SACKLER FACULTY OF MEDICINE

DEPARTMENT OF HUMAN MOLECULAR GENETICS AND BIOCHEMISTRY

Understanding of Developmental and Physiological Conditions of the Inner Ear using Transcriptome and Proteome Analysis

THESIS SUBMITTED FOR THE DEGREE "DOCTOR OF PHILOSOPHY" BY

KOBI PERL

SUBMITTED TO THE SENATE OF TEL AVIV UNIVERSITY

DECEMBER 2017

The work was carried out under the supervision of

Prof. Karen B. Avraham and Prof. Ron Shamir

ACKNOWLEDGMENTS

I would like to take this opportunity to thank both of my mentors, who have helped me to become the scientist that I am, contributed substantially to my training, provided me with important tools and clever comments, sharpened my writing skills, went over my writing over and over again, and overall, walked with me every step of the way.

Ron, your meticulous work and strong working ethics motivated me greatly. You were always available for bouncing off ideas, very open to my suggestions, and you responded to all my emails in a timely manner, which is exceptional for a PI. Your input had a major influence on my work, as well as the feedback from the many experts with whom I met on your initiative.

Karen, without your enthusiasm I would never have made it through. There were times when we sat down to edit my paper, and only got up by dark, once all figures were in place. You emphasized the importance of making my work comprehensible outside the bioinformatics community, inspired me with confidence, and helped me with my presentations.

To all current and former lab members in both labs, thank you for making this period unforgettable. Your constructive criticism had an impact on my work, and I was influenced by many of your amazing projects. I wish you all the best of luck.

I am grateful to my collaborators and teachers. Among them, Prof. Tami Geiger, who helped me take my first steps in the world of proteomics. I would like to express my sincere gratitude to the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and I-CORE Centers of Excellence, which supported my work.

Finally, I owe thanks to my family. To my mother and my father, Tali and Avi, and my siblings, Guy, Shani, and Itay – I could not have made it without you. You pushed me to aim high in my studies and in my research. You were my sunshine on a cloudy day.

I have never tried that before, so I think I should definitely be able to do that.

Astrid Lindgren, Pippi Longstocking

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

List of Abbreviations

BS: Brier Score

CPM: Counts Per Million mapped reads

DE: Differentially Expressed

DG: Deafness gene

E: Embryonic day

FACS: Fluorescence-activated cell sorting

FDR: False Discovery Rate

FC: Fold Change

GO: Gene Ontology

HC: Hair cell

HL: Hearing loss

IE: Inner ear

MA: Major Axis

MS: Mass Spectrometry

MDS: Multi-Dimensional Scaling

mRNA: messenger RNA

LCL: Lymphoblastoid Cell Line

MMT: Multiple Mouse Tissues

OLS: Ordinary Least Square

P: Post-natal day

PTR: Protein-transcript ratio

PU: Positive-Unlabeled

(R)FCB: (Relaxed) FC Based

RPKM: Reads Per Kilobase per Million mapped reads

(R)MSE: (Root) Mean Square Error

SILAC: stable isotope labeling with amino acids in cell culture

SC: Supporting cell

TF: Transcription factor

(W)AP: (Weighted) Average Protein

(W)APTR: (Weighted) Average Protein-Transcript Ratio

SUMMARY

This work aimed at analyzing the transcriptomic and proteomic repertoire of the auditory and vestibular systems to define the mechanisms of deafness and balance disorders. For this purpose, mRNA and protein was produced for the cochlea and the vestibule in mice of ages embryonic day (E)16.5 and post-natal day (P)0.

The inner ear is composed of two major systems: the hearing or auditory system, and the balance or vestibular system. While these systems have extensive similarities, there are structural and functional differences. The mouse has long been a model for studying human inner ear structure and function, due in part to the ability to breed and select offspring with desired traits, including those affecting hearing and balance. We took an interest in the cochlear and vestibular tissues in mice of ages E16.5 and P0, as they correspond to ages before and during the acquisition of mechanosensitivity. To date, only limited work has been done to compare the transcriptome of the two tissues before and after this developmental stage. We applied systematic transcriptomic approaches to decipher the regulatory pathways of the auditory system, with the primary goal of identifying transcription factors that serve as key regulators of proliferation and differentiation

Most research articles comparing expression levels do so for a single omics technique, most commonly RNA-seq for transcriptomics and protein mass spectrometry for proteomics. By analyzing a single omics type, one reduces the ability to identify post-transcriptional regulation mechanisms. Integrated analyses show that the correlation between expression levels of protein and mRNA in mammals is relatively modest, with a Pearson correlation coefficient of ~0.40. Suggested explanations for this low correlation include post-transcriptional regulation and measurement noise. We obtained mRNA and protein expression levels for the inner ear tissues at P0, and used them, along with other datasets of RNA and protein, to identify a pattern of post-transcriptional regulation that exists in non-proliferating tissues. A subsequent

analysis, comparing enrichments in the protein and mRNA domains, offered a possible biological advantage for this mechanism.

Exploring the transcriptomics in the dimensions of age and tissue expanded our knowledge about the development of the inner ear. We found the cochlea to be more enriched in neurological functions, and to contain a higher percentage of hair cells than the vestibule, but also to have a delayed development of its sensory perception compared with the vestibule. The vestibule, on the other hand, was found to be more vascular and more accessible to the immunological system. The majority of transcription factors that we predicted to be key regulators of the differentiation process have known functions that agree with this characterization. Some of these were further suggested as possible candidates in inducing hair cell regeneration.

Focusing on known deafness genes, we found that they tend to be differentially expressed between the tissues. During development, they increase both in expression, and in cochlea-to-vestibule expression ratio. We showed how this can be leveraged to build a classifier to identify candidate genes for deafness.

A joint analysis of the mRNA and protein data for P0 was used to demonstrate that the protein-to-mRNA ratio in steady state varies in a direction that lessens the change in protein levels as a result of changes in the transcript abundance. This trend was also shown in two other datasets, one of mouse organ tissues, and another of lymphoblastoid primate samples. A fourth dataset, of human cancer cell lines, failed to show this trend.

We suggest that partial buffering between transcription and translation ensures that proteins can be made rapidly in response to a stimulus, and we show that accounting for the buffering can improve the prediction of protein levels from mRNA levels.

1 INTRODUCTION

1.1 Inner ear

The inner ear (IE) is composed of two major systems: the hearing or auditory system, and the balance or vestibular system. While these systems have extensive similarities, there are structural and functional differences. In the auditory system, the organ of Corti in the cochlea contains the sensory epithelium responsible for hearing. The vestibular system contains five organs, including the three semicircular canals with cristae sensory epithelium that detect angular acceleration by fluid motion and the saccule and the utricle, which contain the macula sensory epithelium that detects linear acceleration due to gravity. The development of the IE requires a complex dynamic process to produce the final sensory organ with both hearing and balance capabilities [1].

1.1.1 Use of mice as models in inner ear studies

The mouse has long been a model for studying human IE structure and function, due in part to the ability to breed and select offspring with desired traits, including those affecting hearing and balance [2]. More recently, the similarities between the genomes, and the ability to manipulate the mouse phenotype by gene-targeted mutagenesis and genome editing, have reaffirmed the mouse as an ideal vehicle for studying human auditory and vestibular dysfunction [3, 4]. As a result, mouse inner ear development has been studied in detail on a molecular level [5, 6].

1.1.2 Gene expression profiling in specific inner ear tissues and cell types

Isolation of biological material from specific tissues and cell populations in the IE is complicated by the paucity of tissue. This, together with the great variety of cell types found

in the IE, makes it difficult to understand the complexity of gene expression within it [7]. In 2005, laser capture microdissection was used for the first time in the analysis of sub-compartments in the IE [8] . This method has led to the identification of hundreds of genes that are uniquely expressed in either hair cells (HCs) or supporting cells (SCs), in mouse, rat, and zebrafish [7]. A finer separation can be achieved using fluorescence-activated cell sorting (FACS). Only a few experiments have employed this expensive method in the IE [7]. The most comprehensive gene profiling study covering IE development in sorted cells used the expression of the transcription factor (TF) *Pou4f3* to select for HCs [9]. This process was performed in mouse cochlea and utricle in the mouse at ages embryonic day (E)16, post-natal days (P)0, P4, and P7. Afterwards, RNA-seq was used to build transcription profiles for HCs and SCs; i.e., all other cells. One important outcome of this research is the demonstration of changes in biological processes with age in each cell population, as they manifest in the expressed genes. One of the major limitations of this study is the lack of repeats for all ages except one.

As mRNA levels do not necessarily reflect protein expression levels, in part due to post-transcriptional regulation; and in light of the broad utility of proteomics for identification of biomarkers and pathophysiological mechanisms in multiple diseases, some proteomic studies were performed in the IE [10]. Still, proteomics platforms are underused in otology because of technical challenges and complex features of auditory and vestibular morbidities. Most large-scale IE proteomics studies were conducted using a multidimensional separation technique, such as two-dimensional difference gel electrophoresis or liquid chromatography, coupled with mass spectrometry (MS) [10]. In some, common add-on labeling techniques such as stable isotope labeling of amino acids in cell culture (SILAC) helped to achieve more accurate quantification. Current proteomic research focus on both profiling the normal inner ear proteome and characterizing protein changes in disease states. Animal models used

include guinea pigs, mice, chicken, zebrafish and chinchilla rodents, as well as bovine models [10]. Similar to the techniques used in the mRNA field to separate cell populations, both microdissection and FACS sorting are used in proteomic studies of the organ of Corti. In [11], for example, microdissection of cochlear and vestibular sensory epithelia was performed in order to attain protein, mRNA and microRNA profiles of the two tissues. In [12], FACS sorting was exploited for the comparison of HCs and non-HCs in the vestibular system.

1.1.3 Gaining mechanosensitivity

We took an interest in the mouse ages E16.5 and P0, as they correspond to ages before and during the acquisition of mechanosensitivity [9]. More precisely, while the vestibule is known to acquire mechanosensitivity between E16 and E17 [13], the cochlea's outer HCs became functional between P0 and P2 [14]. In terms of structural development, both E16.5 and P0 are ages after the formation of the cochlear organ of Corti (at E14.5 [15]) and the vestibular semi-circular and utriculo-saccular canals (E12 and E15), and the octonia (E16) [16]. Subsequent to the formation of the organ of Corti, a morphological differentiation takes place, with the opening of the tunnel of Corti composed of one row of inner HCs, three rows of outer HCs and supporting cells (SCs) [15]. In comparison, the differentiation of HCs to type I and type II in the vestibule occurs later, between E16 and E18 [16]. Between E15.5 and E17.5, the IE grows and extend, forming one and three-quarter turns of the cochlea and the semicircular canals [17]. At P0, the cellular patterning of the cochlear duct is essentially complete [6]. Both organs are morphologically well developed, but continue to mature. In the cochlea, the period of onset and maturation of acoustically evoked signal processing is between P12 and P14 [18]. In the vestibule, the type I cells are only partly surrounded by calyces until birth. The first calyces with adult type appear at P4, and the innervation is comparable to the adult at P10 [16].

During early development, transcriptional pathways have been elucidated that govern the differentiation of the otocyst towards sensory or nonsensory regions (reviewed in [17]). A number of temporal and spatial triggers of development and maturation have been characterized, including the molecular controls on the patterning, hair bundle heights and numbers of stereocilia. Knowledge about transcriptional pathways has laid the groundwork for establishing early and late developmental pathways of the IE. Mutations in some of these critical developmental genes lead to mouse [19] and human IE defects and deafness [20], although in most cases null mutations of these critical genes would lead to lethality due to their crucial role in early development in other organs.

1.1.4 Gene therapy for deafness

WHO estimates 466 million people worldwide have a disabling hearing loss (HL), and 34 million of these are children (http://www.who.int/deafness/estimates/en/; updated 06/23/18). Despite its widespread effects, the medical disability currently has no cure. Nonetheless, gene therapy is an emerging treatment, designed to tackle the root causes of this morbidity. Gene therapy can be employed to either fix a genetic problem of improperly functioning HCs and/or to promote proliferation of SCs in the cochlea, and their transdifferentiation into HCs [21]. In the last decade dozens of new deafness genes have been discovered [22]. In parallel, there were advances in the field of reprogramming and regeneration of HCs, including a clinical study, in which ATOH1's potential in causing transdifferentiation in SCs is being used to improve hearing function (CGF166). In order to extend the applicability of gene therapy to problems of HL and balance, the scientific community is focusing its efforts on both identifying new mutations underlying these conditions, as well as discovering other factors that can be manipulated in a coordinated manner, in order to improve the efficacy of HC regeneration in vivo. These topics will be discussed in detail in the next two sections.

1.1.5 Hair cell regeneration

Regeneration after cellular damage shares some similarities with normal organ development. In birds, regeneration of HCs involves proliferation of nearby epithelial supporting cells, which then differentiate to form replacement HCs and SCs [23, 24]. However, while mature mammalian vestibular organs are also able to regenerate at least a subpopulation of HCs after damage [18, 25, 26], the adult cochlea is incapable of any regeneration. It should be noted that there is some evidence that the cochlea may contain supporting cells with the ability to form new HCs in very young animals [27] or upon misexpression of *Atoh1* [28]. Given the limitations in the mammalian systems, the resemblance of the auditory sensory epithelia and cochlea between birds and mammals [5], and the ability of birds to regenerate HCs in the cochlea and vestibule, it is relevant to compare the gene expression profiles of the mammalian and avian inner ears. To this end, we applied systemic transcriptomic approaches to decipher the regulatory pathways of the auditory system and to make relevant comparisons to the avian transcriptome.

Sensorineural HL most commonly results from degeneration of cochlear HCs. As mentioned, if these are lost through damage or the natural aging process, they are not replaced. Gene therapy could potentially be used to induce HC regeneration [21]. For many tissues, reprogramming and regeneration is achieved by coordinated manipulation of multiple factors. Initial evidence shows this approach might be successful in the cochlea. In embryonic and neonatal mouse cochlear tissue, ectopic expression of *ETV4*, *TCF3*, *GATA3*, *MYCN*, or *ETS2* in combination with *ATOH1* yielded more HC-like cells than did overexpression of *ATOH1* alone [29, 30]. Another promising method for inducing cochlear cell regeneration, included a temporal modification of the expression of the retinoblastoma-1 (*Rb1*) gene in mice [31]; however, the response to *Rb1* inactivation was shown to dependent on the differentiation stages of HCs, with mature post-natal HCs re-enter cell-cycle but rapidly die afterwards [32].

The efficacy of these interventions is partial, rendering the search for other TFs that can be manipulated to enhance this process extremely relevant. As the number of TFs in human is estimated to be in the range of a few thousands [33], one cannot perform an exhaustive experimental search on all possible manipulations of TFs and their combinations. Instead one should focus its efforts on TFs that are more likely to participate in tissue differentiation. In some of the aforementioned studies [29, 30], the manipulation was performed on TFs that have conserved binding sites near *ATOH1* on the *POU4F3* gene. Here, we suggest yet another method to identify these candidate TFs, which focus on the concordance between TFs involved in tissue identity in early stages of development, and those participating in avian HC regeneration.

A different problem holds for the mammalian vestibular system that do possess restorative capacity, but for which an external intervention, in the form of growth factors infusion or gene therapy, is needed to induce the renewal of HCs [34]. While regeneration using *Atoh1* gene transfer is a promising method, the discovery of new genes whose replacement may restore IE function, can improve the treatment of balance disorders.

1.1.6 Deafness gene discovery

About 50%–60% of HL cases have a genetic etiology [35]. Approximately 80% of genetic deafness is nonsyndromic, i.e. not associated with other clinical features. HL is a recognized feature of more than 400 syndromes, the most common of which are Usher syndrome, Pendred syndrome (PS), and Jervell and Lange-Nielson. Nonsyndromic HL is classified by the inheritance pattern, and relatively common clinical features have been noted for each inheritance pattern. For autosomal recessive HL, the most frequent causative genes in order of frequency are *GJB2*, *SLC26A4*, *MYO15A*, *OTOF*, *CDH23*, and *TMC1*. Autosomal dominant

common mutations include *WFS1*, *MYO7A*, and *COCH*. Several of these genes are also implicated in syndromic HL.

Other classifications of HL are based on the impaired structure and phenotypic features such as time of onset (prelingual or postlingual), severity, and the affected frequencies [36]. Abnormalities of the external ear and/or the ossicles of the middle ear causes conductive HL, malfunction of inner ear structures leads to sensorineural HL, mixed HL is a combination of the aforementioned HL types, and damage or dysfunction at the level of the eighth cranial nerve results in central auditory dysfunction.

The search of disease genes is performed mainly by studying familial segregation [22]. Examples for mutations identified in families of patients include those found in the genes *MYO15A*, *TMC1*, and *COCH* [35]. Using the technique of whole-genome linkage analysis, or the less general method of homozygosity mapping, critical chromosomal intervals are mapped [22]. These approaches typically identify large chromosomal regions that include hundreds of genes. Before the availability of the next-generation sequencing technique, candidate genes were selected after positional cloning. Since 2010, techniques such as massively parallel sequencing and exome sequencing were used in conjunction with these loci-mapping approaches, removing the need for positional cloning. Such next-generation sequencing powered methods can be used to some degree in small families without the need for linkage analysis. However, the analysis of the data obtained, in particular through whole exome sequencing, is a complicated process, and filters must be applied to prioritize candidate variants [37].

Importantly, there are dozens of loci published in peer-reviewed journals without a causative gene assigned, most of which are from before the era of next-generation sequencing [22]. The mutation analysis of all genes encoded by a large genomic interval is extremely labor-

intensive. In [38], a bioinformatic approach was used to reduce the number of candidate genes in regions associated with nonsyndromic HL. The filtering criteria were based on evidence of cochlear expression, in human or in the orthologous gene in mouse, and a known interaction of the gene's product with a gene involved in IE development or function or, alternatively, with a candidate gene from a different locus. A list of 2378 genes mapping to various genomic intervals were narrowed down to 92 genes as candidates. Unfortunately, the authors did not provide any measure for estimating the accuracy of their prediction.

As outlined above, prioritization of candidate genes for deafness is important when multiple candidates arise from a familial segregation study, even in the more recent works that use whole exome sequencing. Here, we present a machine learning method that offers such a prioritization.

1.2 mRNA versus protein levels

1.2.1 Steady-state versus non-steady-state (perturbed) systems

The correlation between expression levels of protein and mRNA in mammals is relatively low, with a Pearson correlation coefficient of ~0.40 [39, 40]. Suggested explanations for this low correlation include post-transcriptional regulation and measurement noise [39]. This low correlation makes it difficult to integrate mRNA and protein data. Tools for this integration are sparse and not yet adopted by the bioinformatics community (reviewed in [41]). Initial findings from such tools suggest that the transcriptional and the translational regulation evolved independently, except in the rare occasions where strong selection in favor of correlation was present [42]. However, such claims are based on data from perturbed systems, where the observed discordance between the transcriptome and the proteome is strongly affected by the lack of temporal synchronization between the transcriptional and

translational regulation levels [43]. In this study we focus on the connection of mRNA and protein levels in non-proliferating tissues.

Virtually all data in "omics" experiments is obtained from systems that can be either described as perturbed (i.e., subjected to a stimulus) or that are said to be in a "steady-state" [44]. It is challenging to rigorously define this latter term, as it is often used for cells undergoing long-term dynamic processes such as continuous proliferation, differentiation, or other types of fate decisions. Molecule concentrations in an individual cell may change substantially. However, as the average concentrations are measured across a population, they remain roughly constant with time, namely, in a "steady-state". Among such systems, the one of non-proliferating cells is especially simple, as the rates of synthesis and degradation of molecules in the tissue are independent of the cell cycle length, as opposed to a system of dividing cells in log phase [45].

1.2.2 Determinants of protein abundance

Protein abundance reflects a dynamic balance among multiple processes, spanning the transcription, processing and degradation of mRNAs to the translation, localization, modification and programmed destruction of the proteins themselves [39]. A large effort was made in order to decipher the relative contributions of these processes to the variation in protein levels. Schwanhäusser et al. [45] determined that about 40% of the variance of protein levels between different proteins could be explained by mRNA levels. A follow-up study re-analyzing the same dataset with a different statistical model concluded that about 56%–84% of the protein variance could be explained by mRNA variance, while the translation rate could only explain 9% of the protein abundance variability [46]. The buffering effect presented later in this thesis supports some coupling between transcriptional and post-transcriptional regulation mechanisms, and challenges the simplistic models used in

the aforementioned studies, in which the contributions of different levels of regulation are independent, and thus sum up to 100%.

1.2.3 Conservation of protein abundances

We will refer to a gene's protein level divided by its transcript level as the gene's protein-transcript ratio or PTR, also called the gene's translation efficiency [47]. We note that this measure is affected by both translation and protein degradation rates, and under steady-state conditions it should be equal to the ratio of the rates [48]. It was observed that across taxa, protein levels are more conserved than mRNA levels [49], although some exceptions exist [50]. Also, it was noticed that differences in protein levels between primates are less common than differences in mRNA levels [51]. While PTR was claimed to be highly conserved between tissues for each given protein [52], it was demonstrated that it somewhat varies between tissues in a direction that buffers or compensates for the change in protein levels from changes in the transcript abundance [48], similar to what was shown across taxa. However, these observations originated from a small number of tissues, and were based mainly on regression coefficients that are affected by regression dilution bias [53].

1.2.4 Protein level prediction

Many experiments only measure transcript abundance in a tissue and use it as a proxy for protein levels. Previous articles that predicted protein levels from mRNA [47, 54] did not use PTR measured in other tissues, and relied mainly on sequence related features; they reached a correlation of 0.75 between the predicted and the observed levels. It has been suggested to use the average PTRs measured in other tissues in order to predict the protein levels for the tissue in question [49]. This assumes the PTR of a gene is constant across tissues. We suggest, instead, a model that assigns a higher PTR in a tissue where the mRNA level is lower.

1.3 Research aims

This project is aimed at analyzing the transcriptomic and proteomic repertoire of the auditory and vestibular systems in order to define the mechanisms of deafness and balance disorders. The aims of the research were as follows:

1. Separate analysis of RNA-seq and MS-based proteomics of cellular components of the inner ear.
2. Integrated analysis of transcriptomics and proteomics.
3. Analysis of transcriptomics prior to and during the acquisition of mechanosensitivity.

Aims 1 and 2 of the research plan were fulfilled to completion. In fact, we achieved much more than what we committed to in the goals, as many of the conclusions made were examined in a broader context, using multiple datasets of RNA and protein

In aim 3 we originally declared that a comparison of the proteomics will be performed as well. However, we were not satisfied with the quality of the proteomic samples, and focused only on the transcriptomics. This change of aims was approved by the PhD committee. Using transcriptomics data alone, we still derived strong biological conclusions regarding the development of the cochlea and the vestibule in the examined period.

We note that the author of this thesis performed the bioinformatics analysis of the data, and did not conduct the biological experiments described below.

Note also that some textual segments of the thesis are taken verbatim from [55] and a manuscript in preparation (Perl K, Shamir R, Avraham KB. mRNA expression profiling in the inner ear reveals candidate transcription factors associated with proliferation and differentiation. 2017;). The original text was written by the author and revised by the supervisors for the paper, and is repeated and often expanded here.

UPDATE (20/06/18): The manuscript in preparation listed above was published after the original submission of the thesis but before its final approval [56] .

2 MATERIALS AND METHODS

2.1 Inner ear mRNA data generation

Cochlear and vestibular sensory epithelia were dissected from 20 inner ears of 10 P0 C57Bl/6J mice, generating 2.4 and 1.5 μg of total RNA, respectively. Tissue dissection of E16.5 mice followed a similar process, and resulted in 2.5 and 1 μg of total RNA, respectively. 450 ng RNA from each sample was used to create libraries with the TruSeq Stranded mRNA Sample Prep Kit (Illumina), followed by high-throughput sequencing at 100 bp paired end (PE) at the Technion Genome Center, Haifa, Israel. Six samples were generated for each developmental age, 3 cochlear and 3 vestibular, for sequencing in triplicate. Read quality was assessed using ShortRead and reads were aligned using tophat2 against a mouse reference genome (Mus_musculus.GRCm38.74). BAM files were manipulated using Samtools and per-gene counts of the reads were computed using htseq-count. edgeR was used for calculating DE, fold changes and RPKM normalized values. Only genes that have one read per million in three or more of the samples were included in the analysis. See [57] for references to each software tool. The mRNA unit of measurement is RPKM (Reads Per Kilobase per Million mapped reads) [58]. For DE analysis using edgeR [59], the read counts were used. The mRNA data for P0 and E16.5 were deposited to the Gene Expression Omnibus (GEO) repository under accession number GSE76149 and GSE97270, respectively. RNA data was also deposited in the gEAR portal (http://umgear.org/).

2.2 Inner ear proteomics data generation

Cochlear and vestibular sensory epithelia were dissected from 15 P0 C57Bl/6J mice, with samples from each set of 5 mice pooled to generate one of 3 replicates of protein from cochlear of vestibular tissues. Protein samples were reduced with DTT and alkylated with

iodoacetamide followed by in-solution digestion with trypsin. Peptides from two replicates were analyzed by single LC-MS runs and one replicate was further separated into six fractions, each analyzed by LC-MS on the EASY-nLC1000 UHPLC coupled to the Q-Exactive MS. Raw MS files were analyzed with MaxQuant and the Andromeda search engine. The label-free algorithm was used for protein quantification with a minimum two ratio counts for normalization. The database search was performed against the Mouse Uniprot database (2013) with 50,807 entries and a list of common contaminants. False discovery rate (FDR) was determined using the forward-reverse approach, and set to 1% FDR on the peptide and protein levels. Database search parameters included Trypsin/P as the proteolytic enzyme, N-terminal acetylation and methionine oxidation as variable modifications, and carbmidomethyl cysteine as a fixed modification. A maximum of two miscleavages and a maximum peptide charge of +7 were allowed. First database search was used for mass recalibration with an error tolerance of 20 ppm followed by the main Andromeda search with mass tolerance of 4.5 ppm for MS spectra and 20 ppm for the MS/MS spectra. Peptide length was set to a minimum of seven amino acids. Analysis of the raw MS data identified 7244 proteins, with correlations of 0.9 and 0.95 between biological replicates of cochlea and vestibule, respectively. The MS proteomics data have been deposited to the ProteomeXchange Consortium [60] via the PRIDE partner repository with the dataset identifier PXD003379. The protein unit of measurement is 2^{LFQ}/MW, where LFQ is a commonly used normalization for protein intensity [61], and MW is the molecular weight in kDa.

2.3 Transcriptomics analysis

2.3.1 Principal component analysis

Principal components were calculated with R, after scaling and centring the log2-transformed RPKM values, and plotted using ggbiplot (http://github.com/vqv/ggbiplot). To test the

association between the principal components and the samples' annotations of age and tissue, swamp (http://CRAN.R-project.org/package=swamp) was used.

2.3.2 Linear mixed models

The following model was used to describe the expression level $E_{g,s}$ of gene g in sample s, which originated from tissue t at age a:

$$E_{g,s} = \alpha_g + \beta_s + T_{g,t} + A_{g,a} + I_{g,t,a}$$

Equation 2.3-1 Expression levels decomposition to tissue and age variance components.

The expression levels are in logged RPKM (Reads per kilo base per million mapped reads). Random variables $T_{g,t}$ correspond to the effect of tissue identity on expression, $A_{g,a}$ correspond to the effect of age on expression, and $I_{g,t,a}$ correspond to a combined effect of tissue and age on expression. The parameters α_g correspond to the base expression levels of genes, and β_s correspond to the normalizing constants of expression between replicates.

Using lmer [62], we fitted this model to our data, and estimated the percentage of variance explained by each variance component. The high number of measurements did not allow fitting the model for all genes at once. Instead, we randomly selected 760 genes (5% of all genes) and fitted the model using their expression data. We performed this process 100 times. In 10 of those times, we used restricted likelihood ratio test to test whether the variance of the random effect $I_{g,t,a}$ is zero [63]. We reported the median p-value in the text.

2.3.3 Differential expression

Differential expression analysis was done using edgeR [59]. The design formula included the combination of age and tissue of each sample. The tested contrasts were the average difference between the two ages across tissues, the average difference between the two tissues across ages, and the difference of the differences at both ages. This last contrast is

sometimes referred to as the interaction term of tissue and age. edgeR detection threshold was q−value≤0.05. FDR correction was applied for each contrast separately.

2.3.4 GO and KEGG enrichment analysis

We performed the enrichment analysis using the Expander software (27), exploring all GO ontologies, 'biological process' (BP), 'molecular function' (MF) and 'cellular component' (CC) (*corrected p−value*≤0.05), and KEGG pathways (*q-value*≤0.01). For each contrast, we looked separately for enrichments in the set of genes up-regulated and down-regulated, using as a background set all the genes that passed the filter and were tested for differential expression.

2.3.5 Illustrating age-tissue interacting GO terms

We calculated the expression ratios between the cochlea and the vestibule for E16.5 and P0 separately, using edgeR. We then z-scored the ratios at each age, to allow a fair comparison of the ages. These ratios were used both to select which GO terms to display, and to calculate a median ratio for each of these terms.

To select GO terms, we began with the lists of terms enriched in genes with increased cochlear to vestibular (C/V) or vestibular to cochlear (V/C) ratios between E16.5 and P0 (see GO and KEGG enrichment analysis). From each of these lists separately, we filtered only the GO terms for which the expression ratios of annotated genes are higher at P0 than at E16.5 (one-sided Wilcoxon signed rank test at the respected direction, q−value≤0.05). FDR correction was applied for each list separately.

2.3.6 Identifying involved transcription factors

2.3.6.1 TF enrichment analysis

We performed the enrichment analysis using PRIMA [65] with detection threshold q−value≤0.1. FDR correction was applied for each list separately.

2.3.6.2 Connecting motifs to target genes

We used the same mapping as PRIMA [65] uses by default. It contains for every gene and for every TRANSFAC motif the number of hits (putative binding-sites) of the motif in the promoter of the gene (spanning from 1000 bp upstream the transcription start site [TSS] to 200 bp downstream the TSS). If there was at least one hit of a motif in the promoter of a gene, we considered the gene to be its target.

2.3.6.3 Integration with avian sensory epithelia regeneration experiment

In [66], TF expression was measured in consecutive time points after the infliction of damage (either laser or Neomycin [NEO]) to an IE tissue (either cochlea [CO] or utricle[UTR]). We adopted the authors' thresholds for declaring a TF as DE in a single time point (FC≥1.2 and p-value≤0.05, compared with the background). Then, we found overlaps between TFs that are DE in at least a single time point in [66], and those arising from the enrichment analysis.

2.3.7 Deafness genes expression patterns

One hundred and forty deafness genes (DGs) were manually curated from http://hereditaryhearingloss.org/ (updated for 3/13/17). Using BioMart [67], we mapped 133 of the genes to mouse orthologs. Three genes were filtered out due to missing expression data of their orthologs. To resolve multiple mapping, we preferably mapped to orthologs for which we have expression data. The DGs were annotated according to the type of deafness as syndromic, nonsyndromic or mitochondrial. Nine genes that were associated with both syndromic and nonsyndromic deafness were treated as if they were syndromic in subsequent analyses. For all five mitochondrial genes we found no homologs. In total, 34 homologs were classified as syndromic and 96 as nonsyndromic.

2.3.8 Classifying deafness genes by expression

We built a classifier in order to categorize each gene as DG or non-DG. We were aware that some of the genes currently categorized as non-DGs are in fact DGs not yet discovered. Our classifier thus learned to distinguish between positive and unlabeled genes. The features for the classifier were, for each gene: (1) the averaged expression over all samples, in log counts per million (CPM), (2) the logarithm of the fold-change (FC) of expression between the ages. (3) the logarithm of FC of expression between the tissues, and (4) the logarithm of the FC of the tissue expression ratio between the ages [i.e., log((cochlea to vestibule expression ratio at P0)/ (cochlea to vestibule expression ratio at E16.5))]. This last feature represents the interaction of age and tissue. All four features were computed using edgeR [59]; the FCs specifically were obtained under the model presented in section 2.3.3. We trained the classifier with 75% of the genes, leaving the other 25% for testing. Our classifier bagged over 1000 decision trees. Down-sampling was used to account for the imbalance in the frequencies of the deafness and non-DGs (130 and 15,076 genes, respectively). That is, to build each decision tree, we chose 130 non-DGs at random and used them together with all DGs in the building process. The R package caret was used for machine learning [68].

For the comparison of the classifier with a classifier using the averaged RPKM values in each condition as features, we used only 25 repeated training/test splits. For assigning genes with probabilities, we used 2000 repeated splits, although internal testing showed the ROC score reaches a plateau after about 150 iterations. In each iteration, we used the classifier to predict the probabilities in the test set, corrected these probabilities for the undersampling bias and corrected them again for the bias caused by the PU scenario. The correction methods are detailed below.

The correction of both biases did not affect the ranking of the genes in that iteration, and was performed in order to produce well-calibrated probabilities. We averaged the probabilities over all iterations. The averaging caused minor differences in ranking between different

methods of calibration, but the ROC score did not change significantly (p>0.05, DeLong's test for two correlated ROC curves [69]).We then assessed the calibration of the probabilities produced by each method. Under the assumption that most DGs are yet to be discovered, calibration curves that treat only known DGs as positive cases will falsely portray the probabilities as inflated. For this reason, we downloaded lists of genes that were associated with hearing loss according to the text mining tools. For the purpose of choosing the correction method that produced the most calibrated probabilities, we assumed that these deafness associated genes together with the known DGs comprise the full list of DGs. The annotation of deafness associated genes and the comparison of the calibration are detailed below. An illustration of the classification process is provided in Figure 2.3-1.

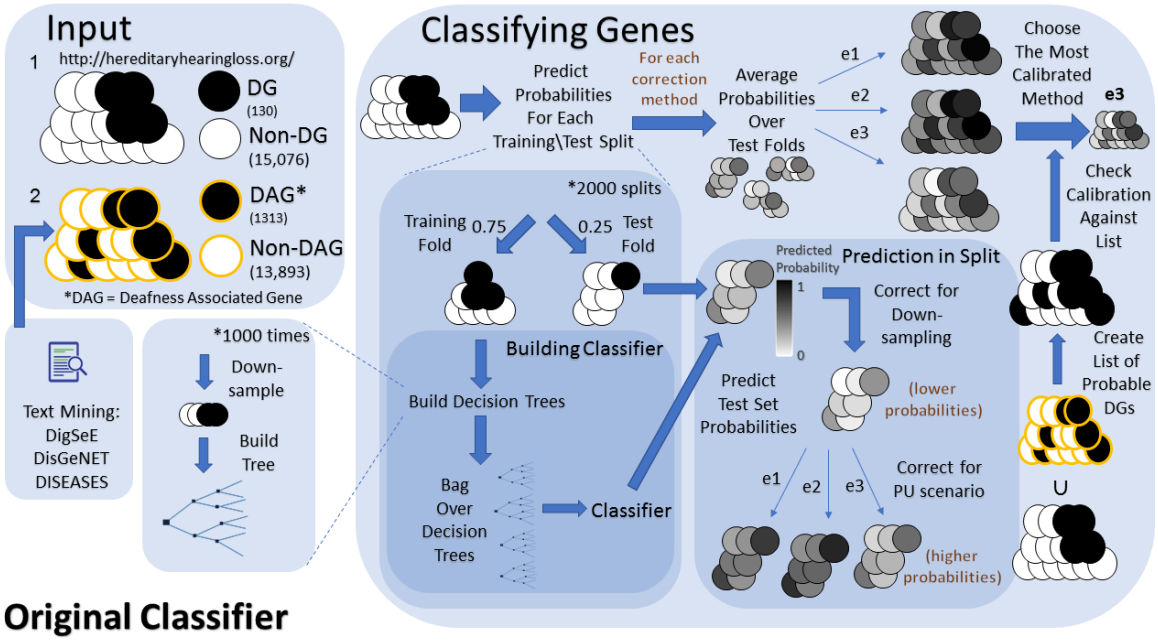

Figure 2.3-1 Illustration of the classification of genes as deafness genes. The input to this process is an assignment of genes to deafness genes and non-deafness genes, expression pattern data (not portrayed in the figure) and annotation of genes as deafness associated according to text mining tools. This last input type is only used in selecting the bias correction method. The output of the process is the predicted probability for each gene to be a DG.

We then used these probabilities to build an improved classifier. Let p_g be our estimation of the probability of gene *g*. We reran our bagging-like algorithm, but this time we chose to treat

a gene g as a positive example with probability p_g, and as a negative example with probability 1 - p_g. This reassignment was performed before each iteration, independently for each gene, and only for the unlabeled genes. Labeled genes were always treated as positive examples. This idea is inspired by [70], where the authors achieve slightly better results by rerunning their classifier with weights based on the initial probabilities learnt, also after fixing for the PU bias. Instead of reweighing the samples, we decided to reassign their classes, as reassignment (of only a few hundred genes) still allows us to perform undersampling. We again used 2000 repeated splits and averaged the probabilities over all iterations. We did not perform any bias correction until the end of the run, when we performed a correction only due to undersampling, as detailed below. We compared the ability to predict deafness associated genes among all unlabeled genes between the initial classifier and the "rerun" classifier using DeLong's test for two correlated ROC curves. We note that the probabilities assigned by the classifiers to known DGs are ignored in this comparison, because the annotation of these genes as positive in the training of the initial classifier can lead to an artificial inflation of the probabilities assigned to them by the "rerun" classifier. An illustration of the classification process improvement is provided in Figure 2.3-2.

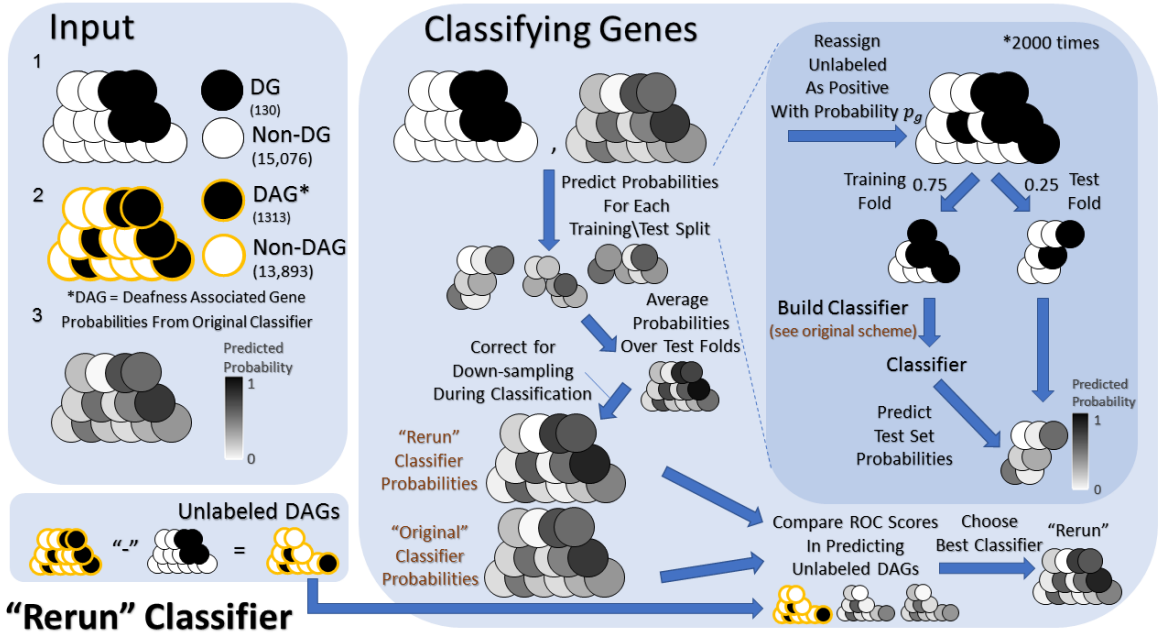

Figure 2.3-2 Illustration of the improved classification process of deafness genes. This "rerun" classifier uses the output probabilities of the original classifier (Figure 2.3-1) and refines them by randomly reassigning genes with unknown role in deafness as deafness genes during the learning process. The probability of a gene to be reassigned as a deafness gene is equal to the probability assigned to it by the original classifier.

Finally, we converted the mouse genes back to human genes. We resolved multiple mapping with averaging of the assigned probabilities.

2.3.8.1 Calibration of the estimator

The calibration of the probabilities was tested using calibration plots produced with the R package caret. The prediction space was discretized into 11 bins. Cases with predicted value between 0 and 0.09 fell in the first bin, between 0.09 and 0.18 in the second bin, etc. For each bin, the mean predicted value was plotted against the true fraction of positive cases, along with the 95% binomial confidence interval. If the model is well calibrated the points should fall near the diagonal line. We also used Brier score (*BS)* to measure probabilities calibration [71]. The lower the BS the more accurate are the probabilistic predictions of a model. Let $\hat{p}(y_i|x_i)$ be the probability estimate of sample x_i to have class $y_i \in \{0,1\}$. BS is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\{y_i - \hat{p}(y_i|x_i)\}^2$$

Equation 2.3-2 Brier score for calibration assessment of DG classifier.

2.3.8.2 Correcting undersampling bias

Undersampling creates an upward bias of the probabilities. To fix for this bias we used the transformation suggested in [72]. p_s is the probability assigned by the model learnt on the balanced training set. p' is the bias-corrected probability obtained from p_s:

$$p' = \frac{\beta p_s}{\beta p_s - p_s + 1}$$

Equation 2.3-3 Correction of undersampling bias in DG classifier.

Where β is the probability of selecting a negative instance with undersampling.

We used this method twice. First, we adapted our PU classifier. Since we know for each gene whether it is positive or unlabeled ("negative"), then the estimation of β is trivial. We set $\beta = \frac{N^+}{N^-}$, with $N^+ = 130$ and $N^- = 15{,}076$. Second, we adapted the "rerun" classifier, which used initial, well-calibrated, probabilities as input. The expected number of DG according to these input probabilities was $E(N^+) = 435$. We thus set $\beta \cong \frac{E(N^+)}{15{,}206 - E(N^+)}$.

2.3.8.3 Correcting positive-unlabeled bias

PU classifiers create a downward bias of the probabilities. Let x be an example and let $y \in \{0,1\}$ be a binary label. Let s = 1 if the example x is labeled, and let s = 0 if x is unlabeled. According to [70], $p(y = 1|x) = p(s = 1|x)/c$ where $c = p(s = 1|y = 1)$. Our PU classifier estimates $p(s = 1|x)$, the probability of the example to be labeled. In order to obtain an estimate for $p(y = 1|x)$, the positivity probability, we need to divide the first probability by an estimate of c. Three estimators were suggested for c:

$$e_1 = \frac{1}{n}\sum_{x \in P} g(x)$$

$$e_2 = \sum_{x \in P} g(x) \Big/ \sum_{x \in V} g(x)$$

$$e_3 = \max_{x \in V} g(x)$$

Equation 2.3-4 Correction methods for positive-unlabeled bias in DG classifier.

Where $g(x) = p(s = 1|x)$ is the posterior probability according to the PU classifier, V is the validation set, and P is the subset of examples in V that are labeled. We used the same set V for validation (estimating c) and for testing (estimating probabilities).

Methods e_1 and e_2 can give estimated probabilities higher than 1. For the calibration plots and calculation of BSs, we truncated them at 1 (1128 and 1897 probabilities exceeded 1 for e_1 and e_2, respectively).

We note that e_1 should theoretically have a lower variance than e_3, since the first is based on averaging over multiple samples instead of using just one [70]. However, we did not assume that e_1 is necessarily more accurate than e_3, especially as the number of positive samples in a validation set used for e_1 calculation is only 32 whereas the e_3 is the maximum of 3801 probabilities, and as such, might be more accurate. In practice, we used all three estimates and chose the one that produced the most calibrated probabilities, which was e_3.

2.3.8.4 Deafness associated genes annotation

We downloaded lists of genes that were associated with hearing loss according to the text mining tools DigSeE [73], DisGeNET [74] and DISEASES [75]. We searched the disease terms 'Hearing Loss' in DigSeE and DisGeNET and the 'Sensorineural Hearing Loss' in DISEASES. All tools returned lists of human genes. We converted them to mouse homologs using BioMart [67]. In DisGeNET and DISEASES an association has a score. In DiGSeE the association of gene

g is characterized by the number of articles $n_{g,a}$ and the number of sentences within articles $n_{g,s}$ supporting it. We assigned this association the score $n_{g,a} + \frac{n_{g,s}}{\max_{x \in G} n_{x,s}+1}$, i.e. the number of sentences served as a tie breaker between associations with the same number of articles. For the computation of ROC scores and BSs we treated association as a binary trait. For demonstrating the effect of choosing different thresholds, we compared the scoring of genes with a probability above the threshold with the scoring of the other genes using Wilcoxon signed-rank test. The score of a gene not associated with deafness was set to 0. In this analysis we also used a combined association score, which is the mean rank across the three lists of scores. We set the minimum "Combined" score to zero.

2.3.9 Deconvolution of heterogeneous tissue samples

Using RNA-seq expression data from [9], we created an expression signature for each combination of tissue (cochlea/utricle), type (GFP+/GFP-) and age (E16/P0). For the process of deconvolution of heterogeneous tissues data, limiting signatures to few hundred genes that best separate the reference cell types, result in good prediction accuracy [76, 77]. Therefore, we selected such subset of genes for each age separately, using the following heuristics. First, we ordered the genes in decreasing order of expression variance across the four reference samples. Then, we took the first *k* genes in the list, with *k* selected to minimize a specific error in the deconvolution on our mixed data (see details below). Once *k* was determined, we built the expression signatures, and used them to assess the proportion of cells, under two different scenarios. In one, the cells composing a tissue were confined to cells originating from that tissue, while in the other we allowed cells composing a tissue to originate also from the other tissue, mimicking a contamination of samples. The property we minimized in the selection of *k* was the estimated percentage of contamination in our mixed data under this second scenario, i.e., the estimated percentage of cochlear cells in vestibular samples plus the

estimated percentage of vestibular cells in cochlear samples. We tested all possible *k*'s in the range 1…1000. For E16.5 we chose the minimizing *k*=453. For P0, we ignored the first local minimum, which was narrow (~5 genes), and instead chose *k*=193 (Figure 2.3-3).

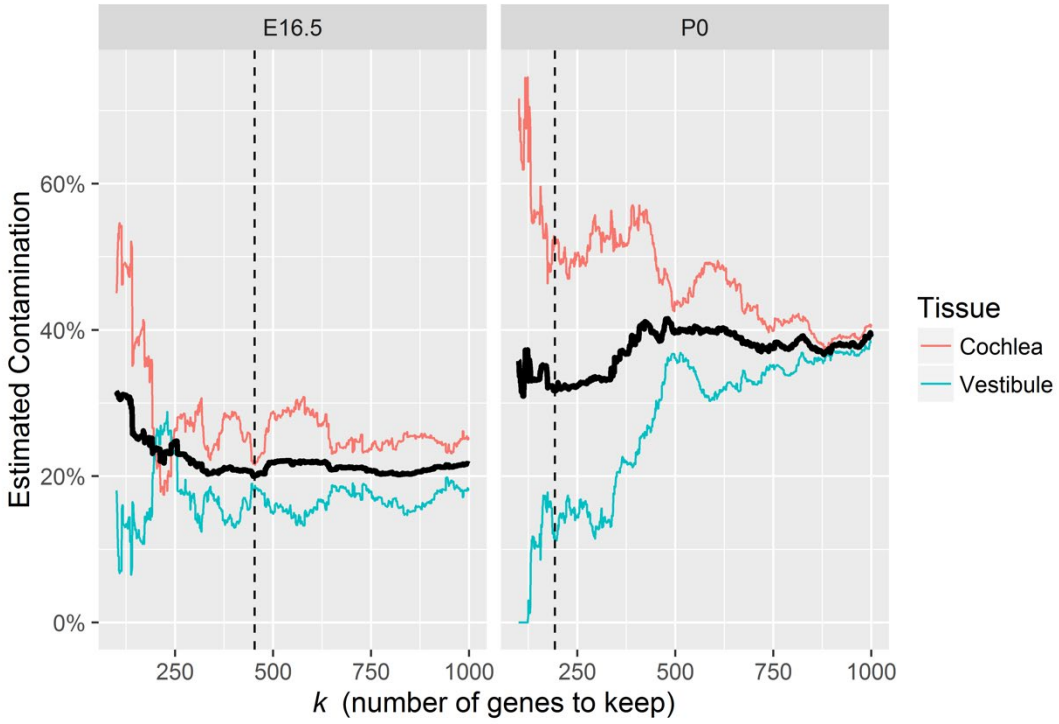

Figure 2.3-3 Choosing *k*: number of genes to keep for deconvolution. The estimated percentage of contamination is plotted against *k* for E16.5 and P0 (solid black line, left and right subfigures, respectively). The contaminations in cochlear and in vestibular samples are marked by the red and blue lines, respectively. The value of *k* chosen to minimize the contamination is marked by the dashed black line.

The expression data of the mixture was given in units of RPKM, and of the reference in counts per million (CPM). We did not normalize the reference data to the gene length, because the technique used in [78] of sequencing the 3' end, is not biased by gene length. Before building a signature, we filtered out genes for which the CPM was less than 1 in any of the conditions (within an age). The calculation of the variance in the expression of a gene was done on log-transformed expression.

We used DeconRNASeq for estimating the mixing proportions [78]. We used the default setting of the *R* package, except we chose not to scale the data. We performed the

deconvolution on the log-transformed expression. This is not generally recommended, specifically of microarray data, as it introduces a bias [79]. However, when we tried to work with the expression in the linear-scale without log-transformation, we got results that deviated extremely from what is known about the ratio of HCs to SCs in both ages. To be specific, the estimated percentage of HCs at E16.5 and P0 were ~12.5% and ~70% in both tissues. The gap is higher than expected, and also, the second estimate is much higher than parallel quantities in other species; in adult human from the age of 27 to 67, 46.5% of the cells of the crista ampullaris are HCs [80], and in posthatch chick 28.2% of the cells in the utricular macula are HCs [81]. Reference samples from E16 were used to estimate the proportions in our E16.5 samples. Also, reference samples originating from the utricle were used to estimate proportion in our whole-vestibule samples. The estimation was done for each sample separately, and afterwards we averaged the predictions across each group.

2.4 Protein and mRNA joint analysis

2.4.1 External protein and mRNA datasets

2.4.1.1 MMT RNA data preprocessing

Multiple Mouse Tissues (MMT) data were downloaded as fastq files from ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-GEOD-30352 and processed into read counts using the same protocol and reference genome as the EAR data. Out of 36,441 genes, only 16,969 genes that have one read per million in three or more of the samples were included in the analysis. We used samples for both wild mice and C57Bl/6J mice. There was clear separation of the samples by tissue and only poor separation by strain (figure not shown). Therefore, we chose to summarize tissue information from both strains to increase the statistical power.

2.4.1.2 MMT protein data preprocessing

Proteomic data was taken from [82]. For each tissue, the study provides two types of measurements, the MS intensity of the light version of the peptide, and the intensity ratio of heavy and light versions of the peptide. The choice of which quantity to use in each analysis is detailed in section 2.4.1.4.

Protein samples of three different brain regions were merged into a single summary sample by computing a weighted mean. This summary sample can be compared to the RNA brain samples that were produced from entire brain except olfactory bulb and cerebellum [83]. The weights used, based on the volume proportions of the regions in an adult C57BL/6J mouse brain [84], were 61.9%, 24.3%, and 13.8% for the cortex, medulla and midbrain respectively. The midbrain volume is computed from the sum of volumes of the superior and inferior colliculi, central gray, and the structure named "the rest of midbrain". Similarly, protein samples of two different kidney regions were merged into a single representing sample. The weights used here were volume proportions of the regions in a newborn Swiss Webster mouse [85] (58.5% and 41.5% for the cortex and medulla, respectively).

2.4.1.3 NCI60 RNA data retrieval

Transcriptomics (series accession GSE32474 [86, 87]) and proteomics [88] data were downloaded from: http://129.187.44.58:7070/NCI60/.

2.4.1.4 Units of measurement in external datasets

In the MMT and PRIMATE datasets proteins were quantified using the SILAC technique, which gives for each protein the ratio of expression between an individual sample to an internal standard (SILAC tissue). In both datasets, we also quantified the protein levels based of the absolute intensity of the peptides in the light version, which corresponds to peptides from the non-SILAC tissue. The absolute levels were used in the production of summary statistics, calculation of correlations, and prediction of protein levels, whereas the SILAC

ratios were used in MDS plotting, DE analysis, and testing whether PTRs vary in a direction that reduces protein divergence. The usage of SILAC ratios was preferred in the last scenarios as it yields a more accurate estimate of protein abundance between two proteomes [89].

MMT: The unit used for absolute protein levels is $Intensity.L/MW$, where Intensity.L is the sum of the measured intensities of the light version of the peptides composing the protein. The unit used for relative protein levels is Ratio.H.L.normalized, where Ratio.H.L.normalized is the ratio of the heavy to light intensities, after applying normalization as in [82]. A mix of SILAC mouse tissues served as an internal standard. The mRNA unit is RPKM. For DE analysis using edgeR, the read counts were used.

NCI60: The protein unit is $LFQ\ Intensity/MW$. The mRNA unit is the intensity level measured from the microarray chip, normalized as in [87].

PRIMATE: The unit used for absolute protein levels is iBAQ [45], based on the intensities of the light version of the peptides composing the protein. The unit used for absolute mRNA levels is RPKM. The unit used for relative protein levels is Ratio.H.L.normalized. A single human SILAC served as an internal standard. The unit used for relative mRNA levels is $RPKM_{sample}/RPKM_{standard}$, using the same reference cell line. The relative mRNA levels were used for the same purposes as the relative protein levels.

2.4.2 MDS plots

Multi-dimensional scaling was used to plot and visualize sample similarity. Plots were calculated using the function cmdscale in package stats (www.R.project.org). For the MMT dataset, the relative protein levels were used.

2.4.3 Measuring correlation between protein and mRNA levels

2.4.3.1 Avoiding biases in correlation measurements

mRNA and protein levels were $\log_2$-transformed, and averaged across all samples from the same group, disregarding missing values. Correlation was measured between pairs of groups, for mRNA and protein separately. Protein-mRNA correlations for each group were also calculated.

For each pair, correlation was measured between all genes that were expressed in at least one sample of protein, and one sample of mRNA, in both groups. Applying this filter is critical, as lowly expressed genes and proteins suffer from low detection rates, with a higher detection threshold in the protein domain. By applying this filter, we reduced the bias that is caused by the difference in the detection abilities.

Another bias that should be accounted for is the different levels of noise in the mRNA and protein domains. The mRNA replicates are more correlated with each other than the protein replicates in all relevant datasets (see Results). The higher noise in the protein domain would cause a larger reduction in the observed correlations compared to the real correlations (i.e. before the induction of noise) for protein. To account for this bias, we used Spearman's method to correct for attenuation of correlation, and obtained better estimators for protein-protein and mRNA-mRNA correlations. The effect of the correction is demonstrated in Figure 2.4-1 for the NCI60 dataset.

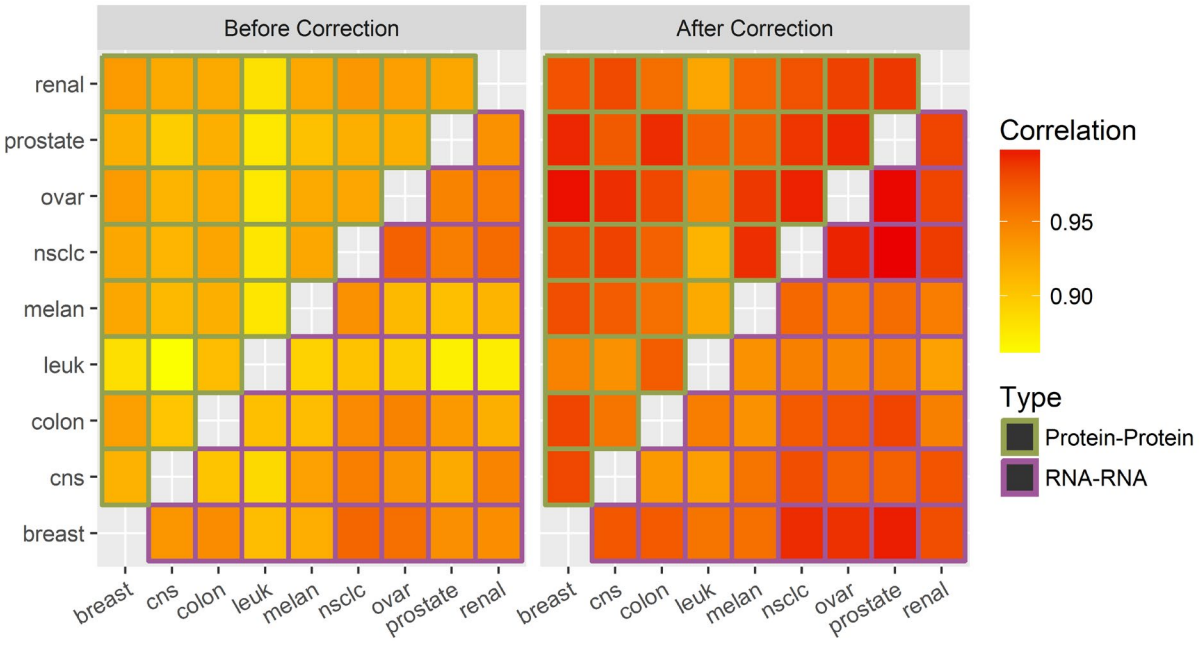

Figure 2.4-1 Corrected correlation plot for the NCI60 dataset. Left: Pearson's correlation (r) before correction. Right: After applying Spearman's method to correct for attenuation of protein-protein and mRNA-mRNA correlations. Note that a corrected correlation is not bound by 1. Before applying the correction only 8/36 pairs had higher correlations in the protein domain. After applying the correction the number increased to 24. See caption in Fig. 1 for the structure of the plot.

2.4.3.2 Spearman's correction

When we wish to compute the correlation between two parameters, measurement errors of each parameter weaken our results. Spearman's correction accounts for this effect and utilizes repeated measurements to correct it. We can infer the Pearson correlation between the latent variables φ and ψ, given N measurements of φ, marked $x_1, \ldots, x_N$, and M measurements of ψ, marked $y_1, \ldots, y_M$. The following estimator for the Pearson correlation between φ and ψ is then used [53]:

$$\hat{r}_{\phi\psi} = \frac{\left(\sum_{i,j}^{N,M} r_{x_i,y_j}\right)^{\frac{1}{N\times M}}}{\left(\sum_{i<i'}^{N} r_{x_i,x_{i'}}\right)^{\frac{1}{N(N-1)}} \left(\sum_{j<j'}^{M} r_{y_j,y_{j'}}\right)^{\frac{1}{M(M-1)}}}$$

Equation 2.4-1 Spearman's correction for attenuation of correlation.

Where r_{x_i,y_j} is the empirical correlation between measurements x_i and y_j. We assume that all the empirical correlations are positive. The estimator is in $[0,\infty)$.

To correct the mRNA correlation between the groups, we treat φ as the levels of mRNA in one group, and ψ as the levels in the other group. We do the same for protein levels. Note that this method can also be used to correct mRNA-protein correlations within a group, treating φ as the levels of mRNA, and ψ as the levels of protein in that group.

2.4.4 Comparing magnitude of differences in protein and mRNA

2.4.4.1 Regressing $\log FC_{protein}$ on $\log FC_{mRNA}$

For all pairs of groups in all datasets, we regressed $\log FC_{protein}$ on $\log FC_{mRNA}$ using ordinary least square (OLS) or a variant of the major axis (MA) regression. For EAR, MMT, and PRIMATE we used regular MA. For NCI60 we used scaled MA (SMA). The choice of which variant of MA to use followed [90] (see next section). We employed three different versions of F−test supplied in the smatr package [91] to test whether the slope is significantly different from 1 for OLS and (S)MA regression. We applied FDR correction for each dataset and method separately. Example for regression result is demonstrated in Figure 2.4-2.

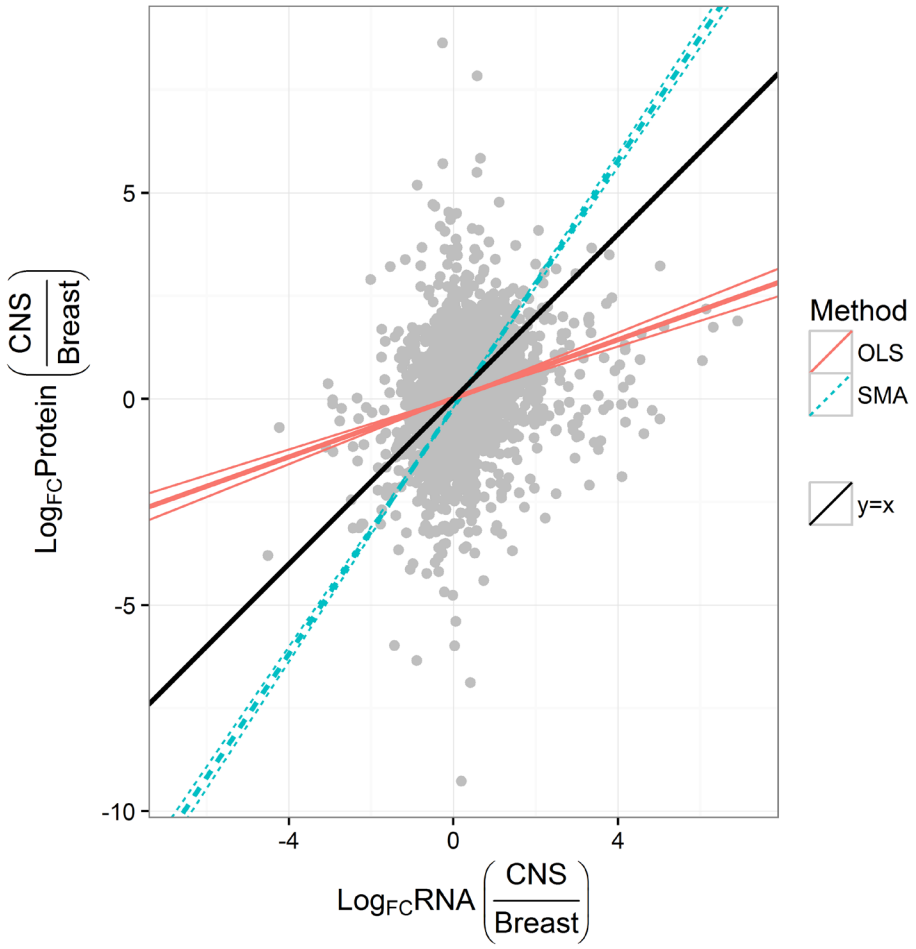

Figure 2.4-2 Discordance between SMA and OLS regression methods. Comparing the CNS and the breast NCI60 cell lines, the protein fold changes (y-axis) were regressed on the mRNA fold changes (x-axis). The fitted regression lines using either OLS (red, solid) or SMA (blue, dashed) were plotted, along with their 95 percent confidence interval (thinner lines). The black line is y=x. While the OLS slope is significantly lower than 1, suggesting range compression, the more reliable SMA slope is significantly higher than 1, in accordance with range expansion.

2.4.4.2 Choice of variant of major axis

We followed the recommendations in [90] in order to choose which variant of major axis (MA) to use. For a pair of groups, the variance of error of log FC_{mRNA}, E_{mRNA}, is approximately the sum of the variances of error of the log-transformed mRNA expression levels in both groups. We make the simplistic assumption that the errors are independently and normally distributed with the same variance across genes and replicates. Hence, within a group, the variance of error in a single measurement can be estimated from replicates, by

taking the mean of gene expression levels' variances. To account for the averaging we do over the replicates when calculating gene expression, we divide the variance by the number of replicates in order to get the variance of error in a group. The same can be done to estimate the variance of error of $\log FC_{protein}$, $E_{protein}$. For protein, the estimator $\widehat{E_{protein}}$ would be an underestimate, due to the large number of missing measurements. The variance of the variable $\log FC_{mRNA}$, V_{mRNA}, can be estimated by reducing E_{mRNA} from the observed variance. The same can be done to estimate the variance of $\log FC_{protein}$, $V_{protein}$. If the ratio $\widehat{E_{mRNA}}/\widehat{E_{protein}}$ is closer to 1 than to the ratio $\widehat{V_{mRNA}}/\widehat{V_{protein}}$, MA is preferred over Scaled MA (SMA) and vice-versa. Performing these calculations on our data, we observed that for the EAR and PRIMATE MA is preferred, and for NCI60 SMA is preferred (for most pairs). For MMT, since there are no replicates in protein, the error cannot be estimated. Assuming that the error in protein for MMT is similar to the error in other MS platforms, the data supports the overall use of MA.

2.4.4.3 Non-parametric approach

We used a nonparametric approach to test whether genes that are up-regulated in one group versus the other in the mRNA domain will show lower PTR in that same group versus the other. This idea allows us to use established tools for differential expression (DE) analysis of mRNA that have high power in discovering such genes.

For each dataset, for each pair of groups, we conducted the following analysis (compare Figure 2.4-3):

1. We separated one mRNA sample from each group for PTR calculation.
2. For each group, we took the matching protein sample for PTR calculation. For unpaired datasets (EAR and MMT) we took the average over all protein samples in the group for PTR calculation.

3. For each group we divided the protein levels in the mRNA levels, to get a PTR vector. We will refer to the difference $\log \mathrm{PTR}_{\mathrm{group2}} - \log \mathrm{PTR}_{\mathrm{group1}}$ as $\log \mathrm{FC}_{\mathrm{PTR}}$.
4. Using the remaining mRNA samples, we performed DE analysis. edgeR [59] was used for the analysis in EAR and MMT RNA-seq datasets and samr (https://cran.r-project.org/web/packages/samr, see [92]) was used for the NCI60 microarray-based data, as well as the PRIMATE RNA-seq dataset (as the counts matrix was not directly available). Both methods also output estimations of $\log \mathrm{FC}_{\mathrm{mRNA}} = \log \mathrm{mRNA}_{\mathrm{group2}} - \log \mathrm{mRNA}_{\mathrm{group1}}$ for every gene, referred to as the DE value. These estimations differ slightly from those calculated by us in the regression analysis. We chose to work with them, in order for the PTR calculation to be consistent with the DE analysis.
5. We tested the significance of the Spearman's rank correlation between the $\log \mathrm{FC}_{\mathrm{PTR}}$ and the $\log \mathrm{FC}_{\mathrm{mRNA}}$ vectors, using the function cor.test in stats package (www.R-project.org). If our decoupling assumption is correct, we expect significant negative correlations. We refer to this test as a global test.
6. We tried two FDR thresholds (0.05, 0.1) for defining which genes are DE. For each FDR threshold, we tested whether the genes up-regulated in the first group have higher $\log \mathrm{FC}_{\mathrm{PTR}}$ than the genes in the other group. We used a one-tailed Wilcoxon rank sum test. We refer to this test as a local test.

In the DE testing, we excluded genes for which one or both protein values were missing and thus the PTR could not be calculated. We repeated this analysis by for every possible pair of repeats selected in step 1 for the PTR calculations. In order to summarize the results we took a median over the results of all pairs. We report both original p-values and FDR corrected q-values. The correction was applied for each dataset and for each test (correlation, Wilcoxon rank sum with FDR 0.05, and with FDR 0.1) separately, over all pairs of groups.

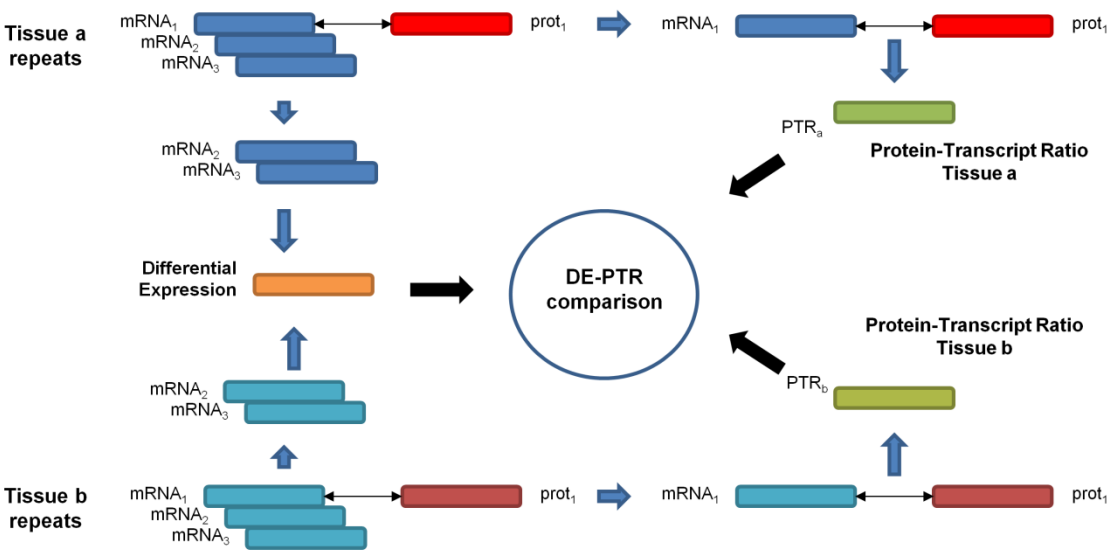

Figure 2.4-3 Illustration of how mRNA samples are split in order to decouple differential expression analysis from protein-transcript ratios calculation. The latter requires integration with protein data, that can be a matched sample when such pairing exists, or averaged protein levels otherwise. This decoupling is needed for the Wilcoxon test to be unbiased.

2.4.5 Protein levels prediction

Assuming we have T−1 groups with matching mRNA and protein profiles, and we want to predict the protein levels in a new group T, using the data from the first T−1 groups and the mRNA levels in group T.

2.4.5.1 Prediction models

We compared three different estimators:

1. **Average PTR (APTR):** It was previously suggested to use the average translational efficiencies measured in the first T−1 groups, and multiply them by the matching mRNA levels in group T [49]. A trivial linear model describing this prediction for a single gene is:

$$\log P_T = \frac{1}{T-1}\sum_{i=1}^{T-1} \log \frac{P_i}{R_i} R_T$$

Equation 2.4-2 APTR protein prediction model.

Where P_i $\mathrm{P_i}$ and R_i are the measured protein and mRNA levels, respectively, in group *i*. This model can be generalized by giving weights to the different groups. The result is called **Weighted Average PTR (WAPTR)** estimator. Weights are obtained by regression.

2. **FC Based (FCB):** A different model assumes linear relationship between log*P* and log*R* (similar to [48]). If for group *i* $\log \mathrm{P_i} = \alpha \log \mathrm{R_i} + \beta$, then for two groups: $\log \frac{\mathrm{P_1}}{\mathrm{P_2}} = \alpha \log \frac{\mathrm{R_1}}{\mathrm{R_2}}$. α is estimated by regression. We expect 0<α<1, in concordance with our previous results. By averaging over all groups, we obtain the following estimator for $\log P_T$:

$$\log P_T = \frac{1}{T-1} \sum_{i=1}^{T-1} \left(\alpha \log \frac{R_T}{R_i} + \log P_i \right)$$

Equation 2.4-3 WAPTR protein prediction model.

Or in a different form, which shows the relation to the APTR estimator:

$$\log P_{T\ FCB} = \log P_{T\ APTR} + \frac{1}{T-1} \sum_{i=1}^{T-1} (1-\alpha) \log \frac{R_i}{R_T}$$

Equation 2.4-4 WAPTR protein prediction model, alternative form.

To generalize the model by allowing group weights, the simplest way assumes an exponential scaling of the protein levels between different groups, that is $\gamma_i \log \mathrm{P_i} = \alpha \log \mathrm{R_i} + \beta$, with $\gamma_T = 1$. This would yield the **Relaxed FCB (RFCB)** estimator:

$$\log P_T = \frac{1}{T-1} \sum_{i=1}^{T-1} \left(\alpha \log \frac{R_T}{R_i} + \gamma_i \log P_i \right)$$

Equation 2.4-5 RFCB protein prediction model.

The group-specific exponents are obtained by regression.

3. **Average Protein (AP):** The simplest estimator is averaging over the protein levels in the other groups, ignoring the mRNA data:

$$\log P_T = \frac{1}{T-1}\sum_{i=1}^{T-1} \log P_i$$

Equation 2.4-6 AP protein prediction model.

This model can also be expanded to give weights for the different groups (**Weighted Average Protein (WAP)** estimator). Weights are obtained by regression.

2.4.5.2 Scoring prediction models

For each dataset we included only the genes for which we had proteomic and transcriptomic data from each of the groups, i.e. a measurement was available for at least one sample belonging to the group (5048, 3514, 3223, and 3394 genes in EAR, MMT, NCI60, and PRIMATE datasets, respectively). We then averaged the data over the repeats in each group. We iterated over the groups, each time setting another one as missing. For each of the aforementioned models we fitted a regression model that allowed scaling of the original estimator and also included an intercept. We performed 10-fold cross-validation on the fitted model, and collected the Root Mean Square Error (**RMSE**), using the DAAG package (cran.r-project.org/web/packages/DAAG). For each group we divided the RMSE by the standard deviation of the protein levels in the group. The result is a dimensionless measure for prediction quality called **NRMSE**, which can be used to compare predictions across datasets.

We followed a different procedure when calculating how much of the variance in protein level is explained by a specific model. We fitted the model for each group separately, and took the median percentage of variance explained. A similar technique [93], which is more appropriate for the evaluation of prediction under a cross-validation scenario, gave results within a range of <1% of the reported results.

For the prediction of protein levels of oncogenes in the NCI60 dataset, we fitted the regression models using data from all genes except the selected oncogenes.

2.4.6 Comparing enrichments in protein and mRNA

2.4.6.1 EAR enrichment analysis

We reran DE and subsequent enrichment analysis using different filters on the genes included in the analysis. The filters were based on the number of measurements in the protein domain: (1) No filter. (2) At least one measurement. (3) At least one, two or three measurements in both tissues (the last one is the complete cases filter). For mRNA DE we used the edgeR package, with a detection threshold of $q\text{–}value \leq 0.05$. For protein we used the samr package, two class unpaired test, with threshold $q\text{–}value \leq 0.1$. We used a less strict FDR threshold for the protein in order to obtain a large enough list of genes for this analysis. The default parameters of samr allow some missing data imputation using the k-nearest neighbor algorithm. Still, for some genes samr could not assign a test score, and they were removed from the background set.

We performed the enrichment analysis using the Expander software [64], checking for enrichment in Gene Ontology (GO) (http://www.geneontology.org) 'biological process' (BP) ontology (*corrected* $p\text{–}value \leq 0.05$). For mRNA and protein, we looked for enrichments in the set of genes up-regulated in the cochlea versus the vestibule and vice-versa, using as a background set all the genes that passed the filter and were tested for DE.

For comparing lists of enrichments terms we used the REVIGO tool [94]. A threshold of 0.7 was used to define similar terms.

2.4.6.2 EAR GOProfile analysis

Comparing the sets of terms emerging in enrichment analysis is sensitive to the significance threshold set for the analysis. In order to gain another perspective on the difference in functions between the mRNA and the protein, we used goProfiles [95], which enabled us to ask whether the functional profiles of the DE genes are different between the domains without presuming a threshold on term significance. A functional profile is defined here as the joint frequencies of annotation in a given set of GO classes.

For each tissue, we compared the functional profile of the genes up-regulated in the protein domain, with the functional profile of the genes up-regulated in the mRNA domain using goProfiles [95]. We followed the methodology described in [95] by starting the comparison in a deep level of the GO ('biological process', 6th level), performing a global test for difference of function profiles, and only if it was significant, doing a class-by-class test to identify significant. If such were not found we restarted the process at a level less deep. For the global test, we used the function 'compareGOProfiles' with the default parameters. For the class-by-class test, we performed Fisher's exact test with FDR correction using the function 'fisherGOProfiles' and the FDR threshold 0.05.

2.4.6.3 MMT enrichment analysis

For each pair of tissues, and each direction of comparison, we calculated $\log \mathrm{FC}_{\mathrm{protein}}$ and $\log \mathrm{FC}_{\mathrm{mRNA}}$ (for $\log \mathrm{FC}_{\mathrm{protein}}$ the relative protein levels were used). We filtered out genes for which we could not calculate either quantity. We created two different orderings of the genes, one by $FC_{protein}$ and the other by FC_{mRNA}. We ran GOrilla [96] on both lists, using the 'biological process' (BP) ontology, with an FDR threshold of 1*10-3 (see 'Choice of Enrichment Analysis Tool' as to why GOrilla was used). This analysis provided us with two lists of terms for each combination of tissues and comparison direction.

In order to determine the semantic similarity between an mRNA list and a protein list of terms, we adopted the method of [97]. The semantic transcriptome specificity is defined as 1 minus the averaged maximal similarities between each term in the mRNA list with any term in the protein list; the semantic translatome specificity is defined as 1 minus the averaged maximal similarities between each term in the protein list with any term in the mRNA list. The semantic similarity between two GO terms is a score between 0 (no similarity) to 1 (full similarity), and it is calculated using the Rel method [98] provided by GOSemSim [99]. The difference between the semantic translatome and transcriptome specificities is a score between -1 to 1.

To calculate the transcriptome versus translatome specificity degree associated to GO slim terms, we followed [97]. For each GO slim term t the transcriptome specificity degree is calculated as the ratio between the number of times a descendant term of t was mRNA specific to the number of times a descendant term of t appeared, across all combinations of tissue pairs and direction. The translatome specificity is calculated similarly, using the protein specific list of terms. These measures are slightly different from [97], in which the aggregation at the level of GO slim was performed after calculating term specificity scores. This modification provides a more accurate estimator in a scenario with relatively low number of terms. In order to set a significance threshold on the term specificity to one of the domains, we assumed that the unique enrichments we see are randomly sampled, with each enrichment coming from the mRNA domain in probability p and the protein domain in probability 1-p. We estimated $p=0.56$ as the proportion of unique enrichments that were found in the mRNA domain. Then, we asked for each GO slim term whether the corresponding proportion is different than p (two-sided proportion test, q-value$\leq$0.1). A proportion larger than p would indicate transcriptome specificity, and vice versa. This simplistic null hypothesis, made for the sake of the test, ignores the inherent dependence that

exists between the uniqueness of mRNA terms and protein terms. The same significant terms were obtained when using a semantic similarity threshold to determine whether a GO term is unique (threshold=0.7, Rel method [98]).

2.4.6.4 Choice of enrichment analysis tool

We used different tools to check for functional enrichment in the EAR and the MMT datasets. For the EAR we used the TANGO algorithm in Expander [64]. This algorithm considers the hierarchical tree-like structure of the gene ontology, using it to provide a good estimation of statistical significance for each term, one that takes multiple testing into account. Also, the reported terms are filtered for redundancy.

For the MMT we decided to forfeit these advantages, in favor of an approach that is cut-off independent. In the MMT dataset there are no replicates (in protein), so a gene DE status cannot be assigned a p-value. If we had chosen to use a cut-off dependent tool like Expander, we would have had to set some arbitrary threshold on the fold-changes in order to define target gene sets. This is different from the analysis in the EAR, where we used a statistical threshold on the corrected p-values, and not on a threshold based on fold-changes. A statistical threshold allows estimating the amount of noise that enters the enrichment analysis, whereas a threshold on the fold-changes would not allow such estimation and would make the enrichment results questionable. For this reason we chose to work with GOrilla [96], to which we entered the genes in order of decreasing fold-changes.

2.4.7 Identifying post-transcriptionally repressed genes

In order to find the post-transcriptionally repressed genes of a group, we ordered the genes in decreasing levels of mRNA expression. We then iterated over the list and calculated the fraction of genes that have a valid measurement in the protein domain, out of those that we already iterated upon. For a given value q, the index of the last iteration, before which we saw

ten fractions in a row above q, was used as a threshold. All genes appearing before this index that have no protein measurements were defined as post-transcriptionally repressed. We note that this definition is more robust than the one used [50], as it avoids inclusion of borderline genes.

We then used the Expander software [64] to perform enrichment analysis on these genes, exploring all ontologies, 'biological process' (BP), 'molecular function' (MF) and 'cellular component' (CC) (corrected p−value ≤0.05). As background we chose all genes with expression above the threshold.

We ran this process for all the groups in datasets EAR, MMT, and NCI60. For PRIMATE, we ran the process on the entire dataset without separation into groups, as an internal test showed that the post-transcriptionally repressed functions in LCLs are similar between species, and such separation would only reduce our power in detecting these functions.

3 RESULTS

3.1 Transcriptomics analysis

Sensory epithelia were dissected from the cochlea and vestibule of mice at two stages of development - embryonic day 16.5 (E16.5) and postnatal day 0 (P0). RNA-seq was performed. Our analysis identified 39,178 Ensembl genes (including non-coding genes and pseudogenes), 15,206 of which have at least one read per million in three or more of the samples and were included in the analysis.

3.1.1 Tissue source and age are associated with differences in transcription

A principal component analysis (PCA) plot demonstrates the four groups are easily separable (Figure 3.1-1). The first principal component (PC1) explains almost half the variance, and is associated with the age of the sample, whereas PC2 explains about a quarter of the variance and is associated with the originating tissue (F-test on associations, *p-values*=1.99×10^{-5}, 1.31×10^{-5} respectively). Additional PCs are not associated with either tissue or age (*p-value*≥*0.05)*. The E16.5 groups show less intra-variability than the P0 groups. This might reflect differences in the rate of development of the different organs between mice from the same population in the period between E16.5 and P0. We stress that this is likely to be a real biological phenomenon and not an artifact of the quality of the dissections, as the dissection of the IE is considered easier and more anatomically accurate at an older age.

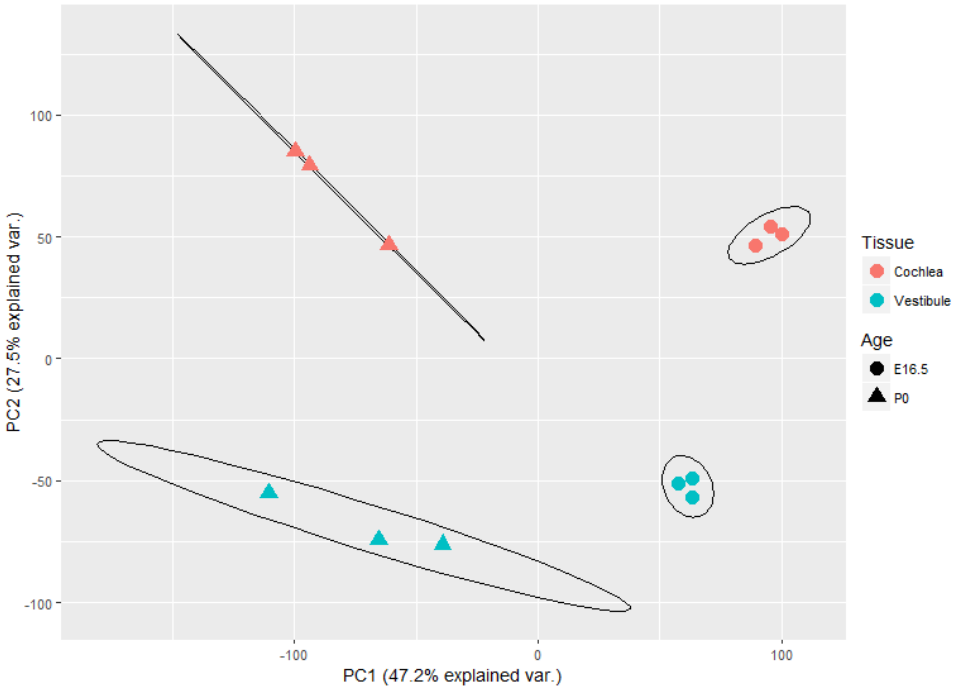

Figure 3.1-1 PCA plot comparing samples in the different ages and tissues according to their mRNA expression. The x- and y-axis are the first and second coordinates, respectively. The samples are colored by their originating tissue, and their shape is determined by age. A normal contour line is drawn for each group for 68% probability.

Using linear mixed models, we estimated the percentage of variance that can be attributed to age, tissue or the interaction of age and tissue. According to our estimates, the majority of variance is attributed to either age (44.0±6.5) or tissue (39.6±5.4) (± stands for standard deviation). Still, a non-negligible percentage is attributed to the interaction term (8.0±1.5), and a model with this interaction term better describes the data according to a restricted likelihood ratio test (*p-value*≤2.2×10^{-16}). Less than 10% of the variance was left unexplained (8.4±1.08).

3.1.2 Change in hair cells proportion in sensory epithelia

When dissecting the sensory epithelia, the HCs are not easily isolated from the adjacent SCs. Thus, all samples contain varying amounts of two roughly defined populations of cells: HCs and SCs. As a result, differences in expression between conditions can be attributed to differences in the expression profiles with a population or to a change in the cell mixture

composition. In order to explore the second option, we used expression signatures of HCs and SCs from a previous experiment [9], in order to estimate the proportion of HCs and SCs in each condition.

We assumed that our P0 cochlear data is a mixture of expression of two types of cells, HCs and SCs, and computed the proportion of each type from the RNA-seq data measured in [9] for in the GFP+ and GFP- P0 cochlear samples, respectively. We made similar assumptions for the other conditions, with the exceptions that we used the E16 samples from the original article as surrogates for E16.5 samples, and utricle samples as surrogate for vestibule samples. We also assessed the proportions under an alternative assumption, in which the cochlear sample contains both cells from the cochlea and cells from the vestibule (or utricle) as contamination, and vice versa.

The subset of genes used to create the signature was different for E16.5 and P0. For each age, we ordered the genes in decreasing order of expression variance across the four reference samples (cochlear and vestibular GFP+ and GFP- samples). We took the expression of the first *k* genes in the list, with *k* equals 453 for E16.5 and 193 for P0. The value of *k* was chosen so that it minimized the estimated percentage of contamination in our mixed data, i.e., the estimated percentage of cochlear cells in vestibular samples plus the estimated percentage of vestibular cells in cochlear samples. We assume this heuristic improved the overall prediction accuracy, although it did not optimize directly the precision of the HCs percentage estimation, which is our main goal. We used DeconRNASeq for estimating the mixing proportions [78].

The estimated proportions of HCs are similar in both scenarios (Figure 3.1-2), so we will focus on the scenario without tissue contamination. The estimated percentages are 32.6(±1.6) and 23.8(± 1.0) in the cochlea and the vestibule at E16.5, and 44.0(±1.1), and

40.1(±0.2) in the cochlea and the vestibule at P0, respectively (± stands for standard deviation). The percentage of HCs is higher in the cochlea at both ages, and increases with development in both tissues, with the increase in the vestibule being more prominent (1.9-fold increase compared to 1.4-fold in the cochlea). Strikingly, in all estimations, the percentage of SCs is higher than 50%, suggesting that they have a dominant influence on the expression profiles.

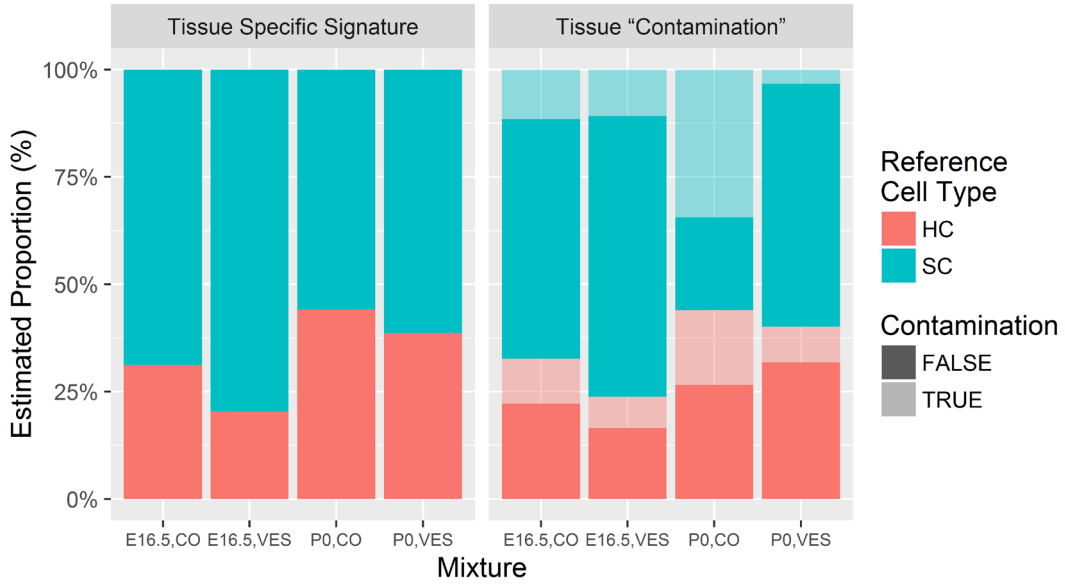

Figure 3.1-2 Estimated proportion of HCs and SCs in samples. This estimated proportion of each cell type in each of the groups is displayed in a stack bar chart, where the color of a stack identifies the cell type. In the left figure, the cells composing a tissue were confined to cells originating from that tissue, without allowing cross-tissue contamination, whereas in the right figure, cross-tissue contamination is assumed to occur. A light color indicates contaminated content. For example, focusing on the cochlear tissue at age P0 (P0.CO), the estimated proportion of HCs, when contamination is not allowed, is 44.1% (in red). When contamination is allowed, the estimated proportion of HCs slightly decreases to 44.0%; a percentage composed of 26.6% cochlear HCs (in dark red) and 17.4% contaminating vestibular HCs (in light red). The three other samples show a majority of non-contaminated tissue (darker colors).

Even with k selected to minimize the contamination, our calculation gives 51.8% contamination in the cochlea at P0. We are unsure how to interpret this high number: If that number does not reflect the reality and is inflated, it may be due to (1) experimental noise, either in our data or the data used to generate the expression signatures at P0, or (2) inaccuracy of the deconvolution method when the signatures are similar. This similarity

manifests in high correlations between signatures of the same cell type in the cochlea and the vestibule (r=0.65, 0.83 for signatures of HCs and SCs, respectively).

3.1.3 Variations in tissue functionalities and developmental timeline

We extracted genes that are differentially expressed between tissues or between ages, and genes for which the interaction of tissue and age is significant in determining expression. 3306 were found to be up-regulated at P0 compared to E16.5, and 6890 down-regulated. 4159 were found to be up-regulated in the vestibule compared to the cochlea, and 2382 down-regulated. For 745 transcripts the cochlea to vestibule expression ratio increases with development, and for 1211 the ratio decreases.

We performed GO and KEGG enrichment analyses on the three lists. The enrichment results are summarized below.

3.1.3.1 Expression change with age

Genes that were upregulated at E16.5 are enriched for terms related to cell cycle, DNA replication, cytoskeleton organization, and other terms that are in accordance with a highly proliferative state (Table 3.1-1, red). In contrast, genes that were upregulated at P0 are enriched for ribosomes, indicating high protein synthesis, mainly of plasma membrane and extracellular matrix proteins (Table 3.1-1, green). The lipid and oxphos-related metabolic activities are also high in this group. The cells at this stage of development are more adhesive, communicate more with one another, and are more responsive to external cues. They are also responsive to a variety of signaling receptors, including calcium signaling, and have high ion transport activity. The upregulated terms are typical of a less proliferative environment, where the highly expressed genes promote homeostatic processes and inhibit peptidase activity. Some terms show signs of cell specialization, in terms of sensory perception, cartilage-related metabolism, and the regulation of ossification; the last might indicate a

cross-talk between sensory epithelium cells and endochondral cells. Another marker for the more differentiated state is an up-regulation of the MHC protein complex. In summary, the enrichment suggests that the inner ear is in a more proliferative state at E16.5 than at P0, whereas at P0 the tissues are more differentiated and exhibit specialization for sensory perception.

KEGG enrichment (Table 3.1-2) generally confirmed the aforementioned differences and provided more details regarding specific metabolic processes activated at P0. For example, we could attribute the enriched lipid metabolism to sphingolipids, arachidonic acid, and retinol, the enriched aminoglycan metabolism to glycan degradation, and the biosynthesis of chondroitin and keratan sulfate. Pathways enriched at P0 suggest that the activity of the immune system increases during development, with leukocytes migrating into the tissue and intercellular communication using cytokines. As the complement and coagulation cascades and the renin-angiotensin system are also enriched at P0, we can hypothesize that the inner ear is more exposed to blood circulation at this age.

Table 3.1-1 GO Enrichments in genes differentially expressed between ages. Enrichments found in genes up-regulated at E16.5 or P0 (marked red and green, respectively, in the 'Set' column) using gene ontology (GO). For each enrichment, we provide the number of genes annotated with term (#genes), the significance of the enrichment (raw and corrected p-values), and the frequency of genes annotated with the term.

Set	Enriched with	#genes	Raw p-value	Corrected p-Value	Frequency in set (%)
Up in E16.5	nucleic acid metabolic process - GO:0090304	1506	3.36E-70	1.00E-04	22
	chromosome - GO:0005694	368	3.58E-38	1.00E-04	5.3
	nuclear lumen - GO:0031981	792	1.52E-31	1.00E-04	11
	non-membrane-bounded organelle - GO:0043228	1406	1.71E-30	1.00E-04	20
	DNA binding - GO:0003677	845	8.60E-30	1.00E-04	12
	DNA repair - GO:0006281	228	6.30E-28	1.00E-04	3.3
	M phase - GO:0000279	256	1.84E-27	0.0001	3.7

	cell cycle phase - GO:0022403	312	5.06E-26	0.0001	4.5
	cell cycle - GO:0007049	501	2.26E-22	0.0001	7.3
	chromatin organization - GO:0006325	272	7.42E-22	0.0001	3.9
	organelle organization - GO:0006996	881	7.93E-22	0.0001	13
	chromatin modification - GO:0016568	251	8.85E-22	0.0001	3.6
	mitotic cell cycle - GO:0000278	258	6.88E-21	0.0001	3.7
	DNA replication - GO:0006260	122	6.88E-21	0.0001	1.8
	ATP binding - GO:0005524	750	4.87E-20	0.0001	11
	mRNA metabolic process - GO:0016071	227	6.11E-16	0.0001	3.3
	methylation - GO:0032259	139	4.68E-13	0.0001	2
	ncRNA metabolic process - GO:0034660	160	3.87E-12	0.0001	2.3
	microtubule cytoskeleton organization - GO:0000226	141	2.71E-11	0.0001	2
	DNA geometric change - GO:0032392	37	2.00E-10	0.0001	0.54
	tRNA metabolic process - GO:0006399	86	1.28E-09	0.0001	1.2
	nucleoside-triphosphatase regulator activity - GO:0060589	216	1.59E-09	0.0001	3.1
	chromatin binding - GO:0003682	161	5.65E-09	0.0002	2.3
	nuclease activity - GO:0004518	83	1.00E-08	0.0002	1.2
	regulation of DNA metabolic process - GO:0051052	113	1.86E-08	0.0003	1.6
	purine NTP-dependent helicase activity - GO:0070035	58	3.41E-08	0.0003	0.84
	gene silencing - GO:0016458	45	1.17E-07	0.0009	0.65

	metal ion binding - GO:0046872	1469	1.27E-07	0.001	21
	regulation of cell cycle process - GO:0010564	164	2.11E-07	0.0011	2.4
	phosphotransferase activity, alcohol group as acceptor - GO:0016773	347	3.51E-07	0.0015	5
	hydrolase activity, acting on acid anhydrides - GO:0016817	345	7.66E-07	0.0025	5
	condensed chromosome, centromeric region - GO:0000779	27	1.02E-06	0.0031	0.39
	negative regulation of DNA metabolic process - GO:0051053	41	1.02E-06	0.0031	0.6
	macromolecule modification - GO:0043412	882	1.20E-06	0.0034	13
	replication fork - GO:0005657	29	1.21E-06	0.0034	0.42
	histone lysine methylation - GO:0034968	35	1.31E-06	0.0034	0.51
	reciprocal meiotic recombination - GO:0007131	18	1.45E-06	0.0039	0.26
	regulation of microtubule cytoskeleton organization - GO:0070507	56	1.90E-06	0.0055	0.81
	spindle pole - GO:0000922	41	5.89E-06	0.0194	0.6
	centrosome organization - GO:0051297	35	9.67E-06	0.0344	0.51
	regulation of GTPase activity - GO:0043087	158	1.19E-05	0.0453	2.3
Up in P0	extracellular region - GO:0005576	505	1.35E-81	0.0001	15
	extracellular space - GO:0005615	238	5.49E-45	0.0001	7.2
	endoplasmic reticulum - GO:0005783	370	2.36E-30	0.0001	11

	proteinaceous extracellular matrix - GO:0005578	141	2.09E-26	0.0001	4.3
	plasma membrane - GO:0005886	800	1.88E-25	0.0001	24
	vacuole - GO:0005773	141	1.49E-24	0.0001	4.3
	biological adhesion - GO:0022610	205	6.31E-17	0.0001	6.2
	plasma membrane part - GO:0044459	363	1.27E-13	0.0001	11
	Golgi apparatus - GO:0005794	300	5.50E-13	0.0001	9.1
	response to wounding - GO:0009611	128	2.04E-12	0.0001	3.9
	chemical homeostasis - GO:0048878	194	3.18E-12	0.0001	5.9
	transmembrane transporter activity - GO:0022857	227	4.63E-12	0.0001	6.9
	negative regulation of peptidase activity - GO:0010466	63	4.67E-12	0.0001	1.9
	response to chemical stimulus - GO:0042221	432	5.58E-12	0.0001	13
	MHC protein complex - GO:0042611	17	1.53E-11	0.0001	0.51
	cytosolic ribosome - GO:0022626	46	2.89E-11	0.0001	1.4
	polysaccharide binding - GO:0030247	65	3.18E-11	0.0001	2
	monovalent inorganic cation transmembrane transporter activity - GO:0015077	94	3.27E-11	0.0001	2.8
	collagen - GO:0005581	41	4.35E-11	0.0001	1.2
	enzyme inhibitor activity - GO:0004857	78	4.43E-11	0.0001	2.4
	carbohydrate binding - GO:0030246	101	5.80E-11	0.0001	3.1
	negative regulation of multicellular	94	1.05E-10	0.0001	2.8

	organismal process - GO:0051241				
	apical part of cell - GO:0045177	97	1.22E-10	0.0001	2.9
	regulation of biological quality - GO:0065008	443	1.25E-10	0.0001	13
	antigen processing and presentation of peptide antigen - GO:0048002	24	2.22E-10	0.0001	0.73
	endopeptidase regulator activity - GO:0061135	44	3.11E-10	0.0001	1.3
	oxidoreductase activity - GO:0016491	187	4.70E-10	0.0001	5.7
	establishment of localization - GO:0051234	649	8.33E-10	0.0001	20
	receptor activity - GO:0004872	202	1.07E-09	0.0001	6.1
	response to organic substance - GO:0010033	273	1.80E-09	0.0001	8.3
	organelle membrane - GO:0031090	324	1.93E-09	0.0001	9.8
	aminoglycan metabolic process - GO:0006022	34	2.82E-09	0.0002	1
	sulfur compound metabolic process - GO:0006790	56	3.75E-09	0.0002	1.7
	extracellular matrix organization - GO:0030198	59	4.01E-09	0.0002	1.8
	vesicle - GO:0031982	218	4.27E-09	0.0002	6.6
	peptide binding - GO:0042277	52	1.06E-08	0.0002	1.6
	cellular cation homeostasis - GO:0030003	87	1.79E-08	0.0003	2.6
	hydrogen ion transmembrane transporter activity - GO:0015078	38	2.05E-08	0.0003	1.1
	growth factor binding - GO:0019838	45	2.56E-08	0.0003	1.4

	cation transport - GO:0006812	167	2.90E-08	0.0003	5.1
	lipid metabolic process - GO:0006629	205	4.42E-08	0.0006	6.2
	negative regulation of molecular function - GO:0044092	177	1.66E-07	0.0011	5.4
	regulation of hormone levels - GO:0010817	51	2.46E-07	0.0013	1.5
	calcium ion binding - GO:0005509	135	7.86E-07	0.0026	4.1
	regulation of response to external stimulus - GO:0032101	92	8.56E-07	0.0027	2.8
	regulation of ossification - GO:0030278	54	1.32E-06	0.0035	1.6
	negative regulation of transport - GO:0051051	79	1.42E-06	0.0037	2.4
	regulation of localization - GO:0032879	317	1.54E-06	0.0041	9.6
	secondary metabolic process - GO:0019748	15	3.24E-06	0.0107	0.45
	regulation of multicellular organismal process - GO:0051239	380	3.86E-06	0.0135	11
	secretory granule - GO:0030141	56	5.36E-06	0.0173	1.7
	positive regulation of cell migration - GO:0030335	70	8.00E-06	0.0281	2.1
	Golgi membrane - GO:0000139	56	8.50E-06	0.0305	1.7
	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen - GO:0016705	44	8.58E-06	0.0305	1.3
	chondroitin sulfate proteoglycan	14	1.00E-05	0.0354	0.42

	metabolic process - GO:0050654				
	sensory perception - GO:0007600	88	1.15E-05	0.0432	2.7
	response to other organism - GO:0051707	89	1.23E-05	0.047	2.7

Table 3.1-2 KEGG Enrichments in genes differentially expressed between ages. Enrichments found in genes up-regulated at E16.5 or P0 using Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment. See caption in **Table 3.1-1** for the structure of the table.

Set	Enriched with	#genes	Raw p-value	Corrected p-Value	Frequency in set (%)
Up in E16.5	Homologous recombination	26	6.51E-08	6.51E-08	2.02
	Mismatch repair	22	8.62E-08	8.62E-08	2.09
	DNA replication	31	1.03E-07	1.03E-07	1.91
	Base excision repair	27	1.99E-05	1.99E-05	1.77
	Cell cycle	79	3.03E-05	3.03E-05	1.39
	Aminoacyl-tRNA biosynthesis	33	4.26E-05	4.26E-05	1.64
	Non-homologous end-joining	10	6.18E-04	0.000618	2.09
	Pyrimidine metabolism	55	2.62E-03	0.00262	1.32
Up in P0	Lysosome	60	6.55E-14	6.55E-14	2.47
	Ribosome	50	2.94E-13	2.94E-13	2.63
	Cell adhesion molecules (CAMs)	50	2.82E-09	2.82E-09	2.18
	ECM-receptor interaction	41	4.23E-09	4.23E-09	2.35
	Oxidative phosphorylation	55	8.05E-09	8.05E-09	2.05
	Metabolic pathways	269	3.30E-07	0.00000033	1.31
	Cytokine-cytokine receptor interaction	52	2.37E-06	0.00000237	1.81
	Parkinson's disease	48	3.82E-06	0.00000382	1.84
	Antigen processing and presentation	25	6.27E-06	0.00000627	2.32
	Sphingolipid metabolism	20	1.29E-05	0.0000129	2.49
	Other glycan degradation	11	1.63E-05	0.0000163	3.43
	Complement and coagulation cascades	19	1.84E-05	0.0000184	2.51
	Glycosaminoglycan degradation	13	3.09E-05	0.0000309	2.98
	Leukocyte transendothelial migration	40	5.60E-05	0.000056	1.78
	Drug metabolism - cytochrome P450	17	6.71E-05	0.0000671	2.47
	Arachidonic acid metabolism	20	1.04E-04	0.000104	2.24
	Graft-versus-host disease	11	1.23E-04	0.000123	3

	Metabolism of xenobiotics by cytochrome P450	15	2.44E-04	0.000244	2.42
	Type I diabetes mellitus	13	5.08E-04	0.000508	2.47
	Allograft rejection	11	5.63E-04	0.000563	2.67
	Asthma	7	7.63E-04	0.000763	3.39
	Systemic lupus erythematosus	20	7.98E-04	0.000798	1.98
	Retinol metabolism	11	1.06E-03	0.00106	2.53
	Autoimmune thyroid disease	11	1.06E-03	0.00106	2.53
	Renin-angiotensin system	9	1.10E-03	0.0011	2.8
	Sulfur metabolism	6	2.62E-03	0.00262	3.27
	Circadian rhythm - mammal	8	3.16E-03	0.00316	2.68
	Alzheimer's disease	49	3.34E-03	0.00334	1.43
	Chondroitin sulfate biosynthesis	11	4.96E-03	0.00496	2.18
	Keratan sulfate biosynthesis	8	5.89E-03	0.00589	2.49
	Cardiac muscle contraction	23	6.55E-03	0.00655	1.64

3.1.3.2 Expression change between tissues

According to the enrichment analysis (Table 3.1-3), a number of the differentially expressed (DE) genes in both the cochlea and vestibule are involved in signal transduction. In the cochlea, the majority of the signaling is mediated by voltage- and ligand-gated ion channels and can be attributed to neuron-neuron synaptic transmission. In agreement with this finding, other upregulated activities are neurogenesis and neuron projection. In contrast, the signaling in the vestibule is probably required for the coordination of both innate and acquired immune responses, an observation that relates to the main function enriched in this tissue. The signaling, some of which involves purinergic receptors, plays a role in the response to external stimulus and stress, and also in taxis. Another function enriched in the vestibule is locomotion, with the cilium and the axoneme being two enriched cellular components related to the movement of the HCs' stereocilia. The vestibule is richer in blood vessel formation and hematopoiesis, and the extracellular matrix is more evolved than in the cochlea. Together with the high immune-related activity, these factors may explain why the vestibular cells are more adhesive. We also detected enrichment for replacement ossification,

suggesting the development of bone. As a generalization, upregulated genes were associated with neurological terms in the cochlear, but to vascular, structural, and immunological terms in the vestibule. This partitioning was not perfect as we could detect enrichment for mesenchymal cell differentiation in the cochlea, and 3.1% of the upregulated genes in the vestibule were annotated for a role in sensory perception.

The KEGG enrichment (Table 3.1-4) data also agreed with the characterization of the cochlea as more neurological versus a more vascular vestibule. In addition, the data provided more information about the typical signaling in each apparatus. Neuroactive ligand signaling was identified in both, although the cochlea was associated with the TGF-beta, MAPK, and ErbB signaling pathways, while cytokine-mediated, calcium, and Toll-like receptor signaling were more important in the vestibule. Three pathways shown to be unique to the cochlea affect cell proliferation, survival, differentiation, and migration [100–102], suggesting that these developmental processes are more activated in the cochlea. Other unique metabolic pathways enriched in the cochlea were O-glycan and chondroitin sulfate biosynthesis. The vestibule, on the other hand, was enriched for glycan degradation and metabolic pathways concerning arachidonic acid, retinol, and glutathione.

Table 3.1-3 GO Enrichments in genes differentially expressed between tissues. Enrichments found in genes up-regulated in the cochlea or in the vestibule (abbreviated 'coch' and 'vest' and marked green and red, respectively, in the 'Set' column) using GO. See caption in **Table 3.1-1** for the structure of the table.

Set	Enriched with	#genes	Raw p-value	Corrected p-Value	Frequency in set (%)
Up in Coch	neuron projection - GO:0043005	195	7.65E-21	0.0001	8.2
	cell morphogenesis involved in differentiation - GO:0000904	114	5.86E-17	0.0001	4.8
	multicellular organismal signaling - GO:0035637	119	8.92E-17	0.0001	5
	cell morphogenesis involved in neuron differentiation - GO:0048667	93	9.76E-17	0.0001	3.9
	synapse - GO:0045202	150	4.38E-16	0.0001	6.3
	signaling - GO:0023052	527	4.19E-15	0.0001	22
	gated channel activity - GO:0022836	81	1.32E-14	0.0001	3.4
	regulation of neurogenesis - GO:0050767	115	3.46E-14	0.0001	4.8

	regulation of cell development - GO:0060284	131	4.10E-14	0.0001	5.5
	cell-cell signaling - GO:0007267	118	2.30E-13	0.0001	5
	cell periphery - GO:0071944	559	8.81E-13	0.0001	23
	regulation of ion transmembrane transport - GO:0034765	74	2.62E-12	0.0001	3.1
	system development - GO:0048731	458	3.96E-12	0.0001	19
	cell development - GO:0048468	213	4.64E-12	0.0001	8.9
	voltage-gated cation channel activity - GO:0022843	44	2.82E-11	0.0001	1.8
	postsynaptic membrane - GO:0045211	58	9.64E-11	0.0001	2.4
	cell projection morphogenesis - GO:0048858	103	1.02E-10	0.0001	4.3
	regulation of multicellular organismal process - GO:0051239	310	1.21E-10	0.0001	13
	regulation of localization - GO:0032879	259	1.64E-10	0.0001	11
	cellular developmental process - GO:0048869	390	4.48E-10	0.0001	16
	regulation of cell differentiation - GO:0045595	192	5.48E-10	0.0001	8.1
	regulation of multicellular organismal development - GO:2000026	209	1.96E-09	0.0001	8.8
	receptor activity - GO:0004872	154	2.71E-09	0.0001	6.5
	behavior - GO:0007610	92	5.93E-09	0.0001	3.9
	regulation of system process - GO:0044057	101	6.77E-09	0.0001	4.2
	cell body - GO:0044297	100	3.28E-08	0.0002	4.2
	signal transducer activity - GO:0004871	158	3.55E-08	0.0002	6.6
	neuron projection terminus - GO:0044306	32	3.86E-08	0.0002	1.3
	transmembrane signaling receptor activity - GO:0004888	106	4.77E-08	0.0002	4.5
	transporter activity - GO:0005215	186	2.11E-07	0.0009	7.8
	extracellular ligand-gated ion channel activity - GO:0005230	21	4.31E-07	0.0015	0.88
	presynaptic membrane - GO:0042734	21	4.31E-07	0.0015	0.88
	neurotransmitter transport - GO:0006836	30	1.01E-06	0.0031	1.3
	anatomical structure morphogenesis - GO:0009653	263	1.05E-06	0.0032	11
	intrinsic to plasma membrane - GO:0031226	112	1.23E-06	0.0037	4.7
	regulation of biological quality - GO:0065008	308	1.51E-06	0.0048	13
	regulation of transmembrane transporter activity - GO:0022898	33	1.55E-06	0.0049	1.4
	sensory perception - GO:0007600	71	1.82E-06	0.0057	3

	regulation of cell communication - GO:0010646	243	2.42E-06	0.008	10
	mesenchymal cell differentiation - GO:0048762	32	5.96E-06	0.0215	1.3
	sensory organ development - GO:0007423	86	6.80E-06	0.0245	3.6
	neuron-neuron synaptic transmission - GO:0007270	18	7.31E-06	0.026	0.76
	cyclic nucleotide metabolic process - GO:0009187	18	7.31E-06	0.026	0.76
	chemical homeostasis - GO:0048878	126	7.93E-06	0.0292	5.3
	regulation of signaling - GO:0023051	306	1.04E-05	0.0385	13
	cyclic nucleotide catabolic process - GO:0009214	9	1.15E-05	0.0424	0.38
	positive regulation of transport - GO:0051050	93	1.21E-05	0.0443	3.9
Up in Vest	extracellular region - GO:0005576	515	7.07E-52	1.00E-04	12
	plasma membrane - GO:0005886	1018	2.33E-39	1.00E-04	24
	plasma membrane part - GO:0044459	500	2.72E-31	1.00E-04	12
	extracellular space - GO:0005615	234	9.16E-27	1.00E-04	5.6
	response to wounding - GO:0009611	179	1.62E-25	1.00E-04	4.3
	inflammatory response - GO:0006954	107	1.20E-20	1.00E-04	2.6
	regulation of multicellular organismal process - GO:0051239	540	1.10E-19	1.00E-04	13
	proteinaceous extracellular matrix - GO:0005578	144	2.34E-18	1.00E-04	3.5
	response to chemical stimulus - GO:0042221	549	7.77E-18	1.00E-04	13
	biological adhesion - GO:0022610	242	8.37E-18	1.00E-04	5.8
	receptor activity - GO:0004872	269	1.65E-17	1.00E-04	6.5
	response to organic substance - GO:0010033	356	6.61E-16	1.00E-04	8.6
	cell activation - GO:0001775	143	1.35E-15	1.00E-04	3.4
	signaling - GO:0023052	848	3.95E-15	1.00E-04	20
	innate immune response - GO:0045087	88	7.56E-15	1.00E-04	2.1
	intrinsic to plasma membrane - GO:0031226	204	2.03E-14	1.00E-04	4.9
	system development - GO:0048731	753	3.14E-14	1.00E-04	18
	signal transduction - GO:0007165	765	3.47E-14	1.00E-04	18
	regulation of immune response - GO:0050776	121	4.23E-14	1.00E-04	2.9
	vasculature development - GO:0001944	162	2.18E-13	1.00E-04	3.9
	regulation of developmental process - GO:0050793	425	3.54E-13	1.00E-04	10
	regulation of cellular component movement - GO:0051270	163	1.80E-12	1.00E-04	3.9
	signal transducer activity - GO:0004871	266	2.04E-12	1.00E-04	6.4

regulation of cell proliferation - GO:0042127	329	2.14E-12	1.00E-04	7.9
defense response to bacterium - GO:0042742	41	2.32E-12	1.00E-04	0.99
response to other organism - GO:0051707	127	2.81E-12	1.00E-04	3.1
positive regulation of cell migration - GO:0030335	100	3.37E-12	1.00E-04	2.4
regulation of response to stimulus - GO:0048583	591	5.41E-12	0.0001	14
regulation of angiogenesis - GO:0045765	69	6.27E-12	0.0001	1.7
apical part of cell - GO:0045177	115	1.27E-11	0.0001	2.8
positive regulation of biological process - GO:0048518	893	2.01E-11	0.0001	21
developmental process - GO:0032502	944	3.27E-11	0.0001	23
immune effector process - GO:0002252	82	3.98E-11	0.0001	2
positive regulation of developmental process - GO:0051094	221	1.54E-10	0.0001	5.3
cellular developmental process - GO:0048869	633	1.98E-10	0.0001	15
anatomical structure morphogenesis - GO:0009653	452	2.01E-10	0.0001	11
regulation of response to external stimulus - GO:0032101	119	4.59E-10	0.0001	2.9
regulation of biological quality - GO:0065008	529	6.36E-10	0.0001	13
anatomical structure formation involved in morphogenesis - GO:0048646	222	1.47E-09	0.0001	5.3
lipid metabolic process - GO:0006629	251	2.37E-09	0.0001	6
negative regulation of multicellular organismal process - GO:0051241	105	2.52E-09	0.0001	2.5
cilium - GO:0005929	95	3.11E-09	1.00E-04	2.3
chemical homeostasis - GO:0048878	215	6.79E-09	1.00E-04	5.2
myeloid leukocyte activation - GO:0002274	38	1.10E-08	2.00E-04	0.91
vacuole - GO:0005773	121	2.83E-08	2.00E-04	2.9
regulation of leukocyte migration - GO:0002685	38	4.04E-08	2.00E-04	0.91
positive regulation of defense response - GO:0031349	52	4.25E-08	2.00E-04	1.3
positive regulation of cytokine production - GO:0001819	64	4.46E-08	2.00E-04	1.5
multicellular organismal homeostasis - GO:0048871	50	6.39E-08	4.00E-04	1.2
locomotion - GO:0040011	211	7.81E-08	4.00E-04	5.1
receptor binding - GO:0005102	298	1.07E-07	5.00E-04	7.2
cell-cell junction - GO:0005911	102	1.60E-07	8.00E-04	2.5

cytokine production - GO:0001816	37	1.61E-07	8.00E-04	0.89
activation of innate immune response - GO:0002218	26	1.73E-07	8.00E-04	0.63
blood circulation - GO:0008015	77	1.76E-07	8.00E-04	1.9
regulation of localization - GO:0032879	390	1.98E-07	9.00E-04	9.4
membrane raft - GO:0045121	84	2.43E-07	1.10E-03	2
hemopoiesis - GO:0030097	125	2.47E-07	1.10E-03	3
response to external stimulus - GO:0009605	191	2.82E-07	1.10E-03	4.6
regulation of response to stress - GO:0080134	175	3.80E-07	1.40E-03	4.2
regulation of plasma lipoprotein particle levels - GO:0097006	19	4.04E-07	1.40E-03	0.46
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains - GO:0002460	40	4.61E-07	1.50E-03	0.96
leukocyte differentiation - GO:0002521	78	4.62E-07	1.50E-03	1.9
vascular process in circulatory system - GO:0003018	36	5.91E-07	1.90E-03	0.87
regulation of cell adhesion - GO:0030155	92	6.83E-07	2.20E-03	2.2
phagocytosis - GO:0006909	34	8.48E-07	2.50E-03	0.82
unsaturated fatty acid metabolic process - GO:0033559	28	1.29E-06	3.70E-03	0.67
carbohydrate binding - GO:0030246	103	1.66E-06	5.30E-03	2.5
lipid binding - GO:0008289	197	1.82E-06	5.70E-03	4.7
positive regulation of signal transduction - GO:0009967	212	1.84E-06	5.80E-03	5.1
antigen processing and presentation - GO:0019882	27	2.84E-06	1.00E-02	0.65
regulation of hormone levels - GO:0010817	56	2.86E-06	1.01E-02	1.3
positive regulation of cell adhesion - GO:0045785	51	2.98E-06	1.09E-02	1.2
production of molecular mediator involved in inflammatory response - GO:0002532	12	3.00E-06	1.10E-02	0.29
leukocyte chemotaxis - GO:0030595	29	3.65E-06	1.32E-02	0.7
polysaccharide binding - GO:0030247	62	3.88E-06	1.37E-02	1.5
calcium ion binding - GO:0005509	157	3.99E-06	1.39E-02	3.8
taxis - GO:0042330	95	4.69E-06	1.59E-02	2.3
cellular cation homeostasis - GO:0030003	93	5.47E-06	1.92E-02	2.2
axoneme - GO:0005930	21	6.45E-06	2.36E-02	0.5

	plasma lipoprotein particle - GO:0034358	15	7.62E-06	2.77E-02	0.36
	purinergic receptor activity - GO:0035586	15	7.62E-06	2.77E-02	0.36
	regulation of protein secretion - GO:0050708	41	8.14E-06	0.0302	0.99
	replacement ossification - GO:0036075	16	8.24E-06	0.0306	0.38
	regulation of tumor necrosis factor production - GO:0032680	30	8.98E-06	0.0342	0.72
	cytokine biosynthetic process - GO:0042089	11	9.59E-06	0.036	0.26

Table 3.1-4 KEGG Enrichments in genes differentially expressed between tissues. Enrichments found in genes up-regulated in the cochlea or in the vestibule using KEGG enrichment. See caption in **Table 3.1-3** for the structure of the table.

Set	Enriched with	#genes	Raw p-value	Corrected p-Value	Frequency in set (%)
Up in Coch	TGF-beta signaling pathway	29	1.84E-06	1.84E-06	2.41
	Neuroactive ligand-receptor interaction	40	2.27E-06	2.27E-06	2.07
	Axon guidance	39	3.42E-05	3.42E-05	1.89
	Long-term depression	21	3.88E-04	3.88E-04	2.12
	O-Glycan biosynthesis	10	8.96E-04	8.96E-04	2.88
	MAPK signaling pathway	56	0.0017	0.0017	1.46
	Fc epsilon RI signaling pathway	20	0.00385	0.00385	1.83
	ErbB signaling pathway	24	0.00386	0.00386	1.73
	Chondroitin sulfate biosynthesis	9	0.00573	0.00573	2.48
	Gap junction	21	0.00608	0.00608	1.74
Up in Vest	Cytokine-cytokine receptor interaction	74	1.26E-12	1.26E-12	2.05
	Leukocyte transendothelial migration	54	4.39E-08	4.39E-08	1.91
	Cell adhesion molecules (CAMs)	53	3.18E-07	0.000000318	1.84
	ECM-receptor interaction	41	3.73E-06	0.00000373	1.87
	Lysosome	52	8.33E-06	0.00000833	1.7
	Hematopoietic cell lineage	28	1.07E-05	0.0000107	2.07
	Arachidonic acid metabolism	22	2.85E-04	0.000285	1.96

	Neuroactive ligand-receptor interaction	50	8.87E-04	0.000887	1.48
	Complement and coagulation cascades	18	1.71E-03	0.00171	1.89
	Focal adhesion	72	1.91E-03	0.00191	1.35
	Retinol metabolism	12	1.93E-03	0.00193	2.19
	Calcium signaling pathway	53	2.27E-03	0.00227	1.41
	Toll-like receptor signaling pathway	33	2.77E-03	0.00277	1.55
	Metabolism of xenobiotics by cytochrome P450	15	3.25E-03	0.00325	1.93
	Drug metabolism - cytochrome P450	16	4.13E-03	0.00413	1.85
	Graft-versus-host disease	10	5.21E-03	0.00521	2.17
	Glutathione metabolism	21	6.08E-03	0.00608	1.65
	Other glycan degradation	9	6.21E-03	0.00621	2.23
	Renin-angiotensin system	9	6.21E-03	0.00621	2.23

3.1.3.3 Tissues expression ratio change with age

Genes for which the cochlea to vestibule expression ratio increased with age $\left(\frac{Cochlea}{Vestibule}\uparrow\right)$, were enriched with processes related to sensory perception and central nervous system development, as well as signaling through G-coupled receptors, ligand-gated ion channels, or calcium (Table 3.1-5). Accordingly, a significant number of genes were annotated to be in the apical part of the cell. Other genes annotated to the extracellular region might mediate the biological adhesion, which increases during development. Another enriched component was identified as the sarcomere, which most closely resembles the stereocilia in the inner ear.

Table 3.1-5 GO Enrichments in genes for which the cochlea to vestibule expression ratio changes with age. Enrichments found in genes for which the cochlea to vestibule expression ratio is increasing or decreasing with age (marked green and red, respectively, in the 'Set' column) using GO. See caption in **Table 3.1-1** for the structure of the table.

Set	Enriched with	#genes	Raw p-value	Corrected p-Value	Frequency in set (%)
$\frac{C}{V}\uparrow$	extracellular region - GO:0005576	127	3.21E-21	1.00E-04	17
	inner ear development - GO:0048839	29	6.77E-11	1.00E-04	3.9
	calcium ion binding - GO:0005509	54	1.28E-10	1.00E-04	7.2
	sensory organ development - GO:0007423	48	1.37E-10	1.00E-04	6.4
	sensory perception - GO:0007600	40	2.78E-10	1.00E-04	5.4
	system development - GO:0048731	163	2.42E-08	1.00E-04	22
	plasma membrane - GO:0005886	190	3.18E-08	1.00E-04	26
	receptor activity - GO:0004872	63	5.19E-08	1.00E-04	8.5
	anatomical structure morphogenesis - GO:0009653	106	7.57E-08	1.00E-04	14
	gated channel activity - GO:0022836	31	1.22E-07	2.00E-04	4.2
	transmembrane signaling receptor activity - GO:0004888	46	1.48E-07	3.00E-04	6.2
	cell-cell signaling - GO:0007267	44	2.38E-07	7.00E-04	5.9
	behavior - GO:0007610	39	2.95E-07	8.00E-04	5.2
	excitatory extracellular ligand-gated ion channel activity - GO:0005231	11	2.98E-07	8.00E-04	1.5
	cation channel activity - GO:0005261	28	4.54E-07	0.0014	3.8
	mechanoreceptor differentiation - GO:0042490	13	7.35E-07	0.0022	1.7
	detection of stimulus involved in sensory perception - GO:0050906	12	8.27E-07	0.0023	1.6
	response to mechanical stimulus - GO:0009612	13	1.19E-06	0.0038	1.7
	G-protein coupled receptor activity - GO:0004930	27	1.61E-06	0.0048	3.6
	chemical homeostasis - GO:0048878	53	3.42E-06	0.0113	7.1
	sarcomere - GO:0030017	17	4.09E-06	0.0123	2.3
	negative regulation of neuron differentiation - GO:0045665	12	5.92E-06	0.0195	1.6
	biological adhesion - GO:0022610	51	7.12E-06	0.0221	6.8
	apical part of cell - GO:0045177	29	7.46E-06	0.0227	3.9
	embryonic morphogenesis - GO:0048598	40	7.81E-06	0.0235	5.4
	central nervous system development - GO:0007417	42	9.24E-06	0.0272	5.6
	cellular developmental process - GO:0048869	132	9.75E-06	0.029	18
	plasma membrane part - GO:0044459	90	1.08E-05	0.0318	12
	negative regulation of astrocyte differentiation - GO:0048712	6	1.31E-05	0.0375	0.81

$\frac{C}{V}\downarrow$					
	plasma membrane - GO:0005886	331	1.80E-18	1.00E-04	27
	neuron projection - GO:0043005	105	1.39E-12	1.00E-04	8.7
	extracellular region - GO:0005576	146	1.97E-11	1.00E-04	12
	multicellular organismal signaling - GO:0035637	67	3.02E-11	1.00E-04	5.5
	transmembrane transporter activity - GO:0022857	105	3.19E-11	1.00E-04	8.7
	synapse - GO:0045202	84	4.35E-11	1.00E-04	6.9
	regulation of multicellular organismal process - GO:0051239	181	4.58E-11	1.00E-04	15
	cell-cell signaling - GO:0007267	70	7.87E-11	1.00E-04	5.8
	plasma membrane part - GO:0044459	155	9.09E-11	1.00E-04	13
	ion transmembrane transporter activity - GO:0015075	86	1.03E-10	1.00E-04	7.1
	regulation of localization - GO:0032879	151	2.39E-10	1.00E-04	12
	extracellular space - GO:0005615	77	3.51E-10	1.00E-04	6.4
	passive transmembrane transporter activity - GO:0022803	58	3.79E-10	1.00E-04	4.8
	regulation of ion transmembrane transport - GO:0034765	44	2.25E-09	1.00E-04	3.6
	signaling - GO:0023052	271	9.90E-09	1.00E-04	22
	regulation of system process - GO:0044057	62	1.53E-08	1.00E-04	5.1
	cellular developmental process - GO:0048869	212	3.62E-08	1.00E-04	18
	system development - GO:0048731	240	5.37E-08	1.00E-04	20
	intrinsic to plasma membrane - GO:0031226	69	3.24E-07	0.0012	5.7
	establishment of localization - GO:0051234	255	3.59E-07	0.0013	21
	proteinaceous extracellular matrix - GO:0005578	46	1.12E-06	0.0038	3.8
	cell development - GO:0048468	108	1.47E-06	0.0047	8.9
	regulation of neurotransmitter secretion - GO:0046928	12	2.04E-06	0.0062	0.99
	cell projection morphogenesis - GO:0048858	54	2.15E-06	0.0066	4.5
	cell morphogenesis involved in neuron differentiation - GO:0048667	42	2.55E-06	0.008	3.5
	locomotion - GO:0040011	76	2.67E-06	0.0085	6.3
	cell body - GO:0044297	56	3.10E-06	0.0107	4.6
	regulation of action potential - GO:0001508	24	3.23E-06	0.0112	2
	pigment granule - GO:0048770	9	4.39E-06	0.0129	0.74
	regulation of biological quality - GO:0065008	169	5.82E-06	0.0193	14
	polysaccharide binding - GO:0030247	27	7.27E-06	0.0223	2.2
	regulation of cell communication - GO:0010646	135	7.28E-06	0.0224	11

	neurotransmitter transport - GO:0006836	19	8.34E-06	0.0246	1.6
	behavior - GO:0007610	49	9.65E-06	0.0285	4

We can envision two possible scenarios for each of these enrichments. The first option is that genes annotated for enrichment were upregulated in the cochlea at E16.5 and the gap between the cochlea and the vestibule increased during development. The second option is that these genes were upregulated in the vestibule at E16.5 and the gap between the cochlea and the vestibule decreased during development. To distinguish between the two, we compared the expression of all genes that are annotated for each GO term. The median expression log-ratio between the cochlea and the vestibule at P0 was plotted against the value of the same parameter at E16.5 (Figure 3.1-3, circles). The plot only contains the terms for which the gap between the cochlea and the vestibule significantly increased with age. More precisely, only terms for which the log-ratios at P0 were larger than their paired values at E16.5 were included (Wilcoxon signed rank test, q values ≤ 0.05).

Interestingly, the vestibule is appeared to be more specialized for sensory perception at E16.5 than the cochlea, as manifested by a negative median log-ratio for terms sensory perception, mechanoreceptor differentiation, and detection of stimulus involved in sensory perception. However, by P0, the cochlea surpassed the vestibule in all of these fields. In contrast, ligand-gated ion channel activity was already higher in the cochlea at E16.5, and the gap only increased with development.

Genes for which the vestibule to cochlea ratio increased with age $\left(\frac{\text{Vestibule}}{\text{Cochlea}} \uparrow\right)$ were enriched for signaling, neuron projection, neurotransmitter transport, and secretion. These are all functions that were higher in the cochlea at E16.5, and for which the difference between the vestibule and the cochlea decreased with time (Figure 3.1-3, triangles).

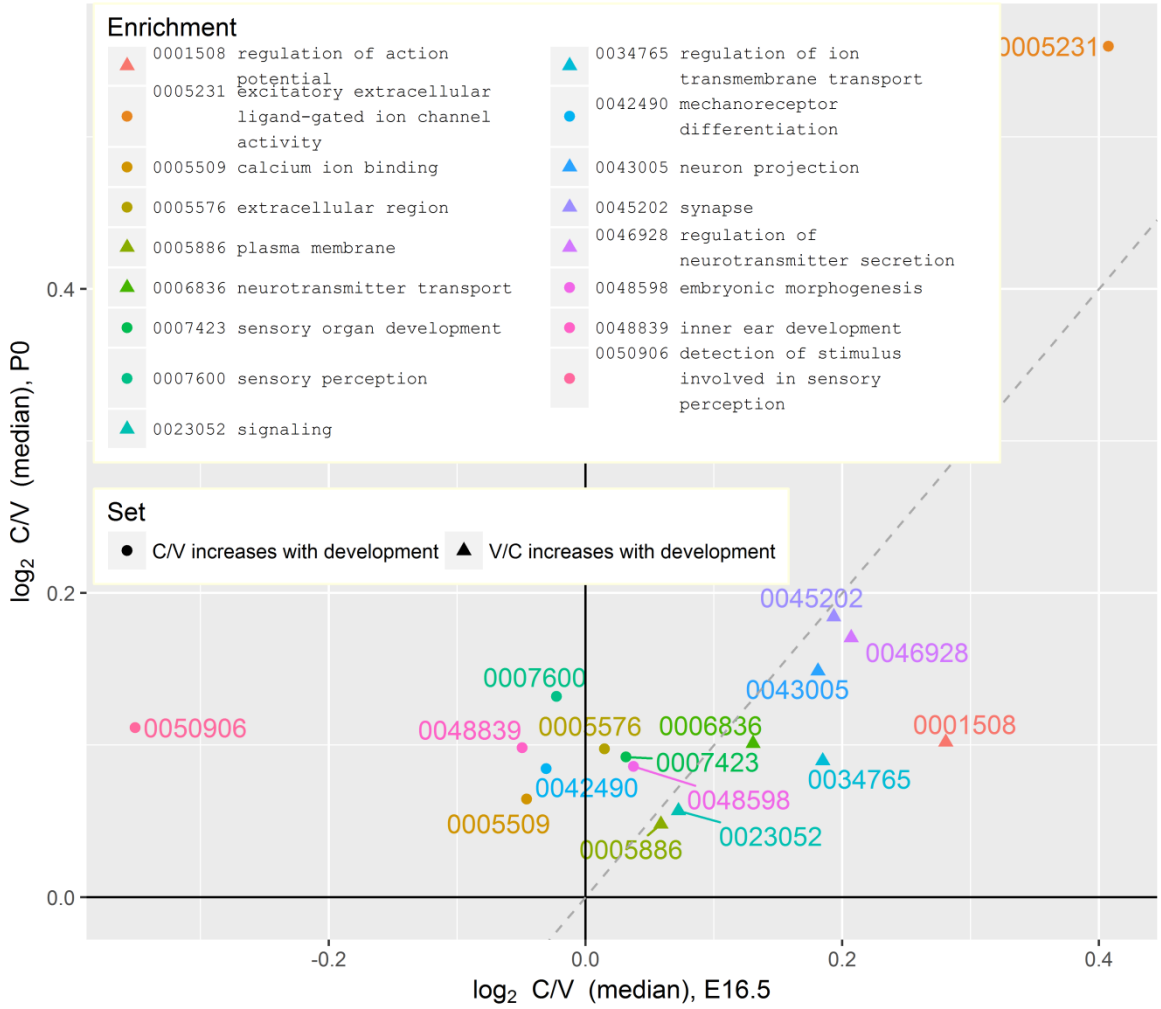

Figure 3.1-3 GO terms enriched with genes affected by age-tissue interaction. The median cochlea to vestibule (C/V) expression ratios of genes annotated for GO terms at P0 (y-axis) against E16.5 (x-axis). Circles mark GO terms enriched in genes with increased C/V ratios between E16.5 and P0, and for which the ratios of all annotated genes are higher at P0 than at E16.5. Triangles mark GO terms with parallel properties for the reciprocal ratio (V/C).

3.1.4 Deafness genes can be predicted using expression patterns

A list of 140 different genes associated with human deafness was compiled from a public dataset (http://hereditaryhearingloss.org/). Expression data for homologous mouse genes of 130 of them is available in our dataset. We observed general patterns of expression for these syndromic and non-syndromic deafness genes (DGs). First, when comparing vestibular and cochlear expression, the fold-changes (FCs) of the DGs are higher in absolute value than the background FCs (p-value = 1.98*10-5, one-sided Wilcoxon rank sum test; Figure 3.1-4, upper subfigure). Also, the FCs of nonsyndromic DGs are slightly higher in absolute value than the

FCs of syndromic DGs (p-value = 7.00*10-2, same test). That is, DGs tend to be tissue-specific, with the nonsyndromic genes being even more specific. Interestingly, the majority of the DE DGs are higher in the vestibule than in the cochlea, in spite of the cochlea's role in hearing (57 out of 76, p-value = 2.19*10-5, two-sided proportion test).

Second, when comparing P0 and E16.5 expressions, DGs tend to have higher FCs compared to background FCs (p-values = 6.32*10-6, one-sided Wilcoxon rank sum test; Figure 3.1-4, middle subfigure). I.e., their expression tends to increase with development. Third, their cochlea to vestibule expression ratio tends to increase with development compared to background (p-values = 5.15*10-6, same test; Figure 3.1-4, lower subfigure). Moreover, the increase in the ratio of nonsyndromic DGs is higher than that for syndromic genes (*p-value* = $3.48*10^{-3}$, same test).

3.1.4.1 Deafness genes prediction

Using the three types of FCs and the averaged expression (see Methods), we built a classifier that can predict whether a gene is a DG. The classifier achieved a ROC score of 0.66±0.04 across repeated training/test splits. A ROC score higher than 0.5 means that these expression data have some predictive value on whether a gene is related to deafness. This classifier performs better than a similar classifier that uses the averaged RPKM values in each condition as features (ROC score: 0.60±0.05). Also, removing one of more of the four feature types from the original classifier resulted in a lower score.

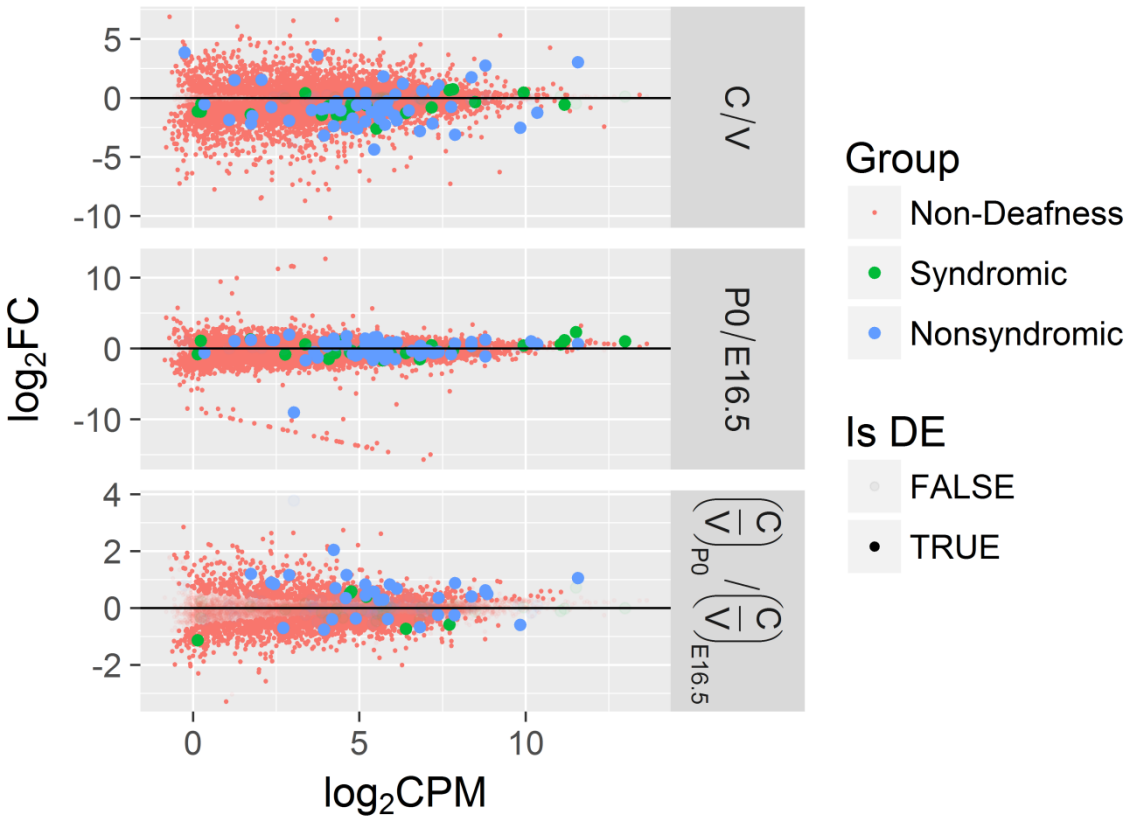

Figure 3.1-4 FCs against average expression for deafness and non-deafness genes. Each point is a gene. The logarithm of the FCs of genes between tissues (upper), ages (middle), and age-tissue combinations (lower), are plotted against their averaged expression across samples (in log counts per million [CPM]). Transparent points correspond to genes that are not differentially expressed in the corresponding comparison. Deafness genes are marked with larger points, and are colored based on the type of deafness they are involved at.

Genes not marked as DGs might still be undiscovered DGs. In this sense our classifier was trained to distinguish between known DGs and genes with unknown role in deafness. We refer to the first group of genes as positive and the second group as unlabeled. We wished to adapt our positive-unlabeled (PU) classifier to output the probability that an unlabeled gene is a positive gene. This type of classification is referred to as transductive PU learning [103]. Suppose that the known DGs are a random subset of all DGs, i.e. the features we explored impose no bias over which of the positive genes is labeled. Then, the probability that the PU classifier assigns to the positivity of new genes [70]: (i) correctly ranks the genes; (ii) the probabilities are only off by a constant factor. We used a bagging-like algorithm similar to the one presented in [103] in order to calculate the probabilities for the set of unlabeled genes,

with some differences detailed in Methods. One main difference between our approach and the one in [103] was that we kept the same proportion of positive (labeled) samples in the training set as in the test set, whereas in [103] all positive samples were included in training. This property allowed us to fix biases in the probabilities at the price of losing some predictive power. One source of bias was due to undersampling in the learning process [72]. A second source of bias was the one described above for a PU classifier. We addressed the latter using methods presented in [70].

To gain some insight about the accuracy of our estimator, in spite of the lack of true labeling for the unlabeled set, we downloaded lists of genes that were associated with HL according to the text mining tools DigSeE [73], DisGeNET [74] and DISEASES [75]. We refer to these genes as *deafness associated genes*. We found 1313 genes associated with deafness by at least one tool, 115 of which are known DGs, covering 82% of all known DGs. The respective numbers for mouse homologs were 1021, 106 and 82%. See Figure 3.1-5 for a comparison of the lists provided by the tools. Assuming a considerable portion of these genes are undiscovered DGs we wished our algorithm to rank those higher than genes that are neither known DGs nor deafness associated.

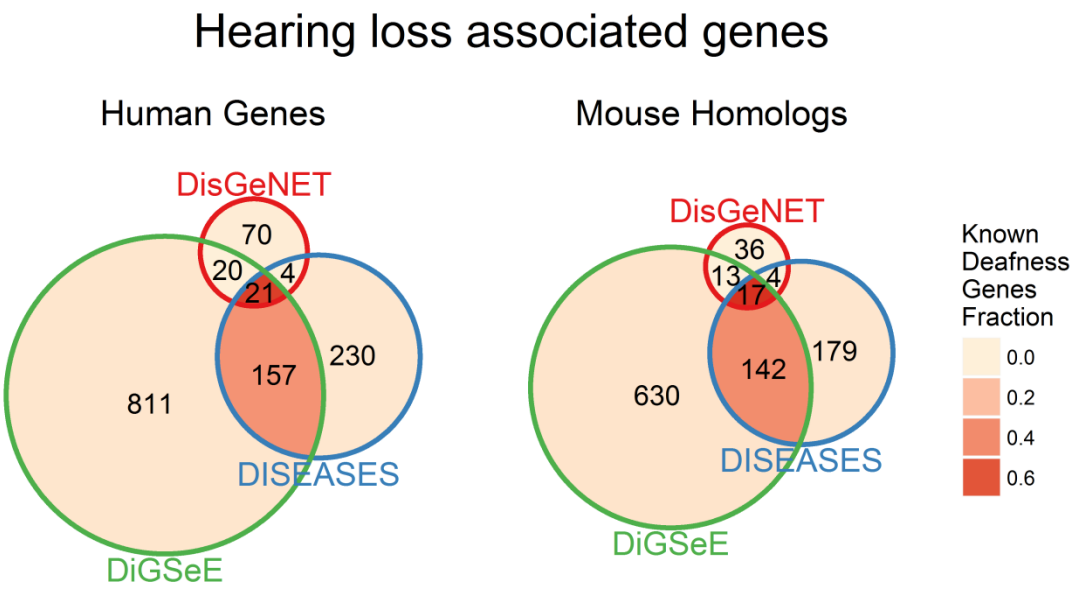

Figure 3.1-5 Number of genes associated with hearing loss. Left: A Venn diagram of the number of human genes associated with HL according to the text mining tools DigSeE, DisGeNET and DISEASES (see Methods). A darker region corresponds to a higher fraction of known deafness genes in the relevant set. Right: similar figure after converting the human genes to their mouse homologs. The text mining tools used showed marked differences in the numbers of associated genes and percentages of known DG within them (DiGSeE: 1009 genes, 10.5% are known DG; DISEASES: 412 genes, 19.4% are known DG; DisGeNet: 115 genes, 13.0%). A higher percentage of known DG indicates a higher specificity. Thus, the largest list provided by DigSeE is relatively nonspecific, but might be more sensitive than the other lists, and the medium size list provided by DISEASES might not be as sensitive, but it is more specific. As expected, the higher the number of text mining tools associating a gene with deafness, the higher the probability of the gene to be a known DG.

Our bagging-like algorithm resulted in a PU classifier with a ROC score of 0.694. The probabilities from this native classifier were biased upward due to undersampling. Correcting for this bias resulted in a better calibration of the probabilities, as demonstrated by a calibration plot (Figure 3.1-6, left), and the lowering of the Brier Score (*BS*) from $2.07*10^{-1}$ to $8.47*10^{-3}$.

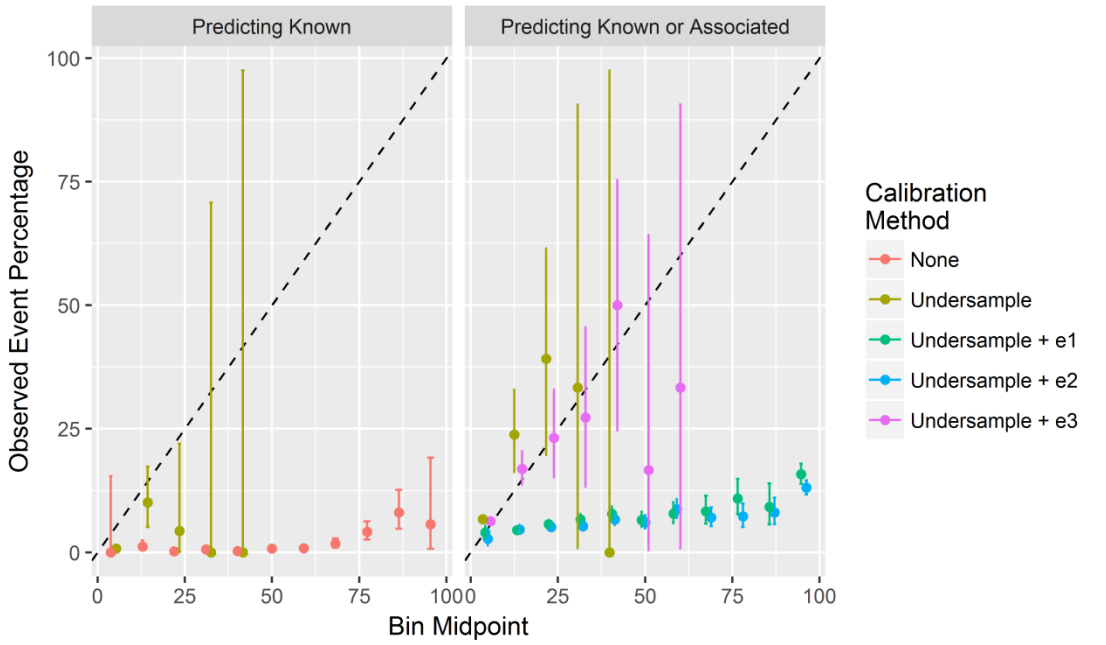

Figure 3.1-6 Probability calibration plots for classification models. The prediction space was discretized into 11 bins. Genes with predicted DG probability between 0 and 0.09 fell in the first bin, between 0.09 and 0.18 in the second bin, etc. For each bin, the mean predicted value was plotted against the true fraction of positive cases, along with the 95% binomial confidence interval. If the model is well calibrated the points should fall near the diagonal line. Left: The predicted probabilities of the PU classifier were either used directly (red) or calibrated for undersampling [72] (yellow). The plot shows how consistent the probabilities are with known deafness genes rates. Right: Three methods (e1, e2, e3) were used in order to calibrate the probabilities of the PU classifier for the classification of deafness genes versus non deafness genes [70]. The calibration was performed after an initial calibration for undersampling. As the true label of an unlabeled gene is unknown, we use as proxy the association of such gene with deafness according to text mining tools. The calibration data for e1, e2, and e3 are colored green, blue, and purple, respectively. For some combinations of bins and calibration methods, there were no samples in the bin, and thus the mean predicted values were not plotted. For example, no gene was predicted to be a DG with a probability over 60% using the e3 method, so no purple points are plotted above this value in the x-axis.

We tried three different methods (e1,e2,e3; see [70]) to correct the bias in the probabilities caused by the PU scenario. All three methods first estimate the probability that a known DG is labeled $p(s = 1|y = 1)$ in order to perform the calibration. The estimates for this probability according to e1, e2, and e3 were 0.032±0.014, 0.022±0.007, and 0.518±0.248, respectively. The estimates made by e1 and e2 support the existence of a few thousands DG, compared with a few hundred according to e3 ($4.1*10^3$, $5.9*10^3$, and $2.5*10^2$, respectively). We believe that given the status of deafness research, the last estimate is more reasonable. To test it further, we assessed the calibration of the probabilities produced by each method. For this, we assumed that *deafness associated genes* are in fact deafness genes. The e3 method resulted in the best calibration, as demonstrated by a calibration plot (Figure 3.1-6, right), and the

lowest *BS* (scores: $6.64*10^{-2}$, $1.20*10^{-1}$, $2.84*10^{-1}$, $6.45*10^{-2}$ for no fix, e1, e2, and e3, respectively). Hence, we decided to use e3 probabilities in the subsequent stages. We mark the probability that gene *g* is positive according to e3 estimate p_g.

We reran our bagging-like algorithm, but this time we chose to treat a gene *g* as a positive example with probability p_g, and as a negative example with probability 1 - p_g. This reassignment was performed before each iteration. Finally, we recalculated the ROC score of our classifier. For this, we ignored known deafness genes in this scoring to allow proper separation of training and test stages. With the rerun we achieved a slightly better ROC score (0.602 vs 0.600, $p<0.05$, DeLong's test for two correlated ROC curves [69]). We chose to continue with the rerun classifier. We performed the calibration due to undersampling on these probabilities as well. The predictions for both human genes and mouse orthologs are available upon request. The twenty mouse genes with the highest predicted probabilities contain the known nonsyndromic DGs *Smpx* and *Ptprq*, seven deafness associated genes (*Gfi1*, *Lhx3*, *Erbb4*, *Ephx1*, *Il33*, *Slc52a3,* and *Ttr*), and nine genes not associated with deafness (*Mlf1, Nell1*, *Espnl*, *Rbm24*, *Lrrc10b*, *Agr3*, *Tgm2*, *Id4Cd164l2,* and *Faim2*).

In order to choose a discrimination threshold for our binary classifier, we offer two assisting plots. Both of which demonstrate how well our classifier predicts deafness associated genes, again ignoring known DGs. The first is a ROC curve, which visualizes the balance between specificity and sensitivity (Figure 3.1-7, top). The threshold maximizing the sum of the two is suggested as a candidate threshold.

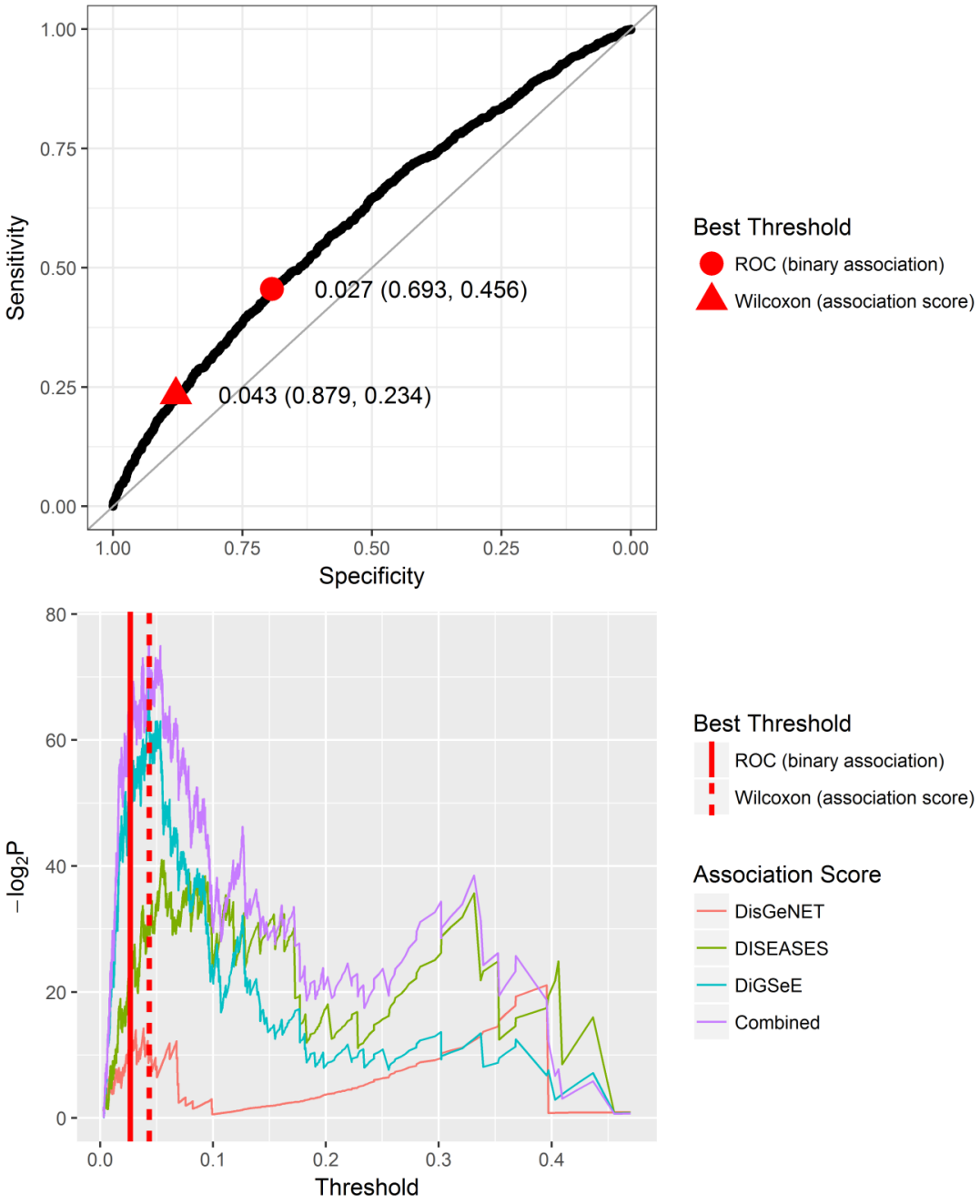

Figure 3.1-7 Choosing a threshold probability for discriminating deafness associated genes. The two plots demonstrate the effect of choosing a threshold on the balance between sensitivity and specificity of prediction. Top: a standard ROC curve. A gene associated with deafness according to any of the text mining tools is considered a positive gene, all others are considered negative. Bottom: Threshold determination based on comparison of association scores. At each threshold, it assigns significance to the difference between the association scores of genes above the threshold and all other genes, according to Wilcoxon rank sum test. A higher $-\log_2 P$ value suggests a more significant difference, in direction of higher association scores for genes above the threshold. Line color indicates the source of the association scores used. Two thresholds are marked in both graphs. The first is the threshold value for which the sum of specificity and sensitivity of the ROC curve is highest (upper: circle shape; lower: solid vertical line). The second is the threshold value which the Wilcoxon test is most significant for the "Combined" association score (upper: triangle shape; lower: dotted line).

A disadvantage of a ROC curve in our context is that it ignores the association scores provided by the text mining tools. In order to account for these scores, we considered the range of values of the threshold and for each one compared the association scores of the genes with probabilities higher than the threshold with all other genes using a non-parametric test (one-tailed Wilcoxon rank sum test). We hypothesized that genes above the 'right' threshold will tend to have higher association scores. We analyzed separately the association scores from each tool and also ran this analysis on scoring based on all three tools (see Methods). We plotted $-\log_2 P-value$ against the threshold (Figure 3.1-7, bottom). The threshold minimizing the p-value for the combined scoring is suggested as a candidate threshold. The suggested thresholds according to the ROC curve and the Wilcoxon test are 0.027 and 0.043, respectively. The respective numbers of genes passing the thresholds are 4764 and 1934. Other thresholds may also be considered, depending on the required number of candidates, specificity and sensitivity. We recommend choosing thresholds that give local maxima using either curve.

3.1.5 Transcription factors affecting expression

We searched for enrichments of transcription factor (TF) binding sites in the three sets of DE genes. We associated 6, 43, and 10 motifs with the expression change across age, tissue, and age-tissue interaction (i.e., the change in the cochlea to vestibule expression ratio throughout development). Considering that the number of genes DE between the two tissues is about 35% less than the number DE between the ages, it is very surprising that the number of motifs regulating tissue differences is almost seven-fold the number regulating age differences. Overall, we found 50 unique motifs across all comparisons, and manually connected them with 64 mouse TFs (i.e., few motifs were associated with multiple TFs).

For each TF, we tested whether the gene itself was DE in the same comparison where its targets were DE. This property interests us for three reasons: (i) It suggests whether the regulation of the TF activity is (at least partially) transcriptional. Knowing how a TF is regulated makes it a better candidate for experimental interventions. (ii) The direction in which a TF is DE implies whether it works as a repressor or an activator. (iii) It strengthens our faith that the associated motif, found in the enrichment analysis, is important for regulation, and not a false-positive. 30/64 of the TF were DE (in at least one comparison).

In order to obtain more insight as to how the levels of the TFs affect their targets, we plotted the median FC (FC) of *all* targets of a specific motif, against the median FC of the TFs associated with that motif (Figure 3.1-8). In all comparisons, we observed a positive, yet insignificant correlation between the two (*Pearson's r=0.51, 0.05, 0.52* for the comparisons across tissue, age, and age-tissue interaction, respectively; *combined p-value* [104]=0.15). Among the factors that contribute to the incomplete correlation is the post-transcriptional regulation of TFs, which reduces the correlation between the transcript levels of a TF and its activity. Also, while most TFs activate transcription of the targets, others repress transcription of some or all of their targets. And third, taking the median FC of the TFs associated with a motif ignores complex relationships between them, such as the ability of a subset to activate transcription without the others (we will later see an example for the motif AHRHIF).

Then, we intersected the TFs found in our developmental experiment, with those shown to change their expression in avian regeneration of IE sensory epithelia [66] (Figure 3.1-9). This was done in the hope of finding common pathways to IE development and regeneration, and specifically revealing those essential for either proliferation of support cells or transdifferentiation to HCs. Out of 712 DE TFs in the regeneration experiment, we mapped

596 to orthologous mouse genes. The intersection with our list of 64 TFs yielded 33 TFs involved in both development and regeneration, 8 of which are also DE.

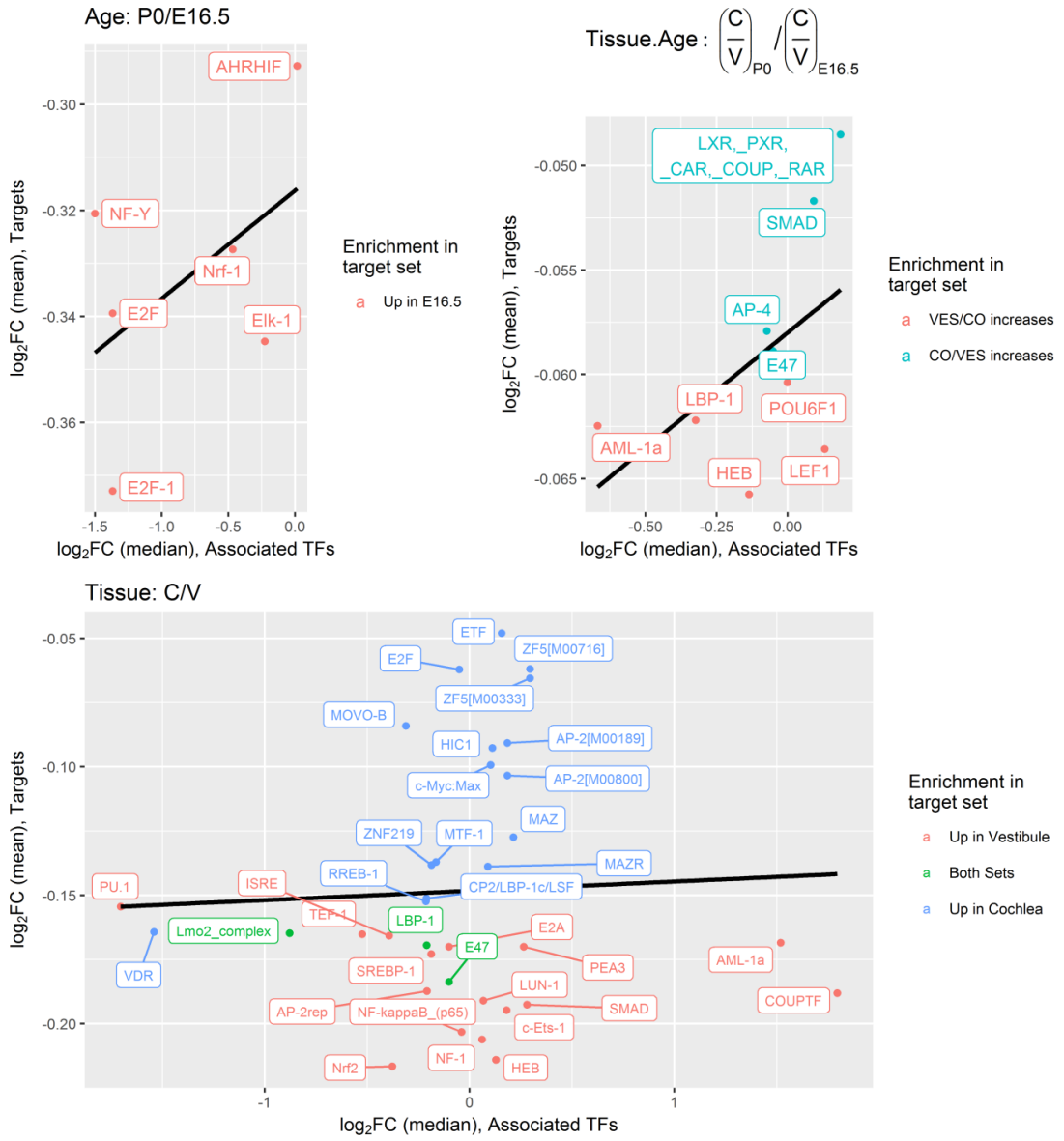

Figure 3.1-8 Expression of Transcription factors and their targets. For the motifs enriched in the differentially expressed genes, we plotted the average log fold-change (FC) of the genes with the motif in their promoter, against the median logFC of the transcription factors associated with the motif. The subset of relevant motifs and the fold-changes were determined separately for each comparison: between ages (upper left), tissues (lower), and the interaction of age and tissue (upper right). The color indicates the gene set in which the motif was enriched. A linear regression line was added to each plot.

Finally, we performed a comprehensive literature search for the motifs found in the context of IE development. For a small subset, the results are detailed in the following sections (3.1.5.1-3) (complete list available upon request).

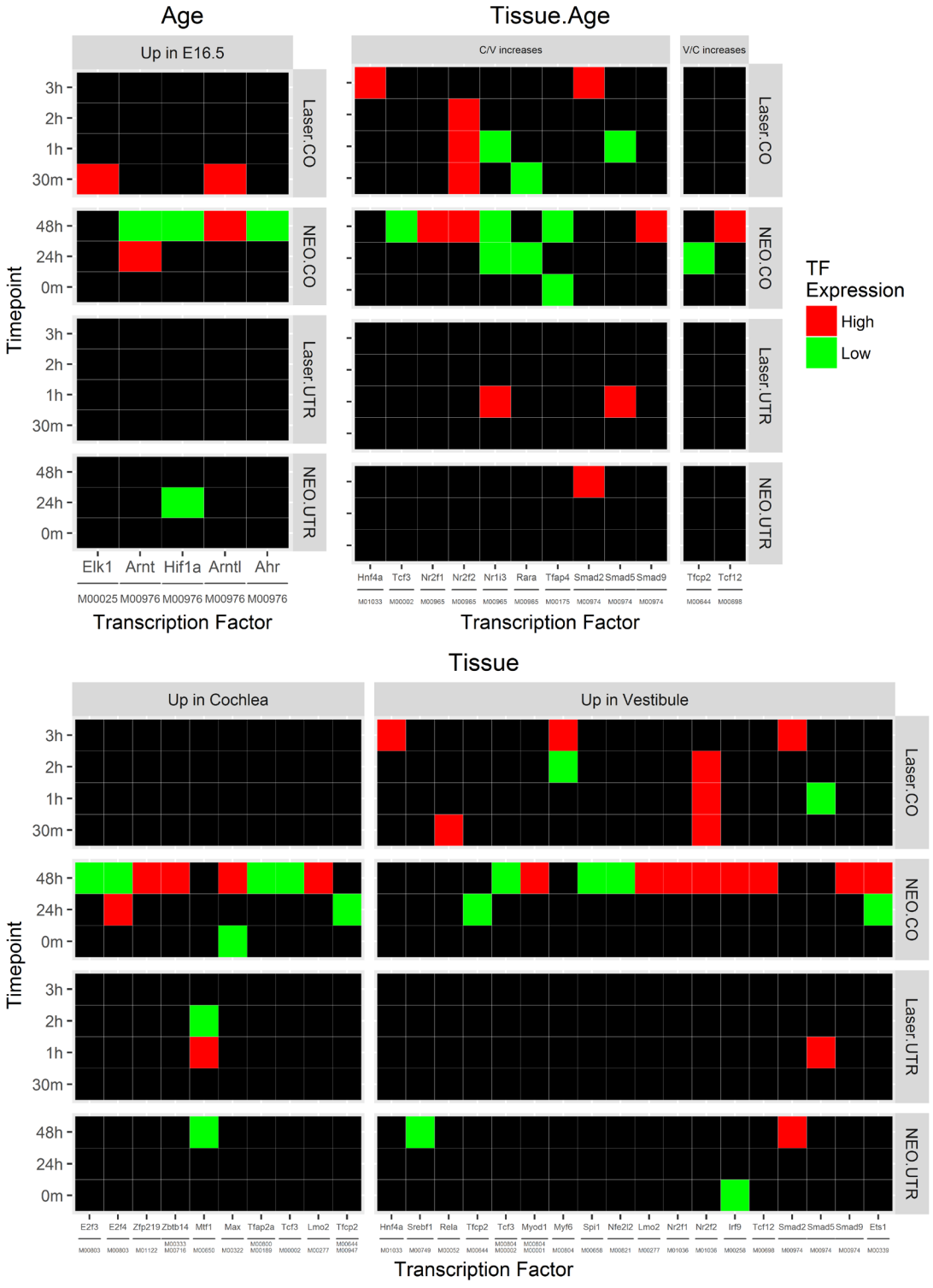

Figure 3.1-9 Transcription factors involved in avian HC regeneration. For the motifs enriched in the differentially expressed genes, we illustrated how their associated transcription factors (TFs) changed during a previous avian sensory epithelia regeneration experiment [66]. In that study, TF expression was measured in consecutive time points after the infliction of damage (either laser or Neomycin [NEO]) to an IE tissue (either cochlea [CO] or utricle [UTR]). Each tile represents a single measurement. The color indicates whether the TF expression was significantly high (red) or low (green) at this time point. The TFs and their associated TRANSFAC

motifs [105] are indicated on the x-axis. The subset of relevant TFs was determined separately for each comparison: between ages (upper left), tissues (lower), and the interaction of age and tissue (upper right).

3.1.5.1 Expression change with age

In the set of genes up-regulated at E16.5, we see enrichments of binding sites for the motifs: Elk-1, Nrf-1, E2F-1, E2F, NF-Y, and AHRHIF (Table 3.1-6). The subset Elk-1, Nrf-1, NF-Y, E2F-1, and some TFs associated with the motifs AHRHIF (*Arnt* and *AhR*) are up-regulated at E16.5 (Table 3.1-7), in concordance with the change of expression of their regulated genes. *Hif1a*, also associated with AHRHIF, is up-regulated at P0, suggesting that the up-regulation of *Arnt* controlled genes, is achieved by an increase in the formation of the heterodimer Arnt:AhR and not Arnt:Hif1a [106].

ELK1 and TFs associated with AHRHIF are changing their expression during regeneration. In [66], the expression of *ELK1* increased 30 min after wounding cochlear HCs with laser, marking an early signaling event that occurs after epithelial damage. As for the expression of the AHRHIF TFs, *ARNT* increased 24 hrs after exposing cochlear HCs to neomycin, and then by 48 hrs it decreased along with *HIF1A* and *AHR*. These time points reflect a change of expression in the SCs [66]. *ARNT* transient increase during regeneration resembles its transient expression pattern during normal IE development, where it expresses between E13 to E17 in mouse cochlear epithelial cells [107]. The three TFs are known to mediate tissue damage caused by a different toxic compound (TCDD [108]).

In the set of genes up-regulated at P0, we do not see enrichments of binding sites.

Table 3.1-6 Motifs enriched in genes differentially expressed between ages. Enrichments for motif binding sites found in genes up-regulated at E16.5. For each motif, we provide the TRANSFAC motif name [105], the number of genes with the motif in their promoter (#genes), the significance of the enrichment, and the ratio of the fraction of genes with the motif in their promoter in the set to the analogous fraction in the background (Enrichment factor). Notably, no enrichments were found in genes up-regulated at P0.

Set	Enriched with	#genes	p-Value	Enrichment factor
Up in E16.5	M00025[Elk-1]	1542	4.80E-13	1.152
	M00652[Nrf-1]	2195	6.77E-08	1.085

	M00940[E2F-1]	717	1.56E-05	1.136
	M00024[E2F]	423	1.30E-02	1.152
	M00287[NF-Y]	1529	1.80E-02	1.058
	M00976[AHRHIF]	1659	2.20E-02	1.068

Table 3.1-7 Levels of transcription factors affecting expression change with age. Each motif in **6** was associated with one or more TFs. For each such TF we provide the average expression across samples (in log counts per million [CPM]), the logarithm of the fold-change of its expression between conditions (P0/E16.5), the q-value for testing of differential expression of the TF between conditions, and an indicator for whether the TF is differentially expressed (DE) using the threshold q-value≤0.05.

Set	Motif	TF	Log CPM	Log FC	FDR	DE
Up in E16.5	M00025[Elk-1]	Elk1	5.43	-0.23	1.95E-02	TRUE
	M00652[Nrf-1]	Nrf1	5.41	-0.47	3.38E-05	TRUE
	M00940[E2F-1]	E2f1	3.68	-1.37	1.38E-32	TRUE
	M00024[E2F]	E2f1	3.68	-1.37	1.38E-32	TRUE
	M00287[NF-Y]	Nfya	5.93	-1.50	6.54E-35	TRUE
	M00976[AHRHIF]	Arnt	6.44	-0.38	6.46E-05	TRUE
	M00976[AHRHIF]	Hif1a	7.85	0.45	3.40E-07	TRUE
	M00976[AHRHIF]	Arntl	4.78	0.41	2.45E-04	TRUE
	M00976[AHRHIF]	Ahr	5.95	-0.46	3.57E-04	TRUE

3.1.5.2 Expression change between tissues

In the set of genes up-regulated in the cochlea, we see enrichments of binding sites for the motifs: HIC1, E2F, ZNF219, ZF5, UF1H3BETA, MOVO-B, MAZ, VDR, MAZR, MTF-1, c-Myc:Max, AP-2, CAC-binding protein, ETF, E47, Lmo2 complex, RREB-1, LBP-1, CP2/LBP-1c/LSF, and Spz1 (Table 3.1-8, green). TFs associated with E2F, ZF5, and MAZ are significantly up-regulated in the cochlea, while TFs associated with MOVO-B, VDR, and Lmo2 complex are up-regulated in the vestibule (Table 3.1-9, green). TFs associated with 10 of these 20 motifs (LBP-1, Lmo2 complex, E47, E2F, ZNF219, ZF5, MTF-1, c-Myc:Max, AP-2, CP2/LBP-1c/LSF) change their expression during the regeneration of IE sensory epithelia [66].

We will focus on the overlap of the lists. The E2F enrichment marks a higher proliferation rate in the cochlea at the relevant period of development. Given the role of this TF family in inducing proliferation, their involvement in HC regeneration is not surprising, and is

currently in active research [31]. *ZF5* is known mainly as a repressor of transcription, specifically regulating cell cycle progression (through *c-myc* [109]) and cognitive development (through *FMR1* [110]). Thus, it is unexpected to see its expression up-regulated in the cochlea, where its targets are up-regulated. This might indicate that *ZF5* has an additional activating role, or that another TF is activating the transcription these targets, and *ZF5* is up-regulated as part of a negative feedback loop. In avian HC regeneration *ZF5* cochlear expression increases in late recovery from neomycin damage, suggesting it has a role in cochlear HC differentiation. Another TF with the same pattern of expression during HC regeneration is *LMO2*. Enrichments for *LMO2* binding sites were found both in the cochlear and the vestibular up-regulated genes. While the results of the regeneration experiment support a function for *LMO2* in the cochlea, the expression of the TF in our experiment is higher in the vestibule. A possible explanation to this duality is that the partners with which LMO2 interacts might be different in the cochlea and the vestibule, and thus a different subset of genes is increased in each. Lmo2 complex typically contains a single GATA factor and a single TAL1/E47 heterodimer, but the GATA factor can be replaced for an additional TAL1/E47 heterodimer, changing the set of regulated genes [111]. As GATA2 is up-regulated in the vestibule, and TAL1 is up-regulated in the cochlea (DE q-values=7.32×10^{-18}, 5.29×10^{-7}, respectively), the complexes formed in each tissue might differ in composition.

In the set of genes up-regulated in the vestibule, we see enrichments of binding sites for the motifs: HNF4, SREBP-1, NF-1, PEA3, TEF-1, AP-2rep, NF-kappaB (p65), LBP-1, LUN-1, E2A, PU.1, MyoD, Nrf2, Lmo2 complex, COUPTF, ISRE, HEB, E47, SMAD, AML-1a, and c-Ets-1 (Table 3.1-8, red). TFs associated with five motifs (TEF-1, PU.1, Nrf2, Lmo2 complex, ISRE) are significantly up-regulated in the vestibule, while TFs associated with four other motifs (Etv4, COUPTF, most SMADs, AML-1a) are up-regulated in the cochlea (Table 3.1-9, red). TFs associated with 15 of these 21 motifs (HNF4, SREBP-1, NF-kappaB (p65), LBP-1, E2A, PU.1,

MyoD, Nrf2, Lmo2 complex, COUPTF, ISRE, HEB, E47, SMAD, c-Ets-1) change their expression during the regeneration experiment [66].

As noted, *SPI1* [PU.1] and *NFE2L2* [Nrf2] are associated with TFs that are up-regulated in the vestibule, supporting their role as inducers of transcription. For both, cochlear expression decreases in late (48h) recovery from neomycin, suggesting their repression is needed for proper differentiation of SCs to cochlear cells. *SPI1* is involved in hematopoietic development and induces proliferation of immune cells [112]; therefore it might up-regulate the immune functions that are enriched in the vestibule. Similarly, *NFE2L2* can up-regulate functions related to stress response, and specifically to antioxidant defense [113]. The expression pattern of *Nr2f1* and *Nr2f2* associated with the COUPTF motif fit their role as repressors of transcription, as they are down-regulated in the vestibule, but the motif is enriched in the genes up-regulated in the vestibule. Following laser damage, the expression of *NR2F2* increases in the cochlea for three hours. An increase in cochlear expression is also evident in late (48h) recovery from neomycin. *Nr2f2* is known to work as a repressor of myogenesis, inhibiting *MyoD* [114], another TF whose targets are up-regulated in the vestibule. Our data suggests that their repressive effect might have a role in cochlea development.

SMADs are intracellular proteins that transduce extracellular signals from transforming growth factor beta (TGF-β) ligands to the nucleus where they active downstream gene transcription [115]. As previously mentioned, TGF-β signaling is an enriched function in the cochlea. Nevertheless, according to this analysis, the downstream targets of this pathway are enriched in the vestibule. In order to settle this controversy, we examined the expression levels of individual SMADs. Most receptor-regulated SMADs (R-SMADs) are up-regulated in the cochlea (*Smad1*, *Smad2*, *Smad5*, *Smad9*), fitting the hypothesis of higher TGF-β activity in the cochlea. However, inhibitors of this signaling pathway (*Smad6*, *Smad7*) are also up-regulated in the cochlea, and with relatively high FCs (1.9 and 1.6, respectively). The

inhibition they induce lower the transcription of the downstream genes in the cochlea compared to the vestibule. The story become more complex when examining the two intracellular pathways SMADs are involved in. The R-SMADS *Smad2* and *Smad3* mediate the response to TGF-β ligands, which participate in the regulation of IE development by retinoic acid [116]. *Smad2* is up-regulated in the cochlea, and *Smad3* is up-regulated in the vestibule. In the regeneration experiment, *SMAD2*'s vestibular expression increases in late response to neomycin damage in the utricle, emphasizing the importance of TGF-β signaling for vestibular differentiation. In a different pathway, the R-SMADS *Smad1*, *Smad5*, and *Smad9* mediate the response to bone morphogenetic proteins (BPMs), which are involved in generation of IE sensory epithelia [117], as well as chondrogenesis [118]. All three are up-regulated in the cochlea, with *Smad9* showing a very impressive FC of 3.4. *SMAD9* also increases in response to late neomycin damage in the cochlea. This, together with its high cochlear levels, implies it has a role in cochlear differentiation.

Table 3.1-8 Motifs enriched in genes differentially expressed between tissues. Enrichments for motif binding sites found in genes up-regulated in the cochlea of in the vestibule (marked green and red, respectively, in the 'Set' column). See caption in **Table 3.1-6** for the structure of the table.

Set	Enriched with	#genes	p-Value	Enrichment factor
Up in Cochlea	M01072 [HIC1]	903	9.38E-13	1.181
	M00803 [E2F]	1146	4.58E-10	1.098
	M01122 [ZNF219]	822	1.04E-09	1.176
	M00716 [ZF5]	977	2.23E-09	1.141
	M01068 [UF1H3BETA]	1122	1.61E-08	1.113
	M01104 [MOVO-B]	681	1.78E-08	1.22
	M00649 [MAZ]	690	2.02E-08	1.218
	M00444 [VDR]	486	1.29E-05	1.187
	M00491 [MAZR]	622	1.05E-04	1.186
	M00650 [MTF-1]	457	2.56E-04	1.203
	M00322 [c-Myc:Max]	613	2.88E-04	1.184
	M00189 [AP-2]	806	9.39E-04	1.106
	M00720 [CAC-binding_protein]	655	1.00E-03	1.163
	M00333 [ZF5]	528	2.00E-03	1.198
	M00695 [ETF]	763	3.00E-03	1.102

	M00002 [E47]	368	3.00E-03	1.203
	M00277 [Lmo2_complex]	349	7.00E-03	1.202
	M00800 [AP-2]	763	1.50E-02	1.115
	M00257 [RREB-1]	590	1.80E-02	1.094
	M00644 [LBP-1]	602	2.50E-02	1.119
	M00947 [CP2/LBP-1c/LSF]	402	4.80E-02	1.195
	M00446 [Spz1]	350	8.10E-02	1.212
Up in vestibule	M01033 [HNF4]	2114	1.51E-08	1.081
	M00749 [SREBP-1]	1033	3.43E-06	1.131
	M00193 [NF-1]	417	6.74E-06	1.253
	M00655 [PEA3]	1308	1.29E-04	1.089
	M00704 [TEF-1]	1639	3.65E-04	1.064
	M00468 [AP-2rep]	871	4.76E-04	1.126
	M00052 [NF-kappaB_(p65)]	277	2.00E-03	1.253
	M00644 [LBP-1]	1071	3.00E-03	1.124
	M00480 [LUN-1]	409	4.00E-03	1.194
	M00804 [E2A]	588	4.00E-03	1.148
	M00658 [PU.1]	541	8.00E-03	1.131
	M00001 [MyoD]	388	8.00E-03	1.194
	M00821 [Nrf2]	365	9.00E-03	1.193
	M00277 [Lmo2 complex]	588	1.10E-02	1.143
	M01036 [COUPTF]	713	1.20E-02	1.118
	M00258 [ISRE]	430	1.50E-02	1.132
	M00698 [HEB]	332	1.70E-02	1.201
	M00002 [E47]	614	2.40E-02	1.133
	M00974 [SMAD]	322	6.00E-02	1.183
	M00271 [AML-1a]	1639	8.30E-02	1.047
	M00339 [c-Ets-1]	682	9.00E-02	1.119

Table 3.1-9 Levels of transcription factors affecting expression change between tissues. Each motif in **6** was associated with one or more TFs. See caption in **Table 3.1-7** regarding which data is provided for each TF. Here, the fold-changes are measured between the cochlea and the vestibule.

Set	Motif	TF	Log CPM	Log FC	FDR	DE
Up in Cochlea	M01072 [HIC1]	Hic1	4.42	0.11	5.89E-01	FALSE
	M00803 [E2F]	E2f1	3.68	-0.11	4.54E-01	FALSE
	M00803 [E2F]	E2f3	5.58	-0.16	2.10E-01	FALSE
	M00803 [E2F]	E2f4	5.70	0.01	9.56E-01	FALSE
	M00803 [E2F]	Tfdp1	6.54	0.24	4.39E-02	TRUE
	M01122 [ZNF219]	Zfp219	7.19	-0.19	1.24E-01	FALSE
	M00333 [ZF5]	Zbtb14	5.53	0.30	2.06E-03	TRUE

	M01068 [UF1H3BETA]	?	NA	NA	NA	NA
	M01104 [MOVO-B]	Ovol2	2.08	-0.31	4.68E-02	TRUE
	M00649 [MAZ]	Maz	7.81	0.21	3.86E-02	TRUE
	M00444 [VDR]	Vdr	3.57	-1.54	7.30E-32	TRUE
	M00491 [MAZR]	Patz1	6.79	0.09	4.09E-01	FALSE
	M00650 [MTF-1]	Mtf1	4.56	-0.16	1.78E-01	FALSE
	M00322 [c-Myc:Max]	Myc	4.72	0.14	2.49E-01	FALSE
	M00322 [c-Myc:Max]	Max	6.10	0.06	6.35E-01	FALSE
	M00800 [AP-2]	Tfap2a	0.35	0.18	4.96E-01	FALSE
	M00720 [CAC-binding protein]	?	NA	NA	NA	NA
	M00716 [ZF5]	Zbtb14	5.53	0.30	2.06E-03	TRUE
	M00695 [ETF]	Tead2	7.74	0.16	2.00E-01	FALSE
	M00002 [E47]	Tcf3	7.76	-0.10	4.25E-01	FALSE
	M00277 [Lmo2 complex]	Lmo2	3.88	-0.88	2.19E-11	TRUE
	M00189 [AP-2]	Tfap2a	0.35	0.18	4.96E-01	FALSE
	M00257 [RREB-1]	Rreb1	5.75	-0.21	3.34E-01	FALSE
	M00644 [LBP-1]	Ubp1	6.54	-0.21	3.37E-01	FALSE
	M00644 [LBP-1]	Tfcp2	5.66	-0.21	2.41E-01	FALSE
	M00947 [CP2/LBP-1c/LSF]	Tfcp2	5.66	-0.21	2.41E-01	FALSE
	M00446 [Spz1]	Spz1	NA	NA	NA	NA
Up in vestibule	M01033 [HNF4]	Hnf4a	NA	NA	NA	NA
	M00749 [SREBP-1]	Srebf1	7.47	-0.19	8.82E-02	FALSE
	M00193 [NF-1]	Nf1	6.70	0.06	7.38E-01	FALSE
	M00655 [PEA3]	Etv4	5.23	0.26	2.01E-02	TRUE
	M00704 [TEF-1]	Tead1	7.20	-0.52	2.44E-07	TRUE
	M00468 [AP-2rep]	Klf12	3.79	-0.21	2.39E-01	FALSE
	M00052 [NF-kappaB (p65)]	Rela	6.15	-0.04	7.64E-01	FALSE
	M00644 [LBP-1]	Ubp1	6.54	-0.21	3.37E-01	FALSE
	M00644 [LBP-1]	Tfcp2	5.66	-0.21	2.41E-01	FALSE
	M00480 [LUN-1]	Topors	5.81	0.07	5.97E-01	FALSE
	M00804 [E2A]	Tcf3	7.76	-0.10	4.25E-01	FALSE
	M00804 [E2A]	Myog	NA	NA	NA	NA
	M00804 [E2A]	Myod1	NA	NA	NA	NA
	M00804 [E2A]	Myf6	NA	NA	NA	NA
	M00658 [PU.1]	Spi1	2.66	-1.70	1.42E-28	TRUE
	M00001 [MyoD]	Myod1	NA	NA	NA	NA
	M00821 [Nrf2]	Nfe2l2	5.31	-0.39	1.68E-02	TRUE
	M00821 [Nrf2]	Mafk	3.55	-0.37	2.59E-03	TRUE
	M00277 [Lmo2 complex]	Lmo2	3.88	-0.88	2.19E-11	TRUE
	M01036 [COUPTF]	Nr2f1	7.02	2.38	7.08E-54	TRUE
	M01036 [COUPTF]	Nr2f2	6.08	1.21	5.82E-22	TRUE

	M00258 [ISRE]	Irf9	4.97	-0.39	1.22E-03	TRUE
	M00698 [HEB]	Tcf12	8.08	0.13	3.79E-01	FALSE
	M00002 [E47]	Tcf3	7.76	-0.10	4.25E-01	FALSE
	M00974 [SMAD]	Smad1	5.27	0.28	1.11E-02	TRUE
	M00974 [SMAD]	Smad2	5.84	0.28	6.90E-03	TRUE
	M00974 [SMAD]	Smad3	7.22	-0.22	4.16E-02	TRUE
	M00974 [SMAD]	Smad4	6.84	0.19	6.05E-02	FALSE
	M00974 [SMAD]	Smad5	7.88	0.27	1.55E-02	TRUE
	M00974 [SMAD]	Smad6	4.63	0.98	2.89E-23	TRUE
	M00974 [SMAD]	Smad7	4.25	0.65	1.32E-06	TRUE
	M00974 [SMAD]	Smad9	3.62	1.77	9.48E-26	TRUE
	M00271 [AML-1a]	Runx1	3.33	1.52	1.59E-23	TRUE
	M00339 [c-Ets-1]	Ets1	6.56	0.18	2.22E-01	FALSE

3.1.5.3 Transcription factors affecting expression ratio change with age

In the set of genes for which the cochlea to vestibule expression ratio increases with age $\left(\frac{Cochlea}{Vestibule}\uparrow\right)$, we see enrichments of binding sites for the motifs: HNF4, E47, a group of nuclear receptors (LXR, PXR, CAR, COUP, RAR), AP-4, and SMAD (Table 3.1-10, green). Out of the associated TFs, the expression ratio of *Nr2f1*, a COUP TF, significantly increases in the same direction as its targets (Table 3.1-11, green); this might have a positive downstream effect on retinoic acid receptor (RAR) signaling [119]. TFs associated with all motifs change their expression during the regeneration of IE sensory epithelia [66].

Retinoid signaling is critical during IE embryonic development, as well as in postnatal maintenance of its function [120]. Both vitamin A deficiency and intake of excess retinoic acid (RA) during pregnancy resulted in malformations in ear development. In rodents, *in utero* exposure of fetuses to RA results in a reduction of the semicircular canals and of the cochlea. Key components in retinoid signaling show spatiotemporal expression patterns, and the interactions that excess RA interferes with are dependent on the developmental stage. KEGG enrichment of our DE genes shows that metabolism of RA is higher in the vestibule and at P0. Taken together with the motif enrichment, we deduce that retinoid signaling is important to

both cochlear and vestibular development, with its role in the cochlea becoming more prominent in the period between E16.5 and P0. In the HC regeneration experiment, the cochlear expression of the retinoid receptor Rara decreases 24h after neomycin damage, and later on by 48h, *NR2F1* expression increases [66]. This later increase might mimic the increase in retinoid signaling seen in normal development.

In the set of genes for which the vestibule to cochlea expression ratio increases with age $\left(\frac{Vestibule}{Cochlea}\uparrow\right)$, we see enrichment of binding sites for: AML-1a, LEF1, LBP-1, HEB, and POU6F1 (Table 3.1-10, red). The expression ratio of *Runx1* [AML-1a] significantly increases in the same direction as its targets (Table 3.1-11, red). TFs associated with LBP-1 and HEB change their expression during the regeneration of IE sensory epithelia.

Table 3.1-10 Motifs enriched in genes for which the cochlea to vestibule expression ratio changes with age. Enrichments for motif binding sites found in genes for which the cochlea to vestibule expression ratio is increasing or decreasing with age (marked green and red, respectively, in the 'Set' column). See caption in **Table 3.1-6** for the structure of the table.

Set	Enriched with	#genes	p-Value	Enrichment factor
$\frac{C}{V}\uparrow$	M01033 [HNF4]	398	1.00E-02	1.136
	M00002 [E47]	131	4.90E-02	1.349
	M00965 [LXR, PXR, CAR, COUP, RAR]	99	5.50E-02	1.363
	M00175 [AP-4]	97	6.10E-02	1.482
	M00974 [SMAD]	74	8.80E-02	1.518
$\frac{C}{V}\downarrow$	M00271 [AML-1a]	528	4.98E-04	1.163
	M00805 [LEF1]	567	1.00E-02	1.106
	M00644 [LBP-1]	336	7.50E-02	1.216
	M00698 [HEB]	107	9.50E-02	1.335
	M00465 [POU6F1]	114	9.70E-02	1.274

Table 3.1-11 Levels of transcription factors affecting expression ratio change with age. Each motif in **Table 3.1-10** was associated with one or more TFs. See caption in **Table 3.1-7** regarding which data is provided for each TF. Here, the fold-changes are measured between the cochlea to vestibule expression ratios in P0 and E16.5.

Set	Motif	TF	Log CPM	Log FC	FDR	DE
$\frac{C}{V}\uparrow$	M01033 [HNF4]	Hnf4a	NA	NA	NA	FALSE
	M00002 [E47]	Tcf3	7.76	-0.05	7.92E-01	FALSE

	M00965 [LXR, PXR, CAR, COUP, RAR]	Nr2f1	7.02	0.50	7.93E-03	TRUE
	M00965 [LXR, PXR, CAR, COUP, RAR]	Nr2f2	6.08	0.17	3.84E-01	FALSE
	M00965 [LXR, PXR, CAR, COUP, RAR]	Nr1i2	NA	NA	NA	FALSE
	M00965 [LXR, PXR, CAR, COUP, RAR]	Nr1i3	NA	NA	NA	FALSE
	M00965 [LXR, PXR, CAR, COUP, RAR]	Rara	5.77	0.07	7.62E-01	FALSE
	M00965 [LXR, PXR, CAR, COUP, RAR]	Rarb	5.72	0.20	1.39E-01	FALSE
	M00175 [AP-4]	Tfap4	3.69	-0.07	7.17E-01	FALSE
	M00974 [SMAD]	Smad1	5.27	-0.19	1.92E-01	FALSE
	M00974 [SMAD]	Smad2	5.84	0.08	6.46E-01	FALSE
	M00974 [SMAD]	Smad3	7.22	0.18	2.09E-01	FALSE
	M00974 [SMAD]	Smad4	6.84	0.09	5.36E-01	FALSE
	M00974 [SMAD]	Smad5	7.88	0.19	2.06E-01	FALSE
	M00974 [SMAD]	Smad6	4.63	0.09	6.13E-01	FALSE
	M00974 [SMAD]	Smad7	4.25	0.14	5.38E-01	FALSE
	M00974 [SMAD]	Smad9	3.62	-0.14	6.41E-01	FALSE
$\frac{C}{V}\downarrow$	M00271 [AML-1a]	Runx1	3.33	-0.67	1.47E-04	TRUE
	M00805 [LEF1]	Lef1	5.77	0.13	5.13E-01	FALSE
	M00644 [LBP-1]	Ubp1	6.54	-0.42	8.44E-02	FALSE
	M00644 [LBP-1]	Tfcp2	5.66	-0.23	3.17E-01	FALSE
	M00698 [HEB]	Tcf12	8.08	-0.14	4.84E-01	FALSE
	M00465 [POU6F1]	Pou6f1	5.61	0.00	9.99E-01	FALSE
	M00465 [POU6F1]	Pou6f1	NA	NA	NA	FALSE

3.2 Protein and mRNA joint analysis

Previous examinations of mRNA-protein relationships were mainly performed in yeast and in cancer cell lines. Aiming to examine these associations in non-transformed cells and differentiated tissue samples, we analyzed four different paired datasets of mRNA and protein. For the first dataset we generated proteomics and transcriptomic data from the cochlea and vestibule of mouse inner ear (dataset termed EAR). The three other datasets were publicly available: (i) multiple mouse tissues (termed MMT; RNA-seq [83] and proteomics [82]); (ii) primate lymphoblastoid cells (PRIMATE; [51]); and (iii) a panel of human cancer cell lines (NCI60; transcription microarrays [87] and proteomics [88]). The

results obtained for the NCI60 dataset were compared with those obtained for datasets of non-transformed cells.

The EAR RNA-seq analysis identified 39,178 Ensembl genes (including non-coding genes and pseudogenes), 14,722 of which have at least one read per million in three or more of the samples and were included in the analysis. MS analysis identified 7244 proteins. 6832 genes were common between the two tissues.

The MMT dataset contains mRNA and protein levels taken from mouse tissues. In the proteomic data [82], the stable isotope labeling with amino acids in cell culture (SILAC) technique was used as an internal standard for relative quantification of proteins across 28 mouse tissues. We used five tissues that had both mRNA and protein data: brain, cerebellum, heart, kidney, and liver. There were three proteomic samples for brain (cortex, medulla, and midbrain) and two for kidney (cortex and medulla), and we weighted the samples' contribution by the volumes of the subregions to obtain the tissue protein levels. mRNA measurements had three replicates per tissue, and six for the brain.

The PRIMATE dataset included transcriptomics (RNA-seq) and proteomics (SILAC-based) data from lymphoblastoid cell lines (LCLs) derived from five human, five chimpanzees, and five rhesus macaques. The species is analogous in the subsequent analysis to the tissue. We downloaded the data from [51], and processed it as described in the article, to obtain expression levels of (orthologous) genes that have at least three measurements from each of the three species, for both mRNA (12,079 genes) and protein (3688 genes). 3394 genes were common between mRNA and protein.

NCI60 is a panel of 59 diverse human cancer cell lines. The type of cancer is analogous in the subsequent analysis to the tissue. We note that we do not necessarily expect to see the same phenomena in cancer cell lines as in healthy tissues, due to the pathological state of the

tissues, and as the cell lines of the same cancer are different samples and not real replicates as the healthy tissues. One manifestation of these differences is a lesser ability to separate NCI60 samples based on their origin, compared to the EAR and MMT datasets. Indeed, multi-dimensional scaling (MDS) plots show better separation of the latter datasets on both mRNA and protein levels, even between very similar tissues (Figure 3.2-1). Moreover, poor results were reported when hierarchical clustering was used to perform such a separation for breast, ovary, renal, and prostate cancers using proteomic data [88].

We refer to the tissue type (in EAR and MMT), species (in PRIMATE), or cancer type (in NCI60) as a *group*. We refer to samples of the same group as *replicates*. We refer to mRNA and protein as *domains*.

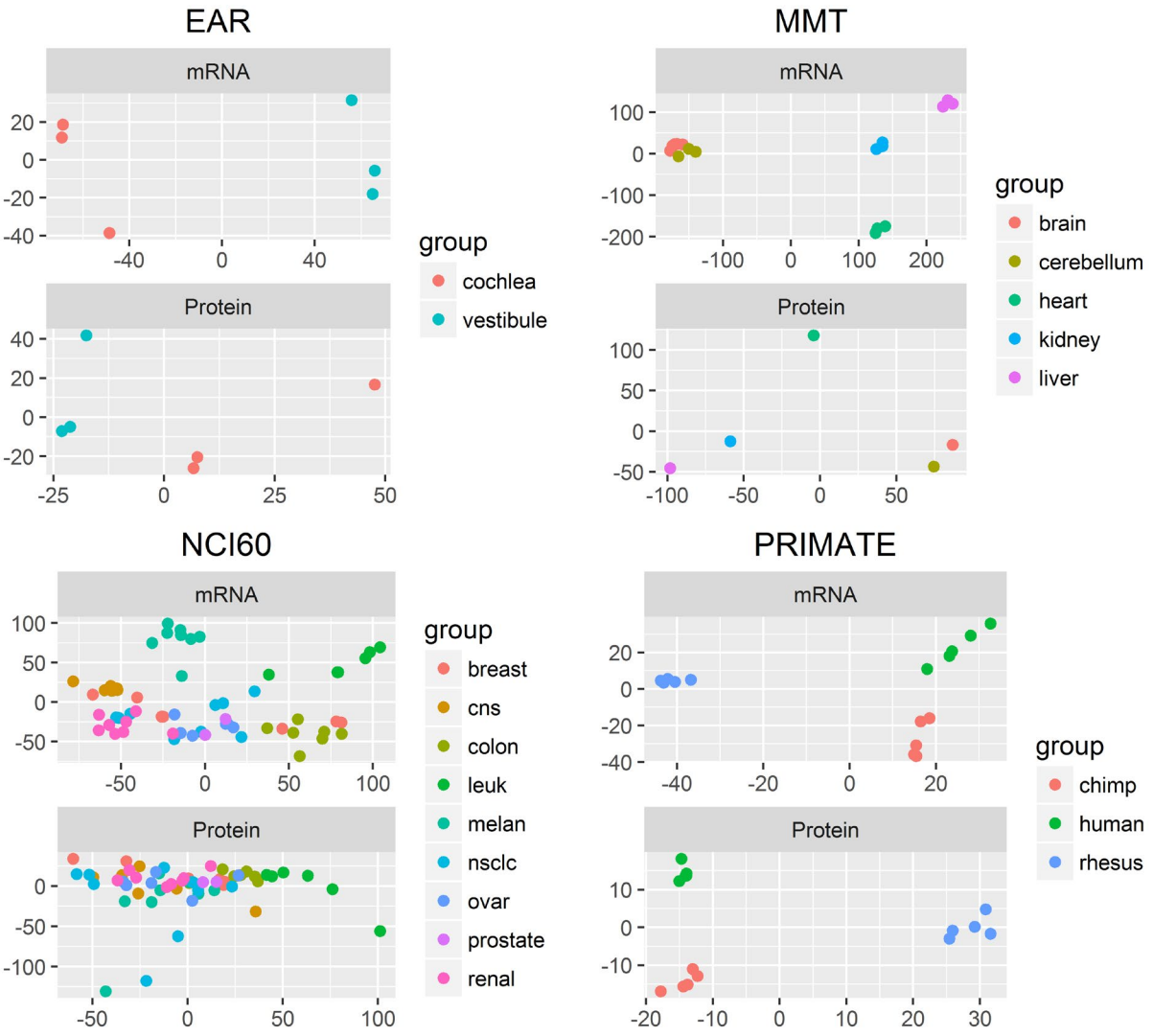

Figure 3.2-1 MDS plots comparing samples in the different datasets according to their mRNA or protein expression. The upper and lower figures for each dataset are the MDS plots according to the mRNA or protein expression, respectively. The x- and y-axis are the first and second coordinates, respectively. The samples are colored by their group. The groups are clearly separable by both mRNA and protein, in all but the NCI60 dataset, where the separation becomes less clear. The MDS plots are based on all the data, and not just the portion of genes for which we find expression in both domains.

3.2.1 Comparison of protocols used to collect mRNA and protein data

Collecting the mRNA and the proteomic data for a dataset from two different published articles, raises the concern that different protocols for sample preparation and source animals will lead to improper results. While this is not the case for the EAR and the PRIMATE datasets, for which the protocols were similar to allow such comparison, the NCI60 and the MMT datasets should be carefully analyzed. In the case of NCI60, this concern is alleviated by

the genetic identity of the cell lines used in the two experiments, and by the fact that a comparison of proteomic and transcriptomic data was previously made and showed a significant degree of correlation between the two [88]. As for MMT, we included only adult mice samples in the analysis, carefully choosing which samples to include, though not all RNA samples were of the same species as the protein data (C57BL/6J). Also, we attentively handled tissues for which we had to merge several protein samples from different sub-regions of a tissue, to create a sample comparable to the RNA data. For the EAR dataset, we stress the use of the same mouse species and the identical growth conditions shared between the mice from which the mRNA material was produced form, and those from which the protein was produced from. This property, together with the assumption of a steady-state in mRNA and protein abundances in non-proliferating tissues, assure that the bias caused by using different mice is minor.

3.2.2 Protein levels are more conserved than mRNA levels

mRNA and protein levels were $\log_2$-transformed, and averaged across all samples from the same group, disregarding missing values. A comparison of the proteomic and transcriptomic data showed, in agreement with previous studies [45], that the overall dynamic range of mRNA is significantly lower than protein, as marked by a higher variability in protein expression compared with mRNA in all datasets (Figure 3.2-2).

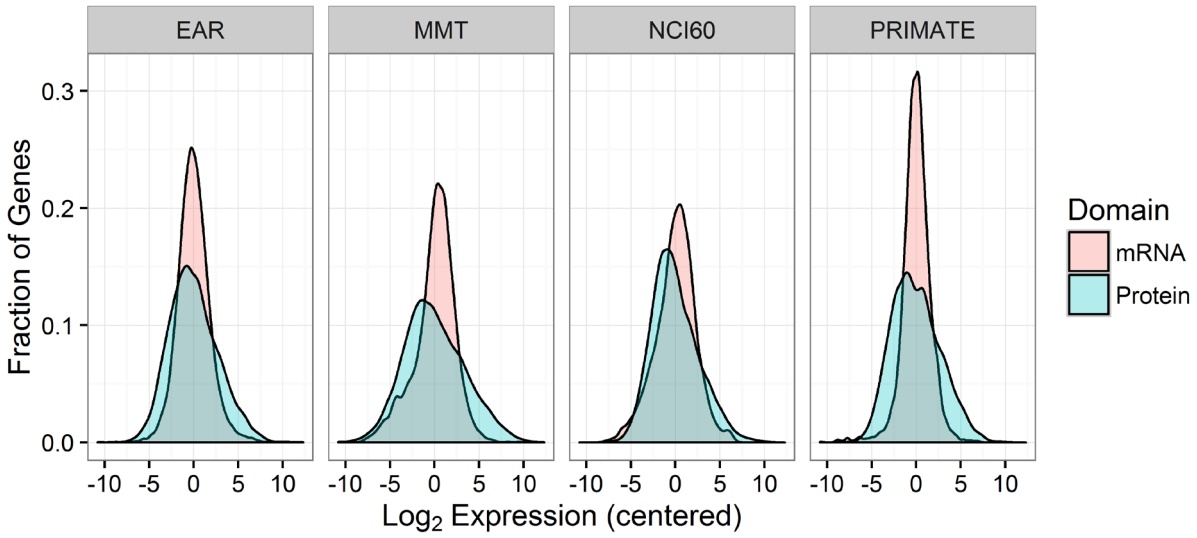

Figure 3.2-2 Dynamic range of expression in mRNA and in protein. The absolute levels of expression in mRNA (red) and protein (blue) are displayed in a density plot. The levels were centered around 0 to allow comparison of the dynamic range between mRNA and protein. For measures of variability (standard deviations and interquartile ranges), see Table S2 in [55].

We calculated protein-mRNA correlations for each group. The average correlations between the two layers were 0.58, 0.44, 0.42, and 0.42 for the EAR, MMT, NCI60 and PRIMATE datasets, respectively, similar to the mRNA-protein correlations reported in literature [39]. Then, we calculated correlations between pairs of groups for mRNA and protein separately. We observed that in all datasets, all the protein-protein and the mRNA-mRNA correlations between groups were higher than the protein-mRNA correlations within each group (Figure 3.2-3). This last trend was somewhat weaker in the MMT dataset, which includes less similar tissues. Using either Pearson correlation or Spearman's rank correlations produced similar results, hence we employed the former throughout.

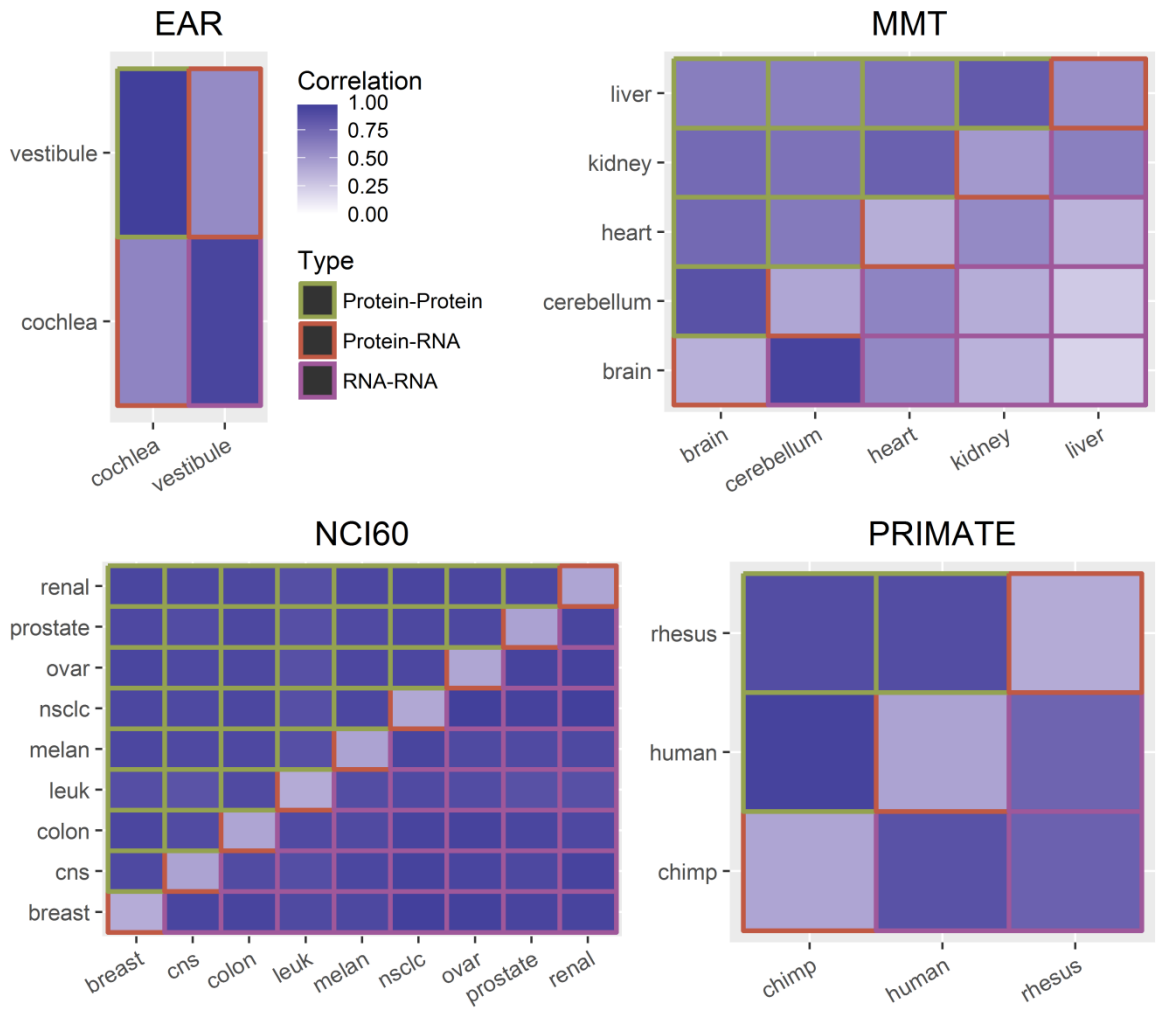

Figure 3.2-3 Protein and mRNA correlations between groups for different datasets. Each subfigure describes the Pearson's correlation (r) between expression levels in one dataset. The upper and lower triangles contain the protein-protein and mRNA-mRNA correlations between pairs of groups, respectively. The diagonal contains the protein-mRNA correlations within each group. Darker color corresponds to higher correlation. The correlations are not Spearman corrected because the correction cannot be applied on the MMT dataset, and this figure is intended for comparison between datasets.

Next, to allow a fair comparison of the correlation between group pairs in each dataset, we had to account for some of the platform differences between RNA-seq and MS, which manifest in higher correlation between replicates of RNA-seq (Figure 3.2-4). We thus applied the Spearman's correction in our calculations, except for MMT where it was inapplicable. The corrected correlations are presented in Figure 3.2-5 (see correction example in Figure 2.4-1).

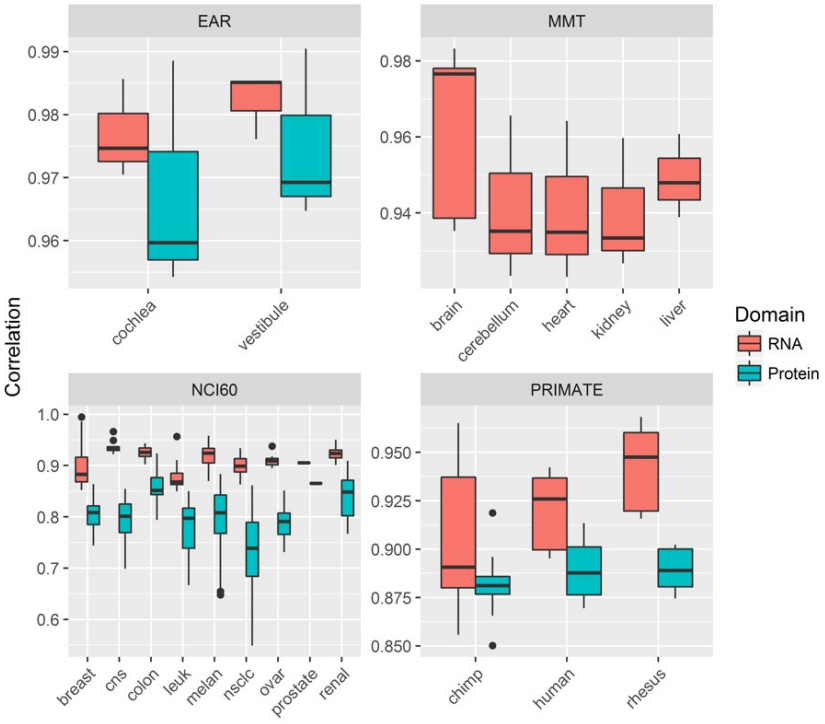

Figure 3.2-4 Correlation between replicates. By 'replicates' we mean samples from the same group. For each dataset we plotted the boxplots of the distribution of Pearson's correlation (r) between replicates in the mRNA levels (pink) and the protein levels (light blue) aggregated by group. For the MMT dataset we had no protein replicates, so only mRNA correlations are presented. Boxplots show median, a box for the middle 50% and whiskers to the largest and smallest values that are not classified as outliers. If the distance of an observation from the box is higher than 1.5 times the box size, it is classified as an outlier. The higher correlations between replicates in the mRNA domain are evident from this figure. This trend was confirmed by ANOVA testing in the two large datasets of NCI60 and PRIMATE (data not shown).

For the EAR dataset the correlation in the protein between the cochlea and the vestibule is higher than the correlation in the mRNA (0.97 versus 0.94). This is also the case for the PRIMATE dataset (3/3 pairs), the MMT dataset (9/10 pairs), and the NCI60 dataset (24/36). For the MMT and NCI60 datasets the protein correlations were significantly higher (*p–value*s=2.9×10^{-3} and 8.0×10^{-3} respectively, Wilcoxon signed-rank test). As the Spearman's correction was not applied on the MMT dataset, we cannot be certain that the higher protein correlations in that dataset are not an artifact caused by different levels of noise in the mRNA and protein domains. Specifically, a higher degree of noise in the mRNA measurements can cause such a bias. However, it is unlikely that the mRNA levels are noisier than the protein levels, as the opposite is true for the EAR and the NCI60 datasets (Figure 3.2-4), which were produced using similar measuring methods. In fact, the protein levels in the MMT dataset are probably even noisier, due to the lack of replicates in the protein, which prevented noise reduction by averaging.

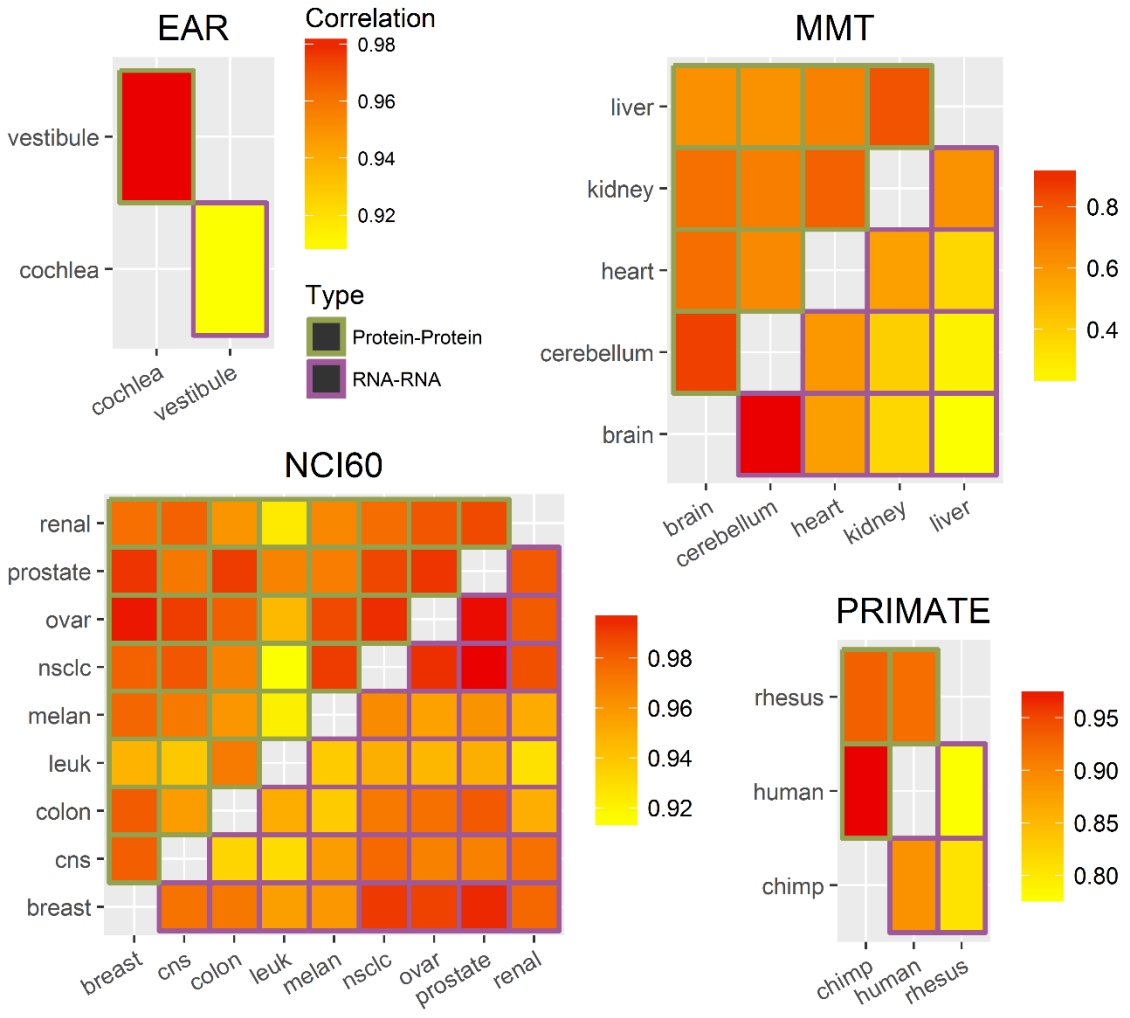

Figure 3.2-5 Protein and mRNA correlation between group pairs. Each subfigure describes the correlation between expression levels of different groups in one dataset. The upper and lower triangles show the protein-protein and mRNA-mRNA correlations between groups, respectively. Darker color corresponds to higher correlation. Pearson's correlation coefficients (r) were corrected using Spearman's method except in the MMT dataset (due to the lack of replicates in protein). See Figure **3.2-3** for intra group protein-mRNA correlations.

3.2.3 PTRs vary in a direction that reduces protein divergence

The higher correlation between pairs of groups in the protein domain suggests that changes in transcription between tissues are coupled to protein-level changes that exert opposite effects on the final protein level, hence producing higher similarity between groups. We call the phenomenon of reduced ("compressed") change in protein levels compared to the change in mRNA levels *buffering*. Spangenberg et al. showed this phenomenon in the initial phases of adipocyte differentiation of adipose-derived human mesenchymal stem cells, by comparing

differentiating cells at two time points [48]. Regressing the fold change (FC) of the protein levels to the FC of the mRNA levels on a log-log scale led to the observation of a slope lower than 1, or, in other words, range compression between protein FC and mRNA FC. They hypothesized that a trend of lower PTR with increasing mRNA levels is the cause.

To test this hypothesis on our data, for all pairs of groups in all datasets, we regressed $\log \mathrm{FC}_{\mathrm{mRNA}}$ on $\log \mathrm{FC}_{\mathrm{mRNA}}$ using a variant of major axis (MA) regression, and tested whether the slope is significantly different from 1. All slopes were significantly less than 1 for the EAR and PRIMATE datasets, and for all except one pair in the MMT dataset (see Figure 3.2-6 for examples). For the NCI60 and brain-cerebellum [MMT] the slopes were significantly higher than 1. When using ordinary least square (OLS) regression, all the slopes calculated were significantly less than 1 ($q\text{–}value \leq 0.01$), consistent with the aforementioned range compression phenomenon (discordance between the regression methods is demonstrated in Figure 2.4-2). However, MA regression is not sensitive to regression dilution bias, which can severely lower the estimate of the slope in OLS regression [121]. Using MA, it appears that the range compression is a common phenomenon for pairs of tissues, or species. For cell lines, an opposite phenomenon of range expansion occurs.

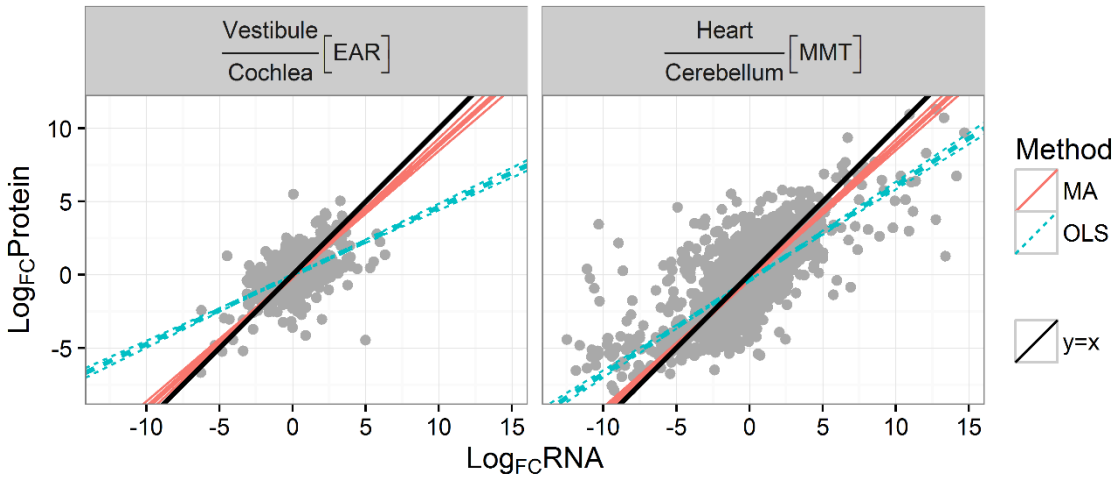

Figure 3.2-6 Examples of range compression. Comparing either the cochlea and the vestibule EAR tissues (left), or the heart and the cerebellum MMT tissues (right), the protein fold changes (y-axis) were regressed on the mRNA fold changes (x-axis). The fitted regression lines using ordinary least squares (OLS, red, solid) and major axis regression (MA, blue, dashed) were plotted, along with their 95 percent confidence interval (thinner lines). The black line is y=x. Both OLS and MA slopes are significantly lower than 1, suggesting range compression.

Next, we used a nonparametric approach to test whether genes that are up-regulated in one group versus the other in the mRNA domain will show lower PTR in that same group versus the other. If this hypothesis is correct, it can explain the compressed ratios in the non-cancerous datasets. We formulated two complementary testing approaches: A global test that considers all the genes ranked by their mRNA *differential expression (DE) values*, and a local test that focuses on those that are DE. Importantly, we separated the repeats on which PTR and DE values are computed in order to avoid bias in the significance evaluation. Figure 3.2-7 provides an example of the DE-PTR comparison in inner-ear tissues. The PTRs in the cochlea were plotted against the PTRs in the vestibule, with the genes DE between the tissues highlighted. We observe that genes up-regulated in one tissue tend to have higher PTRs in the other tissue. This property is tested by the local approach.

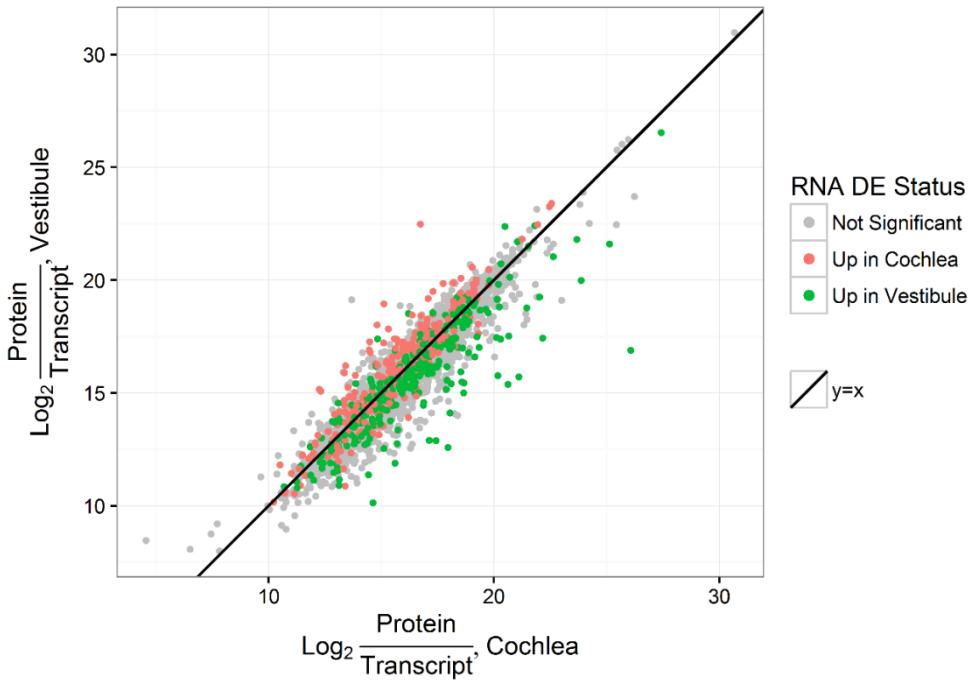

Figure 3.2-7 Protein-transcript ratio (PTR) and differential expression between two inner-ear tissues. The PTRs in the cochlea (x-axis) are plotted against the PTRs in the vestibule (y-axis), where the PTRs were calculated using mRNA data of samples SA623 and SA626 respectively. Marked in red are genes that are up-regulated in the cochlea, and in green are genes that are up-regulated in the vestibule (edgeR, q−value≤0.05). Samples SA623 and SA626 were excluded from the differential expression analysis. The black line is y=x. There is a clear tendency for the genes that are up-regulated in the cochlea (red points) to have higher PTR in the vestibule (be above the black line), and vice versa. Note that to emphasize the DE status, significant (colored) genes are drawn at the front and may occlude some non-significant ones.

The global tests were significant for all group pairs in the EAR, MMT, and PRIMATE datasets ($q\text{−}value \leq 0.01$). The results were in complete agreement with those of the local approach. The positive results support the buffering observation for all these datasets, and those of the local approach specifically indicate that within these datasets reduced protein expression changes have a major effect on the DE genes. For the NCI60 dataset, none of the pairs were significant, and all the correlations were very close to zero. Therefore, we cannot determine the presence of a compression or an amplification effect based on this approach. As mentioned before, the different cell lines have very similar expression profiles, and this might cause a low signal-to-noise ratio.

3.2.4 Predicting protein abundance from mRNA levels

Next, we examined whether we can predict protein levels based on the mRNA data. More specifically, the problem we aimed to solve was the prediction of protein levels in a group, when given the mRNA levels in that group, and matching mRNA and protein profiles from other groups. We compared three estimators all of which are trained on a subset of each dataset, and examined their ability to predict the protein level in the rest of the dataset.

The first estimator was built on the average PTR (APTR); the second estimator, which is fold change based (FCB), assumes a constant compression ratio of the fold changes between protein and RNA; the third infers the protein levels from the average protein (AP) levels in other tissues. AP and APTR also have a weighted version, which gives higher weight to the tissues with higher similarity, and FCB has a relaxed version (RFCB) that allowed for protein levels to change exponentially between groups, independent of change in mRNA. This accounts for differences between groups in the activity of the translational mechanisms and in protein stability.

In all datasets, the FCB and RFCB models achieved better results than the others (Figure 3.2-8). For all models, the weighted/relaxed versions achieved better results than their unweighted counterparts. The difference was very apparent for the MMT dataset, where the presence of two related tissues, brain and cerebellum, lowered the prediction error dramatically for those tissues; analysis of this dataset after the removal of one of the two still showed an advantage for the weighted versions, albeit smaller (data not shown). These findings support the use of a weighted estimator, which gives higher weights to tissues that are closer in their protein levels and PTRs.

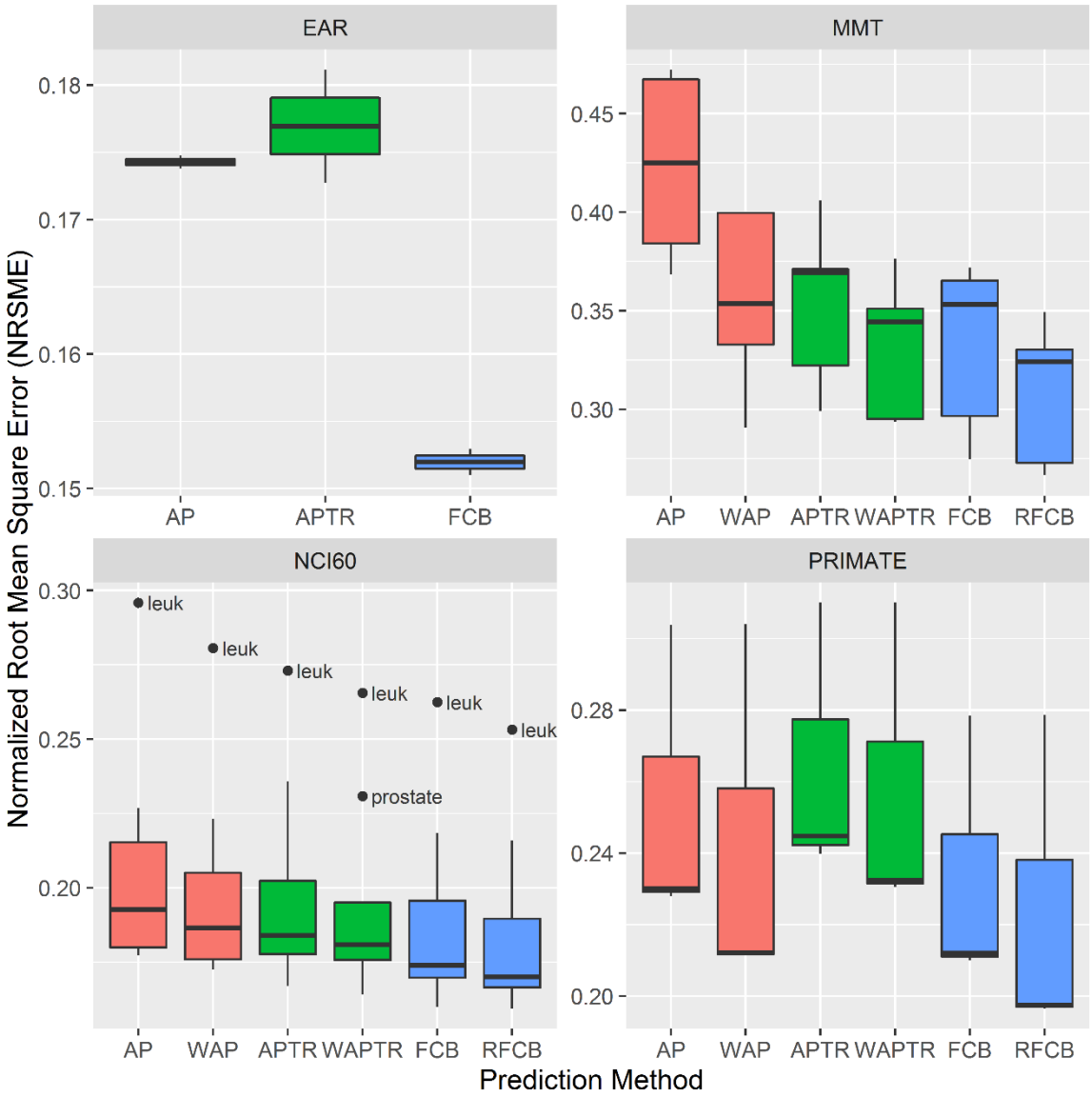

Figure 3.2-8 Performances of methods for protein level prediction. Boxplots show the distribution of the normalized root mean square error (NRMSE) in the prediction of protein levels, using six described methods: Averaged Protein (AP), Weighted Average Protein (WAP), Average PTR (APTR), Weighted Average PTR (WAPTR), FC Based (FCB), and Relaxed FCB (RFCB). In each tissue, RMSE values are divided by the standard deviation of the protein levels in that tissue. The error sizes are averages over tissues of 10-fold cross validation. In the EAR dataset there are only two groups, so the weighted/relaxed versions are irrelevant. See Figure 3.2-4 for box plot structure. Outliers are labeled.

The average improvement in the Mean Square Error (MSE) using the RFCB model over the next best weighted/relaxed model was 24.0%, 15.2%, 14.3%, 8.9% in the EAR, MMT, PRIMATE, and NCI60 datasets. Overall, the superiority of the FCB and RFCB supports the model of constant compression or expansion ratio between mRNA and protein fold-changes.

3.2.4.1 Compression parameter

Our previous analysis supports compression, at least for the EAR, MMT, and PRIMATE. The value of the compression parameter, α, of the FCB model is directly linked to the extent of compression. High variance between datasets and between groups was observed in the estimated value of this parameter (Figure 3.2-9). We thus conclude that this parameter should be adjusted separately for each protein level prediction task.

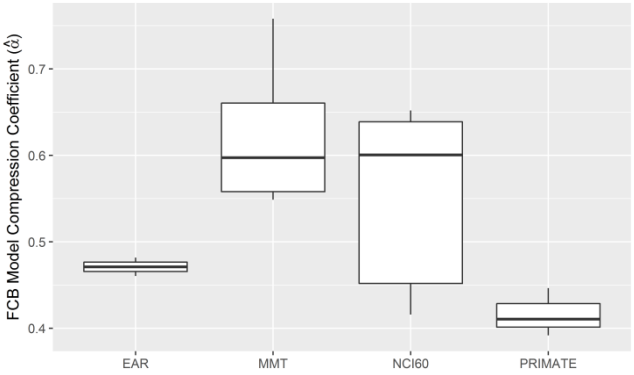

Figure 3.2-9 Estimated FCB model compression coefficient α. The FCB model was fitted separately for each group using linear regression. The boxplots show the distribution of α in the different datasets. See Figure 3.2-4 for box plot structure.

3.2.4.2 Protein expression prediction power in different datasets

The error measure presented in Figure 3.2-8 is normalized to allow the comparison of prediction quality between datasets. According to this measure, all models perform best on the EAR dataset, then on the NCI60, PRIMATE and MMT datasets in decreasing order of performance. In addition, we scored these differences by measuring the extent of variance in protein levels that is explained by the RFCB model in each of the datasets. In the EAR and NCI60 datasets 95.7% and 93.2% were explained respectively, decreasing to 89.8% in the PRIMATE dataset, and only 82.4% in the MMT dataset. The ranking is also the same using the AP model to score the datasets (94.3%, 91.6%, 87% and 65.7% respectively). We can conclude that the task of predicting protein levels, where one is given expression data from a similar tissue (EAR), or under the scenario of cancerous cell lines (NCI60, see next section regarding outliers), is easier than predicting using data from the same tissue but in different

species that were separated millions of years ago (PRIMATE, [122]), or from less similar tissues (MMT).

3.2.4.3 RFCB model is superior across groups and genes

So far, our analysis showed the superiority of the RFCB method at the level of a dataset. This superiority still holds when moving to the level of a group, as in all groups the MSE of the RFCB prediction is the lowest among all methods. Focusing on the NCI60 dataset, the greatest improvement in predictions in terms of normalized MSE is achieved for the leukemia and prostate, these cell lines having the lowest protein prediction power to begin with (Figure 3.2-10).

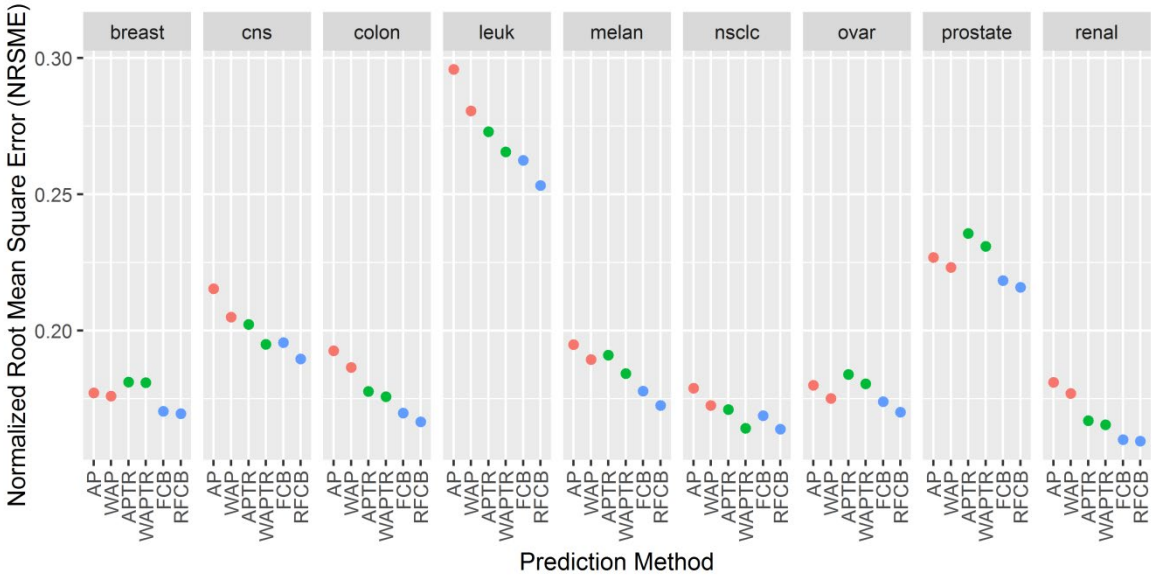

Figure 3.2-10 Quality of protein level prediction methods in NCI60 groups. For each group, we plotted the normalized root mean square error (NRMSE) in the prediction of protein levels, using six described methods: Averaged Protein (AP), Weighted Average Protein (WAP), Average PTR (APTR), Weighted Average PTR (WAPTR), FC Based (FCB), and Relaxed FCB (RFCB). Each box contains the prediction quality for a single group.

Next, we focused on the gene level, checking how well our prediction performs in predicting oncogene levels in cancer cell lines. Out of the 24 oncogenes surveyed in [123], we had full protein and mRNA data for CTNNB1, NRAS, and RB1. Using the six described methods, we predicted their protein levels in each NCI60 group, and compared the results to the measured

protein levels (Figure 3.2-11). For 21 out of 27 combinations of gene and group, all six predictions method performed well, with less than 2-fold difference between the expected and predicted levels. In the few cases where the difference was greater than 2-fold, the six methods were biased in their prediction in the same direction. An exception to this agreement was found in the prediction of NRAS expression in breast and prostate cell lines, where the predictions of the AP and APTR methods suffered from ~1.4-fold prediction biases in opposite directions. In both cell lines the FCB and RFCB methods had a nearly perfect prediction.

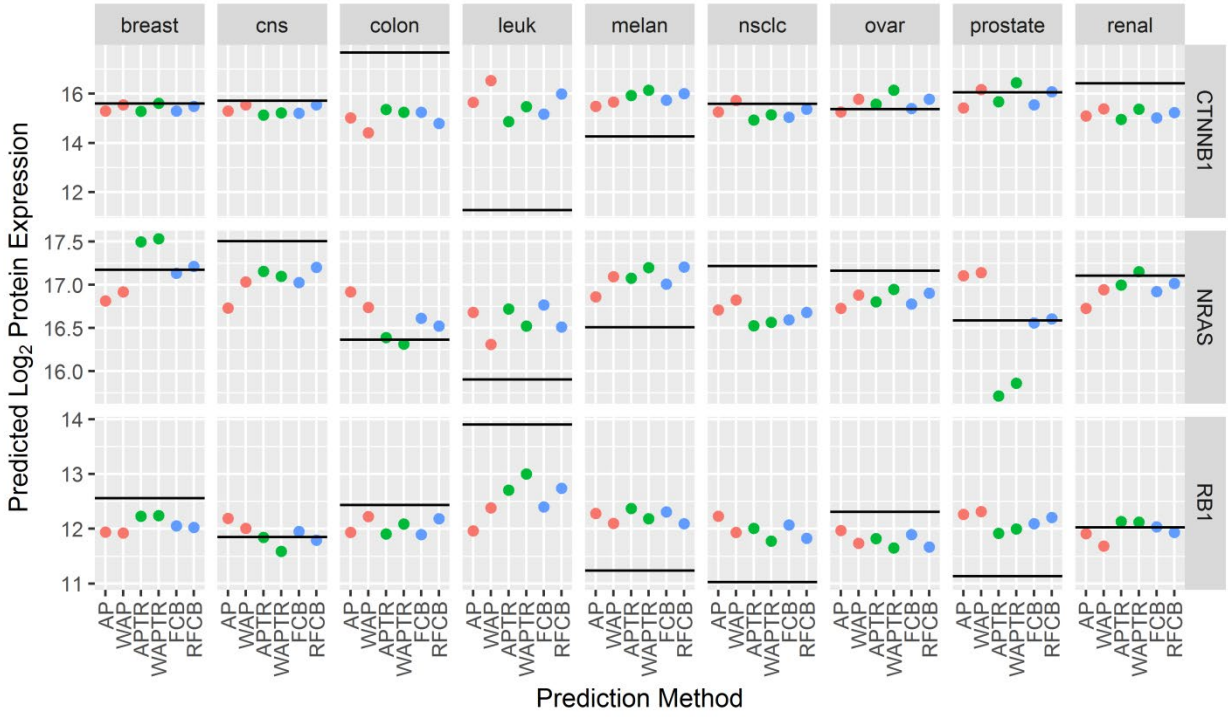

Figure 3.2-11 Quality of protein level prediction methods for oncogenes in the NCI60 dataset. For each group (columns) and gene (rows), we plotted the prediction of protein levels for that gene, using six described methods (see legend for Figure 3.2-10). The true (measured) protein levels are marked by horizontal lines.

3.2.5 Comparing differentially expressed genes in protein and in mRNA

We compared the DE genes between tissues in the EAR dataset, both at the protein and the mRNA domain. This type of comparison, as well as the comparison of the functional

enrichment of the DE genes on the mRNA and protein levels, can suffer from several biases that must first be addressed.

3.2.5.1 Biases make it more difficult to compare data between protein and mRNA

One source of bias is the different levels of noise in each domain, which affects the overall accuracy of the comparison. As demonstrated in Figure 3.2-4, protein levels are noisier in all datasets for which noise levels can be estimated. Resorting to enrichment analysis alleviates this last problem, because a failure to see a single gene DE in one domain but not the other, due to noise, is more probable than missing an entire group of genes that share a function and are coexpressed, because of noise.

The detection bias against lowly expressed proteins poses a more complex problem. Such proteins tend to have more missing measurements, and so our power to detect DE for a lowly expressed protein is lower. Consequently, the power to detect up-regulated functions that are performed mainly by lowly expressed proteins is lower.

The problem of missing data was evident in our data for the protein domain. Out of the 7018 proteins that have at least one measurement, only 5101 have at least one measurement in both cochlea and vestibule, 4443 have at least two in each tissue, and 3678 were measured in all six samples. The data are clearly not missing at random, as indicated by the increase in median logarithmized protein level across the four groups (19.91, 20.11, 20.28, and 20.63, respectively), consistent with known literature [53]. For the mRNA data this problem is negligible. One of the filtering stages performed in the preprocessing of this data is including only the genes that have one read per million in three or more of the samples. After this stage, 14,722 genes remain, out of which 14,693 genes have full data.

Only 201 genes have some measurements in protein, but not in mRNA. For the rest, we can compare the mRNA levels distribution of the genes that have some measurements in protein, to those without a measurement. We did so separately for the cochlea and vestibule (Figure 3.2-12), and observed that the levels of genes that have protein measurements is higher than the levels of genes without a measurement (p−value<2.2×10-16 for both, one-sided Two-Sample Kolmogorov-Smirnov test). The mRNA levels that we are comparing do not correlate perfectly with the protein levels, yet this still supports the 'missing not at random' quality of the protein data.

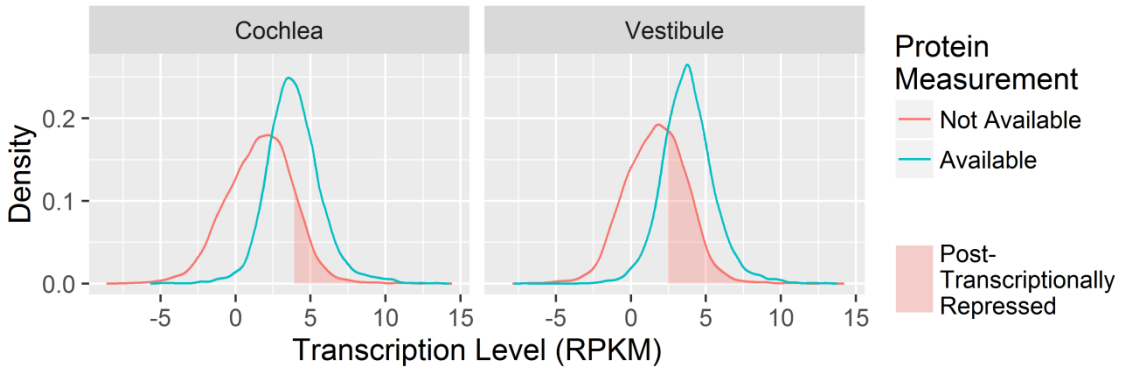

Figure 3.2-12 Distribution of mRNA levels in the cochlea (right) and vestibule (left). Density plots are shown for the mRNA levels of genes for which measured protein is available (blue), or not available (red). The red area marks the fraction of the distribution where genes with no protein measurement are considered post-transcriptionally repressed (see section 0).

To account for this effect, we reran DE using different filters on the minimum number of measurements in the protein domain. We focus here on the results when analyzing only proteins for which all measurements were available.

3.2.5.2 Differential expression indicates protein profiles are more similar than their RNA counterparts

Plotting the RNA and protein fold-changes of the DE genes (Figure 3.2-13), we observed that (i) more DE genes were found in the mRNA domain (235 versus 46 and 358 versus 156, upregulated in the cochlea and vestibule, respectively), (ii) genes found to be DE in protein

were usually DE also in mRNA in the same direction (in the cochlea, of the genes upregulated in protein, 78% were upregulated in mRNA and only 2.2% were downregulated; in the vestibule, the corresponding numbers were 76% and 2.6%, respectively), and (iii) genes found to be DE in both domains had more extreme mRNA fold changes than those found to be DE only in mRNA (median FC: 2.90 versus 1.62 and 2.37 versus 1.69 for genes upregulated in the cochlea and vestibule, respectively; q−values=9.4×10-11, 4.8×10-20, one-sided Wilcoxon rank sum-test). These observations imply that we expect the similarity between protein profiles to be higher than between their mRNA counterparts.

The results above were obtained using different DE tools and thresholds for protein and mRNA, in attempt to use the optimal tool for each domain. However, to reaffirm the above conclusions, we reran the analysis using the same tool (samr) and the same FDR threshold of 0.1 for both protein and mRNA. We found 752 and 956 genes up-regulated in the cochlea and vestibule respectively in the mRNA domain, and 46 and 156 genes in the protein domain. 85.1% of the genes found to be differentially expressed (DE) in protein were also DE in the same direction in mRNA, and 3.5% in the opposite direction. The FC_{mRNA} of genes that were up-regulated in one tissue in both mRNA and protein domains, was significantly more extreme than the FC_{mRNA} of the genes that are DE only in the mRNA domain (q−value=4.8×10-16 , 5.5×10-40, where group2 is the cochlea and vestibule respectively, one-sided Wilcoxon rank sum-test; Median FC: 2.72 versus 1.29 in cochlea, 2.29 versus 1.35 in vestibule). To summarize, all the three observations made above using the different tools, remain valid in this additional analysis.

We note that these results also remain valid when using other filters, which allow missing measurements in the protein domain (data not shown). We could not perform this type of analysis on the MMT dataset as statistically reliable DE techniques require replicates.

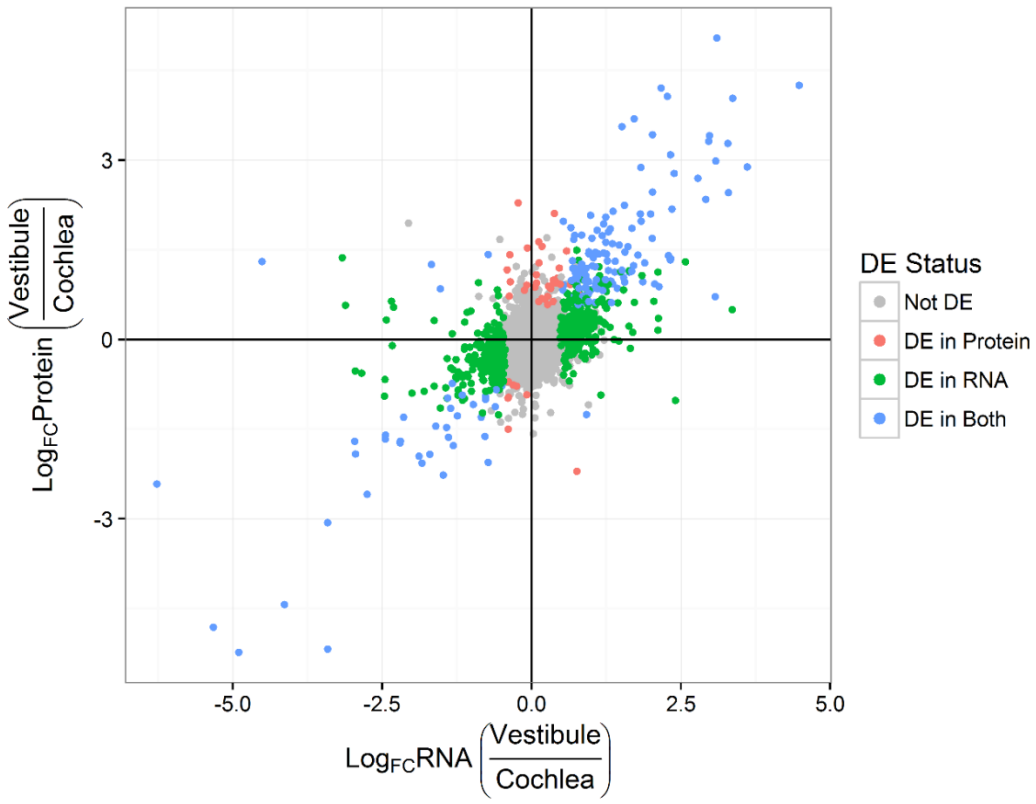

Figure 3.2-13 RNA and protein expression fold changes between inner ear tissues. For mRNA differential expression and fold-change estimation we used the edgeR package, with a detection threshold of q−value≤0.05. For protein we used the samr package (two class unpaired test) with threshold q−value≤0.1. Only proteins with measurements in all samples were included. Note that to emphasize the DE status, significant (colored) genes are drawn at the front and may occlude some non-significant ones.

3.2.6 Some tissue-functionalities coded in mRNA are not manifested in protein

For mRNA and protein, we looked for GO enrichment in the set of genes up-regulated in the cochlea versus the vestibule and vice-versa (for the full list of enrichments see Table S5 in [55]). We observed that the terms found in the mRNA domain represent a far broader list of functions than those found in the protein domain, when summarizing over the enrichments found using all filters. However, when comparing only the lists of enrichment terms found in the full data filter (i.e., using only the proteins with measurements values in all samples), the

lists were similar in size, yet quite distinct in content. Only three terms overlapped in the vestibule, representing 33% and 30% of the enrichments in the mRNA and protein, respectively, and none overlapped in the cochlea. The similar size of the two lists was surprising, considering the much higher number of DE genes in the mRNA domain. It was also unexpected to see so little overlap between the lists, as 77% of the genes found to be DE in protein were also DE in the same direction in mRNA in this analysis.

The analysis in the cochlea captured the functions of cell morphogenesis and nucleobase catabolic process in the mRNA domain, and the function of sensory perception in the protein domain. Importantly, the functions enriched in the protein domain were found in the mRNA domain when using less stringent filters, but not vice versa.

The analysis of the vestibule identified functions related to cell development and morphogenesis, biological adhesion, and response to wounding in both domains. Responses to general stimulus and chemicals, localization and cellular component movement, and renal system development, known to be related to ear development [124], were functions observed only in mRNA enrichments. Terms relating to anatomical structure morphogenesis, and specifically to the process of endochondral bone morphogenesis, were enriched in the protein, as was the less expected term of phagocytosis. Here also, all the functions enriched in the protein domain were either found, or similar terms to them were found, in the mRNA domain with less stringent filters. In contrast, none of the functions unique to the mRNA domain were found in the protein domain when using less stringent filters. These observations fit the hypothesis that some functionalities coded in mRNA are not manifested in protein.

An exception to this behavior, that is, a function that is relatively more 'active' in the protein domain, was found using a different approach for detecting post-transcriptional regulated

functionalities, in which we compared the functional profiles [95] of the DE genes between protein and mRNA. Using this approach, we concluded that the function of cell adhesion is post-transcriptionally controlled in the vestibule, with a relatively large number of genes that are not DE in the mRNA, but are so in the protein. In detail, the GOProfiles analysis showed a difference in the functional profiles of the genes up-regulated in the vestibule in the two domains, and managed to pinpoint the difference in the cell adhesion category (GO:0007155, q−value=1.75×10-2), for which 26% of the genes in the protein were annotated, and only 16% of the genes in the mRNA domain. In the cochlea, the possibility that the functional profile of up-regulated genes is the same for mRNA and protein could not be rejected.

We performed enrichment analysis on the MMT dataset as well, by ranking the genes according to their fold-changes in protein and mRNA, and using a cut-off independent approach [96] to identify enrichments in both domains (for the full list of enrichments see Table S7 in [55]). Inspired by [97], we scored each pair of tissues according to how specific the terms that arise from the enrichment analysis are, to either the protein or the mRNA domain (Figure 3.2-14). For most pairs of tissues, this analysis showed that there are more functions unique to the mRNA than to the protein. This was very prominent in functions upregulated in the heart compared to the liver. In contrast, functions up-regulated in the cerebellum, compared to the liver and kidney, were more specific to the protein domain. Next, we pooled the unique terms from all pairs, to determine which functions are uniquely enriched in one of the domains. After aggregating the results at the level of 'GO slim' [125], we observed that protein modification and amino acid metabolism, as well as transport, including vesicle-mediated transport, tend to be unique in the protein domain (Figure 3.2-15). In contrast, lipid metabolism and catabolic processes, along with stress response, are more transcriptome-specific functions. Terms related to cell death, cell adhesion, and

immune system response, all appeared multiple times (≥5) and only in the mRNA comparisons.

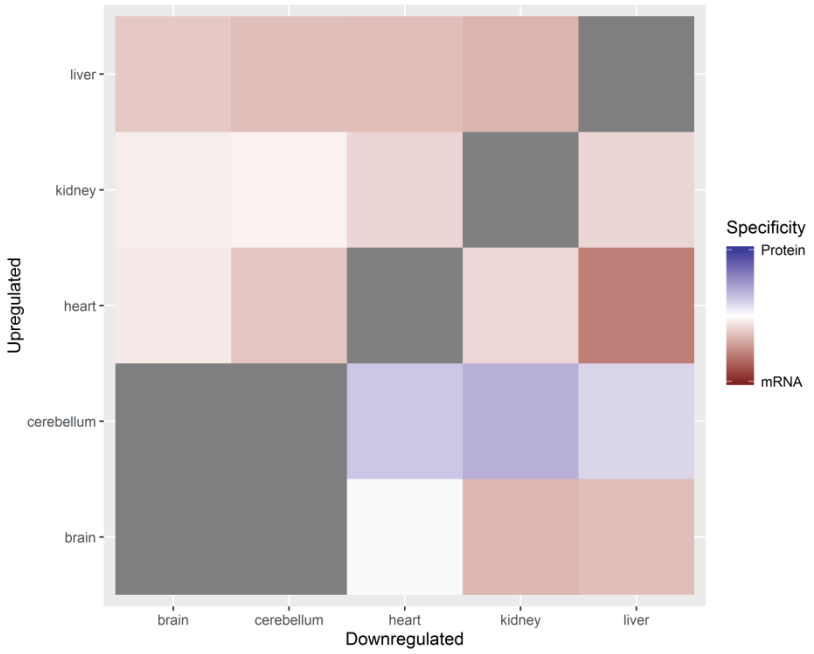

Figure 3.2-14 Semantic specificity of enrichments to protein or mRNA in the comparison of pairs of tissues [MMT]. Each tile represents a comparison between a pair of tissues in a certain direction, i.e., the terms emerged from genes that are up-regulated in the tissue of the y-axis compared to the tissue of the x-axis. The color of the tile indicates how much the enrichment terms tend to be specific to a single domain, ranging from full specificity to protein (blue) to full specificity to mRNA (red). Pairs, for which no terms were found, are colored in gray. The tiles on the diagonal do not represent a valid comparison.

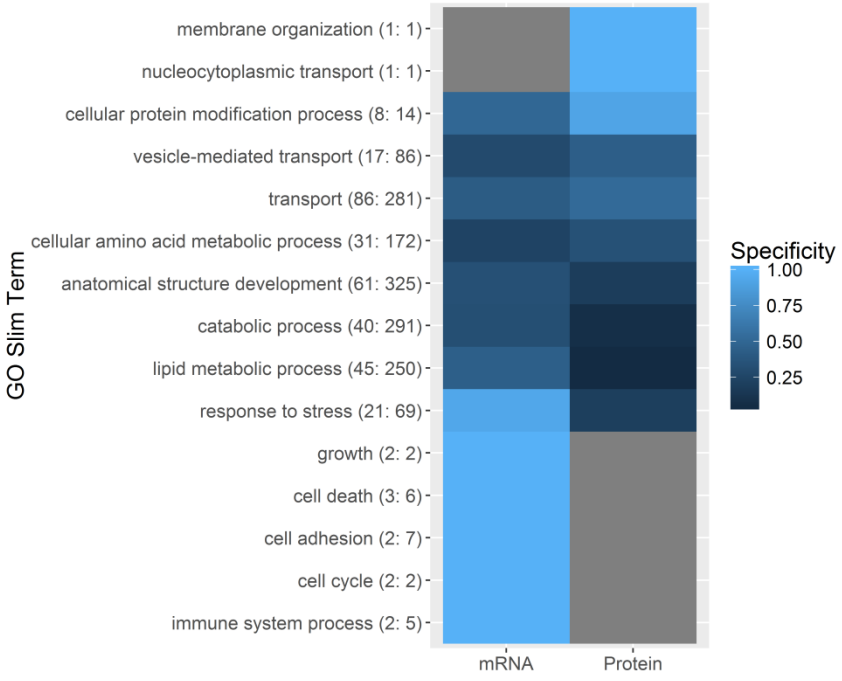

Figure 3.2-15 Transcriptome versus translatome specificity degrees associated with GO slim terms [MMT]. The heatmap illustrates the specificity of GO slim terms (rows) to the transcriptome and the translatome (columns), where a specificity of 1 indicates that all appearances of the GO term and its descendants are unique to the respective domain, and a specificity of 0 indicates that all these appearances are present in the other domain as well. For each term, the number of unique GO terms that were aggregated and the overall count of their appearances are listed in parenthesis. Included in this figure are only the GO terms for which the proportion of RNA unique terms, out of all unique terms, is significantly different than a background probability of 0.56 (two-sided proportion test, q-value≤0.1), or that were spotted only in one of the domains (in such a case the tile of the other domain is colored in gray).

3.2.6.1 Post-transcriptionally repressed genes

To complete the analysis, we also analyzed genes that show relatively high expression in the mRNA, but their measurements are completely missing from the protein.

Kwon et al. analyzed genes with high mRNA expression, with not even a single measurement in the protein domain [50]. To conduct a similar test on our datasets, for each group separately, we found the lowest mRNA expression level above which at least a fraction q of the genes have a valid measurement in the protein domain, with some correction for robustness. We defined the post-transcriptionally repressed genes as those above the mRNA threshold but with no protein measurement. We then performed enrichment analysis for each group, comparing the post-transcriptionally repressed genes to the background of all genes above the mRNA threshold.

We wanted to set a single fraction q for all datasets and groups. In Kwon et al. q=0.8 was used [50]. To accommodate the sparsity of our data we chose q=0.65. Figure 3.2-12 shows the EAR thresholds. For the cochlea and vestibule the thresholds were 3.97 and 2.58 RPKM, respectively, and 3603 and 7880 genes (out of 14,722) had higher expression, respectively. In both tissues, about one third of these genes were identified as post-transcriptionally repressed, as dictated by the parameter q. We note that between datasets and between groups within the same dataset, there is high variability in the number of genes with mRNA expression above the threshold (Table 3.2-1), suggesting that q could be set per group.

Table 3.2-1 Number of post-transcriptionally repressed genes. Number of genes for which: both mRNA and protein were detected (Detected), only mRNA was detected (Not Detected), mRNA was above 'detectability' expression threshold (Above Threshold), and the number of genes that fit the definition of post-transcriptionally repressed.

Dataset	Group	Detected	Not Detected	Above Threshold	Post-Transcriptionally Repressed
EAR	cochlea	5169	9553	3603	1260
EAR	vestibule	6638	8084	7880	2755
MMT	brain	5142	11827	668	230

MMT	cerebellum	5017	11952	986	341
MMT	heart	4173	12796	1453	507
MMT	kidney	5296	11673	3390	1185
MMT	liver	4791	12178	2493	870
NCI60	breast	5121	17633	2576	900
NCI60	CNS	4998	17756	2806	979
NCI60	colon	5116	17638	3105	1085
NCI60	leukemia	4690	18064	2103	734
NCI60	melanoma	5241	17513	3122	1092
NCI60	NSCLC	5367	17387	3220	1123
NCI60	ovary	4958	17796	2522	880
NCI60	prostate	4198	18556	1254	438
NCI60	renal	5066	17688	2473	865
PRIMATE		3394	8685	2320	810

We checked the post-transcriptionally repressed genes for enrichments, against the background of all genes exceeding the threshold (for the full list of enrichments see Table S8 in [55]). Some recurrent terms were intuitively explained as artifacts of detectability bias (data not shown). In the NCI60 data, other terms, such as 'cellular response to interferon-gamma - GO:0071346' in the colon, were more group specific. It was previously shown that colon cancer cell lines with Ki-ras mutations display reduced expression of interferon (IFN)-responsive genes [126]. It is known that IFN-γ regulates the mRNA translation of components of the immune system [127, 128]. Combining the two, we can explain why the cellular response to IFN-γ is post-transcriptionally repressed in colon cancer. The regulation of the immune response is also post-transcriptionally repressed in ovarian cancer according to our analysis. Several pathways are known to cause immune suppression in this cancer type [129–131]. Our findings suggest that some of the suppression is post-transcriptionally mediated. Other interesting terms are 'structural constituent of ribosome - GO:0003735' and 'organellar ribosome - GO:0000313' in leukemia [NCI60]. Indeed, in leukemia [132], as well as in ovarian cancer [133], the levels of expression of some ribosomal protein genes were found to be positively correlated with favorable clinical course. This suggests that reduction in certain

ribosomal proteins is contributing to the progression of cancer. We conclude that this reduction is achieved by post-transcriptional repression. The term ’embryonic organ development - GO:0048568’ appears in multiple cancer lines, such as breast, CNS, colon and NSCLC. It was reported that in poorly differentiated cancers, including breast and CNS tumors, the gene expression signature is similar to the one in embryonic stem cells [134]. We conclude that post-transcriptional repression is involved in this process. Many of the enrichments found in the PRIMATE datasets are related to either the mitotic cell cycle or to lymphocyte homeostasis and the immune response. This suggests the involvement of post-transcriptional repression in the immortalization process used to establish LCLs [135]. The EAR and MMT data suggest several other functions that are post-transcriptionally repressed, but additional evidence is required to support these hypotheses, as they are based on small number of genes or the detectability of the relevant proteins might be reduced (data not shown).

In conclusion, for some of the cancerous cell lines, we found tumor related functionalities that are controlled through post-transcriptional repression, namely, functionalities that are coded in mRNA but are less 'active' in protein.

4 CONCLUSIONS

In this work we generated RNA-seq and protein MS data for the sensory epithelia of the inner ear at mouse ages E16.5 and P0, which correspond to ages before and during the acquisition of mechanosensitivity. Except for the proteomics data for E16.5, the data were previously published, and are available for download in raw and processed forms. The transcriptomics data were also deposited in the gEAR portal, where it can be easily compared with data from other studies of the auditory and vestibular systems.

Our conclusions from the data can be predominantly separated into those resulting from the transcriptomic analysis across dimensions of age and tissue, and those resulting from the comparison of mRNA and protein levels at P0. Still, this last group of conclusions has an important implication to the interpretation of the transcriptomic analysis. Specifically, we learn that functionalities that are enriched in the cochlea versus the vestibule or vice versa, according to mRNA levels comparison, do not necessarily manifest in different protein levels, or they manifest to a lesser extent.

4.1 Transcriptomics Analysis

Exploring the transcriptomics data in the dimensions of age and tissue expanded our knowledge about the development of the IE. Moreover, we found transcription factors that are involved in transcriptional regulation, and focused on those that are also involved in regeneration of the avian IE after damage, based on previous work [66].

4.1.1 Major differences between the cochlea and the vestibule

The sensory epithelium constitutes a heterogeneous tissue. It contains two roughly defined populations of cells, HCs and SCs, which cannot be easily separated during dissection. In some experiments (e.g. [9]), separation is done using FACS sorting. Still, the analysis of data from

the native tissue has the advantage of summarizing the expression of the HCs and the milieu they interact with. Though we did not separate the cells by type, to better understand how the cell type proportions effected our results, we estimated them using expression deconvolution. Our estimated proportions showed a higher HC content in the cochlea compared to the vestibule, as well as an increase of HC content with development in both tissues, with a relatively larger increase in the vestibule.

Nearly 75% of the variation in the gene expression of our samples was explained by principal components associated with age (~47.5%) or tissue (~27.5%). Our analysis focused, accordingly, on the comparison of expression across age and across tissue. We also analyzed the harder-to-isolate interaction of age and tissue. In the comparison across age, we affirmed that both tissues become less proliferative and more differentiated with development, showing specialization for their roles in sensory perception. According to our estimations, this specialization is accompanied by an increase in the HC proportion in the sensory epithelia.

More surprising enrichments were obtained from the comparison between tissues. While the cochlea was characterized mainly by neurological GO terms, the vestibule was shown to be enriched in vascular, structural and immunological functions. Some of these differences could be attributed to the differences in HC proportion, which is presumably higher in the cochlea at both ages. This finding has medical implications, as the higher vascularization of the vestibule, and its accessibility to immune cells, might imply different susceptibility of the tissues to ototoxic medications and IE infections.

Examining the interaction of tissue and age, one notable finding is the delay in the development of sensory perception in the cochlea versus the vestibule. This finding is supported by delayed acquirement of mechanosensitivity in the cochlea (between P0 and P2

[14]) compared to the vestibule (between E16 and E17 [13]). On the other hand, at E16.5 the vestibule is less developed in neuron projection and signaling compared to the cochlea, and by P0 the gap between the two decreases. This decrease can be attributed to the relatively larger increase in HC proportion in the vestibule compared to the cochlea.

In conclusion, in our comparison of the cochlea and the vestibule, we found the cochlea to be more enriched in neurological functions, and to contain a higher percentage of HCs than the vestibule, but also to have a delayed development of its sensory perception compared with the vestibule. The vestibule, on the other hand, was found to be more vascular and more accessible to the immunological system.

4.1.2 Deafness genes prediction

Known DGs tend to be differentially expressed between the tissues. From E16.5 to P0, they increase both in expression, and in cochlea to vestibule expression ratio. We built a classifier that used expression features to predict the probabilities of genes to be yet undiscovered DGs. This classifier achieved a ROC score of 0.602 in predicting which genes are associated with deafness according to text mining tools.

A previous attempt to find candidate deafness genes using bioinformatics [38] was limited to a search within genomic regions linked to various nonsyndromic hereditary HL phenotypes, did not use machine learning, and did not provide an estimation of the accuracy of the predictions. While our classifier achieves a poor ROC score, our work demonstrates many advantages over this previous attempt: our classifier considers all genes in the genome, it learns from patterns of known DGs, and we provide some estimation of its performance. Notably, this estimation may be biased downwards as the classifier learned to classify DGs, but it was tested on the task of classifying deafness-**associated** genes. Ideas for improvement of the classifier are presented in section 0

Suggestions for future research.

The list of deafness associated genes might not accurately reflect the true list of DGs, i.e. all genes that are essential for hearing. Still, it served to estimate the performance of the classifier, because we believe it to be a good proxy for the true list. Our ordering of the genes according to their estimated probability of being DGs can be used to prioritize candidate DGs in real world scenarios, e.g., when multiple candidates arise from a familial segregation study or when exploring very large genomic regions associated with deafness (as done in [38], a list of loci associated with deafness is available at http://hereditaryhearingloss.org/).

4.1.3 Transcription factors in inner ear development

The main purpose of this research was to elucidate transcriptional pathways that govern auditory versus vestibular specification or control cell cycle exit. We used enrichment analysis to identify TFs that are responsible for differences in expression between across tissues or ages, respectively. Some of the TFs we identified as controlling expression were already known, e.g. the E2F family of TFs, which is promoting proliferation in the sensory epithelia and is controlled by retinoblastoma 1 during this stage of development [17], or the retinoic acid nuclear receptors, which are essential for the proper morphogenesis of the ear [120]. Our analysis not only strengthens the evidence connecting these known TFs to IE development, but also emphasizes their additional roles in HC regeneration in birds (see below). Other TFs found do not have a known function in the IE. Such are *Arnt*, which activate the transcription of its target genes in E16.5; COUP TFs, which we speculate to have a dual role, with *Nr2f2* inhibiting myogenesis in the cochlea and *Nr2f1* promoting retinoid signaling; and the hemopoiesis agent *Lmo2* [111], which we believe cooperates with different coactivators in the vestibule and the cochlea.

It might be beneficial to mimic the transcriptional regulation during a response to IE damage in birds, as avian cochlear HCs regenerate [66]. In a previous experiment, thousands of TFs changed significantly during such a response. We intersected them with our list of developmental TFs, in order to highlight genes that are more likely to be involved in either proliferation or differentiation. We found dozens overlapping TFs, but we focused only on those that are DE between conditions, because we can more easily interpret how they are regulated, and influence this regulation through interventions. We highlighted the complex Arnt:AhR, which we believe is important in early development, and its transient increase is observed during avian HC regeneration. We also emphasized that an increase in the genes *Zbtb14* [ZF5], *Lmo2*, *Nr2f1*, *Nr2f2*, and *Smad9*, and a decrease in *Spi1* [PU.1], *Nfe2l2* [Nrf2], and *Mafk* [Nrf2], is needed for proper differentiation of the cochlea during HC regeneration. An increase in *Smad2* is involved in the same process in the vestibule.

To conclude, the majority of TFs we predicted to be key regulators of the differentiation process, have known functions that agree with this dichotomist characterization. Some of which were further selected as possible candidates in inducing HC regeneration.

4.1.4 Summary of transcriptomics analysis

Our work highlighted differences in biological processes activity, developmental timeline, and HC content between the cochlea and the vestibule, as these are manifested in the mRNA expression profile of the two tissues. Differential expression of certain TFs was identified as a driving force for differentiation into one tissue and not the other, while others were associated with proliferation of cells during development. The intersection of the two groups with a list of TFs involved in avian HC regeneration provides strong candidates for future intervention in HC damage in mammals. Apart from HC damage, there are many other

mechanisms of genetic deafness. In our analysis we noted that DGs show a unique pattern of spatial and temporal expression, which we used to build a classifier for undiscovered DGs.

4.1.5 Limitations of the current study

Any conclusions drawn from naïve enrichment analysis are limited by faulty assumptions made by the statistical framework used, inherent error rates (especially of type I, also known as "false positives"), biases in the annotations in genes towards more extensively studied biological entities and disorders [136], and, of course, errors in the preliminary mRNA data and the differential expression analysis. Another source of inaccuracy lies in the interpretation of the results, as we made some careful hypotheses about the function of the tissues based on the composition of the enrichments that can profit from further biological validation.

4.1.6 Suggestions for future research

Protein-protein interaction (PPI) networks are used to represent physical binding events measured between protein pairs [137]. It was shown that these networks possess a "community structure" that groups together proteins that interact more frequently with one another and share common functions. This property, among others, is used in research to detect relevant regulated genes and pathways in individual samples or disease states. A PPI network usually reflects an aggregate of likely networks, which are based on measurements across many different conditions that together cover a variety of protein expression profiles [137]. This may pose a disadvantage in interpreting a specific biological condition in light of a PPI network, as the network contains many interactions that are not active in the condition due to some of the interacting proteins not being expressed. Because of this, different attempts were made to incorporate PPI networks with the more flexible technologies of mRNA and protein expression profiling. A simplistic approach for finding functional modules

of interacting proteins that are active in one state and not the other, is to filter the network for DE genes between the states, find clusters of highly connected proteins in the resulted network and search for functional enrichment within them (for examples in disease states see [138, 139]). A similar approach can be used in our research to find modules involved in differentiation and proliferation in the IE. Even more importantly, by combining PPI data with TF binding site data, we can refine our list of the master regulators of IE regeneration, using the established method suggested in [137].

The development of a classifier for deafness genes is a unique contribution of this study. While better than everything else available, it does not perform very well on absolute terms. Its development is a first step toward achieving a good classifier. Problems facing the development of such a classifier are the very small number of positive samples (around 140 known deafness genes); and the heterogeneity of hereditary deafness, as HL can be classified as conductive, sensorineural, or mixed (a combination of both); syndromic or nonsyndromic; and prelingual or postlingual [36]. This latter quality suggests a potential for using multi-label classification algorithms [140].

A new classifier might extend our use of gene expression to data from organisms other than mouse, other developmental ages, isolated cell populations, tissues exposed to noise or pharmacological treatment, and perhaps even data from non-ear tissues that can implicitly suggest ear specific roles for genes (for available inner ear datasets, see [7]). It might also benefit from the availability of protein MS data, published by our lab [55]. Another feature that can be used is PPIs. The key assumption in using this feature is that a network-neighbor of a disease-causing gene is more likely to cause either the same or a similar disease [141]. This suggestion is inspired by [38], where the bioinformatic search for candidate genes for deafness included a filtering over genes according to the interaction of their products with proteins involved in inner ear development or function, as well as from other works that used

these interactions to predict disease genes, alone [142] or with the use of co-expression [143] and/or gene ontology annotations [144, 145]. In this regard, it might be beneficial to use the ensemble method proposed by Yang et al. in the field of positive unlabeled learning for disease genes [144]. That method was used to incorporate gene expression, protein interactions and gene ontology annotations.

4.2 Protein and mRNA joint analysis

4.2.1 Changes in transcription levels are buffered on the protein level

In this analysis we compared mRNA and protein expression across diverse datasets: mouse inner ear tissues, mouse organs, cancer cell lines and primate lymphoblastoids. We observed that the correlations in protein expression between groups are higher than the correlations in mRNA expression, across all datasets. It was previously observed that across *taxa* protein levels are more conserved than mRNA levels [49]. We showed this phenomenon across *tissues* as well, and explained it by changes in the transcript level that are attenuated at the protein levels. A direct outcome of this phenomenon is the compression of large differences in mRNA expression to smaller ones in the protein domain. This is the first observation of this phenomenon for non-proliferating tissues, though it was previously seen in proliferative ones [48]. Moreover, the aforementioned studies used OLS regression, which is known to suffer from a strong dilution bias [53]. Using the more robust MA regression instead, we provided evidence for such compression in EAR, PRIMATE and in MMT (except for one tissue pair). In NCI60 and the brain-cerebellum pair [MMT] the regression results supported expansion, instead of compression.

When comparing tissues that are very similar in level of expression, small biases can render the regression invalid. In order to solve this issue, we tried a non-parametric approach, which can be less powerful but is not dependent on an underlying linear model. Using this approach,

we showed buffering for all datasets except NCI60. We therefore conclude that a partial buffering between translation and transcription exists in the MMT, EAR, and PRIMATE datasets. For NCI60, the results were insignificant, and supported neither compression nor its opposite, signal amplification. Perhaps a more powerful test (for example, a random effects model [53]) may provide the answer. For the PRIMATE dataset such an observation was made previously [51]. In this study, by addressing some of the limitations of that statistical analysis, we reaffirmed the correctness of the observation. Notably, both the parametric and non-parametric approaches were statistically robust to the different levels of noise in protein and mRNA, which in our study, manifested in the higher correlations between mRNA replicates compared to between protein replicates in all datasets (a property that was observed also in [53], where the authors reported the median correlation between mRNA replicates to be higher than the median in correlation between protein replicates in 13 and 11 separate studies in yeast, respectively).

We did not necessarily expect to see the same phenomena in cancer cell lines as in healthy tissues, for obvious reasons: cell lines are programmed to proliferate, whereas cells in healthy tissues divide slowly, if at all; cell lines somewhat lose their resemblance to their tissue of origin, thus becoming more similar to a "global cancer pattern"; and cell lines of the same origin may diverge in their transcriptomic and proteomic profiles as they follow different paths of cancer evolution. In addition, the post-transcriptional regulation may be altered or even damaged in cancer. We showed one manifestation of these biological differences, namely the lesser ability to separate NCI60 samples based on their origin, compared to the EAR and MMT datasets. Since the cell lines are more similar to each other in their expression profiles, the compression effect is expected to be less dominant in cancer.

A translational model has been proposed, where transcriptional signals are amplified by translational regulation [53]. The existence of an amplifying mechanism might appear to

contradict the buffering suggested here. However, the authors studied budding yeast, a single cell type. In this model an increase in the mRNA level of a transcript would translate into an exponential increase of the matching protein, while our analysis is based on multiple tissues. In each tissue the transcriptional, translational and post-translational regulations are fine-tuned to enable the correct function of the tissue. Both mechanisms can coexist, i.e. the expression profiles that we observe are the result of a balance between compressing and amplifying mechanisms. The first is related to the tissue identity (perhaps through epigenetic marks), and the second is connected to the way the translational apparatus of a cell functions. A very similar argument was made in [53], in the context of different species. We speculate that the contradicting evidence we observe for buffering in groups that are more similar to one another might be the result of such balance. I.e., in such groups, the balance between the two mechanisms leans towards amplification.

4.2.2 Possible mechanisms for buffering

What biological mechanism explains the buffering observation? Decoupling is achieved by changing the translation rates, the protein degradation rates, or both. We cannot distinguish between these three options using our analysis, yet according to the literature, protein translation is assumed to be the major contributor to the variance of protein concentration [45], and was shown to change through tissue differentiation [43]. Hence we can speculate that the translation rate is the factor that is changing between the two tissues, although in a different context, of expression quantitative trait loci in LCLs, the buffering observed between protein and mRNA was attributed mainly to protein degradation [146].

It has been suggested that translational efficiency decreases with increased mRNA levels due to competition for scarce resources, e.g., ribosomes [48]. However, as ribosomes are part of a nonspecific translation machinery, that would not work slower in translating a specific gene

if it is over-transcribed. Another explanation as to how the coordination of translation and transcription is achieved, is that certain proteins, which participate in replication and transcription (e.g., Rap1 and Abs1 in yeast), could be incorporated into the mRNA, exported from the nucleus, and differentially affect the rate of translation at the ribosome [47]. However, to date no proof was provided for this mechanism.

We propose a third option, inspired by a study on the correlation of transcription and mRNA degradation rates in yeast [147]. The authors demonstrated that these rates are negatively correlated, and showed that the mutations responsible for this effect usually influence both transcription and mRNA degradation. Analogously, perhaps epigenetic changes are involved in coordinating the rates of transcription and translation in our system.

4.2.3 Range compression assumption improves protein levels prediction

We demonstrated how the prediction of protein can be improved by taking the range compression into account. Models that allow PTR to vary between tissues in a direction that buffers the change in protein levels (R\FCB), performed better than models that did not allow this variation or ignored RNA levels altogether. The improvement in the prediction error was between 9% and 24%, depending on the dataset. The largest improvement was achieved in the EAR, but in this dataset the prediction was very good to begin with. In the PRIMATE dataset the smaller improvement of 14% can make a large difference in the prediction quality. This enhanced ability to predict protein levels can be utilized, for example, to better predict disease status using machine learning. The higher accuracy exhibited by the RFCB method in the prediction of the NRAS protein level in breast cancer cell lines, supports its usage in disease status evaluation, as overexpression of NRAS is associated with poor prognosis in breast cancer [148].

In the future, as understanding of mRNA-protein relationship improves, more sophisticated prediction tools can be developed that will be aware of this mechanism and explore different features of it (for example, whether it saturates in higher mRNA expression levels). Notably, the ability of sequence transcription features to explain 30% of protein abundance in addition to what can be accounted for by mRNA concentrations [149] suggests a large influence of gene sequence over post-transcriptional processes such as translation and degradation. It is tempting to speculate that they have some impact on the buffering mechanism as well; and that this impact can be modeled to improve accuracy.

If buffering worked in the linear fashion captured by the FCB model, and the noise level was similar in the measurements of protein and mRNA, then we would expect the correlations between tissue pairs in the protein and the mRNA domains to be almost equal. We observed, however, that the correlations in the protein domain were higher. This is a surprising finding, especially in light of the higher noise level in protein, suggesting that a more powerful nonlinear buffering model could be described. Another support for a stronger buffering comes from the number of DE genes we found, which was much higher in the mRNA domain. As mentioned, the protein measurements are slightly noisier, though probably not to the extent that justifies these high differences.

4.2.4 Suggested role for buffering mechanism in stress response

In the enrichment analysis we observed that the functionalities represented at the protein domain were, by and large, a subset of the functionalities represented at the mRNA domain, which were far more numerous. The fact that we find less enrichment categories in protein is partially explained by the missingness pattern in the protein measurements: we have less chance to detect categories in which some or all of the genes are lowly expressed in the protein domain (or characterized by low detectability by MS). Focusing on the subset of

genes with full measurements in protein allows a fairer comparison, but nearly ignores the possible differences between those 'low expression' categories. In that comparison we found a similar number of enrichment categories for protein and mRNA. The lists differ greatly; however, we notice that the categories that were found in the protein and not in the mRNA, were represented in the analysis of the full, non-filtered, mRNA data. We can conclude that all the functionalities that are represented in the protein are also evident in the mRNA data. For the opposite direction it is much harder to tell; to accurately answer this question we need to somehow predict the missing values in the protein, or develop an enrichment analysis tool that is aware of the 'missing not at random' nature of the data [150].

Why does one tissue maintain higher mRNA levels but the same protein levels compared to another, where such practice requires more energy from the cell? We suggest that functionally distinct tissues possess different mRNA profiles but similar protein profiles, in rest, as part of a preparation for a stimulus. Under some stimulus a translational inhibition is removed from a gene (or group of genes) that is DE between the tissues only at the mRNA domain, so that the tissue that possesses higher levels of the gene's transcript will synthesize the protein faster. Indeed, one of the virtues attributed to translational control is the possibility of rapid response to external stimuli [151]. Moreover, when exposing mammalian cells to stress induced by dithiothreitol, mRNA- and protein-level regulation contribute equally to the change in protein expression [152], demonstrating the importance of protein-level regulation under stress. If our suggestion is correct, it might be beneficial to measure both mRNA and protein levels in order to deduce functionality of genes. If a gene is DE at the protein domain, then the protein is important to the function of the resting tissue. If a gene is DE only at the mRNA domain, then it is required for the tissue functionality under some stimulus.

The fact that the vestibular up-regulated genes are enriched for response to stimulus and chemicals only in the mRNA domain might be a manifestation of this hypothesis, as a role for these responses in the normal development of the ear is not known. Also fitting this hypothesis are the multiple immune related terms found in the mRNA domain, in the analysis of the non-filtered data. Nevertheless, the lack of these terms from the protein analysis might be related to a relatively low expression of the genes in these categories. In the MMT analysis we see a similar pattern. Response to stress terms are enriched in mRNA data and not in protein, and those of immune system response are unique only to mRNA. In the literature we can find examples where the translational regulation of genes changes in response to heat shock [153], hypoxic stress [154], changes in iron concentration [155], and exposure to EGF [97]. It is interesting to explore whether the genes activated in these responses are highly expressed in the mRNA domain, compared to a tissue that is not normally subjected to these types of stress, even before the actual exposure.

4.2.5 Summary of protein and mRNA joint analysis

Our work demonstrates that protein levels are more conserved between tissues than mRNA levels. We employed this observation to improve the prediction of protein levels in a non-proliferating tissue based on the mRNA levels measured in that tissue, by using data from several other tissues. A biological explanation is proposed as to why tissues maintain different levels of mRNA and similar levels of protein, by providing examples where this phenomenon serves as a preparation for a stimulus.

4.2.6 Limitations of the current study

The number of proteins detected in different proteomic experiments ranges between 3,000 and 7,000 per sample, whereas the expressed mRNA transcripts covers a much larger portion of the genome. As shown in section 3.2.5.1, the data are not missing at random, that is, there is

a detection bias against lowly expressed proteins. In our analysis, whenever two groups were compared with one another, we filtered genes that were expressed in at least one sample of protein and one sample of mRNA, in both groups, in order to reduce the bias that is caused by the difference in the detection abilities between protein and mRNA. This is a reasonable solution, although it is not as complete as a statistical modelling of this pattern of missingness [53]. More importantly, some of our conclusions regarding the buffering effect cannot be generalized to the part of the genome that encodes proteins that are lowly expressed. Specifically, results indicating a higher conservation of protein levels compared to mRNA levels, and others supporting the existence of a buffering mechanism, were based on biased measurements.

Similarly, the manifestation of certain tissue-functionalities in mRNA but not in protein can be explained by a detection bias against functionalities that are performed mainly by lowly expressed proteins. This problem was somewhat alleviated by rerunning the enrichment analysis with different filters on the genes included according to the minimum number of measurements in the protein domain, and then comparing the enrichment terms between the reruns, yet we cannot reject this explanation altogether.

4.2.7 Suggestions for future research

As outlined in the previous section, the problem of missing protein measurements interferes with our ability to generalize our finding to the entire genome. Perhaps in the future, the technology of MS will advance to a point, where the problem of missing measurements will become a non-issue. A more immediate, yet partial, solution using current datasets is to examine whether these phenomena exist specifically for the portion of the genes with the lowest protein measurements within the datasets. If so, this would provide some support for their relevance for the entire genome. Alas, such an analysis might be more prone to effects

caused by the detection bias. Another alternative is the use of ribosome profiling data as a proxy for protein abundance. In contrast to MS, ribosome profiling is not limited as much by a detection threshold, as it is based on deep sequencing ribosome-protected mRNA fragments [140]. Therefore, it can be used to investigate whether our results also apply to lowly expressed proteins. However, if the buffering mechanism described earlier works by changing protein degradation rates, then its effect would not be evident from ribosome profiling data, as this method provides a snapshot only of the translational activity of the cell.

This property of ribosome-seq leads us to yet another proposal for future research. As ribosome-seq results are only affected by the part of the buffering mechanism mediated by protein degradation, this tool can be used to determine the relative contribution of the degradation to the mechanism, and, by subtraction from the full effect as it is measured by MS, to estimate the translation component as well. The contribution of protein degradation can also be assessed, to some limited degree, by the existing MS data. According to Schwanhäusser et al., the protein production rate appears to saturate for very highly expressed proteins [45]. Therefore, any buffering seen for these proteins can be ascribed primarily to differences in degradation rates.

Another interesting question relates to the connection between the buffering mechanism described here, which reduces variability in proteins level between tissues, and the observed buffering of protein levels in the context of "noisy" mRNA levels at other scales. One example is the buffering on evolutionary timescales of inter-species variation [49, 51]. This was demonstrated here through the PRIMATE dataset. Other examples, of buffering observed in the intra- and inter-individual scale are reviewed in [44]. One can ask whether all these buffering phenomena share a molecular mechanism or have similar traits.

5 REFERENCES

1. Petit C, Richardson GP. Linking genes underlying deafness to hair-bundle development and function. Nat Neurosci. 2009;12:703–10.

2. Friedman LM, Dror AA, Avraham KB. Mouse models to study inner ear development and hereditary hearing loss. Int J Dev Biol. 2007;51:609–31.

3. Platt RJ, Chen S, Zhou Y, Yim MJ, Swiech L, Kempton HR, et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell. 2014;159:440–55.

4. Zou B, Mittal R, Grati M, Lu Z, Shu Y, Tao Y, et al. The application of genome editing in studying hearing loss. Hear Res. 2015;327:102–8.

5. Groves AK, Fekete DM. Shaping sound in space: the regulation of inner ear patterning. Development. 2012;139:245–57.

6. Wu DK, Kelley MW. Molecular mechanisms of inner ear development. Cold Spring Harb Perspect Biol. 2012;4:a008409.

7. Schimmang T, Maconochie M. Gene expression profiling of the inner ear. J Anat. 2016;228:255–69.

8. Cristobal R, Wackym PA, Cioffi JA, Erbe CB, Roche JP, Popper P. Assessment of differential gene expression in vestibular epithelial cell types using microarray analysis. Brain Res Mol Brain Res. 2005;133:19–36.

9. Scheffer DI, Shen J, Corey DP, Chen Z-YZ-Y. Gene expression by mouse inner ear hair cells during development. J Neurosci. 2015;35:6366–80.

10. Alawieh A, Mondello S, Kobeissy F, Shibbani K, Bassim M. Proteomics studies in inner ear

disorders: pathophysiology and biomarkers. Expert Rev Proteomics. 2015;12:185–96.

11. Elkan-Miller T, Ulitsky I, Hertzano R, Rudnicki A, Dror AA, Lenz DR, et al. Integration of transcriptomics, proteomics, and microRNA analyses reveals novel microRNA regulation of targets in the mammalian inner ear. PLoS One. 2011;6:e18195.

12. Herget M, Scheibinger M, Guo Z, Jan TA, Adams CM, Cheng AG, et al. A simple method for purification of vestibular hair cells and non-sensory cells, and application for proteomic analysis. PLoS One. 2013;8:e66026.

13. Géléoc GSG, Holt JR. Developmental acquisition of sensory transduction in hair cells of the mouse inner ear. Nat Neurosci. 2003;6:1019–20.

14. Lelli A, Asai Y, Forge A, Holt JR, Géléoc GSG. Tonotopic gradient in the developmental acquisition of sensory transduction in outer hair cells of the mouse cochlea. J Neurophysiol. 2009;101:2961–73.

15. Li S, Mark S, Radde-Gallwitz K, Schlisner R, Chin MT, Chen P. Hey2 functions in parallel with Hes1 and Hes5 for mammalian auditory sensory organ development. BMC Dev Biol. 2008;8:20.

16. Jamon M. The development of vestibular system and related functions in mammals: impact of gravity. Front Integr Neurosci. 2014;8.

17. Kelley MW. Regulation of cell fate in the sensory epithelia of the inner ear. Nat Rev Neurosci. 2006;7:837–49.

18. Sonntag M, Englitz B, Typlt M, Rubsamen R. The Calyx of Held develops adult-like dynamics and reliability by hearing onset in the mouse in vivo. J Neurosci. 2011;31:6699–709.

19. Kiernan AE, Ahituv N, Fuchs H, Balling R, Avraham KB, Steel KP, et al. The Notch ligand Jagged1 is required for inner ear sensory development. Proc Natl Acad Sci USA. 2001;98:3873–8.

20. Vahava O, Morell R, Lynch ED, Weiss S, Kagan ME, Ahituv N, et al. Mutation in transcription factor POU4F3 associated with inherited progressive hearing loss in humans. Science. 1998;279:1950–4.

21. Chien WW, Monzack EL, McDougald DS, Cunningham LL. Gene therapy for sensorineural hearing loss. Ear Hear. 2015;36:1–7.

22. Vona B, Nanda I, Hofrichter MAH, Shehata-Dieler W, Haaf T. Non-syndromic hearing loss gene identification: A brief history and glimpse into the future. Mol Cell Probes. 2015;29:260–70.

23. Corwin JT, Cotanche DA. Regeneration of sensory hair cells after acoustic trauma. Science. 1988;240:1772–4.

24. Ryals BM, Rubel EW. Hair cell regeneration after acoustic trauma in adult Coturnix quail. Science. 1988;240:1774–6.

25. Forge A, Davies S, Zajic G. Characteristics of the membrane of the stereocilia and cell apex in cochlear hair cells. J Neurocytol. 1988;17:325–34.

26. Kawamoto K, Izumikawa M, Beyer LA, Atkin GM, Raphael Y. Spontaneous hair cell regeneration in the mouse utricle following gentamicin ototoxicity. Hear Res. 2009;247:17–26.

27. Cox BC, Chai R, Lenoir A, Liu Z, Zhang L, Nguyen D-H, et al. Spontaneous hair cell regeneration in the neonatal mouse cochlea in vivo. Development. 2014;141:816–29.

28. Kawamoto K, Ishimoto S-I, Minoda R, Brough DE, Raphael Y. Math1 gene transfer generates new cochlear hair cells in mature guinea pigs in vivo. J Neurosci. 2003;23:4395–400.

29. Masuda M, Pak K, Chavez E, Ryan AF. TFE2 and GATA3 enhance induction of POU4F3 and myosin VIIa positive cells in nonsensory cochlear epithelium by ATOH1. Dev Biol. 2012;372:68–80.

30. Ikeda R, Pak K, Chavez E, Ryan AF. Transcription factors with conserved binding sites near ATOH1 on the POU4F3 gene enhance the induction of cochlear hair cells. Mol Neurobiol. 2015;51:672–84.

31. Tarang S, Doi SMSR, Gurumurthy CB, Harms D, Quadros R, Rocha-Sanchez SM. Generation of a Retinoblastoma (Rb)1-inducible dominant-negative (DN) mouse model. Front Cell Neurosci. 2015;9.

32. Weber T, Corbett MK, Chow LML, Valentine MB, Baker SJ, Zuo J. Rapid cell-cycle reentry and cell death after acute inactivation of the retinoblastoma gene product in postnatal cochlear hair cells. Proc Natl Acad Sci USA. 2008;105:781–5.

33. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10:252–63.

34. Silviu A, Muresanu DF. Vestibular regeneration–experimental models and clinical implications. J Cell Mol Med. 2012;16:1970–7.

35. Angeli S, Lin X, Liu XZ. Genetics of hearing and deafness. Anat Rec Adv Integr Anat Evol Biol. 2012;295:1812–29.

36. Kochhar A, Hildebrand MS, Smith RJH. Clinical aspects of hereditary hearing loss. Genet

Med. 2007;9:393–408.

37. Atik T, Bademci G, Diaz-Horta O, Blanton SH, Tekin M. Whole-exome sequencing and its impact in hereditary hearing loss. Genet Res. 2015;97:e4.

38. Alsaber R, Tabone CJ, Kandpal RP. Predicting candidate genes for human deafness disorders: a bioinformatics approach. BMC Genomics. 2006;7:180.

39. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13:227–32.

40. Kosti I, Jain N, Aran D, Butte AJ, Sirota M. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. Sci Rep. 2016;6:24799.

41. Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr Genomics. 2013;14:91–110.

42. Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B, et al. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. Bioinformatics. 2008;24:2894–900.

43. Kristensen AR, Gsponer J, Foster LJ. Protein synthesis rate is the predominant regulator of protein expression during differentiation. Mol Syst Biol. 2013;9:689.

44. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016;165:535–50.

45. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011;473:337–42.

46. Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ. 2014;2:e270.

47. Tuller T, Kupiec M, Ruppin E. Determinants of protein abundance and translation efficiency in *S. cerevisiae*. PLOS Comput Biol. 2007;3:e248.

48. Spangenberg L, Correa A, Dallagiovanna B, Naya H. Role of alternative polyadenylation during adipogenic differentiation: an in silico approach. PLoS One. 2013;8:e75578.

49. Laurent JM, Vogel C, Kwon T, Craig SA, Boutz DR, Huse HK, et al. Protein abundances are more conserved than mRNA abundances across diverse taxa. Proteomics. 2010;10:4209–12.

50. Kwon T, Huse HK, Vogel C, Whiteley M, Marcotte EM. Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains. J Proteome Res. 2014;13:2370–80.

51. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013;342:1100–4.

52. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–7.

53. Csárdi G, Franks A, Choi DS, Airoldi EM, Drummond DA. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. PLoS Genet. 2015;11:e1005206.

54. Mehdi AM, Patrick R, Bailey TL, Bodén M. Predicting the dynamics of protein abundance. Mol Cell Proteomics. 2014;13:1330–40.

55. Perl K, Ushakov K, Pozniak Y, Yizhar-Barnea O, Bhonker Y, Shivatzki S, et al. Reduced changes in protein compared to mRNA levels across non-proliferating tissues. BMC Genomics. 2017;18:305.

56. Perl K, Shamir R, Avraham KB. Computational analysis of mRNA expression profiling in

the inner ear reveals candidate transcription factors associated with proliferation, differentiation, and deafness. Hum Genomics. 2018;12:30.

57. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8:1765–86.

58. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

59. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

60. Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol. 2014;32:223–6.

61. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics. 2014;13:2513–26.

62. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48.

63. Scheipl F, Greven S, Küchenhoff H. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Comput Stat Data Anal. 2008;52:3283–99.

64. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, et al. Expander: from expression microarrays to networks and functions. Nat Protoc. 2010;5:303–22.

65. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res. 2003;13:773–80.

66. Hawkins RD, Bashiardes S, Powder KE, Sajan SA, Bhonagiri V, Alvarado DM, et al. Large scale gene expression profiles of regenerating inner ear sensory epithelia. PLoS One. 2007;2:e525.

67. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.

68. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008;28.

69. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.

70. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08. New York, New York, USA: ACM Press; 2008. p. 213–20.

71. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78:1–3.

72. Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence. IEEE; 2015. p. 159–66.

73. Kim J, Kim J, Lee H. An analysis of disease-gene relationship from Medline abstracts by DigSee. Sci Rep. 2017;7:40154.

74. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015;2015:bav028.

75. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2015;74:83–9.

76. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.

77. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, et al. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS One. 2011;6:e27156.

78. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29:1083–5.

79. Zhong Y, Liu Z. Gene expression deconvolution in linear space. Nat Methods. 2011;9:8–9.

80. Lopez I, Ishiyama G, Tang Y, Tokita J, Baloh RW, Ishiyama A. Regional estimates of hair cells and supporting cells in the human crista ampullaris. J Neurosci Res. 2005;82:421–31.

81. Goodyear RJ, Gates R, Lukashkin AN, Richardson GP. Hair-cell numbers continue to increase in the utricular macula of the early posthatch chick. J Neurocytol. 1999;28:851–61.

82. Geiger T, Velic A, Macek B, Lundberg E, Kampf C, Nagaraj N, et al. Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. Mol Cell Proteomics. 2013;12:1709–22.

83. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343–8.

84. Ma Y, Hof PR, Grant SC, Blackband SJ, Bennett R, Slatest L, et al. A three-dimensional digital atlas database of the adult C57BL/6J mouse brain by magnetic resonance microscopy. Neuroscience. 2005;135:1203–15.

85. Cebrián C, Borodo K, Charles N, Herzlinger DA. Morphometric index of the developing murine kidney. Dev Dyn. 2004;231:601–8.

86. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.

87. Pfister TD, Reinhold WC, Agama K, Gupta S, Khin SA, Kinders RJ, et al. Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. Mol Cancer Ther. 2009;8:1878–84.

88. Moghaddas Gholami A, Hahne H, Wu Z, Auer FJ, Meng C, Wilhelm M, et al. Global proteome analysis of the NCI-60 cell line panel. Cell Rep. 2013;4:609–20.

89. Ong S-E, Mann M. Mass spectrometry–based proteomics turns quantitative. Nat Chem Biol. 2005;1:252–62.

90. Legendere P. Model II regression user's guide, R edition. R Vignette. 1998.

91. Warton DI, Duursma RA, Falster DS, Taskinen S. smatr 3- an R package for estimation and inference about allometric lines. Methods Ecol Evol. 2012;3:257–9.

92. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA. 2001;98:5116–21.

93. Consonni V, Ballabio D, Todeschini R. Evaluation of model predictive ability by external validation techniques. J Chemom. 2010;24:194–201.

94. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6:e21800.

95. Salicrú M, Ocaña J, Sánchez-Pla A. Comparison of lists of genes based on functional profiles. BMC Boinformatics. 2011;12:401.

96. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.

97. Tebaldi T, Re A, Viero G, Pegoretti I, Passerini A, Blanzieri E, et al. Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. BMC Genomics. 2012;13:220.

98. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 2006;7:302.

99. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010;26:976–8.

100. Cargnello M, Roux PP. Activation and Function of the MAPKs and Their Substrates, the MAPK-Activated Protein Kinases. Microbiol Mol Biol Rev. 2011;75:50–83.

101. Wieduwilt MJ, Moasser MM. The epidermal growth factor receptor family: Biology driving targeted therapeutics. Cell Mol Life Sci. 2008;65:1566–84.

102. Derynck R, Zhang YE. Smad-dependent and Smad-independent pathways in TGF-β family signalling. Nature. 2003;425:577–84.

103. Mordelet F, Vert J-P. A bagging SVM to learn from positive and unlabeled examples. Pattern Recognit Lett. 2014;37:201–9.

104. Han C-P. Combining tests for correlation coefficients. Am Stat. 1989;43:211.

105. Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34 Database issue:D108-10.

106. Jain S, Maltepe E, Lu MM, Simon C, Bradfield CA. Expression of ARNT, ARNT2, HIF1 alpha, HIF2 alpha and Ah receptor mRNAs in the developing mouse. Mech Dev. 1998;73:117–23.

107. Aitola MH, Pelto-Huikko MT. Expression of Arnt and Arnt2 mRNA in developing murine tissues. J Histochem Cytochem. 2003;51:41–54.

108. Mimura J, Fujii-Kuriyama Y. Functional role of AhR in the expression of toxic effects by TCDD. Biochim Biophys Acta. 2003;1619:263–8.

109. Sobek-Klocke I, Disqué-Kochem C, Ronsiek M, Klocke R, Jockusch H, Breuning A, et al. The human gene ZFP161 on 18p11.21-pter encodes a putative c-myc repressor and is homologous to murine Zfp161 (Chr 17) and Zfp161-rs1 (X Chr). Genomics. 1997;43:156–64.

110. Orlov S V, Kuteykin-Teplyakov KB, Ignatovich IA, Dizhe EB, Mirgorodskaya OA, Grishin A V, et al. Novel repressor of the human FMR1 gene - identification of p56 human (GCC)(n)-binding protein as a Krüppel-like transcription factor ZF5. FEBS J. 2007;274:4848–62.

111. Chambers J, Rabbitts TH. LMO2 at 25 years: a paradigm of chromosomal translocation proteins. Open Biol. 2015;5:150062.

112. Celada A, Borràs FE, Soler C, Lloberas J, Klemsz M, van Beveren C, et al. The transcription factor PU.1 is involved in macrophage proliferation. J Exp Med. 1996;184:61–9.

113. Ma Q. Role of nrf2 in oxidative stress and toxicity. Annu Rev Pharmacol Toxicol. 2013;53:401–26.

114. Bailey P, Sartorelli V, Hamamori Y, Muscat GE. The orphan nuclear receptor, COUP-TF II,

inhibits myogenesis by post-transcriptional regulation of MyoD function: COUP-TF II directly interacts with p300 and myoD. Nucleic Acids Res. 1998;26:5501–10.

115. Massagué J, Wotton D. Transcriptional control by the TGF-beta/Smad signaling system. EMBO J. 2000;19:1745–54.

116. Butts SC, Liu W, Li G, Frenz DA. Transforming growth factor-beta1 signaling participates in the physiological and pathological regulation of mouse inner ear development by all-trans retinoic acid. Birth Defects Res A Clin Mol Teratol. 2005;73:218–28.

117. Li H, Corrales CE, Wang Z, Zhao Y, Wang Y, Liu H, et al. BMP4 signaling is involved in the generation of inner ear sensory epithelia. BMC Dev Biol. 2005;5:16.

118. Yoon BS, Ovchinnikov DA, Yoshii I, Mishina Y, Behringer RR, Lyons KM. Bmpr1a and Bmpr1b have overlapping functions and are essential for chondrogenesis in vivo. Proc Natl Acad Sci USA. 2005;102:5062–7.

119. Lin B, Chen GQ, Xiao D, Kolluri SK, Cao X, Su H, et al. Orphan receptor COUP-TF is required for induction of retinoic acid receptor beta, growth inhibition, and apoptosis by retinoic acid in cancer cells. Mol Cell Biol. 2000;20:957–70.

120. Romand R, Dollé P, Hashino E. Retinoid signaling in inner ear development. J Neurobiol. 2006;66:687–704.

121. Smith RJ. Use and misuse of the reduced major axis for line-fitting. Am J Phys Anthropol. 2009;140:476–86.

122. Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, et al. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 1998;9:585–98.

123. Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, et al. Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. Mol Cancer Ther. 2006;5:2606–12.

124. Torban E, Goodyer P. The kidney and ear: emerging parallel functions. Annu Rev Med. 2009;60:339–53.

125. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32:258D–261.

126. Klampfer L, Huang J, Corner G, Mariadason J, Arango D, Sasazuki T, et al. Oncogenic Ki-ras inhibits the expression of interferon-responsive genes through inhibition of STAT1 and STAT2 expression. J Biol Chem. 2003;278:46278–87.

127. Su X, Yu Y, Zhong Y, Giannopoulou EG, Hu X, Liu H, et al. Interferon-γ regulates cellular metabolism and mRNA translation to potentiate macrophage activation. Nat Immunol. 2015.

128. Goñalons E, Barrachina M, García-Sanz JA, Celada A. Translational control of MHC class II I-A molecules by IFN-gamma. J Immunol. 1998;161:1837–43.

129. Barnett B, Kryczek I, Cheng P, Zou W, Curiel TJ. Regulatory T cells in ovarian cancer: biology and therapeutic potential. Am J Reprod Immunol. 2005;54:369–77.

130. Yang R, Cai Z, Zhang Y, Yutzy WH, Roby KF, Roden RBS. CD80 in immune suppression by mouse ovarian carcinoma-associated Gr-1+CD11b+ myeloid cells. Cancer Res. 2006;66:6807–15.

131. Patankar MS, Jing Y, Morrison JC, Belisle JA, Lattanzio FA, Deng Y, et al. Potent suppression of natural killer cell response mediated by the ovarian tumor marker CA125. Gynecol Oncol. 2005;99:704–13.

132. Dürig J, Nückel H, Hüttmann A, Kruse E, Hölter T, Halfmeyer K, et al. Expression of

ribosomal and translation-associated genes is correlated with a favorable clinical course in chronic lymphocytic leukemia. Blood. 2003;101:2748–55.

133. Dua K, Williams TM, Beretta L. Translational control of the proteome: relevance to cancer. Proteomics. 2001;1:1191–9.

134. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet. 2008;40:499–507.

135. Hussain T, Mulherkar R. Lymphoblastoid cell lines: a Continuous in vitro source of cells to study carcinogen sensitivity and dna repair. Int J Mol Cell Med. 2012;1:75–87.

136. Tipney H, Hunter L. An introduction to effective use of enrichment analysis software. Hum Genomics. 2010;4:202–6.

137. Padi M, Quackenbush J. Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators. BMC Syst Biol. 2015;9:80.

138. Kotni MK, Zhao M, Wei D-Q. Gene expression profiles and protein-protein interaction networks in amyotrophic lateral sclerosis patients with C9orf72 mutation. Orphanet J Rare Dis. 2016;11:148.

139. Sun W, Qiu Z, Huang W, Cao M. Gene expression profiles and protein-protein interaction networks during tongue carcinogenesis in the tumor microenvironment. Mol Med Rep. 2017.

140. Zhang M-L, Zhou Z-H. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng. 2014;26:1819–37.

141. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18:644–52.

142. Oti M. Predicting disease genes using protein-protein interactions. J Med Genet.

2006;43:691–8.

143. Li J, Wang L, Guo M, Zhang R, Dai Q, Liu X, et al. Mining disease genes using integrated protein-protein interaction and gene-gene co-regulation information. FEBS Open Bio. 2015;5:251–6.

144. Yang P, Li X, Chua H-N, Kwoh C-K, Ng S-K. Ensemble positive unlabeled learning for disease gene identification. PLoS One. 2014;9:e97079.

145. Yang P, Li X-L, Mei J-P, Kwoh C-K, Ng S-K. Positive-unlabeled learning for disease gene identification. Bioinformatics. 2012;28:2640–7.

146. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. Impact of regulatory variation from RNA to protein. Science. 2015;347:664–7.

147. Dori-Bachash M, Shema E, Tirosh I. Coupled evolution of transcription and mRNA degradation. PLoS Biol. 2011;9:e1001106.

148. Suter R, Marcum JA. The molecular genetics of breast cancer and targeted therapy. Biologics. 2007;1:241–58.

149. Vogel C, de Sousa Abreu R, Ko D, Le S-Y, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. 2010;6.

150. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087–91.

151. Mathews MB, Sonenberg N, Hershey JW. Origins and principles of translational control. Cold Spring Harb Monogr Arch. 2000;39:1–31.

152. Cheng Z, Teo G, Krueger S, Rock TM, Koh HW, Choi H, et al. Differential dynamics of the

mammalian mRNA and protein expression response to misfolding stress. Mol Syst Biol. 2016;12:855–855.

153. Reiter T, Penman S. “Prompt” heat shock proteins: translationally regulated synthesis of new proteins associated with the nuclear matrix-intermediate filaments as an early response to heat shock. Proc Natl Acad Sci USA. 1983;80:4737–41.

154. Blais JD, Filipenko V, Bi M, Harding HP, Ron D, Koumenis C, et al. Activating transcription factor 4 is translationally regulated by hypoxic stress. Mol Cell Biol. 2004;24:7469–82.

155. Hentze M, Caughman S, Rouault T, Barriocanal J, Dancis A, Harford J, et al. Identification of the iron-responsive element for the translational regulation of human ferritin mRNA. Science. 1987;238:1570–3.

6 APPENDICES

6.1 List of publications

- **K. Perl**, R. Shamir, K. B. Avraham. Computational analysis of mRNA expression profiling in the inner ear reveals candidate transcription factors associated with proliferation, differentiation, and deafness. Human Genomics 12.1 (2018): 30.
- O. Yizhar-Barnea, C. Valensisi, K. Kishore, N. D. Jayavelu, C. Andrus, T. Koffler-Brill, K. Ushakov, **K. Perl,** Y. Noy, Y. Bhonker, M. Pelizzola, D. Hawkins. DNA methylation dynamics during embryonic development and postnatal maturation of the mouse auditory organ of Corti. bioRxiv 262832 (2018).
- K. Ushakov, T. Koffler-Brill, A. Rom, **K. Perl**, I. Utilsky, K. B. Avraham. Genome-wide identification and expression profiling of long non-coding RNAs in auditory and vestibular systems. Scientific Reports 7:8637 (2017).
- **K. Perl**, K. Ushakov, Y. Pozniak, O. Yizhar-Barnea, Y. Bhonker, S. Shivatzki, T. Geiger, K. B. Avraham, R. Shamir. Reduced changes in protein compared to mRNA levels across non-proliferating tissues. BMC Genomics 18:305 (2017).

תקציר

מטרת עבודה זו היא ניתוח מולקולות הרנ"א והחלבונים המתבטאים במערכות השמיעה ושיווי-המשקל, במטרה לזהות מנגנוני מחלה בחירשות ובהפרעות שיווי משקל. לצורך זה, הופקו נתונים על רנ"א שליח וחלבון בשבלול האוזן ובוסטיבולה בעכבר בגיל העוברי 16.5 ימים ובעכבר ביום ה-0 הבתר-לידתי (הגילאים יסומנו E16.5 ו-P0 בהתאמה).

האוזן הפנימית מורכבת משתי מערכות עיקריות: מערכת השמיעה, או בשמה האחר המערכת האודיטורית, ומערכת שיווי המשקל, המכונה גם הווסטיבולרית. על-אף הדמיון הרב ביניהן, קיימים הבדלים מבניים ותפקודיים בין השתיים. העכבר משמש זה זמן רב חיית מודל לצורך לימוד המבנה והתפקוד של האוזן הפנימית האנושית, וזאת, בין היתר, בשל היכולת להרביע אותו ולבחור צאצאים בעלי תכונות רצויות, כולל כאלה המשפיעות על שמיעה ושיווי משקל. אנו התעניינו ברקמות שבלול האוזן והוסטיבולה בעכברים בגילאי E16.5 ו-P0. אלו מתאימים לנקודות התפתחות של לפני ואחרי הרכישה של הרגישות לגירוי מכני. עד כה, הייתה רק עבודה מדעית מועטה שעסקה בהשוואת רמות ביטוי הגנים בשתי רקמות האוזן הפנימית, שבלול האוזן והוסטיבולה, לפני ואחרי שלב התפתחותי זה. השתמשנו בגישות של ניתוח מערכתי של שעתוק כדי לפענח את מסלולי הבקרה במערכת השמיעה, כשהמטרה העיקרית הייתה זיהוי גורמי שעתוק המשמשים כבקרים ראשיים של שגשוג והתמיינות.

רוב מאמרי המחקר המשווים רמות ביטוי עושים זאת עבור מידע שמופק בטכניקת אומיקס (omics) יחידה. עבור פרופיל שעתוק (transcriptomics), נהוג להשתמש במדידות של ריצוף רנ"א, ואילו עבור פרופיל חלבון (proteomics), במדידות של ספקטרומטר מסות. בניתוח מידע מסוג אומיקס יחיד מוגבלת יכולת הזיהוי של מנגנוני בקרה של תהליכים הקורים אחרי שעתוק. ניתוחים משולבים מראים כי המתאם בין רמות הביטוי של חלבון ובין רמות ביטוי של רנ"א שליח ביונקים הוא יחסית נמוך, עם קבוע מתאם של פירסון בסביבות 0.40. הסברים אפשריים למתאם נמוך זה כוללים בקרה אחרי שעתוק ורעש במדידה. אנו השתמשנו במדידות של רמות ביטוי של חלבון ורנ"א שליח ברקמות האוזן הפנימית בגיל P0, לצד אוספי נתונים אחרים של רנ"א שליח וחלבון, כדי לזהות תבנית של בקרה אחרי שעתוק המתקיימת ברקמות שאינן בשלב של שגשוג (proliferation). בניתוח עוקב השווינו העשרות ברמת החלבון וברמת הרנ"א שליח, והצענו יתרון ביולוגי אפשרי למנגנון זה.

ניתוח פרופיל השעתוק בממדים של גיל ורקמה הרחיב את הידע שלנו בנוגע להתפתחות האוזן הפנימית. מצאנו שהשבלול עשיר יותר בתפקודים נוירולוגיים ומכיל אחוז גבוה יותר של תאי שערה לעומת הוסטיבולה, ומנגד, ההתפתחות של התפיסה החושית בו מעוכבת לעומת זו בוסטיבולה. הוסטיבולה, מאידך, התגלתה כעשירה יותר בכלי דם וחדירה יותר למערכת

החיסונית. רוב גורמי השעתוק שחזינו כבקרים מרכזיים בהתמיינות הם בעלי תפקידים ידועים שתואמים את האפיון הדיכוטומי הנזכר לעיל. חלק מאלו גם נבחר בהמשך כמועמדים אפשריים בהשראת התחדשות של תאי שערה.

בהתמקדות בגנים ידועים של חרשות, מצאנו כי אלו נוטים להתבטאות מובדלת (differential expression) בין הרקמות. במהלך ההתפתחות, הן רמות הביטוי והן יחס הביטוי בין השבלול לוסטיבולה עולים בהם. הדגמנו כיצד ניתן לנצל זאת כדי לבנות מסווג (classifier) המזהה גנים נוספים כמועמדים בגרימת חרשות.

ניתוח משולב של רנ"א שליח וחלבון בגיל P0 עזר להדגים שיחס החלבון לרנ"א שליח במצב שיווי משקל משתנה בכיוון המפחית את השינוי ברמות החלבון הנגרם עקב שינויים בכמות התעתיק. מגמה זו נראתה בשני אוספי מידע נוספים, האחד של רקמות מאיברי עכבר, והשני של דגימות לימפובלסטואידים מפרימטים. באוסף מידע רביעי, של שורות תאים מסרטן הומני, לא הופיעה נטייה זו.

אנו מציעים שקיום אפקט משכך (buffering) חלקי בין השעתוק לתרגום מבטיח שחלבונים יוכלו להיבנות במהירות כתגובה לגירוי חיצוני, ומדגימים כיצד התחשבות באפקט המשכך יכולה לשפר את החיזוי של רמות חלבון מתוך רמות רנ"א שליח.

תוכן עניינים

הפקולטה לרפואה על שם סאקלר

החוג לגנטיקה מולקולארית של האדם וביוכימיה קלינית

הבנת ההתפתחות והפיזיולוגיה של האוזן הפנימית באמצעות ניתוח רמות הביטוי של גנים וחלבונים

חיבור לשם קבלת התואר "דוקטור לפילוסופיה"

מאת

קובי פרל

הוגש לסנאט של אוניברסיטת תל-אביב

דצמבר 2017

העבודה התבצעה בהנחייתם של

פרופ' קרן אברהם ופרופ' רון שמיר