Author's response to reviews

Title: Combined examination of sequence and copy number variations in human deafness genes better explained the cause of genetic deafness

Authors:

Haiting Ji (haiting123456@163.com)
Jingqiao Lu (jlu30@emory.edu)
Jianjun Wang (jwang33@emory.edu)
Huawei Li (hwli@shmu.edu.cn)
Xi Lin (xlin2@emory.edu)

Version: 4 Date: 20 June 2014

Author's response to reviews: see over
June 11, 2014
Xi Lin, Professor
Dept of Otolaryngology
Emory University School of Medicine
615 Michael Street
Whitehead Building, Rm# 543
Atlanta, GA 30322

Dr. Ciara Ni Dhubhghaill
e-mail: editorial@biomedcentral.com
BMC Ear, Nose and Throat Disorders
Re: manuscript# 9871574111269921

Dear Dr. Dhubghaill:

We are submitting the revised manuscript (manuscript# 9871574111269921) for consideration by the BMC Ear, Nose and Throat Disorders for publication. We have made changes suggested by the Editors (e.g., format of authors’ contributions, copy editing by a native English speaker, etc.), and checked the formatting of the paper for the requirements of your journal. A list of point-by-point answers to all the questions raised by the Reviewers is given in the next few pages.

We hope that we have satisfactorily answered all the questions and concerns from the last review. We are looking forward to hearing a decision from the Editors at BMC Ear, Nose and Throat Disorders.

Sincerely

Xi Lin, Professor
404-727-3723
xlin2@emory.edu
Point-by-point answers to Reviewers’ comments:

**First Reviewer**

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Not suitable for publication unless extensively edited

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:** I am part of the non-profit Molecular Otolaryngology & Renal Research Laboratories at the University of Iowa that offers a similar genetic testing panel for deafness. Our laboratory is non-profit and I do not have a conflict of interest in evaluating this manuscript.

**MAJOR COMPULSORY REVISIONS**

1. The authors describe a new method for CNV analysis ...... Several other groups have created similar methods. There are more than 20 other methods described for determining CNVs from MPS data in Zhao, Min, et al. ...... BMC Bioinformatics 14.Suppl 11 (2013): S1. The authors should ideally compare their novel method to one of these methods or at least indicate for what reason they did not use a previously published method.

---Answer: We have cited the paper (BMC Bioinformatics 14.Suppl 11 (2013): S1.) suggested by this Reviewer in the revised manuscript. We have added the following writings to address this question:

**Page7 line176 (Method Section):** We added “Compared to previous methods used to detect CNVs from whole human genome or exome NGS data [9], this is the first method used to specifically detect CNVs from NGS data obtained in a disease (e.g., deafness) panel. …..”.

To further clarify the differences between our method and published CNV detection methods, we have added (starting from page14 line318) following writing in the Discussion: “Most published methods on CNV detections from NGS data are designed for examining CNVs on the whole genome or whole exome scale, with specific mathematical models applied in the algorithms [9]. For example, JointSLM and ExoCNVT are designed to detect common CNVs shared among many samples. CoNIFER and XHMM are made to detect rare CNVs using population data. In our approach, we used the average read-depth of samples hybridized together for any particular exon covered, which greatly reduced the variations introduced by GC content, capture efficiency, and alignment biases (Fig. 1). The optimal detection length for CNVs also depends on the specific program used. Pindel is effective for the detection of small deletions less than 300 bp; in contrast, most prior read-depth based bioinformatic tools are used to identify large CNVs (e.g., >10,000 bps). The resolution of our method of detecting CNVs is the size of single exon, which varies between about 100 to a few thousand bps. The average size of targeted exons for the 80 deafness genes is 4491 bps (supplemental Table1). Drawbacks of our approach include the inability to detect the breakpoints in CNVs, as well as common CNVs presented in many samples. …..”.

2. There is a lack of information regarding the patients, ...... Please provide a table that lists each patient, clinical data (for example, age, onset of hearing loss, type of hearing loss, family history, etc), candidate variants, and causative variants.

---Answer: A new table (supplemental Table 5) was added, in which we provided information about patient age at the onset of hearing loss, results of pure-tone audiogram, family history of hearing loss, point mutations found and CNVs identified, and other information summarized in the table.

3. The overall results were confusing. What was the overall diagnostic rate? The number seemed to vary between 21-29. Please also more clearly indicate the overall results of the study with either a table or a simple graph showing total number of patients, patients given a diagnosis, and by what type of variant (CNV, SNV, indel). …..
---Answer: Overall diagnostic rate changes depending on criteria used. In our statements, we always define the criteria (Supplemental Table 2) used first. The information demanded by this Reviewer is now provided in the new Supplemental Table 5 for 79 patients studied.

4. Please explain in the Methods section why 4 different analysis methods were used. …… Please describe how the analysis methods differed and what the overall differences were between the methods. …… What tools in Galaxy were used? In addition, on page 6 the authors refer to a “custom knowledge database.” Please describe what this is.

---Answer: To answer the question of “Please explain in the Methods section why 4 different analysis methods were used”, we added (page5 line124) “Systematic comparison of different tools in population-scale genomic CNV analysis found notable differences between different methods used in identifying genomic regions ascertained, size-range and breakpoint [8]. Since only partially-overlapping CNVs are identified using different methods, some studies used as many as 36 CNV call-sets to help improve the accuracy of finding high-confidence CNVs [8]. In our study, FASTQ data files generated after sequencing with the Illumina HiSeq2000 were processed using four independent bioinformatic data processing pipelines, since the use of multiple independent bioinformatic pipelines help improve accuracy in finding high-confidence CNVs [9]. ……”

Answers to this question “Please describe how the analysis methods differed and what the overall differences were between the methods.” was provided previously when answering Quesiton#1.

To address “…… the Galaxy platform described in the Methods section is a framework for analysis of MPS. …… What tools in Galaxy were used?”, we added (staring from page6 line133): “……A web-based open source platform called Galaxy (https://usegalaxy.org/), which runs on a local Linux server. The tools we used on Galaxy Platform were: (a) BWA (ver. 0.7.4), to generate SAM (Sequence Align/Map) files; (b) Samtools (ver.0.1.19), which were used to transform Binary SAM into BAM and were sorted with samtools; (c) Picard (ver. 1.79), this was used to remove PCR duplicates in the sorted BAM files; (d) GenomeAnalysisTK-1.6 (GATK-1.6). The duplicate-removed BAM files were used as inputs of GATK-1.6 for InDel re-alignment and base quality recalibration using known InDels from dbSNP137 and the 1000 Genome project. Target region coverage and VCF (Variant Calling Format) files of SNP/InDel calling were generated by GATK, based on processed BAM files.”.

To address: “…… on page 6 the authors refer to a “custom knowledge database.” Please describe what this is. “, we added (starting from page7 line158): “……Variations contained in VCF reports were filtered by a custom knowledge database in order to identify candidate disease-causing mutations. Information in this database was collected from Human Genome Mutation Database (HGMD), data on unpublished mutation and normal hearing control NGS data the authors, and consensus predications made by both PolyPhen and Shift bioinformatic algorithms [10]. ”.

5. In the Methods section, line 133, the authors indicate that they used Sanger sequencing to validate the platform. What does N=8 refer to? …… Was this validation for CNVs or SNVs or both? The authors indicate that a “high percentage of detected sequence variations . . . agreed with each other.” A formal sensitivity, specificity, positive predictive value, and negative predictive value analysis should be performed along with data indicating which variants were screened for this analysis.

---Answer: N=8 refer to the number of samples we did both NGS and Sanger sequencing. The validation was for SNPs only. CNVs were validated by independent qPCR method described later in the paper. To make this more clearly, we modified the text (starting from page6 line149): “……Sequencing results obtained by the NGS method for the coding region of the GJB2 gene were compared to those obtained by the Sanger method (N=8) for validation purposes. A high percentage (99.2%) of detected sequence variations (SNPs) in the GJB2 gene as identified by the two methods agreed with each other.”.
The purpose of this section of method description was to show that NGS methods in our hands created credible sequencing results, as validated by the independent Sanger sequencing method. Because NGS is a well-established method for high-fidelity sequencing and a complete analysis and validation (e.g., a formal sensitivity, specificity, positive predictive value, and negative predictive value analysis, as suggested by this Reviewer) will require significantly more effort in terms of data collection and analysis, we believe it is far beyond the scope of this paper.

6. I am concerned that the DSPP variants the authors identified are platform-specific errors. .... I think that these variants in DSPP should be examined more closely. .... Do they segregate with deafness in families? DSPP is considered a dominant deafness gene. If these were sporadic cases do the authors assume that all of these DSPP mutations arose de novo? Please address this surprising finding.

---Answer: The findings for DSPP variants are unlikely to be platform-specific errors, because we didn’t find a similar high percentage in other batches of samples we processed (unpublished data). Those samples included more than 1000 patients diagnosed with sensorineural hearing loss and about 800 normal hearing control samples. Since samples in this study are sporadic samples, we can’t answer whether they segregated with deafness in families or whether these DSPP mutations arose de novo.

To address this question, we have added (starting from page 13 line 297) in the Discussion: “The high incidence of DSPP variants we found in category III (Table 1) is unlikely to be caused by platform-specific errors, because we didn’t find similar results in other batches of samples we processed (unpublished data). Those samples included more than 1000 patients diagnosed with sensorineural hearing loss and about 800 normal-hearing control samples. ...."

7. The method for targeted capture is not described adequately on line 123. I understand that the authors have previously published it, but it would be helpful for clarity to provide a brief summary of the method. Also, on Line 146, the authors state “we normally hybridize 20 samples together.” Is this how many were done in this study? This is not formal scientific English and the sentence should be revised and clarified in order for the method to be reproducible.

---Answer: A brief summary of the method for targeted capture is given and other details mentioned by this Reviewer are also given (starting from page 7 line 168): “....In the hybridization step for capturing the targeted deafness genes, we normally hybridized 20-30 samples together. Samples were differentiated by different barcodes. In-house generated cDNA capture probes with lengths ranging from 100 to ~5000 bps were used. Biotinylated probes were hybridized overnight with fragmented genomic DNA (gDNA) at 47°C in a thermocycler (Bio-Rad T100, Hercules, CA). Captured gDNA fragments were enriched using streptavidin dynabeads (Beckman Coulter, Brea, CA). ....”

MINOR ESSENTIAL REVISIONS

1. Please have the manuscript proofread by a native English speaker. Here are some of the grammatical/spelling errors I have identified:
   a. Title: should read something like “Combined examination of sequence and copy number variations in human deafness genes improves genetic diagnosis for genetic deafness”
   b. Line 54: “supplement” should be “supplemental”
   c. Line 68: “mega bps” should be “mega base pairs (MB)”
   d. line 73 and line 75: “in causing” should be “to” and “that cause” respectively

---Answers: The revised manuscript was proofread by a native English speaker, who is acknowledged in the Acknowledgment. The title was changed to “Combined examination of sequence and copy number variations in human deafness genes improves diagnosis for cases of genetic deafness”. Other changes (b-d) were made as suggested.
2. Please provide a reference for line 71
   ---Answer: a reference [6] is provided in the revised manuscript.

3. Line 119: InPE1.0 primer is used two times?
   ---Answer: this was changed (page5 line118) to: “……The ligated product (20 ng) was amplified for 14 PCR cycles with Illumina PCR primers InPE1.0 and indexing primer following the manufacturer’s instructions. The PCR products were purified again with QIAquick MinElute column and eluted into 50 µl of hybridization buffer (HB, Roche NimbleGen, Madison, WI). ……”

4. Please describe briefly your targeted capture method on line 123.
   ---Answer: this was answered for a previous question. The brief method description is provided (starting from page7 line168): “……In the hybridization step for capturing the targeted deafness genes, we normally hybridized 20-30 samples together. Samples were differentiated by different barcodes. In-house generated cDNA capture probes with lengths ranging from 100 to ~5000 bps were used. Biotinylated probes were hybridized overnight with fragmented genomic DNA (gDNA) at 47°C in a thermocycler (Bio-Rad T100, Hercules, CA). Captured gDNA fragments were enriched using streptavidin dynabeads (Beckman Coulter, Brea, CA). For targeted NGS projects, on-target coverage after the capturing step is a more appropriate index for describing the quality of sequencing, not the number of overall reads. The average depth of coverage for each exon of targeted deafness genes was used in the calculation of CNVs. ……”

5. Line 135: “sequencing quality score,” do you mean Phred-like quality score?
   ---Answer: yes, the sentence is changed to (page7 line155): “After controlling for data quality (coverage≥20 and Phred-like quality score ≥30), we obtained the VCF reports for the coding regions and the exon-intron boundaries of the targeted deafness genes (supplemental Table 1). ……”

6. Line 144: There is a formatting issue – is the square supposed to be a “<”?
   ---Answer: yes, this was caused by online conversation form Word to PDF files. In the Word version the symbol shows as “<”. The correct symbol should be used in the final production of the manuscript.

7. Line 162: “putatively” should be “putative”
   ---Answer: change was made as suggested.

8. Line 215: “compare” should be “compared”
   ---Answer: change was made as suggested.

9. Please revise the sentence on line 243-246; due to poor grammar it is confusing.
   ---Answer: we modified the writing (page11 line252) to: “Among the 79 patients we studied, the total number of deafness genes affected by CNVs is twenty seven. Among these 27, only five have corresponding pseudogenes. We detected the most CNVs on the DIAPH1 gene, which is located on chromosome 5. ……”

10. Line 249: “structure” should be “structural”
    ---Answer: change was made as suggested.

11. Line 253: “about how” is not formal scientific English, please revise
    ---Answer: We changed the sentence to (page12 line290): “We performed a combined analysis of DNA sequence mutations and CNVs to assess how protein functions known to be essential for hearing may be disrupted. ……”
12. Line 294: “investigated” should be “investigating”  
---Answer: change was made as suggested.

13. Please move the description of Figure 1 from the text to the Figure legend.  
---Answer: Description of the results of Figure 1 was summarized in the legend for the Figure 1 now.

14. How did the authors ensure that the qPCR primers used for STRC were specific? This gene is notoriously difficult to amplify without amplifying the pseudogene. This may be the reason for the false positive result.  
---Answer: This issue was discussed in the Discussion section. Our conclusion was that CNV findings for the STRC gene need to be cautiously interpreted. We have modified the discussion (page14 line334) to: 
“Among the 27 genes that showed CNVs (Table 3), we found that five of them have corresponding pseudogenes (MTAP, MYO15A, OTOA, SOX2 and STRC), according to database www.pseudogene.org. The CNV data of a small group of genes (5 out of 80), especially those on CNV gains including qPCR validation results, need to be cautiously interpreted. ……”.

Second Reviewer

Level of interest: An article of outstanding merit and interest in its field  
Quality of written English: Acceptable  
Statistical review: Yes, but I do not feel adequately qualified to assess the Statistics

Specific comments:

Minor Essential Revision:  
1. In page 3, line 83, “otitis media were ruled out”. Perhaps, “excluded” may be better choice of word to describe patients with otitis media were not included in the study.  
---Answer: the change was made as suggested (page3, line83), “were ruled out” was changed to “were excluded”.

Minor Discretionary Revisions:  
2. In page 6, line 144, a symbol indicating greater than or less than is missing in front of 0.005. This is most likely due to the web-based manuscript retrieval.  
---Answer: in the word file of the manuscript the sentence is “and novel mutations with low population allele frequencies (˂0.005) of unknown significance.”. I agree with the Reviewer that this is likely to be an error caused by web-based manuscript conversion from Word file to PDF file, which should not appear in the final version.

3. In page 7, line 171, “Ct” could be plainly explained or defined for unsophisticated readers in genetics or statistics.  
---Answer: page8 line201: “Calculating Ct differences between the target gene ……” was changed to “Calculating threshold cycle (Ct) differences between the target gene ……”.

Third Reviewer

Reviewer: Sivakumaran A Theru Arumugam  
Reviewer’s report: Please see the attached word document  
Level of interest: An article of importance in its field  
Quality of written English: Acceptable  
Statistical review: No, the manuscript does not need to be seen by a statistician.
Specific comments:

(1) Authors did not provide any information on how many reads were generated per sample. Authors may want to provide the quality metrics summary table.

---Answer: Unlike whole genome NGS projects in which no targeted capture step is involved in preparing NGS libraries, overall reads for each sample is generally not a sensitive quality index for disease panel NGS projects. For any targeted NGS project, on-target coverage after capturing step is more appropriate for describing the quality of sequencing. Poor enrichment leaves too many off-target gDNA fragments, which will still be sequenced by Illumina HiSeq later. However, these off-target reads don’t help the purpose of disease-panel NGS at all.

In this project we have used average on-target coverage for deafness genes analyzed in this study (given in Supplemental Table 1). To better clarify these points, we changed the relevant Method description (starting from page7 line172) to: “Captured gDNA fragments were enriched using streptavidin dynabeads (Beckman Coulter, Brea, CA). For targeted NGS projects, on-target coverage after the capturing step is a more appropriate index for describing the quality of sequencing, not the number of overall reads. ……”. In addition to average coverage, we also used depth of coverage for individual bp, Phred scores in NGS data analysis (starting from page7 line155): “After controlling for data quality (coverage≥20 and Phred-like quality score ≥30), we obtained the VCF reports for the coding regions and the exon-intron boundaries of the targeted deafness genes (supplemental Table 1). ……”.

(2) It is also not clear what percentage of coding exons were covered by certain number of reads? Are there any exons that were not covered by NGS? A summary table/supplementary table would be better.

---Answer: We have provided a supplemental Table7 to show an example of the number of coverage for targeted exons of deafness genes. For this particular sample there was a small percentage (57/1248=4.6%) of exons covered by x1 or less. Percentage of exons with zero coverage was 1.2% (15/1248). Percentage of exons with average coverage at and below 10 & 20 was 6.8% (82/1248) and 10.0% (125/1248), respectively.

We have analyzed 10 samples and added following description (page6 line152): “All samples processed in this project had an on-target average coverage of greater than 100 (Supplemental Fig. 1). Distribution of average coverage (n=10) at 0x, 1x, 10x & 20x was 1.1±0.5, 4.8±1.1, 5.6±1.9 and 11.2±2.9, respectively. ……”.

(3) In the section titled “combined sequence variation and CNV analysis for the cause of genetic deafness” (page 10), the authors mentioned that they could detect pathogenic mutations that may lead to deafness in 29 of the 79 samples. But it is unclear to me from the manuscript in how many patients the genetic cause of deafness was detected.

---Answer: We have provided a new Supplemental Table 5 summarizing results obtained for all 79 patients. In these patients, we detected 22 pathogenic point mutations (categories I, III and V) in 79 patients. Four patients carried CNVs deletions (without point mutation) that could directly lead to deafness. One patient (ot3285) had a combined heterozygous point mutation and copy number reduction in one of the two chromosomes, possibly may also result in deafness. Therefore, the combined detect rate is 34.2% (27/79) in the 79 sporadic patients clinically diagnosed with sensorineural hearing loss. If the criteria are relaxed to include category VII mutations, the overall detection rate is increased to 43.0% (34/79).

We have modified following description (page11 line270) to better clarify these points: “One patient (ot3285) had a heterozygous deafness mutation and a copy number reduction in the same gene (GJB3, supplemental Table 5). Since GJB3 has only one coding exon, this combination may potentially result in compounded effects which cause disruption of GJB3 protein functions, leading to deafness. The other 11 carriers had CNVs affecting deafness genes different from those affected by DNA sequence mutations, and the significance of such combinations is currently unknown. Overall, combined sequence
variation and CNV analysis increased the detection rate from 36.7% (29/79) to 43.0% (34/79) among 79 sporadic patients clinically diagnosed with sensorineural hearing loss.

(4) It is clear from Table 1 that the category II mutations found in DSPP may cause autosomal dominant deafness, but is unclear whether category I mutations found in 15/16 patients (mutations in GJB2 and SLC26A4 cause autosomal recessive deafness) are homozygous or a single heterozygous mutations.

---Answer: By our definition the mutations belonging to the category I are homozygous for recessive, or heterozygous for dominant mutations. Patients bearing type II mutations are carriers of type I mutations. Table 1 didn’t list any category II mutations.

(5) It is unclear if the combination of both sequence variants and CNV increased the detection rate of genetic causes.

---Answer: page12 line275, we have clearly stated that “……Overall, combined sequence variation and CNV analysis increased the detection rate from 36.7% (29/79) to 43.0% (34/79) among 79 sporadic patients clinically diagnosed with sensorineural hearing loss.”.

Line 234 states that 12, who are carrier for deafness mutation also bear CNVs, which were not presented in detail in any Tables.

---Answer: We have added a new Supplemental Table 6 to specifically answer this question. Information for all 12 patients who are carriers for deafness mutation and also bear CNVs is given in this new Table. Corresponding description was changed (starting from page11 line268): “We also found 12 patients to be carriers of deafness mutations (belonging to the categories II or IV as defined in supplemental Table 2), who also bear CNVs (Supplemental Table 6). ……”.

Line 236 states that patient 3742 had heterozygous mutation and a heterozygous CNV in GJB3, which are not found in any Tables or Supplementary Tables. Moreover, I don’t see any CNV in GJB3 in Table 3, which list all CNVs found in this study.

---Answer: The correct patient ID should be ot3285, and we apologize for the mistake we made. Corrections were made in the Table 3 and results were presented in the new Supplemental Table 5. Description was changed (page11 line270): “One patient (ot3285) had a heterozygous deafness mutation and a copy number reduction in the same gene (GJB3, supplemental Table 5). Since GJB3 has only one coding exon, this combination may potentially result in compounded effects which cause disruption of GJB3 protein functions, leading to deafness. ……”.

Authors may want to provide the all CNVs in the supplementary Table

---Answer: All CNVs found in this study are given in the new Supplemental Table 5 now, along with sequence mutation(s) and clinical information of patients.

Line 100 on page 4 – I believe the whole genomic DNA was used to amplify GJB2 and fragmented genomic DNA was used for NGS. If this correct, a couple of sentences “Highmolecular …….200 cycles per burst” should be moved to “NGS protocol and data analysis” section below.

---Answer: Changes were made as suggested. The sentence “High-molecular weight gDNA (5 µg) was fragmented ultrasonically with the Covaris E210 DNA shearing instrument (Covaris Inc., Woburn, MA) to an average size of 300 bps for subsequent construction of Illumina NGS libraries. The Covaris protocol is set at 3 minutes total duration, duty cycle 10%, intensity 5, and 200 cycles per burst.” was moved to the beginning of “NGS protocol and data analysis” section (page4 line110).
Line 124 on page 5 – reads as “FASTQ data files generated after sequencing with the Illumina HiSeq2000 were processed using four independent bioinformatics data processing pilelines”. I wonder if the authors found a significant difference between these four pipelines? What was the consensus rate?  
---Answer: We found that the differences among these four pipelines were insignificant, and consensus rates were high among different bioinformatic pipelines. We added following description (staring from page6 line147 ) to clarify this issue: “The variation and CNV results presented here are consensus results of the four bioinformatic pipelines. The consensus rates among different platforms were high, with the lowest being 94.5% between the GATK and Galaxy pipelines. ……”.

Line 135 on page 6 – “average coverage ≥ 20” – does this mean a threshold of 20 x coverage was set at target bases for variant calling?  
---Answer: Yes, we filtered out the variant calls if the depth of coverage for that particular base pair is below 20. To make this point out more clearly, we changed (starting from page7 line155) the sentence to: “After controlling for data quality (coverage≥20 and Phred-like quality score ≥30), we obtained the VCF reports for the coding regions and the exon-intron boundaries of the targeted deafness genes (supplemental Table 1). ……”.

Line 158 on page 7 – Bioinformatic analysis...... had an average resolution of 4491 bps. Does this mean your algorithm is less sensitive in detecting CNVs less than 4491 bps?  
---Answer: No, sensitive of our algorithm in detecting CNVs affecting small or larger exons should be the same. The resolution of our algorithm in CNV detections is the same as the size of each exon, and the sensitivity is the same for smaller and bigger exons. We changed this sentence (starting from page8 line189) from “Bioinformatic analysis of the gene structure showed that our method of detecting CNVs at single exon resolution of the 80 deafness genes (supplemental Table1) had an average resolution of 4491 bps. ……” to “The resolution of our method used to detect CNVs is the size of a single exon as determined by comparing the average depth of coverage for each exon. Bioinformatic analysis of the gene structure showed that the average exon size for the 80 deafness genes (supplemental Table1) is 4491 bps. ……”.

Line 182 on page 8 – as defined in the supplementary table 1 should be Supplementary Table 2.  
---Answer: this was corrected as suggested.

Line 190 on page 8 – protein fucntions should be protein functions.  
---Answer: this was corrected as suggested.