GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

GENE ONTOLOGY-BASED FRAMEWORK TO ANNOTATE GENES OF HEARING

Dissertation

for the award of the degree

“Doctor rerum naturalium”

Division of Mathematics and Natural Sciences

of the Georg-August-Universität Göttingen

Submitted by

Guvanchmyrat Ovezmyradov

From Ashgabat, Turkmenistan

Göttingen, 2012

Members of Thesis Committee

Prof. Dr. Martin Göpfert (Supervisor/Reviewer)

Georg-August-University of Göttingen, Dept. of Cellular Neurobiology

Prof. Dr. André Fiala

Georg-August-University of Göttingen, Dept. of Molecular Neurobiology of Behaviour

Prof. Dr. Tobias Moser

Georg-August-University Göttingen, Dept. of Otorhynolaryngology

Additional reviewer:

Prof. Dr. Burkhard Morgenstern (Reviewer)

Georg-August-University Göttingen, Dept. of Bioinformatics

Date of the oral examination:23.10.2012

I, Guvanchmyrat Ovezmyradov, herewith declare that my doctoral thesis entitled “Gene Ontology-based framework to annotate genes of hearing” was written independently. No other sources and aids than the quoted were used.

Guvanchmyrat Ovezmyradov

Göttingen, September 2012

Table of Contents

1. Summary

As the number of discovered deafness genes and our knowledge about their role in hearing continuously increases, the need to accurately manage this genetic information is becoming more apparent. Gene Ontology (GO) is a golden standard for gene annotation and the GO database is central to bioinformatics. This thesis demonstrates how GO terms can be used to annotate auditory genes and their gene products in a systematic and sustainable manner. The GO-based framework proposed in this study facilitates the comprehensive annotation of single genes as well as the integrated analysis of high-throughput data related to hearing in multiple species. This practical framework approaches hearing on various levels of molecular complexity comprising genes, transcripts and gene networks: Combining emerging data about deafness genes with the adequate bioinformatics infrastructure helps to leverage the value of gene annotations and can thus contribute to a better understanding of hearing and hearing disorders. In addition to utilizing existing bioinformatics tools, a novel web-based application for mining the GO database with complex queries is presented and applied to extract information about auditory relevant genes. Future perspectives include further extension of this flexible framework to other data sources, such as expression databases and phenotype ontologies, to build powerful platforms for integrative data analysis. This study is a step forward towards systems biology approaches that are expected to broaden hearing research horizons and ultimately translate into novel strategies for fighting deafness.

2. Introduction

How can we conquer deafness? Comprehensive understanding of the auditory system and its mechanisms is a prerequisite for successfully fighting this disease. Structural integrity and proper function of the auditory system underlies hearing processes: Maintenance of the intricate structures, such as stereocilia, on one side, and coordinated activity of molecular players, such as the mechanotransduction machinery components, on the other [Steel & Kros, 2001]. Most of our knowledge about the genetic aspects of this complex system comes from studying deafness mutations and associated phenotypes [Resendes *et al*., 2001]. The discovery of deafness genes and the characterization of their products have provided insights how hearing is accomplished on molecular level [Dror & Avraham, 2010]. Sequencing of genomes and high-throughput screens have significantly contributed to the progress in the genomics of hearing. One of the challenges in the so-called “post-genomic” or “omics” era is to handle the massively produced genetic data. Being one of the many “omics” sciences, auditory genomics also faces this problem.

Computation, databases and the internet are central concepts in bioinformatics [Brusic, 2007]. As more time is spent by biologists for analyzing their data, online access to and analysis of biological information becomes a crucial step in biological research. Especially the analysis of data from “omics” experiments rely on the computational power provided by bioinformatics tools. In this thesis, the focus will be on the feasibility of using the Gene Ontology (GO) database [Ashburner *et al*., 2000] to catalogue known deafness genes to facilitate discoveries. After introducing the existing GO-based tools, the rationale behind the novel application for mining GO data (AGENDA) is discussed. After illustrating possible applications of GO in hearing research, broader bioinformatics strategies including multiple databases and ontologies will be discussed in context of systems biology. As more efficient genetic techniques become available and powerful bioinformatics tools to support them are developed, the search for deafness genes is expected to accelerate and augment our understanding of hearing on the molecular level.

2.1. Gene Ontology

2.1.1 Genetic databases

Genetic databases archive genetic information and enable online access to it. Archiving includes the annotation, formatting and storage of the target data. Enabling access to the database allows the user to assess its contents and, in most cases, to mine the data. For example, a user interested in the *Drosophila nompC* gene can use AmiGO [Carbon *et al*., 2009], the official browser of the Gene Ontology (GO) database [Ashburner *et al*., 2000]. There, information about its gene product can be accessed and its protein sequence can be used for BLAST [Altschul *et al*., 1990] queries. In this thesis, archiving and accessing issues related to the GO database will be examined in the context of hearing research.

From a programmatic point of view, biological databases tend to be similar in their architecture: They are usually built using three layers of software (Figure 1A) [Stein, 2003]. Bottom layer includes the database management system (DBMS) that handles database queries and presents it to the middle layer. This middle layer consists of the database access software and the web server. These programs act as a middleware by exchanging data across bottom and top layers. The middle layer can also be considered as the “brain of the database“ because it prepares the results of the user’s request. The top layer consists of a web interface that interacts with the user’s web browser using HTML web pages and thus mediates data transmission between the user and the middle layer. A competent database is achieved with the successful interplay between these integral parts for the purpose of delivering the requested information to the user. This common architectural approach was also employed in designing the Application for Mining Gene Ontology Data (AGENDA), developed within the course of this thesis.

From a programmatic point of view, the models describing the design of web applications (Figure 1B) [Aravindhan *et al*., 2009] can also be applied to the common biological databases. In the classical model, the interaction between the web browser and the web server is based solely on HTML and CSS. In addition to HTML and CSS, the novel web applications employ a technique called AJAX (Asynchronous Javascript and XML). The AJAX engine enables more user-friendly websites and more powerful data transfer. On the other hand, developing AJAX web applications is more challenging and time-consuming when compared to the classical web applications. Currently, more and more biological database are starting to employ the AJAX web applications model.

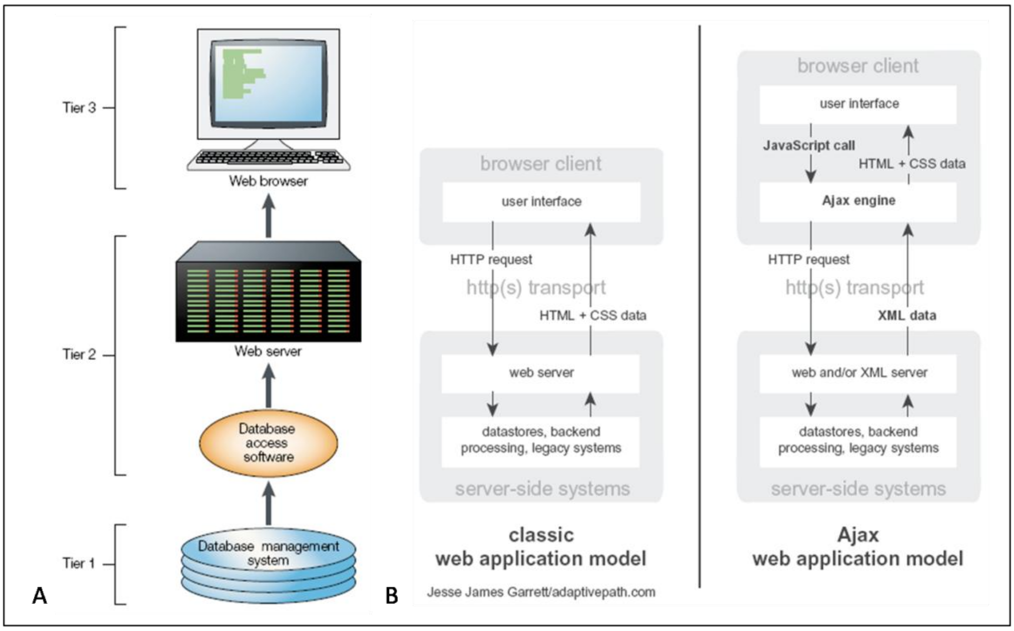


Figure 1: Design of biological databases and web applications.

A. 3-layered structure of biological databases. Biological databases usually follow an architectural pattern of 3 consequent software layers. These layers are the back-end (where the database management system, DBMS, handles data), the middle layer (where the database access software and the web server perform performes computations) and the front end (where the web interface interacts with a user) [Stein, 2003]. B. Classical vs. AJAX-based architecture of the web applications. While classical web applications are based solely on HTML and CSS, AJAX web applications also employ the AJAX (Asynchronous Javascript and XML) engine [Aravindhan et al., 2009].

2.1.2 Gene annotation

Gene annotation is the process of screening and recording literature and findings about genes. Usually gene annotation is achieved through the manual annotation that is based on the work of biocurators or experts in the field. In other cases, gene annotation can be achieved through the automatically annotation based on computational analysis or predictions. There are also web resources that combine both types of gene annotation (for example, the GO database). The term “Gene annotation” is used in GO database for simplicity and represents virtually information about gene products. The term “gene products” in the GO database mostly stand for proteins and in some cases for non-coding RNAs. As result, the terms “gene” and “gene product” will be sometimes used interchangeably in this thesis.

Web sources can be based solely on a web interface (a website based solely on the HTML pages) or constitute a fully functional biological database, based on the three-layered architecture (Figure 1A). In both cases gene annotations are presented to the visitors by the web interfaces. In some cases, gene annotations can be also extracted programmatically without web browser by using SQL or API (application programming interface) queries (for example, the GO database).

2.1.3 GO project

The GO Consortium develops and applies controlled vocabularies with the purpose of recording and providing gene annotations in a standard manner [Gene Ontology Consortium, 2001]. Standard GO terms that serve as controlled vocabularies belong to one of the three main GO categories: Molecular Function, Cellular Component, and Biological Process [Ashburner *et al.*, 2000]. For example, the human protein Myosin-VIIa is among the gene products annotated to the GO term “sensory perception of sound”. This is the name for the GO term that belongs to the Biological Process category. This GO term is assigned the ID “GO:0007605” and its synonym names are “hearing” and “perception of sound”). Every GO term has a name, an accession number and sometimes synonym names. In addition to “hearing”, Myosin-VIIa is annotated to many other GO terms. For example, this protein is associated with “motor activity” (GO term of the Molecular Function category) and “plasma membrane” (GO term of the Cellular Component category). Using such GO terms, the GO database summarizes findings (molecular functions, cellular components, and biological processes) about a certain gene products. These findings constitute gene annotations in the GO database.

Just like a single gene product can be related with multiple GO terms, a GO term can be associated with multiple gene products. Thus, the GO database contains information describing gene products, GO terms and their relationships (if any) to each other. In the GO database, a record about a gene product includes among others its symbol, full name and synonyms (if any). For instance, the human protein Myosin-VIIa is known as “MYO7A” (symbol), “Myosin-VIIa” (full name), and “USH1B” (one of many synonyms).

In the seminal paper about the GO project, the GO prototype was described as the “tool for the unification of biology” [Ashburner *et al.*, 2000]. This is due to the universal nature of the GO database: From the start, it was designed to enable describing genes from all possible species by linking them to all areas of biology with relevant GO terms. Thus, the great genetic diversity and broad biological context of the database is achieved by including numerous genomes and GO terms.

2.1.4 GO annotations

GO annotations are results of associating gene products with particular GO terms in the GO database. This association is based on and described by specific evidence. The evidence is specified by its source and type. The source of the evidences can be accessed using the evidence reference. There can be many types of evidences depending on the nature of the work or analysis that links a gene product to a specific GO term. Thus, a suitable evidence codes is selected to designate which type of the evidence supports the annotation. Therefore, the basic constituents of a gene annotation are a gene product-GO term association, evidence reference and an evidence code. If there are more than one evidence supporting the association of a gene product with a particular GO term, this will result in multiple annotations. For example, fruit fly protein Atonal is associated with hearing (GO:0007605) using two annotations (Table 1). Each of these annotations is based on distinct evidences that are described by their own references and respective evidence types. In this case, a PubMed IDs was used as the evidence reference and IMP (Inferred from Mutant Phenotype) was used as the evidence code for each annotation.

Table 1: Annotations describing association of fruit fly Atonal protein with hearing (GO:0007605).

GO term	Gene product	GO evidence code	Evidence reference
GO:0007605	ato (Atonal)	IMP	PMID:10934246
GO:0007605	ato (Atonal)	IMP	PMID:12203727

Source: GO database as of June 2012.

2.1.5 GO evidence codes

GO evidence codes in the GO database can be assigned both manually and automatically. Only "Inferred from Electronic Annotation" (IEA) is assigned automatically. All other evidence codes are assigned manually by a biocurator and belong to one of the four groups: experimental, computational analysis, author statement, and curator statement. "Not Recorded" (NR) is an obsolete evidence code. The full list of evidence codes is shown in the table 2. Detailed user guide about GO evidence codes is available in the related web page at Gene Ontology website [Gene Ontology website, "Guide to GO Evidence Codes" web page].

Table 2: GO evidence codes.

No.	GO evidence code	GO evidence code group
1	Inferred from Experiment (EXP)	Experimental
2	Inferred from Direct Assay (IDA)	Experimental
3	Inferred from Physical Interaction (IPI)	Experimental
4	Inferred from Mutant Phenotype (IMP)	Experimental
5	Inferred from Genetic Interaction (IGI)	Experimental
6	Inferred from Expression Pattern (IEP)	Experimental
7	Inferred from Sequence or Structural Similarity (ISS)	Computational analysis
8	Inferred from Sequence Orthology (ISO)	Computational analysis
9	Inferred from Sequence Alignment (ISA)	Computational analysis
10	Inferred from Sequence Model (ISM)	Computational analysis
11	Inferred from Genomic Context (IGC)	Computational analysis
12	Inferred from Biological aspect of Ancestor (IBA)	Computational analysis

13	Inferred from Biological aspect of Descendant (IBD)	Computational analysis
14	Inferred from Key Residues (IKR)	Computational analysis
15	Inferred from Rapid Divergence (IRD)	Computational analysis
16	Inferred from Reviewed Computational Analysis (RCA)	Computational analysis
17	Traceable Author Statement (TAS)	Author statement
18	Non-traceable Author Statement (NAS)	Author statement
19	Inferred by Curator (IC)	Curator statement
20	No biological Data available (ND)	Curator statement
21	Inferred from Electronic Annotation (IEA)	Automatically-assigned
22	Not Recorded (NR)	Obsolete

Source: Gene Ontology website, "Guide to GO Evidence Codes" web page. URL: http://www.geneontology.org/GO.evidence.shtml. Accessed on 02 September 2012.

2.1.6 Specialized GO annotation projects

GO terms and gene annotations may not be always as representative and up-to-date for some areas of biology as desired. In that case, the GO database may fail in fulfilling expectations of biologists interested in that field. A number of GO-associated annotations projects, initiated by related special interest groups, have addressed this caveat and specifically improved the area-specific content of the GO database [GO and GO associated projects website]. The exact scope of improving the area-specific GO content can be different. While some of these projects concentrate on a single organism, others can be dealing with several organisms in the course of their work. In some cases these projects are limited to updating the list of genes annotated to certain GO terms and supporting literature. In other cases, these projects also re-design target structured vocabularies and related GO terms themselves. While some of these projects are still active, others are apparently finished. Results of the project usually appear in the new GO database release and are described in details in a publication (Table 3).

Table 3: GO-associated annotation projects.

GO-associated annotation projects	Publications
Cardiovascular Annotation	[Lovering *et al.*, 2008; Lovering *et al.*, 2009; Alam-Faruque *et al.*, 2011]
Immune System	[Diehl *et al.*, 2007; Lovering *et al.*, 2008]
Muscle Biology	[Feltrin *et al.*, 2009]
Renal Annotation	[Alam-Faruque *et al.*, 2010; Alam-Faruque *et al.*, 2011]
Reference Genome Annotation Project	[Reference Genome Group of the GO Consortium, 2009]

2.1.7 GO and Bioinformatics

The need for powerful bioinformatics tools became more pressing with the advent of novel genetic techniques and the exponential increase of genomic data [Kumar & Dudley, 2007; Baxevanis, 2009]. Historically, the establishment of the GO Consortium coincided with the onset of whole-genome sequencing strategies and high-throughput expression profiling approaches, making GO annotations especially valuable for processing and interpreting the massively produced genomic data [Ashburner *et al.*, 2000]. As bioinformatics research continues to relate with new areas of biomedicine [Brusic, 2007], GO becomes part of many bioinformatics-driven methods. For example, GO has been implemented in studies related with disease gene prioritization [Schlicker *et al.*, 2010], gene function prediction [Mitrofanova *et al.*, 2011], genetic network analysis [Costanzo *et al.*, 2010], biomedical text mining [Rebholz-Schuhmann *et al.*, 2008], and the Semantic Web technology [Chen *et al.*, 2009]. In this thesis, interpreting candidate genes obtained from microarray screens using GO data will be one of the main themes. Functional annotation using AmiGO GO term enrichment tool [Carbon *et al.*, 2009], followed by interactome mapping using Cytoscape [Shannon *et al.*, 2003] and GOlorize plugin [Garcia *et al.*, 2007] will be performed and proposed as a part of the GO-based annotation framework.

2.1.8 GO Slims

GO Slims are charts created using GO annotations to summarize the properties of gene lists. GO Slimmers are tools that analyze the user's input with GO data and generate GO Slims as the output. They are usually used for the annotation of genomes (see, for example, figure 2) and the functional annotation of microarray data.

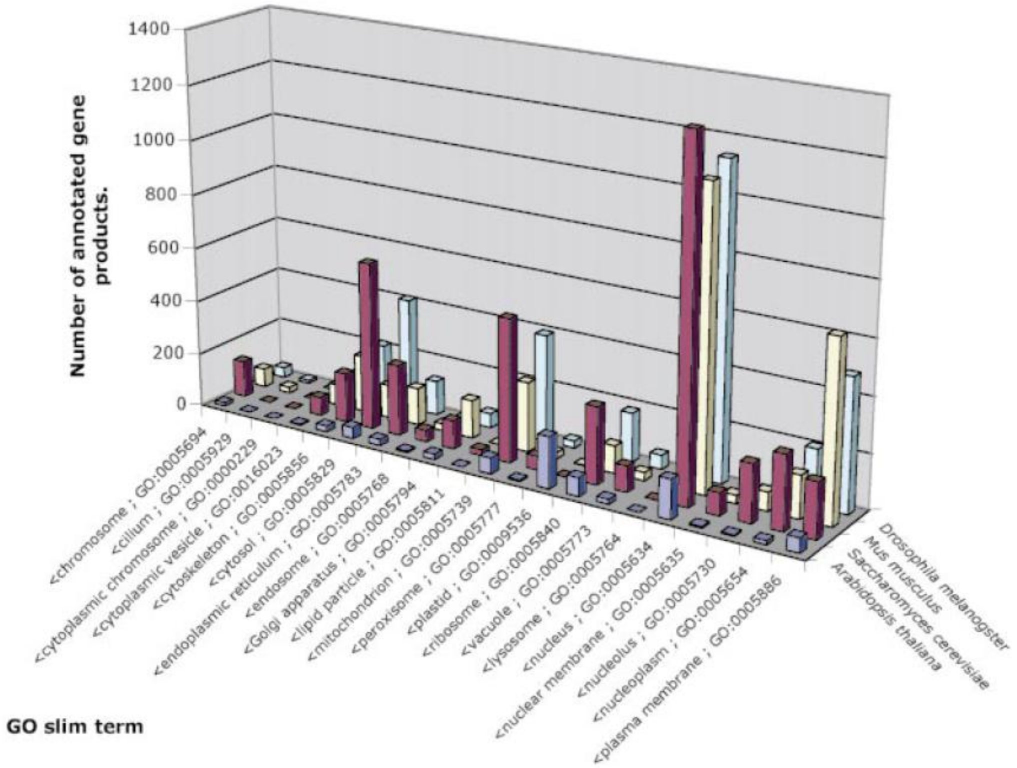


Figure 2: Genome annotation of 4 genetic model organisms using GO Slim.

Columns show how many genes are associated with each GO term in the species-specific manner. All GO terms chosen for this GO Slim set belong to the Cellular Component category. Data presented as of 1st August 2003 [Harris *et al*., 2004].

2.1.9 GO tools

Many tools have been created by the members of GO Consortium and by third parties to enable the searching, browsing and analyzing of the GO database [Gene ontology tools website]. However, some of them accept only a single GO term or gene product as an input. Since some complex biological questions cannot be answered by one GO term solely, this requires that two or more GO categories are simultaneously taken into account. Similarly, while elucidating a certain biological mechanism, a set of genes instead of a unique gene is frequently the focus of the study. Thus, using multiple GO terms as the query input shall be an important feature for users of the GO database. Most of the tools that enable gene set input perform directly the GO term enrichment analysis that produces a short list of the most significantly enriched GO terms [Beissbarth, 2006; van den Berg *et al*., 2009; Gene Ontology website, “GO Tools: Term Enrichment” web page] and summarize the output in a GO Slim. While this approach proved to be powerful in analyzing especially microarray data, it usually does not allow to query user-defined GO terms, regardless of their enrichment. In addition, vast majority of these tools do not allow performing Boolean queries using GO terms. While a previous study acknowledged the usability of Boolean operators in mining GO data [Berriz *et al*., 2003], a more straightforward way is needed for constructing the powerful queries and obtaining biologically meaningful results. Moreover, all results should be supplemented with related evidences. Finally, there is also a shortage in the graphical representation of the query results. Graphs and diagrams would significantly complement understanding of the output from the program. Thus there is a need for a program that beside the standard query modes would allow analysis of multiple GO terms independent of enrichment with a GO Slimmer or with a number of Boolean operators, summarize the results in the graphical overview, and provide links for the evidences supporting the output.

2.1.10 Bio-ontologies, data integration and Systems Biology

Managing complex biological data in a computer-readable manner and ensuring interoperability across numerous data sources can be achieved using biological ontologies [Bard & Rhee, 2004; Mi & Thomas, 2011]. Thus, the availability of relevant ontologies is a prerequisite for the biological data integration [Bodenreider, 2008]. When it comes to Systems Biology, the integration of omics data is a central concept (Figure 3) [Ge *et al*., 2003]. Relating independent datasets to each other is vital for interpreting available results en masse. Accordingly, the development of data standards has become essential for enabling integrated data analysis in systems biology [Brooksbank & Quackenbush, 2006]. GO has already established itself as the "golden standard" for describing genes products [Brazma *et al.*, 2006], serves as a model for other biomedical ontologies [Lewis, 2005], and has proved to be extremely useful in the context of Systems Biology [Costanzo *et al.*, 2010].

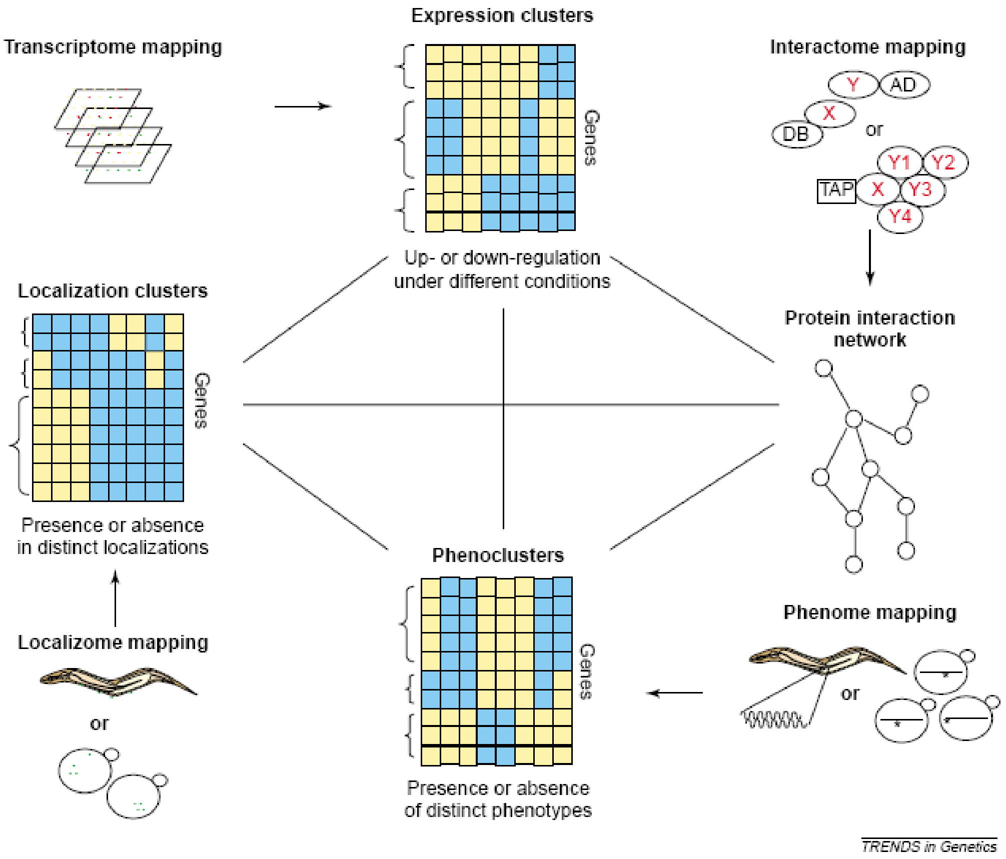


Figure 3: Integration of various omics data obtained by different high-throughput methods.

Various types of functional genomic and proteomic data from *Saccharomyces cerevisiae* and *Caenorhabditis elegans* are shown as example [Ge *et al.*, 2003].

2.2. Genes for hearing

2.2.1 Human hereditary hearing impairment

Hearing impairment is regarded as the most prevalent human sensory disease [Dror & Avraham, 2010]. Genetic factors are responsible for about half of the cases of congenital deafness [Eisen & Ryugo, 2007]. Progress in identifying and characterizing human deafness genes has yielded insights into the wide range of functions accomplished by their products in the auditory system [Dror & Avraham, 2010; May be a better reference]. Mutations in deafness genes cause hereditary hearing loss of syndromic or non-syndromic nature (Figure 4) [Resendes *et al.*, 2001]. Terms “deafness gene”, “auditory gene” and “gene for hearing” will be used in this thesis interchangeably.

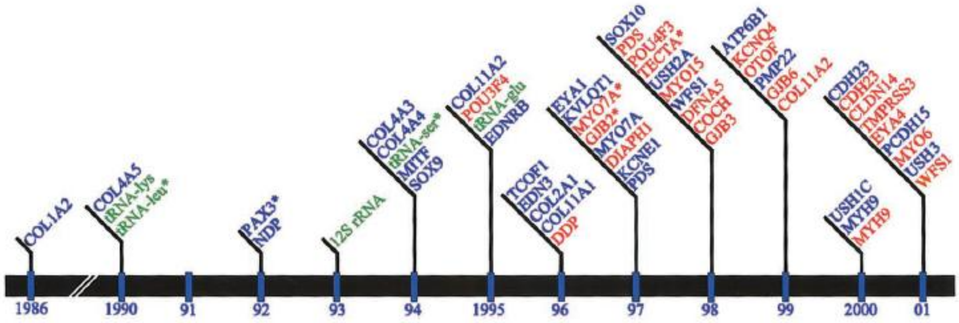


Figure 4: Chronological table for the deafness genes discovery.

Genes are classified according to their involvement in syndromic (blue), non-syndromic (red), or mitochondrial (green) deafness. Genes associated with multiple forms of deafness are marked with an asterisk (*). [Resendes *et al*., 2001]

2.2.2 From human to genetic model organisms of deafness

In parallel with the studies of the human hereditary hearing impairment, the genetics of deafness has been also intensively investigated in the mouse [Brown *et al.*, 2008], the zebrafish [Nicolson, 2005] and the fruit fly [Lu *et al.*, 2009]. In this thesis, these organisms will be collectively referred to as “genetic model organisms of deafness”. The model organisms of deafness have played a key role in dissecting molecular mechanisms underlying hearing in its normal and disease state. This undertaking was particularly accelerated by initial findings obtained from the sequencing of the fruit fly [Adams *et al.*, 2000], human [Lander *et al.*, 2001; Venter *et al.*, 2001], mouse [Mouse Genome Sequencing Consortium *et al.*, 2002] and partially zebrafish [Ekker *et al.*, 2007] genomes. The genomes provided the valuable platform for developing novel experimental methods (eg, high-throughput techniques) and investigating hearing and other biological processes. Each of these genomes has a dedicated genetic database (FlyBase for fruit fly, UniProt for human, MGI for mouse and ZFIN for zebrafish) that makes its genomic contents available and presents gene-specific findings including the chromosomal localization, molecular function, and expression pattern. (Table 4). Being a member of the GO Consortium, all these databases submit their gene annotations (records related with Cellular Component, Molecular Function, and Biological Process) to the GO database. Thus, the GO database is the universal source combining findings obtained from humans and genetic model organisms of deafness.

The focus of this thesis will be on tracing deafness genes in humans as well as in genetic model organisms of deafness using their corresponding databases within a novel GO-based annotation framework. Since its very beginning, the GO database, among others, was meant to become a much needed link for comparative genomic analyses [Ashburner *et al.*, 2000]. As the search for deafness genes gains momentum, the need for accessing and comparing findings across species becomes more apparent. This thesis will approach this objective in the context of hearing research and demonstrate the usability of the GO database in transferring knowledge about deafness across target genomes.

Table 4: Genetic databases for human, mouse, zebrafish and fruitfly.

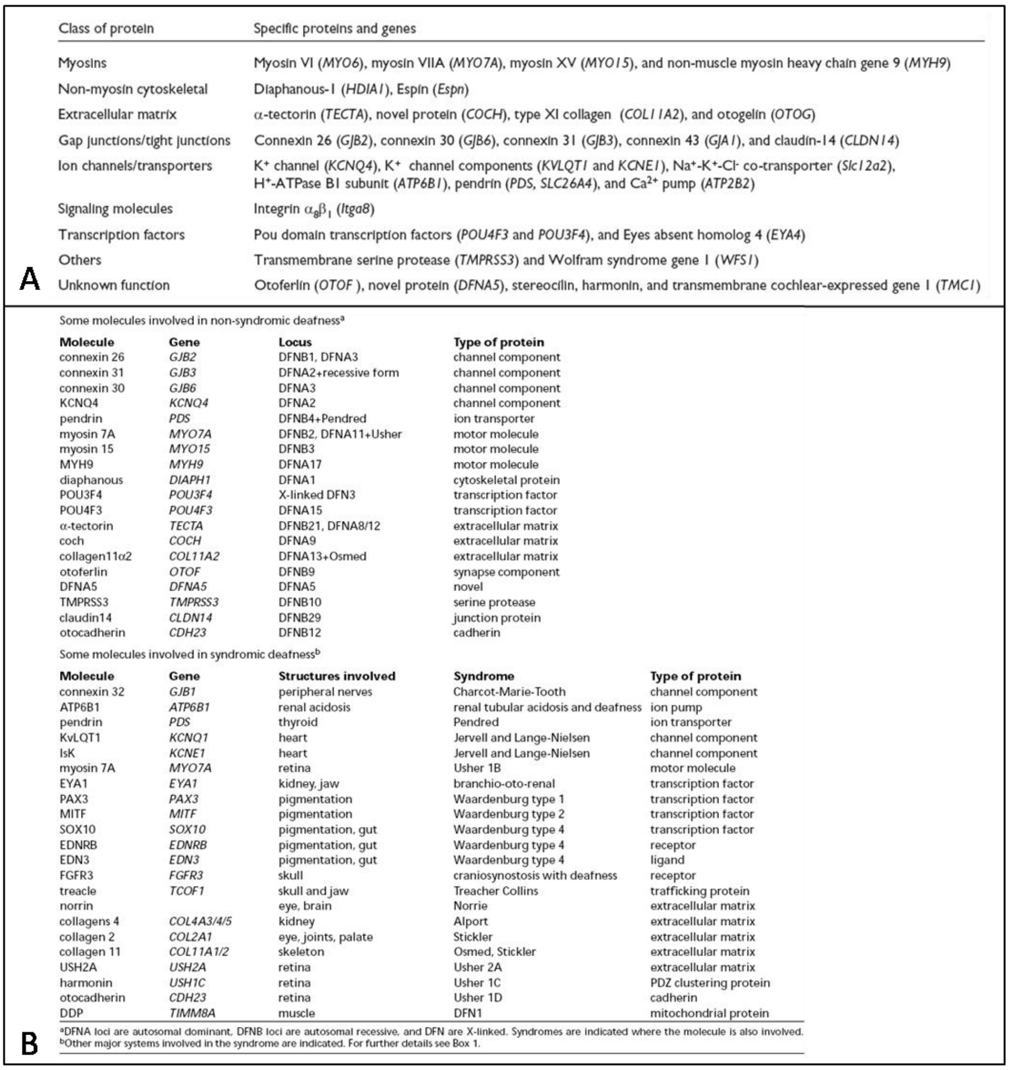
Species	**Related database and the reference**	**URL**
Human	Universal Protein Resource (UniProt) [UniProt Consortium, 2012]	http://www.uniprot.org/
Mouse	Mouse Genome Informatics (MGI) [Blake *et al.*, 2011]	http://www.informatics.jax.org/
Zebrafish	Zebrafish Information Network (ZFIN) [Bradford *et al.*, 2011]	http://zfin.org/
Fruit fly	FlyBase [McQuilton *et al.*, 2012]	http://flybase.org/

URLs as of 3 September 2012.

2.2.3 Functional categorization of auditory genes

Following the identification and characterization of novel deafness genes, many reviews about progress in this field have been published. These reviews usually describe known deafness genes, related forms of human deafness (if any), and other details. Historically, most of the research on genetic deafness has concentrated on human and mouse. Currently, most of the findings about genetic deafness derive from the studies on these organisms. As a result, when it comes to the genetic basis of deafness, their genomes are the most understood. In concordance with this, most of the reviews about genetic deafness present findings related with these two species. These reviews can either target entire forms of deafness (for example, a publication by Resendes *et al*. [Resendes *et al*., 2001]) or focus only on specific forms of deafness such as non-syndromic deafness (for example, a publication by Hilgert *et al*. [Hilgert *et al*., 2009]). Many of these reviews not only present deafness genes but group them into certain categories. This classification can be based on various criteria such as discovery year (for example, a publication by Resendes *et al*. [Resendes *et al*., 2001]), chromosomal location (for example, a publication by Dror & Avraham [Dror & Avraham, 2010]), expression pattern (for example, a publication by Hilgert *et al*. [Hilgert *et al*., 2009]) and functional characteristics (for example, a publication by Steel & Kros [Steel & Kros, 2001]). In one review, the author commented "Grouping the genes discovered to be etiologic in deafness disorders into functional categories begins the process of understanding their role in hearing" [Morton, 2002]. Accordingly, the focus in this thesis will be primarily on the functional classification (Table 5) and secondarily on the chronological classification (Figure 4). Reviews that provide functional classification of deafness genes employ certain representative categories (Table 6). These categories vary in their specificity and encompass diverse molecular aspects related with hearing. Although the usage of such functional classification varies across the reviews in types and numbers of the chosen categories, their diversified usage in general is an established practice of presenting an overview of the genetic basis of deafness.

Table 5: Functional classification of deafness genes.



Class of protein	Specific proteins and genes
Myosins	Myosin VI (*MYO6*), myosin VIIA (*MYO7A*), myosin XV (*MYO15*), and non-muscle myosin heavy chain gene 9 (*MYH9*)
Non-myosin cytoskeletal	Diaphanous-1 (*HDIA1*), Espin (*Espn*)
Extracellular matrix	α-tectorin (*TECTA*), novel protein (*COCH*), type XI collagen (*COL11A2*), and otogelin (*OTOG*)
Gap junctions/tight junctions	Connexin 26 (*GJB2*), connexin 30 (*GJB6*), connexin 31 (*GJB3*), connexin 43 (*GJA1*), and claudin-14 (*CLDN14*)
Ion channels/transporters	K$^+$ channel (*KCNQ4*), K$^+$ channel components (*KVLQT1* and *KCNE1*), Na$^+$-K$^+$-Cl$^-$ co-transporter (*Slc12a2*), H$^+$-ATPase B1 subunit (*ATP6B1*), pendrin (*PDS, SLC26A4*), and Ca^{2+} pump (*ATP2B2*)
Signaling molecules	Integrin $\alpha_8\beta_1$ (*Itga8*)
Transcription factors	Pou domain transcription factors (*POU4F3* and *POU3F4*), and Eyes absent homolog 4 (*EYA4*)
Others	Transmembrane serine protease (*TMPRSS3*) and Wolfram syndrome gene 1 (*WFS1*)
Unknown function	Otoferlin (*OTOF*), novel protein (*DFNA5*), stereocilin, harmonin, and transmembrane cochlear-expressed gene 1 (*TMC1*)

Some molecules involved in non-syndromic deafness[a]

Molecule	Gene	Locus	Type of protein
connexin 26	*GJB2*	DFNB1, DFNA3	channel component
connexin 31	*GJB3*	DFNA2+recessive form	channel component
connexin 30	*GJB6*	DFNA3	channel component
KCNQ4	*KCNQ4*	DFNA2	channel component
pendrin	*PDS*	DFNB4+Pendred	ion transporter
myosin 7A	*MYO7A*	DFNB2, DFNA11+Usher	motor molecule
myosin 15	*MYO15*	DFNB3	motor molecule
MYH9	*MYH9*	DFNA17	motor molecule
diaphanous	*DIAPH1*	DFNA1	cytoskeletal protein
POU3F4	*POU3F4*	X-linked DFN3	transcription factor
POU4F3	*POU4F3*	DFNA15	transcription factor
α-tectorin	*TECTA*	DFNB21, DFNA8/12	extracellular matrix
coch	*COCH*	DFNA9	extracellular matrix
collagen11α2	*COL11A2*	DFNA13+Osmed	extracellular matrix
otoferlin	*OTOF*	DFNB9	synapse component
DFNA5	*DFNA5*	DFNA5	novel
TMPRSS3	*TMPRSS3*	DFNB10	serine protease
claudin14	*CLDN14*	DFNB29	junction protein
otocadherin	*CDH23*	DFNB12	cadherin

Some molecules involved in syndromic deafness[b]

Molecule	Gene	Structures involved	Syndrome	Type of protein
connexin 32	*GJB1*	peripheral nerves	Charcot-Marie-Tooth	channel component
ATP6B1	*ATP6B1*	renal acidosis	renal tubular acidosis and deafness	ion pump
pendrin	*PDS*	thyroid	Pendred	ion transporter
KvLQT1	*KCNQ1*	heart	Jervell and Lange-Nielsen	channel component
IsK	*KCNE1*	heart	Jervell and Lange-Nielsen	channel component
myosin 7A	*MYO7A*	retina	Usher 1B	motor molecule
EYA1	*EYA1*	kidney, jaw	branchio-oto-renal	transcription factor
PAX3	*PAX3*	pigmentation	Waardenburg type 1	transcription factor
MITF	*MITF*	pigmentation	Waardenburg type 2	transcription factor
SOX10	*SOX10*	pigmentation, gut	Waardenburg type 4	transcription factor
EDNRB	*EDNRB*	pigmentation, gut	Waardenburg type 4	receptor
EDN3	*EDN3*	pigmentation, gut	Waardenburg type 4	ligand
FGFR3	*FGFR3*	skull	craniosynostosis with deafness	receptor
treacle	*TCOF1*	skull and jaw	Treacher Collins	trafficking protein
norrin		eye, brain	Norrie	extracellular matrix
collagens 4	*COL4A3/4/5*	kidney	Alport	extracellular matrix
collagen 2	*COL2A1*	eye, joints, palate	Stickler	extracellular matrix
collagen 11	*COL11A1/2*	skeleton	Osmed, Stickler	extracellular matrix
USH2A	*USH2A*	retina	Usher 2A	extracellular matrix
harmonin	*USH1C*	retina	Usher 1C	PDZ clustering protein
otocadherin	*CDH23*	retina	Usher 1D	cadherin
DDP	*TIMM8A*	muscle	DFN1	mitochondrial protein

[a] DFNA loci are autosomal dominant, DFNB loci are autosomal recessive, and DFN are X-linked. Syndromes are indicated where the molecule is also involved.
[b] Other major systems involved in the syndrome are indicated. For further details see Box 1.

A. Table focusing mostly on human and mouse genes involved in non-syndromic deafness [from Parkinson & Brown, 2002]. B. Table describing human deafness genes and associated details [from Steel & Kros, 2001].

Table 6: Functional categories applied to deafness genes.

Functional category	Publications
Myosins	[Parkinson & Brown, 2002]
Non-myosin cytoskeletal; Cytoskeletal protein	[Steel & Kros, 2001; Parkinson & Brown, 2002]
Extracellular matrix	[Hilgert *et al.*, 2009; Resendes et al., 2001; Steel & Kros, 2001; Parkinson & Brown, 2002;]

Gap junctions/tight junctions; Junction protein; Cadherin; Gap junction proteins: the connexins	[Steel & Kros, 2001; Morton, 2002; Parkinson & Brown, 2002; Eisen & Ryugo, 2007]
Ion channels/transporters; Channel component; Ion transporter	[Steel & Kros, 2001; Parkinson & Brown, 2002]
Signaling molecules	[Parkinson & Brown, 2002]
Transcription factors	[Resendes *et al*., 2001; Steel & Kros, 2001; Parkinson & Brown, 2002; Hilgert *et al*., 2009]
Others; Miscellaneous	[Resendes *et al*., 2001; Parkinson & Brown, 2002]
Unknown function	[Parkinson & Brown, 2002; Eisen & Ryugo, 2007]
Motor molecule	[Steel & Kros, 2001]
Synapse component	[Steel & Kros, 2001]
Novel	[Steel & Kros, 2001]
Serine protease	[Steel & Kros, 2001]
Ion pump	[Steel & Kros, 2001]
Receptor	[Steel & Kros, 2001]
Ligand	[Steel & Kros, 2001]
Trafficking protein	[Steel & Kros, 2001]
PDZ clustering protein	[Steel & Kros, 2001]
Mitochondrial protein	[Steel & Kros, 2001; Morton, 2002; Hilgert *et al*., 2009]
Hair-cell structure; maintenance of hair cell function	[Resendes *et al*., 2001; Morton, 2002]
Ion homeostasis; Endolymph ion homeostasis; Hair cell ion homeostasis	[Resendes *et al*., 2001; Hilgert *et al*., 2009]
Modifier genes	[Morton, 2002]
Tectorial membrane anchoring	[Eisen & Ryugo, 2007]
Stereocilia	[Eisen & Ryugo, 2007]
Outer hair cell electromotility	[Eisen & Ryugo, 2007]
Hair cell exocytosis	[Eisen & Ryugo, 2007]

Cell surface proteolytic enzyme	[Eisen & Ryugo, 2007]
Endolymph potassium secretion	[Eisen & Ryugo, 2007]
Melanocyte	[Eisen & Ryugo, 2007]
Hair bundle morphogenesis proteins	[Hilgert *et al*., 2009]
Proteins with poorly understood function	[Hilgert *et al*., 2009]

Similar categories are presented as a single unit. 5 underlined categories (Cytoskeletal protein, Extracellular matrix, Gap junctions, Ion channels, Synapse component) were used as an example input for the GO-based data mining program (See the Results part, figure 14).

2.2.4 Annotation of auditory genes

The most prominent and up-to-date web resources that annotate auditory genes are the Hereditary Hearing Loss Homepage (HHH), the Homepage of Hereditary Hearing Impairment in Mice (HHHM), Online Mendelian Inheritance in Man (OMIM), and the Gene Ontology (GO) database (Table 7). It is possible to come across these web sources while reading reviews about the genetics of hearing. While the goal and scope of these resources is different, they resemble each other in one basic feature: recording and providing a list of deafness genes accompanied by links for the supporting literature. While HHH and HHHM are species specific, OMIM includes data on humans and mice. All of these web sources provide information about the diseases in which the genes are involved. In contrast, the GO database encompasses auditory gene annotations from many genomes but doesn’t include information about the associated diseases. These properties are among the main advantages and disadvantages of using the GO database for the auditory gene annotation (See the Results part, table 14). Since this thesis was about investigating hearing in humans and genetic model organisms of deafness, the GO database was the only suitable web resource. For this reason, despite its disadvantages, the GO database was chosen the as the basis of this study.

Table 7: The most prominent and up-to-date web resources providing information about deafness genes.

Website name	URL	Species
The Hereditary Hearing Loss Homepage (HHH)	http://hereditaryhearingloss.org	Human
The Hereditary Hearing Impairment in Mice (HHIM)	http://hearingimpairment.jax.org/index.html	Mouse
Online Mendelian Inheritance in Man (OMIM)	http://www.ncbi.nlm.nih.gov/omim	Human, mouse
Gene Ontology (GO) database	http://www.geneontology.org	Numerous

URLs as of 3 September 2012.

3. Aim of this study

3.1. Genome-level investigation

3.1.1 Development of Application for mining Gene Ontology data (AGENDA) and its usage in hearing research

3.1.1.1 Development of AGENDA

There are numerous bioinformatics programs focusing on the GO database with a common purpose of enabling effective usage of this source [Gene ontology tools website]. Their difference lies in their specific functional aspects and approaches. Still, there is a need for a program that beside the standard query options would enable analysis of multiple GO terms with GO Slimmer and Boolean queries independent of enrichment, while being able to present the graphical overview of results and provide links to the related evidences. To address this issue, a novel web-based tool AGENDA (Application for mining Gene Ontology data) was developed. While the name "AGENDA" is used here as the abbreviation, it also implies the ability to mine GO data in accordance with the user's agenda using the user-specified GO terms. AGENDA simultaneously accesses multiple GO terms and executes complex queries to compare lists of associated gene products using GO Slimmer and Boolean operators. The goal of this application was not to replace the existing GO-based tools, but to complement them with a new interface that offers new options for mining the GO database. In this way, AGENDA is anticipated to facilitate efficient usage of GO information, including, but not limited to, auditory gene annotations. To demonstrate this usage, AGENDA was applied to mine GO data associated with hearing.

3.1.1.2 Functional categorization of auditory genes with AGENDA

Using AGENDA, human and mouse auditory gene products were functionally categorized with GO Slimmer. In addition to the species-specific functional categorization of auditory products (gene products annotated to the GO term "hearing", denoted by "GO:0007605"), an interspecies (between humans and mice) comparison of categorizations was performed.

3.1.2 Manual gene annotation with the Auditory Gene Ontology Annotation (AGOA) project

3.1.2.1 Improving lists of genes annotated to hearing in the GO database

As the number of discovered auditory genes steadily grows, so does the need to properly record their accumulating annotations in genetic databases. The stored shall be accessible not only by web browsers but also by data mining programs. The GO database satisfies these criteria and contains annotations about auditory genes in multiple species. Bioinformatics applications enable automated access to – and queries of – this database. This thesis added AGENDA (Application for Mining Gene Ontology data) to the list of these applications and applied this novel tool to access GO annotations related with hearing. While investigating the genetics of hearing, this thesis limited its scope to humans and the model organisms of hearing (namely, mouse, zebrafish and fruit fly). However, it became apparent that the related gene lists annotated to hearing (GO:0007605) in the GO database do not include some of the known auditory genes. This situation has a negative effect on the usability of the GO database in hearing research in general and on the results obtained with AGENDA in this thesis. Still, there is need for increasing the quality of GO annotations related with hearing became self-evidence.

Extraction of findings about genes from the literature and storing it in the GO database is an elaborate process. This also applies for auditory genes, it is that some of them evade annotation. Each species-specific database participating in the GO Consortium is dedicated to a distinct species and responsible for the gene annotations in the respective genome. Due to the time constraints, it is difficult for biocurators of these databases to capture all information available in the literature about genes. In addition, due to the broad scope of the GO database, it is impossible for them to be experts in every biological field. Continuous publication of new studies and rapid accumulation of genetic data makes annotation of each gene in the GO database a constantly active process with many challenges. The constraints described above are among numerous obstacles that result in the absence of some annotations in the GO database. This issue is also relevant to the part of the GO database related with hearing (GO:0007605). Auditory Gene Ontology Annotation (AGOA) project was initiated to address this issue by supporting biocurators in annotation of auditory genes and to provide an overview about the state of the art in the field using the resulting up-to-data GO data. This was expected to be achieved by the joint endeavor of the research community and biocurators.

The starting aim of the AGOA project was to improve the lists of the human, mouse, zebrafish and fruit fly auditory genes in the GO database. The work included checking the pre-existing lists of auditory genes and adding new ones. Since the related data in the GO database originates from the species-specific database, the effort directly focused on the involved databases. Updated information from these databases was expected to ensure as much as possible the accuracy and completeness of gene lists annotated to hearing in the GO database.

3.1.2.2 Revision of the evidences for genes annotated to hearing in the GO database

Since there can more than one study showing involvement of a certain gene in hearing, it is important to accurately record as much as possible the available evidence. Otherwise, some genes in the GO database can be correctly annotated to hearing but still lack some of the important references. An improved gene list can be biologically meaningful only with adequate evidences. For this reason, the AGOA was also aimed to revise the related evidences in order to include more complete and up-to-date references. This work was done in parallel to improving the list of the auditory genes themselves. The revision included checking pre-existing references for auditory genes and adding new evidences (if any).

3.1.2.3 Chronological overview of the auditory gene discoveries

Auditory genes can be classified according to their discovery years (for example, see a publication by Resendes *et al*. [Resendes *et al*., 2001]). This classification results in the chronological overview of the discoveries (Figure 4). Improvement of the auditory gene list and revision of the related evidences in the GO database was expected to provide chronological data required for such a classification. As result, the final aim of the AGOA project was to use gathered GO annotations to obtain a timeline of discoveries in the target species. This perspective was expected to provide an up-to-date review of achievements covering all species and insights into factors involved, along with the interspecies comparison of the progress.

In summary, the AGOA project included improvement of auditory genes lists, revision of relevant evidences and chronological overview of related discoveries. The study was conducted separately for each species. While the tasks related with fruitfly and zebrafish auditory genes were essentially finished and described in this thesis, the parts including human and mouse auditory genes are still in progress.

3.2. Transcriptome-level investigation

3.2.1 Functional annotation of candidate auditory genes using the AMIGO GO Term Enrichment tool

Recent microarray screen by Senthilan *et al*. [Senthilan *et al*., 2012] resulted into the identification of 274 candidate *Drosophila* auditory organ genes. Functional annotation of these genes using the AMIGO GO term enrichment tool [Carbon *et al*., 2009] was performed as the last step in the microarray data analysis that included normalization, significance testing and clustering. This GO term enrichment analysis was expected to support the evaluation of the findings by providing a list of the significantly overrepresented GO terms in the candidate gene list.

3.3. Interactome-level investigation

3.3.1 Reconstruction of the auditory gene network using Cytoscape

Reconstruction of the auditory gene network was aimed to gain insights into the interactions between molecular components of the auditory system. Moreover, this study was expected to result into the identification candiate auditory genes.

3.3.2 Annotation of the auditory gene network using the GOlorize plugin

In addition to reconstruction of the auditory gene network, its subsequent annotation using the GOlorize plugin was expected to provide more in-depth understanding of the network.

3.4. Developing Gene Ontology-based framework to annotate genes of hearing

3.4.1 Structure of the GO-based framework to annotate genes of hearing

The aims described above approached hearing on different omics levels and shaped the workflow of the thesis (Figure 5). This workflow describes information transfer throughout the thesis and reminds the central dogma of molecular biology [Crick, 1970]. The GO database played a central role in this work and provided useful means for deciphering the auditome. Final step of this thesis was to integrate the methods employed in previous steps within the "Gene Ontology-based framework to annotate genes of hearing". The rationale was to relate different GO-based annotations methods to each other by demonstrating how results of a conducted study (for example, microarray screen) could be used effectively as input for a subsequent study (for example, gene network analysis). Regarding method as components of a single annotation framework was expected to streamline their evaluation in the light of the sequential procedures associated with hearing research.

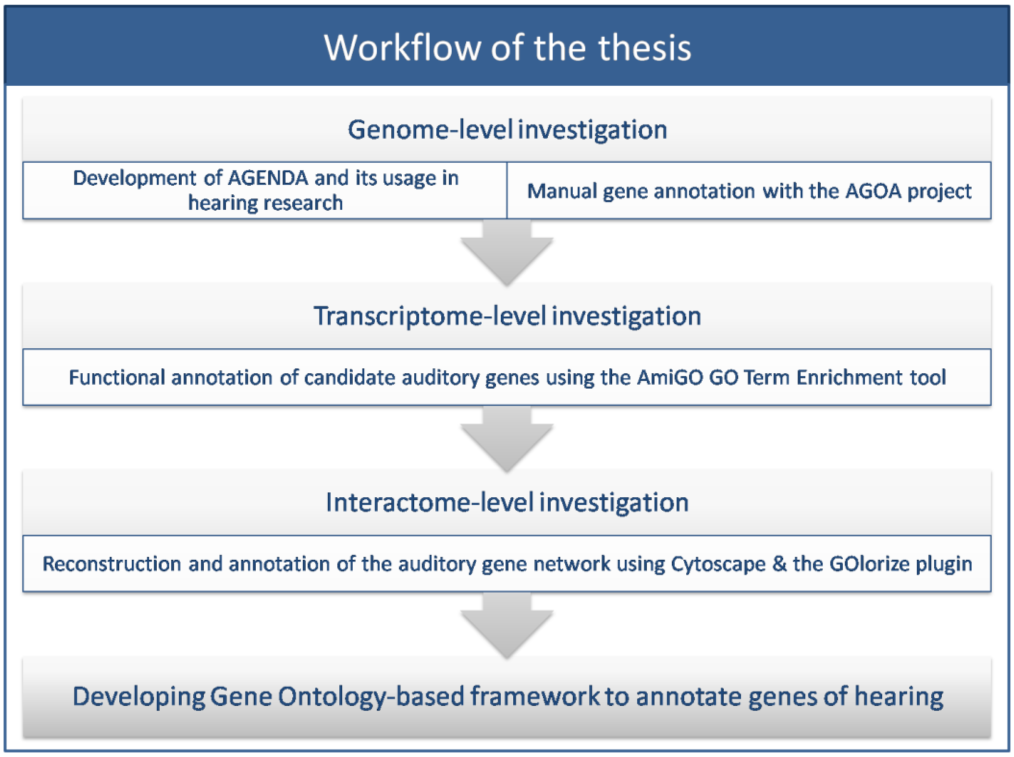

Workflow of the thesis
Genome-level investigation
Development of AGENDA and its usage in hearing research
Manual gene annotation with the AGOA project
Transcriptome-level investigation
Functional annotation of candidate auditory genes using the AmiGO GO Term Enrichment tool
Interactome-level investigation
Reconstruction and annotation of the auditory gene network using Cytoscape & the GOlorize plugin
Developing Gene Ontology-based framework to annotate genes of hearing

Figure 5: The workflow of the thesis.

This thesis demonstrates and evaluates the role of the GO database in hearing research while dealing with auditomics on three different levels (genomics, transcriptomics and interactomics). The workflow illustrates the order of the conducted studies, reminiscent of the central dogma of molecular biology [Crick, 1970]. In addition to applying established methods for different types of gene annotation, a novel web-based tool “AGENDA” was developed and applied as a part of this thesis. Combining various GO-based methods used in this work within the “Gene Ontology-based framework to annotate genes of hearing” culminates the thesis and opens a door for more powerful bioinformatics approaches to investigate hearing.

Following the terminology used in naming kinome [Manning *et al*., 2002], olfactome [Galizia *et al*., 2010] and ion channelome [Gabashvili *et al*., 2007], this thesis suggests to call the entire set of auditory genes within a genome an “auditome”. Hearing research covering entire auditomes is named accordingly “auditomics”. This designation would be consistent with the paradigm shift observed in the field with the advent of high-throughput screens relating to thousands of genes at once. Aim of this thesis was to use the GO database as means of annotating and investigating complete auditomes of several species and to apply various GO-based bioinformatics tools to aid auditomics.

3.4.2 Evaluating usability of the GO database in hearing research

The strong and weak points of the GO database were expected to directly affect the usability of the GO-based framework to annotate genes. That is why another aim in this thesis was to define the benefits and complications of using the GO database in gene annotation in general and particularly in hearing research.

3.4.3 Approaching challenges and potential of Systems biology of hearing

The GO-based annotation framework to annotate genes of hearing was also planned to be a step towards to the systems biology of hearing. Thus, the final aim of this thesis was to discuss the framework in the context of the systems biology and discuss related implications for hearing hearing. All tasks in this thesis are summarized in the table 8.

Table 8: Tasks in this thesis.

1.	**Genome-level investigation**
1.1.	Development of Application for mining Gene Ontology data (AGENDA) and its usage in hearing research
1.2.	Manual gene annotation with the Auditory Gene Ontology Annotation (AGOA)
2.	**Transcriptome-level investigation**
2.1.	Functional annotation of candidate auditory genes using the AMIGO GO Term Enrichment tool
3.	**Interactome-level investigation**
3.1.	Reconstruction of the auditory gene network using Cytoscape
3.2.	Annotation of the auditory gene network using the GOlorize plugin
4.	**Developing Gene Ontology-based framework to annotate genes of hearing**
4.1.	Structure of the GO-based framework to annotate genes of hearing
4.2.	Evaluating usability of the GO database in hearing research
4.3.	Approaching challenges and potential of Systems biology of hearing

4. Material and methods

4.1. Genome-level investigation

4.1.1 Development of Application for mining Gene Ontology data (AGENDA) and its usage in hearing research

4.1.1.1 Development of AGENDA

AGENDA (Application for mining Gene Ontology data) was developed using the XAMPP software suite 1.7.1 for Linux [XAMPP software website]. This platform-independent software suit that combines the power of Apache, MySQL, PHP, and Perl software is widely used for developing web applications. The GO database is obtained from the GO database archive [Gene Ontology database archive] as a MySQL dump file and deployed in the internal MySQL server of AGENDA. While server-side scripting that uses complex SQL queries is accomplished using PHP, JavaScript serves for client-side scripting. Web pages are created using HTML and CSS. Cross-browser compatibility of the web interface was successfully verified on common web browsers. While query results can be downloaded as CSV files, Google Chart Tools [Google Chart Tools website] are used to dynamically generate the charts.

Summary of the architecture and system requirements

Project name: AGENDA (Application for mining Gene Ontology data)

Project homepage: http://sourceforge.net/projects/bioagenda

Online version: http://bioagenda.uni-goettingen.de

Operating systems: platform independent

Programming languages: PHP and JavaScript

Compliance with web standards: Valid XHTML 1.0 Strict and CSS level 2.1

Browser compatibility: all common web browsers supported

Software requirements: Apache, PHP, and MySQL

Other requirements: Google Chart API and the local GO database

The source code and the documentation of the software: freely available in the project homepage

License: GNU GPL version 3

Any restriction to use by non-academics: license needed

4.1.1.2 Functional categorization of auditory genes with AGENDA

Using the GO Slimmer page of AGENDA, a table describing functional categories related with human gene products implicated in hearing (GO:0006915) (Table 6) was produced. 5 representative functional classes (Cytoskeletal protein, Extracellular matrix, Gap junctions, Ion channels, Synapse component), each specified by a distinct GO term (GO:0005856, GO:0031012, GO:0005921, GO:0005216, GO:0045202), were used to categorize gene products annotated to hearing. GO Slimmer was used to retrieve data from the GO database and to calculate how many of auditory gene products are associated with each functional class. The resulting GO Slim summarizes the categorization and can be used for follow-up Boolean queries.

4.1.2 Manual gene annotation with the Auditory Gene Ontology Annotation (AGOA) project

4.1.2.1 Improving lists of genes annotated to hearing in the GO database

Currently, the part of the GO database related to hearing (GO:0006915) is incomplete since some auditory genes are missing there. Inspired by previous GO-associated annotation projects [GO and GO Associated Projects Website] (Table 3), the Auditory Gene Ontology Annotation (AGOA) project was started to increase the quality of the gene annotations in the GO database associated with hearing. More precisely, the objective of this project was to annotate auditory genes in humans and the model organisms (including *Drosophila*, zebrafish and mouse) of deafness using the Gene Ontology annotation best practices [Gene Ontology website, “GO Annotation Policies and Guidelines” web page].

The first step in the plan was first to prepare a comprehensive list of auditory genes for each mentioned species. The author of this thesis was in charge of this task and conducted extensive literature review to obtain a preliminary gene list. The second step was to discuss this preliminary list with the experts in these fields and agree on a consensus gene list. Third step was to use this species-specific gene list to revise together with the biocurators of the related genetic database (Table 4) the preexisting gene list annotated to hearing (GO:0006915) in that database (Table 9). This was expected to make each of the revised gene lists available in the form of the updated gene list in the species-specific database. Since all these databases supply periodically their updated GO annotations to the GO database, this synchronization was expected to ultimately improve the overall auditory-relevant content of the GO database.

Table 9: Collaborators in the AGOA project.

Species	Expert in the experimental field	Biocurator in the related genetic database
Fruit fly	Prof. Martin Göpfert (University of Göttingen), Prof. Daniel Eberl (University of Iowa)	Susan Tweedie (FlyBase)
Mouse	Prof. Tobias Moser (University of Göttingen), Prof. Ulrich Mueller (Scripps Research Institute, La Jolla, California), Prof. Karen Steel (Welcome Truss Sanger Institute, Hinxton, UK)	Harold Drabkin (MGI)
Zebrafish	Prof. Teresa Nicholson (OHSU)	Doug Howe (ZFIN)
Human	Pending.	Emily Dimmer (UniProt)

4.1.2.2 Revision of the evidences for genes annotated to hearing in the GO database

In addition to the auditory gene products themselves, the references used in the GO evidences to link them to hearing were subject to revision and update. In case of fruit fly auditory gene products, these pre-existing references were always in forms of publications. Thus, the corresponding list of auditory gene products was based on the manual annotation. However related GO data in other target species also gene annotations that were based on computational analysis and as result were automatically assigned a GO evidence code such as Inferred from Sequence or Structural Similarity (ISS). Evidences such as ISS can be confusing and questionable for some users of the GO database. In addition, tracing and evaluating the source of such evidences is proved to be extremely difficult. The aim was to check the consistency of the pre-existing evidences, add new evidences (if available), and substitute automatically assigned GO evidence codes with those assigned manually (if available and applicable).

4.1.2.3 Chronological overview of the auditory gene discoveries

The improved auditory lists and the revised evidences were used as input data to obtain a chronological overview of the auditory gene discoveries. Auditory genes which annotation was based solely on evidences obtained from computational analyses were excluded from the input. As result, only genes with at least one shown association to hearing (GO:0006915) based on experimental evidence were taken into consideration. This evidence was also required to be traceable to the related publication. The date of the first publication with experiments showing involvement of a gene in hearing was taken as the discovery date. In another words, the discovery date was linked to the first study that resulted in the identification of a gene as an auditory gene. (This discovery date was related specifically to hearing and did not necessarily have to be the date when the gene was discovered in the related genome, although both dates could in some cases coincide.) As result, these publication dates were used to obtain a chronological overview of the auditory gene discoveries. This overview was similar to the one shown for human genes by Resendes *et al*. [Resendes *et al*., 2001]). While being inspired by the figure 4 from their review publication, the novelty of the approach employed in the AGOA project comes from including four species and using GO data together with actual publications for obtaining the overview.

In summary, the AGOA project included improvement of auditory genes lists, revision of relevant evidences and chronological overview of related discoveries. The work was performed separately for each species. While goals related with fruitfly and zebrafish auditory genes are essentially accomplished and presented in this thesis, the work concentrating on human and mouse auditory genes is still in progress. Achievements so far in the AGOA project are results of the collaboration between the author of this thesis, the experts in the related species and the biocurators of the dedicated databases. In contrast to the usual review publications, this study allowed direct storage of results in the GO database that could make them easily accessible online and usable for future bioinformatics analyses. In addition, the results were expected to be easily reproducible since the GO database saves its releases in the archive [Gene Ontology database archive].

4.2. Transcriptome-level investigation

4.2.1 Functional annotation of candidate auditory genes using the AMIGO GO Term Enrichment tool

The AMIGO GO term enrichment tool version 1.7 [Carbon *et al.*, 2009] was applied used for the functional annotation of candidate Drosophila auditory organ genes using the GO database release dated 2010.11.20.

4.3. Interactome-level investigation

4.3.1 Reconstruction of the auditory gene network using Cytoscape

Cytoscape software (version 2.4.1) was used as the network analysis and visualization tool [Shannon *et al.*, 2003]. Interaction data was obtained from the human gene interaction network, preinstalled inside the Cytoscape software. This network was obtained from the BIND (Biomolecular interaction network database) database version 10.10.2006 [Gilbert, 2005]. Human auditory genes were downloaded from the GO database (version 2012.06) and saved as a gene list into a text file. This list was used as an input to map the auditory genes into the BIND gene network. This initial gene network was used to produce a sub-network specific for hearing that constituted the initial auditory gene network.

First and second neighbors of the auditory genes mapped into the network together and their corresponding interactions were used to generate a sub-network of the initial network. This sub-network would be the auditory network and most of the mapped auditory genes were expected to be within this new network. All unrelated genes within the initial network and auditory genes that did not locate to this network (collectively called in this thesis as “outsider genes”) were discarded. The resulting polished auditory gene network was used for further analysis.

4.3.2 Annotation of the auditory gene network using the GOlorize plugin

The Cytoscape GOlorize plugin [Garcia *et al.*, 2007] was used to annotate gene products inside the network using GO annotations. Gene products were colored according to their GO term associations. The table below summarizes all steps in the generation and analysis of auditory gene network.

Table 10: Steps in the gene network-based analysis of human auditome.

Step	**Operation**	**Resulting gene network**
1. Mapping	Auditory genes are mapped into the BIND gene network.	Initial BIND network
2. Pre-processing	Mapped auditory genes plus their first and second neighbors are selected. The remaining genes (“outsider genes”) are included.	Initial auditory gene network with outsider
3. Polishing	Genes that are not connected to the main network (“outsider genes “) are discarded. Layout is applied.	Polished auditory gene network
4. Annotation	Use the GO database to annotate genes within the network.	Annotated auditory gene network

4.4. Developing Gene Ontology-based framework to annotate genes of hearing

4.4.1 Structure of the GO-based framework to annotate genes of hearing

Each of the GO-based methods applied in this thesis so far was developed originally to meet a specific need related with GO annotations. After showing separately application of these methods in hearing research, the idea of combining applying them altogether has emerged. In quest of the optimal way of integrating these methods into a comprehensive annotation procedure of auditory gene products, the "Gene Ontology-framework to annotate genes of hearing" was developed. Two properties of the methods provided the conceptual basis for incorporating each method as a separate step in the framework: The sequential interdependence between the described methods and the constant utilization of the GO database. While the former can be linked to the central dogma of molecular biology, the latter can be attributed to the versatility of the GO database. The structure of the framework and its applicability in hearing research was discussed in this thesis.

4.4.2 Evaluating usability of the GO database in hearing research

Results obtained in this thesis and literature search was used to define pros and contras of using the GO database in gene annotation, with emphasis on hearing.

4.4.3 Approaching challenges and potential of Systems biology of hearing

The future perspectives of the GO-based framework to annotate genes of hearing were approached in the context of systems biology. This included discussing the need for - and potential of - hearing systems biology together with the associated challenges. Thus, this thesis can also be considered as an effort to fill the gap between the areas of hearing research and systems biology by demonstrating benefits of approaching the auditome in a broad interdisciplinary context.

5. Results

5.1. Genome-level investigation

5.1.1 Development of Application for mining Gene Ontology data (AGENDA) and its usage in hearing research

5.1.1.1 Development of AGENDA

Here, a novel web-based tool called "AGENDA" (Application for mining Gene Ontology data) is presented. In addition to Simple queries, based on a single gene product or GO term, AGENDA allows comparison of gene lists related with multiple GO terms. This comparison is based on GO Slimmer or Boolean query and achieved with complex queries that evaluate at once multiple GO terms. In contrast to Simple query, GO Slimmer and Boolean query will be collectively referred to in this thesis as "batch queries". AGENDA generates data-driven charts supporting the results of batch queries. All three query options, represented by distinct pages, are interlinked to each other. Thus, it is possible to elaborately interpret data provided in one query page by importing it as input into another query page. Furthermore, evidences page provides information about the evidences used in GO annotations underlying the results of both simple and batch queries. While all query options are described in the User guide page, additional information about AGENDA is also provided in the Imprint page (Table 11).

Table 11: Pages in AGENDA.

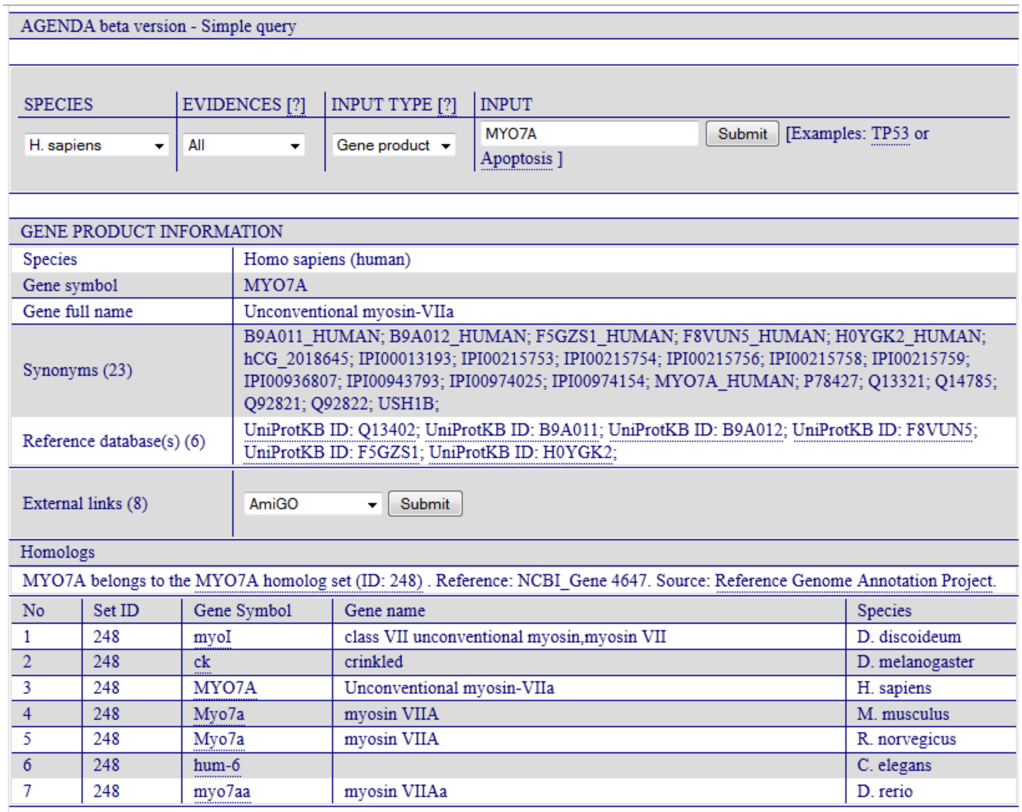
Page	Feature
Simple query	Provides detailed data about one gene product or GO term at a time.
GO Slimmer	Allows generating GO Slims using multiple GO terms as input.
Boolean query	Allows Boolean queries using multiple GO terms as input.
Evidences	Provides evidence(s) for gene product – GO term association(s).
User guide	Contains documentation about query options of AGENDA.
Imprint	Contains contact details, copyright information, references, etc.

Genomes of 12 species are accessible with AGENDA: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila* *melanogaster*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. These genomes are target of the ongoing Gene Ontology's Reference Genome Project. This project's goal is to provide comprehensive GO annotation of these genomes together with the homology information about the involved genes [Reference Genome Group of the Gene Ontology Consortium, 2009]. This homology information can also be accessed with AGENDA by querying a specific gene product.

The interface of AGENDA includes many user-friendly features and was designed to enable convenient navigation along with the intuitive use of query options. A query page in AGENDA consists of an input field and an output field (accompanied in batch queries by a chart). Symbol or full name as well as synonyms (if any) are all valid as input for querying gene products. Likewise, it is possible to query GO terms using their GO term accession numbers, names or synonyms (if any). This is achieved with a query expansion that supports all input types listed above. Stepwise refinement of searches is possible by using results of the initial query as the input for new queries. User can perform even more specific queries by applying GO evidence filters (Table: GO evidence codes). In addition, the drop-down menu for selecting the species enables resubmitting the query to view the corresponding findings in other species. Internal as well as external links for retrieving the target data are provided. It is also possible to bookmark AGENDA pages (together with results of queries) for revisiting and to export obtained gene lists as CSV files.

Due to the biological universality of the GO database in terms of considered species and topics [Ashburner *et al.*, 2000], AGENDA can be used as a generic bioinformatics tool to answer various questions. For example, AGENDA can be used while dealing with heart contraction (GO:0060047) in mouse, fruit development (GO:0010154) in plants or response to drug (GO:0042493) in bacteria. Examples related with hearing (GO:0006915) will be presented here to demonstrate different scenarios of AGENDA usage.

The simple query page in AGENDA allows querying a single gene product or a GO term. For example, it is possible to retrieve information about the human protein Myosin-VIIa (Figure 6 and 7) or about hearing (GO:0006915) (Figure 8). In both cases, the output is also specified by the selected species and the GO evidence code. The output for gene product query includes the protein's description and the list of the associated GO terms. Vice versa, the output for gene product query includes the GO term's description and the list of the associated gene products.



AGENDA beta version - Simple query

SPECIES	EVIDENCES [?]	INPUT TYPE [?]	INPUT
H. sapiens	All	Gene product	MYO7A Submit [Examples: TP53 or Apoptosis]

GENE PRODUCT INFORMATION

Species	Homo sapiens (human)
Gene symbol	MYO7A
Gene full name	Unconventional myosin-VIIa
Synonyms (23)	B9A011_HUMAN; B9A012_HUMAN; F5GZS1_HUMAN; F8VUN5_HUMAN; H0YGK2_HUMAN; hCG_2018645; IPI00013193; IPI00215753; IPI00215754; IPI00215756; IPI00215758; IPI00215759; IPI00936807; IPI00943793; IPI00974025; IPI00974154; MYO7A_HUMAN; P78427; Q13321; Q14785; Q92821; Q92822; USH1B;
Reference database(s) (6)	UniProtKB ID: Q13402; UniProtKB ID: B9A011; UniProtKB ID: B9A012; UniProtKB ID: F8VUN5; UniProtKB ID: F5GZS1; UniProtKB ID: H0YGK2;
External links (8)	AmiGO Submit

Homologs

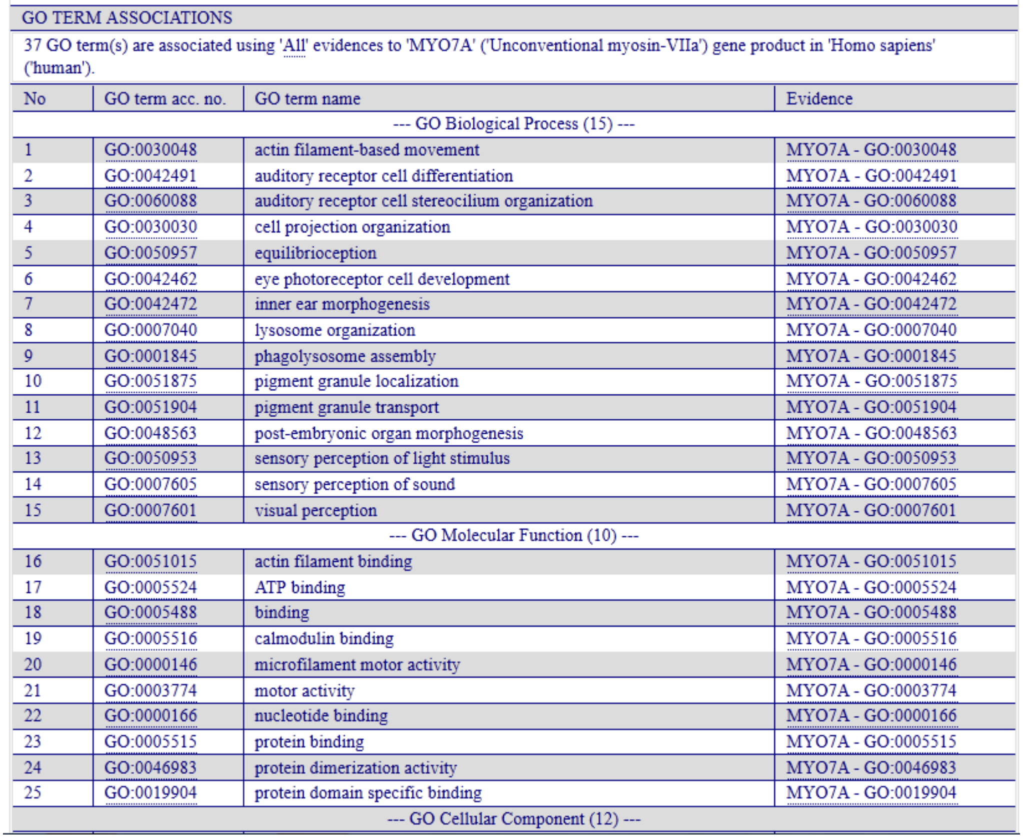
MYO7A belongs to the MYO7A homolog set (ID: 248) . Reference: NCBI_Gene 4647. Source: Reference Genome Annotation Project.

No	Set ID	Gene Symbol	Gene name	Species
1	248	myoI	class VII unconventional myosin,myosin VII	D. discoideum
2	248	ck	crinkled	D. melanogaster
3	248	MYO7A	Unconventional myosin-VIIa	H. sapiens
4	248	Myo7a	myosin VIIA	M. musculus
5	248	Myo7a	myosin VIIA	R. norvegicus
6	248	hum-6		C. elegans
7	248	myo7aa	myosin VIIAa	D. rerio

Figure 6: Simple query page in AGENDA, gene product query, results for the human MYO7A protein, screenshot 1.

Query result for the human MYO7A gene product includes the description of the gene, its homologues and list of associated GO terms. The upper part of the web page displaying the gene description homologues is shown.

The list of auditory gene products obtained by a simple query with the GO term "hearing" can be exported as a CSV file or used directly as the input for batch queries by clicking the related links. Consequently, the "GO term 1" input field in the batch query page corresponds to the GO term (in this case “hearing”) associated with the queried gene list. Batch queries facilitate a further annotation of this dataset using other user-defined GO terms. Thus, it is for example possible to delineate auditory-relevant genes that are associated with a second GO term such as the plasma membrane (GO:0005886). Using “hearing” as input for the “GO term 1” in batch queries allows mining of the GO data associated with the human auditory gene products. Batch queries are performed in the GO Slimmer page and the Boolean query page of AGENDA.



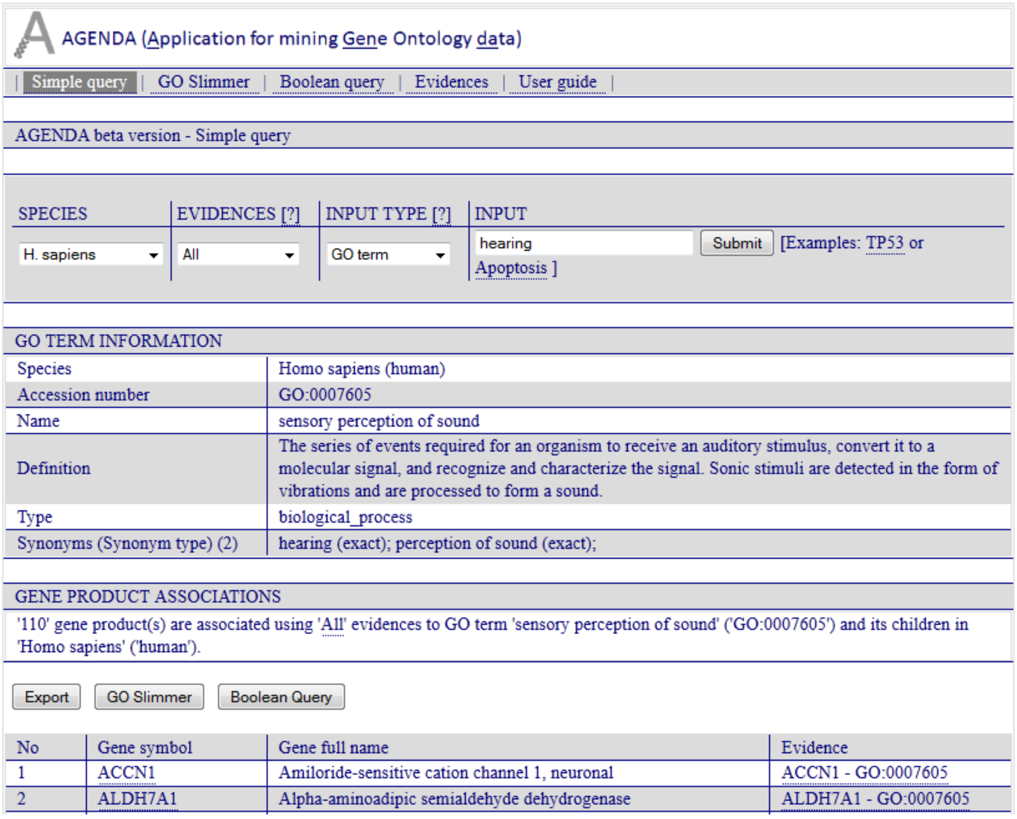
GO TERM ASSOCIATIONS

37 GO term(s) are associated using 'All' evidences to 'MYO7A' ('Unconventional myosin-VIIa') gene product in 'Homo sapiens' ('human').

No	GO term acc. no.	GO term name	Evidence
--- GO Biological Process (15) ---			
1	GO:0030048	actin filament-based movement	MYO7A - GO:0030048
2	GO:0042491	auditory receptor cell differentiation	MYO7A - GO:0042491
3	GO:0060088	auditory receptor cell stereocilium organization	MYO7A - GO:0060088
4	GO:0030030	cell projection organization	MYO7A - GO:0030030
5	GO:0050957	equilibrioception	MYO7A - GO:0050957
6	GO:0042462	eye photoreceptor cell development	MYO7A - GO:0042462
7	GO:0042472	inner ear morphogenesis	MYO7A - GO:0042472
8	GO:0007040	lysosome organization	MYO7A - GO:0007040
9	GO:0001845	phagolysosome assembly	MYO7A - GO:0001845
10	GO:0051875	pigment granule localization	MYO7A - GO:0051875
11	GO:0051904	pigment granule transport	MYO7A - GO:0051904
12	GO:0048563	post-embryonic organ morphogenesis	MYO7A - GO:0048563
13	GO:0050953	sensory perception of light stimulus	MYO7A - GO:0050953
14	GO:0007605	sensory perception of sound	MYO7A - GO:0007605
15	GO:0007601	visual perception	MYO7A - GO:0007601
--- GO Molecular Function (10) ---			
16	GO:0051015	actin filament binding	MYO7A - GO:0051015
17	GO:0005524	ATP binding	MYO7A - GO:0005524
18	GO:0005488	binding	MYO7A - GO:0005488
19	GO:0005516	calmodulin binding	MYO7A - GO:0005516
20	GO:0000146	microfilament motor activity	MYO7A - GO:0000146
21	GO:0003774	motor activity	MYO7A - GO:0003774
22	GO:0000166	nucleotide binding	MYO7A - GO:0000166
23	GO:0005515	protein binding	MYO7A - GO:0005515
24	GO:0046983	protein dimerization activity	MYO7A - GO:0046983
25	GO:0019904	protein domain specific binding	MYO7A - GO:0019904
--- GO Cellular Component (12) ---			

Figure 7: Simple query page in AGENDA, gene product query, results for the human MYO7A protein, screenshot 2.

Query result for the human MYO7A gene product includes the description of the gene, its homologues and list of associated GO terms. The middle part of the web page displaying GO term annotations under GO Biological Process and Molecular Function categories is shown.



AGENDA (Application for mining Gene Ontology data)

| Simple query | GO Slimmer | Boolean query | Evidences | User guide |

AGENDA beta version - Simple query

SPECIES	EVIDENCES [?]	INPUT TYPE [?]	INPUT
H. sapiens	All	GO term	hearing Submit [Examples: TP53 or Apoptosis]

GO TERM INFORMATION	
Species	Homo sapiens (human)
Accession number	GO:0007605
Name	sensory perception of sound
Definition	The series of events required for an organism to receive an auditory stimulus, convert it to a molecular signal, and recognize and characterize the signal. Sonic stimuli are detected in the form of vibrations and are processed to form a sound.
Type	biological_process
Synonyms (Synonym type) (2)	hearing (exact); perception of sound (exact);

GENE PRODUCT ASSOCIATIONS

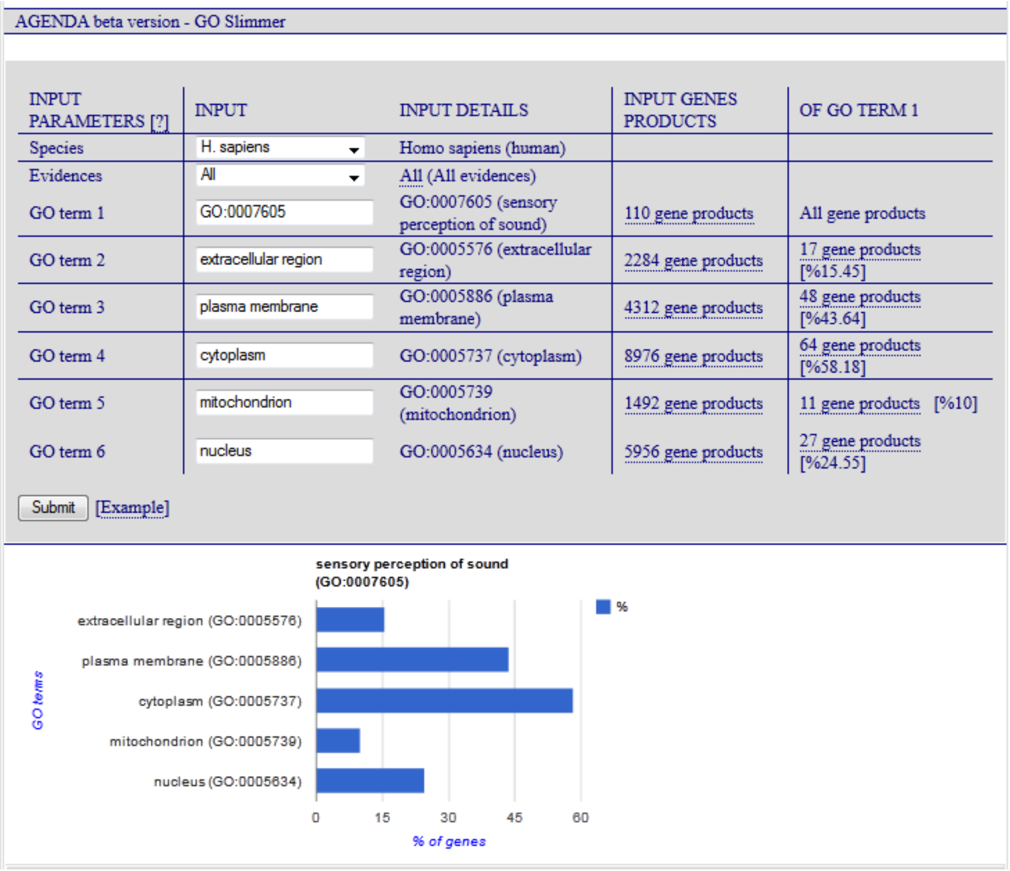
'110' gene product(s) are associated using 'All' evidences to GO term 'sensory perception of sound' ('GO:0007605') and its children in 'Homo sapiens' ('human').

No	Gene symbol	Gene full name	Evidence
1	ACCN1	Amiloride-sensitive cation channel 1, neuronal	ACCN1 - GO:0007605
2	ALDH7A1	Alpha-aminoadipic semialdehyde dehydrogenase	ALDH7A1 - GO:0007605

Figure 8: Simple query page in AGENDA, GO term query, results for human hearing.

Query result for the GO term "hearing" in human includes the description of the GO term, further options, and the list of associated gene products. The upper part of the web page is shown. It displays the GO term description, further query options and the first 2 gene products from the total list of 110 human auditory gene products. While the "Export" button downloads this gene list as CSV file, the "GO Slimmer" and "Boolean Query" buttons allow using this list as input for batch queries.

With GO Slimmer, it is also possible to get an idea about the distribution of human auditory gene products in various cellular components. This assay will be referred in this thesis as "cellular components analysis". The user can customize this data mining analysis by specifying the GO term of interest in this batch query. In addition to the GO term 1 (in this case "hearing"), further GO terms belonging to the GO Cellular Component category (for example, the GO term "plasma membrane") can be used as the input (Figure 9). GO Slimmer calculates how many auditory genes (GO:0007605) are associated with each functional category. The resulting GO Slim includes numerical data in the "OF GO TERM 1" column and a chart summarizing the findings. Numerical results consist of the numbers and percentages of auditory gene products annotated to each cellular component. In order to obtain the names of the actual gene products, the user can click on the corresponding number in the "OF GO TERM 1" column. This will open the Boolean query page that allows using related gene products for follow-up Boolean queries (for example, see figure 10).



AGENDA beta version - GO Slimmer

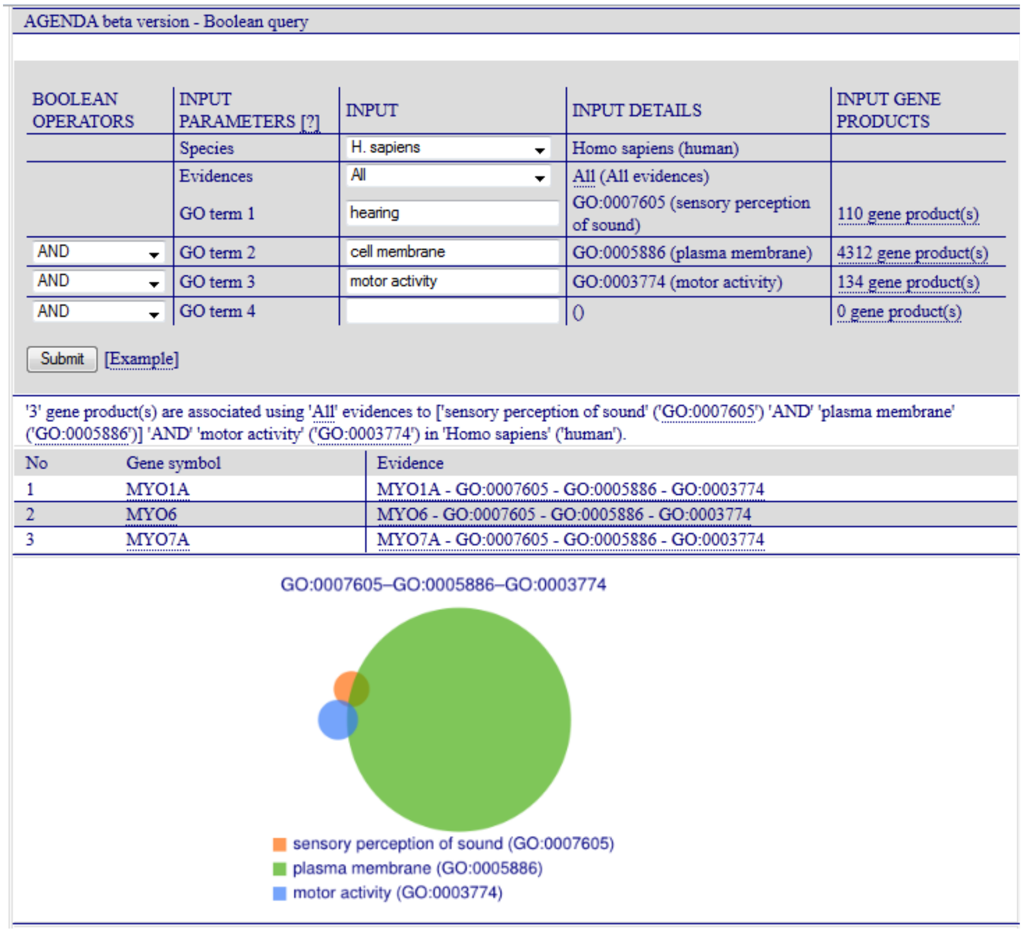
INPUT PARAMETERS [?]	INPUT	INPUT DETAILS	INPUT GENES PRODUCTS	OF GO TERM 1
Species	H. sapiens	Homo sapiens (human)		
Evidences	All	All (All evidences)		
GO term 1	GO:0007605	GO:0007605 (sensory perception of sound)	110 gene products	All gene products
GO term 2	extracellular region	GO:0005576 (extracellular region)	2284 gene products	17 gene products [%15.45]
GO term 3	plasma membrane	GO:0005886 (plasma membrane)	4312 gene products	48 gene products [%43.64]
GO term 4	cytoplasm	GO:0005737 (cytoplasm)	8976 gene products	64 gene products [%58.18]
GO term 5	mitochondrion	GO:0005739 (mitochondrion)	1492 gene products	11 gene products [%10]
GO term 6	nucleus	GO:0005634 (nucleus)	5956 gene products	27 gene products [%24.55]

Submit
[Example]
sensory perception of sound (GO:0007605)
%
extracellular region (GO:0005576)
plasma membrane (GO:0005886)
cytoplasm (GO:0005737)
mitochondrion (GO:0005739)
nucleus (GO:0005634)
GO terms
0
15
30
45
60
% of genes

Figure 9: GO Slimmer query page, cellular components analysis of human auditory gene products.

Human gene products that are implicated in hearing and their relation with a number of cellular components are shown. 110 human gene products are identified that are annotated to hearing (GO:0007605). 11 of the respective gene products are associated, for example, with the mitochondrion (GO:0005739), 27 with the nucleus (GO:0005634), and 48 with the plasma membrane (GO:0005886).

The Boolean query page of AGENDA allows retrieving human gene products that are associated with both hearing (GO:0007605) and the plasma membrane (GO:0005886). This is achieved with the Boolean operator "AND". It is also possible to include further GO terms and Boolean operators in the query. For example, the human gene products that are implicated in hearing (GO:0007605), associated with the plasma membrane (GO:0005886), and annotated with motor activity (GO:0003774) can be obtained. In this case, the Boolean query is performed by using three GO terms and two “AND” Boolean operators. (Figure 10). Another example for a Boolean query is a search for gene products annotated to hearing (GO:0007605), vision (GO:0007601) and the cilium (GO:0005929) using two “AND” Boolean operators (Figure 11).



AGENDA beta version - Boolean query

BOOLEAN OPERATORS	INPUT PARAMETERS [?]	INPUT	INPUT DETAILS	INPUT GENE PRODUCTS
	Species	H. sapiens	Homo sapiens (human)	
	Evidences	All	All (All evidences)	
	GO term 1	hearing	GO:0007605 (sensory perception of sound)	110 gene product(s)
AND	GO term 2	cell membrane	GO:0005886 (plasma membrane)	4312 gene product(s)
AND	GO term 3	motor activity	GO:0003774 (motor activity)	134 gene product(s)
AND	GO term 4		0	0 gene product(s)

Submit [Example]

'3' gene product(s) are associated using 'All' evidences to ['sensory perception of sound' ('GO:0007605') 'AND' 'plasma membrane' ('GO:0005886')] 'AND' 'motor activity' ('GO:0003774') in 'Homo sapiens' ('human').

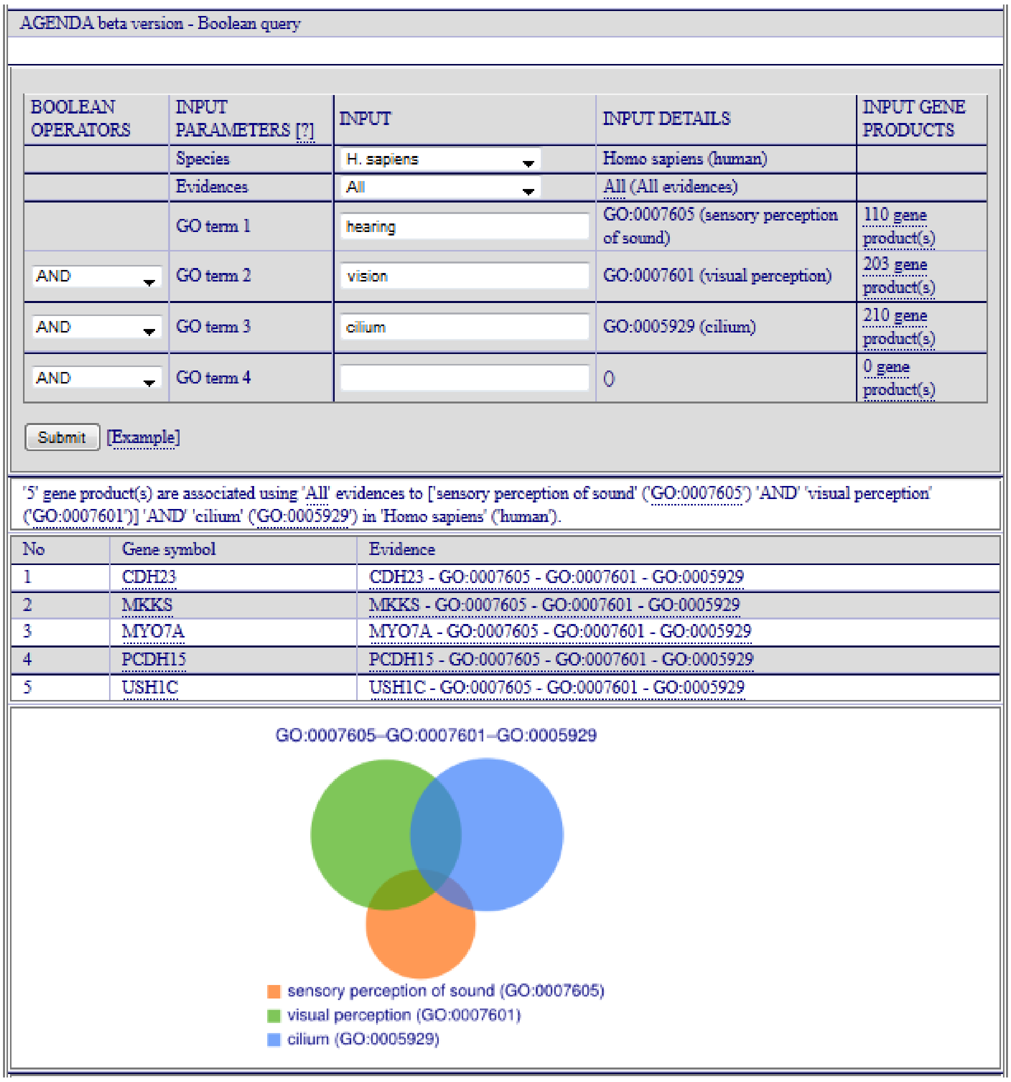
No	Gene symbol	Evidence
1	MYO1A	MYO1A - GO:0007605 - GO:0005886 - GO:0003774
2	MYO6	MYO6 - GO:0007605 - GO:0005886 - GO:0003774
3	MYO7A	MYO7A - GO:0007605 - GO:0005886 - GO:0003774

GO:0007605–GO:0005886–GO:0003774
🟧 sensory perception of sound (GO:0007605)
🟩 plasma membrane (GO:0005886)
🟦 motor activity (GO:0003774)

Figure 10: Boolean query page in AGENDA, human gene products annotated simultaneously to hearing, the plasma membrane and the motor activity.

Result of a Boolean query delineating human gene products annotated to hearing (GO:0007605), the plasma membrane (GO:0005886) and the motor activity (GO:0003774). The three GO terms are combined using two Boolean operators “AND”. 3 gene products are identified that are associated simultaneously with all three GO terms. A Venn diagram illustrates relationship between the GO terms based on their common gene products.

The result of a typical Boolean query includes a list of retrieved gene products and a Venn diagram. Clicking on a gene product’s name in this list will direct user’s web browser to the Simple query page dedicated to this gene product. For each gene product there is a link to the evidences page that provides information supporting association of the gene product with the GO terms at issue. In addition to a list of gene products, it is possible to generate a Venn diagram depicting the relationship between two or three GO terms based on their common gene products. Each GO term is represented by a circle with a different color. Size of circles and their overlapping areas are proportional to the number of corresponding gene products.



AGENDA beta version - Boolean query

BOOLEAN OPERATORS	INPUT PARAMETERS [?]	INPUT	INPUT DETAILS	INPUT GENE PRODUCTS
	Species	H. sapiens	Homo sapiens (human)	
	Evidences	All	All (All evidences)	
	GO term 1	hearing	GO:0007605 (sensory perception of sound)	110 gene product(s)
AND	GO term 2	vision	GO:0007601 (visual perception)	203 gene product(s)
AND	GO term 3	cilium	GO:0005929 (cilium)	210 gene product(s)
AND	GO term 4		0	0 gene product(s)

'5' gene product(s) are associated using 'All' evidences to ['sensory perception of sound' ('GO:0007605') 'AND' 'visual perception' ('GO:0007601')] 'AND' 'cilium' ('GO:0005929') in 'Homo sapiens' ('human').

No	Gene symbol	Evidence
1	CDH23	CDH23 - GO:0007605 - GO:0007601 - GO:0005929
2	MKKS	MKKS - GO:0007605 - GO:0007601 - GO:0005929
3	MYO7A	MYO7A - GO:0007605 - GO:0007601 - GO:0005929
4	PCDH15	PCDH15 - GO:0007605 - GO:0007601 - GO:0005929
5	USH1C	USH1C - GO:0007605 - GO:0007601 - GO:0005929

GO:0007605–GO:0007601–GO:0005929
🟧 sensory perception of sound (GO:0007605)
🟩 visual perception (GO:0007601)
🟦 cilium (GO:0005929)

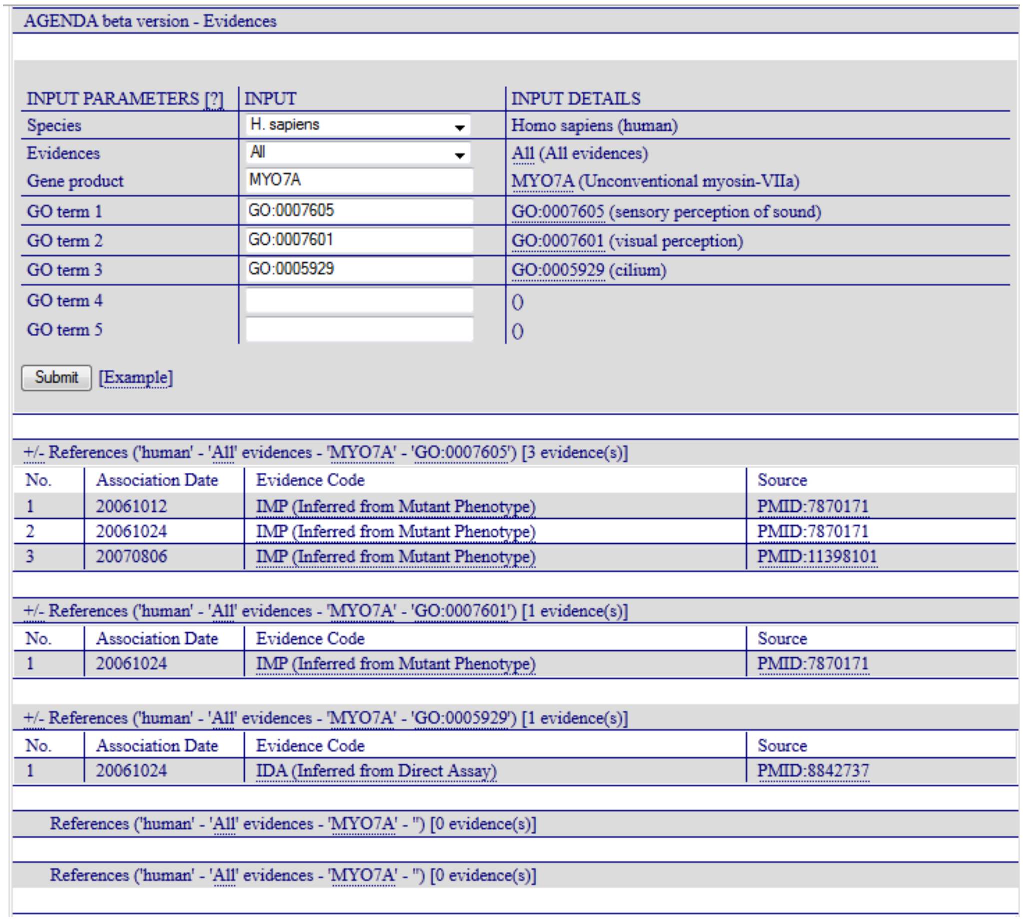
Figure 11: Boolean query page in AGENDA, human gene products annotated simultaneously to hearing, vision and the cilium.

Result of a Boolean query delineating human gene products annotated to hearing (GO:0007605), vision (GO:0007601) and the cilium (GO:0005929). The three GO terms are combined using two Boolean operators “AND”. 5 gene products are identified that are associated simultaneously with all three GO terms. Venn diagram illustrates relationship between the GO terms based on their common gene products.

The evidences page of AGENDA helps to pursue evidence underlying distinct GO annotations. This page can be also accessed directly from the query pages described above to obtain information about evidences substantiating each of the

"

used GO annotations. The evidences page assesses association of a gene product with up to five different GO terms and displays separately data for each association. If there is more than one evidence record, all of them are shown and ordered according to the association date. Information linking a gene product to a certain GO term includes the association date, the evidence code and the name of evidence source. Additional evidence can be obtained by clicking on the links in the Source column (Figure 12).



AGENDA beta version - Evidences

INPUT PARAMETERS [?]	INPUT	INPUT DETAILS
Species	H. sapiens	Homo sapiens (human)
Evidences	All	All (All evidences)
Gene product	MYO7A	MYO7A (Unconventional myosin-VIIa)
GO term 1	GO:0007605	GO:0007605 (sensory perception of sound)
GO term 2	GO:0007601	GO:0007601 (visual perception)
GO term 3	GO:0005929	GO:0005929 (cilium)
GO term 4		()
GO term 5		()

+/- References ('human' - 'All' evidences - 'MYO7A' - 'GO:0007605') [3 evidence(s)]

No.	Association Date	Evidence Code	Source
1	20061012	IMP (Inferred from Mutant Phenotype)	PMID:7870171
2	20061024	IMP (Inferred from Mutant Phenotype)	PMID:7870171
3	20070806	IMP (Inferred from Mutant Phenotype)	PMID:11398101

+/- References ('human' - 'All' evidences - 'MYO7A' - 'GO:0007601') [1 evidence(s)]

No.	Association Date	Evidence Code	Source
1	20061024	IMP (Inferred from Mutant Phenotype)	PMID:7870171

+/- References ('human' - 'All' evidences - 'MYO7A' - 'GO:0005929') [1 evidence(s)]

No.	Association Date	Evidence Code	Source
1	20061024	IDA (Inferred from Direct Assay)	PMID:8842737

References ('human' - 'All' evidences - 'MYO7A' - '') [0 evidence(s)]

References ('human' - 'All' evidences - 'MYO7A' - '') [0 evidence(s)]

Figure 12: Evidences page of AGENDA, evidences linking the human MYO7A protein with hearing, vision and the cilium.

Result of querying the evidences page of AGENDA for information supporting the association of the human MYO7A protein with hearing (GO:0007605), vision (GO:0007601) and the cilium (GO:0005929). The results show that MYO7A is associated with all of these GO terms. To support this finding, the web page also provides evidences for each gene product – GO term association. Here, all of the evidences are based on the PubMed publications as their source.

Results obtained from GO Slimmer can be used for an interspecies comparison. It was already demonstrated above how AGENDA’s GO Slimmer page can be used for the species-specific cellular components analysis of auditory gene products (Figure 9). In addition, it is possible to manually compare the results of different GO Slimmer queries in different species (This feature is not supported in AGENDA). For example, it is possible to compare the numbers of the human and the mouse auditory gene products that are annotated to a specific cellular component. For this purpose, 5 demonstrative GO terms (same as in the figure 9) were selected. Furthermore, results of the cellular components analyses of the whole human and mice genomes were also included as references (Figure 13). The comparison between humans and mice showed that the percentages of gene products associated with the plasma membrane were higher in the auditome (41.74% and 39.84%, respectively) when compared to the genome (25.52% and 14.81%, respectively). This suggests that there could be an enrichment of the plasma membrane proteins among auditory gene products and that this pattern could be conserved in both human and mice auditomes. As a further step, the GO term enrichment test needs to be applied to evaluate the significance of these findings.

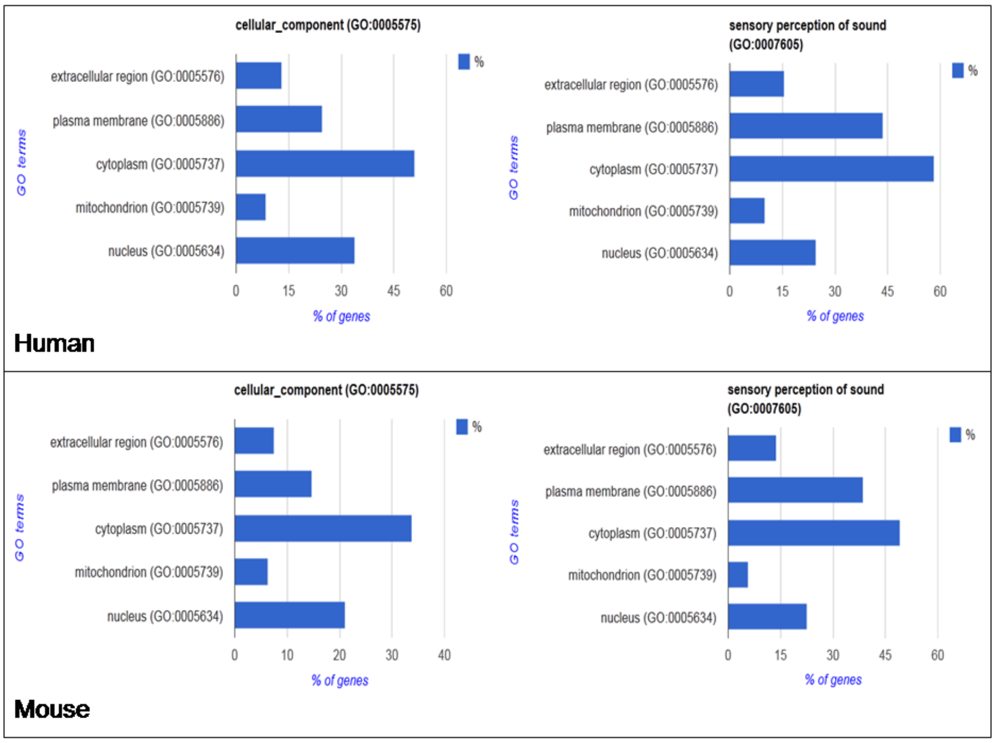
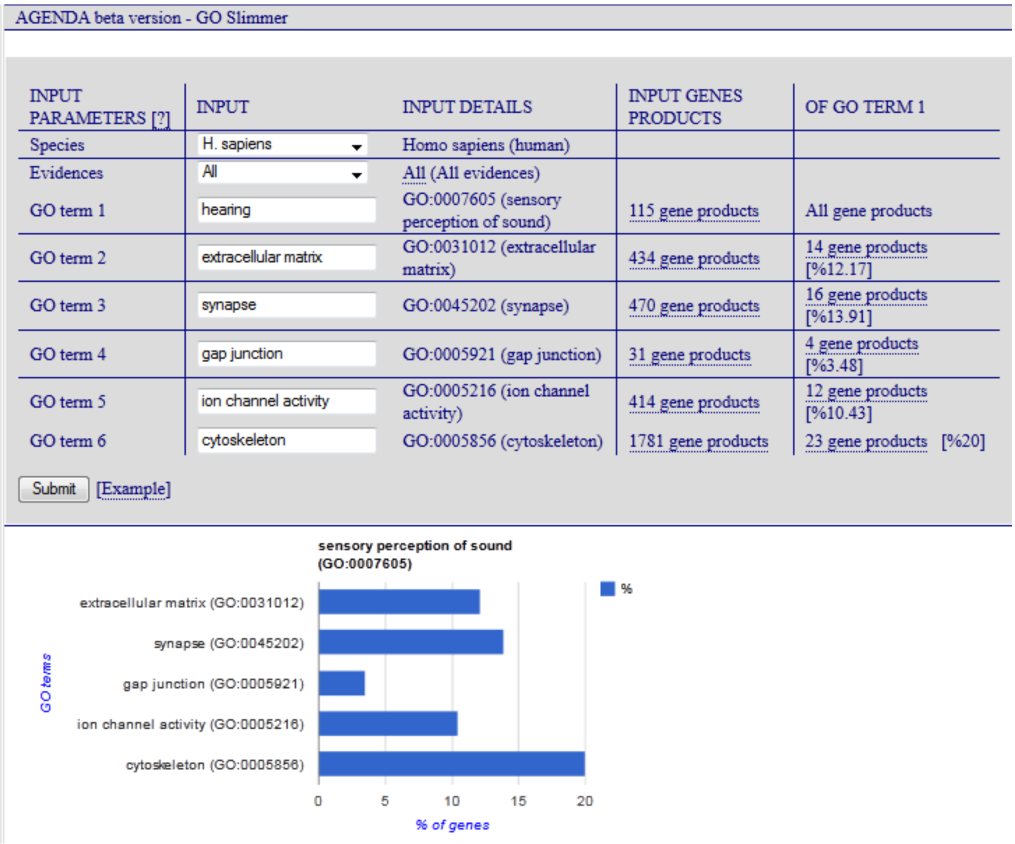


Figure 13: Comparing results of the cellular components analyses of human and mouse auditory gene products.

Comparisons of GO Slims representating results of the cellular components analyses in the genomes (left) and auditomes (right) of humans (top) and mice (bottom). While the GO term “sensory perception of sound” stands for all auditory gene products, the GO term “cellular component” is used as the reference set representing the whole genome. Percentage of gene products associated with the plasma membrane is higher in the auditome when compared to the genome in both in humans and mice.

5.1.1.2 Functional categorization of auditory genes with AGENDA

GO Slimmer was successfully applied to partially reconstruct the table describing functional categories related with human auditory gene products (Table 5) (Figure 14). It is also possible to manually compare (not with AGENDA) the results of the functional categorization of auditory gene products between human and mice. Thus, GO Slims can be useful in the comparative genomics analysis of the different auditomes.



AGENDA beta version - GO Slimmer

INPUT PARAMETERS [?]	INPUT	INPUT DETAILS	INPUT GENES PRODUCTS	OF GO TERM 1
Species	H. sapiens	Homo sapiens (human)		
Evidences	All	All (All evidences)		
GO term 1	hearing	GO:0007605 (sensory perception of sound)	115 gene products	All gene products
GO term 2	extracellular matrix	GO:0031012 (extracellular matrix)	434 gene products	14 gene products [%12.17]
GO term 3	synapse	GO:0045202 (synapse)	470 gene products	16 gene products [%13.91]
GO term 4	gap junction	GO:0005921 (gap junction)	31 gene products	4 gene products [%3.48]
GO term 5	ion channel activity	GO:0005216 (ion channel activity)	414 gene products	12 gene products [%10.43]
GO term 6	cytoskeleton	GO:0005856 (cytoskeleton)	1781 gene products	23 gene products [%20]

Submit
[Example]
sensory perception of sound (GO:0007605)
%
GO terms
extracellular matrix (GO:0031012)
synapse (GO:0045202)
gap junction (GO:0005921)
ion channel activity (GO:0005216)
cytoskeleton (GO:0005856)
0
5
10
15
20
% of genes

Figure 14: GO Slimmer query page, functional classification of human auditory gene products.

Human gene products that are implicated in hearing and their relation with a number of GO terms (each representing a specific functional category) are shown. 110 gene products are identified that are annotated to hearing (GO:0007605). 4 of the respective gene products are associated, for example, with the gap junction (GO:0005921), 12 with ion channel activity (GO:0005216), and 23 with the cytoskeleton (GO:0005856).

5.1.2 Manual gene annotation with the Auditory Gene Ontology Annotation (AGOA) project

5.1.2.1 Improving lists of genes annotated to hearing in the GO database

The Auditory Gene Ontology Annotation (AGOA) project is the first comprehensive annotation effort to catalogue genes human, mouse, zebrafish and fruitfly genes involved in hearing. While the improving of the annotations related with fruit fly and zebrafish auditory genes are finished (this work was achieved together with the collaborators in the AGOA project.), the mammalian part (mouse and human) of work is still in progress.

5.1.2.2 Revision of the evidences for genes annotated to hearing in the GO database

While the revision of evidence for fruit fly and zebrafish auditory genes is finished (this work was accomplished together with the collaborators in the AGOA project), the work related with the mouse and human is still ongoing. Since FlyBase already updated its database using the improved auditory gene list and the revised evidences, the results for fruit fly were made available online in the GO database (See Appendix, table 15, 16). Although annotation of the zebrafish auditory genes and revision of the related evidences are finished, the update of the ZFIN using the data obtained from the AGOA project did not occur yet. The status of the AGOA project is summarized in the table below.

Table 12: Progress in the AGOA project.

Species	Annotation	Before	Deleted	Added	Now
Zebrafish	Genes	4	0	8	12
Zebrafish	Publications	5	1	16	20
Fruit fly	Genes	20	1	11	30
Fruit fly	Publications	30	5	23	48
Mouse	Genes	113	In progress	In progress	In progress
Mouse	Publications	In progress	In progress	In progress	In progress
Human	Genes	94	In progress	In progress	In progress
Human	Publications	In progress	In progress	In progress	In progress

Source: GO database as of September 2012.

5.1.2.3 Chronological overview of the auditory gene discoveries

The approach employed in the AGOA project also resulted in a chronological overview of the auditory gene discoveries in the past decade (Figure 15). Results of the forward screens published in 2000 and 1998 accounted for the discoveries of many auditory genes in the fruit fly and in zebrafish, respectively (Figure 15 A, C). Once the AGOA project is finishing the ongoing AGOA project is expected to complete this figure by including human and mouse data.

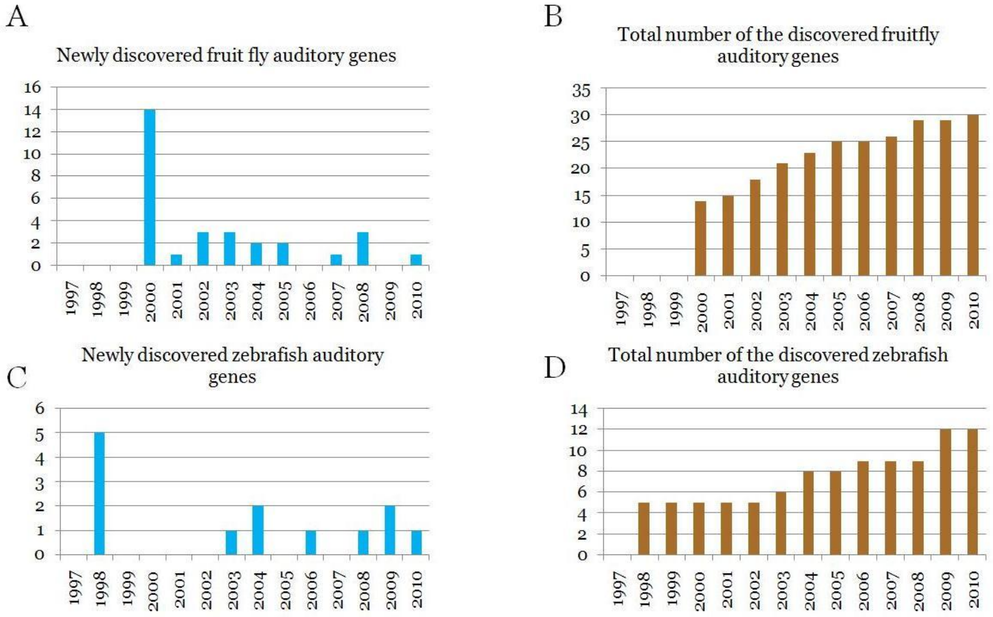


Figure 15: Chronological overview of the auditory gene discoveries in fruit fly and zebrafish.

This figure was produced as result of the AGOA project and summarizes the progress in identification of auditory genes in fruit fly and zebrafish.

In summary, the AGOA project included improvement of auditory genes lists, revision of relevant evidences and chronological overview of related discoveries. The study was conducted separately for each species. While the tasks related with fruitfly and zebrafish auditory genes were essentially finished and described in this thesis, the parts including human and mouse auditory genes are still in progress. Experts in the related species and biocurators of thed dedicated atabases contributed a lot to the achievements of the AGOA.

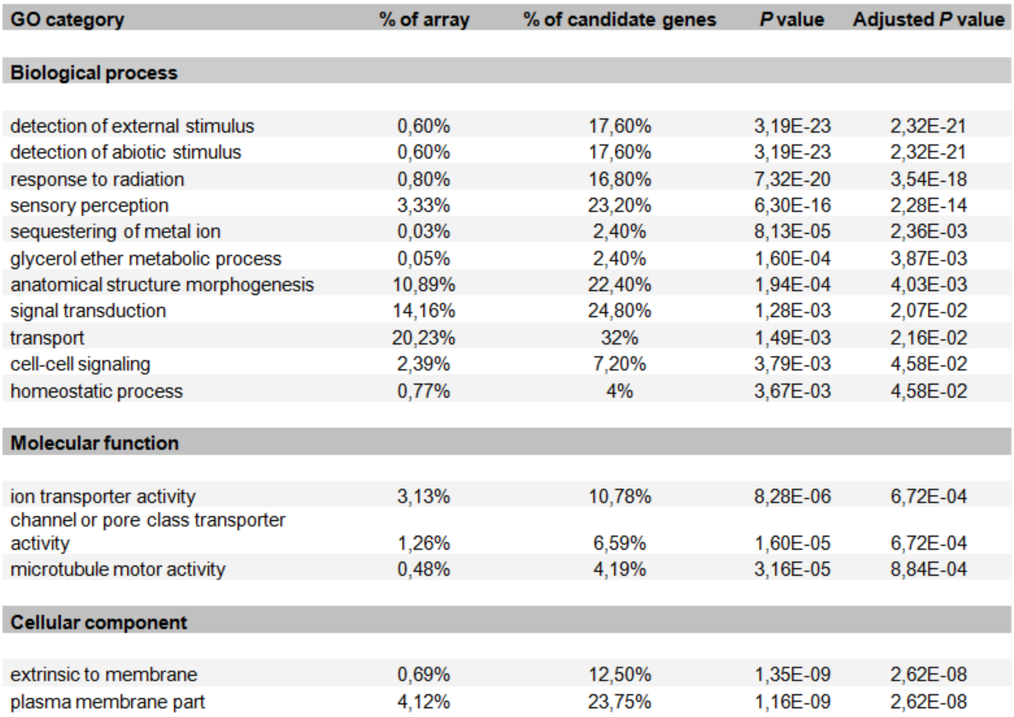
5.2. Transcriptome-level investigation

5.2.1 Functional annotation of candidate auditory genes using the AMIGO GO Term Enrichment tool

Functional annotation using the AMIGO GO term enrichment tool version revealed significantly enriched GO terms in the analyzed list of candidate

Drosophila auditory organ genes (Table 13). These GO terms were significantly overrepresented among the candidate JO genes when compared to the whole genome.

Table 13: Significantly enriched GO terms in the list of candidate *Drosophila* auditory organ genes identified by the microarray analysis.



GO category	% of array	% of candidate genes	*P* value	Adjusted *P* value
Biological process				
detection of external stimulus	0,60%	17,60%	3,19E-23	2,32E-21
detection of abiotic stimulus	0,60%	17,60%	3,19E-23	2,32E-21
response to radiation	0,80%	16,80%	7,32E-20	3,54E-18
sensory perception	3,33%	23,20%	6,30E-16	2,28E-14
sequestering of metal ion	0,03%	2,40%	8,13E-05	2,36E-03
glycerol ether metabolic process	0,05%	2,40%	1,60E-04	3,87E-03
anatomical structure morphogenesis	10,89%	22,40%	1,94E-04	4,03E-03
signal transduction	14,16%	24,80%	1,28E-03	2,07E-02
transport	20,23%	32%	1,49E-03	2,16E-02
cell-cell signaling	2,39%	7,20%	3,79E-03	4,58E-02
homeostatic process	0,77%	4%	3,67E-03	4,58E-02
Molecular function				
ion transporter activity	3,13%	10,78%	8,28E-06	6,72E-04
channel or pore class transporter activity	1,26%	6,59%	1,60E-05	6,72E-04
microtubule motor activity	0,48%	4,19%	3,16E-05	8,84E-04
Cellular component				
extrinsic to membrane	0,69%	12,50%	1,35E-09	2,62E-08
plasma membrane part	4,12%	23,75%	1,16E-09	2,62E-08

Table shows percentages of the array genes (genome) and the candidate genes annotated to a certain GO term is shown. In addition, statistics of the GO term enrichment analysis (P values and adjusted P values) is shown.

5.3. Interactome-level investigation

5.3.1 Reconstruction of the auditory gene network using Cytoscape

Reconstruction of the auditory gene network generated an auditory gene network that included both auditory genes and their neighbors (Figure 16). The neighboring genes were regarded as candidate auditory gene that could be used for further investigation.

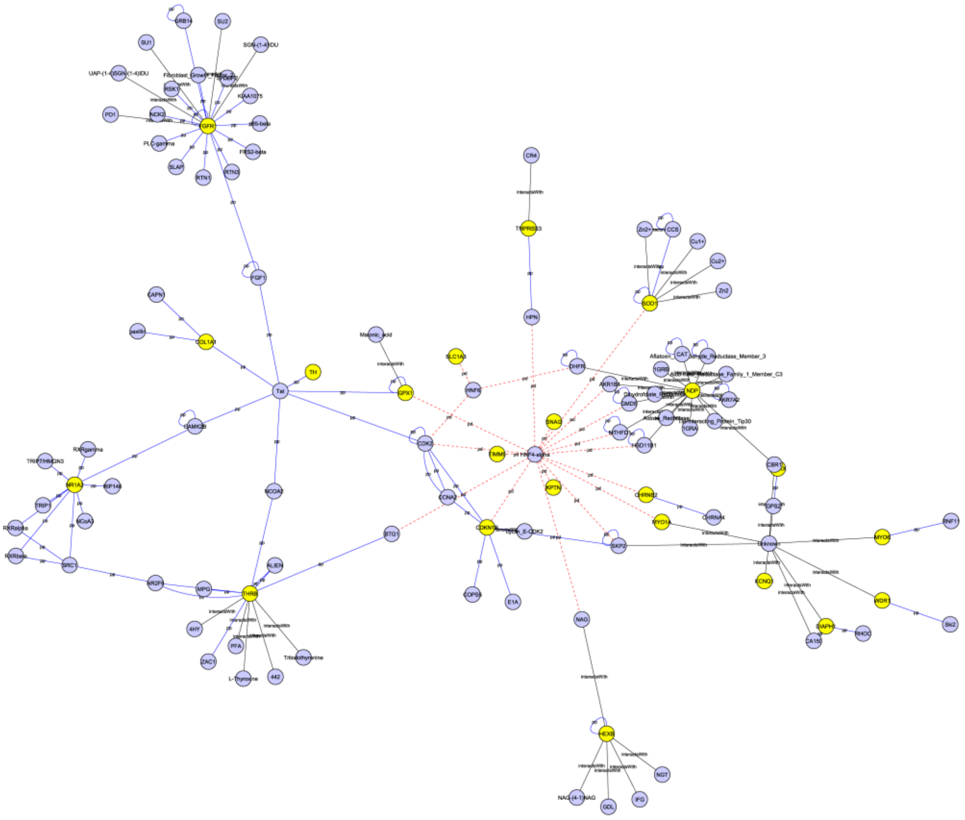
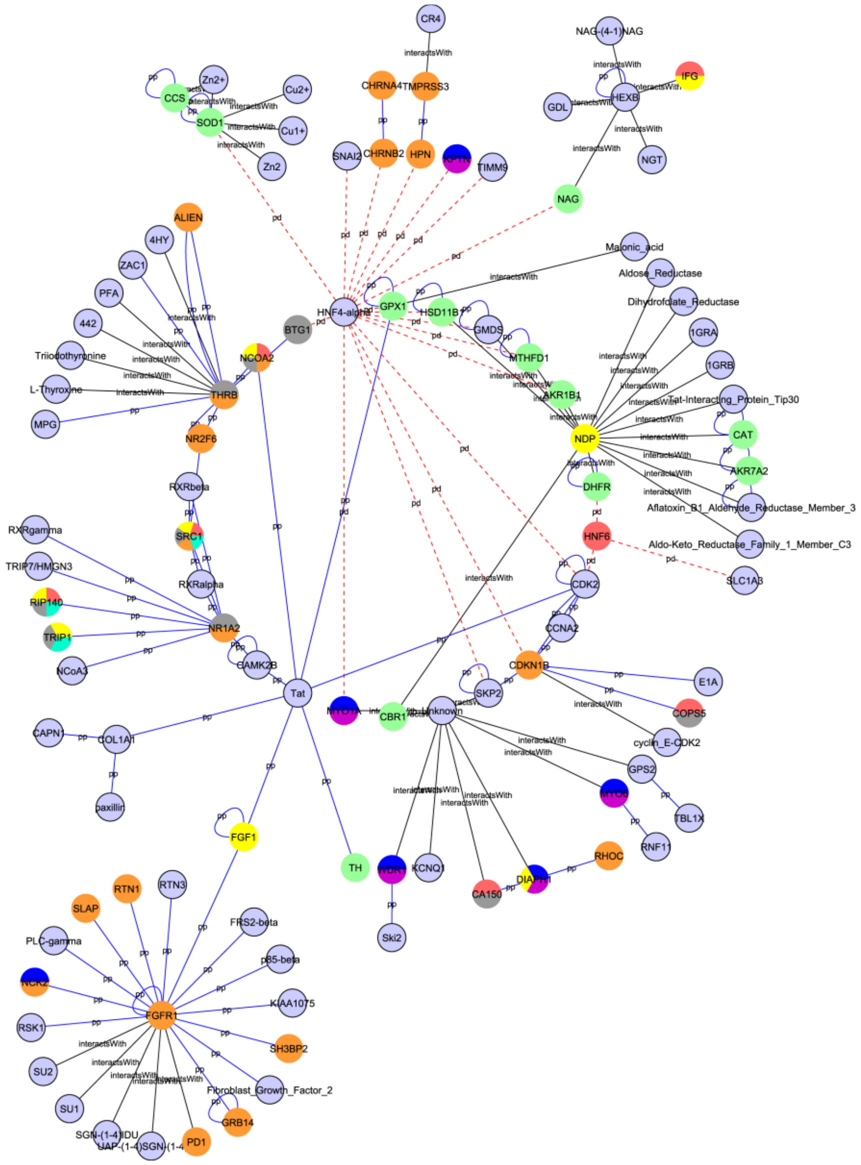


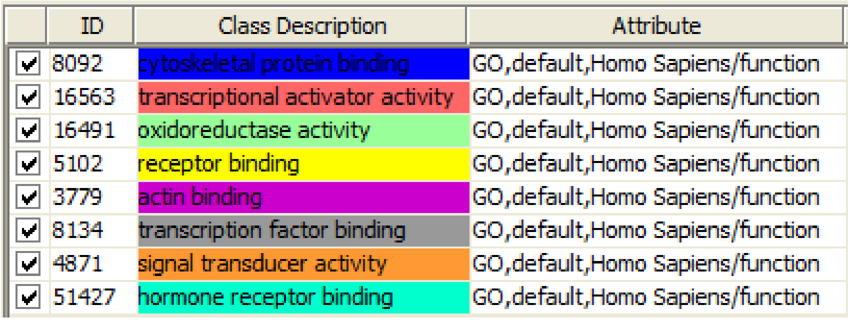
Figure 16: Auditory Gene Network.

Yellow genes represent auditory genes. Blue genes are non-auditory genes that represent first and second neighbors of the auditory genes.

5.3.2 Annotation of the auditory gene network using the GOlorize plugin

GO annotations were applied to gain more information about the auditory gene network. Some of the GO terms significantly enriched were highlighted. This annotation provided a better picture about the specific roles of the network components (Figure 17).





	ID	Class Description	Attribute
☑	8092	cytoskeletal protein binding	GO,default,Homo Sapiens/function
☑	16563	transcriptional activator activity	GO,default,Homo Sapiens/function
☑	16491	oxidoreductase activity	GO,default,Homo Sapiens/function
☑	5102	receptor binding	GO,default,Homo Sapiens/function
☑	3779	actin binding	GO,default,Homo Sapiens/function
☑	8134	transcription factor binding	GO,default,Homo Sapiens/function
☑	4871	signal transducer activity	GO,default,Homo Sapiens/function
☑	51427	hormone receptor binding	GO,default,Homo Sapiens/function

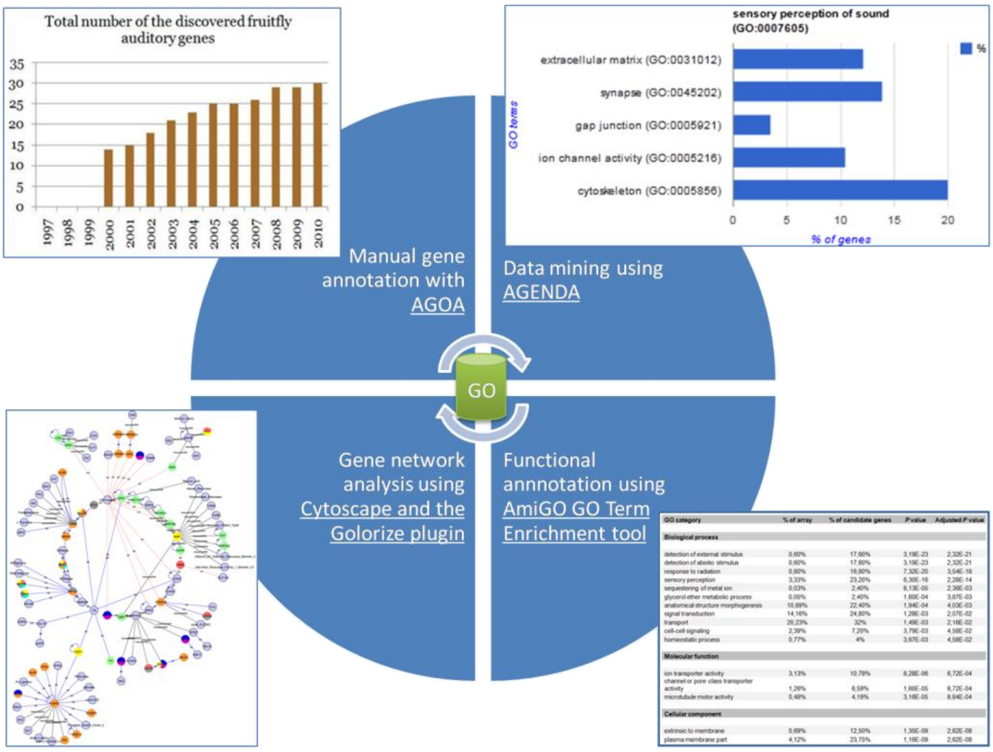
Figure 17: Auditory Gene Network with GO annotations.

GO term enrichment analysis was applied to the list of genes within the auditory network. Only GO terms related with the Molecular Function were used in the enrichment analysis. 8 representative GO terms were chosen manually among significantly enriched GO terms. Genes are colored according to their associations with these GO terms. Circular layout was applied.

5.4. Developing Gene Ontology-based framework to annotate genes of hearing

5.4.1 Structure of the GO-based framework to annotate genes of hearing

The usage of the GO database is the common aspect of methods applied in this thesis. Another important point is that the steps in the analysis of omics data were parallel to the central dogma of molecular biology. The AGOA project is expected to add novel auditory genes obtained from these analyses into the GO database. Improved content of the GO database, in turn, can be used as the basis for new hypotheses and as a better data source for future analysis of high-throughput data. This functional link between different types of studies makes it possible to combine them into a logical sequence. Therefore, the rationale in this thesis was to relate all this investigations to each other within a single framework that was named the “GO-based framework to annotate genes of hearing” (Figure 18). With such a framework it is also possible to complete the research loop consisting of the experimental and computational studies that starts from the experiment and spans through multiple steps to the new hypothesis. It is widely acknowledged that the GO database is especially useful when applied to studies involving high-throughput screens. Because of this, the GO-based framework to annotate genes of hearing fits ideally especially into the context of microarray screens.


Total number of the discovered fruitfly auditory genes
35
30
25
20
15
10
5
0
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
sensory perception of sound (GO:0007605)
extracellular matrix (GO:0031012)
synapse (GO:0045202)
gap junction (GO:0005921)
ion channel activity (GO:0005216)
cytoskeleton (GO:0005856)
GO terms
% of genes
0
5
10
15
20
%
Manual gene annotation with AGOA
Data mining using AGENDA
GO
Gene network analysis using Cytoscape and the Golorize plugin
Functional annotation using AmiGO GO Term Enrichment tool

GO category	% of array	% of candidate genes	P value	Adjusted P value
Biological process				
detection of external stimulus	0,60%	17,60%	3,19E-23	2,32E-21
detection of abiotic stimulus	0,60%	17,60%	3,19E-23	2,32E-21
response to radiation	0,80%	18,80%	7,32E-20	3,54E-18
sensory perception	3,33%	23,20%	6,30E-16	2,28E-14
sequestering of metal ion	0,03%	2,40%	8,13E-05	2,36E-03
glycerol ether metabolic process	0,05%	2,40%	1,60E-04	3,87E-03
anatomical structure morphogenesis	10,89%	22,40%	1,94E-04	4,03E-03
signal transduction	14,16%	24,80%	1,28E-03	2,07E-02
transport	20,23%	32%	1,49E-03	2,16E-02
cell-cell signaling	2,39%	7,20%	3,79E-03	4,58E-02
homeostatic process	0,77%	4%	3,67E-03	4,58E-02
Molecular function				
ion transporter activity	3,13%	10,78%	8,28E-06	6,72E-04
channel or pore class transporter activity	1,26%	6,59%	1,60E-05	6,72E-04
microtubule motor activity	0,48%	4,19%	3,16E-05	8,84E-04
Cellular component				
extrinsic to membrane	0,69%	12,50%	1,35E-09	2,62E-08
plasma membrane part	4,12%	23,75%	1,16E-09	2,62E-08

Figure 18: Gene-Ontology based framework to annotate genes of hearing.

GO-based methods (underlined) for annotating auditory genes applied in this thesis were integrated into the Gene Ontology-based framework to annotate genes of hearing. This aim of this novel framework was to provide a functional link between various methods to facilitate consequent usage of related findings.

5.4.2 Evaluating usability of the GO database in hearing research

The GO-based framework to annotate genes of hearing showed how the GO database can be used in hearing research. Although the GO database provides a powerful infrastructure for the gene annotation described above, it also has certain limitations. These limitations are also directly inherited by the Gene Ontology-based framework to annotate genes of hearing. Thus, defining and discussing pros and contras of using the GO database in annotating auditory genes was one of the aims of this thesis. It was that the GO database has four strong and four weak points (Table 14). Results of the evaluation GO database are summarized below. For some features, specific comments for hearing research are also provided.

Table 14: Pros and cons of using the GO database in gene annotation.

Pros	Contras
Universality	Absence of temporal information
Compatibility	Absence of spatial information
Versatility	Experimental vs. electronic annotations
Broad biological scope	Missing annotations
Sustainability	Genes vs. phenotypes

Pros of using the GO database in gene annotation:

1. Universality: Concerns numerous species including human and genetic model organisms. This property is ideal for hearing research as many of the auditory genes are investigated in the model organisms of deafness.
2. Compatibility: The database can be programmatically accessed and conveniently implemented in data mining and genomic data analysis. This facilitates summarizing current findings (for example, cellular component analysis of auditory gene products with AGENDA) and evaluating results from high-throughput experiments (for example, microarrays).
3. Versatility: In addition to the existing bioinformatics tools implementing the GO database, many new applications are being continuously developed. Parts of this huge toolkit can be useful while investigating genetics of deafness.
4. Broad biological scope: The GO database contains data about different areas of biology. Thus, this source does not only list of auditory genes (primary information), but also secondary information about these genes, such as, their associated Cellular Compartments, Molecular Functions and Biological Processes. This feature can be easily used to functionally annotate and categorize auditory genes.
5. Sustainability: The GO project is the long term project backed by many bioinformatics databases and communities. In addition, since all releases of the GO database are maintained in the repository, GO data is always traceable.

Contras of using the GO database in gene annotation:

1. Absence of temporal information: GO annotations do not contain tissue-specific information. Thus, it is not possible to define where (in which tissue or cell type) the annotated properties (Cellular Compartments, Molecular Functions and Biological Processes) of a gene product were observed. This fact and the absence of spatial information (see below), were reported earlier by Hildebrand *et al*. [Hildebrand *et al*., 2007] as the limitations of the GO database with regard to hearing research.

2. Absence of spatial information: GO annotations do not contain time-specific information. Thus, it is not possible to define when (at which developmental stage) the annotated properties (Cellular Compartments, Molecular Functions and Biological Processes) of a gene product were observed.

3. Experimental vs. electronic annotations: Users should be aware and cautious about the evidences codes while dealing with GO annotations. Especially handling electronic annotations can be sometimes problematic or confusing for biologists. For example, many human auditory genes are inferred from mouse and as result have the ISS (Inferred from Sequence Similarity) GO evidence code.

5. Missing annotations: Some findings about genes (including genes for hearing) have not been yet recorded by the GO database curators. As result, this knowledge cannot be found in the GO database. Since GO annotations don’t capture all auditory genes there are many auditory genes that are not associated with hearing in the GO database. As result, the user will not see these genes while retrieving the list of auditory genes. The AGOA project was aimed to address this issue and improve the list of auditory genes in human and model organisms of deafness. Still, this project is still not completed and there are many other species not covered.

4. Genes vs. phenotypes: In the GO database, there are annotations for gene products which gene sequence is still unknown. In this case, the name of the gene and its annotations can be based on the mutant phenotype. The corresponding gene was not identified yet. This is the case, for example, in the fly auditory gene list. User should be aware and cautious while dealing with such data.

5.4.3 Approaching challenges and potential of Systems biology of hearing

Due to the limited scope of the GO database (Table 14), it has been impossible until now to answer many questions related with hearing using solely GO data. Extending the GO-based methodology described in this thesis to a broader bioinformatics toolkit and applying it to a wider range of biological data is expected to yield to more comprehensive results. The full power of GO data will be probably revealed by using it in combination with other databases and ontologies. Such an integrative approach is a challenging issue in computational systems biology. Similarly, because of the enormous complexity of the auditory system, it has been impossible until now to answer many questions in this field with a single experimental technique. Thus, experimental systems biology advocates combined usage of multiple experimental techniques and concentrates on finding optimal solutions for this task. Ultimately, unifying power of different computational and experimental methods is the major challenge facing systems biology of hearing. System biology strategies that combine advanced experimental and computational methods have a breakthrough potential for solve these questions (Figure 19).

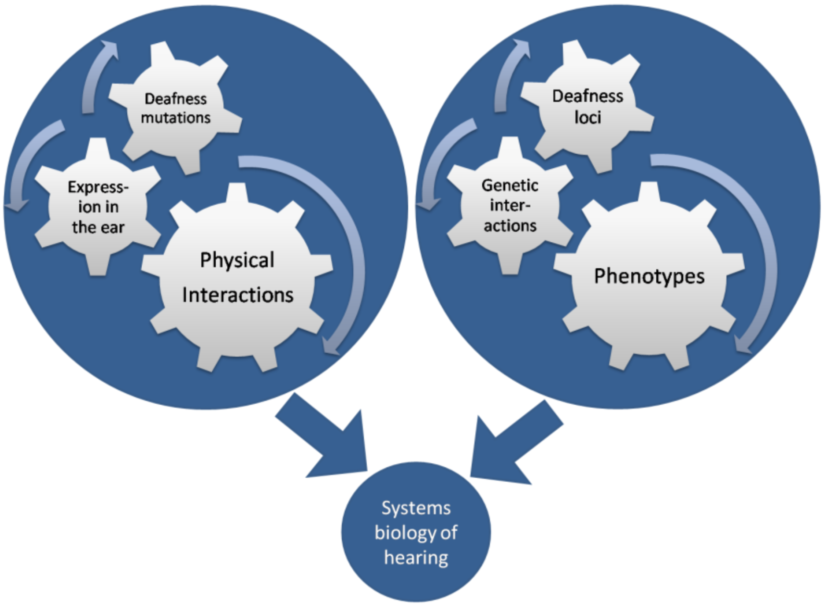


Figure 19: From the GO-based annotation framework towards systems biology of hearing.

Data types from the top left circle were implemented in the "GO-based framework to annotate genes of hearing" and examples of their usage were demonstrated in this thesis. Future perspectives include integration of other data types (examples are shown in the top right circle) to complement the mentioned framework in building the comprehensive data analysis infrastructure. This ultimate bioinformatics infrastructure, together with advanced experimental methods, is expected to enable systems biology approaches in hearing research.

Systems biology can be divided into computational and experimental systems biology. In some case, experiments precede computation and provide data for statistical analysis (for example, high-throughput screens). In other cases, computation is followed by experiments that test validity of computation results (in the form of predictions, models, simulations, prioritized candidate genes, etc.). Systems biology of hearing is expected to improve our understanding of the auditory system by unraveling how its components cooperate functionally and its constituents interact structurally on molecular level to accomplish hearing. (Figure 20).

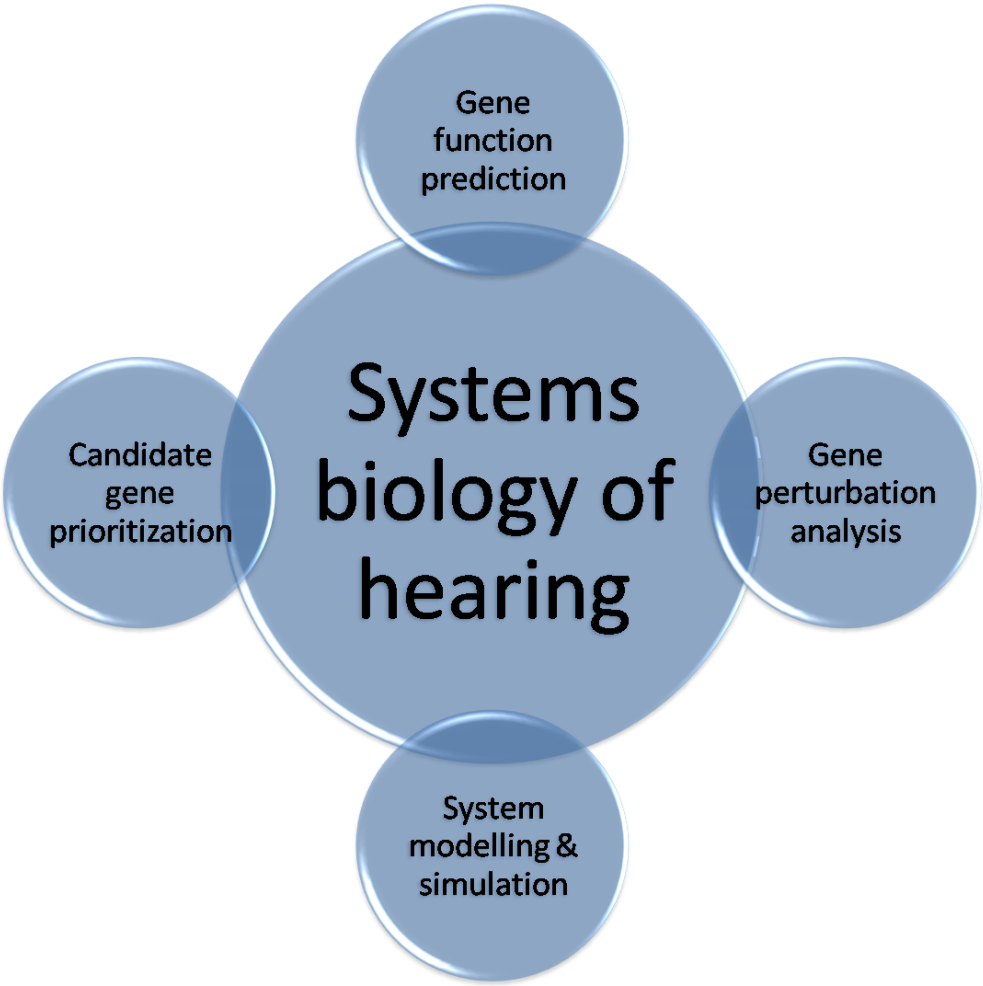

Gene function prediction
Candidate gene prioritization
Systems biology of hearing
Gene perturbation analysis
System modelling & simulation

Figure 20: Systems biology of hearing and *in-silico* analysis.

Systems biology of hearing can be linked to many types of *in-silico* analysis such as prediction of gene function, gene perturbation analysis, candidate gene prioritization, system modeling and simulation. System biology approaches are expected to facilitate auditomics studies by merging different kind of computational and experimental procedures. In this way, new insights into the auditory systems can be gained.

6. Discussion

6.1. Genome-level investigation

6.1.1 Development of Application for mining Gene Ontology data (AGENDA) and its usage in hearing research

6.1.1.1 Development of AGENDA

The Gene Ontology (GO) database is a widely used source in bioinformatics. Availability of robust, powerful tools is crucial for the efficient usage of the GO database. Developing AGENDA (Application for Mining Gene Ontology Data) in order to make the access to - and usage of - target GO data easier was one of the main goals in this thesis. This goal was realized through a new interface that provides diverse query options and enables visualization, download and bookmarking of the obtained results. AGENDA resembles the official GO browser AmiGO [Carbon *et al*., 2009] in the simple queries aimed at single GO terms and gene products. What makes AGENDA novel is how it combines the power of batch queries (GO Slimmer and Boolean query) in data mining and supports the results with data-driven charts and traceable evidences. Batch queries are used to simultaneously access multiple, user-specified GO terms and to compare lists of the associated gene products. Mining GO data with AGENDA can be used while investigating genetic basis of various biological processes such as hearing.

GO Slimmer of AGENDA currently can generate only one GO Slim at a time. Could it beneficial for users to query the GO database simultaneously using multiple GO Slims? From the data mining perspective, comparison of the GO Slims obtained from various species can be of particular interest. The example given in this thesis showed that in both humans and mice the percentages of gene products associated with the plasma membrane were higher in the auditome (41.74% and 39.84%, respectively) than in the genome (25.52% and 14.81%, respectively) (Figure 13). These results hint at the possible enrichment of the plasma membrane proteins in the respective auditomes. The GO term enrichment test still needs to be applied to assess these results. If the enrichment proves to be significant in both species, this could be indicative of a conserved pattern of this enrichment. In this case, extending the comparison to other species (for example, zebrafish and fruit fly) can offer deeper insight into this subject. While the anticipated figure showing the results of the comparison is expected to bear similarities with the GO Slimmer generated by Harris *et al*. [Harris *et al*., 2004] (Figure 2), it is also expected to provide additional features such as viewing the statistical data and clicking on columns for further information. The methodology of the interspecies comparisons described here can also be used for assessing the consistency of results obtained from other GO Slimmer queries (for example, figure 14). Currently, the GO term enrichment test and comparison of different GO Slims (for example, figure 13) are not supported in AGENDA. Therefore, these operations require usage of a third-party tool (the former), such as Excel, and manual processing of the results (the latter). These tasks frequently turn out to be challenging and time-consuming. Enabling statistical testing and handling of multiple GO Slims by GO Slimmer is a prerequisite of the statistically verified and automated GO Slim comparison across species. This, for example, can allow automated generation of the figure 13 including additionally the statistical results. Given the promising outlook of this described approach, empowering GO Slimmer with the needed properties is one of the possible directions of the future development of AGENDA. This enhancement can facilitate comparative genomic studies by taking AGENDA to the next level.

Apart from the GO Slimmer-related issues, future perspectives in the AGENDA development include increasing the functionality of searches: On one side, expanding queries, currently limited to twelve species, to all species covered by the GO database. On another side, adding the upload option that makes it possible for users to provide their own lists of gene products as input for batch queries. These steps may result in the acceptance of AGENDA by a broader user community and can facilitate functional annotation of candidate genes obtained from high-throughput experiments. In addition to the measures concentrating on query refinement, there is also a room for improvement in the interface design. The future prospect includes adding new user-friendly features (such as an autocomplete function in searches) by implementing AJAX technology. All these improvements are expected to increase usability of AGENDA.

Keeping pace with advancement in internet technologies, novel techniques for making fast and powerful web-based applications become available. This progress is also successfully reflected in bioinformatics. For example, implementing the Google Application Program Interfaces (APIs) proved to be useful in several bioinformatics applications [Klekota *et al*., 2006; Obayashi *et al*., 2008; Arakawa *et al*., 2009; Kono *et al*., 2009]. In this thesis, so far as its author is aware, Google Chart API was used for the first time in bioinformatics for the dynamic visualization of GO data. From this perspective, this work and similar studies can inspire proliferation of the "Google-powered bioinformatics" in future.

The rapidly accumulating knowledge about diverse cellular mechanisms continuously increases the size of the GO database and the relevance of the associated bioinformatics tools. Examples shown in this thesis demonstrate the applicability of AGENDA in the context of the hearing research. While AGENDA is expected to contribute to the existing toolkit of GO-based bioinformatics tools [Gene ontology tools website], it is also essential for the GO-based framework to annotated genes of hearing, discussed later in this thesis. AGENDA is available as open source software and can be used freely for non-commercial purposes. AGENDA can be accessed online (URL: http://bioagenda.uni-goettingen.de/) or downloaded, along with the source code and documentation, from the project homepage (URL: http://sourceforge.net/projects/bioagenda).

6.1.1.2 Functional categorization of auditory genes with AGENDA

In this thesis, mining the GO database with AGENDA was used to gain novel insights into the genetics of hearing by analyzing GO annotations related with auditory gene products. For example, it was shown that GO Slimmer of AGENDA provides powerful means for the functional categorization of human auditory gene products. As a next step, respective results from the model organisms of hearing can be obtained and compared. The methodology of the comparison would be similar to the one used while comparing results of the cellular compartments analyses in humans and mice (Figure 13). Such a crossspecies comparison would clarify whether there is a common recognizable signature for the distribution of functions in the auditomes of these organisms.

6.1.2 Manual gene annotation with the Auditory Gene Ontology Annotation (AGOA) project

6.1.2.1 Improving lists of genes annotated to hearing in the GO database

Improving lists of auditory genes is expected to result into a comprehensive annotation of auditory genes in the GO database and provide a general overview of the field. This work is especially valuable while it bridges the findings in different species and facilitates exchange of information between the experts working on them. Preliminary results demonstrated that the GO database is a valuable data source for hearing research.

6.1.2.2 Revision of the evidences for genes annotated to hearing in the GO database

Completing the revision of the evidences for auditory genes is expected to make the GO database even more attractive data source. In addition, this can also lead to better results in future bioinformatics analyses.

6.1.2.3 Chronological overview of the auditory gene discoveries

The chronological overview of the auditory gene discoveries demonstrates the role of successful screens as the major driving forces in the discovery of auditory genes that despite the difficulty of the subsequent positional cloning and characterization. This notion is supported by a recent study by Senthilan et al. [Senthilan *et al.*, 2012] that used microarrays for screening and doubled the number of fly auditory genes (These novel auditory genes are still to be included into the GO database and into the chronological overview). This finding emphasizes the potential of high-throughput analyses such as the gene expression microarrays in auditory research.

In conclusion, the AGOA project included improvement of auditory genes lists, revision of relevant evidences and chronological overview of related discoveries.

The work was performed separately for each species. While goals related with fruitfly and zebrafish auditory genes are essentially accomplished and presented in this thesis, the work concentrating on human and mouse auditory genes is still in progress. Preliminary results presented in this thesis showed the feasibility of the employed approach.

6.2. Transcriptome-level investigation

6.2.1 Functional annotation of candidate auditory genes using the AMIGO GO Term Enrichment tool

Functional annotation using the AMIGO GO term enrichment tool revealed significantly enriched GO terms in the list of candidate gene obtained from the screen. The GO term enrichment analysis provided a better understanding of the gene list. This analysis also showed that the AmiGO GO term enrichment tool is a suitable tool and the GO database is a valuable data source for the functional annotation of microarray data. Future perspective includes gene network-based analysis of the candidate genes using the standard protocol [Cline MS *et al.*, 2007].

6.3. Interactome-level investigation

6.3.1 Reconstruction of the auditory gene network using Cytoscape

The explorative study and resulted in the generation of the auditory gene. Current findings suggests the applicability of the approach but more investigation is needed to evaluate this preliminary data in order to obtain significant biological insights.

6.3.2 Annotation of the auditory gene network using the GOlorize plugin

The annotation of the auditory gene network helped in interpreting the interactions and demonstrates the usability of the GO database together with Cytoscape for this purpose. Future perspective includes cellular component analysis and functional categorization of genes within the auditory gene network using the

Cytoscape Cerebral plugin [Barsky *et al.*, 2007]. GO data can be manually used as input for this plugin to obtain more information about the auditory gene network.

6.4. Developing Gene Ontology-based framework to annotate genes of hearing

6.4.1 Structure of the GO-based framework to annotate genes of hearing

Over the last decade, deafness research has witnessed a considerable progress in the discovery and characterization of genes involved in hearing. This breakthrough in the field is accompanied by novel opportunities and unprecedented challenges. Archiving current findings about auditory genes and using this data in the analysis of new experimental tasks are main issues in this thesis. Results described above demonstrate that the GO database is a valuable data source for the gene annotation (with the AGOA project), functional categorization (with AGENDA), functional annotation (with AmiGO GO term enrichment tool) and network analysis (with Cytoscape and GO-based plugins) of auditory genes. Thus, the GO database can be used in hearing research both as a repository of auditory genes (with the AGOA project) and as a data source for the analysis of genomics, transcriptomics and interactomics data.

"GO-based framework to annotate genes of hearing" was developed by combining different GO-based methods used so far in this thesis. This framework describes gene annotation on three "omics" levels (genomics, transcriptomics and interactomics) and offers practical links between these annotation types. This results into the functional integrity of the consecutive annotation steps and facilitates interpretation of available data. As result, it is possible to a obtain research workflow spanning from an experiment to a new hypothesis through the parts of this framework. This can be especially evidence in microarray screens (Figure 21). Results in this thesis suggests that the GO database can be used successfully within this practical framework for annotating the auditome in a systematic and sustainable manner.

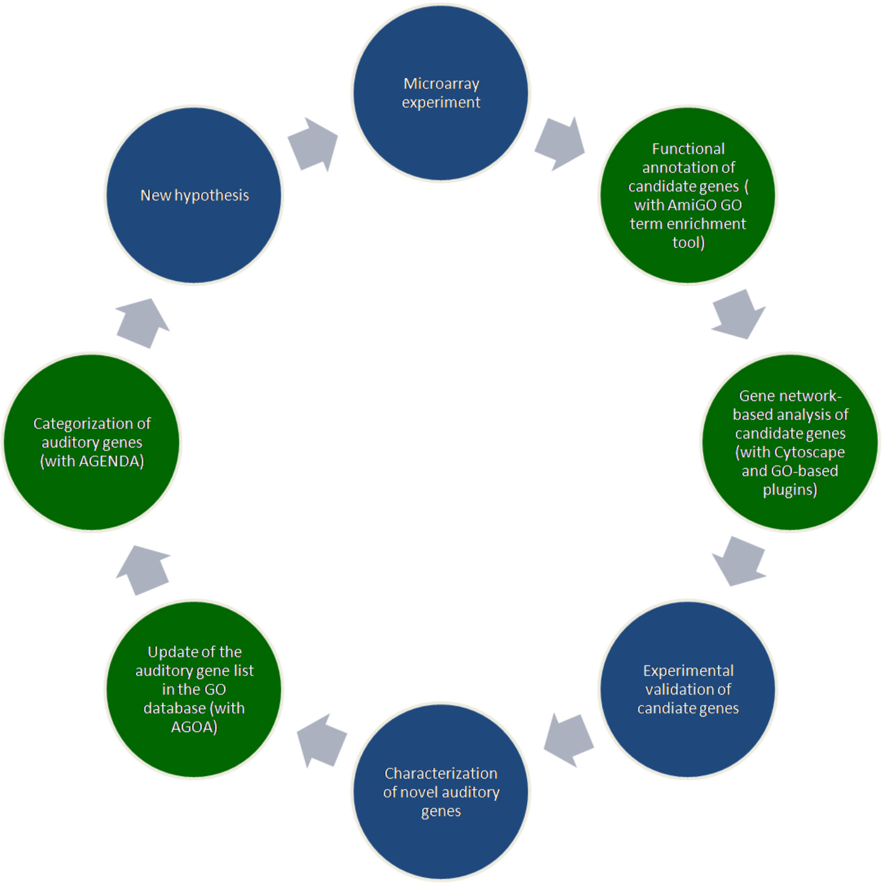

Microarray experiment
Functional annotation of candidate genes (with AmiGO GO term enrichment tool)
Gene network-based analysis of candidate genes (with Cytoscape and GO-based plugins)
Experimental validation of candiate genes
Characterization of novel auditory genes
Update of the auditory gene list in the GO database (with AGOA)
Categorization of auditory genes (with AGENDA)
New hypothesis

Figure 21: GO-based framework to annotate genes of hearing, adapted for microarray screens.

The GO-based framework to annotate genes of hearing can be useful in microarray studies. The loop consisting of the experimental, theoretical and computational studies leads from the microarray experiment to the new hypothesis. Examples described in this thesis demonstrate how the effective usage of the GO database within the proposed framework facilitates auditomics by accelerating the loop and leveraging the results. In auditomics, the GO database can be used to interpret microarray results and helps to review auditory genes. Interpreting candidate gene lists obtained from screens is achieved by functional annotation (using, for instance, AmiGO GO term enrichment tool) and gene network-based analysis (using, for example, Cytoscape and GO-based plugins). Reviewing auditory genes includes update (for example, with the AGOA project) and characterization (using, for example AGENDA tool) of the auditory gene list in the GO database. Steps represented by the green circle are parts of the “GO-based framework to annotate genes of hearing” and examples for each of these steps were demonstrated in this thesis. Blue circles show steps that include the experimental design, the microarray experiment and follow-up procedures such validation of candidate auditory genes and their subsequent characterization.

6.4.2 Evaluating usability of the GO database in hearing research

Despite its usability, possible limitations of the GO database also constrain the power of the Gene Ontology-based framework to annotate genes of hearing. In this thesis, pros and contras of using the GO database in gene annotation were defined. Four main strong and four weak points summarized the current situation (Table 14). Specific comments for hearing research were provided for some features.

6.4.3 Approaching challenges and potential of Systems biology of hearing

Since both bioinformatics and auditomics develop very rapidly, scientists continuously face new challenges. The GO-based framework to annotate genes of hearing has a potential of inspiring novel solutions for some of the complex problems scientist working on auditomics will face in future. Therefore, this thesis can be regarded as a bridge towards more powerful frameworks that will probably be on demand in future.

6.4.4 Outlook

The framework presented in this thesis describes various applications of the GO database in auditomics and shows how they can be utilized together in order to interpret more efficiently relevant experimental findings. While the framework itself can be optimized and developed further to meet the specific and growing needs of scientist, there is still room for improvement in the projects and tools described within the framework. All studies conducted in this these, except the transciptome-level investigation, were designed and initiated with this thesis. From the genome-level studies presented in this thesis, the AGENDA project is completed [Ovezmyradov *et al.*, 2012] and the resulting web-based application has taken its place among other GO-based tools. The AGOA project is still in progress and expected to provide much needed annotations of auditory genes. The improved auditory content can make the GO database more attractive for the auditomics community and increase the relevance of AGENDA and the GO-based annotation framework. In addition to these two projects, the transcriptome-level investigation of auditory genes in *Drosophila* is finished [Senthilan *et al.*, 2012] and the interactomics-level study of auditory gene networks is still ongoing.

Although this thesis concentrated solely on the GO-based solutions, the author is aware that there are many technologies in bioinformatics that can be very useful in auditomics. The ultimate goal is to combine these technologies and corresponding data sources in order to lay down the computational infrastructure for the systems biology of hearing. The latter is expected to enable better understanding of hearing and novel approaches to combat deafness. Given all this, this work can be regard as a step further towards systems biology of hearing.

Finally, the structure of the GO-based annotation framework here can be modified to suit the needs of the scientists working on other “omics” fields. While the scope of the AGOA project was to improve auditory content of the GO database, a scientific community interest in another subject can develop its specific GO annotation project for their own needs. Due to the variety of biological areas within the GO database and the versatility of the implemented GO-based tools, the framework described in this thesis can serve as model for other annotation studies.

7. Abbreviations

AGOA :	Auditory Gene Ontology Annotation
AGENDA:	Application for Mining Gene Ontology Data
BIND:	Biomolecular interaction network database
DBMS:	Database management system
GO:	Gene Ontology

8. Literature

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC.. The genome sequence of Drosophila melanogaster. *Science*. Mar 24; 287(5461):2185-95 (2000).

Alam-Faruque Y, Dimmer EC, Huntley RP, O'Donovan C, Scambler P, Apweiler R. The Renal Gene Ontology Annotation Initiative. *Organogenesis*. Apr-Jun; 6(2):71-5 (2010).

Alam-Faruque Y, Huntley RP, Khodiyar VK, Camon EB, Dimmer EC, Sawford T, Martin MJ, O'Donovan C, Talmud PJ, Scambler P, Apweiler R, Lovering RC. The impact of focused Gene Ontology curation of specific mammalian systems. *PLoS One*. 6(12):e27541 (2011).

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. Oct 5; 215(3):403-10 (1990).

Arakawa K, Tamaki S, Kono N, Kido N, Ikegami K, Ogawa R, Tomita M. Genome Projector: zoomable genome map with multiple views. *BMC Bioinformatics*. Jan 23; 10:31 (2009).

Aravindhan G, Ramesh Kumar G, Sathish Kumar R, Subha K. AJAX interface: a breakthrough in bioinformatics web applications. *Proteomics Insights*. 2009:2 1–7 (2009).

Ashburner M, Ball CA, JA Blake, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 25(1):25-9, (2000). URL: http://www.geneontology.org. Accessed on 2 September 2012.

Bard JB and Rhee SY. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*. Mar 5(3):213-22 (2004).

Barsky A, Gardy JL, Hancock RE, Munzner T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*. Apr 15; 23(8):1040-2 (2007).

Baxevanis AD. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*. Chapter 1:Unit 1.1 (2009).

Beissbarth T. Interpreting experimental results using gene ontologies. *Methods Enzymol*. 411:340-52 (2006).

Berriz GF, White JV, King OD, Roth FP. GoFish finds genes with combinations of Gene Ontology attributes. *Bioinformatics*. Apr 12; 19(6):788-9 (2003).

Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT; Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res*. Jan; 39(Database issue):D842-8 (2011).

Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*. 2008:67-79 (2008).

Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K, Howe DG, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, Ramachandran S, Ruef BJ, Ruzicka L, Bauer Schaper H, Schaper K, Shao X, Singer A, Sprague J, Sprunger B, Van Slyke C, Westerfield M. ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res*. Jan; 39(Database issue):D822-9 (2011).

Brazma A et al. Standards for systems biology. *Nat Rev Genet*. Aug; 7(8):593-605 (2006).

Brooksbank C and Quackenbush J. Data standards: a call to action. *OMICS*. Summer; 10(2):94-9 (2006).

Brown SD, Hardisty-Hughes RE, Mburu P. Quiet as a mouse: dissecting the molecular and genetic basis of hearing. *Nat Rev Genet*. Apr; 9(4):277-90 (2008).

Brusic V. The growth of bioinformatics. *Brief Bioinform*. Mar; 8(2):69-70 (2007).

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S; AmiGO Hub; Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 25(2):288-9 (2009). URL: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi. Accessed on 2 September 2012.

Chen H, Ding L, Wu Z, Yu T, Dhanapalan L, Chen JY. Semantic web for integrated network analysis in biomedicine. *Brief Bioinform*. Mar; 10(2):177-92 (2009).

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*. 2(10):2366-82 (2007).

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin ZY, Liang W, Marback M, Paw J, San Luis BJ, Shuteriqi E, Tong AH, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pál C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras AC, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C. The genetic landscape of a cell. *Science*. Jan 22; 327(5964):425-31 (2010).

Crick F. Central dogma of molecular biology. *Nature*. Aug 8; 227(5258):561-3 (1970).

Diehl AD, Lee JA, Scheuermann RH, Blake JA. Ontology development for biological systems: immunology. *Bioinformatics*. (2007).

Dror AA, Avraham KB. Hearing impairment: a panoply of genes and functions. *Neuron*. Oct 21; 68(2):293-308 (2010).

Eisen MD, Ryugo DK. Hearing molecules: contributions from genetic deafness. *Cell Mol Life Sci*. Mar; 64(5):566-80 (2007).

Ekker SC, Stemple DL, Clark M, Chien CB, Rasooly RS, Javois LC. Zebrafish genome project: bringing new biology to the vertebrate genome field. *Zebrafish*. Winter; 4(4):239-51 (2007).

Feltrin E, Campanaro S, Diehl AD, Ehler E, Faulkner G, Fordham J, Gardin C, Harris M, Hill D, Knoell R, Laveder P, Mittempergher L, Nori A, Reggiani C, Sorrentino V, Volpe P, Zara I, Valle G, Deegan J. Muscle Research and Gene Ontology: New standards for improved data integration. *BMC Med Genomics*. (2009).

Gabashvili IS, Sokolowski BH, Morton CC, Giersch AB. Ion channel gene expression in the inner ear. *J Assoc Res Otolaryngol*. Sep; 8(3):305-28 (2007).

Galizia CG, Münch D, Strauch M, Nissler A, Ma S. Integrating heterogeneous odor response data into a common response model: A DoOR to the complete olfactome. *Chem Senses*. Sep; 35(7):551-63 (2010).

Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B, Aittokallio T. GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*. Feb 1; 23(3):394-6 (2007).

Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*. Oct; 19(10):551-60 (2003).

Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research*. 11(8):1425-33 (2001).

Gene Ontology website, “Guide to GO Evidence Codes” web page. URL: http://www.geneontology.org/GO.evidence.shtml. Accessed on 2 September 2012.

Gene Ontology website, “GO Annotation Policies and Guidelines” web page. URL: http://www.geneontology.org/GO.annotation.shtml. Accessed on 2 September 2012.

Gene Ontology tools website. URL: http://www.geneontology.org/GO.tools.shtml. Accessed on 2 September 2012.

Gene Ontology database archive. URL: http://archive.geneontology.org. Accessed on 2 September 2012.

Gilbert D. Biomolecular interaction network database. *Brief Bioinform*. Jun; 6(2):194-8 (2005).

GO and GO associated projects website. URL: http://www.geneontology.org/GO.contents.projects.shtml. Accessed on 2 September 2012.

Gene Ontology website, “GO Tools: Term Enrichment” web page. URL: http://www.geneontology.org/GO.tools_by_type.term_enrichment.shtml.
Accessed on 2 September 2012.

Google Chart Tools website. URL: https://developers.google.com/chart/?hl=de-DE. Accessed on 2 September 2012.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. Jan 1; 32(Database issue):D258-61 (2004).

Hildebrand MS, de Silva MG, Klockars T, Campbell CA, Smith RJ, Dahl HH. Gene expression profiling analysis of the inner ear. *Hear Res*. Mar; 225(1-2):1-10 (2007).

Hilgert N, Smith RJ, Van Camp G. Function and expression pattern of non-syndromic deafness genes. *Curr Mol Med*. Jun; 9(5):546-64 (2009).

Klekota J, Roth FP, Schreiber SL. Query Chem: a Google-powered web search combining text and chemical structures. *Bioinformatics*. Jul 1; 22(13):1670-3 (2006).

Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M. Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS One*. Nov 11; 4(11):e7710 (2009).

Kumar S and Dudley J. Bioinformatics software for biologists in the genomics era. *Bioinformatics*. 23(14):1713-7 (2007).

Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. Feb 15; 409(6822):860-921 (2001).

Lewis SE. Gene Ontology: looking backwards and forwards. *Genome Biol*. 6(1):103 (2005).

Lovering RC, Camon EB, Blake JA, Diehl AD. Access to immunology through the Gene Ontology. *Immunology*. Oct; 125(2):154-60 (2008).

Lovering RC, Dimmer E, Khodiyar VK, Barrell DG, Scambler P, Hubank M, Apweiler R, Talmud PJ. Cardiovascular GO annotation initiative year 1 report: why cardiovascular GO? *Proteomics*. May; 8(10):1950-3 (2008).

Lovering RC, Dimmer EC, Talmud PJ. Improvements to cardiovascular gene ontology. *Atherosclerosis*. Jul; 205(1):9-14 (2009).

Lu Q, Senthilan PR, Effertz T, Nadrowski B, Göpfert MC. Using Drosophila for studying fundamental processes in hearing. *Integr Comp Biol*. Dec; 49(6):674-80 (2009).

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science*. Dec 6; 298(5600):1912-34 (2002).

McQuilton P, St Pierre SE, Thurmond J; FlyBase Consortium. FlyBase 101--the basics of navigating FlyBase. *Nucleic Acids Res*. Jan; 40(Database issue):D706-14 (2012).

Mi H, Thomas PD. Ontologies and standards in bioscience research: for machine or for human. *Front Physiol*. Feb 21; 2:5 (2011).

Mitrofanova A, Pavlovic V, Mishra B. Prediction of protein functions with gene ontology and interspecies protein homology data. *IEEE/ACM Trans Comput Biol Bioinform*. May-Jun; 8(3):775-84 (2011).

Morton CC. Genetics, genomics and gene discovery in the auditory system. *Hum Mol Genet*. May 15; 11(10):1229-40 (2002).

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail

M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial sequencing and comparative analysis of the mouse genome. *Nature*. Dec 5; 420(6915):520-62 (2002).

Nicolson T. The genetics of hearing and balance in zebrafish. *Annu Rev Genet*. 39:9-22 (2005).

Obayashi T, Hayashi S, Shibaoka M, Saeki M, Ohta H, Kinoshita K. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res*. Jan; 36(Database issue):D77-82 (2008).

Ovezmyradov G, Lu Q, Göpfert MC. Mining Gene Ontology Data with AGENDA. *Bioinform Biol Insights*. 6:63-7 (2012).

Parkinson N, Brown SD. Focusing on the genetics of hearing: you ain't heard nothin' yet. *Genome Biol*. 3(6):COMMENT2006 (2002).

Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text processing through Web services: calling Whatizit. *Bioinformatics*. Jan 15; 24(2):296-8 (2008).

Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Computational Biology*. 5(7):e1000431 (2009).

Resendes BL, Williamson RE, Morton CC. At the speed of sound: gene discovery in the auditory system. *Am J Hum Genet*. Nov; 69(5):923-35 (2001).

Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*. Sep 15; 26(18):i561-7 (2010).

Senthilan PR, Piepenbrock D, Ovezmyradov G, Nadrowski B, Bechstedt S, Pauls S, Winkler M, Möbius W, Howard J, Göpfert MC. Drosophila Auditory Organ Genes and Genetic Hearing Defects. *Cell*. August 31; 150, 1–13 (2012).

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. Nov; 13(11):2498-504 (2003). URL: http://www.cytoscape.org. Accessed on 3 September 2012.

Steel KP, Kros CJ. A genetic approach to understanding auditory function. *Nat Genet*. Feb; 27(2):143-9 (2001).

Stein LD. Integrating biological databases. *Nat Rev Genet*. May; 4(5):337-45 (2003).

UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. Jan; 40(Database issue):D71-5 (2012).

van den Berg BH, Thanthiriwatte C, Manda P, Bridges SM. Comparing gene annotation enrichment tools for functional modeling of agricultural microarray data. *BMC Bioinformatics*. Oct 8; 10 Suppl 11:S9 (2009).

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. *Science*. Feb 16; 291(5507):1304-51 (2001).

XAMPP software website. URL: http://www.apachefriends.org/en/xampp.html. Accessed on 3 September 2012.

9. Appendix

9.1. Fruitfly auditory genes

Table 15: Fruit fly auditory genes

Gene no.	Gene symbol	Gene full name
1	ato	Atonal
2	btv	Beethoven
3	ck	Crinkled
4	cp309	cp309
5	ct	Cut
6	DCX-EMAP	Doublecortin-domain-containing echinoderm-microtubule-associated protein ortholog
7	dia	Diaphanous
8	Eb1	Eb1
9	f	Forked
10	iav	Inactive
11	Kap3	Kinesin associated protein 3
12	Klp64D	Kinesin-like protein at 64D
13	nan	Nanchung
14	nompA	no mechanoreceptor potential A
15	nompB	no mechanoreceptor potential B
16	nompC	no mechanoreceptor potential C
17	nompE	no mechanoreceptor potential E
18	nompF	no mechanoreceptor potential F
19	nompI	no mechanoreceptor potential I
20	nompJ	no mechanoreceptor potential J
21	rempA	reduced mechanoreceptor potential A
22	rempD	reduced mechanoreceptor potentials D
23	Rfx	Rfx
24	salm	spalt major
25	salr	spalt-related
26	tilB	touch insensitive larva B
27	tko	technical knockout

28	Tmhs	Tetraspan membrane protein in hair cell stereocilia ortholog
29	unc	Uncoordinated
30	uncl	uncoordinated-like

Source: GO database as of June 2012.

9.2. References for fruit fly auditory genes

Table 16: References for fruit fly auditory genes.

Gene no.	Gene symbol	GO evidence code	Reference
1	ato	IMP	PMID:10934246
1	ato	IMP	PMID:12203727
2	Btv	IMP	PMID:10934246
2	Btv	IMP	PMID:12642657
3	ck	IMP	PMID:15886106
4	cp309	IMP	PMID:15184400
5	ct	IMP	PMID:18820445
6	DCX-EMAP	IMP	PMID:20975667
7	dia	IMP	PMID:19102128
7	dia	IMP	PMID:20624953
8	Eb1	IMP	PMID:15591130
9	f	IMP	PMID:19102128
10	iav	IMP	PMID:15483124
10	iav	IMP	PMID:16819519
10	iav	IMP	PMID:19666538
11	Kap3	IMP	PMID:14521834
12	Klp64D	IMP	PMID:14521834
13	nan	IMP	PMID:12819662
13	nan	IMP	PMID:16819519
13	nan	IMP	PMID:19666538
14	nompA	IMP	PMID:10934246
14	nompA	IMP	PMID:11239432
14	nompA	IMP	PMID:12642657
15	nompB	IMP	PMID:10934246
15	nompB	IMP	PMID:14521833

16	nompC	IMP	PMID:10934246
16	nompC	IMP	PMID:12642657
16	nompC	IMP	PMID:16819519
16	nompC	IMP	PMID:19666538
17	nompE	IMP	PMID:10934246
18	nompF	IMP	PMID:10934246
19	nompI	IMP	PMID:10934246
20	nompJ	IMP	PMID:10934246
21	rempA	IMP	PMID:10934246
21	rempA	IMP	PMID:19097904
22	rempD	IMP	PMID:10934246
23	Rfx	IMP	PMID:12403718
24	salm	IGI	PMID:11934862
24	salm	IMP	PMID:12925729
25	salr	IGI	PMID:11934862
25	salr	IMP	PMID:12925729
26	tilB	IMP	PMID:10934246
26	tilB	IMP	PMID:12642657
26	tilB	IMP	PMID:20215474
27	tko	IMP	PMID:11560901
28	Tmhs	IMP	PMID:19102128
29	unc	IMP	PMID:10934246
29	unc	IMP	PMID:15226257
30	uncl	IMP	PMID:10934246

Source: GO database as of June 2012.

9.3. List of Figures

9.4. List of Tables

10. Acknowledgement

I thank to

- Prof. Martin Göpfert for his supervision that made this work possible. He showed me the meaning of the “doctoral father”. It was an invaluable experience to work with him.
- My doctoral thesis committee and Prof. Morgenstern for guidance and helpful discussions.
- Simon Qianhao Lu for suggestions about the AGENDA project and manuscript.
- GWDG team members for their technical support related with the AGENDA webserver.
- Gene Ontology Consortium for providing the GO database
- Google for providing the Charts API infrastructure
- Professors and biocurators that significantly contributed to the AGOA project.
- Pingkalai R. Senthilan, David Piepenbrock for their cooperation in the Screen project.
- Bart Geurten for designing the AGENDA logo and giving feedback on my doctoral project.
- IGS-GFG, DFG and GGNB for funding.
- GGNB for providing an excellent doctoral program.
- Everybody in our lab for the team work, help and friendship.
- All members of my family for their help and encouragement.
- Many other colleagues and friends who are not included here.

11. Curriculum vitae

Personal Information

Name	Guvanchmyrat Ovezmyradov
Date of birth	21th April 1984
Nationality	Turkmen
Place of birth	Ashgabat, Turkmenistan

Education

2008-Present

Doctoral Thesis: Gene Ontology-based framework to annotate genes of hearing
Supervisor: Prof. Dr. Martin C. Göpfert
Department of Cellular Neurobiology
Georg-August-University Göttingen
Göttingen, Germany

2005-2007

Master of Science in Molecular Biology and Genetics. Master Thesis: *In silico* analysis of mutant p53(R249S) oncogenicity in hepatocellular carcinoma
Supervisor: Assist. Prof. Dr. Rengül Çetin Atalay
Department of Molecular Biology and Genetics
Bilkent University, Ankara, Turkey

2001-2005

Bachelor of Science in Biology
Hacettepe University, Ankara, Turkey

Publications

[1] Ovezmyradov G, Lu Q, Göpfert MC. Mining Gene Ontology Data with AGENDA. *Bioinform Biol Insights*. 6:63-7 (2012).

[2] Senthilan PR, Piepenbrock D, Ovezmyradov G, Nadrowski B, Bechstedt S, Pauls S, Winkler M, Möbius W, Howard J, Göpfert MC. Drosophila Auditory Organ Genes and Genetic Hearing Defects. *Cell*. August 31; 150, 1–13 (2012).