic causes of *STRC*-related SNHI in 100 Taiwanese HHI patients unresolved by NGS-based diagnostics. Through marker-mediated refinement and MLPA validation, we identified disease-causing *STRC* variants, including CNVs and SNVs. Our LRS assay detected bi-allelic *STRC* variants in 11 patients, establishing a diagnostic yield of 11% for DFNB16 in this cohort. Considering both bi-allelic and mono-allelic variants, our results indicate a high allele frequency (19%) of *STRC* variants in 27 HHI patients. Notably, 81.6% of these variants were identified in patients with mild-to-moderate SNHI. This is the first study to report the genetic epidemiology of *STRC*-related SNHI using LRS, which demonstrated a prevalence of DFNB16 in the Taiwanese HHI cohort comparable to other East Asian populations [12, 47].

Previous studies across various populations have highlighted the significant contribution of pathogenic *STRC* variants to HHI [17]. In particular, NGS-based studies in Asian populations, complemented by CNV validation methods, have reported a high frequency of CNVs in patients with mild-to-moderate SNHI [12, 13]. While NGS effectively detects homozygous SNVs and deletions within the highly divergent region of *STRC* (exons 18-29) [6, 48, 49], the presence of the homologous pseudogene *STRCP1* and potential gene-to-pseudogene conversion events can complicate variant interpretation [18]. These challenges can lead to reduced coverage and hinder accurate variant annotation and analysis [50, 51]. Although MLPA is a widely used tool for CNV detection [6, 13, 52], it may not effectively capture the CNVs within the *STRC/STRCP1* highly-homologous region

spanning from exons 1 to 18 [13, 51].

LRS technologies, with their ability to sequence long genomic regions (kilobases to megabases) [53], offer a potential solution for the comprehensive detection of both SNVs and CNVs [28, 54]. While earlier LRS methods suffered from high error rates (~5-20% base calling errors)[55], recent advances in the PacBio and ONT platforms have significantly improved accuracy, with error rates now down to <1% [29] and <5% [56], respectively. This increased accuracy makes LRS suitable for clinical applications.

In this study, we used a PacBio-based LRS approach, following the recommended amplicon-based workflow, to decipher hidden *STRC* variants in a Taiwanese HHI cohort. Although pseudogene contamination in our LRS assays initially hindered variant interpretation, we developed a marker-mediated refinement strategy based on *STRC/STRCPI* divergent marker clusters to address it. This approach exploits the concept of genetic linkage [57], which suggests that proximal genetic loci are more likely to occur on the same haplotype. The performance of our LRS-based refinements has been validated using NIST reference samples, with results consistent with external sources and our own experiments.

For CNV detection, our refined copy number (rCN) ratios derived from the divergent marker cluster (M1 to M14) effectively addressed pseudogene contaminations and provided normalized results integrated with *STRCPI* screening data. Our rCN ratios also cover a broader range (intron 15 to exon 29) than conventional MLPA assays (exon 19 to exon 28). For SNV detection, the confirmed rSNVs in our study (**Table 3**) represent high-confidence genotyping in *STRC*. We also identified false-

positive pathogenic variants resulting from pseudogene contaminations in our LRS assays due to low-confidence allele ratios of alternative nucleotides (**Table S3**). These variants are characterized by opposite REF/ALT nucleotides between *STRC* and *STRCPI*. Compared to the short reads of NGS, LRS offers the advantage of long-range sequence information, allowing the inclusion of more divergent markers near the targeted SNV for refinement.

To our knowledge, this is the first study to apply LRS technology to detect genetic causes in a large HHI cohort. Our study demonstrates the power of LRS in addressing challenging variants in *STRC* and provides valuable insights into the genetic etiology of HHI that remain unresolved by conventional NGS diagnostics. However, several limitations of this study deserve discussion. First, the limitations of long-range PCR, coupled with the recommended length of the CCS approach for HiFi reads, restricted amplicon generation to 10 kb instead of 20 kb, limiting the full coverage of the entire *STRC* region. Second, the lack of divergent markers in the highly homologous region spanning exons 1 to 15, which was previously reported to have nearly 100% identity between *STRC* and *STRCPI* [9], may hinder the refinement process for *STRC* variants within in this region. To improve clinical performance in the future, the long-range PCR workflow targeting the *STRC/STRCPI* divergence sites [8, 9] could be combined with amplicon-based LRS assays to reduce pseudogene contaminations.

Conclusion

Our study highlights the diagnostic potential of LRS for detecting challenging variants in *STRC* and resolving the genetic etiology of HHI that remain unresolved by conventional NGS. The high allele frequency (19%) of *STRC* variants observed in this cohort emphasizes the importance of comprehensive *STRC* screening in HHI patients, especially those with mild-to-moderate SNHI. Future studies incorporating LRS-based *STRC* screening in clinical genetic testing are needed to further evaluate its diagnostic utility.

Acknowledgements

We sincerely thank the A1 Laboratory of Genetic Testing of National Taiwan University Hospital (NTUH), GenePhile Bioscience Laboratory of Ko's Obstetrics and Gynecology Clinic, Blossom Biotechnologies, Inc. (Pacific Biosciences distributor in Taiwan), and Taiwan Genome Industry Alliance Inc. (Taipei, Taiwan) for their invaluable experimental resources and technical support. We also thank all the subjects and their families for their generous contributions to this study.

Ethics approval and consent to participate

All the patients and/or their families signed an informed consent form before participating in the study. All procedures used in the study were approved by the Research Ethics Committee of the National Taiwan University Hospital (201104025RC).

Consent for publication

All patients have provided written informed consent.

Data availability:

The resultant datasets in this study are included within the article. Full datasets are available from the corresponding authors upon reasonable request.

Funding Statements:

This study was supported by research grants from the National Science and Technology Council of the Executive Yuan of Taiwan (NSTC 110-2314-B-002-189-MY3, Chen-Chi Wu), National Health Research Institutes grant (NHRI-EX111-10914PI, Chen-Chi Wu), National Taiwan University Hospital & National Taiwan University Joint Program grants (111-UN0048, Chen-Chi Wu & Jacob Shu-Jui Hsu), and National Taiwan University Hospital Hsin-Chu Branch & National Health Research Institutes Joint Program grant (NHRI-113-B02, Chen-Chi Wu & Shih-Feng Tsai)

Competing interests:

The authors declare that they have no competing interests.

Authors' contributions

Conceptualization: *Cheng-Yu Tsai, Pei-Lung Chen, Jacob Shu-Jui Hsu, and Chen-Chi Wu*

Investigation: *Cheng-Yu Tsai and Yu-Ting Chiang*

Validation: *Cheng-Yu Tsai* and *Yue-Sheng Lu*

Formal analysis: *Cheng-Yu Tsai*

Visualization: *Cheng-Yu Tsai*

Resources: *Pei-Hsuan Lin*, *Chuan-Jen Hsu*, and *Pei-Lung Chen*

Project administration: *Cheng-Yu Tsai*, *Yue-Sheng Lu*, and *Ming-Yu Lo*

Funding Acquisition: *Shih-Feng Tsai*, *Jacob Shu-Jui Hsu*, and *Chen-Chi Wu*

Supervision: *Jacob Shu-Jui Hsu*, and *Chen-Chi Wu*

Writing (Original Draft): *Cheng-Yu Tsai*, and *Chen-Chi Wu*

Writing (Review/Editing): *Cheng-Yu Tsai*, *Pei-Lung Chen*, *Jacob Shu-Jui Hsu*, and *Chen-Chi Wu*

Reference

1. Shearer, A.E., et al., *Genetic Hearing Loss Overview*. GeneReviews® [Internet], edited by Margaret P Adam et. al. University of Washington, Seattle., 1999. **Available from: <https://www.ncbi.nlm.nih.gov/books/NBK1434/>**.
2. Tsai, C.-Y., et al., *Implementing next-generation sequencing for diagnosis and management of hereditary hearing impairment: a comprehensive review*. Expert Review of Molecular Diagnostics, 2024: p. 1-13.
3. Verpy, E., et al., *Stereocilin connects outer hair cell stereocilia to one another and to the tectorial membrane*. Journal of comparative neurology, 2011. **519**(2): p. 194-210.
4. Zhang, Y., et al., *Sensorineural deafness and male infertility: a contiguous gene deletion syndrome*. Journal of medical genetics, 2007. **44**(4): p. 233-240.
5. Markova, T., et al., *Clinical features of hearing loss caused by STRC gene deletions/mutations in Russian population*. International Journal of Pediatric Otorhinolaryngology, 2020. **138**: p. 110247.
6. Morgan, A., et al., *Lights and shadows in the genetics of syndromic and non-syndromic hearing loss in the Italian population*. Genes, 2020. **11**(11): p. 1237.
7. Plevova, P., et al., *STRC deletion is a frequent cause of slight to moderate congenital hearing impairment in the Czech Republic*. Otology & Neurotology, 2017. **38**(10): p. e393-e400.
8. Vona, B., et al., *DFNB16 is a frequent cause of congenital hearing impairment: implementation of STRC mutation analysis in routine diagnostics*. Clinical genetics, 2015. **87**(1): p. 49-55.
9. Mandelker, D., et al., *Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology*. The Journal of Molecular Diagnostics, 2014. **16**(6): p. 639-647.
10. Amr, S.S., et al., *Allele-specific droplet digital PCR combined with a next-generation sequencing-based algorithm for diagnostic copy number analysis in genes with high homology: proof of concept using stereocilin*. Clinical Chemistry, 2018. **64**(4): p. 705-714.
11. Ito, T., et al., *Rapid screening of copy number variations in STRC by droplet digital PCR in patients with mild-to-moderate hearing loss*. Human Genome Variation, 2019. **6**(1): p. 1-6.
12. Kim, B.J., et al., *Significant Mendelian genetic contribution to pediatric mild-to-moderate hearing loss and its comprehensive diagnostic approach*. Genetics in Medicine, 2020. **22**(6): p. 1119-1128.
13. Nishio, S.-y. and S.-i. Usami, *Frequency of the STRC-CATSPER2 deletion in STRC-associated hearing loss patients*. Scientific Reports, 2022. **12**(1): p. 634.
14. Sloan-Heggen, C.M., et al., *Comprehensive genetic testing in the clinical evaluation of 1119 patients with hearing loss*. Human genetics, 2016. **135**(4): p. 441-450.
15. Ho, S.S., A.E. Urban, and R.E. Mills, *Structural variation in the sequencing era*. Nature Reviews Genetics, 2020. **21**(3): p. 171-189.

16. Abbasi, W., et al., *Evaluation of copy number variants for genetic hearing loss: a review of current approaches and recent findings*. Human Genetics, 2022: p. 1-14.
17. Han, S., et al., *Prevalence and characteristics of STRC gene mutations (DFNB16): a systematic review and meta-analysis*. Frontiers in genetics, 2021. **12**: p. 707845.
18. Shearer, A.E., et al., *Copy number variants are a common cause of non-syndromic hearing loss*. Genome medicine, 2014. **6**(5): p. 1-10.
19. Mahmoud, M., et al., *Structural variant calling: the long and the short of it*. Genome biology, 2019. **20**(1): p. 1-14.
20. van Belzen, I.A., et al., *Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology*. NPJ Precision Oncology, 2021. **5**(1): p. 1-11.
21. Shearer, A.E., et al., *Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing*. Proceedings of the National Academy of Sciences, 2010. **107**(49): p. 21104-21109.
22. Hoppman, N., et al., *Genetic testing for hearing loss in the United States should include deletion/duplication analysis for the deafness/infertility locus at 15q15*. 3. Molecular Cytogenetics, 2013. **6**(1): p. 1-5.
23. Moteki, H., et al., *Detection and confirmation of deafness-causing copy number variations in the STRC gene by massively parallel sequencing and comparative genomic hybridization*. Annals of Otology, Rhinology & Laryngology, 2016. **125**(11): p. 918-923.
24. Os, P.G.E.-V. and J.P. Schouten, *Multiplex Ligation-dependent Probe Amplification (MLPA®) for the detection of copy number variation in genomic sequences*. PCR Mutation detection protocols, 2011: p. 97-126.
25. Mochizuki, T., et al., *Mutation analyses by next-generation sequencing and multiplex ligation-dependent probe amplification in Japanese autosomal dominant polycystic kidney disease patients*. Clinical and Experimental Nephrology, 2019. **23**: p. 1022-1030.
26. Rhoads, A. and K.F. Au, *PacBio sequencing and its applications*. Genomics, Proteomics and Bioinformatics, 2015. **13**(5): p. 278-289.
27. Wang, Y., et al., *Nanopore sequencing technology, bioinformatics and applications*. Nature biotechnology, 2021. **39**(11): p. 1348-1365.
28. Oehler, J.B., et al., *The application of long-read sequencing in clinical settings*. Human genomics, 2023. **17**(1): p. 73.
29. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome*. Nature biotechnology, 2019. **37**(10): p. 1155-1162.
30. Lee, Y.-H., et al., *Revisiting Genetic Epidemiology with a Refined Targeted Gene Panel for Hereditary Hearing Impairment in the Taiwanese Population*. Genes, 2023. **14**(4): p. 880.
31. Zook, J.M., et al., *Extensive sequencing of seven human genomes to characterize benchmark reference materials*. Scientific data, 2016. **3**(1): p. 1-26.

32. PacBio, *pbbm2*. <https://github.com/PacificBiosciences/pbbm2>, 2023. **ver 1.13**.
33. Poplin, R., et al., *A universal SNP and small-indel variant caller using deep neural networks*. Nature biotechnology, 2018. **36**(10): p. 983-987.
34. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data*. Bioinformatics, 2011. **27**(21): p. 2987-2993.
35. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data*. Nucleic acids research, 2010. **38**(16): p. e164-e164.
36. Li, H., et al., *The sequence alignment/map format and SAMtools*. bioinformatics, 2009. **25**(16): p. 2078-2079.
37. Masella, A.P., et al., *BAMQL: a query language for extracting reads from BAM files*. BMC bioinformatics, 2016. **17**: p. 1-6.
38. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function*. Nucleic acids research, 2003. **31**(13): p. 3812-3814.
39. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2*. Current protocols in human genetics, 2013. **76**(1): p. 7.20. 1-7.20. 41.
40. Liu, X., et al., *dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs*. Genome medicine, 2020. **12**: p. 1-8.
41. Schubach, M., et al., *CADD v1. 7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions*. Nucleic acids research, 2024. **52**(D1): p. D1143-D1154.
42. Jaganathan, K., et al., *Predicting splicing from primary sequence with deep learning*. Cell, 2019. **176**(3): p. 535-548. e24.
43. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans*. Nature, 2020. **581**(7809): p. 434-443.
44. Wei, C.-Y., et al., *Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese*. NPJ genomic medicine, 2021. **6**(1): p. 1-10.
45. Azaiez, H., et al., *Genomic landscape and mutational signatures of deafness-associated genes*. The American Journal of Human Genetics, 2018. **103**(4): p. 484-497.
46. *Genome in a Bottle (GIAB)*. National Institute of Standards and Technology (NIST). <https://www.nist.gov/programs-projects/genome-bottle> (last accessed on 2024/10/16).
47. Yokota, Y., et al., *Frequency and clinical features of hearing loss caused by STRC deletions*. Scientific Reports, 2019. **9**(1): p. 4408.
48. Schrauwen, I., et al., *A sensitive and specific diagnostic test for hearing loss using a microdroplet PCR-based approach and next generation sequencing*. American Journal of Medical Genetics Part A, 2013. **161**(1): p. 145-152.

49. Kannan-Sundhari, A., et al., *Screening consanguineous families for hearing loss using the miamiotogenes panel*. Genetic testing and molecular biomarkers, 2020. **24**(10): p. 674-680.
50. Imizcoz, T., et al., *Next-generation sequencing improves precision medicine in hearing loss*. Frontiers in Genetics, 2023. **14**: p. 1264899.
51. Downie, L., et al., *Exome sequencing in infants with congenital hearing impairment: a population-based cohort study*. European Journal of Human Genetics, 2020. **28**(5): p. 587-596.
52. Shearer, A.E., et al., *Advancing genetic testing for deafness with genomic technology*. Journal of medical genetics, 2013. **50**(9): p. 627-634.
53. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis*. Genome biology, 2020. **21**(1): p. 1-16.
54. Logsdon, G.A., M.R. Vollger, and E.E. Eichler, *Long-read human genome sequencing and its applications*. Nature Reviews Genetics, 2020. **21**(10): p. 597-614.
55. Xiao, T. and W. Zhou, *The third generation sequencing: the advanced approach to genetic diseases*. Translational pediatrics, 2020. **9**(2): p. 163.
56. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads*. Nature biotechnology, 2018. **36**(4): p. 338-345.
57. Pulst, S.M., *Genetic linkage analysis*. Archives of neurology, 1999. **56**(6): p. 667-672.

Graphical Abstract

