Identification of the genetic determinants of hearing loss by means of genetic isolates

MED/03 Genetica Medica

Ph.D. student Dragana Vuckovic
Ph.D. program Director Prof.ssa Giuliana Decorti
Thesis Supervisor Prof. Paolo Gasparini

ANNO ACCADEMICO 2013-2014
Abstract

The auditory system is one of the most complex sensory organs, involving hair cells, cochlear neurons and many other components, strongly regulated by the underneath molecular bases. The genetic analysis of complex traits/diseases such as normal hearing function (NHF) and Age-Related Hearing Loss (ARHL) could shed light on these molecular mechanisms. Until now, only few genes are known to contribute to the variability of NHF. Understanding NHF from the genetic point of view can be very interesting and also provide important clues to solving ARHL genetic causes. This issue is of fundamental importance for our ageing society. In fact, Age-Related Hearing Loss is the most prevalent sensory impairment in the elderly affecting 30% of people aged over 60. The disease is not directly life threatening but it contributes to loss of autonomy and is associated with anxiety, depression, and cognitive decline largely compromising the quality of life.

Moreover, interactions with lifestyle and environmental determinants should be taken into account. In order to overcome these confounding factors, we focused on isolated populations, where people share the same environment and similar occupations/habits.

The main aim of the thesis is the understanding of the molecular bases of variation of normal hearing function using: a) Genome Wide Association Studies (GWAS) to identify new genes/loci; b) Statistical association to replicate new candidate findings; c) expression studies to evaluate their role in the mouse cochlea.

Firstly we replicated 9 previously published genes (CSMD1, ANK2, CDH13, DCLK1, ARSG, EVI5, GRM8, RIMBP2, SLC16A6) by running a candidate association analysis in the Silk Road cohort (SR) including several isolated communities in Caucasus and Central Asia. Afterwards a large collaboration with the TwinsUK cohort led to a GWAS meta-analysis, which identified a significant SNP in SIK3 gene. Finally, with the availability of 1000Genomes Project imputation data another GWAS meta-
analysis allowed us to identify two more loci close to PCDH20 and SLC28A3 genes, which we were able to replicate in two independent cohorts from UK (B58C) and Finland (FITSA). All the identified genes were expressed in mouse cochlea.

As regards new contribution in GWAS methodology, we developed an R package software called *MultiMeta*. It is a novel statistically efficient method to perform meta-analysis in a multivariate setting with a flexible approach and tools for easily visualizing results. The method can be applied to any multivariate set of results and will be very useful for analysing hearing data and hopefully leading to novel association discovery, by taking into account the underlying correlations between phenotypes.
Abstract (italian)

Il sistema uditivo è uno degli organi sensoriali più complessi e coinvolge cellule ciliate, neuroni cocleari, e molti altri elementi regolati da basi molecolari. L'analisi genetica dei tratti complessi/malattie come la funzione uditiva e la presbiacusia potrebbe far luce su questi meccanismi molecolari. Fino ad ora solo pochi geni sono noti per contribuire alla variabilità della funzione uditiva. Capirla dal punto di vista genetico può essere molto interessante e può anche fornire indizi importanti per individuare le cause genetiche della presbiacusia. La questione è di fondamentale importanza per la nostra società che invecchia sempre più. La presbiacusia è la perdita sensoriale più diffusa negli anziani e colpisce il 30 % delle persone di età superiore ai 60 anni. La malattia non dà direttamente pericolo di vita, ma contribuisce alla perdita di autonomia ed è associata all'ansia, alla depressione e alla perdita di funzioni cognitive che compromettono seriamente la qualità della vita. Inoltre, le interazioni con fattori ambientali dovrebbero essere prese in considerazione. Per tenere conto di questi fattori, ci siamo concentrati su popolazioni isolate, dove le persone condividono lo stesso ambiente e occupazioni/abitudini analoghe. L'obiettivo principale della tesi è la comprensione delle basi molecolari della variazione della funzione uditiva utilizzando: a) Genome-Wide Association Studies (GWAS) per identificare nuovi geni/loci; b) l'associazione statistica per replicare i candidati identificati; c) studi di espressione per valutare il ruolo dei candidati nella coclea di topo.

In primo luogo abbiamo replicato 9 geni precedentemente pubblicati (CSMD1, ANK2, CDH13, DCLK1, ARSG, EVI5, GRM8, RIMBP2, SLC16A6) con un'analisi di associazione per geni candidati nella coorte della Silk Road (SR) che comprende diverse comunità isolate nel Caucaso e in Asia centrale. In seguito, una grande collaborazione con la coorte britannica TwinsUK ha portato ad una nuova meta-analisi di GWAS, in cui abbiamo identificato uno SNP significativo nel gene SIK3. Infine, con la disponibilità dei dati del progetto 1000Genomes per l'imputazione,
un’ulteriore meta-analisi ci ha permesso di individuare altri due loci vicino ai geni PCDH20 e SLC28A3, che siamo stati in grado di replicare in due coorti indipendenti dal Regno Unito (B58C) e dalla Finlandia (FITSA). Tutti i geni identificati sono risultati espressi nella coclea.
Per quanto riguarda contributi originali alla metodologia GWAS, abbiamo sviluppato un pacchetto R chiamato MultiMeta. Si tratta di un metodo statisticamente efficace per eseguire la meta-analisi in un contesto multivariato, con un approccio flessibile e strumenti per gestire facilmente la visualizzazione dei risultati. Il metodo può essere applicato a qualsiasi insieme di risultati multivariati e sarà molto utile per analizzare i dati uditivi e possibilmente per individuare nuove associazioni, sfruttando la correlazione tra fenotipi.
# TABLE OF CONTENTS

**ABSTRACT**  
2

**ABSTRACT (ITALIAN)**  
4

**GENERAL INTRODUCTION**  
8

1.1 Hearing function and hearing loss  
9

1.2 Genetics of hearing loss  
11

1.3 Isolated populations  
14

1.4 Genome-wide association studies (GWAS)  
15

1.5 Aim of the thesis and contents  
16

**EXPRESSION AND REPLICATION STUDIES TO IDENTIFY NEW CANDIDATE GENES INVOLVED IN NORMAL HEARING FUNCTION**  
18

2.1 Background  
19

2.2 Materials and Methods  
19

2.2.1 Ethics statement  
19

2.2.2 Sample collection, preparation and genotyping of the replication cohort  
19

2.2.3 Imputation  
20

2.2.4 Audiometric evaluation  
20

2.2.5 Selection of candidate genes  
21

2.2.6 Expression studies  
22

2.2.7 Statistical analysis  
22

2.2.8 Genotype-phenotype correlations  
23

2.3 Results  
24

2.4 Discussion and conclusions  
30

**SALT-INDUCIBLE KINASE 3, SIK3, IS A NEW GENE ASSOCIATED WITH HEARING**  
34

3.1 Background  
35

3.2 Materials and Methods  
36

3.2.1 Subjects  
36

3.2.2 Phenotypes  
36

3.2.3 Genotyping  
37

3.2.4 Imputation  
37

3.2.5 Meta analysis and statistical analysis  
37

3.2.6 Immunohistochemistry for SIK3 and Confocal Microscopy  
38

3.3 Results  
38

3.3.1 Subjects  
38

3.3.2 Meta-analysis  
39

3.3.3 Expression analysis of SIK3  
46

3.4 Discussion and conclusions  
46

**GENOME-WIDE ASSOCIATION ANALYSIS ON NORMAL HEARING FUNCTION IDENTIFIES PCDH20 AND SLC28A3 AS CANDIDATES FOR HEARING FUNCTION AND LOSS**  
51

4.1 Background  
52

4.2 Materials and Methods  
52

4.2.1 Sample collection, preparation and genotyping  
52

4.2.2 Imputation and filtering  
54

4.2.3 Audiometric evaluation  
55

4.2.4 Association analysis  
55
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.5</td>
<td>RNA EXTRACTION, REVERSE TRANSCRIPTASE-POLYMERASE CHAIN REACTION (RT-PCR) AND REAL-TIME QUANTITATIVE RT-PCR</td>
<td>56</td>
</tr>
<tr>
<td>4.2.6</td>
<td>RNA-Seq expression profiling</td>
<td>56</td>
</tr>
<tr>
<td>4.2.7</td>
<td>Sequencing</td>
<td>56</td>
</tr>
<tr>
<td>4.3</td>
<td>RESULTS</td>
<td>57</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Statistical association</td>
<td>57</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Sequencing analysis</td>
<td>62</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Gene expression analysis</td>
<td>62</td>
</tr>
<tr>
<td>4.4</td>
<td>DISCUSSION AND CONCLUSIONS</td>
<td>64</td>
</tr>
</tbody>
</table>

**MULTIVARIATE GWAS**  
5.1 BACKGROUND  
5.2 OVERVIEW OF MULTIVARIATE GWAS METHODOLOGY  
5.3 STATISTICAL POWER AND META-ANALYSIS  
5.4 METHODS  
5.5 RESULTS  
5.6 CONCLUSIONS  

**APPENDIX - MAPS OF GENETIC ISOLATES**  
77

**REFERENCES**  
78
CHAPTER 1

General Introduction
1.1 Hearing function and hearing loss

The hearing system is characterized by three structures: a) the outer ear, b) the middle ear, and c) the inner ear, that all play a role in hearing function. Sound waves pass through the outer ear and cause vibrations at the eardrum. The eardrum and three small bones of the middle ear amplify the vibrations as they travel to the inner ear. There, the vibrations pass through fluid in a snail-shaped structure in the inner ear (cochlea). In particular, the organ of Corti, which is located in the cochlea, performs a fundamental step in hearing function, consisting of translation of sound from mechanical waves to neural impulses. This is achieved thanks to the presence of hair cells. They deflect by the vibrations of sound and this activity opens ion channels modulating potential within the cell and releasing neurotransmitters to synaptic junctions between hair cells and neural fibres of the auditory nerve (Steel et al. 2001). The neural spike subsequently propagates in the auditory nerve fibre, it is perceived by the brain, mainly in the temporal lobe where it can be processed and interpreted. Thus the correct functioning of the hair cells and the proper regulation of potential in the cochlear liquid are particularly important for hearing function.

Figure 1. The anatomy of the ear
Hearing loss can occur due to several reasons regarding each part of the ear, such as ruptured eardrum, earwax obstruction or infections. However, permanent and severe hearing loss is often due to inner ear malfunctioning. This in turn can be caused by ageing, noise exposure and/or genetic predisposition. Hearing function can be evaluated through medical examination of the ear and a variety of tests, the most common being a pure-tone hearing measurement, recorded on an audiogram.

Figure 3. Audiogram: the figure shows three example audiometric curves. The green area is the normality range for hearing function.
The hearing system is difficult to study through biochemical routes, due to the small amounts of tissue available for analysis and to key molecules that may be present in only a few tens of copies per cell (Steel et al. 2001). A genetic approach can be more useful and precise for uncovering molecular mechanisms underlying hearing function and hearing loss.

1.2 Genetics of hearing loss

Given the complexity of the hearing mechanism, it should come as no surprise that many genes are involved in hearing. To date more than 140 loci and 80 genes have been mapped as involved in hearing loss. Proteins coded by these genes belong to different functional categories: 1) hair bundle morphogenesis, 2) ion homeostasis, 3) extracellular matrix composition and 4) transcription factors (Dror & Avraham 2010). In particular, ion channels and transporters are fundamental for allowing the flow of ions and preserving the endolymph space in the inner ear (see Figure 4) (Lang et al. 2007), while several adhesion proteins, motor proteins and others are involved in stereocilia (hair-like apices of hair cells, Figure 5) (Frolenkov et al. 2004).

**Figure 4. Ion channels and transporters in the inner ear (A), in the inner hair cells (B) and in the outer hair cells (C)**
Figure 5. Stereocilia and proteins involved
Thus dysfunctional proteins are involved in impaired molecular-physiologic processes of potassium and calcium homeostasis, apoptosis signalling, stereocilia linkage, mechano-electric transduction, electromotility, and others (Van Laer et al. 2003). In particular, the study of mutant mice revealed that often hearing and balance problems are due to deficient hair cells differentiation (Quint et al. 2003) and about 230 genes have been described to cause mouse inner ear malformations or dysfunction (http://hearingimpairment.jax.org/index.html).

Despite these interesting results the gap between the number of loci and the number of genes identified for hearing loss indicates that many more genes are yet to be discovered. Moreover the molecular basis of variation of normal hearing function and complex traits, such as Age-Related Hearing Loss (ARHL) or Noise-Induced Hearing Loss (NIHL), is still largely unknown. These
traits strongly depend on genetic bases, but also on environmental variables, which is why they are more difficult to study. In this work we try to overcome many of the environment-related issues by focusing on isolated populations.

1.3 Isolated populations

Isolated populations are subpopulations derived from a small number of individuals (Hatzikotoulas et al. 2014). Often they are either geographically isolated, for example on islands, such as the Sardinian, Icelandic and Orkney populations, or linguistically isolated, such as the Saami (Huyghe et al. 2011, Francalacci et al. 2003) or the inhabitants from Resia, a village located in the Friuli Venezia Giulia (FVG) genetic park (Esko et al. 2012). Moreover, a population isolate can exist without physical or linguistic barriers, as is the case of Kuusamo and Southern Ostrobothnia communities (Jakkula et al., 2008, Hovatta et al., 1997). The consequences of isolation can be observed at a genetic level: a) Linkage Disequilibrium (LD) is extended to longer distances, thus reducing haplotypes' complexity; b) homozygosity also increases due to inbreeding and c) rare variants (<1% minor allele frequency in the general population) can be enriched giving higher prevalence for certain diseases, such as multiple sclerosis in Sardinia, Southern Ostrobothnia and Orkney islands (Pugliatti et al. 2005; Tienari et al., 2004; Rothwell et al. 1998).

As regards environmental factors, the inhabitants of these populations/communities share the same geographical setting (often a restricted area), the same nutritional habits, a small range of occupations, and the same cultural patterns. Given all these genetic and lifestyle characteristics, isolated populations are ideal samples for studying the genetics of complex diseases, such as in our case hearing function and ARHL. They provide a unique genetic structure, while allowing controlling for environmental factors. For example, it is possible to detect the effect of rare variants even in relatively small isolates thanks to their enrichment since these variants raise the risk in carriers at a higher rate and have larger effect sizes than common variants (Wray et al. 2001).
1.4 Genome-wide Association studies (GWAS)

Association analyses have become increasingly popular for mapping genes involved in complex traits and diseases (Wang et al. 2012). A particular marker in the genome is tested for statistical association with a trait of interest in a large cohort. For example, if the sample is a case-control sample for a disease, the allele frequencies are compared between cases and controls for any significant differences. In the case of GWAS, the same test is repeated for each marker in the genome and, if statistical power is sufficient, gives us the information of all the loci in the genome involved in the analysed trait.

Thanks to the development of high-throughput sequencing and genotyping technologies, a large number of genetic markers (usually Single Nucleotide Polymorphisms - SNPs) are available even for large cohorts of thousands of subjects. The main aim of GWAS design is to capture all common genetic variation across the genome and to relate this variation to disease risk. GWAS requires three essential items: 1) a large number of samples accurately phenotyped from populations that effectively provide genetic information regarding the research question, 2) polymorphic alleles that can be inexpensively and efficiently genotyped and cover the whole genome adequately, and 3) analytic methods that are statistically powerful and can be employed to identify the genetic associations in an unbiased fashion.

In particular, as regards the first item, a large use of Meta-analyses studies, which combine results across different cohorts, has been applied. Therefore, by combining association studies the sample size grows, which in turn increases the statistical power to detect association signals. Several Meta-analysis methods exist, that differ for example in the way they weigh the contribution of each population (Nelis et al. 2009; Consortium WTCC, 2007; Devlin et al. 1999; Pritchard et al. 2000).

Regarding the second item many different arrays are available from companies such as Illumina and Affymetrix. Genome-wide arrays cover up to 1 million SNPs and custom arrays exist if a specific disease is analysed, such as the Illumina Cardio-Metabo Chip. The SNPs selected for genotyping arrays are often tag-SNPs summarizing information for larger regions surrounding
them, thanks to LD. Thus imputation techniques can easily and precisely increase the amount of information up to tens millions of markers (Howie et al. 2012). However, once a significant association is detected it may be a challenging process to find out the real causative variant. Performing association on sequencing data easily overcomes this issue. Although sequencing cost is still much higher than genotyping, large projects of sequencing samples already exist and will possibly become more and more common in the future (UK10K Project).

Concerning the third item different approaches are possible based on the trait and the population characteristics. The most reliable method performs a Mixed Model regression (both linear LMM, and generalized mixed model GLMM) in order to account for population stratification and relatedness (Bolker et al. 2009). This model accounts for both fixed and random effects. Fixed effects include the candidate SNP and optional covariates such as gender or age, while random effects usually include a genomic relatedness matrix called kinship matrix (Price et al. 2010).

Finally due to the large number of statistical testing performed in a GWAS, it is important to apply a correction to the p-value limit for significance, while keeping the false discovery rate at 5%. This limit has been set to 5*10^{-8}, assuming ~1,000,000 independent association tests and applying Bonferroni correction (Risch et al. 1996).

1.5 Aim of the thesis and Contents

The aim of this work has been a comprehensive study of the genetics of hearing function and hearing loss. The understanding of the vast heterogeneity characterizing these traits has been a fundamental step towards the identification of new genetic variants playing a role in the auditory system. Thanks to the availability of a large collection of individuals from isolated populations, here we describe population-based studies, mainly relying on GWAS. A follow-up study statistically replicating previous results and validating genes at the expression level was performed (Chapter 2) together with two different GWAS identifying new candidate genes involved in
hearing function (SIK3 in Chapter 3) and (PCDH20 and SLC28A3 in Chapter 4). In order to increase statistical power and seek for other interesting associations in our dataset, we decided to focus our attention on multivariate GWAS, which would allow us to analyse all the hearing traits simultaneously. In particular, as described in Chapter 5, our efforts concerned improving existing study designs by applying and developing new methodology. We studied the generalization of standard GWAS to multiple phenotypes (multivariate mixed model regression), which has been recently implemented in GEMMA software (Zhou & Stephens 2014). We then developed an R package to perform meta-analysis of these multivariate results, which is a necessary step when dealing with more than one cohort. To our knowledge this analysis was not previously available.

Present research has many possible future developments. A further exploration of statistical associations will be possible by means of the multivariate methodology developed, which exploits the strong correlation between phenotypes. Furthermore with the availability of Next Generation Sequencing data, it will be interesting to explore rare functional or regulatory elements contributing to NHF and ARHL.
CHAPTER 2

Expression and Replication Studies to Identify New Candidate Genes Involved in Normal Hearing Function
2.1 Background
Results from a GWAS on normal hearing function in humans were first published by our group in 2011 (Girotto et al. 2011), giving an insight into several candidate genes that may be involved in the trait. Audiometric measurements were tested as quantitative traits and hundreds of SNPs had suggestive p-values (~10^-5 or less). In order to further investigate these SNPs we designed a follow-up study. The approach was based on four sequential steps: a) selection of a short list of candidate genes from the previous study, b) evaluation of their expression patterns in the mouse inner ear, c) genetic association replication of those genes clearly expressed in the cochlea in a new independent cohort from the Silk Road, d) examination of genotype/phenotype relationships for each of the replicated genes or SNPs.

2.2 Materials and Methods

2.2.1 Ethics statement
Human. Consent forms for clinical and genetic studies were signed by each participant and all research was conducted according to the ethical standards defined by the Helsinki declaration. The study was approved by the Institutional Review Board of IRCCS Burlo Garofolo PROT CE/v-78.
Mice. Mouse studies were carried out in accordance with UK Home Office regulations and the UK Animals (Scientific Procedures) Act of 1986 (ASPA) under a UK Home Office licence, and the study was approved by the Wellcome Trust Sanger Institute's Ethical Review Committee. Mice were culled using methods approved under this licence to minimize any possibility of suffering.

2.2.2 Sample collection, preparation and genotyping of the replication cohort
We collected samples from 493 people aged >18 years old from several rural communities located along the Silk Road during the Marco Polo Scientific Expedition (Girotto et al. 2011 B) (see Appendix). Saliva samples were
collected using the Oragene kit (DNA Genotek Inc.) and DNA extracted according to supplier’s protocols. DNA samples were genotyped with the Illumina 700K platform. In order to perform a consistent replication study, the collection of genotype and phenotype data followed the same protocols as reported for the previously published work (Girotto et al. 2011). After standard quality control, all genotypes were checked to ensure that they were reported with the coordinates of the 1000 Genome Project (releases build 37) reference data and on the forward strand of the reference genome (1000 Genomes Project Consortium 2010).

2.2.3 Imputation
Genotype imputation was conducted using SHAPEIT2 (Delaneau et al. 2012) for the phasing step and IMPUTE2 (Howie et al. 2012) for the imputation to the 1000 Genomes phase I v3 reference set (1000 Genomes Project Consortium 2010). After imputation SNPs with minor allele frequency (MAF) < 0.01 or imputation quality score (Info) < 0.4 were excluded from the statistical analyses.

2.2.4 Audiometric evaluation
Audiometric tests and a careful clinical examination (i.e. psychological, neurological, cardiological, etc.) were carried out for each individual of the Silk Road cohort. Thresholds for different frequencies (0.25, 0.5, 1, 2, 4, 6, 8 kHz) were measured and the pure-tone averages of air-conduction thresholds (PTA at the lower 0.25, 0.5 and 1 kHz, middle 0.5, 1 and 2 kHz, and high frequencies 4, 8 kHz) were calculated. Furthermore the first three Principal Components (PC1, PC2 and PC3) for all the frequencies combined were computed in order to summarize the audiometric data. Single frequencies, PTAs and PCs were used to run the association studies. Although many different measurements were collected and analysed, they were not independent; in fact PC1 alone accounted for approximately 80% of the total variance so we are not analysing multiple phenotypes. Familial forms of
inherited hearing loss as well as subjects affected by diabetes or other systemic disorders facilitating the development of hearing loss were excluded from the study.

2.2.5 Selection of candidate genes
From our earlier GWAS analysis, we started with a list of 683 SNPs that were associated with the hearing trait with a p-value of 10^-5 or lower. This number of SNPs was reduced to 84 by a) examining a region 250kb upstream and downstream of each SNP and selecting the most significantly-associated SNP, removing the less significant SNPs in the region; b) removing SNPs that marked gene deserts, where there was no annotated gene in the region +/- 250kb; c) removing SNPs that marked only genes with unknown functions such as LOC or FAM designations; and d) removing SNPs that marked genes already known to be involved in diseases unrelated to hearing function. The 84 SNPs were then reduced to a total of 19 genes prioritized by the availability of antibodies for expression studies plus at least one of the following criteria a) the lowest p-values of 10^-7; b) genes belonging to gene families previously associated with hereditary hearing loss; c) expression in the inner ear from genome-wide approaches reported in databases including Eurexpress, MGI and the NCBI EST list; or d) genes with potential functional links to the genes selected by the above criteria occurring in the same in silico pathways. For each SNP the closest gene was selected, but in some cases, more than one gene was included based on linkage to a single SNP (see Table 1).

Table 1. List of candidate genes derived from previously published Meta-Analysis (Girotto et al. 2011) with inclusion criteria satisfied
Legend: Sugg. p-value† = suggestive p-value: p~10^-5,10^-6; H.Sugg. p-value‡ = highly suggestive p-value: p~10^-7 [3].

<table>
<thead>
<tr>
<th>SNP</th>
<th>GENE</th>
<th>SNP-GENE distance OR SNP position in the gene</th>
<th>Similarity with HHL genes</th>
<th>Sugg. p-value†</th>
<th>H.Sugg. p-value‡</th>
<th>Availability of antibodies</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs8077384</td>
<td>AMZ2</td>
<td>~39 kb</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>rs17045059</td>
<td>ANK2</td>
<td>NM_001127493.1:c.-40+35627A&gt;T</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>rs8077384</td>
<td>ARSG</td>
<td>~92 kb</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>rs17195859</td>
<td>CDH13</td>
<td>NM_001220488.1:c.778-59471G&gt;A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>rs989667</td>
<td>CIMP</td>
<td>NM_030629.2:c.18+37681C&gt;T</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10091102</td>
<td>CSMD1</td>
<td>NM_033225.5:c.4985-1219C&gt;G</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9574464</td>
<td>DCLK1</td>
<td>~30 kb</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6673959</td>
<td>DFFB</td>
<td>~321 kb</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12758887</td>
<td>EVI5</td>
<td>~136 kb</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs11159133</td>
<td>FOS</td>
<td>~10 kb</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7182802</td>
<td>GABRG3</td>
<td>NM_001270873.1:c.270+75238C&gt;T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>rs2687481</td>
<td>GRM8</td>
<td>~210 kb</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>rs6673959</td>
<td>KIAA0562</td>
<td>~350 kb</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10815873</td>
<td>PTPRD</td>
<td>NM_002839.3:c.4086+398C&gt;A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>rs10848114</td>
<td>RIMBP2</td>
<td>NM_015347.4:c.-4+11838G&gt;A</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12758887</td>
<td>RPL5</td>
<td>~136 kb</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs8077384</td>
<td>SLC16A6</td>
<td>~100 kb</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>rs12758887</td>
<td>SNORA66</td>
<td>~87 kb</td>
<td>X</td>
<td>in situ probes used</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs12758887</td>
<td>SNORD21</td>
<td>~91 kb</td>
<td>X</td>
<td>in situ probes used</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2.6 Expression studies
Expression studies in mouse cochlea were performed in collaboration with prof. Karen Steel’s lab by Dr. Girotto. For details see (Girotto et al. 2014).

2.2.7 Statistical Analysis
Genes showing robust expression patterns in the cochlea were further analysed for replication. Association analysis of GWAS-type was carried out in the Silk Road cohort using the GRAMMAR-Gamma method as implemented in the GenABEL suit (Aulchenko et al. 2007) for genotyped SNPs and MixABEL (Svichcheva et al. 2012) for imputed data. Data were adjusted to account for sex, age and genomic kinship as previously described (Girotto et al. 2011) and different genetic models were tested (additive, recessive, dominant and overdominant). Moreover, in order to account for
different population-specific effects, that are difficult to replicate at the SNP level, a gene-based test was performed as follows: 1) all the hearing traits were adjusted for covariates and genomic relatedness, using GRAMMAR-Gamma; 2) for each gene only intragenic SNPs were selected; 3) principal components (PC) of the genotypes coded as numerical values (0,1,2) were computed and only those explaining the largest amount of the genetic variance were taken into account (i.e. if N is the total number of intragenic SNPs then PCs explaining more than 1/N variance were chosen); 4) the selected PCs were used as multiple regressors for the corrected traits. This methodology was described and tested by (Wang and Abbott 2010). Significance level considered was nominal significance (5%), which is usually accepted for replication studies, especially when generalizing findings to non-European populations as in our case (Bryant et al. 2013). Furthermore we did not correct for the different genetic models tested because these are not four independent tests (Parsons et al. 2013).

2.2.8 Genotype-phenotype correlations
The audiometric phenotype was compared in a subset of 2904 subjects from Italy, Caucasus and Central Asia (Silk Road cohort) by focusing on the three genotypes (AA, AB and BB) of each replicated top SNP to study any differences among them. Among these SNPs, the homozygotes for the minor allele group ranged from 34 to 642 people. After computing mean values and standard errors for each frequency (0.25, 0.5, 1, 2, 4, 8 kHz) and for each genotype, results were plotted to form three different audiometric curves. For each plot, the Y-axis represents age-adjusted threshold values measured in decibels sensation level (from 50 dB SL to -10 dB SL) and the X-axis represents frequency measured in kHz. The points represent the mean values for each genotype at each frequency and standard errors of the traits are shown. The profiles were then visually inspected.
2.3 Results

In this study, a list of 19 genes were selected from a list with suggestive significance recently reported in a GWAS meta-analysis of different quantitative hearing traits (Girotto et al. 2011), for further investigation of their possible role in hearing function. Those with evident staining within the cochlea (expression step) underwent a replication study in an independent cohort from the Silk Road and, for the replicated ones, genotype-phenotype correlations were carried out (validation step) (Fig.1).

Figure 1. Analysis flow chart
The diagram illustrates the 4 steps defining our strategy. Relevant details for each step are given: GWAS meta-analysis description, expression studies in the mouse cochlea, replication association study in Silk Road cohort and genotype-phenotype relationship.

The list of 19 genes was selected following the criteria described in the methods. Details of each gene are given in Tables 1 (above) and 2, including information about expression reported in several databases and descriptions of any mouse mutant phenotypes available.
### Table 2. Further details of shortlisted candidate genes


**Legend:** SHIELD*=SHIELD - enriched in hair cells compared to supporting cells; SHIELD**=SHIELD - auditory/vestibular ganglia; DFN=within a human deafness locus?; N/A=data not available.

<table>
<thead>
<tr>
<th>Gene</th>
<th>MGI</th>
<th>Morton</th>
<th>SHIELD*</th>
<th>SHIELD**</th>
<th>DFN</th>
<th>Mouse knockout phenotype (MGI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMZ2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
</tr>
<tr>
<td>ANK2</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Homozygous mutation of this gene results in death by postnatal day 8, although some animals survive to P20. Mutant animals display reduced body size, impaired balance and locomotion, brain structure dysmorphologies, abnormal lens, and optic nerve degeneration.</td>
</tr>
<tr>
<td>ARSG</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Mice homozygous for a null mutation display lysosomal storage pathology in the nervous system and peripheral tissues, including the liver and kidneys, resulting in Purkinje cell loss and age dependent cognitive impairment.</td>
</tr>
<tr>
<td>CDH13</td>
<td>No</td>
<td>No</td>
<td>Yes (p&lt;0.089)</td>
<td>No</td>
<td>No</td>
<td>Mice homozygous for a null allele exhibit decreased retinal neovascularization and increased adiponectin levels.</td>
</tr>
<tr>
<td>CMIP</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>OTSC4</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>CSMD1</td>
<td>No</td>
<td>No</td>
<td>Yes (p&lt;0.076)</td>
<td>No</td>
<td>DFNW2 WS2C</td>
<td>No abnormal phenotype reported; Mice exhibit normal pre-pulse inhibition, social interaction, succrose preference and d-amphetamine sensitivity.</td>
</tr>
<tr>
<td>DCLK1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Mice homozygous for a null allele lack the corpus callosum and hippocampal commissure and show aberrant interhemispheric axonal projections. Mice homozygous for a different null allele have normal gross brain architecture but show axonal and dendritic defects following knockdown of Dcx expression.</td>
</tr>
<tr>
<td>DFFB</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Mice homozygous for a knock-out allele are viable, fertile and developmentally normal; however, mutant thymocytes and other cell types fail to undergo apoptotic DNA fragmentation in response to dexamethasone or other apoptotic stimuli.</td>
</tr>
<tr>
<td>EVI5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
</tr>
<tr>
<td>FOS</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Diminished responses to sharp noises. Null mutants are growth-retarded, most dying perinatally. Survivors have osteopetrosis and abnormal tooth eruption, gametogenesis, hemopoiesis, behavior and photoreceptor apoptosis. Hippocampal-specific mutants have seizures and highly excitable neurons.</td>
</tr>
<tr>
<td>GABRG3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
</tr>
<tr>
<td>GRM8</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>DFNB14</td>
<td>DFNB17</td>
<td>Mice homozygous for a knock-out allele are overweight and mildly insulin resistant, and display increased anxiety-related responses and reduced exploration in a new environment. Mice homozygous for a different knock-out allele exhibit altered excitatory responses in the dentate gyrus.</td>
</tr>
<tr>
<td>KIAA0562</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
</tr>
<tr>
<td>PTPRD</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Homozygotes for a targeted null mutation exhibit impaired learning of spatial tasks, enhanced long-term potentiation at hippocampal synapses, and high mortality associated with reduced food intake.</td>
</tr>
<tr>
<td>RIMBP2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>DFNA25 DFNA41</td>
<td>N/A</td>
</tr>
<tr>
<td>RPL5</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
<td>No</td>
<td>N/A</td>
</tr>
<tr>
<td>SLC16A6</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
</tr>
</tbody>
</table>
In order to define their expression pattern in the mouse cochlea, which has not been reported before, immunohistochemistry studies in mice at P4-P5 were performed. Results from these studies demonstrated specific cochlear expression patterns for 12 of them (63%) at this age. All 12 genes gave the same levels of staining from the apex to the base of the cochlea. Furthermore, in 9 cases clear labelling was also evident in the vestibular system (Girotto et al. 2014).

A summary of the distribution of labelling is given in Table 3.

Table 3. Summary of expression results
The Table shows a list of genes with specific labelling patterns in the cochlea.
Legend: ihc, inner hair cells; i/obc, inner and outer border cells; ohc, outer hair cells; rc, root cells; sp, spiral prominence; tm, tectorial membrane; bm, basilar membrane; sv, stria vascularis; pc, pillar cells; Dc, Deiter cells; cC, cells of Claudius; Hc, Hensen cells; Ko, Kölliker’s organ; sg, spiral ganglion; N/A, not available.

<table>
<thead>
<tr>
<th>Gene</th>
<th>ihc</th>
<th>ohc</th>
<th>rc</th>
<th>sp</th>
<th>tm</th>
<th>bm</th>
<th>sv</th>
<th>pc</th>
<th>Dc</th>
<th>cC</th>
<th>Hc</th>
<th>Ko</th>
<th>sg</th>
<th>maculae</th>
<th>cristae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ank2</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Arsg</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cdh13</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Csmd1</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dclk1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evi5</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Gabrg3</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GlyBP</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Grm8</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Ptprd</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rimbp2</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Slc16a6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These 12 genes expressed in specific patterns in the mouse cochlea were then included in a replication step by analysing their genetic association using an independent cohort from the Silk Road, genetically and geographically very
distant from the European population. A total of 9 genes (75%) were replicated; 4 genes were replicated at the SNP level (CSMD1, ANK2, CDH13, DCLK1) while an additional five (ARSG, EVI5, GRM8, RIMBP2, SLC16A6) were replicated at the gene level (see Table 4 and 5).

Table 4. Results of the replication association study for the candidate SNPs of expressed genes

Legend: Eff.All=effect allele; * from published work (Girotto et al. 2011); Repl.=replication; Maf=minor allele frequency; Str=strand; imp=imputed; gen=genotype; Info Score=imputation quality score form IMPUTE2 (Howie et al. 2012); Conc=concordance for the region encompassing the SNP (chunk dimension=5 Mb); HWE=Hardy-Weinberg Equilibrium; add=additive; dom=dominant.

<table>
<thead>
<tr>
<th>Trait</th>
<th>PC1</th>
<th>4 kHz</th>
<th>6 kHz</th>
<th>0.25 kHz</th>
<th>Maf</th>
<th>Repl. p-value</th>
<th>Beta±SE</th>
<th>Repl. Beta±SE</th>
<th>Meta-analysis Beta±SE</th>
<th>Eff. All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Trait</td>
<td>PC1</td>
<td>4 kHz</td>
<td>6 kHz</td>
<td>0.25 kHz</td>
<td>Maf</td>
<td>Repl. p-value</td>
<td>Beta±SE</td>
<td>Repl. Beta±SE</td>
<td>Meta-analysis Beta±SE</td>
<td>Eff. All</td>
</tr>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
</tbody>
</table>

Table 4 continued...

<table>
<thead>
<tr>
<th>Trait</th>
<th>PC1</th>
<th>4 kHz</th>
<th>6 kHz</th>
<th>0.25 kHz</th>
<th>Maf</th>
<th>Repl. p-value</th>
<th>Beta±SE</th>
<th>Repl. Beta±SE</th>
<th>Meta-analysis Beta±SE</th>
<th>Eff. All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Trait</td>
<td>PC1</td>
<td>4 kHz</td>
<td>6 kHz</td>
<td>0.25 kHz</td>
<td>Maf</td>
<td>Repl. p-value</td>
<td>Beta±SE</td>
<td>Repl. Beta±SE</td>
<td>Meta-analysis Beta±SE</td>
<td>Eff. All</td>
</tr>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
</tbody>
</table>

Table 4 continued...

<table>
<thead>
<tr>
<th>Trait</th>
<th>PC1</th>
<th>4 kHz</th>
<th>6 kHz</th>
<th>0.25 kHz</th>
<th>Maf</th>
<th>Repl. p-value</th>
<th>Beta±SE</th>
<th>Repl. Beta±SE</th>
<th>Meta-analysis Beta±SE</th>
<th>Eff. All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Trait</td>
<td>PC1</td>
<td>4 kHz</td>
<td>6 kHz</td>
<td>0.25 kHz</td>
<td>Maf</td>
<td>Repl. p-value</td>
<td>Beta±SE</td>
<td>Repl. Beta±SE</td>
<td>Meta-analysis Beta±SE</td>
<td>Eff. All</td>
</tr>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
</tbody>
</table>

Table 4 continued...

<table>
<thead>
<tr>
<th>Trait</th>
<th>PC1</th>
<th>4 kHz</th>
<th>6 kHz</th>
<th>0.25 kHz</th>
<th>Maf</th>
<th>Repl. p-value</th>
<th>Beta±SE</th>
<th>Repl. Beta±SE</th>
<th>Meta-analysis Beta±SE</th>
<th>Eff. All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Str</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>gen</td>
<td>0.022</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
<tr>
<td>Type</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>imp</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>0.98106</td>
<td>G</td>
</tr>
<tr>
<td>Maf</td>
<td>0.25</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04±0.02</td>
<td>0.04±0.02</td>
<td>0.08±0.03</td>
<td>0.07±0.03</td>
<td>C</td>
</tr>
</tbody>
</table>
In particular, the following SNPs were replicated: rs10091102-C (CSMD1, lowest p-value at PC1=0.022) and rs17045059-G (ANK2, lowest p-value at 4 kHz=0.010) under an additive model; rs17195859-A (CDH13, lowest p-value at 6 kHz =0.024) and rs9574464-G (DCLK1, lowest p-value=0.008 at 0.25 kHz) under a dominant model. The strongest associations from the gene-based test were obtained for GRM8 at 4 kHz (p-value=0.006), and EVI5 (p-value=0.002) at 6 kHz.

Finally the 9 genes with positive replication data were investigated to check for the possible presence of genotype/phenotype relationships by comparing the average audiometric curves. As shown in Fig.2, seven genes (77%) (GRM8, ANK2, CDH13, SLC16A6, ARSG, RIMBP2 and DCLK1) showed a distinct audiometric pattern for each genotype, with differences at some specific frequencies and intensities. GRM8, ANK2 and CDH13, expressed in multiple cells type in the cochlea, showed a downward slope towards high
frequencies for the homozygous BB individuals with a deterioration in all the thresholds compared with AA and AB genotypes (see Fig.2A). Three other genes SLC16A6, ARSG and DCLK1 with a specific expression pattern (in the hair cells in the organ of Corti or marginal cells of the stria vascularis) displayed a slope in which the homozygous BB subjects present an improvement in thresholds (ranging from 0.5 to 8 dB) at all frequencies, in particular for SLC16A6 and ARSG genotypes. A slight improvement at all frequencies for the homozygous genotype AA was present for the RIMBP2 gene (see Fig.2B). As shown in Fig.2C, CSMD1 and EVI5 did not show any specific pattern of genotype/phenotype relationship.

Figure 2. Genotype-phenotype relationship. A, B, C) The figure displays genotype-phenotype relationship for genes with evident differences among the three profiles. The x-axis represents frequencies (kHz), while the y-axis represents sex and age adjusted hearing thresholds (dB SL) with standard deviations, three different colours represent the corresponding genotypes (AA=black, AB=red, BB=green) and the number of subjects for each genotype is reported in brackets. A) GRM8, ANK2 and CDH13 show a deterioration of all the threshold levels for the homozygous BB. B) SLC16A6, ARSG and DCLK1 display an improvement (ranging from 0.5 to 8 dB) at all frequencies for the homozygous BB. Similarly, RIMBP2 presents an improvement for the homozygous genotype AA. C) CSMD1 and EVI5 do not display any particular pattern.
2.4 Discussion and conclusions

Despite recent progress, almost nothing is known about the molecular bases of variation of normal hearing, apart from genes identified as being directly involved in hereditary hearing loss and one gene (GRM7) recently described associated with age-related hearing loss (ARHL) in humans (Friedman et al. 2009; Newman et al. 2012). Here, we report a strategy based on a) selection of genes from a previous GWAS (Girotto et al. 2011), b) evaluation of their expression in the mouse, c) genetic replication of those showing a specific inner ear expression pattern in different and distant populations and, d) genotype-phenotype relationship of the replicated genes. This strategy proved to be an effective filtering and discovery approach, allowing us to uncover
novel weakly-associated genetic variants. However, we cannot discard the remaining genes which could also be interesting candidates.

After the selection of 19 genes (see Materials and Methods) with suggestive GWAS p-values plus other criteria described above, expression studies demonstrated that 12 of them showed a clear expression pattern in the mouse cochlea. Nine of these were successfully replicated in an independent cohort of people geographically and genetically distant from those used to perform the first GWAS analysis (Girotto et al. 2011). This finding is extremely relevant taking into consideration that in four cases the replication was obtained using the same SNPs thus strongly suggesting that the genetic effect observed is of general relevance to multiple human ethnic groups (Bush & Moore 2012). In line with the criteria used by other large consortia (Dumitrescu et al. 2011; Plenge et al. 2011), we used a nominal significance of p<0.05 for replication using non-European populations.

Finally, the clinical step looking at genotype-phenotype relationships highlighted, in 7 cases, the presence of different audiometric profiles of the three genotypes. Two of them (ARSG, SLC16A6) showed striking expression at the top of sensory hair cells in the cochlea (SLC16A6 only at the top of outer hair cells), DCLK1 was expressed only in marginal cells of the stria vascularis, while the others (CDH13, GRM8, ANK2 and RIMBP2) were expressed more widely in multiple cell types in the cochlea.

DCLK1 is involved in several different cellular processes, including neuronal migration in the developing brain and in maturation of the nervous system (Lin et al. 2000). Dclk1 expression within the cochlea is localized to the marginal cells of the stria vascularis, a structure that is essential for the secretion of K+ into the endolymph and for maintaining its associated endocochlear potential, a feature that enhances the electrochemical gradient across the top of hair cells, making them more sensitive. Dclk1 could be involved in the development of the extensive baso-lateral processes extended by the marginal cells. In this light, the hearing improvement present in homozygous BB subjects might be due to a variation in endolymph homeostasis.
SLC16A6 and ARSG, which are located in the same locus overlapping each other and were both replicated under the additive model, display a downward slope for the homozygous BB genotype very similar to that shown by DCLK1. SLC16A6 belongs to the solute carrier (Slc) family, whose members have already been associated with different forms of deafness (Liu et al. 2003), and ARSG is involved in hormone biosynthesis, modulation of cell signalling, and degradation of macromolecules. Both are expressed at the top of the hair cells, where the mechanical forces evoked by sound are translated into an electrical signal, so any variants in these genes could alter the sensitivity of hearing.

A genotype-phenotype difference, with a slight improvement of threshold in the homozygous AA subjects, is evident for RIMBP2; interestingly, it maps within the DFNA41 locus associated with dominantly inherited progressive hearing loss (Blanton et al. 2002). RIMBP2 is a member of a family of proteins that act as binding partners of the presynaptic active zone proteins RIMs (Mittelstaedt & Schoch 2007) as well as for voltage-gated Ca2+-channels, such as CACNA1B and CACNA1D, the latter already known to be involved in deafness (Baig et al. 2011). In this light, the strong staining of this protein in different cell types in the cochlea and the possible interaction with CACNA1D could indicate an important role in hearing function.

A similar audiometric slope for the three genotypes and in particular for the homozygous BB subjects is also evident for three genes expressed in multiple cell types in the cochlea: GRM8, ANK2 and CDH13 (this last one also replicated under a dominant model, while the first two were replicated under a standard additive model). These three genes belong to families whose members have been already reported as expressed in the hearing system. CDH13 belongs to the cadherin superfamily genes (El-Amraoui & Petit 2013). ANK2 is a member of the Ankyrin family of proteins with protein domains found in TRP channels which may be important for the mechano-sensitive channel responsible for hearing (Gaudet 2008). GRM8 is a member, together with GRM7, of the group III metabotropic glutamate receptors (GRMs), which are neurotransmitter receptors (Cartmell & Schoepp 2000). Furthermore,
GRM8 and GRM7 proteins show 87% homology and 76% identity to each other (Schulz et al. 2002) and GRM7 has been recently reported to be involved in ARHL (Friedman et al. 2009; Newman et al. 2012). Consequently the strong expression of Grm8 in the mouse cochlea supports a role for this gene.

As regards CSMD1 and EVI5, despite their strong expression in the inner ear and their replication in the Silk Road cohort (under the additive model), they didn’t show any specific genotype-phenotype audiometric pattern. Previous analysis of Csmd1 mRNA expression by in situ hybridization and immunolabelling of neurons indicated that the primary sites of synthesis are the developing CNS and epithelial tissues (Kraus et al. 2006), while we have shown that the staining is localized at the top of sensory hair cells, including in stereocilia. EVI5 encodes a protein which binds to alpha and gamma tubulin, essential components of microtubules of the inner hair cells (Faitar et al. 2005; Tannenbaum & Slepecky 1997).

As regards genes that were not replicated (GABRG3, PTPRD and GlyBP), their expression patterns suggest they are worthy of further investigation, and GABRG3 and PTPRD both belong to families previously implicated in auditory function (Shahin et al. 2010; Nayak et al. 2011; Goodyear et al. 2003).

In conclusion, our approach confirms the usefulness of a multi-step method, combining several known techniques in order to further investigate the role of genes identified after a GWAS for hearing function, and increases our knowledge of the genes involved in normal hearing function that might also play a role in hearing loss including presbycusis.
CHAPTER 3

Salt-inducible kinase 3, SIK3, is a new gene associated with hearing
3.1 Background

The following study was our second large investigation of hearing function in a GWAS meta-analysis. As already mentioned in §1.2, not only hearing loss but also hearing function heavily rely on molecular bases, and thus have a strong genetic component. Twin studies estimate a heritability of up to 70-75% for hearing ability (Viljanen et al. 2007; Wolber et al. 2012; Karlsson et al. 1997). Several environmental risk factors for diminished hearing ability have been identified, including noise exposure (Gates et al. 1999; Fransen et al. 2008), cardiovascular disease (Gates et al. 1993), ototoxic medication (Rybak et al. 2007) and smoking (Fransen et al. 2008). Here is a brief summary of the state of the art for complex hearing traits, at the time of the present study. Candidate gene studies supported associations of proteins involved in reactive oxygen species removal including N-acetyl-transferase 2 (NAT2); glutathione S-transferase theta 1 (GSST1) and glutathione S-transferase Mu 1 (GSTM1) with presbycusis (Van Eyken et al. 2007; Unal et al. 2005). Various studies tried to elucidate the genetic background of adult hearing function and age-related hearing loss in genome wide linkage (De Stefano et al. 2003; Garringer et al. 2008; Huyghe et al. 2008) and association studies (Huyghe et al. 2008; Friedman et al. 2009; Van Laer et al. 2010; Girotto et al. 2011). In these studies, several genes for age-related HL were identified, including metabotropic glutamate receptor 7 (GRM7), the deafness, autosomal dominant 18 locus (DFNA18), and IQ motif containing GTPase activating protein 2 (IQGAP2). Furthermore, a meta-analysis of hearing function revealed associations with metabotropic glutamate receptor 8 (GRM8) and protein tyrosine phosphatase receptor type D (PTPRD) (Girotto et al. 2011). However, many reported associations did not meet the criteria for genome-wide significance and replication or follow up in model organisms has been reported for only a few studies (see Chapter 2) (Friedman et al. 2009; Girotto et al. 2014).

In order to optimise power by maximising sample size (Manolio et al. 2009), we performed a GWAS meta-analysis of 8 European population samples,
having hearing ability measured by pure-tone audiometry and summarised by principal components. Genetic variants found associated with hearing ability were examined further for expression of the gene product in mouse cochlear structures.

3.2 Materials and Methods

3.2.1 Subjects
All samples included in this analysis were from the G-EAR consortium or TwinsUK (www.twinsuk.ac.uk). Both collections have been described in depth elsewhere (Girotto et al. 2011; Moayyeri et al. 2012). Recruitment took place in 6 different Southern European communities in Italy and Croatia (Carlantino, Friuli Venezia, Korcula, Split, Cilento, Talana), a twin cohort in the UK and six countries from the Silk Road combined as one cohort. Participants underwent a standard pure-tone audiogram as recommended by the local authorities (British Society of Audiology (BSA 2011) for TwinsUK) and gave blood or saliva samples for DNA extraction. Participants with a reported family history of hearing loss or previous ear diseases causing conductive HL were excluded from the analysis. All research was conducted according to the ethical standards as defined by the Helsinki declaration. The study was approved by the Institutional Review Board of IRCCS-Burlo Garofolo, Trieste, Italy and the National Research Ethics service London-Westminster (REC reference number: 07/H0802/84) as well as by the other involved members. Fully informed written consent was obtained from participants prior to their inclusion in the study.

3.2.2 Phenotypes
Principal components were calculated from pure-tone thresholds measured in decibel hearing loss (dB HL) (250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz and 8 kHz) as previously described (Girotto et al. 2011). The principal components represented the overall horizontal threshold shift, slope and concavity of the audiogram, respectively. Principal components (PCs) were calculated
separately for each cohort and adjusted for age and gender using linear regression. Age- and gender-adjusted residuals for PC1, 2 and 3 were rank transformed to normality and used as traits for hearing ability.

3.2.3 Genotyping
Participants from the G-EAR consortium were genotyped on the Illumina 370k chip or on the Affymetrix 500k genotyping array. Genotyping was performed separately for each population. Subjects from the TwinsUK cohort were genotyped on either Illumina HumanHap300 Bead Chip or Illumina HumanHap610 Quad Chip (Moayyeri et al. 2012). Genotype calling was performed using the appropriate software and imputed to the HapMap Phase 2 CEU sample. Genotyped SNPs were excluded from further analysis based on minor allele frequency (MAF), call rate and significance of violation from Hardy-Weingberg equilibrium (pHWE).

3.2.4 Imputation
Imputation was performed separately for each of the 8 populations based on haplotypes of the CEU Hapmap Phase II reference population. Populations from the G-EAR cohort (Italy, Croatia and Silk Road) were imputed using MACH 1.0 (Markov Chain based haplotyper) (Li et al. 2010; Li et al. 2009), for TwinsUK using Impute version 2 (Howie et al. 2009). Imputed SNPs with a quality score below 0.4 (info in Impute vs 2) or 0.3 (Rsq in MACH 1.0), respectively, were excluded from the meta-analysis.

3.2.5 Meta analysis and statistical analysis
GWAS was run by each group using age and gender adjusted rank-transformed PC residuals. A linear mixed model regression was performed, assuming an additive genetic model. Population substructure was accounted for in the genome-wide association per population, if present. The meta-analysis was performed by two centres independently using METAL (Willer & Abecasis 2010) with the weighted Z-score option. This method calculates a
signed Z-score dependent on the p-value and direction of effect per allele per population. Z-scores for each allele and sample are then weighted according to the sample size and summed to give a combined score. Extreme Z-scores (positive or negative) indicate a small p-value, while a positive score indicates increased, and a negative score a decreased disease risk (Willer & Abecasis 2010). Meta analysis was performed for PCs 1, 2 and 3.

3.2.6 Immunohistochemistry for Sik3 and Confocal Microscopy

Expression studies in mouse cochlea were performed in Prof. Karen Steel’s lab, for details see Wolber et al. 2014.

3.3 Results

3.3.1 Subjects

The characteristics of the 4939 individuals are summarized by sample in Table 1.

<table>
<thead>
<tr>
<th>Population</th>
<th>Country</th>
<th>Sample size</th>
<th>gender (%females)</th>
<th>age range (years)</th>
<th>Mean age (± sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlantino</td>
<td>Italy</td>
<td>280</td>
<td>56.87%</td>
<td>18-89</td>
<td>53.29 (18.1)</td>
</tr>
<tr>
<td>Friuli Venezia</td>
<td>Italy</td>
<td>1097</td>
<td>60.19%</td>
<td>18-89</td>
<td>51.47 (16.3)</td>
</tr>
<tr>
<td>Guilia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korcula</td>
<td>Croatia</td>
<td>804</td>
<td>63.30%</td>
<td>18-98</td>
<td>56.3 (13.7)</td>
</tr>
<tr>
<td>Split</td>
<td>Croatia</td>
<td>497</td>
<td>56.00%</td>
<td>18-79</td>
<td>49.0 (14.6)</td>
</tr>
<tr>
<td>Cilento</td>
<td>Italy</td>
<td>421</td>
<td>56.67%</td>
<td>18-91</td>
<td>56.3 (17.6)</td>
</tr>
<tr>
<td>Talana</td>
<td>Italy</td>
<td>470</td>
<td>59.00%</td>
<td>18-92</td>
<td>50.82 (18.5)</td>
</tr>
<tr>
<td>Silk Road</td>
<td>Azerbaijan Giorgia Kazakistan Tajikistan Uzbekistan</td>
<td>348</td>
<td>54.60%</td>
<td>18-82</td>
<td>41.59 (15.5)</td>
</tr>
<tr>
<td>Twins UK</td>
<td>United Kingdom</td>
<td>1022</td>
<td>100%</td>
<td>29-86</td>
<td>61.06 (9.1)</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>4939</td>
<td>-</td>
<td>18-98</td>
<td>-</td>
</tr>
</tbody>
</table>
Age for the different study samples ranged from 18 to 98 years. There was a preponderance of female participants (55% - 100%) across the groups. The first three PCs of the pure-tone audiogram accounted for 87% of the variance in the data for TwinsUK and so were used as phenotypes for the GWAS. PC1 represented a measure of the horizontal threshold shift in the audiogram and, in TwinsUK, is highly correlated with a pure-tone average calculated over all frequencies (r=0.98). PC2 represented the slope of the audiogram, and PC3 gave a measure of its concavity.

3.3.2 Meta-analysis
The most significantly associated SNPs identified in GWAS meta-analysis of PC1, PC2 and PC3 are shown in Tables 2-4 respectively.

Table 2. Meta analysis results for PC1
Five single nucleotide polymorphisms (SNPs) were suggestive genome-wide (p<0.5x10^-7), associated with PC1. The table shows results for PC1 characterised by non-effect allele (all1) and effect allele (all2), the number of individuals with genotyping or imputation data for the respective SNP (N), the resulting Z-score and significance of association (p-value). The direction of effect (minus or plus) is indicated for each of the 8 included populations. If a SNP did not pass QC criteria for a certain cohort, this is indicated by a question mark (?) in the direction column. Mapping information for each SNP is specified by chromosome (chr), base-pair position (position), genes at this locus (gene) and the feature of the SNP position within a gene (feature).

<table>
<thead>
<tr>
<th>SNP</th>
<th>all1</th>
<th>all2</th>
<th>N</th>
<th>Z-score</th>
<th>p-value</th>
<th>direction</th>
<th>chr</th>
<th>position</th>
<th>gene</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7182802</td>
<td>t</td>
<td>c</td>
<td>4725</td>
<td>-5.3</td>
<td>1.16E-07</td>
<td>+----------</td>
<td>15</td>
<td>24929952</td>
<td>GABRG3</td>
<td>intron</td>
</tr>
<tr>
<td>rs1400460</td>
<td>a</td>
<td>c</td>
<td>4680</td>
<td>-5.164</td>
<td>2.41E-07</td>
<td>---------</td>
<td>8</td>
<td>128048873</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rs2455561</td>
<td>t</td>
<td>g</td>
<td>4797</td>
<td>-5.008</td>
<td>5.51E-07</td>
<td>---------</td>
<td>15</td>
<td>84127352</td>
<td>KHL25</td>
<td>intron</td>
</tr>
<tr>
<td>rs717136</td>
<td>a</td>
<td>g</td>
<td>4449</td>
<td>4.929</td>
<td>8.27E-07</td>
<td>+++++++</td>
<td>11</td>
<td>44481121</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rs17575933</td>
<td>t</td>
<td>c</td>
<td>4791</td>
<td>-4.926</td>
<td>8.41E-07</td>
<td>---------</td>
<td>15</td>
<td>84060543</td>
<td>AKAP13</td>
<td>intron</td>
</tr>
</tbody>
</table>

Table 3. Meta analysis results for PC2
The table displays 6 suggestive and 1 significant SNPs associated with this trait. The legend is the same as for Table 2.

<table>
<thead>
<tr>
<th>SNP</th>
<th>all1</th>
<th>all2</th>
<th>N</th>
<th>Z-score</th>
<th>p-value</th>
<th>direction</th>
<th>chr</th>
<th>position</th>
<th>gene</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs681524</td>
<td>t</td>
<td>c</td>
<td>4322</td>
<td>-5.505</td>
<td>3.69E-08</td>
<td>---------</td>
<td>11</td>
<td>116253524</td>
<td>SIK3</td>
<td>intron</td>
</tr>
<tr>
<td>rs1393902</td>
<td>a</td>
<td>g</td>
<td>4829</td>
<td>5.119</td>
<td>3.07E-07</td>
<td>+++++++++</td>
<td>8</td>
<td>68584119</td>
<td>CPA6</td>
<td>intron</td>
</tr>
<tr>
<td>rs1503369</td>
<td>t</td>
<td>c</td>
<td>4829</td>
<td>5.116</td>
<td>3.12E-07</td>
<td>+++++++++</td>
<td>8</td>
<td>68584552</td>
<td>CPA6</td>
<td>intron</td>
</tr>
<tr>
<td>rs1827524</td>
<td>a</td>
<td>g</td>
<td>4831</td>
<td>5.104</td>
<td>3.32E-07</td>
<td>+++++++++</td>
<td>8</td>
<td>68587796</td>
<td>CPA6</td>
<td>intron</td>
</tr>
<tr>
<td>rs1393901</td>
<td>t</td>
<td>c</td>
<td>4830</td>
<td>5.1</td>
<td>3.40E-07</td>
<td>+++++++++</td>
<td>8</td>
<td>68587888</td>
<td>CPA6</td>
<td>intron</td>
</tr>
<tr>
<td>rs6472312</td>
<td>t</td>
<td>g</td>
<td>4791</td>
<td>-4.994</td>
<td>5.92E-07</td>
<td>---------</td>
<td>8</td>
<td>68572052</td>
<td>CPA6</td>
<td>intron</td>
</tr>
<tr>
<td>rs1503363</td>
<td>a</td>
<td>g</td>
<td>4766</td>
<td>-4.94</td>
<td>7.80E-07</td>
<td>---------</td>
<td>8</td>
<td>68569258</td>
<td>CPA6</td>
<td>intron</td>
</tr>
</tbody>
</table>

Table 4. Meta analysis results for PC3
The table displays 1 significant SNP associated with this trait. The legend is the same as for Table 2.

<table>
<thead>
<tr>
<th>SNP</th>
<th>all1</th>
<th>all2</th>
<th>N</th>
<th>Z-score</th>
<th>p-value</th>
<th>direction</th>
<th>chr</th>
<th>position</th>
<th>gene</th>
<th>feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12289510</td>
<td>a</td>
<td>g</td>
<td>4489</td>
<td>-4.906</td>
<td>9.31E-07</td>
<td>---------</td>
<td>11</td>
<td>1.24E+08</td>
<td>SLC37A2</td>
<td>intron</td>
</tr>
</tbody>
</table>
After quality control, 2.3 million SNPs common to all study groups were examined. There was a single genome-wide significant SNP (p<0.5 x 10^{-8}) on chromosome 11 found associated with PC2, the principal component representing the slope of the audiogram (Figure 1, locus zoom of SNP ±400 kB).

**Figure 1. Locus zoom plot of the associated genetic markers on chromosome 11q23.3**
The locus zoom depicts the location of genetic markers versus their significance of association with PC2. Significance of association is measured as the negative logarithm of the p-value. Genes located in the area 400kb up-and downstream of rs681524 are displayed below the x-axis. The colour of each genetic marker indicates its correlation with the reference SNP rs681524. A legend for the correlation colour scheme is shown in the upper left corner of the locus zoom. Recombination rate is highlighted as light blue peaks.

A forest plot of the results for this SNP rs681524 is shown in Figure 2 with corresponding data in Table 5.

**Figure 2. Forest plot for the PC2 meta analysis result for rs681524 in SIK3**
The forest plot shows the effect sizes (beta± standard error) at rs681524 on PC2 for each population in comparison to a combined meta-analysis effect (total). A consistent negative effect for the minor C allele at rs681524 for all European populations could be determined. The beta± standard error only crosses the zero line only for subjects collected from the Silk Road. Corresponding data is listed in Table 5.
Table 5. GWAS and meta-analysis results for rs681524

The GWAS results for each population at rs681524 are characterised by the number of subjects with genotyping or imputation data at this SNP (n), the direction of effect for the C allele (beta), standard error of the direction of effect (se), significance of association and Z-score calculated for the meta analysis. The combined meta-analysis results are listed as Total. A graphical interpretation of this data can be seen in Figure 2. SNP rs681524 did not pass quality control criteria for the population of Talana.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>beta</th>
<th>se</th>
<th>p-value</th>
<th>Z-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlantino</td>
<td>280</td>
<td>-0.2438</td>
<td>0.2215</td>
<td>0.2710</td>
<td>-1.101</td>
</tr>
<tr>
<td>Cilento</td>
<td>419</td>
<td>-0.2206</td>
<td>0.1958</td>
<td>0.2599</td>
<td>-1.127</td>
</tr>
<tr>
<td>Fruili Venezia Guilia</td>
<td>1097</td>
<td>-0.1836</td>
<td>0.0728</td>
<td>0.0117</td>
<td>-2.522</td>
</tr>
<tr>
<td>Korcula</td>
<td>794</td>
<td>-0.1695</td>
<td>0.1153</td>
<td>0.1415</td>
<td>-1.470</td>
</tr>
<tr>
<td>Split</td>
<td>497</td>
<td>-0.1666</td>
<td>0.1276</td>
<td>0.0011</td>
<td>-3.265</td>
</tr>
<tr>
<td>Silk Road</td>
<td>255</td>
<td>-0.0579</td>
<td>0.1705</td>
<td>0.7342</td>
<td>-0.340</td>
</tr>
<tr>
<td>Talana</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TwinsUK</td>
<td>980</td>
<td>-0.3423</td>
<td>0.0914</td>
<td>1.8 x10^-4</td>
<td>-3.745</td>
</tr>
<tr>
<td>Total</td>
<td>4322</td>
<td>-0.2439</td>
<td>-</td>
<td>3.7x10^-8</td>
<td>-5.505</td>
</tr>
</tbody>
</table>

The forest plot in Figure 2 shows the estimated effect sizes (beta± standard error) of the C allele at rs681524 for different samples and a combined meta-analysis effect (total beta = -0.24). This SNP was genotyped in TwinsUK and imputed in the other samples but failed quality control in the sample from Talana, so meta-analysis was performed on 7 out of the 8 groups. The forest plot shows a consistent direction of effect of this allele across all 7 samples, which is statistically significant in all but one sample (from the Silk Road). To control for differences in allele frequency between the different samples, data
from the human genome diversity project (Cavalli-Sforza 2005) was used. No significant differences could be detected for rs681524 in the samples studied here. The locus zoom in Figure 1 shows little in the way of supporting SNPs associated with the hearing phenotype. Further evidence that this finding was not a false positive was sought. Exploration of the linkage disequilibrium (LD) pattern surrounding SNP rs681524 was made using data from the UK10k project. This collection of UK samples including twins and singletons (n=3621) having whole genome-sequencing data confirmed that there were no adjacent SNPs in strong LD. Visual inspection of the cluster plots of rs681524 in TwinsUK revealed clear separation of the alleles. Taken together, these observations suggest the association of rs681524 with hearing PC2 is likely to be a real one.

For PC1, Z-scores ranged from -5.3 to 4.9 (Table 2). The most significant association was detected on chromosome 15 for SNP rs7182802 (p=1.16 x10^-7, Z-score=-5.3) located in an intron of the gamma-aminobutyric acid A receptor gamma 3 (GABRG3) gene (Figure 3).

**Figure 3. Locus Zoom plot of the associated genetic markers on chromosome 15**
The locus zoom depicts the location of genetic markers versus their significance of association with PC1. Significance of association is measured as the negative logarithm of the p-value. Genes located in the area 400kb up-and downstream of rs7182802 are displayed below the x-axis. The colour of each genetic marker indicates its correlation with the reference SNP rs7182802. A legend for the correlation colour scheme is shown in the upper left corner of the locus zoom. Recombination rate is highlighted as light blue peaks.
A forest plot of the association of PC1 with the C allele at rs7182802 shows a consistent negative effect for each population (Table and Figure 4). However, this effect was not significant in samples from Talana and the Silk Road.

Table 6. GWAS and meta analysis results for rs7182802

The GWAS results for each population at rs7182802 with PC1 are characterised by the number of subjects with genotyping or imputation data at this SNP (n), the direction of effect for the C allele (beta), standard error of the direction of effect (se), significance of association and Z-score calculated for the meta analysis. The combined meta-analysis results are listed as Total. A graphical interpretation of this data can be seen in Figure 4.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>beta</th>
<th>se</th>
<th>p-value</th>
<th>Z-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talana</td>
<td>468</td>
<td>0.111963</td>
<td>0.125408</td>
<td>0.372</td>
<td>0.893</td>
</tr>
<tr>
<td>Silk Road</td>
<td>255</td>
<td>-0.031278871</td>
<td>0.067273</td>
<td>0.642</td>
<td>-0.465</td>
</tr>
<tr>
<td>Split</td>
<td>497</td>
<td>-0.133231</td>
<td>0.075792</td>
<td>0.07877</td>
<td>-1.758</td>
</tr>
<tr>
<td>Korcula</td>
<td>794</td>
<td>-0.216977</td>
<td>0.069865</td>
<td>0.001899</td>
<td>-3.106</td>
</tr>
<tr>
<td>Fruili Venezia Guilia</td>
<td>1097</td>
<td>-0.110822</td>
<td>0.051406</td>
<td>0.0311</td>
<td>-2.156</td>
</tr>
<tr>
<td>Cilento</td>
<td>419</td>
<td>-0.166712</td>
<td>0.102418</td>
<td>0.1036</td>
<td>-1.628</td>
</tr>
<tr>
<td>Carlantino</td>
<td>280</td>
<td>-0.397804</td>
<td>0.112783</td>
<td>0.00042</td>
<td>-3.527</td>
</tr>
<tr>
<td>TwinsUK</td>
<td>915</td>
<td>-0.180350185</td>
<td>0.063587</td>
<td>0.004565</td>
<td>-2.836</td>
</tr>
<tr>
<td>Total</td>
<td>4725</td>
<td>-0.140085201</td>
<td>0.000000116</td>
<td>0.000000116</td>
<td>-5.3</td>
</tr>
</tbody>
</table>

Figure 4. Forest plot for the PC1 meta analysis result for rs7182802 in GABRG3

The forest plot shows the effect sizes (beta± standard error) at rs7182802 on PC1 for each population in comparison to a combined meta-analysis effect (total). A consistent negative effect for the minor C allele at rs7182802 for most European populations could be determined. The beta± standard error only crosses the zero line for subjects collected from Talana and the Silk Road. Corresponding data is listed in Table 6.
GABRG3 has previously been associated with PC1 (p=9.9×10−6) in a meta-analysis incorporating the data from the G-EAR cohort (Carlantino, Friuli Venezia, Korcula, Split, Cilento and Talana), which is also used in this analysis (Girotto et al. 2011). GABRG3 encodes a receptor for gamma-aminobutyric acid, the major inhibitory neurotransmitter in the vertebrate brain (Watanabe et al. 2002).

Z-scores determined for associations with PC3 ranged from -4.906 to 4.706 (Table 5). Suggestive significant association with PC3 were determined for rs12289510, an intronic SNP of the solute carrier family 37 member 2 (SLC37A2) gene (p=9.31×10−7, Z-score = -4.906), Figure 5.

**Figure 5. Locus Zoom plot of the associated genetic markers on Chromosome 11**
The locus zoom depicts the location of genetic markers versus their significance of association with PC3. Significance of association is measured as the negative logarithm of the p-value. Genes located in the area 400kb up-and downstream of rs12289510 are displayed below the x-axis. The colour of each genetic marker indicates its correlation with the reference SNP rs12289510. A legend for the correlation colour scheme is shown in the upper left corner of the locus zoom. Recombination rate is highlighted as light blue peaks.
A forest plot of the association of PC3 with the G allele at rs12289510 shows a consistent negative effect for each population. However, this effect was not significant in samples from Talana and the Silk Road (Table 7 and Figure 6).

**Table 7. GWAS and meta-analysis results for rs12289510**
The GWAS results for each population at rs12289510 with PC3 are characterised by the number of subjects with genotyping or imputation data at this SNP (n), the direction of effect for the C allele (beta), standard error of the direction of effect (se), significance of association and Z-score calculated for the meta analysis. The combined meta-analysis results are listed as Total. A graphical interpretation of this data can be seen in Figure 6.

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>beta</th>
<th>se</th>
<th>p-value</th>
<th>Z-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talana</td>
<td>467</td>
<td>-0.10221</td>
<td>0.113812</td>
<td>.3692</td>
<td>-0.898</td>
</tr>
<tr>
<td>Silk Road</td>
<td>255</td>
<td>-0.02702</td>
<td>0.083579</td>
<td>.7465</td>
<td>-0.323</td>
</tr>
<tr>
<td>Split</td>
<td>497</td>
<td>-0.10963</td>
<td>0.066939</td>
<td>.1015</td>
<td>-1.638</td>
</tr>
<tr>
<td>Korcula</td>
<td>794</td>
<td>-0.09505</td>
<td>0.042067</td>
<td>.02386</td>
<td>-2.259</td>
</tr>
<tr>
<td>Fruili Venezia Giulia</td>
<td>1097</td>
<td>-0.11837</td>
<td>0.039413</td>
<td>.002671</td>
<td>-3.003</td>
</tr>
<tr>
<td>Cilento</td>
<td>419</td>
<td>-0.20148</td>
<td>0.088663</td>
<td>.02306</td>
<td>-2.272</td>
</tr>
<tr>
<td>Carlantino</td>
<td>280</td>
<td>-0.15858</td>
<td>0.086336</td>
<td>.06624</td>
<td>-1.837</td>
</tr>
<tr>
<td>TwinsUK</td>
<td>680</td>
<td>-0.05906</td>
<td>0.056598</td>
<td>.2967</td>
<td>-1.043</td>
</tr>
<tr>
<td>Total</td>
<td>4489</td>
<td>-0.10769</td>
<td>0.000000931</td>
<td>4.906</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6. Forest plot for the PC3 meta-analysis for rs12289510 in SLC37A2**
The forest plot shows the effect sizes (beta± standard error) at rs12289510 on PC3 for each population in comparison to a combined meta-analysis effect (total). A consistent negative effect for the minor G allele at rs12289510 for all European populations could be determined. The beta± standard error only crosses the zero line for subjects collected from Talana and the Silk Road. Corresponding data is listed in Table 7.
SLC37A2 has been reported to act as a glucose-to-phosphate antiporter involved in blood glucose homeostasis (Pan et al. 2011). Furthermore, SLC37A2 has been suggested as a candidate gene for human deafness disorders based on a bioinformatics approach (Alsaber et al. 2006).

3.3.3 Expression analysis of Sik3
Sik3 was highly expressed at all three developmental stages (Wolber et al. 2014).

3.4 Discussion and conclusions
We have performed the largest GWAS meta-analysis to date of hearing ability using 8 samples from across Europe, giving total sample size of 4765 adults. We have identified a single SNP in SIK3 (p=3.7x10^-8) as significantly associated with hearing PC2.
Concern that the finding of the single associated SNP on chromosome 11 represented a false positive was addressed in a number of ways. Examination of the region of interest revealed an unusual pattern of LD: UK10K data using whole genome sequencing were found to be consistent with HapMap CEU
data, with very low LD across the region. Further support for the validity of our findings on chromosome 11 is lent by genome-wide suggestive hits (p=5x10^{-5}-5x10^{-8}) which fell within genes GABRG3 and SLC37A2 - both of which represent excellent candidates for hearing ability: expression studies are planned for these genes.

The SIK3 gene was considered a good candidate for hearing ability – and thus suitable for immunohistochemical follow-up - because of the involvement of other salt-inducible kinases (SIKs) in the inner ear. Immunohistochemistry was performed using mouse cochlea in order to localize expression of the protein to a single cell type, rather than to the cochlea in general. Sik3 was found expressed in the apex of the hair cells of murine cochlea during early development and in cells of the spiral ganglion throughout development and adulthood. Furthermore it was expressed in macrophage-like melanocytes in the stria vascularis and perilymph facing cells at the Reissner's membrane. Confocal microscopy confirmed the expression of Sik3 in both the apex of the hair cells around the bases of the stereocilia of 5-day old wildtype mice. Expression of Sik3 in the apex of the hair cells was limited to early stages of development. Mice pups are born deaf and develop the ability to hear at around day 12 (Beutner & Moser 2001). Our data suggest that Sik3 is important for hair cell or stereocilia development but is not required for maintenance of these structures in adulthood. In contrast Sik3 was consistently expressed in spiral ganglia throughout all developmental stages. The striking expression of Sik3 in the murine cochlea supports a role during development and maintenance of the inner ear.

We compared the expression of Sik3 to that of molecular markers for Type 1 and Type 2 spiral ganglion neurons and glial cells. Sik3 expression in the spiral ganglion did not overlap with neurons but with smaller, distinct cells interspersed surrounding them. Comparison of Sik3 and Gfap, a marker for glial cells, in adjacent cochlear sections showed Sik3 limited to the nucleus of labelled cells, while Gfap was expressed in cytoplasmic structures of non-neuronal cells of the spiral ganglion. Sik3 and Gfap expression overlapped partially in the spiral ganglion. In conclusion, Sik3 is expressed in non-
neuronal cells of the spiral ganglion, which we assume to be an undefined subgroup of glial cells.

SIKs belong to the family of cAMP activated serine threonine kinases. Three family members have been described so far (Katoh et al. 2004), and isoforms 1 and 2 have been the main focus of research. Sik1 and 2 show high expression in adrenal glands of mice on a high sodium diet, thus the name “salt-inducible” kinases, while Sik3 was reported to be expressed ubiquitously (Katoh et al. 2004). Sik1 has been linked to various processes, including sodium sensing and cardiomyogenesis (Jaitovich & Bertorello 2010, Romito et al. 2010). In addition, Sik1 has been shown to be expressed in the inner ear, in the sensory epithelium of the vestibular system, where it is thought to be involved in the formation of hydrops via interaction with phosphodiesterases (Degerman et al. 2011). Both Sik1 and Sik3 control histone deacetylases via phosphorylation and nuclear export (Berdeaux et al. 2007; Sasgawa et al. 2012; Walkinshaw et al. 2013). Sik3 knockout mice (Sik3-/-) have a high mortality rate at birth. Surviving pups show skeletal abnormalities, reduced bodyweight and dwarfism but no details of their hearing ability have been reported yet. Sik3-/- mice show reduced energy storage, which is associated with hypoglycaemia and hyper-insulin sensitivity. Lipid metabolism disorders in (Sik3-/-) mice were partially restored after 9-cis-retionionic acid supplementation (Uebi et al. 2012). Sik3 is assumed to regulate cholesterol bile acid homeostasis and lipid storage size and is essential for chondrocyte hypertrophy (Sasgawa et al. 2012; Uebi et al. 2012). In addition, salt-inducible kinases have recently been shown to be involved in the formation of regulatory macrophages. In fact, Salt-inducible kinases were responsible for phosphorylation of the CREB-regulated transcriptional coactivator 3 and could therefore inhibit the formation of regulatory macrophages from proinflammatory macrophages (Clark et al. 2012). Despite having identified a new candidate gene for hearing function, there are details that were beyond the scope of this study. Although we have shown functional expression of Sik3 in mice, Sik3 expression and function in humans awaits confirmation. Replication the effect of SIK3 rs681524 genotypes on
hearing function would be useful. The samples studied in this meta analysis of GWAS were of different ethnic populations which increases the risk of population stratification. To adjust for this, population sub-structure was controlled for in the GWAS performed for each sample separately, where applicable. Different ethnic backgrounds should also be considered when determining LD patterns of meta analysis results. LD pattern displayed in the locus zoom plot was based on LD as observed in the HapMap CEU population, however, this might not represent LD pattern in Samples from the Silk Road. To explore this possible source of error, we compared allele frequencies for the respective SNPs in data from the Human Genome Diversity project (Cavalli-Sforza 2005). Furthermore, it should be mentioned that the cohorts examined here had, overall, better health and hearing ability for the higher age group than previously reported (Mathers et al. 2008). We assume that this recruitment bias is influenced by the fact that subjects had to visit the test centre, which might hinder participation of less healthy volunteers. In addition, while the age range for the cohorts was very broad (18-98 years) it did vary between the study-groups, and the mean age of the samples (41.6 - 61.1) varied by 20 years. In addition, PCS were adjusted for age at hearing test prior to the GWAS. Previous studies reported a gender difference in hearing ability particularly for older individuals (Helzner et al. 2005). To control for possible gender differences in hearing ability, we adjusted for this in the GWAS of the mixed gender cohorts (all but TwinsUK). Furthermore, PCs rather than standard measures of hearing ability (i.e. pure-tone averages) were used to capture hearing ability. PCs might be more complex in their interpretation in respect to hearing, but they enable representation of both overall threshold shift and slope of the audiogram. In addition, PCs have been applied to capture pure-tone audiograms previously (Huyghe et al. 2008; Girotto et al. 2011).

In conclusion, this study reports the findings of the largest meta-analysis of hearing function to date. A SNP in the SIK3 gene was genome-wide significantly associated with the slope of the audiogram. This association as well as the specific expression pattern of Sik3 throughout the mouse cochlear
highlight Sik3 as playing a role in hair cell development as well as adult hearing ability.
CHAPTER 4

Genome-wide association analysis on Normal Hearing Function identifies PCDH20 and SLC28A3 as candidates for hearing function and loss
4.1 Background
In our previous GWAS meta-analysis, using European isolated populations, we detected several suggestive associations (Girotto et al. 2011). A later follow-up of these results highlighted 12 genes, which were replicated and validated at the expression level (Chapter 2). Furthermore, a larger GWAS meta-analysis identified SIK3 (Chapter 3), a member of the salt-inducible kinase family known to be involved in the inner ear. In a candidate gene study, the estrogen-related receptor gamma (ESRRG) was suggested as a potential player in the maintenance of hearing in both humans (mainly females) and mice (Nolan et al. 2013).

Considering these findings and the extremely intricate mechanisms underlying the auditory process, we reasoned that GWAS might still help in the detection of new genes involved in the variability of the hearing process. In particular we were able to impute genetic data more accurately, thanks to the release of the 1000 Genomes Project reference panel (The 1000 Genomes Project Consortium, 2010). The final information available consisted in approximately 8 million SNPs per cohort, compared to about 2 million used in previous GWAS (HapMap project-based imputation).

We therefore performed a GWAS meta-analysis based on the new imputation data in our Italian isolated populations, and identified two statistically significant candidates for modulating normal hearing function. These hits were then replicated in the Silk Road cohort and in two independent cohorts from UK and Finland, as well as validated at expression level in the mouse cochlea.

4.2 Materials and Methods

4.2.1 Sample collection, preparation and genotyping
All studies had ethical approval obtained from the Institutional Review Board of IRCCS – Burlo Garofolo, Trieste, Italy and other involved institutions. Consent forms for clinical studies were signed by each participant and all
research was conducted according to ethics standards defined by the Helsinki declaration. Detailed information about each cohort is shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Description of the analysed cohorts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort Name</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>DNA collection</td>
</tr>
<tr>
<td>SNP chip</td>
</tr>
<tr>
<td>Genotype calling software</td>
</tr>
<tr>
<td>Genotype filtering exclusion criteria</td>
</tr>
<tr>
<td>Call rate&lt;0.97</td>
</tr>
<tr>
<td>HWE exact test p&lt;1e-08</td>
</tr>
<tr>
<td>Imputation filtering exclusion criteria</td>
</tr>
<tr>
<td>Info score&lt;0.4</td>
</tr>
<tr>
<td>Total SNPs after imputation</td>
</tr>
<tr>
<td>Age range</td>
</tr>
<tr>
<td>Sex females (%)</td>
</tr>
</tbody>
</table>

**Italian isolated populations**

Altogether 2155 subjects from four different cohorts (INGI-FVG, INGI-CARL, Talana, and Cilento) from several isolated communities located in Northeast Italy, South Italy and Sardinia were analysed. These are the samples already described in previous work (Girotto et al. 2011) and in Chapters 2 and 3. Moreover The Silk Road cohort (SR), was considered in this study for replication purpose (as in Chapter 2), for a total of 481 people.

**1958 UK Birth Cohort (B58C)**
The B58C and the collection of hearing data have been described previously (Wolber et al. 2014; Strachan et al. 2007). In brief, participants were drawn up from 17,638 individuals born in England, Scotland, and Wales within one week of March 1958. Of the original cohort, 9377 members were revisited by a research nurse for a biomedical follow-up in 2002–2004. The hearing measure consisted of pure tone audiometry at 1 and 4 kHz at age 44–45 years and were adjusted for sex, conductive loss, hearing loss in childhood and nuisance variables (noise at test, nurse performing test, audiometer used in test). Genome-wide data for the 1958BC were obtained through several sub-studies, using them as population controls and genotyping a total of 6099 individuals (for details see http://www2.le.ac.uk/projects/birthcohort/1958bc/available-resources/genetic). These include the Welcome Trust Case Control Consortium study (WTCCC1 and 2) (The Wellcome Trust Case Control Consortium 2007), the Type 1 diabetes genetics consortium (T1DGC)(Barrett et al. 2009) and the GABRIEL Large-Scale Genome-Wide Association Study of Asthma (Moffatt et al. 2010).

The Finnish Twin Study on Ageing (FITSA)
The FITSA cohort includes 217 female twin pairs, aged 63-76 years, recruited in order to investigate the contribution of genetic and environmental factors to the disablement process, and to investigate the effects of ageing on several physiological and functional traits (Viljanen et al. 2007). In the current study, a total of 270 subjects from 196 twin pairs with hearing phenotype, covariates, and GWAS data available were included in the analysis (91 subjects from 91 monozygotic pairs and 179 subjects from 105 dizygotic pairs, of which 74 included both co-twins and 31 included only one co-twin).

4.2.2 Imputation and filtering
In order to obtain a homogenous set of markers among all the cohorts and to increase the statistical power, genotypes were imputed following the same protocol in each cohort. In particular, SHAPEIT2 (Delaneau et al. 2012) was used for the phasing step and IMPUTE2 for the imputation step to the 1000-
Genomes phase I v3 reference set (Howie et al. 2012). After imputation, SNPs with MAF < 0.01 or imputation quality score (Info) < 0.4 were excluded from the statistical analyses. Table 1 shows the final number of SNPs analysed after imputation and filtering.

4.2.3 Audiometric evaluation
Audiometric data for the Italian cohorts and SR, as well as for the FITSA cohort, were the same as in previous Chapters. For the B58C cohort, only 1 and 4 kHz were available. Overall, nine hearing traits were analysed: 0.25, 0.5, 1, 2, 4, 8 kHz, PTAL, PTAM, PTAH. Familial forms of inherited hearing loss were excluded from the study, as well as subjects affected by diabetes or other systemic disorders producing hearing loss.

4.2.4 Association analysis
Genome wide association studies were separately performed in each cohort, testing individual hearing thresholds and the PTAs (after adjustment for sex and age) in a linear mixed model in order to account for genomic kinship. Four different genetic models were tested: additive, dominant, recessive and overdominant. All the analyses were carried out using the GRAMMAR-Gamma method as implemented in the GenABEL suit (Svishcheva et al. 2012) for genotyped SNPs and MixABEL (Aulchenko et al. 2007) for imputed data. We did not consider any p-value correction for the different genetic models tested because these are not four independent tests (Parsons et al. 2013).

In order to increase the statistical power, a two-stage meta-analysis approach was used. The first stage included only Italian cohorts, using a fixed-effects meta-analysis with inverse variance weights as implemented in METAL software (Willer et al. 2010). Resulting SNPs with p-value<1e-06 were carried on by joining the SR cohort and repeating the analysis (second stage). The final significance level was set to 5e-08.
Finally, following the same analysis protocol, two independent cohorts (B58C and FITSA) were investigated for additional replication, at a nominal level of significance (p-value<0.05).

For each GWAS and meta-analysis, population stratification was checked by computing the lambda coefficient of inflation. The phenotypic variance explained by single SNPs was computed from the mixed model regression score estimates.

4.2.5 RNA extraction, reverse transcriptase-polymerase chain reaction (RT-PCR) and real-time quantitative RT-PCR

RT-PCR in mouse cochlea was performed in collaboration with Dr. Di Stazio. Details are available upon request.

4.2.6 RNA-Seq expression profiling

RNA-Seq for hair cells in mouse cochlea and vestibular system were obtained in collaboration with Dr. Scheffer. The data are available to the community on a dedicated database: Shared Harvard Inner-Ear Laboratory Database (SHIELD), https://shield.hms.harvard.edu/.

4.2.7 Sequencing

We selected a pool of 48 subjects from the INGI-FVG cohort for sequencing the candidate genes. The subjects were divided in four groups based on their genotypes at the top two associated SNPs. In particular, following the genetic models of the associations detected, the groups consisted of alternative homozygotes versus reference homozygotes and/or heterozygotes for each allele (see Results). For the PCDH20 gene the selection was done according to the dominant model of association as follows: (a) 12 carriers of the effect allele of rs78043697 (CC or CT) and (b) 12 homozygotes for the non-effect allele (TT). For the SLC28A3 gene, the selection was done under the additive model as follows: (c) 12 homozygotes for the effect allele of rs7032430 (AA) and (d) 12 homozygotes for the non-effect allele (CC). The coding sequence
and the intron-exon boundaries were analysed by Sanger sequencing. A standard PCR was carried out at 60°C, for 35 cycles using KAPA2G Fast ReadyMix PCR Kit (Kapa Biosystems), according to the manufacturer's protocol. Sequencing was performed on a 3500Dx Genetic Analyzer (Life Technologies, CA, USA), using 3.1 Big Dye terminator chemistry (Life Technologies, CA, USA) according to the manufacturer's instructions. Primer sequences are available upon request.

4.3 Results

4.3.1 Statistical association
The first stage meta-analysis was performed on 2155 subjects from four Italian cohorts, testing nine hearing measurements (six single frequencies and three PTAs). There were no signs of population stratification (all lambda coefficients ≤ 1.01) and a total of 223 SNPs reached a p-value < 1e-06. These 223 SNPs were further analysed by combining data from a completely independent cohort of isolated communities (Silk Road cohort, N=481), resulting in a total sample size of 2636. In the combined analysis, 35 out of the 223 SNPs were replicated on the same traits and genetic models, reaching a genome-wide significant p-value < 5e-08. Among the 35 significant SNPs, 32 were located within one large locus on 13q21.31 (chr13:62382925–chr13:62520492 according to build 37) and located close to PCDH20 (Figure 1).

Figure 1. Locus-zoom plot for chromosome 13 locus
The plot shows the associated locus on chromosome 13, highlighting the top SNP rs78043697 and several other significant SNPs in the same region being in high linkage disequilibrium
The top SNP in this region is rs78043697-C which reached a p-value=4.71E-10 on the 2kHz trait under the dominant model. Subjects homozygous for the non-effect allele T (N=1964) have a better hearing compared to the other groups (N=191). The remaining SNPs are in very high linkage disequilibrium with the top SNP (r²>0.8), suggesting that their observed effect was due to linkage disequilibrium with the top associated SNP. To verify this hypothesis we reran the association analysis in the INGI-FVG cohort, which showed the strongest association among the samples included in the first stage meta-analysis (Figure 2), conditioning on the top SNP (rs78043697).

**Figure 2. Forest plot for rs78043697**

The figure shows the forest plot for the top SNP associated to 2 kHz threshold on chromosome 13. The x-axis represents effect sizes (beta coefficients) for trait residuals after sex age and kinship adjustments. The strongest association can be observed in the INGI-FVG cohort.
As expected, no significant residual signal remained (p-values>0.05) (data not shown), suggesting that there is only one independent signal in this locus. The phenotypic variance of the 2kHz hearing explained by rs78043697-C was 1.8%.

The second highest SNP was rs7032430-A (p=2.39E-09 and p=4.22E-09 for 0.5 kHz under additive and dominant models, respectively) located on 9q21.32 and explaining 1.3% of the phenotypic variance of the 0.5 kHz hearing. Subjects carrying the effect allele A (N=1007) have worse hearing compared to those homozygous for the non-effect allele C (N=1622). Several
SNPs in the same locus show association with the same trait. Among these, rs35664751-A, in high linkage disequilibrium ($r^2 > 0.6$) with the top SNP (rs7032430-A), reached genome-wide significance ($p$-value=$1.42\times10^{-8}$) (Figure 3).

**Figure 3. Locus-zoom plot for chromosome 9 locus**

The plot shows the associated locus on chromosome 9 and the genes within the region, including *SLC28A3*, the strongest candidate for hearing function.

These SNPs lie close to several genes: RMI1, HNRNPK, C9orf64, KIF27, GKAP1, UBQLN1, and SLC28A3. Prioritization for follow-up studies was based on known functions, belonging to the same gene family as known hearing genes and distance from the associated SNP. Based on these criteria, SLC28A3 emerged as the strongest candidate, belonging to the “solute carriers family” whose members are already known to be involved in hearing.

Finally, rs2243805-A on 2q14.3, located within LIMS2 and GPR17, reached a genome-wide significant $p$-value of $1.74\times10^{-8}$ at 4kHz under the dominant model. As no additional evidence of association was found in this region, this SNP was not further studied.
Table 2A shows summary statistics for the above-mentioned loci, while Table 2B displays imputation quality (IMPUTE2 info score) across cohorts for the top SNPs.

Table 2A. Two-step meta-analysis results: top SNPs for each locus
Legend: SNP=Single Nucleotide Polymorphism; add=additive; dom=dominant; Chr=Chromosome; EA=Effect Allele; AA=Alternative Allele (non-effect allele); N=number of samples; Beta=beta coefficients; SE=Standard Error; Dir=Direction; + = positive effect; - = negative effect; ? = SNP not present/analysed; N. pop=Number of populations; Orig. p=Original p-value (first-stage meta-analysis).

<table>
<thead>
<tr>
<th>SNP</th>
<th>Trait and model</th>
<th>Chr</th>
<th>Position</th>
<th>EA</th>
<th>AA</th>
<th>N</th>
<th>Beta</th>
<th>SE beta</th>
<th>Final p-value</th>
<th>Dir</th>
<th>N pop.</th>
<th>Orig. p</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs78043697</td>
<td>2kHz dom</td>
<td>13</td>
<td>62467039</td>
<td>C</td>
<td>T</td>
<td>2155</td>
<td>0.119997443</td>
<td>0.019266322</td>
<td>4.71E-10</td>
<td>+++?+</td>
<td>4</td>
<td>1.83E-08</td>
</tr>
<tr>
<td>rs7032430</td>
<td>0.5kHz add</td>
<td>9</td>
<td>86714002</td>
<td>A</td>
<td>C</td>
<td>2629</td>
<td>0.035277526</td>
<td>0.00591061</td>
<td>2.39E-09</td>
<td>+++++</td>
<td>5</td>
<td>1.27E-08</td>
</tr>
<tr>
<td>rs2243805</td>
<td>4 kHz dom</td>
<td>2</td>
<td>128407499</td>
<td>G</td>
<td>A</td>
<td>2632</td>
<td>0.068598345</td>
<td>0.012170387</td>
<td>1.74E-08</td>
<td>+++++</td>
<td>5</td>
<td>6.90E-08</td>
</tr>
<tr>
<td>rs35664751</td>
<td>0.5 kHz add</td>
<td>9</td>
<td>86717006</td>
<td>G</td>
<td>A</td>
<td>2630</td>
<td>0.035460097</td>
<td>0.006253427</td>
<td>1.42E-08</td>
<td>+++++</td>
<td>5</td>
<td>2.06E-07</td>
</tr>
</tbody>
</table>

Table 2B. Imputation quality score across cohorts for the two best hits
Legend: INFO=Imputation Info Score (IMPUTE2)

<table>
<thead>
<tr>
<th>COHORT</th>
<th>INFO rs78043697</th>
<th>INFO rs7032430</th>
</tr>
</thead>
<tbody>
<tr>
<td>INGI-FVG</td>
<td>0.996</td>
<td>0.825</td>
</tr>
<tr>
<td>INGI-CARL</td>
<td>0.995</td>
<td>0.766</td>
</tr>
<tr>
<td>TALANA</td>
<td>NA</td>
<td>0.795</td>
</tr>
<tr>
<td>CILENTO</td>
<td>0.994</td>
<td>0.737</td>
</tr>
<tr>
<td>SR</td>
<td>0.998</td>
<td>0.998</td>
</tr>
</tbody>
</table>

To further prove the validity of our findings, a replication step was performed using the B58C cohort (N=5892), with only two traits available (1 and 4 kHz), and the FITSA cohort (N=270) from Finland. Among the 35 SNPs, 4 SNPs within the 13q21.31 locus were nominally replicated (p<0.05) at the 1kHz frequency in the B58C cohort (Table 3). For the 9q21.32 locus the second best hit rs35664751-A was also nominally replicated (p=0.02761) in the FITSA cohort (Table 3).

Table 3. Independent replication results
Legend: SNP=Single Nucleotide Polymorphism; add=additive; EA=Effect Allele; AA=Alternative Allele; N=number of samples; Beta= beta coefficients; SE=Standard Error; P=p-value

<table>
<thead>
<tr>
<th>SNP</th>
<th>Trait and model</th>
<th>EA</th>
<th>AA</th>
<th>N</th>
<th>Beta</th>
<th>SE beta</th>
<th>P</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs111887296</td>
<td>1kHz add</td>
<td>T</td>
<td>C</td>
<td>5892</td>
<td>-0.0109</td>
<td>0.005435</td>
<td>0.04495</td>
<td>B58C</td>
</tr>
<tr>
<td>rs113274536</td>
<td>1kHz add</td>
<td>T</td>
<td>C</td>
<td>5892</td>
<td>-0.0111</td>
<td>0.005431</td>
<td>0.04072</td>
<td>B58C</td>
</tr>
</tbody>
</table>
4.3.2 Sequencing analysis
The top genes located on chromosome 13 and 9 (PCDH20 and SLC28A3 respectively) were selected for further analyses. To investigate any functional genetic variants marked by the associated tag SNPs (rs78043697 for PCDH20 and rs7032430 for SLC28A3) that contribute to the variation in this trait, a pool of 48 people (24 individuals for each gene) from the INGI-FVG cohort were sequenced. Nine variants within the PCDH20 gene (3 in 3'-UTR; 5 exonic, of which 2 are synonymous, 2 nonsynonymous, and 1 frameshift; and 1 in 5'-UTR) and 42 within SLC28A3 (1 downstream; 18 in 3'-UTR; 5 exonic, of which 2 synonymous and 3 nonsynonymous; 14 intronic; and 4 in 5'-UTR) were identified. However, none of the variants identified was associable with the selected genotypes (i.e. tag SNPs). This finding suggests that GWAS-associated SNPs mark potential regulatory regions more than functional coding variants, as expected for quantitative traits.

4.3.3 Gene expression analysis
RT-PCR showed positive expression for both genes in mouse cochlea (Figure 4).

Figure 4. (a) qReal-time analysis of the selected genes (PCDH20, SLC28A3) (b) Semi quantitative RT-PCR
The plot shows expression for Pcdh20 and Slc28a3, compared to Myo7a and Myo6 (a). The y-axis represents a scale whose unit is the expression of the reference gene Myo7a. Pcdh20 displays a strong expression, while Slc28a3 shows lower levels. Semi-quantitative RT-PCR demonstrates that both genes are expressed (b).
RNA-Seq analysis of FACS-sorted cells in the developing mouse cochlea and utricle demonstrated expression of both genes from embryonic day E16 to postnatal day P7 (Figure 5a,b).

**Figure 5. RNA-seq Expression Profiles for PCDH20 (a) and SLC28A3 (b)**

Histograms display the normalized number of reads in hair cells (green) and surrounding cells (purple), in samples from the cochlea (C; dark colours) and utricle (U; light colours) at E16, P0, P4 and P7. The fold change (FC) representing the GFP+/GFP- counts ratio and the multiple test adjusted false discovery rate (FDR) by the Benjamini-Hochberg procedure are indicated.
4.4 Discussion and Conclusions

Genes involved in hearing loss have been identified through many different techniques, beginning with positional cloning and linkage studies. In the past few years, high-throughput and “omics” technologies, including GWAS and Next Generation Sequencing studies, have led to the identification of new genes/mutations in both Mendelian and complex forms of hearing loss. Despite these successes, there is a need for a better understanding of the molecular bases of hearing in complex forms, and as a quantitative trait. In this field, identification of association of candidate genes at genome-wide statistically significant p-values, replication in independent cohorts, and demonstration of expression in the inner ear can suggest a fundamental role for the gene products in the auditory system (Williamson et al. 2008).

Following this strategy, we performed a GWAS meta-analysis followed by replication in several cohorts of European and Central Asian ancestry. This is the largest collaborative effort established on hearing function so far, with a final sample size of 8707. Data were imputed to the 1000 Genomes
reference panel, resulting in a dense marker map and increased statistical power and precision (Sung et al. 2012). The two-stage meta-analysis approach led to the discovery of two genome-wide significant loci (close to PCDH20 and SLC28A3). This approach was first proposed in order to minimize the required sample size and genotyping without losing power (Hirschhorn & Daly 2005) and has been successfully applied in different studies (Soranzo et al. 2009; Evangelou et al. 2013; Lambert et al. 2013). It has been argued that this is an efficient way of performing a single study and not a proper replication step (Thompson et al. 2011). Thus we sought further nominal replication and we successfully replicated both loci in independent cohorts, implying that, despite genetic heterogeneity of the study populations, genes with global importance can be found.

The analysis of individual frequency thresholds as well as the PTAs allowed us to comprehensively portray the normal hearing function. Although multiple phenotypes were tested, these measurements are not independent, thus reducing the concern about multiple testing; in fact the first principal component of the hearing data accounted for approximately 80% of the total variance, illustrating how the biological trait is essentially unique. Statistical power was also increased by testing four different genetic models, a step of fundamental importance when sample size is limited (Tsepilov et al. 2013).

The top associated SNPs for the detected loci explained 1.3% and 1.8% of the phenotypic variance. These data show that the associating variants had relatively large effect sizes when compared to other known associations, such as height associated markers with variance explained ranging from 0.04% to 1.13% (Park et al. 2010). We acknowledge the possible limitation of the present study in detecting lower effect variants due to lack of power. In this light, larger studies could be useful in the future to overcome this problem.

The top genes identified in this study (PCDH20 and SLC28A3) are both members of gene families already implicated in the inner ear. PCDH20 belongs to the cadherin superfamily, whose members have multiple roles in cellular adhesion (Yagi & Takeichi 2000; Sotomayor et al. 2014) and have been previously implicated in murine hearing loss as well as in the human
Usher syndrome (Liu et al. 2012). Although very little is known about PCDH20, a phylogenetic grouping based on N-terminal sequence suggests that has potential for trans homotypic and heterotypic interactions (Sotomayor et al. 2014). We used RNA-Seq and qRT-PCR for validation of PCDH20 expression, finding it to be expressed especially in the sensory hair cells. This gene is also reported by the GENSAT atlas as expressed in the spiral ganglion of the inner ear at E15.5 (http://www.gensat.org/index.html).

SLC28A3 is a member of the solute-carrier family and is thought to concentrate nucleosides including cytidine, uridine, adenosine and guanosine in a cell by coupling translocation of two H+ or Na+ ions per nucleoside. Its apical location in absorptive epithelia drives vectorial flux of nucleosides, which are essential for nucleotide biosynthesis (Errasti-Murugarren et al. 2009). Three other members of the broad solute carrier family are implicated in inherited deafness (Liu et al. 2004; Li et al. 1998; Ruel et al. 2008). Both qRT-PCR and RNA-Seq showed clear expression of SLC28A3 in mouse cochlea and utricle. Although other genes were present in the same locus, none of them (nor their gene-family members) have previously been implicated in the auditory system in the most common databases (http://omim.org/; http://www.ncbi.nlm.nih.gov/pubmed), nor characterized by a function clearly related to hearing. Thus the best candidate on 9q21.32 for modulating hearing function was SLC28A3. Moreover, GTEx (http://www.gtexportal.org/) eQTL database revealed a significant association between the top SNP rs78043697-C and SLC28A3 expression in stomach tissue. These findings together support the hypothesis that rs78043697-C could affect the expression of SLC28A3, although eQTL data for the inner ear tissue were not available.

No causal alleles were identified in the sequencing of the coding regions, supporting the involvement of regulatory regions underlying these associations. Such a mechanism is common in GWAS studies and it has been recently addressed by combining sequencing, epigenetic and transcription-factor data sets (Chen et al. 2014), which could be a follow-up of the present study.
In conclusion, this study reports the discovery of two new genes significantly associated with auditory function in humans and hypothesizes a specific role in the sensory epithelium of the utricle and cochlea. Although this should be confirmed with further experiments, these results increase our knowledge of the molecular basis of normal hearing function by identifying new important candidates for further investigation.
CHAPTER 5

Multivariate GWAS
5.1 Background
Genome-wide association studies (GWAS) have been a powerful tool for genetic discovery for almost a decade. Results have shed light on many different biological processes from lipid metabolism to blood composition as well as social and behavioural patterns (Hindorff et al. [accessed 2014]). In previous chapters, we reported GWAS on hearing traits, discovering several candidate genes involved in modulating hearing function (Chapters 2 and 3). Starting with six different variables (0.25, 0.5, 1, 2, 4, 8 kHz), we either considered them singularly or combined them in order to summarize the information. For example we used average values such as Pure Tone Averages (PTA) for low frequencies (0.25, 0.5, 1 kHz), medium frequencies (0.5, 1, 2 kHz) and high frequencies (4 and 8 kHz) and/or the first three principal components (PCs) of the phenotype. However it is clear that the overall hearing ability depends on a combination of these traits and that they are correlated with a specific pattern. Thus addressing multiple phenotypes can give great increase in power, by taking into account the underlying correlations between variables. A part from hearing, many other biological features are better described by a combination of several variables, such as blood parameters, lipid related traits, psychological evaluations etc. In this light multivariate analyses are of great interest in genetic association studies.

5.2 Overview of Multivariate GWAS methodology
Several generalizations of standard association studies are available. One approach is based on dimensional reduction of the dataset. Either, traits are analysed by univariate GWAS and results are combined afterwards (van der Sluis et al. 2013), or traits are transformed in order to be summarized. For example PCHAT (Klei et al. 2008) calculates the optimal linear combination of the phenotypes, maximizing heritability, while PLINK performs canonical correlation analysis i.e. selects the linear combination of the traits maximizing the covariance between each marker and all traits (Ferreira & Purcell 2009). Another possible approach consists in reversing the usual regression and
regressing the genotype to a set of phenotypes. This method is implemented in MultiPhen software (O’Reilly et al. 2012).

For a long time multivariate linear regression was computationally infeasible when applied to genome-wide testing. Bayesian multivariate regression analysis overcomes this problem and is available e.g. in SNPTEST software (Marchini et al. 2007). However, one of the most efficient univariate GWAS methods consists in mixed linear regression, which can model fixed effects, such as the genotype and the covariates, as well as random effects which are usually given in form of a kinship matrix accounting for relatedness and population structure (Hoffman 2013). This model has been derived in a two-trait setting, in software called MTMM, which considers both the within-trait and between-trait variance components (Korte et al. 2012). Finally, a new set of algorithms included in the GEMMA software (Zhou & Stephens, 2014) allow for a fast multivariate mixed model fitting and testing up to ten phenotypes in large sample sizes. The method avoids repeated computation for each SNP and instead performs an initial eigendecomposition of the relatedness matrix, reducing computational complexity. It is very fast and performs well for up to 10 phenotypes as well as being a straightforward generalization of univariate linear mixed regressions.

5.3 Statistical power and Meta-analysis

Statistical power for GWAS consists in the probability of observing a significant (or suggestive) result, when a true genotype-phenotype association is present. Due to the very large number of tests performed and to the small effect sizes for each SNP, especially in complex traits, power in GWAS is usually limited. A way to improve its efficiency is to increase sample size. Figure 1 shows the estimates of the probability of detection (power) for different sample and effect sizes (Visscher 2008).

Figure 1: Statistical power of detection in GWAS for variants that explain 0.1–0.5% of the variation at a type I error rate of 5 x 10^{-7}, calculated using the Genetic Power Calculator (Purcell et al. 2003; Visscher 2008).
Simulated data were used to estimate the power increase thanks to multivariate analysis for several methods described in the previous section (Galesloot et al. 2014).

Figure 2. Power of detection for the case in which one of three traits is associated with the SNP
The explained variance of the SNP is set to 0.1%, the sample size was simulated to 1000 subjects (Galesloot et al. 2014).

Legend: MAF=minor allele frequency; MV=multivariate; PCHAT=Principal Component of Heritability Association Test; QTL=quantitative trait locus; rE=residual correlation; rG=genetic correlation induced by the QTL; TATES=Trait-based Association Test that uses Extended Simes procedure; UV-MA=meta-analysis of univariate results; UV-PCA=univariate analysis of the first principal component; UV T1=univariate analysis of trait 1; UV T2=univariate analysis of trait 2; UV T3=univariate analysis of trait 3.
As shown in Figure 2, for a sample size of 1000 individuals, even with multivariate methods the detection is below 30%. Hence, as for the univariate GWAS, it is strongly desirable to be able to combine multivariate results from different studies in order to further increase statistical power in genotype-phenotype associations. At the time of the present study, to our knowledge there was no software for meta-analysing multivariate results. In this work we implemented a novel statistically efficient method to perform meta-analysis in a multivariate setting. It is an inverse-variance based method that allows different weights for each cohort in order to take into account the accuracy of each effect estimate. The inverse-variance method has been successfully used for single-trait association testing (Sanna et al. 2008) and is suitable for n-dimensional generalization. Finally it is implemented as part of the R package MultiMeta to benefit from flexible environment and open access, as well as extra plotting functions for results visualization.
5.4 Methods

The multivariate setting implies that results for each single nucleotide polymorphism (SNP) include several effect sizes (or beta-coefficients, one for each trait), as well as related variance and covariance values, since beta coefficients can be correlated.

Let \( p \) be the number of phenotypes analysed and \( n \) the number of cohorts to include in the meta-analysis. For each cohort \( i \in \{1, \ldots, n\} \) let \( \beta^i \) be the vector of effect sizes and \( \Sigma^i \) be a \( pxp \) variance-covariance matrix.

Input data:

\[
\beta^i = \begin{pmatrix}
\beta^i_1 \\
\vdots \\
\beta^i_p
\end{pmatrix}; \Sigma^i = \begin{pmatrix}
\sigma^i_{11} & \ldots & \sigma^i_{1p} \\
\vdots & \ddots & \vdots \\
\sigma^i_{p1} & \ldots & \sigma^i_{pp}
\end{pmatrix}
\]

Our method to combine effect sizes is an inverse-variance based method. It is an \( n \)-dimensional generalization of the single trait meta-analysis, such as the one implemented in METAL software (Willer, Li, & Abecasis, 2010). In particular, each vector \( \beta^i \) is weighted by the inverse of its variance-covariance matrix \( (\Sigma^i)^{-1} \), then the final effect size \( B \) is computed as the weighted mean of all the beta coefficients.

\[
B = \frac{\sum_i (\Sigma^i)^{-1} \beta^i}{\sum_i (\Sigma^i)^{-1}}
\]

The variance of \( B \) is:

\[
Var(B) = \sum_i (\Sigma^i)^{-1}
\]

The resulting beta divided by its standard error follows a multivariate normal distribution.

\[
S = \frac{B}{SE(B)} \sim \mathbb{N}_p(0, I)
\]

\[
S = \left[ \sum_i (\Sigma^i)^{-1} \right]^{1/2} \sum_i (\Sigma^i)^{-1} \beta^i
\]
Finally, significance in the multivariate association is estimated from the chi-squared statistics with $p$ degrees of freedom.

$$SS^t \sim \chi^2_p$$

5.5 Results

The method was implemented as part of the R package MultiMeta, together with plotting functions useful for visual representation of the results, including Manhattan plot, quantile-quantile plot and an overview plot of effect sizes for a chosen SNP. The default options for the meta-analysis function multi_meta are set to work with GEMMA file format (multivariate analysis option). The plotting functions work with output files from the multi_meta function by default. However both can be easily adapted to deal with different file formats by changing options, such as field separators, or by changing column names, as specified in the manual. Furthermore the plotting functions can be run by passing files as well as objects in input. This choice is meant to provide more flexibility and avoid unnecessary opening of large files. Example datasets with only few SNPs are included in the package and can be analysed as detailed in the instruction manual. Figure 1 shows an example summary plot obtained with the betas_plot function for a SNP tested on six traits. This is a custom plot, meant to show combined beta coefficients with 95% confidence intervals, as well as correlations between them.

Figure 1. Examples of effect sizes summary plot

This plot provides a visually simple representation of results for a chosen SNP of interest: the left hand panel displays effect sizes with 95% confidence intervals for each trait; the right panel shows a heat-map representing correlations between effect sizes. A) An example of default options in scales of grey; B) For a small number of original cohorts, it is possible to plot effect sizes for each one (light blue and pink on the left panel). It is also possible to choose colours for the heat-map (here: blue, white and pink) and specify the trait names.
The meta-analysis runs on very low memory with default options (e.g. RAM < 250 MB for two cohorts). Computation time is ~0.07 sec/SNP for two cohorts and it grows linearly with the addition of other cohorts in input. As each SNP is analysed singularly, overall computation time depends linearly on the total number of SNPs. By changing settings to increase the dimension of regions in which the genome is divided (see manual), it is possible to increase the performance, while allocating more memory. The package is freely available on CRAN repository and can be run on any operating system.

Finally, in order to test our software, we analysed UK10K Project data (2 cohorts, total N ~ 3000 samples, http://www.uk10k.org/) on six lipid traits: apolipoprotein A, apolipoprotein B, High-Density Lipids, Low-Density Lipids, Triglycerides and Total Cholesterol. Despite limited sample size, 6 previously reported associations were confirmed with highly significant p-value (lowest
Interestingly, some of these loci were not detected with univariate GWAS meta-analysis on the same sample. Furthermore, two possibly new loci reached genome-wide significance on chromosome 6, highlighting the increase in power and the strong potential for new discovery of the multivariate approach.

5.6 Conclusions
Combining results from different cohorts is particularly important for GWAS, where large sample sizes are required to reliably detect alleles with small effects. The R package MultiMeta provides a flexible approach to meta-analysing multivariate GWAS and easily visualizing results. As regards hearing function, single-trait GWAS meta-analyses on hearing thresholds already identified and replicated a number of candidate genes for both NHF and ARHL (Girotto et al. 2011 and previous Chapters). Thanks to the availability of efficient multivariate GWAS and meta-analysis described above, in future projects we will analyse the six hearing thresholds simultaneously. Moreover, the method allows combining different traits that are known or hypothesized to interact, which represents a new direction for association studies. In our case we plan to explore association between auditory function and Body Mass Index (BMI), Blood Pressure (BP), Glycaemia and possibly other traits. Findings could lead towards a better understanding of the molecular mechanisms underlying hearing function and of the interactions between the auditory system and other metabolic processes.
Appendix - maps of genetic isolates
REFERENCES


Li Y. et al. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34: 816-834.


Walkinshaw D.R. et al. (2013) The tumor suppressor kinase LKB1 activates SIK2 and SIK3 to stimulate nuclear export of class IIa histone deacetylases. Journal of Biological Chemistry.

Wang S.J. et al. (2009) Inactivation of fibroblast growth factor receptor signaling in myelinating glial cells results in significant loss of adult spiral ganglion neurons accompanied by age-related hearing impairment. Journal of Neuroscience Research 87: 3428-3437.


