A Big Data Perspective on the Genomics of Hearing Loss

Authors
Barbara Vona, Marcus Müller, Saskia Dofek, Martin Holderried, Hubert Löwenheim, Anke Tropitzsch

Affiliation
Eberhard Karls Universität, Universitäts-Hals-Nasen-Ohren-Klinik Tuebingen, Germany

Key words
big data, genetics, genomics, GJB2, hearing loss diagnostics, high-throughput sequencing, variant interpretation

Bibliography
DOI https://doi.org/10.1055/a-0803-6149
Online-Publikation: 2019
Laryngo-Rhino-Otol 2019; 98: S58–S81

ISSN 0935-8943

Correspondence
Prof. Dr. med. Hubert Löwenheim
Univ. HNO-Klinik
Elfriede-Aulhorn-Str. 5
D-72076 Tübingen
Germany
hubert.loewenheim@med.uni-tuebingen.de

ABSTRACT

The completion of the human genome, the most fundamental example of big data in science and medicine, is the remarkable product of multidisciplinary collaboration and is regarded as one of the largest and most successful undertakings in human history. Unravelling the human genome means not only identifying the sequence of its more than 3.2 billion nucleotide bases, but also understanding disease-associated variations and applying this knowledge to patient-tailored precision medicine approaches. Genomics has moved at a remarkable pace, with much of the propelling forces behind this credited to technological developments in sequencing, computing, and bioinformatics, that have given rise to the term “big genomics data.” The analysis of genetics data in a disease context involves the use of several big data resources that take the form of clinical genetics data repositories, *in silico* prediction tools, and allele frequency databases. These exceptional developments have cultivated high-throughput sequencing technologies that are capable of producing affordable high-quality data ranging from targeted gene panels to exomes and genomes. These new advancements have revolutionized the diagnostic paradigm of hereditary diseases including genetic hearing loss.

Dissecting hereditary hearing loss is exceptionally challenging due to extensive genetic and clinical heterogeneity. There are presently over 150 genes involved in non-syndromic and common syndromic forms of hearing loss. The mutational spectrum of a single hearing loss associated-gene can have several tens to hundreds of pathogenic variants. Moreover, variant interpretation of novel variants can pose a challenge when conflicting information is deposited in valuable databases. Harnessing the power that comes from detailed and structured phenotypic information has proven promising for some forms of hearing loss, but may not be possible for all genetic forms due to highly variable clinical presentations. New knowledge in both diagnostic and scientific realms continues to rapidly accumulate. This overwhelming amount of information represents an increasing challenge for medical specialists. As a result, specialist medical care may evolve to take on new tasks and facilitate the interface between the human genetic diagnostic laboratory and the patient. These tasks include genetic counselling and the inclusion of genetics results in patient care.

This overview is intended to serve as a reference to otolaryngologists who wish to gain an introduction to the molecular genetics of hearing loss. Key concepts of molecular genetic diagnostics will be presented. The complex processes underlying the identification and interpretation of genetic variants in particular would be inconceivable without the enormous amount of data available. In this respect, "big data" is an indispensable prerequisite for filtering genetic data in specific individual cases and making it clear and useful, especially for clinicians in contact with patients.

Inhaltsverzeichnis

ABBREVIATIONS

A	adenine
C	cytosine
CADD	Combined Annotation Dependent Depletion
COL11A2	collagen type XI, alpha-2
DFNA2A	deafness, autosomal dominant 2 A locus
DFNA3A	deafness, autosomal dominant 3 A locus
DFNA6/14/38	deafness, autosomal dominant 6/14/38 locus
DFNA13	deafness, autosomal dominant 13 locus
DFNB1A	deafness, autosomal recessive 1 A locus
DFNB16	deafness, autosomal recessive 16 locus
ddNTP	dideoxynucleotide
dNTP	deoxynucleotide
DVD	Deafness Variation Database
E	embryonic day
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium Browser
G	guanine
Gb	gigabase
GJB2	gap junction protein beta 2
GJB6	gap junction protein beta 6
GME	Greater Middle Eastern Variome
gnomAD	genome aggregation database
HGMD	Human Gene Mutation Database
HPO	Human Phenotype Ontology
KCNQ4	potassium channel voltage gated KQT-like subfamily member 4
LOVD	Leiden Open Variation Database
MAF	minor allele frequency
mRNA	messenger ribonucleic acid
MYO1A	myosin IA
P	postnatal day
PCR	polymerase chain reaction
SHIELD	Shared Harvard Inner-Ear Laboratory Database
SIFT	Sorting Intolerant from Tolerant
STRC	stereocilin
T	thymine
WFS1	wolframin ER transmembrane glycoprotein

1. Glossary

Autosome: Non-sex chromosome.
Baits: Capture probes that are made out of oligonucleotides complimentary to a region of interest for sequencing.
Copy number variation: Deletions or duplications of chromosomal regions that affect the number of gene copies.
Coverage: The collection of aligned sequencing reads across a nucleotide or region of interest.
Dideoxynucleotides: Modified deoxynucleotides that lack a 3′ hydroxyl group to inhibit chain elongation in Sanger sequencing.
DNA library: A collection of amplified DNA fragments for high-throughput sequencing.
Exome: The part of the genome that is composed of exons that are translated into proteins.
Exome sequencing: Sequencing of all exons in coding genes.
Exon: A region of a gene that encodes a protein.
Gene panel diagnostics: Sequencing of selected genes relevant to a specific disease.
Genome: The complete set of DNA in an organism.
Gigabase: 10^9 nucleotide bases.
High-throughput sequencing: A scalable and relatively cheap sequencing method that can range from gene panels to genome sequencing.
Indel: A term for the insertion or deletion of one or more bases in a genome.
***In silico* gene panel:** A computational filter applied to exome or genome sequencing data that restricts the variants for analysis in a selected sub-set of genes.
***In silico* pathogenicity prediction:** Computational tools that predict the pathogenicity of variants.
Intron: A non-coding region of a gene between two coding exons.
Kilobase: 1,000 nucleotide bases.
Megabase: 1,000,000 nucleotide bases.
Minor allele frequency: The frequency of the less common allele (MAF).
Missense variant: A nucleotide substitution that changes an amino acid.

Moore's law: An observation that the number of transistors on a dense integrated circuit doubles every two years, thus cutting costs of transistors in half.
Non-synonymous variant: A nucleotide substitution that alters the amino acid sequence.
Nonsense variant: A nucleotide substitution that results in a premature stop codon during transcription.
Phenome: The comprehensive description of the phenotype and course of disease in an individual.
Read: A short fragment of sequence.
Sanger sequencing: A type of sequencing that uses a chain-termination method with chemically modified dideoxynucleotides that determines the nucleotide sequence.
Secondary findings: A genetic test result that is unrelated to the primary disease indication.
Sequencing gap: A region that is poorly covered or missed during sequencing usually due to technical reasons.
Splice site variant: A variant that impacts normal gene splicing during translation.
Start gain variant: A variant that causes a new translation initiation site.
Start loss variant: A variant that disrupts the normal translation initiation site.
Stop gain variant: A variant that results in a premature stop codon during transcription.
Stop loss variant: A variant that removes the terminator codon and results in an elongated transcript.
***STRC*:** A gene that encodes stereocilin, a structural protein in the stereocilia of the outer hair cells of the inner ear, and causes autosomal recessive hearing loss (DFNB16).
Synonymous variant: A nucleotide substitution that does not alter the amino acid sequence.
Terabase: 10^{12} nucleotide bases.
Variant: A deviation from the reference sequence.

2. Big Data in the Age of Genomics

"Big data" is an increasingly ubiquitous concept in healthcare. The fields of genetics and genomics have embraced the big data revolution particularly well, so much that it is impossible to extract and interpret meaningful results without benefitting from the masses of genomic information stored in numerous data repositories. The most fundamental example of big data in this field is the human genome sequence, which in the most basic sense, serves as a DNA sequence blueprint for the more than 20,000 genes in the human genome. The completion of the Human Genome Project (HGP) in 2003 delivered a reference human genome sequence. The success of the HGP represents a remarkable milestone that has empowered and accelerated the understanding of variation in the human genome. This fundamental knowledge has also revolutionized sequencing technologies, as well as our understanding of normal and disease-associated human variation.

Human genome variation not only accounts for our unique characteristics, but also determines the chances for targeted treatment in the event of disease. Our human genome is the product of generations of migration, selection, and adaptation. Naturally occurring errors in the germline or somatic cells can introduce both small and large changes, termed genetic variation, into our genomes, much of which can be considered benign or polymorphic, while other changes are associated with disease states. These changes can affect single nucleotides or bases (adenine (A), thymine (T), guanine (G), and cytosine (C)) or several million nucleotides in the genome (e.g. large duplications or deletions). However, changes can also involve whole chromosomes (e. g. monosomy, trisomy) or involve the exchange of genetic material within different parts of a single chromosome (intrachromosomal rearrangement) or between different chromosomes (interchromosomal rearrangement).

2.1. Genetic variation—benign or pathogenic?

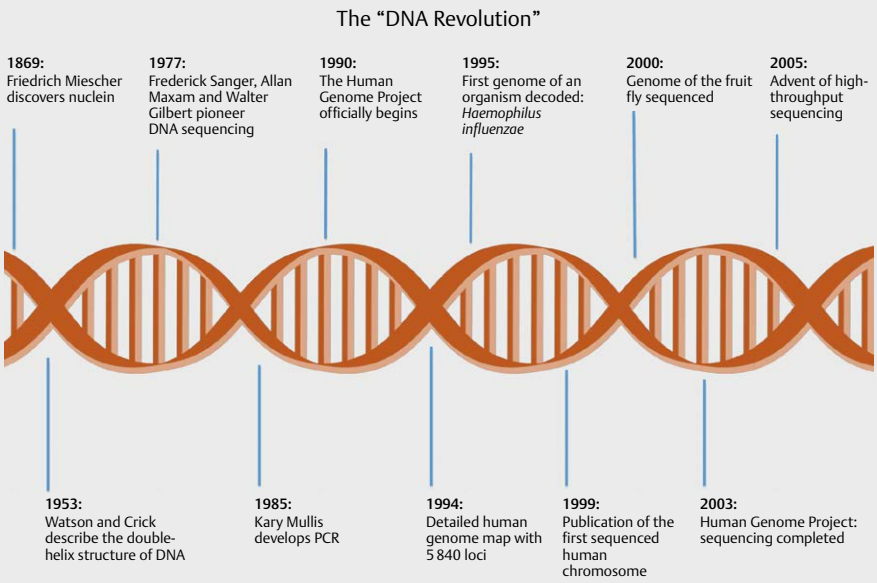
One of the largest sequencing studies to date analysed human variation in 60,706 individuals and estimated that the protein coding sequence (exome sequence) of human genes harbours the equivalent of one variant every eight nucleotide positions across the collective pool of individuals studied [1]. Deciphering patterns in this variation is difficult, as it follows a non-uniform distribution and the density of variation is influenced by mutational properties and selective pressures.

Another important form of genetic variation is called copy number variation (CNV). CNVs are defined as large duplications and deletions ranging between 50 and 3,000,000 base pairs. Based on this type of genetic variation, up to 9.5 % of the genome can vary in healthy individuals and be involved in "gains" (gene duplications) or "losses" (gene deletions) [2]. Consequently, the 3.2 billion nucleotides normally in our genome can vary by ± 9.5 %. This impressively demonstrates some of the resilience of the human genome to large variation.

One of the basic tasks of clinical interpretation of genomic data is the differentiation between normal and pathogenic variation [3–9]. In the last 15 years since the completion of the human genome sequence, genomics research has begun to characterize variation tolerance and intolerance by studying healthy and affected individuals and comparing the changes and genes that are repeatedly affected by diseases. The genome's remarkable complexity makes the field of genetics particularly interesting and extremely dynamic as our knowledge continues to build at an impressive rate, much to the credit of rapid technological advancements in sequencing technologies.

3. Paving a Path for the Genomics Revolution

The path to the human genome (▶ **Fig. 1**) has some of its early roots at the University of Tübingen. In 1869, Friedrich Miescher, a physician from a well-known medical family, discovered and isolated the nucleus (▶ **Fig. 2**) from the nuclei of white blood cells [10]. After his medical studies in Basel, Miescher first had to undergo clinical training before he started his career. However, due to his childhood hearing loss, Miescher deliberately refrained from clinical work and turned to research in Tübingen [10]. Although Miescher did not fully recognise the importance of his discovery, he nevertheless assumed that the substance he had isolated was the molecule of hereditary. This was confirmed 75 years later, in 1944, by the classical experiments of Avery, MacLeod, and McCarty [11, 12]. In 1953, the struc-



▶ **Fig. 1** The DNA revolution. A timeline of selected milestones throughout history that gave rise to modern molecular genetics. It all began in 1869 with the discovery of nucleic acid by the hearing impaired physician Friedrich Miescher in Tübingen.

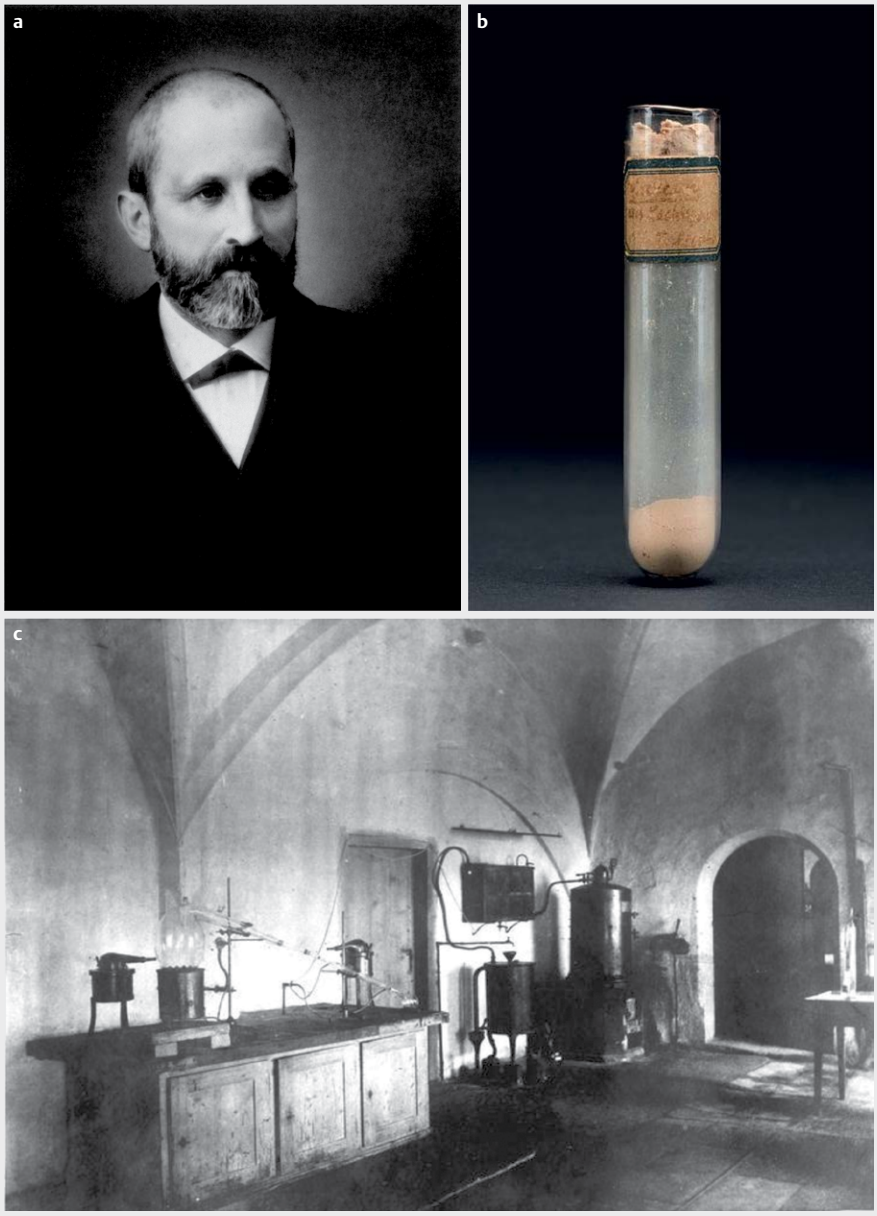
ture of DNA was resolved by Watson and Crick, profiting from data generated from Rosalind Franklin and Maurice Wilkins, giving rise to the field of molecular biology [13, 14]. It was more than two decades after the structure of DNA was uncovered that the first two “robust” sequencing methods emerged. Maxam-Gilbert sequencing [15] uses a chemical-cleavage based method. This technique uses radioactive labelling of DNA fragments that are chemically cleaved at each of the nucleotides (A, T, G, C) to determine the sequence [16]. An alternative form of sequencing called “Sanger sequencing” was named after one of the developers, Fredrick Sanger, and is based on altered ribose sugars [17]. This method is also referred to as a “chain-termination” or “di-deoxy technique” because it uses dideoxynucleotides (ddNTPs) that lack a 3’-hydroxyl, thus halting extension of a growing nucleotide chain. Using 4 different reactions with four different dNTP/ddNTP mixtures, 1 corresponding to each nucleotide, a reaction integrates both normal dNTPs that allows for extension of a growing DNA strand, but also ddNTPs that causes the DNA strand to terminate randomly (▶ **Fig. 3**). The sequence fragments are run on a gel and the nucleotide order can be determined. The Maxam-Gilbert sequencing method was widely used for many decades because it directly analysed DNA fragments, while the early Sanger sequencing methods required clonal amplification of a DNA fragment. However, after further development, the popularity of Sanger sequencing surpassed that of Maxam-Gilbert sequencing, so much that it dominated sequencing methods for a quarter century and is still widely used today for its reliability.

In 1985, polymerase chain reaction (PCR) was developed with the idea of using two primer pairs flanking a region of DNA to be copied using DNA polymerase, an idea that pioneered modern molecular biology [18]. This method was a key part of the workhorse for the HGP, which was in planning stages in the late 1980s and officially launched in 1990 [19]. Ironically, there was much opposition to this project in the West German government that was deeply rooted in ethical concerns [20]. Nevertheless, West Germany was one of only six countries that collectively performed nearly all of the sequencing in the Human Genome Project [21].

In 1994, the first high-density human genome map with 5,840 loci was published that served as a major leap forward in genetic physical maps, greatly enhancing efforts for gene identification [22]. The following year, the first organism *Haemophilus influenzae*, was sequenced [23] that followed four years later by the first human chromosome, and second smallest of the autosomes, chromosome 22, [24]. The *Drosophila melanogaster* genome was sequenced in the year 2000 [25], paving the way for exploration of conserved genes responsible for hereditary diseases in humans [26]. The completion of the human genome sequence in 2003 not only opened a new era in medicine, but also promoted significant developments in DNA sequencing and computational technologies. Less than two years later in 2005, the first high-throughput sequencing method emerged from George Church’s group [27], that used a novel cyclic array and multiplex sequencing approach that dropped the cost of sequencing using this method to roughly one-ninth the cost of Sanger sequencing. This transformative method was voted “Method of the Year” in 2007 [28].

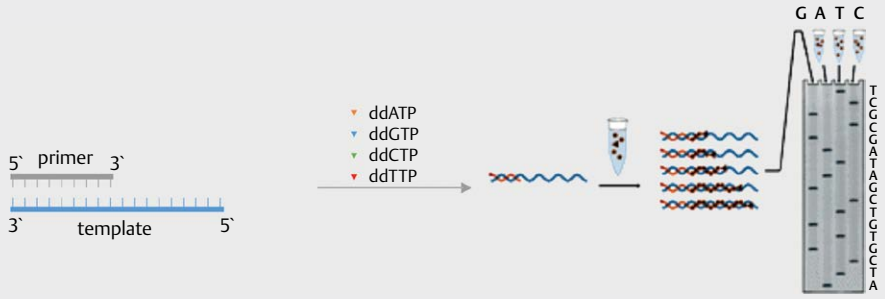
4. Development of High-throughput Sequencing Technologies

What took the Human Genome Project nearly 15 years and $3 billion USD can presently be done in as little as 19.5 h and $1,000 [29, 30]. DNA sequencing technology has existed since the 1970s and has quickly integrated as an essential technology in molecular genetic diagnostics. Sanger sequencing, which was used to sequence the first human genome, is still considered the "gold standard" of se-



▶ **Fig. 2** Friedrich Miescher and the discovery of nucleic acids. **a** Friedrich Miescher (born August 13, 1844, died August 26, 1895) was a Swiss physician. After studying medicine, Miescher searched for a subject without patient interaction due to his hearing loss. Therefore, he decided to devote his career to medical research and went to Tübingen to visit Felix Hoppe-Seyler at the "Cradle of Biochemistry." There, in 1869, he discovered nucleic acid, the basic substance of DNA and RNA. **b** Test tube of salmon sperm nucleic acid, inscribed by Friedrich Miescher and bearing his name (around 1871) **c** Tübingen castle laboratory (German *Schlosslabor*) "Cradle of Biochemistry" in which Felix Hoppe-Seyler discovered haemoglobin and Friedrich Miescher, nucleic acid. (Courtesy of the Museum of the University of Tübingen; MUT).

quencing due to its reliability and accuracy of up to 99.999 % [31]. A modern version of this sequencing method is still used today [17, 32]. The years immediately following the completion of the human genome were marked by the development of commercialized high-throughput sequencers (note: the terms high-throughput sequencing, next generation sequencing, and massively parallel sequencing are used synonymously) (▶ **Fig. 4**). These sequencing technologies have scaled up data output by several orders of magnitude and propelled a massive cost reduction over a relatively short period of time (▶ **Fig. 5**). Since about 2007, the reduction in sequencing cost has substantially outpaced Moore's law for computing costs, which is an observation that every 2 years, computing power tends to double, thus halving the cost. In 1998, the ABI 3730xl (Thermo Fisher Scientific) sequencer generated 84 kilobases of data per run [21], that was then scaled up to 1 gigabase per run with the 2005 debut of the Genome Analyzer (Illumina) system that could sequence 1.3 human genomes per year (Illumina) [33]. This technological leap was further developed and resulted in an



▶ **Fig. 3** Sanger sequencing. A depiction of the modified modern Sanger method. A primer binds to an amplified template and is extended by a single nucleotide. The extension with standard deoxynucleotides (dATP, dGTP, dCTP, an dTTP, not shown) is carried out until the integration of a fluorescently labelled dedeoxynucleotide (ddATP (orange), ddGTP (blue), ddCTP (green), an ddTTP (red)), which breaks the growing DNA chain. After several cycles, the DNA fragments are separated with a gel according to their length and the sequence of the nucleotides is determined according to the sorting of fluorescently-labelled fragments. (Gel image courtesy of Smith RJ).

improvement in sequence performance from 10^2 kilobases per day to 10^{12} kilobases per day [34]. Another notable advancement occurred in 2014 with the emergence of the HiSeqX Ten System (Illumina), that generated 1.8 gigabases per sequencing reaction and broke the $1,000 USD barrier for a human genome. Most recently, in 2017, the NovaSeq 6000 System (Illumina) can generate up to 6 terabases of sequence data in less than two days. Looking to the future, it seems very likely that this sharp reduction in cost will continue, with the sequencing company Illumina aiming to usher in the $100 USD genome over the next 10 years [35].

These sequencing technologies have been accompanied by other developments in digitalisation such as data storage, parallel computing, further developments in CPU architecture, and the invention of the world wide web, which have also contributed to cost reductions. The surge in data output and the sharp decline in cost make these methods widely accessible to individual patients and enable research and clinical laboratories to generate large datasets containing the sequence of hundreds of thousands of individuals. These datasets are crucial for uncovering novel disease associations and supporting the annotation of the complete catalogue of human pathogenic variants. The shift to "big data" in genome research has enormous implications for diagnostics and treatment of patients in all disease areas. Due to its genetic complexity, hearing loss is a particularly interesting and challenging example.

5. The Genetics of Hearing Loss

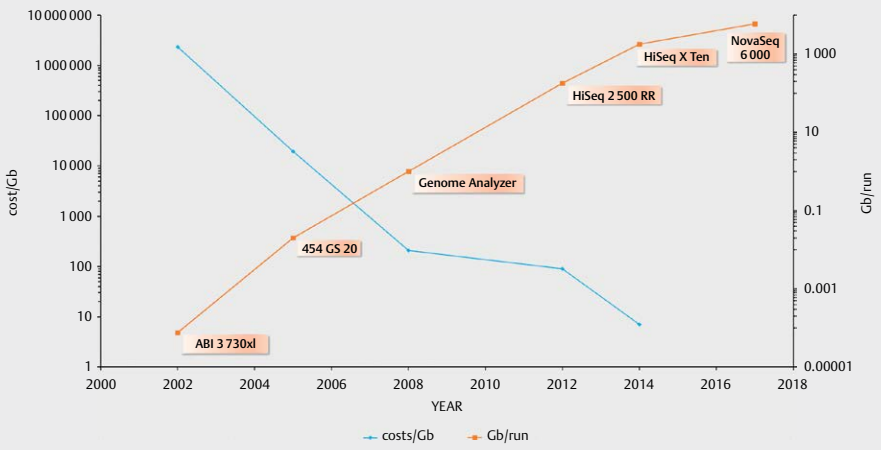
According to the World Health Organization, genetic disorders occur with a prevalence of 10 per 1000 births [36] and affect as many as 1 in 17 individuals throughout life [37]. Hearing loss is the most common congenital sensory disorder affecting 1–2 out of 1000 newborns [38]. More than half of sensory hearing disorders have underlying genetic factors (▶ **Fig. 6**). Hearing loss is predominantly non-syndromic (70 %), but it can also take the form of a syndromic clinical presentation (30 %) [39]. Hearing loss is classically regarded as a Mendelian, or single-gene disorder, that exhibits autosomal recessive (77 %), autosomal dominant (22 %), X-linked (1 %) and mitochondrial (< 1 %) modes of inheritance [39].

Efforts to unravel the molecular genetics of hearing loss have already annotated thousands of variants in the currently recognized genes involved in non-syndromic and syndromic [40] forms of hearing loss (▶ **Fig. 7**). For example, the Deafness Variation Database (DVD) has presently curated over 8,100 pathogenic or likely pathogenic variants in a gene set that includes 152 genes [41, 42]. For comparison, the Human Gene Mutation Database (HGMD) is a comprehensive collection of all known germline variants that are associated with human diseases. This database (HGMD Professional 2018.2) currently contains approximately 225,000 variants with the vast majority of these annotated as pathogenic [43]. The relatively high proportion of pathogenic variants for hearing loss alone underscores the genetic complexity of this sensory disorder.

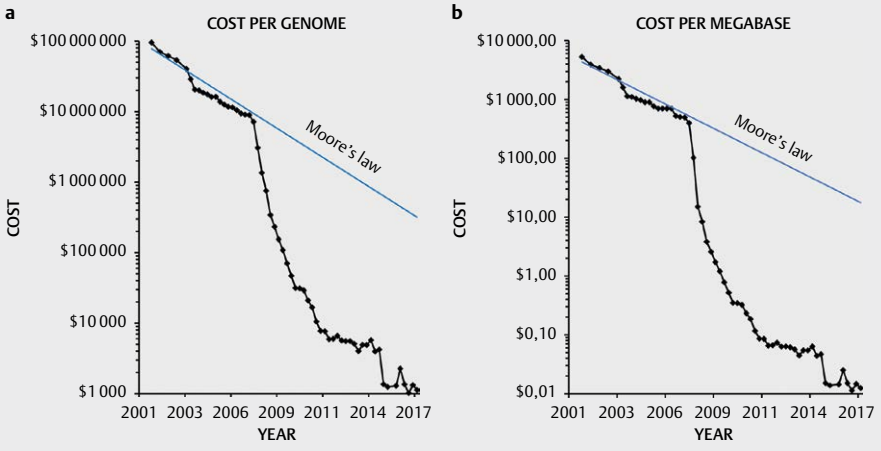
The Hereditary Hearing Loss Homepage presently contains 175 hearing loss associated genes (▶ **Fig. 7a**). These include 161 non-syndromic hearing loss loci, of which 122 genes have been identified. Additionally, non-syndromic hearing loss can be classified according to inheritance to include approximately 70 autosomal recessive, 40 autosomal dominant, and 5 X-linked hearing loss-associated genes, as well as 7 mitochondrial hearing loss-associated variants. Furthermore, there are 53 syndromic-associated hearing loss genes and mitochondrial variants presently documented in this database (▶ **Fig. 7c**). The genetic heterogeneity of hearing loss makes genetic interpretation extremely difficult, not only because of the sheer number of genes involved, but also because each gene can contain a number of pathogenic sequence alterations in the several tens to hundreds range. Significant leaps forward have been made in uncovering this complexity that is much to the credit of the development of big data repositories such as databases and bioinformatics tools, including those that are developed specifically for the genetics of hearing loss.

6. Altering the Diagnostic Paradigm for Hearing Loss

Before the widespread availability of high-throughput sequencing, conventional clinical examinations involved a series of medical tests in order to obtain the most detailed phenotypic picture possible

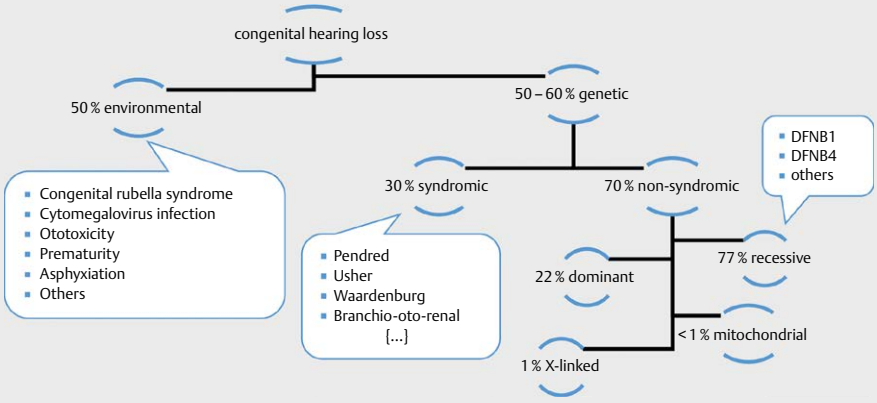


▶ **Fig. 4** Development of sequencing costs over the past decades. Comparison of cost (in US$) and the amount of data generated per year during the continued development of sequencing instruments. As instrument capacity increased in gigabases (Gb) (Gb/run, red, right y-axis), it corresponded to a sharp decrease in sequencing costs over time (cost/Gb, blue, left y-axis). The cost estimate is only available until 2014.

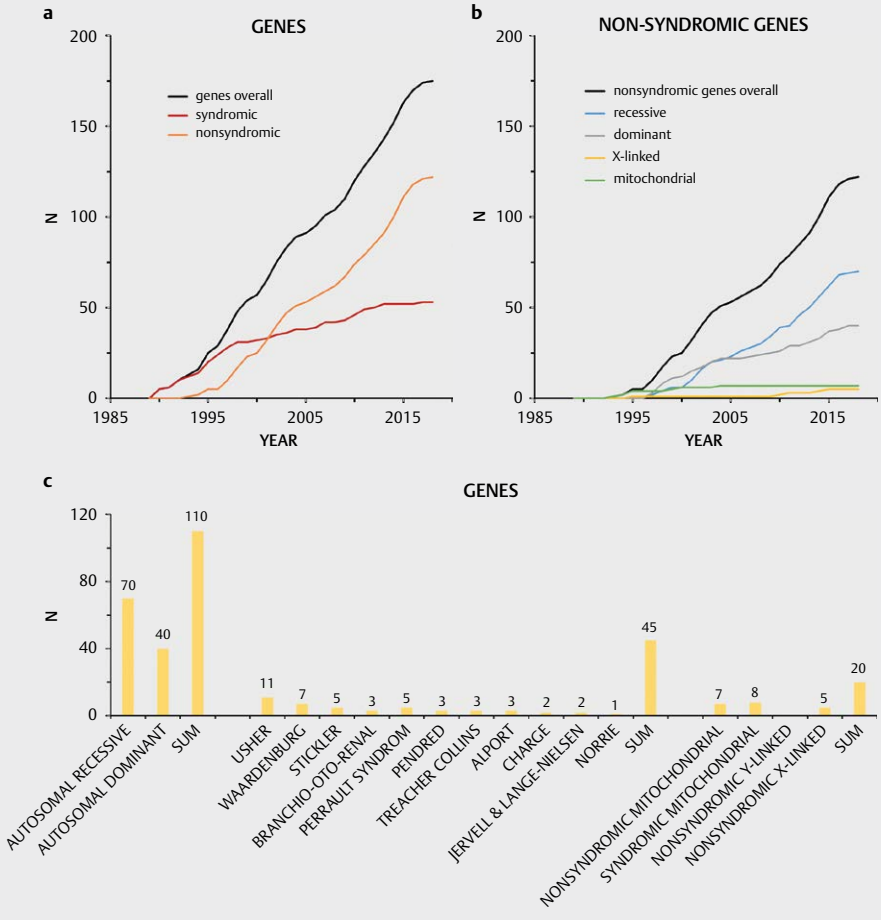


▶ **Fig. 5** Representation of DNA sequencing costs and Moore's law. Moore's law states that costs are reduced every two years by doubling computing power of integrated circuits in computers. **a** Costs per genome and **b** costs per megabase, from 2011 to 2018 follow an almost identical profile. In both analyses, the decline in costs has significantly outpaced Moore's law since 2007.

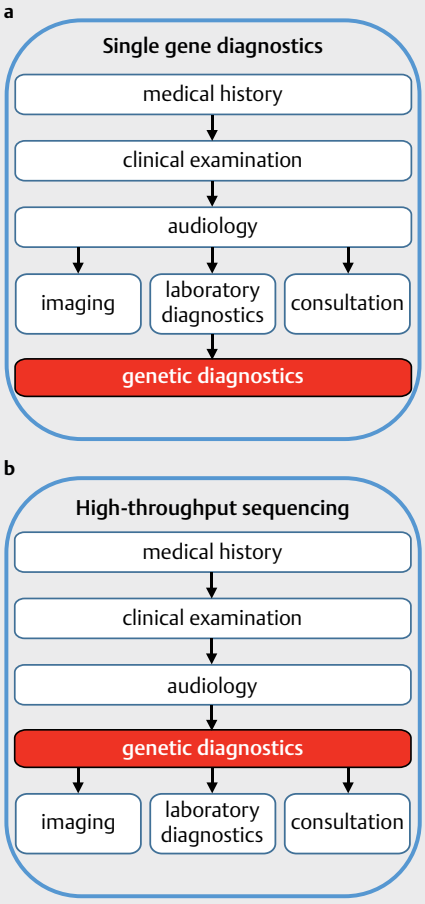
that may be useful to direct molecular genetic analysis (▶ **Fig. 8**) [44, 45]. The most common molecular genetic test to be integrated into clinical care was testing of the connexin 26 encoding gene, *GJB2*, that is primarily associated with an autosomal recessive form of non-syndromic hearing loss. In Germany, this single gene is responsible for the diagnosis of roughly one in five patients with hearing loss [46]. In a separate study, we identified pathogenic variants in *GJB2* that diagnose approximately 17 % of cochlear implant candidates undergoing molecular genetic diagnostic testing for hearing loss [47]. The success of this screening procedure has been supported by the short length of the *GJB2* gene, making it simple to sequence, and the relatively high rate of diagnosis [48]. If the clinical evaluation pointed to a certain form of syndromic hearing loss, an attempt was made to carry out a targeted sequencing of candidate genes on the basis of this clinical suspicion. However, the management of the genetic analysis by phenotypic data could offer only limited success in a genetically heterogeneous and phenotypically variable disease such as hearing loss and was therefore limited to a few genes with a clear genotype-phenotype correlation. These single-gene molecular genetic testing approaches were slow, labour-intensive, expensive, and often yielded uninformative results [49]. Additionally, as most hearing loss is non-syndromic, follow-up beyond exclusionary *GJB2* testing was challenging, as it is nearly impossible to establish a pre-diagnostic hypothesis through clinical examination and audiological findings. The screening of



▶ **Fig. 6** Proportion of environmental and genetic factors in congenital hearing loss. In about half of patients, the hearing loss is due to a genetic cause.



▶ **Fig. 7** Number of non-syndromic and syndromic genes causing hearing loss (Hereditary Hearing Loss Homepage [40]). **a** Identification of 175 genes discovered over time (black), divided into non-syndromic (orange) and syndromic (red) categories. **b** Consideration of the number of genes per year that cause non-syndromic hearing loss (black), divided into inheritance patterns, recessive (blue), dominant (grey) and X-linked (yellow), and mitochondrial (green). **c** An overview of non-syndromic and syndromic genes presented on the "Hereditary Hearing Loss Homepage." Genes involved in autosomal dominant and recessive as well as non-syndromic and syndromic hearing loss are presented individually in each category. N is the number of identified genes.



▶ **Fig. 8a** Classical procedure of hearing loss diagnostics. The process begins with anamnesis, clinical examination, audiological examination, imaging, additional examination (e. g. laboratory diagnostics, thyroid function test, ECG), consultation (e. g. ophthalmology, neurology, cardiology) and ends with a genetic diagnostic test on the basis of individual genes. Genetic causes can usually only be assumed after this procedure by exclusionary diagnostics, but can usually not be proven directly. **b** Chronology of comprehensive hearing loss diagnostics that includes molecular genetic testing. Direct proof of a genetic cause is sought after the anamnesis, clinical examination and audiological examination. The clinical phenotype can provide valuable information for the genetic evaluation of findings. In the case of evidence of a syndromic form of hearing loss, further targeted additional examinations and consultations can then be initiated. Modified from Löwenheim, 2014 [45].

other genes was often prohibitively expensive, slow, and often limited in number.

The past decade has seen a remarkable transition away from single-gene sequencing to high-throughput sequencing approaches for the genetic diagnosis of hereditary disorders [50]. This is particularly important in the case of hearing loss molecular genetic diagnostics, as this technology can be applied to generate previously unimaginable amounts of data to overcome the remarkable genetic heterogeneity in a short amount of time and at a low cost. Current diagnostic approaches utilise either gene panels or exome sequencing. The exome comprises about 1–2 % of the entire genome that encodes proteins. The strategic transition to high-throughput sequencing methods is altering the paradigm of patient management and care.

High-throughput sequencing provides several decisive advantages over single-gene approaches. Many 10's of patients are tested in a single laboratory procedure, simplifying work flows. In particular, all known genes related to hearing loss can be sequenced in a single reaction and analysed in parallel, allowing for a hypothesis-free approach to patient diagnostics. Because certain syndromes, such as Usher syndrome, do not become clinically apparent until after hearing loss onset, it is impossible to accurately diagnose a pre-symptomatic hearing-associated syndrome despite thorough clinical examination. This is one area that can be improved by molecular genetic diagnostic testing. Furthermore, different variants in many genes may lead to multiple clinical outcomes. One such example is the gene *MYO7A* that is responsible for autosomal dominant (DFNA11) or recessive (DFNB2) non-syndromic hearing loss, as well as Usher syndrome (USH1B). It is not always clear about the expected outcomes in patients with homozygous or compound heterozygous variants, especially in very young children, with respect to whether retinitis pigmentosa will develop or not. This has tremendous implications in genetic counselling and downstream medical care. Providing a patient with a pre-symptomatic diagnosis can prevent unnecessary testing, provide useful prognostic value, and also reveal important heritability information [50]. Furthermore, patients who are diagnosed with pathogenic variations in genes that are clinically well-characterized in the literature can benefit from information concerning the possibility of progression and selection of the most beneficial rehabilitation paths.

6.1. Gene panel diagnostics in hearing loss

Gene panels represent a selective and specific approach to molecular genetic diagnostics as they enrich custom gene content for a targeted sequencing approach in a specific disease area. Panel design involves selecting genes based on current knowledge for custom "bait" design. These baits are made out of oligonucleotides that are complementary to targeted regions/exons of interest. "Targeted Genomic Enrichment" or "Sequence Capture" are terms that describe the selection of the desired DNA regions for amplification and enrichment during preparation of a sequencing library. A library contains the complete set of targeted and amplified fragments of interest for sequencing. As gene panels undergo a gene selection and design step and the sequencing data are initially subjected to validation and optimization for quality and uniformity before use in a diagnostic setting, the sequencing coverage (or number of times a sequencing read covers a single base in a gene) across the set of genes has greater uniformity with fewer "gaps" (or bases with poor or no coverage) in sequence coverage. Obtaining high coverage sequencing is important for comprehensive sequence analysis of the variants that may reside in these regions.

6.2. Exome diagnostics in hearing loss

Exome sequencing enriches all presently recognized genes and gene isoforms and is not limited to known genes in a specific disease area. There are many commercially available off-the-shelf exome library preparation kits that are continuously improving.

Many providers also allow users to "spike in" custom bait content to help improve sequencing coverage as desired or to target known variants implicated in human diseases that are not residing in exonic regions and would otherwise be completely missed. While exomes have had a long reputation of delivering sequencing coverage that was of substandard diagnostic quality, this has drastically changed, and now exome-based diagnostics have been successfully integrated into clinical settings for a number of years [51–54]. Sanger sequencing of regions of interest that are poorly covered can complement this method well.

Exome sequence analysis in a diagnostic setting is most efficiently guided by a so-called *in silico* gene panel. Similar to gene panels, the analysis is restricted to clinically relevant genes in order to save time and quickly make a diagnosis. Detailed clinical information about the patient is essential for selecting the most appropriate genes for analysis. This gene selection process enables an analysis that includes all genes that are clinically relevant to the specific phenotype of the patient that meet a certain coverage threshold. This opens up improved possibilities for analysis of patient-tailored gene sets as opposed to gene panels that are always limited to a fixed gene set.

6.3. Advantages and disadvantages of gene panels and exome-based diagnostics

There are several advantages of selecting hearing loss gene panels over exome sequencing. One strong argument is that the data produced by gene panel sequencing are specific to the primary disease. This means that genes related to other disorders will not be sequenced and the analysis and genetic results are restricted to the primary indication. In other words, laboratories that utilise gene panel diagnostics do not have to discuss the potential of secondary findings which are of clinical significance to the patient but unrelated to the primary indication. Expert groups from the American College of Medical Genetics and Genomics have recommended guidelines for reporting secondary findings in a minimum of 59 medically actionable genes be reported in clinical genomic sequencing [55]. The vast majority of these genes involve autosomal dominant conditions that typically involve late onset (adulthood) disorders with only a few having a pediatric onset. In 2013, the German Society of Human Genetics published guidelines for returning secondary findings that emphasized the consenting procedure and the patient's right not to know, or to decline receiving, these results [56]. These guidelines do not specify secondary findings found in a particular set of genes, but define four categories in which a variant may fall. In particular, it is encouraged to report additional findings for which treatments exist. This means that the diagnostic application of exome sequencing may go far beyond the original question of existing hearing loss, for example. For the specialist who initiates a genetic examination, the possible findings then potentially go far beyond his or her own specialist area.

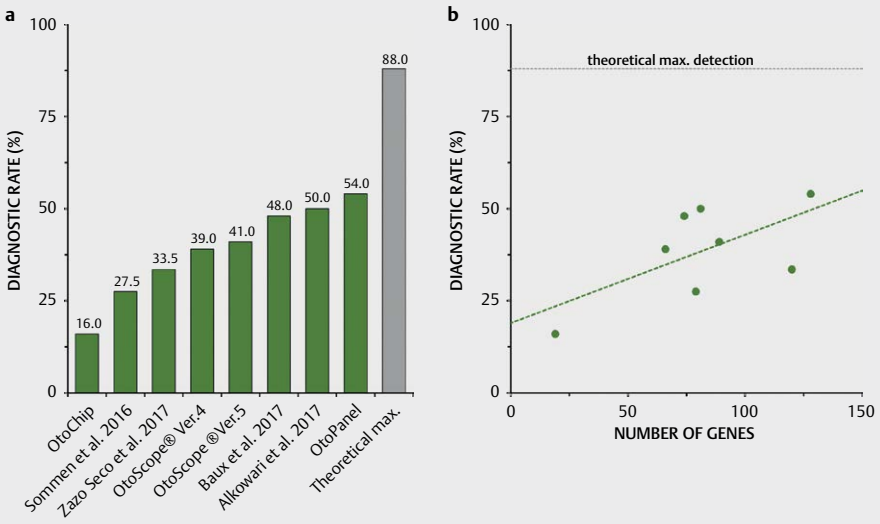
Since the sequencing of predefined gene panels enriches a smaller and disease-specific group of genes than exome sequencing, the coverage of genes is usually much higher, and the specific bait design can target regions that are difficult to sequence (i. e. GC-rich regions, repetitive DNA sequences called tandem repeats, and unevenly fragmented DNA regions). This means that the sensitivity (false-negative rate) and specificity (false-positive rate) of detecting variants can be improved. At least in the case of false positive results from high-throughput sequencing data, these can be validated using Sanger sequencing and are generally not a major issue. The uniform coverage also supports a more consistent copy number variation detection.

Exome sequencing in turn offers some advantages over gene panels. One benefit is the possibility for data re-analysis as new genes are identified, which may be of potential value to patients remaining without a genetic diagnosis after analysis of the known genes. The rapid pace of the field means that novel genes are being identified quickly. Gene panels require periodic updating of gene content and then are subjected to validation. Only after these steps can the DNA from the patient be re-tested, which is more laborious and expensive than the single test that is required to generate an exome dataset. A benefit for laboratories offering a single exome test for molecular genetic diagnostic testing as opposed to several different gene panels is that patients with a variety of different disorders can be tested in parallel, which can significantly decrease the turn-around time for laboratory testing. Depending on the laboratory and number of testing requisitions, laboratories must often wait several weeks or even months before enough DNA from hearing loss patients has been received for a hearing loss gene panel to be prepared and sequenced.

6.4. Diagnostic Rates

While the comprehensive collection of genes involved in hearing loss is presently incomplete [57], molecular genetic diagnostic rates of hearing loss patients in Germany [47, 58, 59] and around the world [60] have experienced a significant improvement since implementing high-throughput sequencing approaches into routine diagnostic care. To illustrate this, we have selected a number of studies reporting diagnostic rates that include *GJB2* (▶ **Fig. 9**). While these studies varied in methodology ranging from gene panels [58, 61–63] to exome sequencing [54], gene content was not always unified, and copy number variation analysis was not consistently performed. However, diagnostic rates ranged from 16 % in a study that included 19 genes on an oligo-hybridization array [64] to 54 % [47] a separate study using a hearing loss gene panel containing 128 genes. Looking closer at the most frequently implicated genes in the various studies (▶ **Table 1**), it can be seen that the five most frequently affected genes differ in part. The exception remains *GJB2*, which is always identified as the most commonly affected gene. The different results can be explained by the different ethnic backgrounds of the patients and the variable number of genes and patients studied.

It is worth emphasizing that a comprehensive molecular genetic diagnosis should include copy number variation analysis [65]. One of the largest studies to date uncovered that roughly 15 % of hearing loss patients carried at least one copy number variation in a hearing loss-associated gene. Moreover, in individuals receiving a diagnosis, 18.7 % had a copy number variation contributing to a diagnosis. In particular, the most prevalent gene implicated in copy number variation is *STRC* (DFNB16), that has a high deletion carrier rate in the European population of approximately 1.6 % that is nearly as high as the well-known c.35delG carrier rate (1.89 %) in the same population [48]. Deletions in *STRC* frequently include a neighbouring gene (*CATSPER2*) encoding a sperm-associated ion



▶ **Fig. 9** Diagnostic rates in selected high-throughput sequencing studies. **a** Overview of the diagnostic rate in patients with hearing loss who undergo a genetic test. **b** There is a positive correlation between the diagnostic rate and the number of genes investigated.

▶ **Table 1** The most commonly affected genes from selected studies.

Autor	Hernandez et al., 2010	Sloan-Hegen et al., 2016	Zazo Seco et al., 2016	Baux et al., 2017	Alkowari et al., 2017	Sommen et al., 2016	Tropitzsch et al., 2013
Country	USA	USA	The Netherlands	France	Qatar	Belgium	Germany
Patients	44	1119	200	207	81	160 families	154
GJB2		1st	1st	1st		1st	1st
MYO15A		5th	3rd	4th		2nd	3rd
SLC26A4		3rd					
MYO7A	1st		4th	3rd		2nd	2nd
CDH23	2nd		4th	5th	1st		
OTOF			5th		3rd		
USH2A	3rd		2nd	4th			
TMC1					2nd	2nd	
MYO6					3rd		5th
TECTA		4th		4th			4th
STRC		2nd	3rd	2nd			
LOXHD1			4th				
TRIOBP			5th				
OTOA				5th	2nd		
GJB6					4th		

channel that causes deafness-infertility syndrome in males. This is an important aspect that should be addressed in pre-diagnostic counselling. *STRC* is an example of a gene that merits analysis for both deletions and pathogenic variants [65–70].

The increasing improvement in molecular genetic diagnostic rates is evidence that high-throughput sequencing should be performed in all patients with hearing loss [49]. Patients without a genetic diagnosis and persistent clinical suspicion of hereditary hearing loss should consider re-testing in subsequent years, since knowledge about genes and variants is still continuing to advance. Taking into account non-synonymous, splice-site, and indel variants (insertions and deletions), as well as copy number variation in the coding regions of a gene panel, a theoretical diagnostic rate of 88 % could be achieved in the future [49] (▶ **Fig. 8b**). Current advances in ge-

nomics assert that non-coding regions of the genome are implicated in hereditary disease [71]. The extent to which changes in the non-coding regions of the genome account for hearing loss remains to be determined. However, disease-associated intronic variants residing more than 20 nucleotides from coding exons and so-called "deep intronic" variants far from exons that would not be captured in standard gene panels and exomes, have already been implicated in a growing number of examples in hearing loss genetics [72, 73]. It is tempting to speculate that a substantial fraction of patients in the future will be diagnosed from variants in non-coding regions that exert effects on gene expression and normal gene splicing.

Molecular genetic diagnostics is dependent on high-quality sequencing, as well as effective bioinformatics analysis strategies that support removal of non-informative variants while keeping important variants for expert analysis. Variant prioritization typically follows the use of a variety of different tools and databases, many of which are described in brief in the next section. It is important to recognize that the number of tools and the data contained within databases is rapidly increasing as sequencing data has become widespread.

7. Computational Resources

Bioinformatics applies computationally intensive methods to process and analyse data to reveal biologically and medically relevant results. Upstream bioinformatics processes involved in high-throughput sequencing data are out of the scope of this article, but involve data pre-processing steps that include alignment of reads, or the fragments of sequencing data, to the human genome reference sequence, as well as post-processing steps that involve removal of duplicate reads and base quality re-calibration. These procedures have also undergone continuous improvement to increase accuracy in variant annotation. Consider that the average exome contains over 20,000 variants, 500 of which are recognized as rare or not yet described in variant frequency databases [74]. The variants that are detected in gene panels can also be extremely rare or remain unclear in interpretation. There are a variety of tools that can be employed to aid in analysis that are summarized in ▶ **Table 2**. The following section describes how these databases and tools, each of which leverages big genomic data resources, are applied to high-throughput sequencing datasets.

Several repositories contain information about human genes, such as GeneCards [75] and the Online Mendelian Inheritance in Man (OMIM) [76] webpage. These resources present summaries about clinical and functional information about the currently characterized genes. With respect to hearing loss, the Hereditary Hearing Loss Homepage [40] lists the loci and genes involved in non-syndromic hearing loss and the most common syndromes with hearing loss as a feature. Many laboratories select gene panel content for custom panel sequencing using these databases.

Variant frequency database repositories have been developed by large networks of international collaborators to present variant frequency information across the exome or genome. Knowing the frequency of a variant can aid interpretation tremendously. For example, if a patient has autosomal dominant hearing loss and a rare variant of interest is present not only in a heterozygous state, but also in a homozygous state in individuals in these databases, then its high frequency in other individuals with presumed normal hearing speaks against pathogenicity in an autosomal dominant disorder. Caution must be used when inferring results, as described in the *GJB2* c.35delG example below, but these are, nonetheless, useful tools for understanding the frequency of a variant, thus providing supporting evidence for or against pathogenicity. However, an essential concept is that just because a variant is not very rare, does not necessarily mean that it is benign or just because a variant is rare or novel does not automatically imply pathogenicity.

One of the first databases developed for documenting genetic variation was the Database of Short Genetic Variants, later abbreviated dbSNP [77], that aims to document all identified genetic variation such as single nucleotide polymorphisms and indels in the genomes of humans and many other species. Other independent databases have developed over the years, such as the exome variant server (EVS), that includes the exome data of 6,500 European American and African American individuals [78]. Even larger databases, such as the Exome Aggregation Consortium Browser (ExAC), which shows variant frequencies from the exome data of 60,706 individuals, which later grew to include 123,136 exomes and 15,496 genomes in an expanded database called the genome aggregation database (gnomAD) [1] were developed that explored variant frequencies in many sub-populations such as Latino, non-Finnish and Finnish European, African, Ashkenazi Jewish, East Asian, South Asian, and "other" individuals not assigned to those populations. During the development of these databases, it became clear that there were many sub-populations that were underrepresented, which triggered the development of a number of other databases, namely the Greater Middle Eastern Variome (GME) that included the exome data from 2,498 individuals from various countries of the Middle East [79], and Iranome that includes the exome data from 800 individuals from eight different ethnic groups in Iran [80]. Continued efforts to capture the genomic variation in rare and isolated human populations will be necessary to understand the unique variants that exist only in these populations. Much can be learned from these rare populations about the human genome and the relationship between variants and human diseases.

One of the most informative strategies to determine the pathogenicity of a variant is through functional validation and experimental testing. However, this is not possible to perform in clinical laboratories that must report genetics results within a restricted timeframe. Molecular biological laboratory specialists have turned to *in silico* pathogenicity prediction tools to analyse the pathogenicity of missense variants. These tools use algorithms that assign variant pathogenicity scores that consider information about evolutionary conservation and the impact of amino acid substitution on protein structure [81]. Since clinical validation is not performed in these programs, specialists usually rely on multiple programs, some of which are presented in ▶ **Table 2**. The programs MutationTaster [82] and PolyPhen-2 [83] assess the effect of amino acid substitutions on protein structure, while SIFT [84, 85] additionally predicts the effects of indel variants on structure. Furthermore, tools such as Combined Annotation Dependent Depletion (CADD) [86] generate weighted single scores from multiple annotations that can be used to quantitatively rank causal variants.

▶ **Table 2** Computational databases and tools commonly used in the interpretation of genetic variants.

Resources for genes and phenotypes				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
GeneCards: The Human Gene Database	Integrative database that contains information on human genes that includes clinical and functional information	✓	✓	[75] https://www.genecards.org/
Hereditary Hearing Loss Homepage	Online resource for genes involved in hereditary hearing loss	✓	✓	[40] http://hereditaryhearingloss.org
Online Mendelian Inheritance in Man (OMIM)	Online resource for human gene and genetic phenotype information	✓	✓	[76] https://www.omim.org/
Allele frequency databases: useful for understanding the frequency of a variant across different ethnicities				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
Database of Short Genetic Variations (dbSNP)	Catalogue of sequence variation	✓	✓	[77] https://www.ncbi.nlm.nih.gov/projects/SNP/
Greater Middle East Variome Project (GME)	Allele frequency reference set from exome sequencing 2,497 individuals from the Middle East	✓	✓	[79] http://igm.ucsd.edu/gme/index.php
Exome Aggregation Consortium (ExAC)	Allele frequency reference set from exome sequencing of 60,706 individuals	✓	✓	[1] http://exac.broadinstitute.org/
Exome Variant Server (EVS)	Allele frequency reference set from exome sequencing of 6,503 individuals	✓	✓	[78] http://evs.gs.washington.edu/EVS/
Genome Aggregation Database (gnomAD)	Allele frequency reference set from individuals that includes 123,136 exomes and 15,496 genomes	✓	✓	[1] http://gnomad.broadinstitute.org/
Iranome	Allele frequency reference set from 800 exomes representing different ethnic groups in Iran	✓	✓	[80] http://www.iranome.com/
***In silico* Pathogenicity Prediction Tools for Variant Analysis**				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
Combined Annotation-Dependent Depletion (CADD)	Integration of many pathogenicity annotations into a single pathogenicity score in the form of C scores to prioritize functional variants	✓	✓	[86] https://cadd.gs.washington.edu/
MutationTaster	Pathogenicity prediction tool for determining the impact of variants on the DNA level	✓	✓	[82] http://www.mutationtaster.org/
PolyPhen-2	Pathogenicity prediction tool for determining the impact of amino acid substitutions on the function of a protein	✓	✓	[83] http://genetics.bwh.harvard.edu/pph2/index.shtml
Sorting Intolerant from Tolerant (SIFT)	Pathogenicity prediction tool for determining the impact of an amino acid substitutions from missense and indel variants on the biological function of a protein	✓	✓	[84, 85] http://sift.jcvi.org http://sift-dna.org/sift4g
Human Splicing Factor	A splicing prediction tool to predict effects of variants on splicing outcomes	✓	✓	[89] http://www.umd.be/HSF3/

▶ **Table 2** Continued...

***In silico* Pathogenicity Prediction Tools for Variant Analysis**				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
GeneSplicer	A splicing prediction tool to predict effects of variants on splicing outcomes	✓	✓	[90] http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml
MaxEntScan	A splicing prediction tool to predict effects of variants on splicing outcomes	✓	✓	[91]
NNSPLICE	A splicing prediction tool to predict effects of variants on splicing outcomes	✓	✓	[92]
Clinically Oriented Databases Aiding with Variant Interpretation				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
ClinVar	Database reporting genetic variants with phenotype associations and supporting evidence	✓	✓	[93] https://www.ncbi.nlm.nih.gov/clinvar/
The Connexin-deafness Homepage	Database of variants for the genes *GJB1*, *GJB2*, *GJB3*, *GJB6*	✓	✓	[109] http://davinci.crg.es/deafness/index.php
Deafness Variation Database (DVD)	Expert-curated catalogue of genetic variation in deafness-associated genes	✓	✓	[41, 42] http://deafnessvariationdatabase.org/
Human Gene Mutation Database (HGMD)	A database that annotates all known variants responsible for human inherited disease	✓	✓	[43] http://www.hgmd.cf.ac.uk/
Leiden Open Variation Database 3.0 (LOVD v.3.0)	Freely accessible database providing a gene-centered collection of DNA variations	✓	✓	[94] https://www.lovd.nl/3.0/home
Evolutionary Conservation Analysis Tools				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
phyloP	Nucleotide evolutionary conservation score	✓	✓	[105]
Grantham distance	A quantification of the physicochemical distance measuring biochemical differences between native and replaced amino acids	✓	✓	[106]
Gene Expression Databases				
Database	**Background**	**Diagnostic tool**	**Research tool**	**Reference/URL**
gEAR Portal	Database showing cell type-specific gene expression based on microarray gene expression and RNAseq data	X	✓	[101] https://gear.igs.umaryland.edu/
Shared Harvard Inner-Ear Laboratory Database (SHIELD)	Mouse and chicken inner ear expression datasets that include RNAseq, ChIP seq, and GeneChip data	X	✓	[100] Shen et al., 2015 https://shield.hms.harvard.edu
Audiogene	A tool to use audiometric data to predict which genes could be affected in patients with autosomal dominant hearing loss	X	✓	[97–99] https://audiogene.eng.uiowa.edu/

Messenger RNA (mRNA) splicing is the process of removing the intronic sequence that does not encode amino acids and splicing together the coding exonic sequence into a single transcript. Genetic variation that disrupts the normal splicing process can substantially influence and contribute to disease by altering gene expression and protein products [87, 88]. Variants that impact proper gene splicing can reside far away from the normal intron-exon sequence boundaries. Understanding the potential impact from this variation is important. Therefore, a number of tools have been developed, such as Human Splicing Factor [89], GeneSplicer [90], MaxEntScan [91], and NNSPLICE [92]. These tools compare the normal and altered sequence for disruption of conserved sequences that are used to guide normal splicing mechanisms.

The interpretation of variants into accurate clinical results can be extremely challenging. Especially in light of the flood of genomics data that are currently cheap and easy to produce, molecular biological laboratory specialists have the task to make sense of many rare variants that represent a mosaic of normal variation and potentially disease-relevant changes. A number of databases such as ClinVar [93], HGMD [43], and the Leiden Open Variation Database (LOVD) [94] document variant interpretations in a clinical context. When variants are contained in these databases, they usually provide an interpretation and link to publication(s) describing this interpretation and clinical information. Many of these databases rely on submitters to share this information or they have a staff of variant curators to do this for data maintenance. There is a risk that the pathogenicity of published variants may not be understood accurately or that the degree of uncertainty may not be communicated correctly. Thus, erroneously included variants can "pollute" these databases with misinformation. While these databases are a very helpful resource, they are also known to contain false-positive results, which can have harmful downstream consequences for patients, lead to inefficient use of resources, and hinder the discovery of true gene and variant associations [95]. Research into this matter uncovered 8.5 % of variants reported in HGMD as disease-associated were also present in a pool of over 1,000 asymptomatic individuals, indicating that these variants were possibly erroneously associated with a disease or penetrance was lower than anticipated [8].

The effects of false variant prioritization also impact gene identification and result in incorrect gene-disease associations. In 2014, the gene *MYO1A* was disqualified as an autosomal dominant nonsyndromic hearing loss gene through the observation of discordant segregation of one missense and two nonsense variants in three different families [96]. In all three families, a molecular genetic diagnosis involving other hearing loss genes was identified that matched the phenotype of the patient, which also highlighted the importance of analysing a comprehensive set of hearing loss-associated genes for a diagnosis. In these families, normal hearing individuals were also detected with suspected *MYO1A* pathogenic variants, arguing against pathogenicity. Falsely associated genes with disease can have major implications on genetic counselling, disease management, and family planning.

Furthermore, clinical laboratories sharing variant and clinical information all use different interpretation criteria, so this information should be carefully considered. The DVD [41] is the only expertly curated database dedicated to the annotation of every variant in every gene associated with hearing loss. As such, rigorous analysis of variants included in this database has re-prioritized previously recognized "pathogenic" variants as "benign" on the basis of frequency of reported variants in multiple populations and considered differences of these variants across multiple populations. The study that gave rise to the DVD found that 93 variants in deafness genes were reclassified from "pathogenic" to "benign", which represented over 4 % of variants identified. This database is also associated with a machine-learning based audiometric profiling tool called AudioGene [97–99] to predict genotypes from audiometric data from autosomal dominant forms of hearing loss. Such tools leverage differences in autosomal dominant audiograms and consider age and progression to prioritize genes based on patient-derived algorithms from studying large datasets of patients with autosomal dominant hearing loss [99]. Several examples from AudioGene are described in the context of phenomics (▶ **Fig. 12**).

Less important in a diagnostic setting, but of substantial importance in research aiming to identify novel genes involved in hearing loss is understanding gene expression in the inner ear. Traditional expression databases contain a variety of tissues, but do not include information about the expression of genes in the inner ear. To overcome this bottleneck, a number of databases have surfaced that specialize in inner ear expression. For example, the Shared Harvard Inner-Ear Laboratory Database (SHIELD) [100] utilises RNA sequencing to provide an overview of the gene expression of four developmental stages (E16, P0, P4, and P7) of the mouse cochlea and utricle. Another database called gEAR Portal [101] contains gene expression information from a variety of mouse developmental stages, as well as zebrafish. The expression pattern in the human inner ear is only available from adults.

8. High-throughput sequencing analysis

For otolaryngologists without practical experience in genetic high-throughput data analysis, it can be a daunting task to understand the procedures involved to obtain potentially useful results. From what we have learned from the controversy surrounding false-positive result reporting from direct-to-consumer genetic testing [102], variant interpretation is highly complex and should be done with a considerable amount of clinical information available to aid with analysis. This next section aims to demystify and simplify the major steps outlining how high-throughput sequencing data are processed and analysed.

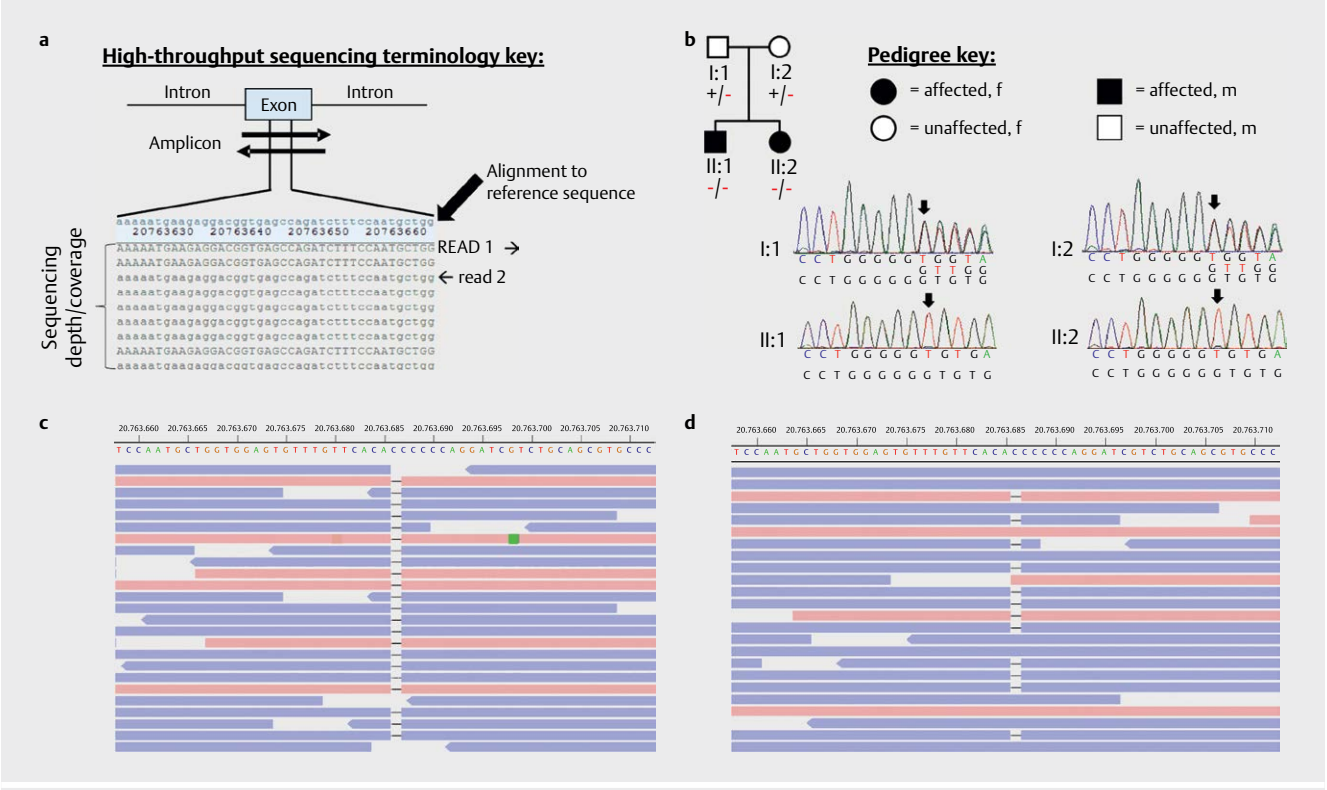
Gene panel, exome, and genome sequencing data are comprised of millions of reads that are contained in a FASTQ file. Each sample has two FASTQ files (read 1, and read 2) representing the bidirectional orientation of sequencing (▶ **Fig. 10**). These files also contain base call and quality information and are used as sequence input for the alignment or mapping to the human reference genome sequence. Alignment organizes the millions of short reads to the correct position of the human reference genome. Visualization of read alignment can show the depth or coverage per base, which is the number of sequencing reads at each base position (▶ **Fig. 10**). Once reads are aligned, variants are called which will then be subjected to what is known as variant "filtering" that sets user-defined parameters to reduce the variants remaining for manual analysis (▶ **Fig. 11**).

Variant filtering can be restricted to a sub-set of genes, for example, those involved in autosomal recessive or autosomal dominant hearing loss genes if the familial inheritance pattern is clear enough to distinguish this. Also of high interest are variants in and adjacent to coding sequence, so a filter is usually applied to remove intronic variants that may not be of initial interest. Despite a number of quality control steps having already been performed in the pre- and post-processing steps, many low quality variants still remain in the data that need to be removed by applying quality cutoff thresholds.

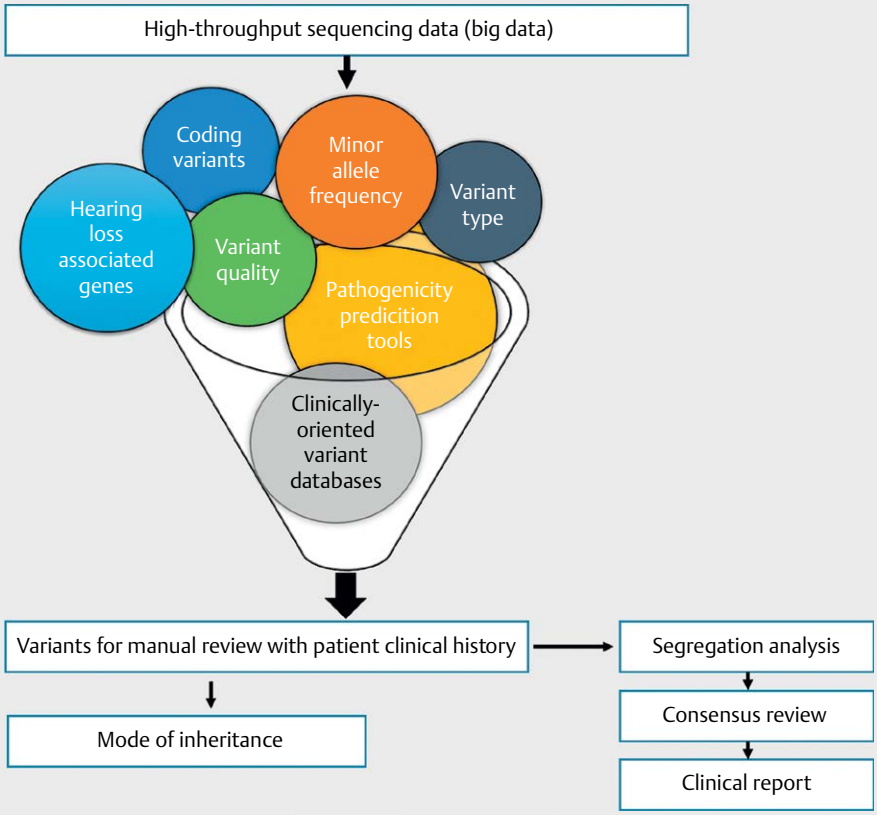
Another important step involves filtering against minor allele frequencies (MAFs). MAFs are calculated as the relative frequency of the less common (minor) allele or variant in the alleles identified in a pool of individuals who have been sequenced. For example, a given population containing 50 individuals identifies one person with a heterozygous variant. Fifty individuals each have two alleles, for a total of 100 alleles. The MAF would be calculated as (1 alternate allele/100 total alleles) for a frequency of 0.01 (1 %) in the individuals tested. Setting optimal MAF thresholds are important for significantly reducing frequent variants that are likely to be benign [41]. Optimal MAF thresholds for hearing loss have been evaluated in large cohorts from multiple laboratories that have enabled expert recommendations. MAF thresholds are recommended as ≤ 0.00007 (0.007 %) for variants in autosomal recessive hearing loss genes and ≤ 0.00002 (0.002 %) for variants in autosomal dominant hearing loss genes [103].

An additional common filtering parameter involves selecting for the type of variant. For example, by selecting for non-synonymous variants (missense, splice, indel, stop gain and stop loss, as well as start gain and start loss), all synonymous variants would be removed, although synonymous variants may be of interest for influencing the splicing landscape, which could profoundly impact protein function. A further filtering step involves analysing each variant using a variety of pathogenicity prediction tools and documenting if the prediction outcomes score each variant as pathogenic or benign. While not every variant in well-studied genes is documented in clinically oriented variant databases, these databases are referenced to identify whether a particular variant has already been interpreted in a patient before.

Nucleotide and amino acid conservation are also considered, as it is thought that variants affecting highly conserved nucleotides that affect highly conserved amino acids are a priori more likely to be damaging [104]. PhyloP scores measure the level of evolutionary conservation in nucleotides by assessing if substitution rates are slower or faster than expected by comparing multiple species [105]. PhyloP scores range from -14 (not conserved) to 6.4 (highly conserved). Grantham distances score the evolutionary distance between two amino acids by considering the biochemical and physical pro-



▶ **Fig. 10** High-throughput sequencing using an example of the *GJB2* c.35delG deletion. **a** A visual representation of some features of high-throughput sequencing. **b** Pedigree of a family with normal hearing parents and two affected children. Females are represented by circles and abbreviated with "f". Males are represented by squares and abbreviated with "m". The symbol + stands for the normal DNA sequence, the symbol - for the deletion. The + /- shows a person who is heterozygous and the -/- shows a person who is homozygous for the c.35delG deletion. Below the pedigree are representative Sanger sequencing images with the heterozygous and homozygous deletion. A visualization of c.35delG with sequencing shows homozygous **c** and heterozygous **d** deletions. The deletion is represented by a gap in the read color sequence. These images were visualized with the Integrative Genomics Viewer of gnomAD [1].



▶ **Fig. 11** Schematic of high-throughput sequencing analysis with high-throughput sequencing data variant filtering subsequent steps.

perties of amino acids and range from 0 to a maximum distance of 215. The more distant two amino acids are, the less "exchangeability" they exhibit, and, therefore, are predicted with a higher probability that the amino acid exchange is pathogenic [106, 107].

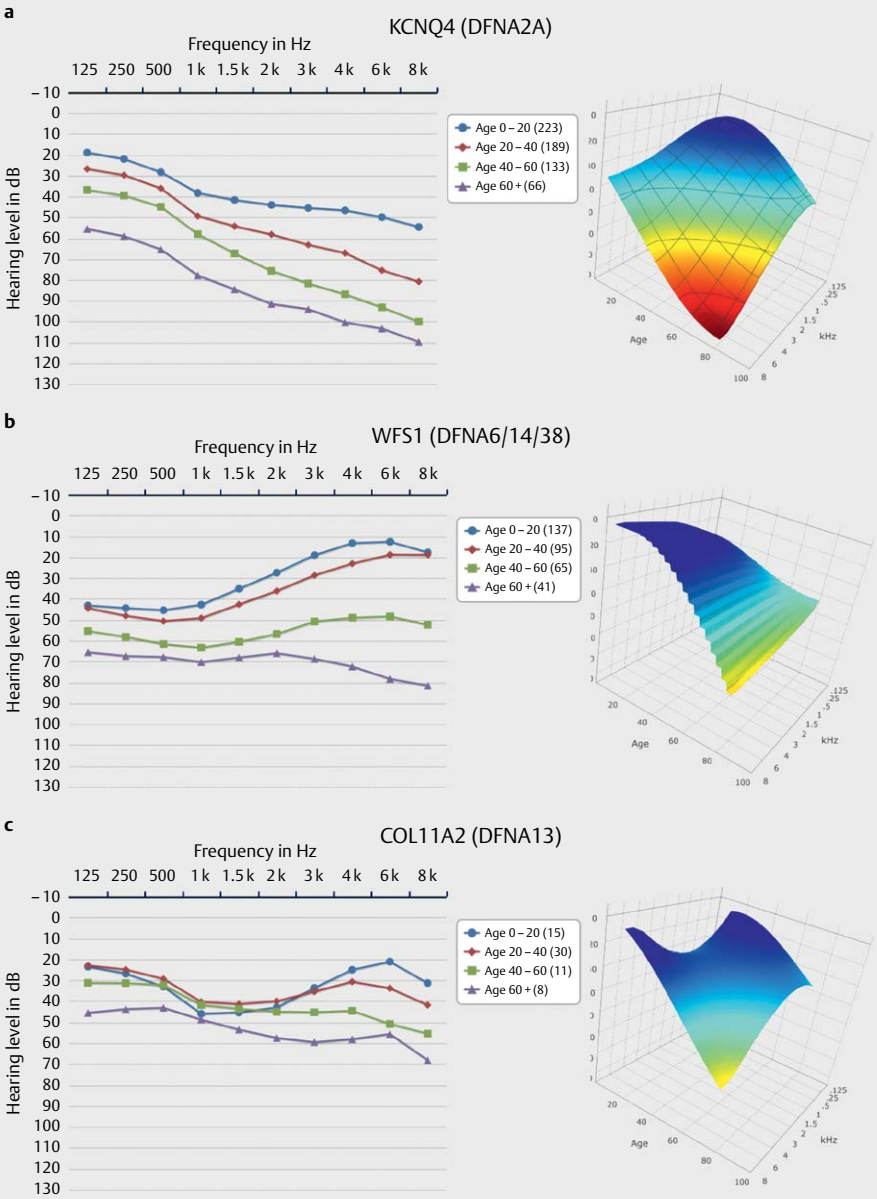
Finally, depending on suspected inheritance, setting pre-defined "allele balance" values (ratio of the number of reads with the variant compared to the number of reads with the reference base) can show variants that appear homozygous or heterozygous. It is expected that a heterozygous variant would be present when approximately 50 % of the total reads show an alternative base, although this allele balance of high quality reads can range broadly. Similarly, a homozygous variant is expected to have 100 % of the reads showing the variant. Sanger sequencing is recommended to validate variants showing allele balances that deviate from accepted cutoffs [49].

Once the variants have been substantially reduced and analysed in the context of the patient's medical history, segregation testing, or testing both unaffected and other additional affected family members is important to avoid erroneous interpretations of variants.

9. An Example of Variant Analysis From *GJB2*

Consider the scenario that a European family with unaffected parents request genetic testing for their two children (▶ **Fig. 10b**), who each report congenital, severe to profound hearing loss. The inheritance pattern in this family appears as autosomal recessive. After genetic testing, it was found that the children have a single nucleotide deletion (c.35delG) in the gene *GJB2* in a homozygous state, while their parents are both heterozygous carriers. This deletion is the single most common cause of hearing loss in Europeans. The high throughput sequencing data show that the affected children are homozygous, with 100 % of their reads showing the deletion (▶ **Fig. 10c**) and the unaffected parents are heterozygous, with roughly half of their reads showing the deletion and the other half with the correct sequence (▶ **Fig. 10d**). *GJB2* gene expression is well studied and is present in supporting cells, hair cells, and in the vestibular and cochlear epithelium throughout several mouse developmental stages. Although this gene and variant are well characterized, it provides a good example about the importance of applying expert guidelines for variant filtering to not only overlook a potentially significant finding, but also to ensure correct variant-disease association.

▶ **Table 3** summarizes the information retrieved from the various resources and tools that are used for variant analysis. *GJB2* encodes the gap junction beta 2 gene that is best known for non-syndromic hearing loss (DFNB1A), but is also associated with autosomal dominant non-syndromic hearing loss (DFNA3A), as well as a number of autosomal dominant syndromes such as Bart-Pumphrey syndrome, hystrix-like ichthyosis-deafness syndrome, keratitis-ichthyo-



▶ **Fig. 12** Two- and three-dimensional audiogram representations created with AudioGene, a program for processing genotypes and audiograms by machine learning. The hearing loss caused by genes **a** *KCNQ4*, **b** *WFS1*, and **c** *COL11A2* manifests with distinctly different audioprofiles Pictures used with permission from Smith RJ [97–99].

sis-deafness syndrome, keratoderma and palmoplantar with deafness, and Vohwinkel syndrome. There are presently over 400 variants in the gene *GJB2* that are annotated in HGMD [43], with the c.35delG being the most commonly implicated variant in non-syndromic hearing loss. This variant has a MAF ranging from 0.002 (0.2 %) to 0.007 (0.7 %), depending on the reference database, and the literature has published carrier rates as high as 0.0189 (1.89 %) in Europeans [48]. If adhering to expert recommendations for filtering for an autosomal recessive disorder, any variant with a MAF ≥ 0.00007 (0.007 %) would be removed from the analysis, which would also remove this important variant. The c.35delG deletion is believed to originate due to a so-called founder effect, wherein the variant originated in a single individual who passed it on to descendants [108]. Many variants involved in hearing loss can be classified as founder mutations and can exhibit particularly high MAFs. These ancient variants that arose many thousands of years ago have been carried through time and space. However, these guidelines also include a list of genes that do not adhere to these MAF cutoff recommendations, and with good reason, *GJB2* is among the list of genes that fit into the exception. Looking closely at the gnomAD variants, there are 1,721 variants reported in total among 275,002 alleles (132,501 individuals) with sequencing data covering this

position. This database also includes ten homozygous individuals who were not supposed to have severe pediatric diseases and whose first-degree relatives were healthy. As hearing loss due to the c.35delG deletion occurs very early in life, it would be expected that these individuals would not be included in this dataset. However, the developers of this database note that some individuals with severe disease may be included in the dataset at a lower frequency than would be would observed in the general population [1] and it is not possible to access clinical histories of these individuals in these large databases to exclude hearing loss. This is an important lesson to bear in mind when using various allele frequency databases, but also opens the possibility for several explanations as to why this occurred. One explanation suggests that it may be possible for individuals to have this deletion and be normal hearing if something called disease penetrance is incomplete, which is unlikely for this particular variant but has been noted for two other variants in *GJB2* (p.Met34Thr and p.Val37Ile) [103]. It could also mean that hearing loss onset in these individuals occurred after time of recruitment, which is not possible to re-trace.

The clinically-oriented databases such as The Connexin-deafness Homepage [109], the DVD [41], HGMD [43], and LOVD v.3.0 [94] unanimously agree that this variant is pathogenic, with one potential caveat. ClinVar showed 27 entries for this deletion. Eleven specified autosomal recessive non-syndromic hearing loss and another 11 entries listed "hearing impairment" or "not provided" meaning the submitters did not provide a condition or mode of inheritance. Three ClinVar submitters state that this variant is involved in autosomal dominant hearing loss (submission accessions: SCV000487402.1, SCV000700274.1, and SCV000536698.1). Included in one of these submissions is mention about several autosomal dominant syndromic forms of hearing loss and autosomal recessive non-syndromic hearing loss (SCV000536698.1) and another submitter listed that this variant is associated with autosomal dominant syndromes. One final entry lists this variant as involved in digenic deafness (*GJB2*|*GJB6*) (SCV000038810.5). According to this information, the possibility of autosomal dominant hearing loss in the carrier parents would also be increased. To the non-expert, the ClinVar entries may add some confusion in interpretation, also raising the possibility for autosomal dominant hearing loss in the carrier parents.

Only two of the four described pathogenicity prediction tools can score this deletion. PolyPhen-2 and SIFT provide predictions about substitutions, not deletions as is the case in this example. The CADD score for this deletion is 24.9, meaning it is roughly in the top 0.5 % of deleterious variations in the human genome. MutationTaster scored this deletion as disease causing. Splicing is not predicted to be significantly impacted by this change. Analysis of conservation on the nucleotide (phyloP) and amino acid (Grantham distance) can only assess substitutions and not deletions, so these are not able to assist with interpretation.

The current literature and clinical reports about the association of the c.35delG deletion in *GJB2* strongly links this variant to hearing loss. Therefore, clinicians can confidently diagnose the children in this example with *GJB2*-associated hearing loss and the parents as carriers, which may be helpful for recurrence calculations if they want to have additional children.

10. From Genome to Phenome

Following the Human Genome Project's extraordinary accomplishment of delivering the human genome reference sequence, many challenges emerged concerning how to effectively apply that knowledge to inherited diseases. Knowing the "anatomy" of the human genome could say nothing directly about the phenotypes encoded in the genotypes. However, as much of the theory and practice of medicine begins with a phenotype, it made sense to introduce the word "*phenome*" shortly after the field had moved beyond the genome [110].

Phenomics captures the natural history of a disease and describes the precise spectrum of disease subclasses, complications and other phenotypic information [111]. On analogous terms, phenomics aims to bring the same centralized well-established, linked, and consolidated strategies for describing the natural history of all phenotypes that genomics already has for annotation, methodologies and standards for the precise description of every genomic element [112]. Effectively implementing phenomics-approaches require novel informatics and data analytic strategies [113]. The development of the Human Phenotype Ontology (HPO) database provides standardized terminology of phenotypic abnormalities to streamline "phenotype-driven" differential diagnostics [114]. The HPO database presently has over 13,000 terms and over 156,000 annotations for hereditary diseases and has proven to be a powerful tool for enhancing exome and genome analysis. For example, by integrating HPO terminology that streamlined "deep phenotyping" of patients, the NIH Undiagnosed Disease Program and Undiagnosed Diseases Network were able to improve molecular diagnosis that entailed the re-analysis of exome sequencing data of previously "undiagnosable" patients, effectively resolving an additional 10 % to 20 % of patients [115, 116]. The HPO database presently contains over 1,600 disease results containing the word "hearing loss." Specific terminology could quickly narrow the list of genetic diseases involving hearing loss from 1,600 to several dozen. Knowing the complete phenotype in streamlined terminology can greatly aid genomics analysis. HPO terms are being integrated into high-throughput sequencing bioinformatics pipelines to greatly increase the speed of analysis in patients with pathogenic variants in genes that have already been identified and characterized.

Hearing loss has the particular challenge of a pronounced clinical heterogeneity, even among individuals from the same family segregating the same variant, that can cloud the precise characterization of hearing loss. This can be reflected by the fact that some genes show extreme clinical heterogeneity and complicate phenotypic correlation. However, by studying many patients with hearing loss due to pathogenic variants in the same gene, researchers have identified several genes that exhibit robust associations. This has been explored in autosomal dominant hearing loss and inspired the development of a tool called AudioGene [97, 99]. This machine-learning based program analyses patient audioprofiles via a computational clustering algorithm and prioritizes the most likely autosomal dominant hearing loss genes for mutational screening. In an experiment comparing the predictive performance of AudioGene and a panel of experts in listing the top likely autosomal dominant genes that could be involved in patients with audiogram data available, Audiogene outperformed expert gene prediction

▶ **Table 3** A variant interpretation example of the *GJB2* c.35delG homozygous deletion.

Database	Information
GJB2 information	
GeneCards	Gap junction protein beta 2; associated with Vohwinkel syndrome and keratoderma, palmoplantar with deafness, as well as autosomal dominant (DFNA3A) and autosomal recessive (DFNB1A) hearing loss
Hereditary Hearing Loss Homepage	DFNA3A, DFNB1A
OMIM	Gap junction protein beta-2; involved in Bart-Pumphrey syndrome, autosomal dominant (DFNA3A) and autosomal recessive (DFNB1A) non-syndromic hearing loss, hystrix-like ichthyosis-deafness syndrome, keratitis-ichthyosis-deafness syndrome, keratoderma and palmoplantar with deafness, and Vohwinkel syndrome
Allele frequency analysis of the c.35delG variant	
dbSNP	MAF = 0.002; Clinical significance: pathogenic
GME	Total allele (variant) count: 5, no homozygous individuals in 1,984 alleles (992 individuals); MAF = 0.00252
ExAC	Total allele (variant) count: 733 including 3 homozygous individuals in 121,352 alleles (60,676 individuals); MAF = 0.00604
EVS	Total allele (variant) count: 93, no homozygous individuals in 12,425 alleles (6,212 individuals); MAF = 0.00748
gnomAD	Total allele (variant) count: 1,721 including 10 homozygous individuals in 275,002 alleles (135,501 individuals); MAF = 0.006258
Iranome	Total allele (variant) count: 3, no homozygous individuals in 1,600 alleles; MAF = 0.001875
Clinically Oriented Databases Aiding with Variant Interpretation	
ClinVar	Clinical significance: Pathogenic from 26 submitters, no conflicts of variant pathogenicity interpretation Conditions: Deafness, autosomal recessive 1A, mutilating keratoderma, hystrix-like ichthyosis with deafness, autosomal dominant keratitis-ichthyosis-deafness syndrome, keratoderma palmoplantar deafness, knuckle pads, deafness and leukonychia syndrome, deafness, autosomal dominant 3a, digenic *GJB2*/*GJB6* deafness, non-syndromic hearing loss and deafness, hearing impairment, bilateral sensorineural hearing impairment, bilateral conductive hearing impairment, severe sensorineural hearing impairment, non-syndromic hearing loss, recessive, deafness
The Connexin-deafness Homepage	Autosomal recessive non-syndromic deafness
DVD	Pathogenic, autosomal recessive non-syndromic hearing loss
HGMD	Deafness, autosomal recessive 1
LOVD v.3.0	Pathogenic
In silico Pathogenicity Prediction Tools for Variant Analysis	
CADD	Score: 24.9
MutationTaster	Disease causing
PolyPhen-2	No score listed
SIFT	No score listed
Splicing Prediction Tools	
Human Splicing Factor	No significant splicing effect predicted
GeneSplicer	No significant splicing effect predicted
MaxEntScan	No significant splicing effect predicted
NNSPLICE	No significant splicing effect predicted
Evolutionary Conservation Analysis	
phyloP	No score listed
Grantham distance	No score listed
Gene Expression Databases	
gEAR Portal	Expressed in P0 mouse hair cells, P1 hair cells, supporting cells, and non-sensory cells, E16.5 and P0 mouse cochlear and vestibular sensory epithelium
SHIELD	FACS-sorted hair cells and ganglion cells: expressed in Utricle and cochlea in embryonic and postnatal stages. (E12, E13, E16, P0, P6, and P15)
Audiometric profiling tool	
AudioGene	Gene not included

The full position of the *GJB2* c.35delG deletion is Chr13(GRCh37):g.20763686, NM_004004.5:c.35del, p.Gly12Valfs * 2

by 33 % [97]. Further development of AudioGene now adds a third dimension, age, to the audioprofile [99]. This additional feature is of clinical significance to autosomal dominant hearing loss, as most forms of dominant hearing loss are progressive. Age is easily visualized by colour on a three-dimensional surface.

The three examples shown in ▶ **Fig. 12** depict the two- and three-dimensional visual representations for the genes *KCNQ4* (DFNA2A), *WFS1* (DFNA6/14/38), and *COL11A2* (DFNA13). The AudioGene profiles in two- and three-dimensional forms of the gene *KCNQ4* show a characteristically progressive, high frequency hearing loss (▶ **Fig. 12a**). Comparing this to the audioprofiles for the genes *WFS1*, with progressive, low-frequency hearing loss (▶ **Fig. 12b**), and *COL11A2* with rather stable mid-frequency and progressive high-frequency hearing loss (▶ **Fig. 12c**), one can imagine how powerful this can be for predicting underlying genetic factors which is helpful in a genetically heterogeneous disorder like hearing loss. This tool provides a diagnostic strategy to support accurate genetic testing and is an example of the pairing of "big audiometric data" with genetics.

Going beyond an understanding of hearing loss on a gene-level, the field would benefit from a phenotype database that correlates the phenotype with underlying variants, for example, the effects of patients with combinations of truncating and non-truncating variants [117]. We now know that different or even the same variant in a single gene may lead to a completely different hearing loss course. Knowledge of variant-oriented hearing loss characteristics is important to establish a baseline understanding of hearing loss, improve genetic counselling, and prospects for future gene- and variant-based therapies.

11. The Outlook of High-throughput Sequencing

Big genomics data has propelled the field into a once unimaginable state. It is important to recognize that sequencing holds great potential to unlock important medical diagnoses that can significantly impact patient care and support patient-tailored medicine. It is also equally important to understand that the field is currently in a setting of great advancements and not every variant in the genome is currently known or understood. A resonating statement from Cynthia C. Morton's 2014 American Society of Human Genetics Presidential Address emphasized that in the current state of genetics, "we find ourselves building the plane as we are flying it" [118]. It is easy to interpret the uncertainties of the field and regard those uncertainties as a sign that the field has little to offer, but the truth is that genomics is likely to play an ever increasing role in patient care, with the hope that one day we will be able to diagnose nearly all patients, even those with ultra-rare genetic disorders.

It is difficult to foresee whether one day every newborn will undergo some form of genetic screening at birth, ranging from targeted gene panels or even genome sequencing in an effort to replace, enhance, or reduce false positives that may be encountered in newborn metabolic and hearing screening to identify high-risk babies before the symptoms are clear. Genomics advocates see great potential of this technology setting the stage for a lifetime of personalized medical care. This could offer additional information in individuals at risk for certain conditions. Rigorous research into the medical and ethical implications of this will hopefully signal the most beneficial paths while respecting the wishes and rights of patients.

It is clear that the genetic landscape of hearing loss has not been fully characterized and for every "known" there seems to be a long list of "unknowns." However, for patients achieving a molecular genetic diagnosis, this information is valuable and is the product of the mass merging of multidisciplinary big data efforts. For patients without a diagnosis, it is important to not give up and to continue exploring genetic testing in the future. With further developments in big genomics data and as more genes are identified and characterized, it may be possible that receiving a genetic diagnosis may one day be the norm.

12. Conclusions for Clinical Practice

A genetic examination in the form of a genetic diagnosis should be made after anamnesis, physical examination and audiological examination for the diagnosis of hearing loss and also considered in families with only one affected individual. Genetic diagnosis can avoid subsequent diagnostic procedures that may be invasive and allows the patient and family to be advised on therapy options and family planning. These form the basis for the development of personalized medicine and, in the future, of possibly customized pharmacotherapy or individual molecular therapy.

13. Acknowledgements

The authors would like to thank Nora Knoblich for support with figure and table preparation. We thank Dr. R. J. E. Pennings (Radboud University Medical Center, Nijmegen, The Netherlands) and Prof. R. J. Smith (University of Iowa, USA) for providing helpful comments on the manuscript. We thank Prof. R. J. Smith (University of Iowa, USA) and Prof. E. Seidel (University of Tübingen, Germany) for providing data for ▶ **Fig. 2, 3**, and **12**.

Conflict of Interest

The author states that there is no conflict of interest.

References

[1] Lek M, Karczewski KJ, Minikel EV et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016; 536: 285–291

[2] Zarrei M, MacDonald JR, Merico D et al. A copy number variation map of the human genome. Nat Rev Genet 2015; 16: 172–183

[3] Xue Y, Chen Y, Ayub Q et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 2012; 91: 1022–1032

[4] Sudmant PH, Rausch T, Gardner EJ et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015; 526: 75–81

[5] Narasimhan VM, Hunt KA, Mason D et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science 2016; 352: 474–477

[6] Narasimhan VM, Xue Y, Tyler-Smith C. Human knockout carriers: Dead, diseased, healthy, or improved? Trends Mol Med 2016; 22: 341–351

[7] Consortium UK, Walter K, Min JL et al. The UK10K project identifies rare variants in health and disease. Nature 2015; 526: 82–90

[8] Cassa CA, Tong MY, Jordan DM. large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. Hum Mutat 2013; 34: 1216–1220

[9] Sulem P, Helgason H, Oddson A et al. Identification of a large set of rare complete human knockouts. Nat Genet 2015; 47: 448–452

[10] Dahm R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet 2008; 122: 565–581

[11] Maderspacher F. Rags before the riches: Friedrich Miescher and the discovery of DNA. Current Biology 2004; 14: R608–R608

[12] Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. J Exp Med 1944; 79: 137–158

[13] Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953; 171: 737–738

[14] Zallen DT. Despite Franklin's work, Wilkins earned his Nobel. Nature 2003; 425: 15

[15] Efstratiadis A, Kafatos FC, Maniatis T. The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. Cell 1977; 10: 571–585

[16] Shendure J, Balasubramanian S, Church GM et al. DNA sequencing at 40: past, present and future. Nature 2017; 550: 345–353

[17] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977; 74: 5463–5467

[18] Saiki RK, Scharf S, Faloona F et al. Enzymatic Amplification of Beta-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle-Cell Anemia. Science 1985; 230: 1350–1354

[19] Rose EA. Applications of the polymerase chain reaction to genome analysis. FASEB J 1991; 5: 46–54

[20] Dickman S. West-Germany Voices Objections to European Genome Project. Nature 1988; 336: 416–416

[21] Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from large-scale biology. Science 2003; 300: 286–290

[22] Murray JC, Buetow KH, Weber JL et al. A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). Science 1994; 265: 2049–2054

[23] Fleischmann RD, Adams MD, White O et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995; 269: 496–512

[24] Dunham I, Shimizu N, Roe BA et al. The DNA sequence of human chromosome 22. Nature 1999; 402: 489–495

[25] Adams MD, Celniker SE, Holt RA et al. The genome sequence of Drosophila melanogaster. Science 2000; 287: 2185–2195

[26] Bier E. Drosophila, the golden bug, emerges as a tool for human genetics. Nat Rev Genet 2005; 6: 9–23

[27] Shendure J, Porreca GJ, Reppas NB et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005; 309: 1728–1732

[28] [Anonymous]. Method of the year. Nat Methods 2008; 5: 1

[29] Sevilla G. New GUINNESS WORLD RECORDS™ Title Set for Fastest Genetic Diagnosis. In Rady Children's Hospital San Diego 2018

[30] Vence T. $1,000 Genome at Last? In, The Scientist 2014

[31] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008; 26: 1135–1145

[32] Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci, USA 1977; 74: 560–564

[33] Illumina. An Introduction to Next-Generation Sequencing Technology. In: Illumina, Inc 2017

[34] Mardis ER. A decade's perspective on DNA sequencing technology. Nature 2011; 470: 198–203

[35] [Anonymous]. Illumina Introduces the NovaSeq Series—a New Architecture Designed to Usher in the $100 Genome. In: Business Wire 2017

[36] [Anonymous]. Genes and Human Disease. In: World Health Organization 2018

[37] Davies SC. Annual Report of the Chief Medical Officer 2016. In, Generation Genome 2017

[38] Morton CC, Nance WE. Newborn hearing screening–a silent revolution. N Engl J Med 2006; 354: 2151–2164

[39] Morton NE, Shields DC, Collins A. Genetic epidemiology of complex phenotypes. Ann Hum Genet 1991; 55: 301–314

[40] Van Camp G, Smith RJH. Hereditary Hearing Loss Homepage. In Shearer AE, Sommen M. (eds) 2018

[41] Shearer AE, Eppsteiner RW, Booth KT et al. Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am J Hum Genet 2014; 95: 445–453

[42] Azaiez H, Booth KT, Ephraim SS et al. Genomic Landscape and Mutational Signatures of Deafness-Associated Genes. Am J Hum Genet 2018; 103: 484–497

[43] Stenson PD, Mort M, Ball EV et al. The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017; 136: 665–677

[44] Bitner-Glindzicz M. Hereditary deafness and phenotyping in humans. Br Med Bull 2002; 63: 73–94

[45] Löwenheim H. [Zukunft der Hördiagnostik] [Article in German]. Zeitschrift für Audiologie Audiological Acoustics 2014; 20: 62–65

[46] Bartsch O, Vatter A, Zechner U et al. GJB2 mutations and genotype-phenotype correlation in 335 patients from germany with nonsyndromic sensorineural hearing loss: evidence for additional recessive mutations not detected by current methods. Audiol Neurootol 2010; 15: 375–382

[47] Tropitzsch A, Friese N, Michels L et al. Next-generation Sequencing in der Diagnostik der genetischen Schwerhörigkeit. 30 Wissenschaftliche Jahrestagung der Deutschen Gesellschaft für Phoniatrie und Pädaudiometrie. 2013Bochum, Germany:

[48] Mahdieh N, Rabbani B. Statistical study of 35delG mutation of GJB2 gene: A meta-analysis of carrier frequency. Int J Audiol 2009; 48: 363–370

[49] Shearer AE, Black-Ziegelbein EA, Hildebrand MS et al. Advancing genetic testing for deafness with genomic technology. J Med Genet 2013; 50: 627–634

[50] Shearer AE, Smith RJ. massively parallel sequencing for genetic diagnosis of hearing loss: The New Standard of Care. Otolaryngol Head Neck Surg 2015; 153: 175–182

[51] Sie AS, Prins JB, van Zelst-Stams WA et al. Patient experiences with gene panels based on exome sequencing in clinical diagnostics: high acceptance and low distress. Clin Genet 2015; 87: 319–326

[52] Sheppard S, Biswas S, Li MH et al. Utility and limitations of exome sequencing as a genetic diagnostic tool for children with hearing loss. Genet Med 2018 [Epub ahead of print]

[53] Guan Q, Balciuniene J, Cao K et al. AUDIOME: A tiered exome sequencing-based comprehensive gene panel for the diagnosis of heterogeneous nonsyndromic sensorineural hearing loss. Genet Med 2018 [Epub ahead of print]

[54] Zazo Seco C, Wesdorp M, Feenstra I et al. The diagnostic yield of whole-exome sequencing targeting a gene panel for hearing impairment in The Netherlands. Eur J Hum Genet 2017; 25: 308–314

[55] Kalia SS, Adelman K, Bale SJ et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. Genet Med 2017; 19: 249–255

[56] [Anonymous]. [Opinion of the German Society for Human Genetics on additional genetic findings in diagnostics and Research] [Article in German]. In, Guidelines and Statements of the German Society for Human Genetics 2013

[57] Bowl MR, Simon MM, Ingham NJ et al. A large scale hearing loss screen reveals an extensive unexplored genetic landscape for auditory dysfunction. Nat Commun 2017; 8: 886

[58] Sommen M, Schrauwen I, Vandeweyer G et al. DNA diagnostics of hereditary hearing loss: A targeted resequencing approach combined with a mutation classification system. Hum Mutat 2016; 37: 812–819

[59] Vona B, Muller T, Nanda I et al. Targeted next-generation sequencing of deafness genes in hearing-impaired individuals uncovers informative mutations. Genet Med 2014; 16: 945–953

[60] Yan D, Tekin D, Bademci G et al. Spectrum of DNA variants for non-syndromic deafness in a large cohort from multiple continents. Hum Genet 2016; 135: 953–961

[61] Baux D, Vache C, Blanchet C et al. Combined genetic approaches yield a 48 % diagnostic rate in a large cohort of French hearing-impaired patients. Sci Rep 2017; 7: 16783

[62] Sloan-Heggen CM, Bierer AO, Shearer AE et al. Comprehensive genetic testing in the clinical evaluation of 1119 patients with hearing loss. Hum Genet 2016; 135: 441–450

[63] Alkowari MK, Vozzi D, Bhagat S et al. Targeted sequencing identifies novel variants involved in autosomal recessive hereditary hearing loss in Qatari families. Mutat Res 2017; 800-802: 29–36

[64] Hernandez AL, Cox S, Kothiyal P et al. The Otochip sequencing array for hearing loss and Usher syndrome. International Symposium on Usher Syndrome and Related Diseases. 2010 Valencia, Spain

[65] Shearer AE, Kolbe DL, Azaiez H et al. Copy number variants are a common cause of non-syndromic hearing loss. Genome Med 2014; 6: 37

[66] Moteki H, Azaiez H, Sloan-Heggen CM et al. Detection and Confirmation of Deafness-Causing Copy Number Variations in the STRC Gene by Massively Parallel Sequencing and Comparative Genomic Hybridization. Ann Otol Rhinol Laryngol 2016; 125: 918–923

[67] Vona B, Hofrichter MAH, Neuner C et al. DFNB16 is a frequent cause of congenital hearing impairment: implementation of STRC mutation analysis in routine diagnostics. Clinical Genetics 2015; 87: 49–55

[68] Plevova P, Paprskarova M, Tvrda P et al. STRC Deletion is a Frequent Cause of Slight to Moderate Congenital Hearing Impairment in the Czech Republic. Otology & Neurotology 2017; 38: E393–E400

[69] Francey LJ, Conlin LK, Kadesch HE et al. Genome-wide SNP genotyping identifies the Stereocilin (STRC) gene as a major contributor to pediatric bilateral sensorineural hearing impairment. American Journal of Medical Genetics Part A 2012; 158a: 298–308

[70] Amr SS, Murphy E, Duffy E et al. Allele-Specific Droplet Digital PCR Combined with a Next-Generation Sequencing-Based Algorithm for Diagnostic Copy Number Analysis in Genes with High Homology: Proof of Concept Using Stereocilin. Clinical Chemistry 2018; 64: 705–714

[71] Ren C, Liu F, Ouyang ZY et al. Functional annotation of structural ncRNAs within enhancer RNAs in the human genome: implications for human disease. Scientific Reports 2017; 7: 15518

[72] Nakano Y, Kelly MC, Rehman AU et al. Defects in the Alternative Splicing-Dependent Regulation of REST Cause Deafness. Cell 2018; 174: 536–548 e521

[73] Khan AO, Becirovic E, Betz C et al. A deep intronic CLRN1 (USH3A) founder mutation generates an aberrant exon and underlies severe Usher syndrome on the Arabian Peninsula. Sci Rep 2017; 7: 1411

[74] Frebourg T. The challenge for the next generation of medical geneticists. Hum Mutat 2014; 35: 909–911

[75] Stelzer G, Rosen N, Plaschkes I et al. The GeneCards Suite: From gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 2016; 54: 1.30.31–31.30.33

[76] [Anonymous]. Online Mendelian Inheritance in Man, OMIM®. In Baltimore MD.USA: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; 2018

[77] Sherry ST, Ward MH, Kholodov M et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29: 308–311

[78] [Anonymous]. NHLBI Exome Sequencing Project (ESP) Exome Variant Server. In 2018

[79] Scott EM, Halees A, Itan Y et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet 2016; 48: 1071–1076

[80] Akbari MR, Fattahi Z, Beheshtian M et al. Iranome: A human genome variation database of eight major ethnic groups that live in Iran and neighboring countries in the Middle East. 67th Annual Meeting of The American Society of Human Genetics, 2017. 2017 Orlando, FL, USA

[81] Walters-Sen LC, Hashimoto S, Thrush DL et al. Variability in pathogenicity prediction programs: Impact on clinical diagnostics. Mol Genet Genomic Med 2015; 3: 99–110

[82] Schwarz JM, Cooper DN, Schuelke M et al. MutationTaster2: Mutation prediction for the deep-sequencing age. Nat Methods 2014; 11: 361–362

[83] Adzhubei IA, Schmidt S, Peshkin L et al. A method and server for predicting damaging missense mutations. Nat Methods 2010; 7: 248–249

[84] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 2009; 4: 1073–1082

[85] Vaser R, Adusumalli S, Leng SN et al. SIFT missense predictions for genomes. Nat Protoc 2016; 11: 1–9

[86] Kircher M, Witten DM, Jain P et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014; 46: 310–315

[87] Xiong HY, Alipanahi B, Lee LJ et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 2015; 9: 1254806

[88] Majoros WH, Holt C, Campbell MS et al. Predicting Gene Structure Changes Resulting from Genetic Variants via Exon Definition Features. LID - . doi:10.1093/bioinformatics/bty324 Bioinformatics 2018 [Epub ahead of print]

[89] Desmet FO, Hamroun D, Lalande M et al. Human Splicing Finder: An online bioinformatics tool to predict splicing signals. Nucleic Acids Res 2009; 37: e67

[90] Pertea M, Lin X, Salzberg SL. GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res 2001; 29: 1185–1190

[91] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 2004; 11: 377–394

[92] Reese MG, Eeckman FH, Kulp D et al. Improved splice site detection in Genie. J Comput Biol 1997; 4: 311–323

[93] Landrum MJ, Lee JM, Benson M et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016; 44: D862–D868

[94] Fokkema IF, Taschner PE, Schaafsma GC et al. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 2011; 32: 557–563

[95] Akle S, Chun S, Jordan DM et al. Mitigating false-positive associations in rare disease gene discovery. Hum Mutat 2015; 36: 998–1003

[96] Eisenberger T, Di Donato N, Baig SM et al. Targeted and Genomewide NGS Data Disqualify Mutations in MYO1A, the "DFNA48 Gene", as a Cause of Deafness. Hum Mutat 2014; 35: 565–570

[97] Hildebrand MS, DeLuca AP, Taylor KR et al. A contemporary review of AudioGene audioprofiling: a machine-based candidate gene prediction tool for autosomal dominant nonsyndromic hearing loss. Laryngoscope 2009; 119: 2211–2215

[98] Taylor KR, Booth KT, Azaiez H et al. Audioprofile Surfaces: The 21st Century Audiogram. Ann Otol Rhinol Laryngol 2016; 125: 361–368

[99] Taylor KR, Deluca AP, Shearer AE et al. AudioGene: Predicting hearing loss genotypes from phenotypes to guide genetic screening. Hum Mutat 2013; 34: 539–545

[100] Shen J, Scheffer DI, Kwan KY et al. SHIELD: an integrative gene expression database for inner ear research. Database (Oxford) 2015; 2015; bav071

[101] Hertzano R, Orvis J. gEAR Portal In 2018

[102] Tandy-Connor S, Guiltinan J, Krempely K et al. False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. LID - . doi:10.1038/gim.2018.38 Genet Med 2018 [Epub ahead of print]

[103] Oza A, DiStefano M, Hemphill S et al. Expert Specification of the ACMG/AMP Variant Interpretation Guidelines for Genetic Hearing Loss. bioRxiv. 2018;

[104] Hu H, Huff CD, Moore B et al. VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genet Epidemiol 2013; 37: 622–634

[105] Pollard KS, Hubisz MJ, Rosenbloom KR et al. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010; 20: 110–121

[106] Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974; 185: 862–864

[107] Tang H, Wyckoff GJ, Lu J et al. A universal evolutionary index for amino acid changes. Mol Biol Evol 2004; 21: 1548–1556

[108] Van Laer L, Coucke P, Mueller RF et al. A common founder for the 35delG GJB2 gene mutation in connexin 26 hearing impairment. J Med Genet 2001; 38: 515–518

[109] Ballana E, Ventayol M, Rabionet R et al. The Connexin-deafness Homepage. In September 8 2018 (ed) 2018

[110] Scriver CR. After the genome--the phenome? J Inherit Metab Dis 2004; 27: 305–317

[111] Oetting WS, Robinson Pn Fau - Greenblatt MS, Greenblatt Ms Fau - Cotton RG et al. Getting ready for the Human Phenome Project: The 2012 forum of the Human Variome Project..

[112] Deans AR, Lewis SE, Huala E et al. Finding Our Way through Phenotypes. Plos Biol 2015; 13: e1002033

[113] Poldrack RA, Congdon E, Triplett W et al. A phenome-wide examination of neural and cognitive function. Sci Data 2016; 3: 160110

[114] Köhler S, Vasilevsky NA-O, Engelstad M et al. The Human Phenotype Ontology. Nucleic Acids Res 2017; 45(D1): D865–D876

[115] Vasilevsky NA, Foster ED, Engelstad ME et al. Plain-language medical vocabulary for precision diagnosis. Nat Genet 2018; 50: 474–476

[116] Gall T, Valkanas E, Bello C et al. Defining disease, diagnosis, and translational medicine within a homeostatic perturbation paradigm: The National Institutes of Health Undiagnosed Diseases Program Experience. Front Med 2017; 4: 62

[117] Hartel BP, Lofgren M, Huygen PL et al. A combination of two truncating mutations in USH2A causes more severe and progressive hearing impairment in Usher syndrome type IIa. Hear Res 2016; 339: 60–68

[118] Morton CC. 2014; Presidential Address: The Time of Our Lives. Am J Hum Genet 2015: 347–351