The Journal of Laryngology & Otology (2014), ***128***, *942–947*.
© JLO (1984) Limited, 2014
doi:10.1017/S0022215114002011

REVIEW ARTICLE

Bioinformatics in otolaryngology research. Part two: other high-throughput platforms in genomics and epigenetics

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

Objectives: This second segment of the two-part review summarises several modern high-throughput methods in genomics, epigenetics and molecular biology. Many principles from nucleotide sequencing and transcriptomics can be applied to other high-throughput molecular biology techniques. Specifically, this manuscript reviews: array comparative genome hybridisation; single nucleotide polymorphism arrays; microarray technology, used to study epigenetics; and methodology applied in proteomics. Finally, the review describes current methods for the integration of multiple molecular biology platforms.

Conclusion: Progress in treating human disease in general will require close collaboration with experts in bioinformatics. Improved understanding, by clinicians and physician-scientists in our field, of the concepts presented in both parts of this review will advance diagnosis and therapy for diseases of the head and neck.

Key words: Bioinformatics; Otolaryngology; Polymorphism, Single Nucleotide; Comparative Genomic Hybridization; DNA Methylation; Epigenomics; Proteomics

Introduction

Part one of this series focused on high-throughput nucleotide sequencing and gene expression analysis. However, there are a multitude of other modern high-throughput molecular biology techniques, each with its own bioinformatics methodology and challenges. Many of the principles from sequencing and transcriptional analysis can be applied to the other molecular biology platforms. Here, we discuss other common high-throughput techniques, and briefly summarise the approach to analysis for each.

Comparative genome hybridisation arrays

These arrays are used to detect copy number changes (or copy number variations) in the genome (i.e. deletions or amplifications of specific regions). This is accomplished using microarrays with thousands of DNA probes (small fragments of complementary DNA) designed to analyse the presence or absence of sequential regions along the genome. Usually, DNA extracted from a test sample and that of a reference sample (for example, a tumour *vs* a normal blood sample in the same patient) are compared on a single array using a two-colour system. The signal intensity at each probe is equivalent to a relative copy number for that small region of the genome.

Just as in gene expression analysis, signals on the array must be normalised and filtered, both internally on an individual chip and across chips, to properly compare samples in an experiment. Copy number variation 'calls' involve using probe expression and that of adjacent probes in the genome to identify regions where there is allelic loss or amplification.[1] Similar to gene expression arrays, comparative genome hybridisation array data must be normalised, filtered and analysed critically in order to determine if variation in probe expression is due to normal variation or experimental error, or truly representative of copy number. The resolution of the array depends on how much 'space' is between each probe; for example, if a copy number variation occurs in between two regions of probed DNA, it will not be detected. Current comparative genome hybridisation arrays have a resolution of between 100 and 10 000 base pairs, depending on the array.

A recent report by Morris and colleagues used comparative genome hybridisation arrays to describe common copy number variations in head and neck

Accepted for publication 12 June 2014 First published online 17 September 2014

https://doi.org/10.1017/S0022215114002011 Published online by Cambridge University Press

squamous cell carcinoma (SCC).[2] In this study, which acts as an example of the application of comparative genome hybridisation arrays in head and neck cancer, several alterations in the EGFR/PI3K pathway were identified, including novel microdeletions in the *PTPRS* gene.

Single nucleotide polymorphism arrays

Single nucleotide polymorphism arrays are microarrays that contain probes specifically examining the presence or absence of single nucleotide polymorphisms, which are single base pair variations in the human genome that are known to occur with regular frequency in the human genome. There are approximately 20 million described single nucleotide polymorphisms,[3] and current arrays can interrogate approximately 1 million of these on a single chip.

Single nucleotide polymorphism arrays are used to genotype an individual for these polymorphisms. The data can be used to carry out linkage analysis and genome-wide association studies. The concept underlying these studies is that inherited polymorphisms that are near a germline mutation or disease-related gene will be inherited in a Mendelian pattern. The relative intensity of signals for polymorphisms on the array can also be used to estimate copy number variations and structural variants along the genome.[1] When utilised in this fashion, bioinformatics approaches to call copy number variations resemble those used with comparative genome hybridisation arrays, with normalisation steps to determine signal thresholds that correspond to copy number.

When using single nucleotide polymorphism data to evaluate copy number variations in cancer studies, it is most appropriate to compare single nucleotide polymorphism data derived from tumour DNA (preferably enriched for tumour cells via pathological assessment and/or laser-capture microdissection) with baseline single nucleotide polymorphism expression in a normal tissue DNA reference from the same patient. A review by Chen and Chen summarises several studies of copy number variations in head and neck SCC, and further discusses the utility of single nucleotide polymorphism arrays.[4]

MicroRNA expression arrays

MicroRNAs are small, non-coding fragments of RNA that regulate gene expression, often silencing genes by binding to specific messenger RNA (mRNA) transcripts leading to their degradation.[5] Disruption of microRNA regulation has been implicated in a multitude of diseases.

Global expression of microRNA can be evaluated with microarrays designed to probe hundreds of known microRNAs at a time. Analysis of these arrays is essentially equivalent to gene expression arrays, except that probes correspond to known microRNAs instead of mRNA transcripts.

Examples of microRNA evaluation in head and neck SCC include studies from our own institution (Childs *et al.*[6] and Harris *et al.*[7]). The latter study used a unique bioinformatics approach. The ratio of tumour versus normal microRNA expression was calculated for each sample, and then a rank consistency score was used to identify which microRNAs were consistently over- or under-expressed among samples. Using this process, miR-375 was identified as the most consistently decreased transcript among head and neck SCC tumours, and low levels of miR-375 were associated with poor survival.[7]

DNA methylation arrays

Another major source of epigenetic regulation is via methylation of DNA. The addition of a methyl group to the 5' region of cytosines found in gene promoter regions typically causes a reduction of gene expression of the associated gene, and modifications are often found at clusters of CpG dinucleotides (commonly referred to as 'CpG islands').[8]

Currently, over 450 000 genome-wide methylation events can be evaluated with methylation arrays such as the Illumina© Infinium HumanMethylation450 BeadChip©. Sample DNA is treated with bisulphite, which converts cytosine bases to uracil, but does not change methylated cytosine. Arrays are constructed with probes that are specific for known CpG sites, and complementary probes contain both the cytosine and uracil versions of each site. Therefore, the methylation status of each CpG site can be evaluated by measuring the hybridisation relative to each complementary probe pair. Bioinformatics methods of analysis are thus similar to comparative genome hybridisation arrays or single nucleotide polymorphism arrays.

As a recent example, a study conducted at our own institution examined DNA methylation events in 118 head and neck SCC tumours, and demonstrated differential methylation events that were unique to the subsite of the tumour and human papilloma virus (HPV) status.[9]

Proteomics

This article has touched on methodologies and bioinformatics in genomics and epigenetics. Proteomics, the comprehensive evaluation of protein expression, structure, modification and function in biological systems, deserves brief mention here.

High-throughput protein analysis techniques include immunohistochemistry (e.g. tissue microarrays), immunoblotting (e.g. enzyme-linked immunosorbent assays, reverse phase protein arrays), and high-throughput techniques using various chromatographic methods combined with mass spectrometry. Detailed review of these methods is beyond the scope of this review.

Bioinformatics approaches focus on the expression, activation and quantification of proteins; these aspects are analysed to delineate protein networks and

https://doi.org/10.1017/S0022215114002011 Published online by Cambridge University Press

signalling pathways, in a manner similar to gene expression. Work by Altelaar and colleagues describes some modern proteomics techniques.[10]

Integrative analyses of high-throughput technology data

In this review, we have summarised modern comprehensive approaches for evaluating the following: DNA structural changes; sequence alterations; gene expression levels; mechanisms of epigenetic regulation; and protein expression, activation and modulation. Table I summarises several of these methods, describing the utility, and advantages and disadvantages, of each. Note that this list is not comprehensive, as several other methods and variations on the listed methods exist.

The bioinformatics approaches used to glean information from each of these platforms individually is complex; however, a greater challenge is to develop methods to integrate these data appropriately, in order to gain a comprehensive understanding of the genetic and molecular underpinnings of human disease. Arguably, the widest applications of integrated analytic approaches have been in the field of cancer biology.

The Cancer Genome Atlas project embodies the modern comprehensive approach to understanding human cancer. The project is an initiative in the USA, supported jointly by the National Human Genome Research Institute and the National Cancer Institute, which aims to comprehensively profile the genetic and epigenetic alterations present in several human cancers. The project studies on glioblastoma, ovarian cancer, colorectal cancer, lung SCC, breast cancer and endometrial cancer have been completed.[11–15] The evaluation of several more cancers is underway. The project investigating head and neck SCC has been completed, and the first report is currently in preparation. Information regarding The Cancer Genome Atlas can be found at the project's website.[16] The project data for head and neck SCC are now publically available.

Methods for the integration of multiple high-throughput technology research platforms are actively being developed. Several programs have been designed to facilitate interfacing and amalgamation of these types of data. A recent review by Berger *et al.* lists several available tools.[17] The methods are not standardised and the approach depends largely on the data available and the experimental questions being asked.

In head and neck SCC, several groups have reported findings gleaned from data combined from two or more high-throughput platforms. A recent study used single nucleotide polymorphism arrays to determine copy number variations, and used gene expression arrays to evaluate 17 tumours without lymph node metastases and 20 lymph node metastases.[18] First, differentially expressed genes between the two groups were selected, with a false discovery rate of less than 5 per cent, leaving 1988 transcripts. The data from single nucleotide polymorphism analysis were then used to filter this list by selecting genes whose relative expression was correlated with regions of copy number loss or gain. This left a 95-transcript signature, which was then evaluated on an independent dataset of 133 patients. In a multivariate analysis, the signature was associated with decreases in overall survival and disease-specific survival. Furthermore, amplified genes in the signature were targeted in an *in vitro* system with a small interfering RNA library, which led to consistent growth suppression in multiple head and neck SCC cell lines.[18]

In a study from our own institution, DNA methylation arrays were performed on 118 head and neck SCC tumour–normal pairs.[9] Differentially methylated genes could distinguish between head and neck SCC and normal tissue. Methylation events specific to tumour subsite and to HPV status were also identified. The expression levels of three highly methylated genes seen in head and neck SCC tumors, *ZNF132, ZNF154*, and *UCHL-1*, were then evaluated in an available corresponding gene expression dataset, and these genes were found to be consistently down-regulated.[9]

Combining data from two platforms is challenging, but a truly integrated approach involves analysing gene networks from data on multiple platforms, exploring genetic alterations, epigenetic regulation and gene transcription in a comprehensive manner. The field of systems biology, as applied to cancer research, aims to use a holistic approach, simultaneously incorporating data from several disciplines in order to evaluate cell signalling networks and model systems, so as to better understand the determinants of cancer cell function, behaviour and phenotype.

Our understanding of how best to integrate multi-platform datasets is in its infancy, but current studies are forging ahead. One recent study by Pickering *et al.* used whole exome sequencing, single nucleotide polymorphism analysis, DNA methylation arrays, microRNA expression and gene expression arrays to evaluate 38 patients with oral SCC.[19] As collective events were considered from these platforms, it appeared that specific cell signalling pathways were consistently altered in distinct subsets of patients. It was noted that very few genes were altered in the majority of samples, even when multiple genomic and epigenomic events were considered. When multiple genomic or epigenomic events were assessed in related genes, an accumulation of disruptions in dominant pathways was revealed, including alterations in TP53-related pathways, notch signalling, cell cycle regulators and the *EGFR*/*PI3K* mitogenic signalling pathways. Furthermore, comprehensive evaluation across the patient set showed amplifications or activating mutations in oncogenes that could be targeted with existing molecular therapeutic agents in the majority of patients, though any single event was largely under-represented in the cohort.[19]

The Cancer Genome Atlas project is currently underway and, consistent with previous reports for other

https://doi.org/10.1017/S0022215114002011 Published online by Cambridge University Press

https://doi.org/10.1017/S0022215114002011 Published online by Cambridge University Press

TABLE I
SUMMARY OF SELECTED HIGH-THROUGHPUT METHODOLOGIES: UTILITY, ADVANTAGES AND DISADVANTAGES

Methodology	Utility	Advantages	Disadvantages
Genomics			
– Array CGH	Used to detect CNVs	Ideal for detecting amplifications or deletions	Cannot reliably detect translocations
– SNP arrays	Evaluates known SNPs in a DNA sample	Ideal for genotyping & GWAS studies	Can be used to detect CNVs & translocations, but ideally if compared against that of a matched normal sample
– Targeted DNA sequencing	Provides exact sequence of genes of interest	Can detect point mutations, insertions & deletions	Only evaluates targeted DNA sequence
– Exome sequencing	Provides sequence information of entire exome	Large amount of genomic information for relatively low cost	Structural variations can only be inferred from available sequence & coverage; no information on unsequenced DNA (97% of human genome)
– Whole genome sequencing	Provides sequence information of entire genome	Provides information for entire genome; can identify all types of mutations & structural variants	Remains costly in comparison with other methods
Transcriptome evaluation			
– Gene expression microarrays	Measures relative expression of thousands of transcribed genes	Evaluates expression of thousands of transcripts at relatively low cost	Does not provide sequence information as signal intensity for each probe is a surrogate for relative expression; generally only evaluates mRNA expression
– RNA-Seq	Provides sequence & relative expression of DNA transcribed to RNA	Provides sequence information of transcripts of interest, including mRNA, miRNA & lncRNA; can identify transcripts from gene fusions	High cost compared with gene expression; expression is inferred from sequence coverage, which can be influenced by factors other than abundance of RNA transcript
Epigenetics			
– DNA methylation	Measures relative methylation of thousands of CpG islands	Can identify methylation sites from thousands of CpGs at relatively low cost	Though promoter CpG methylation often results in decreased gene expression, consequences of DNA methylation at specific sites require validation
– MiRNA expression arrays	Measures relative expression of miRNAs	Evaluates miRNA expression at relatively low cost	MiRNA expression must be subsequently validated; functional importance evaluation requires follow-up studies
Proteomics			
– RPPA	Measures relative protein expression of phosphorylated & unphosphorylated proteins	High-throughput method for evaluating several potential signalling networks	Results depend on quality & affinity of antibodies, & stability of phosphorylation events; any findings must be validated, e.g. with Western blot analysis
– Tissue microarrays	Evaluates in situ expression of protein of interest across many samples	High-throughput method for evaluating protein of interest directly in tissues, allowing evaluation of level of expression & localisation of protein target	Results depend on quality & affinity of antibody; immunohistochemical evaluation is highly subjective
– MALDI-TOF MS	A laser is used to ionise a protein sample, & components are measured using 'time-of-flight' mass spectroscopy	Sample preparation is relatively simple, & samples can be evaluated in a high-throughput fashion	Limitations based on ionisation properties & molecular weight of specific proteins

CGH = comparative genome hybridisation; CNV = copy number variation; SNP = single nucleotide polymorphism; GWAS = genome-wide association study; mRNA = messenger RNA; miRNA = microRNA; lncRNA = long non-coding RNA; RPPA = reverse phase protein arrays; MALDI-TOF MS = matrix-assisted laser desorption/ionisation time-of-flight mass spectroscopy

TABLE II
PUBLICALLY AVAILABLE GENOMICS, EPIGENETICS AND MOLECULAR BIOLOGY DATABASES

Database type	Name & website	Description
Nucleotide databases	EMBL (European Molecular Biology Laboratory). In: http://www.ebi.ac.uk/	Genome project groups & individual authors are the major contributors to this database
	GenBank. In: http://www.ncbi.nlm.nih.gov/genbank/	The primary nucleic acid public database for which genomic data are directly submitted from individual laboratories & large scale sequencing projects, by & for the scientific community
Protein sequence databases	PIR (Protein Information Resource). In: http://pir.georgetown.edu/	Established to assist researchers in identification & interpretation of protein sequence information
	SwissProt. In: http://www.ebi.ac.uk/uniprot	An annotated protein sequence database
Structure database	Protein Data Bank. In: http://www.rcsb.org/pdb/home/home.do	Contains 3D structures of proteins, nucleic acids & some carbohydrates
Genome databases	Ensembl Genome. In: http://useast.ensembl.org/index.html	The Ensembl project produces genome databases on vertebrates & other eukaryotic species, freely available online
	UCSC Genome. In: http://genome.ucsc.edu/	Contains reference sequence & working draft assemblies for a large collection of genomes
Microarray databases	Array Express. In: http://www.ebi.ac.uk/arrayexpress/	A database of functional genomics experiments
	Gene Expression Omnibus. In: http://www.ncbi.nlm.nih.gov/geo/	A public functional genomics data repository supporting MIAME-compliant data submissions
Pathway databases	KEGG (Kyoto Encyclopedia of Genes & Genomes). In: http://www.genome.jp/kegg/pathway.html	A collection of manually drawn pathway maps
	BioCyc. In: http://biocyc.org/	A collection of 2988 pathway/genome databases
RNA databases	MiRBase. In: http://www.mirbase.org/	A searchable database of published miRNA sequences & annotations
	Ribosomal Database. In: http://rdp.cme.msu.edu/	The Ribosomal Database Project provides ribosome-related data & services, including online data analysis, & aligned & annotated RNA sequences
Specialised databases	HGMD® (Human Gene Mutation Database). In: http://www.hgmd.cf.ac.uk/ac/index.php	Represents an attempt to collate known (published) gene lesions responsible for human inherited disease
	OMIM (Online Mendelian Inheritance in Man). In: http://www.omim.org/	A comprehensive, authoritative compendium of human genes & genetic phenotypes, freely available & updated daily
	COSMIC (Catalogue Of Somatic Mutations In Cancer). In: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/	Designed to store & display somatic mutation information & related details; contains information relating to human cancers
	DbSNP (Single Nucleotide Polymorphism Database). In: http://www.ncbi.nlm.nih.gov/SNP/	A free public archive of genetic variations within & across different species
	Database of Genomic Variants. In: http://dgv.tcag.ca/dgv/app/home	Provides a comprehensive summary of structural variation in the human genome

3D = three-dimensional; UCSC = University of California Santa Cruz; MIAME = Minimum Information About a Microarray Experiment standard; miRNA = microRNA

https://doi.org/10.1017/S0022215114002011 Published online by Cambridge University Press

tumour sites, the head and neck SCC study should provide a comprehensive evaluation of this disease on a large cohort of patients. As the data are publically available, mining this resource using several bioinformatics strategies should greatly advance our understanding of head and neck SCC, and lead to improved therapeutic strategies.

Several open-source databases have been created to make the ever-increasing amounts of biological data available to the public via the internet. Table II presents several online resources that the authors have found useful for a wide range of data analyses, including genome evaluation, gene expression and cell signalling pathway analysis databases. These tools are invaluable for understanding results from research using high-throughput technology, and for exploring relationships between findings in both single- and multi-platform analyses.

Conclusion

The goal of this review was to introduce the reader to current high-throughput assays available for translational research in otolaryngology, with a focus on the analytical principles of the bioinformatics methods used to understand data derived in such studies. Progress in treating human disease in general will require close collaboration with experts in bioinformatics. Improved understanding of these concepts by clinicians and physician-scientists in our field will advance diagnosis and therapy for diseases of the head and neck.

References

1 Wineinger NE, Kennedy RE, Erickson SW, Wojczynski MK, Bruder CE, Tiwari HK. Statistical issues in the analysis of DNA copy number variations. *Int J Comput Biol Drug Des* 2008;**1**:368–95
2 Morris LG, Taylor BS, Bivona TG, Gong Y, Eng S, Brennan CW *et al*. Genomic dissection of the epidermal growth factor receptor (EGFR)/PI3K pathway reveals frequent deletion of the EGFR phosphatase PTPRS in head and neck cancers. *Proc Natl Acad Sci U S A* 2011;**108**:19024–9
3 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
4 Chen Y, Chen C. DNA copy number variation and loss of heterozygosity in relation to recurrence of and survival from head and neck squamous cell carcinoma: a review. *Head Neck* 2008;**30**:1361–83
5 Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 2007;**8**: 93–103
6 Childs G, Fazzari M, Kung G, Kawachi N, Brandwein-Gensler M, McLemore M *et al*. Low-level expression of microRNAs let-7d and miR-205 are prognostic markers of head and neck squamous cell carcinoma. *Am J Pathol* 2009;**174**:736–45
7 Harris T, Jimenez L, Kawachi N, Fan JB, Chen J, Belbin T *et al*. Low-level expression of miR-375 correlates with poor outcome and metastasis while altering the invasive properties of head and neck squamous cell carcinomas. *Am J Pathol* 2012;**180**:917–28
8 Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003;**33**(suppl):245–54
9 Lleras R, Smith RV, Adrien LR, Schlecht NF, Burk RD, Harris T *et al*. Unique DNA methylation loci distinguish anatomic site and HPV status in head and neck squamous cell carcinoma. *Clin Cancer Res* 2013;**19**:5444–55
10 Altelaar AF, Munoz J, Heck AJ. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* 2013;**14**:35–48
11 Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**:1061–8
12 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70
13 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**:609–15
14 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7
15 Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;**489**:519–25
16 The Cancer Genome Atlas. In: http://cancergenome.nih.gov [22 September 2013]
17 Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet* 2013;**14**:333–46
18 Xu C, Wang P, Liu Y, Zhang Y, Fan W, Upton MP *et al*. Integrative genomics in combination with RNA interference identifies prognostic and functionally relevant gene targets for oral squamous cell carcinoma. *PLoS Genet* 2013;**9**:e1003169
19 Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM *et al*. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discov* 2013;**3**:770–81

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/S0022215114002011 Published online by Cambridge University Press