The Journal of Laryngology & Otology (2014), ***128***, *848–858*.

doi:10.1017/S002221511400200X

REVIEW ARTICLE

Bioinformatics in otolaryngology research. Part one: concepts in DNA sequencing and gene expression analysis

T J OW[1,2], K UPADHYAY[2], T J BELBIN[2], M B PRYSTOWSKY[2], H OSTRER[2,3], R V SMITH[1,2,4]

Departments of [1]Otorhinolaryngology – Head and Neck Surgery, [2]Pathology, [3]Pediatrics, and [4]Surgery, Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, New York, USA

Abstract

Background: Advances in high-throughput molecular biology, genomics and epigenetics, coupled with exponential increases in computing power and data storage, have led to a new era in biological research and information. Bioinformatics, the discipline devoted to storing, analysing and interpreting large volumes of biological data, has become a crucial component of modern biomedical research. Research in otolaryngology has evolved along with these advances.

Objectives: This review highlights several modern high-throughput research methods, and focuses on the bioinformatics principles necessary to carry out such studies. Several examples from recent literature pertinent to otolaryngology are provided. The review is divided into two parts; this first part discusses the bioinformatics approaches applied in nucleotide sequencing and gene expression analysis.

Conclusion: This paper demonstrates how high-throughput nucleotide sequencing and transcriptomics are changing biology and medicine, and describes how these changes are affecting otorhinolaryngology. Sound bioinformatics approaches are required to obtain useful information from the vast new sources of data.

Key words: Bioinformatics; Otolaryngology; Sequencing Analysis, DNA; High-Throughput Nucleotide Sequencing; Gene Expression; Transcriptome

Introduction

The twenty-first century is proving to be an era driven by information. The increase in computing power and speed, the overwhelming impact of the internet, and the abundance of data available during the last two decades has changed nearly every aspect of life. The staggering amount of readily available information has led to the establishment of 'informatics', a discipline that focuses on the storage, retrieval and processing of data.

The biological sciences have been greatly impacted by the 'information age', which has led to the development of the field of bioinformatics. Advances in genetics and molecular biology have been driven recently by methodologies that generate massive amounts of data in a short time. For example, next-generation sequencing techniques can provide the base pair sequence of an entire human genome in a matter of weeks. Microarray technology can assess the gene expression or methylation status of tens of thousands of genes, probe a million single nucleotide polymorphisms, or capture DNA fragments composing the entire human exome on a single array chip.

These advances have led to an equivalent explosion in published scientific information that is readily accessible to the medical and scientific community worldwide. Bioinformatics is a multidisciplinary field that integrates a vast array of subjects, including computer science, mathematics, statistics and biology, with the goal of optimising the acquisition, storage, analysis, interpretation and application of biological data. An important goal of translational research in medicine today is to utilise and interpret this massive amount of data to yield information that is applicable to patient care, ultimately resulting in improved strategies for diagnosis, prognostication and treatment.

The advances highlighted above have already had a great impact on medical practice and the field of otolaryngology. Otolaryngology is a subspecialty that focuses on several congenital, inflammatory, immunological, infectious and neoplastic disorders, and discoveries in biology gleaned from modern bioinformatics

Accepted for publication 12 June 2014 First published online 16 September 2014

approaches will have an impact on the way several of these disorders are diagnosed and treated in the near future. This review highlights several of the most relevant applications of bioinformatics in otolaryngology. The goal of this review is to provide a basic understanding of bioinformatics to the clinical otolaryngologist. This paper briefly explains common modern molecular biology techniques, describes bioinformatics approaches for data analysis and interpretation, and provides several contemporary references as examples of these applications in otolaryngology.

The review has been divided into two parts. The first instalment details bioinformatics approaches to high-throughput nucleotide sequencing and gene expression analysis. Many principles involved in the bioinformatic approaches to these two types of data can be applied to other high-throughput molecular biology techniques. The second part of this review summarises several other modern genomic, epigenetic and molecular biology platforms, highlighting the considerations in bioinformatics for each. A glossary of terms specific to molecular biology, genomics and bioinformatics is provided as a reference for the reader (Table I).

Common bioinformatics applications in ENT

Bioinformatics is an immense field, and a full description of the breadth of applications that fall under the discipline of bioinformatics is well beyond the scope of this article. The first part of this review will focus on next-generation sequencing data and gene expression analysis, as several principles from these two platforms can be applied to other modern high-throughput systems. In the second part of the review, other important methodologies are summarised. The series concludes with a discussion of recent approaches used for integration of these data, and developments anticipated in the near future.

Bioinformatics in DNA sequencing

Over the course of approximately 50 years, our understanding of the human genome has grown exponentially. In 1953, Watson and Crick first reported their discovery of the double-helix structure of DNA, and proposed a mechanism for how heritable information was passed on from generation to generation.[1] It was not until Frederick Sanger developed the chain-termination method of DNA sequencing in 1977 that a rapid and reliable method of DNA sequencing became feasible.[2] Sanger sequencing techniques were used to carry out the Human Genome Project,[3,4] which was completed at the turn of this century after more than a decade of work across several institutions.

As the Human Genome Project neared completion, the demand for high-throughput sequencing that required less time and cost to produce large amounts of DNA sequence increased dramatically. During the last decade, modern 'next-generation' sequencing techniques have evolved. These techniques 'parallelise' sequencing reactions on an enormous scale, such that several sequences from smaller fragments of the DNA of interest are sequenced simultaneously and then computationally aligned in order to arrive at the final sequence result.[5,6] These techniques have now made it possible to sequence an entire human genome in a matter of weeks at a cost that is several orders of magnitude less than that of the Human Genome Project. Details of the methodology of next-generation sequencing are not discussed here, but for the purposes of this review it is important to emphasise that it has become affordable and efficient to produce large amounts of genomic sequencing data in a short time. Analysis and interpretation of these data requires distinct bioinformatics approaches.

Analysis of DNA sequences

Whether one wants to sequence a single fragment of DNA, hundreds of genes at a time, or sequence the entire human exome or genome, the data generated must undergo specific processing and analysis in order to be interpretable and to ultimately produce useful information. Single, small fragments of DNA (of the order of tens to hundreds of base pairs) can be aligned to a specific reference sequence. Several tools exist for this process.

Perhaps the most frequently utilised tool is the Basic Local Alignment Search Tool ('BLAST'), publically available via the National Center for Biotechnology Information.[7] This alignment tool was created in 1990,[8] and remains one of the most highly utilised bioinformatics applications. Using this tool, one can enter a DNA sequence or protein sequence, which is then compared, using a heuristic algorithm, against several databases to create a library of sequence matches. The Basic Local Alignment Search Tool has numerous applications, but one particularly useful function is to align and map a DNA sequence to the reference database for the human genome.

While the Basic Local Alignment Search Tool is useful to evaluate single or small numbers of DNA sequences of short length, more complex sequencing projects, such as those applying next-generation sequencing platforms, require alignment tools that are much more complex and efficient. When analysing large amounts of DNA sequence, the goals are generally two-fold: (1) to obtain the sequence itself, and (2) to identify the position of the examined sequences in a reference genome (e.g. the human genome). Next-generation sequencing techniques typically provide millions of copies of relatively short lengths of sequence reads, which must be assembled into the final target sequence. Sequence assembly can follow a *de novo* process, or it can be accomplished via alignment to a reference.[9] Figure 1 presents a simplified model of these two methods of sequence assembly.

Traditional alignment programs are interfaced with tools such as the Basic Local Alignment Search Tool. They operate in a manner similar to that described above, but on a much larger scale, using an automated

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press

TABLE I
GLOSSARY OF TERMS

Term	Description
Chain-termination (Sanger) sequencing	DNA sequencing methods that copy a DNA sequence of interest using randomly inserted nucleotides which halt the copying process when they are incorporated. This produces fragments of DNA that represent each nucleotide along the ultimate sequence. Serial evaluation of each fragment allows one to build the entire sequence
Chromosomal translocations	A rearrangement of parts between non-homologous chromosomes. Translocations can be balanced (an equal exchange of material, with no loss of genetic information) or unbalanced (an unequal exchange, where material is lost or gained in excess)
Deletion	A mutation that results from loss of nucleotides from the DNA sequence
Differentially expressed genes	Genes that are expressed at different levels between 2 groups (consistently increased or decreased)
DNA alignment	The process of matching a DNA sequence to a reference that is complementary to the sequence of interest
DNA copy number variation	The human genome often carries 2 copies of each gene within each cell. A deletion of 1 or both copies, or a duplication (amplification) creating more than 2 copies of a specific region of the genome, is called a copy number variation (CNV)
DNA sequence library	In next-generation sequencing, the DNA sequence library refers to the small regions of the experimental DNA that are sequenced in parallel. All of these smaller sequences in the library are then assembled to generate the final complete sequence
DNA sequence reads	The term 'reads' is often used to refer to the DNA sequences obtained from next-generation sequencing
DNA structural variants	Alterations in the genome that result in the loss or gain of chromosomes (deletions or amplifications), or chromosomal translocations
Exome	All regions of the genome that are eventually transcribed into mRNA (exons) & code for expressed genes
Gene expression	The process by which a coding DNA is ultimately expressed as a functional product (usually protein). Gene expression studies often examine the abundance of specific mRNA
Gene expression signature	In gene expression studies, this is a list of genes that are consistently expressed at different levels between experimental groups, as determined after statistical evaluation of all gene expression data across the groups. The 'signature' can then be used to identify a particular group based strictly on gene expression information
Genome	All of an organism's hereditary information. In humans, this includes DNA that codes for genes & non-coding DNA
Genome-wide association analysis	Studies that examine many common genetic variants in large samples of subjects that either do or do not possess the phenotype of interest (e.g. disease), in order to determine if the phenotype can be linked to a specific heritable trait
Genomic variants	The term 'variant' is used to refer to elements in a DNA sequence that differ from the reference sequence. These can be single nucleotide differences, deletions or amplifications (copy number variants), or structural changes (e.g. translocations)
Heuristic algorithm	Methods that trade accuracy & completeness for speed using approximations to problem solve; beneficial when an exhaustive approach would be highly inefficient
High-throughput	A term used to describe scientific methods that make large numbers of observations or collect several data points simultaneously
Hybridisation	The process of joining DNA fragments to their complementary sequences. Microarray research (e.g. gene expression microarray, DNA methylation array, comparative genome hybridisation array) often utilises hybridisation as a means to 'capture' the DNA of interest onto the array
Indels	A term in genetics & genomics commonly used to refer to both insertions & deletions collectively
Insertion	A mutation that results from extra nucleotides placed within the DNA sequencing
Linkage analysis	A method in genetics where one determines the location of a gene of interest by associating a phenotype with a genetic region of interest that is already known. For example, with a large enough cohort (e.g. a family with multiple members afflicted with a disorder passed along in a Mendelian fashion), a disease of interest can be associated with specific SNPs (defined below) that are known & always found in the individuals who are afflicted, thus mapping the region where the disease-related gene can be found
Mendelian pattern	Phenotypes (e.g. diseases or disorders) that are inherited in a pattern similar to classic Mendelian genetics, i.e. a single gene associated with the phenotype that carries 2 different alleles, 1 of which has a dominant influence over the other
Microarray	Any 2D substrate (usually glass or silicon) that allows evaluation of several analytes simultaneously
Microarray chip	The term 'chip' is often used to refer to a single microarray slide
Microarray probes	Most often, these are short specific nucleotide sequences that are spotted on the array & designed to pair with an analyte of interest (e.g. a cDNA reverse transcribed from a specific mRNA)
Missense mutation	A single nucleotide change that results in an alteration in a codon which is translated into a different amino acid
Next-generation sequencing	Modern methods of DNA sequencing that parallelise sequencing, i.e. several small & overlapping sections of the longer segment of DNA of interest are sequenced simultaneously, greatly increasing the speed of arriving at the ultimate sequence
Non-parametric statistics	Statistical methods that assume data do not follow a particular distribution. Often refers to statistics that examine categorical data, e.g. the chi-square statistic
Nonsense mutations	Missense mutations that change a codon coding for an amino acid into a stop codon, terminating the protein sequence
Non-synonymous mutation	Mutation that results in an alteration in the amino acid sequence that is translated

Parametric statistics	A set of statistical methods that assume data come from a probability distribution, e.g. *t*-tests (common parametric statistics which assume that data follow a Student's *t*-distribution)
Polymerase chain reaction (PCR)	A molecular biology technique that uses DNA polymerase (an enzyme that copies DNA) to amplify a specific region of interest of DNA by several orders of magnitude (thousands to millions more copies)
Reverse-transcription	The process of synthesising DNA from an RNA template
Reverse-transcription PCR (RT-PCR)	RNA is transcribed to DNA, creating a complementary DNA sequence with several orders of magnitude more copies. RT-PCR can be done quantitatively, & can be used to measure the relative abundance of mRNA present for a given transcript, thus estimating the level of expression of a specific gene
Ribosomal RNA	RNA that forms the structural component of the ribosome. Ribosomal RNA is not translated into protein
RNA transcript	RNA sequence that has been transcribed from a DNA coding region
Single nucleotide polymorphism (SNP)	Variations at a single nucleotide in the genome observed within a population. Millions of these variations occur in the human genome, the majority of which are not associated with disease. SNPs can occur within non-coding DNA (the majority) or coding DNA, & they can have either no functional impact or significant impact
Synonymous mutation	Mutation that results in a different nucleotide sequence which causes no alteration in the translated amino acid sequence
Translational research	Research that endeavours to apply information gleaned from basic science research to improve clinical medicine & human health
Type 1 error	Generally speaking, this is an experimental 'false-positive', where the null hypothesis is rejected in error
Variant annotation	In the analysis of high-throughput next-generation sequencing results, this is the process of characterising variants identified between the sequence derived & the reference; for instance, an annotation tool may determine the location of variants (e.g. exonic, intronic, splice site) & predict functional impact (e.g. synonymous, non-synonymous)
Variant call format ('VCF') file	In next-generation sequencing, VCF format is the standard text format used to report variations between the sequence result & the reference sequence
Variant calling	The process of identifying differences between sequence results & the reference in next-generation sequencing

MRNA = messenger RNA; 2D = two-dimensional; cDNA = complementary DNA

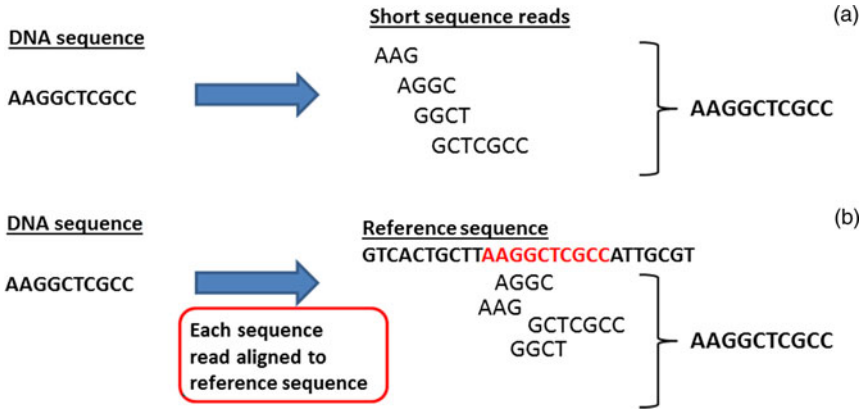

FIG. 1
Simplified schematic of (a) *de novo* versus (b) alignment sequence assembly.

and iterative process. A significant amount of computing power and time is required to accomplish the ultimate alignment.[10] Newer programs, such as Maq or Bowtie, use highly efficient algorithms to perform sequence alignment. A summary of these methods as well as other common open-source alignment tools are discussed in a review by Trapnell and Salzberg.[10]

The alignment method for sequence assembly, rather than a *de novo* process, is the most commonly used approach for next-generation sequencing applications in medicine. *De novo* sequencing assembly is a process by which reads are aligned based on overlapping regions of the analysed fragments. Sequences generated with overlapping common regions are 'aligned' to piece together the ultimate read for the complete sequence of interest. These methods obviously require assembly programs that perform different algorithms than the alignment tools. *De novo* assembly is commonly applied to smaller genomes, such as those of bacteria. *De novo* sequencing of a large target sequence, such as that of the human genome, is still a major challenge.[9] *De novo* sequencing methods and tools are further discussed in a review by Nagarajan and Pop.[11]

For most DNA sequencing applications in otolaryngology, and medicine in general, the goal is to identify abnormalities in the DNA sequence, namely germline or somatic mutations. Once sequence data are aligned to a reference (for example, the latest build of the human genome), one can generate a catalogue of any differences between the DNA sequences and the reference. These are often referred to as 'variants'; this generic term is used because these differences can be due to either mutation or known polymorphisms (i.e. normal differences in the DNA sequence that occur with a defined frequency in the population). When localised to one nucleotide, these polymorphisms are referred to as single nucleotide polymorphisms. The term 'variants' can also be used to describe changes in the copy number of genomic elements (known as copy number variations) or other structural changes, such as chromosomal translocations.

Variant calling (the process of identifying differences between the experimental sequence and the reference to which the sequence is aligned) might seem simple in principle, but there are several barriers to accurate variant calling when analysing a large volume of sequence reads (for example, when examining an exome or a genome). Differences between the DNA sequence and the reference can be secondary to error or misalignment, which can be influenced by several factors. For example, false variant calls are common when: insertion/deletion mutations (indels) are present; regions of DNA containing a high volume of repeated guanine/cytosine nucleotides (guanine-cytosine rich regions) are evaluated; or there are errors due to polymerase chain reaction artefacts in library construction or poor sequence signal, which is common at the ends of each sequence read.[12]

Variant calling for next-generation sequencing data requires alignment, adjustments for quality control and probability algorithms to define the confidence with which a variant is 'called'. Several software tools exist for this process, such as the Genome Analysis Toolkit,[13] VarScan or VarScan2,[14] SOAPsnp (a member of the Short Oligonucleotide Analysis Package),[15] and Atlas 2.[16] The outputs from these analyses are commonly presented in a 'variant call format', or 'VCF' file, which is a text format that has been developed in order to standardise these data for use in large-scale projects, such as the 1000 Genomes Project. A description of the variant call format can be found on the 1000 Genomes website.[17] Another step in analysing these data, particularly for whole genome or whole exome sequencing, is the determination of regions of minor or major structural variation (e.g. insertions, deletions or break points of translocations), which is much more complicated than the identification of single nucleotide variants. Programs that carry out these processes include Pindel,[18] Dindel[19] and some features in the Genome Analysis Toolkit.

The subsequent steps after variant calling are filtering and annotation steps. The goal of this part of the

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press

process is to identify sequence variants of interest. In otolaryngology research, these would be variants that are likely to cause functional alteration in a protein or phenotypic changes at the cellular level, and those that ultimately contribute to disease.

There are several methods of filtering and characterising genomic variants, and the process is tied to the specific scientific question. For example, if the goal is to identify significant mutations in a tumour sample, one might compare the variants called in the tumour DNA with the DNA from a sample of normal tissue (e.g. blood or adjacent mucosa) from the patient. This would improve the identification of acquired mutations in the tumour sample as opposed to variants that were normal polymorphisms harboured by the genome of the individual. Variants can be referenced to large single nucleotide polymorphism databases, such as the Single Nucleotide Polymorphism Database ('dbSNP') on the National Center for Biotechnology Information website[20] or the 1000 Genomes Project[21,22] database,[23] in order to filter out normal polymorphisms known to exist in the population from novel mutations.

The functional consequences of variants identified can be evaluated in a high-throughput fashion using variant annotation tools. These programs rapidly scan all reported variants in a project in order to determine if they occur in coding regions and if the variants cause significant alteration of the amino acid sequence in the translated protein (e.g. synonymous *vs* non-synonymous mutations). Again, several programs are available to annotate variants, including SIFT (Sorting Intolerant from Tolerant),[24] PolyPhen[25] and Annovar,[26] though there are many others.

A review by Dolled-Filhart and colleagues summarises the process for analysing next-generation sequencing data, and highlights several additional tools available for alignment, variant calling and annotation.[12] Table II provides a list and brief description of several analysis tools that the authors have found useful for the processing and analysis of next-generation sequencing data.

To date, next-generation sequencing in otolaryngology has arguably had the largest impact in the fields of cancer biology and congenital deafness. In 2011, two articles reported results from whole exome sequencing of head and neck squamous cell carcinoma (SCC), and each independently discovered that mutations in the gene *NOTCH1* frequently occurred in these tumours.[27,28] Sequence details from both studies showed frequent missense and nonsense mutation, generally occurring upstream of the transmembrane domain of the protein. These studies also confirmed several mutations in genes known to be frequently altered in head and neck SCC (e.g. *TP53*, *CDKN2A*),[27,28] and identified other mutations in novel genes (e.g. *CASP8*, *FAT1*).[28] Since these results were published, results from whole exome sequencing have been reported for both medullary thyroid carcinoma[29] and adenoid cystic carcinoma.[30] Each of these studies used next-generation sequencing to evaluate the whole exome of matched tumour–normal pairs, and used similar bioinformatics approaches to arrive at variant calls unique to these cancer types.

Research has also led to significant discoveries in the genetics of hearing loss. A number of genetic factors have been linked to both syndromic and non-syndromic deafness. Several of these factors are summarised in a recent review by Shearer and Smith.[31] Next-generation sequencing techniques have been utilised to identify novel mutations associated with hearing loss. For example, a recent study using whole exome sequencing identified a novel mutation in *COCH* in a Chinese family with progressive autosomal dominant hearing loss.[32] Another study, which evaluated 13 Korean families with autosomal recessive non-syndromic hearing loss, identified frequent variants in the gene *MYO15A*.[33] Along with discovery research, next-generation sequencing is also being advocated as a diagnostic tool. These methods have been proposed as means to: identify patients that carry mutations known to be associated with hearing loss, and discover novel variants in known genetic loci that have previously been shown to be critical for hearing.[34]

New discoveries and applications using next-generation sequencing are changing our understanding of human disease, and improving approaches to diagnosis and treatment. Applications in otolaryngology are in their infancy. The methods in bioinformatics described above are crucial for the advancement of translational genomics.

Bioinformatics for analysis of high-throughput gene expression data

The ability to evaluate the global expression of thousands of genes in one experiment has been available for approximately two decades, since the advent of microarray technology.[18,35] The standard method for high-throughput gene expression has been via microarray analysis. In these techniques, RNA, usually messenger RNA (mRNA), is either copied or converted to complementary DNA using reverse transcription. These are then fluorescently labelled and hybridised to an array carrying thousands of probes corresponding to specific genes. The arrays are scanned with a laser that causes the hybridised samples to fluoresce; the relative intensity at each probe corresponds to the relative abundance of each transcript. These data can then be used to compare the relative expression of genes that are represented by the probes on the array.

RNA-Seq is a new technology that utilises next-generation sequencing to evaluate RNA expression using a different approach.[36] With RNA-Seq, RNA is harvested, and the RNA of interest is isolated (e.g. ribosomal RNA is often discarded, or perhaps only mRNA is isolated). The RNA is then reverse transcribed, and the transcripts are sequenced using next-generation techniques. Once the reads are assembled, the 'transcriptome'

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press

TABLE II
SELECTED ANALYSIS TOOLS FOR EVALUATION OF DNA SEQUENCING DATA

Tool type	Description
Sequence alignment	
Database search	Tools that align a short sequence to a reference database
– BLAST	BLAST (Basic Local Alignment Search Tool) allows alignment of shorter sequence reads to a reference using a local search with a fast k-tuple heuristic
– FASTA	For alignment of shorter sequences. Involves a local search with a fast k-tuple heuristic; slower but more sensitive than BLAST
Pairwise alignment	Tools that compare 2 sequences to each other
– Bioconductor Biostrings: pairwise alignment	Bioconductor is managed by the Fred Hutchinson Cancer Center, which uses R programming language to provide several bioinformatics tools, including this tool for pairwise sequence alignment
– BioPerl dpAlign	BioPerl offers several open-source bioinformatics tools written using various versions of Perl programming language, including this pairwise alignment tool
– JAligner	Open-source program written in Java programming language that utilises the Smith–Waterman algorithm for alignment
Multiple sequence alignment	Tools that align multiple sequences concurrently
– ClustalW	Uses a progressive alignment algorithm to align multiple sequences (aligns sequences in a hierarchical manner)
– Sequence Alignment/Map ('SAM')	Uses a hidden Markov method (probabilistic algorithm) to align multiple sequences
Short read sequence alignment	Tools used to align vast numbers of short sequence reads, applicable for next-generation sequencing
– Bowtie	Uses the Burroughs–Wheeler transform to index the human genome, & provides rapid & efficient alignment of sequences to the reference
– Burroughs–Wheeler Alignment ('BWA') tool	Also uses the Burroughs–Wheeler transform to index the human genome. Slower than Bowtie but allows for insertions & deletions in the alignment
– Maq	Alignment tool used for Illumina® & ABI SOLiD™ (Sequencing by Oligonucleotide Ligation and Detection) platforms of next-generation sequencing (not compatible with 454 or capillary sequencing). Used for ungapped sequences (cannot align sequences with insertions or deletions). Provides a probability score for each alignment
Variant calling	Tools that identify sequences which differ from a reference sequence
– Genome Analysis Toolkit ('GATK')	Offers a wide variety of tools, with a primary focus on variant discovery & genotyping. Also strongly considers data quality
– VarScan	Platform-independent, technology-independent software tool for identifying SNPs & indels in massively parallel sequencing of individual & pooled samples
– SNVer	A statistical tool for calling common & rare variants in the analysis of pooled or individual next-generation sequencing data
– SAMtool	Provides short DNA sequence read alignments, supporting complex tasks like variant calling & alignment viewing, as well as sorting, indexing, data extraction & format conversion
– CRISP	CRISP (Comprehensive Read analysis for Identification of SNPs from Pooled sequencing) can be used to identify both rare & common variants
Variant annotation	Evaluates variants for position (e.g. exonic region, intronic region, splice site) & probable effect (e.g. amino acid change)
– SIFT	SIFT (Sorting Intolerant From Tolerant) predicts whether an amino acid substitution affects protein function based on the degree of amino acid conservation in related sequences noted against the position-specific iterated BLAST (PSI-BLAST) database
– PolyPhen-2	Similar to SIFT, PolyPhen-2 assesses the functional impact resulting from an amino acid change in a protein, & uses several algorithms to evaluate sequence, structure & predicted function against several databases
– Annovar	It can be used to: annotate genetic variants and predict whether SNPs or indels cause protein coding changes; compare variants against the dbSNP (Single Nucleotide Polymorphism Database) & 1000 genomes database to determine if a variant is a reported polymorphism; & evaluate functional impact based on SIFT score

SNP = single nucleotide polymorphism

sequences can be aligned to the reference genome, and relative expression can be estimated from the coverage depth of each read.[36] Subsequent analysis of gene expression microarray data and RNA-Seq transcriptome data requires distinct bioinformatics approaches to glean useful information.

The first step in analysing microarray data is to normalise the data, in order to eliminate measurement noise that is systemic to the platform being utilised for the study. With tens of thousands of signals from each chip analysed, there are several sources of variation: there is normal variation between signals on an individual chip, between data gathered from different chips and between experimental batches.[37] Methodologies and programs used to normalise array data are a subject for an entire article. In general, normalisation processes employ several strategies, including methods that use internal controls built into each array (e.g. duplicate probes placed in different locations of an array) and approaches that measure 'housekeeper' probes that generally show low variability between samples. Normalisation also considers variation due to batch effect. Several statistical methods (e.g. global normalisation (which equalises the mean or median values for each array in the experiment), parametric linear regression methods and non-parametric methods) aim to adjust the array output values so that differences identified are not due to inherent or experimental variation.[38] There are many approaches and methods that can be employed to normalise microarray data. Reviews by Fan and Ren,[39] and Gusnanto and colleagues,[38] provide further details of some standard methods.

Along with normalisation, a filtering process is often employed. The goal of filtering is to disregard any probes that may not provide meaningful data and would thus add 'noise' that could dilute the data of interest.[38] For example, if a fraction of probes show very low expression across all of the samples in the experiment, one may wish to eliminate these from the analysis. Conversely, one may want to eliminate probes that show extreme variation between samples. The filtering process depends on the goals of the experiment; if conducted appropriately, the filtering process can improve the signal-to-noise ratio, which is an inherent problem in microarray research.

After normalisation and filtering, data can be analysed to seek an answer to the experimental question. The methodology for interpreting microarray data is diverse, and largely depends on the goals of the experimental design. The following are some basic principles and approaches.

One starting point is to determine if a supervised or an unsupervised approach would be best. Unsupervised approaches are methods that examine inherent patterns within the entire dataset without outside bias. Unsupervised methods include feature determination (a common technique is called a principle component analysis), cluster determination and network determination.[37] Supervised analyses are methods that are used to determine which genes or groups of genes on the array can best differentiate between pre-determined groups (e.g. tumour tissue *vs* normal tissue, treated *vs* untreated). Statistical methods to perform a supervised analysis range from approaches that are fairly simple, such as basic parametric or non-parametric comparisons between the mean or median values of each probe between each group, to extremely complex, such as the use of support vector machines or Bayesian methods to segregate groups.[37] The authors point the reader to the concise review by Butte,[37] who highlights some common supervised and unsupervised methods for gene expression array analysis.

One concept that is important to discuss is that of the 'false discovery rate'. This is a measure of error, which was developed approximately two decades ago coinciding with the increased application of microarray technology.[40] In brief, when one performs multiple statistical tests in an iterative fashion, there exists an inherent error with each test performed (generating a *p*-value for each individual test, which describes the level of confidence that the result was not due to a type 1 error). As more tests are performed, the chances of 'discovering' a positive result simply by chance increases with each additional test. Therefore, the stringency for calling 'true positives' over 'false positives' should increase as the number of tests increases. Traditional methods for adjusting the accepted error rate (namely, the family-wise error rate, the most common method of which is the Bonferroni correction) are very stringent. When applied to the thousands of data points generated with microarray data, these methods have a high likelihood of dismissing pertinent results. The false discovery rate is a less stringent correction than family-wise error rate methods, and is commonly reported as a *q*-value, which can be interpreted as similar to a *p*-value generated in statistics and used for single-hypothesis testing.

The steps highlighted above involve complex bioinformatics approaches to arrive at the end result in microarray analysis; namely, a list of genes corresponding to array probes that appear to be associated with the experimental condition of interest. These could be genes that cluster together in an unsupervised analysis, forming what appear to be biologically distinct entities, or genes that are differentially expressed between predefined categories (i.e. in a supervised analysis, groups are predetermined and genes that are expressed differently between the different categories are then identified). An example of a signature generated from a supervised analysis would be a list of genes which are differentially expressed in a set of tumours that respond to a specific therapy versus a set of tumours that do not respond.

Whether attained via unsupervised or supervised analyses, the results of these studies generate a list of genes of interest, which must subsequently be validated

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press

both internally and externally. Internal validation examines the expression of individual genes in the original test samples (with reverse-transcription polymerase chain reaction, for example). This is done in order to verify that what was seen on the array was truly from gene expression levels in the samples, as opposed to artefacts on the array or artefacts secondary to experimental methods. Additionally, the results should be validated externally; findings in the test set must often be verified on an additional sample set (e.g. a 'repeated' experiment on an independent cohort) in order to deem the results generalisable.

One important concept to mention is the problem of 'overfitting' when developing a gene expression signature. Because multiple variables are tested while developing these signatures (i.e. comparisons between thousands of probes), there is a relatively high likelihood that a statistical model generated from the dataset to describe a given cohort will only be applicable to the experimental cohort; the model is therefore 'overfitted' internally. This is the main reason why external validation is a required part of the process in generating gene expression signatures. The difficulty in replicating and validating results generated from gene expression microarray experiments is arguably one of the greatest barriers to the progress of these research endeavours.

Gene expression analysis has been applied in several areas of otolaryngology. By far the widest application has been in the field of head and neck cancer biology. A number of interesting bioinformatics approaches have been implemented in these studies. For example, in an early study by Chung *et al.*, 60 tumour samples were examined using complementary DNA gene expression arrays.[41] First, a pooled subset of 30 randomly selected samples from the test set was used as the reference (early generation gene expression arrays used a two-colour fluorescence system to compare a test sample to a reference sample). The gene expression data were normalised and filtered in an interesting manner. As part of the filtering process, in a subset of 10 samples, RNA was extracted from 2 different regions of the same tumour. When these paired samples were examined, genes that showed little intrinsic variance (i.e. variance between matched pairs) but high variance across unmatched samples were selected for further evaluation. An unsupervised hierarchical clustering analysis identified four distinct subtypes of tumours, which appeared to have distinct molecular characteristics.

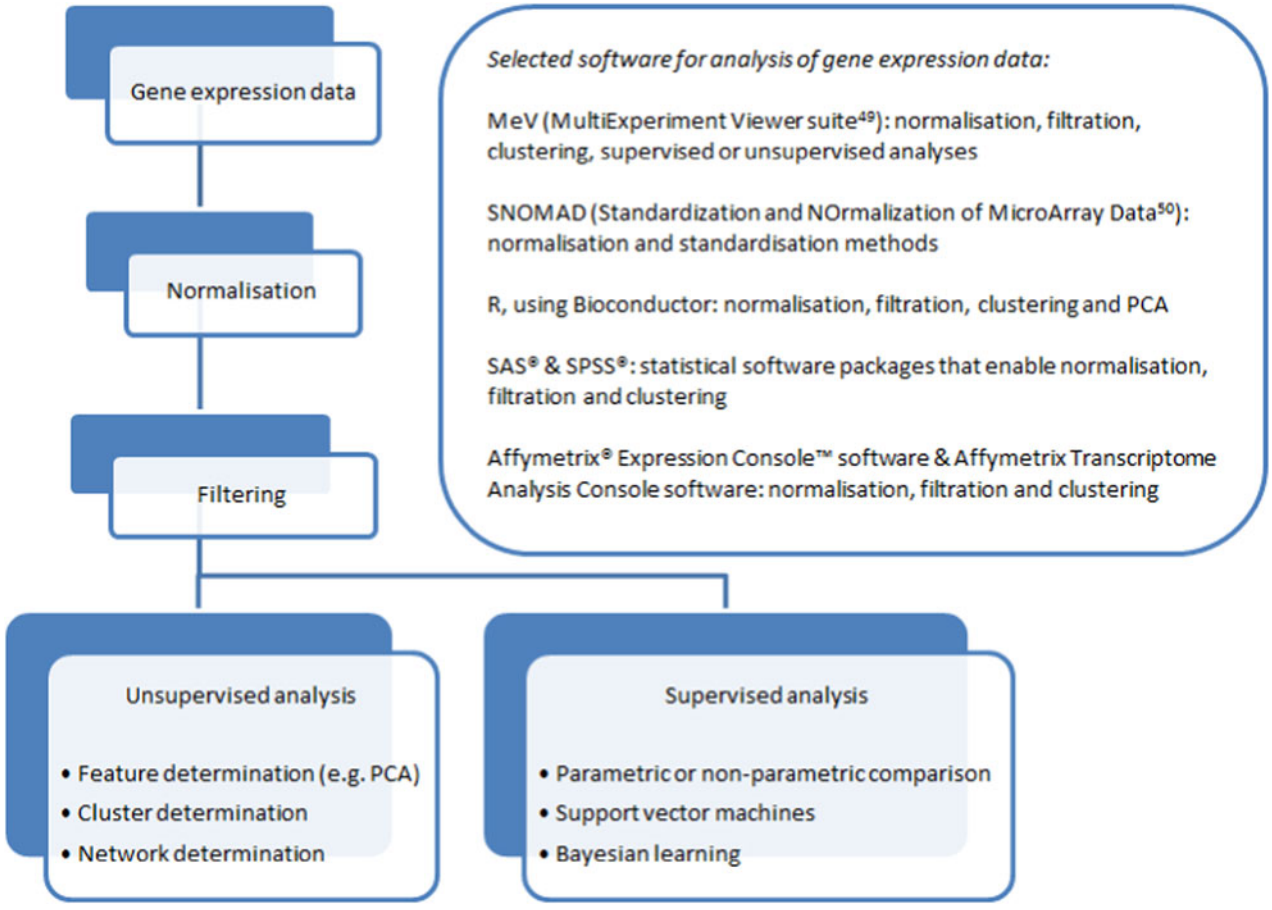

FIG. 2

Steps for processing gene expression data, and list of selected tools for normalisation, filtering and analysis. PCA = principle component analysis

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press

A supervised analysis using two specific statistical methods (*k*-nearest neighbour method and prediction analysis of microarrays) was carried out to develop a model that was 80 per cent accurate at predicting pathological lymph node status.[41]

In another early study, Belbin and colleagues examined 17 head and neck SCC samples.[42] A gene expression profile derived using an unsupervised analysis was predictive for overall survival within the patient set.

In more recent examples of gene expression studies in head and neck SCC, Schlecht and colleagues used a supervised analysis of gene expression microarray data to demonstrate that human papilloma virus (HPV) positive and HPV-negative head and neck SCC have distinct gene expression profiles.[43] Two studies examining HPV-negative oral cavity SCC demonstrated that a gene signature could be a better predictor of survival than clinicopathological factors.[44,45] In the latter study, the authors notably developed a predictive gene expression signature using an iterative supervised method on a 97-patient training set, and validated the result on an external dataset.[45] These examples highlight how the bioinformatics principles described in this review have been employed in head and neck cancer research.

Gene expression has not been limited to cancer biology alone. A study by Klenke and colleagues compared gene expression in cholesteatoma with that in ear canal skin, which demonstrated expression patterns consistent with chronic inflammation and characteristics similar to those of invasive tumours.[46] A study by Stankovic *et al*. reported a distinct transcriptional signature in nasal polyps associated with chronic sinusitis, compared with those in patients with aspirin-sensitive asthma.[47] Another interesting study, by Vambutas and colleagues, used gene expression arrays to evaluate peripheral blood mononuclear cells extracted and cultured from patients with autoimmune hearing loss.[48] That study revealed that stimulation of cultured peripheral blood mononuclear cells with perilymph extracted at the time of cochlear implantation led to differential expression of interleukin 1 receptor type II.[48] These studies demonstrate the wide potential applications of gene expression in otolaryngology, and these studies exemplify both the creative approaches and the bioinformatics challenges in analysing and applying these data.

The bioinformatics approaches to analysing gene expression data are complex. Figure 2 summarises the basic approaches to gene expression analysis and lists some tools that the authors have found useful.

Conclusion

As evidenced by this review, high-throughput nucleotide sequencing and transcriptomics are changing biology and medicine. Gleaning useful information from these vast data sources requires sound bioinformatics approaches. Several of the bioinformatics principles described for next-generation sequencing and gene expression analysis are applicable to other high-throughput molecular biology techniques. The next part of this review highlights several other high-throughput platforms, and discusses recent approaches that seek to integrate data from multi-platform projects.

References

1 Watson JD, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 1953;**171**:737–8
2 Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;**74**:5463–7
3 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921
4 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG *et al*. The sequence of the human genome. *Science* 2001;**291**:1304–51
5 Shendure J, Lieberman Aiden E. The expanding scope of DNA sequencing. *Nat Biotechnol* 2012;**30**:1084–94
6 Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402
7 Basic Local Alignment Search Tool (BLAST). In: http://blast.ncbi.nlm.nih.gov/Blast.cgi [22 September 2013]
8 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10
9 Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet* 2010;**11**:31–46
10 Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. *Nat Biotechnol* 2009;**27**:455–7
11 Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet* 2013;**14**:157–67
12 Dolled-Filhart MP, Lee M Jr, Ou-Yang CW, Haraksingh RR, Lin JC. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *ScientificWorldJournal* 2013;**2013**:730210
13 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303
14 Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER *et al*. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;**25**:2283–5
15 Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008;**24**:713–14
16 Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C *et al*. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* 2012;**13**:8
17 1000 Genomes. VCF (Variant Call Format) version 4.1. In: http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41 [22 September 2013]
18 Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**:2865–71
19 Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. *Genome Res* 2011;**21**:961–73
20 dbSNP home page. In: http://www.ncbi.nlm.nih.gov/projects/SNP/ [2 September 2013]
21 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73
22 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65
23 1000 Genomes FTP directory. In: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/ [2 September 2013]
24 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**:1073–81

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press

25 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9
26 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164
27 Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ *et al*. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* 2011;**333**:1154–7
28 Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A *et al*. The mutational landscape of head and neck squamous cell carcinoma. *Science* 2011;**333**: 1157–60
29 Agrawal N, Jiao Y, Sausen M, Leary R, Bettegowda C, Roberts NJ *et al*. Exomic sequencing of medullary thyroid cancer reveals dominant and mutually exclusive oncogenic mutations in RET and RAS. *J Clin Endocrinol Metab* 2013;**98**:E364–9
30 Ho AS, Kannan K, Roy DM, Morris LG, Ganly I, Katabi N *et al*. The mutational landscape of adenoid cystic carcinoma. *Nat Genet* 2013;**45**:791–8
31 Shearer AE, Smith RJ. Genetics: advances in genetic testing for deafness. *Curr Opin Pediatr* 2012;**24**:679–86
32 Gao J, Xue J, Chen L, Ke X, Qi Y, Liu Y. Whole exome sequencing identifies a novel DFNA9 mutation, C162Y. *Clin Genet* 2013;**83**:477–81
33 Woo HM, Park HJ, Baek JI, Park MH, Kim UK, Sagong B *et al*. Whole-exome sequencing identifies MYO15A mutations as a cause of autosomal recessive nonsyndromic hearing loss in Korean families. *BMC Med Genet* 2013;**14**:72
34 Yan D, Tekin M, Blanton SH, Liu XZ. Next-generation sequencing in genetic hearing loss. *Genet Test Mol Biomarkers* 2013; **17**:581–7
35 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA. *Science* 1995;**270**:467–70
36 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63
37 Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov* 2002;**1**:951–60
38 Gusnanto A, Calza S, Pawitan Y. Identification of differentially expressed genes and false discovery rate in microarray studies. *Curr Opin Lipidol* 2007;**18**:187–93
39 Fan J, Ren Y. Statistical analysis of DNA microarray data in cancer research. *Clin Cancer Res* 2006;**12**:4469–73
40 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* 1995;**57**:289–300
41 Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D *et al*. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* 2004;**5**:489–500
42 Belbin TJ, Singh B, Barber I, Socci N, Wenig B, Smith R *et al*. Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. *Cancer Res* 2002;**62**:1184–90
43 Shlecht NF, Burk RD, Adrien L, Dunne A, Kawachi N, Sarta C *et al*. Gene expression profiles in HPV-infected head and neck cancer. *J Pathol* 2007;**213**:283–93
44 Mendez E, Houck JR, Doody DR, Fan W, Lohavanichbutr P, Rue TC *et al*. A genetic expression profile associated with oral cancer identifies a group of patients at high risk of poor survival. *Clin Cancer Res* 2009;**15**:1353–61
45 Lohavanichbutr P, Mendez E, Holsinger FC, Rue TC, Zhang Y, Houck J *et al*. A 13-gene signature prognostic of HPV-negative OSCC: discovery and external validation. *Clin Cancer Res* 2013;**19**:1197–203
46 Klenke C, Janowski S, Borck D, Widera D, Ebmeyer J, Kalinowski J *et al*. Identification of novel cholesteatoma-related gene expression signatures using full-genome microarrays. *PloS One* 2012;7:e52718
47 Stankovic KM, Goldsztein H, Reh DD, Platt MP, Metson R. Gene expression profiling of nasal polyps associated with chronic sinusitis and aspirin-sensitive asthma. *Laryngoscope* 2008;**118**:881–9
48 Vambutas A, DeVoti J, Goldofsky E, Gordon M, Lesser M, Bonagura V. Alternate splicing of interleukin-1 receptor type II (IL1R2) in vitro correlates with clinical glucocorticoid responsiveness in patients with AIED. *PloS One* 2009;**4**:e5293
49 TM4 Microarray Software Suite. In: http://www.tm4.org/ [6 August 2014]
50 SNOMAD: Standardization and NOrmalization of MicroArray Data. In: http://pevsnerlab.kennedykrieger.org/snomad.php [6 August 2014]

Address for correspondence:
Dr Thomas J Ow,
Department of Otorhinolaryngology – Head and Neck Surgery,
Montefiore Medical Center,
3rd Floor MAP Building,
3400 Bainbridge Avenue,
Bronx,
New York
10467, USA

E-mail: thow@montefiore.org

Dr T J Ow takes responsibility for the integrity of the content of the paper
Competing interests: None declared

https://doi.org/10.1017/S002221511400200X Published online by Cambridge University Press