Application of omic technologies in cancer research

Sarah Wagner, Graham R. Ball, A. Graham Pockley, Amanda K. Miles

John van Geest Cancer Research Centre, Nottingham Trent University, Nottingham, United Kingdom

Abstract

Understanding the biology of health and diseases such as cancer, generating insight into the triggers and potentiators of disease and the development of therapeutic approaches to counter and treat disease requires detailed interrogation of inherited genes, and the dynamic positioning of the transcriptome and proteome. In the last 10 years, significant technological developments and increases in sample throughput capabilities have led to a dramatic increase in the size and complexity of the datasets that can be generated. A key challenge now is to develop robust approaches for analysing and interpreting these, and converting data into biologically- and clinically-relevant information. Herein, we provide an overview of approaches for acquiring, integrating and interpreting complex datasets generated using multiple omic platforms, with a focus on the field of cancer research, and highlight key successful data handling and integration applications.

Introduction

It is well established that insertions and deletions within the BRCA1 and BRCA2 genes associate with a greatly enhanced risk for the development of breast cancer.[1-3] Altered expression of three breast cancer associated receptors (ER, PR, HER2) are also present in breast cancer,[4,5] and these can act as targets for therapeutics such as Trastuzumab and Trastuzumab emtansine which bind to HER2.[6,7] Another area which has attracted a great deal of interest in the recent past relates to microRNA (miRNA), which are small non-coding RNA molecules of about 22 nucleotides in length. An example of this are the miR-141 and miR-375, which have been detected at significantly greater levels in metastatic compared to non-metastatic prostate cancer.[8,9] Although investigating the transcriptome can provide some valuable insight into biology (*phenotype*), the so-called *machinery* driving the biology of the cell and organism are the proteins (the *proteome*). The link between transcriptional activity and the proteome is not necessarily direct, and so the definition of the biology and the identification of drivers of disease and therapeutic resistance require multiple *omics*-based approaches.

Current challenges in omics research

Understanding biological complexity by generating and analysing low-dimensional datasets is limited. Improving our understanding of biological systems in health and disease therefore requires robust approaches for mining, integrating and interpreting the large and complex multiple biological (omics) datasets that can now be readily generated. These approaches need to identify indicative changes within the given data and pinpoint alterations to a subset of molecules. Although the literature on the use of big data sets and multi-omics integration has progressively increased in recent years, the published approaches vary considerably and are typically individually tailored to each experimental question (Figure 1). Notwithstanding this, the power of these approaches can be significant, in that they have the capacity to provide unprecedented insight into the status of key pathways[10] and underpin the discovery of novel biomarkers of disease and therapeutic resistance across a profile of diseases, including cancer.[11]

Correspondence: Amanda K. Miles, John van Geest Cancer Research Centre, Nottingham Trent University, Nottingham NG11 8NS, United Kingdom.
Tel.: +44.115.848.3754.
E-mail: amanda.miles@ntu.ac.uk

Key words: Cancer; Omics; Data mining; Integration.

Contributions: SW was involved in the conception and design of the review and wrote the manuscript. GRB and AGP provided intellectual input and contributed to the drafting and finalising of the manuscript. AKM was involved in the conception of the review, provided intellectual input and contributed to the drafting and finalising of the manuscript.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: this work was funded by the John and Lucille van Geest Foundation and a Nottingham Trent University Vice Chancellor PhD Student Bursary to SW.

Received for publication: 6 November 2017.
Revision received: 6 March 2018
Accepted for publication: 28 March 2018

Translational Medicine Reports 2018; 2:7176
doi:10.4081/tmr.7176

Sources of omics data

Genomics

The completion of the first sequenced human genome[12] triggered an era of major developments in the understanding of the living world. The genome represents the building blocks of a system. The DNA, and more specifically the nucleotides, code for single genes, which are translated (*via* the *transcriptome*) into functional proteins. However, not all genetic information is translated into proteins, and research into the role of non-coding regions in disease is coming under greater scrutiny.[13] Although non-coding RNA was initially disregarded as being *data junk*, it is now known to provide insight into a hidden layer of internal signalling pathways that orchestrate highly specific nucleic acid recognition and RNA modifications.[14,15] Wang *et al.* performed a meta-analysis on the impact of the long non-coding RNA SPRY4-IT1 on cancer prognosis. The results highlighted the significant association between increased levels of SPRY4-IT1 expression and overall survival and the development of metastasis in patients with gastric and ovarian cancer.[16] Sequencing of the whole genome or targeted sequencing of specific genes can highlight alterations, such as the well-known example of the BRCA1[17] and BRCA2[18] genes.

Epigenomics

The Epigenome[19] reflects chemical changes that influence DNA and histones. A common type of epigenetic modification is the hypo- or hypermethylation of promotor regions, which can be caused through the aberrant behaviour of DNA methyltransferases. Such alterations have been shown to be present in many cancers, including breast cancer[20] and cholangiocarcinoma.[21] Hypermethylation of the paired-like homeodomain transcription factor 2 gene has been shown to be a prognostic indicator of disease recurrence in prostate cancer.[22] The Epigenome can also be influenced by altered chromatin regulation resulting from histone modifications.[23] Methylation of histones can alter the response and regulate the invasiveness of cells, which has been demonstrated in a study on the protein arginine methyltransferase 5 complexed with MEP50/WDR77.[24]

Transcriptomics

The transcriptome reflects the complete set of ribonucleic acid transcripts that are present in a sample at a defined time point. Although the analysis of the transcriptome can be focussed solely on the coding RNAs, the inclusion of non-coding RNAs presents additional crucial information. Commonly used gene expression microarrays, and more recently developed RNA-sequencing approaches, offer the ability to analyse the complete transcriptome. Microarrays are commonly used because of their affordability and robustness, however the value of this approach is limited to an *a priori* knowledge of genes.

Microarray analysis was successfully used for the screening of various biological specimens, including clinical material. Lapointe *et al*.[25] used microarray technology for the screening of ~26000 genes in ~100 clinically derived samples of prostate cancer and matching healthy prostate tissue. This study firstly highlighted the differences in genetic profiles between healthy and diseased tissue based on hierarchical clustering and secondly the identification of genetic profiles associated with aggressiveness. Iorio *et al*. applied a similar approach by screening miRNAs of healthy and diseased breast tissue, which enabled the clustering of patients into their respective groups based on the miRNA profiles. Such approaches enable the identification of strongly influential genes for the classification and the discovery of potential underlying pathways.

In contrast, RNA-sequencing offers the advantage of analysing RNA independently of knowledge relating to the sequence, and offers a larger dynamic range in the measurement of gene expression. It is also possible to analyse samples for which no full genome sequence exists.[26] RNA-sequencing platforms can also be used to focus on specific elements of the transcriptome, such as miRNAs, and can enable a closer interrogation of actively translated genes by screening transcripts that are bound to ribosomes using ribosome profiling.[27] Recent advances have enabled the analysis of the transcriptome of single cells,[28] which can improve the understanding through selection of unique entities within a heterogeneous cell population. The development of RNA-sequencing approaches for the measurement of whole transcriptomes[26] enables users to analyse a sample for approximately £500. A novel development enables the user to analyse the transcriptome of single cells,[29] which can be applied in a broad field of research such as the study of circulating tumour cells[28] or induced pluripotent stem cells.[30] Ren *et al*.[31] used the RNA-seq technology for the profiling of Chinese prostate cancer patients and healthy tissue to elucidate underlying variations between races. It enabled them to identify differences in the profiles of gene-fusions, somatic mutations, alternative splicing and the expression of non-coding RNA (ncRNA). Furthermore, the analysis highlighted differential expression of ncRNA between the specimens. An intermediate option to large-scale whole transcriptome analysis is the nanoString nCounter™ gene profiling platform.[32-34] This platform can simultaneously analyse up to 800 genes/miRNAs and around 40 proteins within a single sample. In contrast to conventional gene array-based approaches, the nanoString technology does not require amplification and so the readout of gene expression is direct and reproducible. This approach allows the use of lower quality RNA than that which is required for gene array, and from sources such as fresh tissue/cells and formalin-fixed, paraffin embedded (FFPE) tissue. Concentrations between 25 and 300 ng of RNA, and material isolated from a single cell is sufficient for a complete analysis.[35] This approach therefore provides a valuable analytical option that delivers results comparable to matched fresh tissue, which is also suitable for archived and partially degraded tissue from FFPE material.[33] Cascione *et al*. used the nanoString nCounter™ platform for the integrated analysis of gene and miRNA expression of triple negative breast cancer (TBNC), and this resulted in the discovery of miRNA expression signatures that can describe phenotypic subtypes within TNBC.[36] A novel promising method for the targeted study of single cells within a tissue sample is offered by nanoString's Digital Spatial Profiling (DSP) technology. The DSP technology, which will be officially released in 2018, enables the quantitative analysis of up to 800 RNA and protein targets from specified regions of interest within the analysed tissue on a single cell basis. Importantly, this approach is non-destructive for the analysed tissue as it is solely based on the use of UV-light for the quantitative measurement of the analytes (https://www.nanostring.com/scientific-content/technology-overview/digital-spatial-profiling-technology).

Proteomics

The proteome represents all proteins within the system of interest at the time

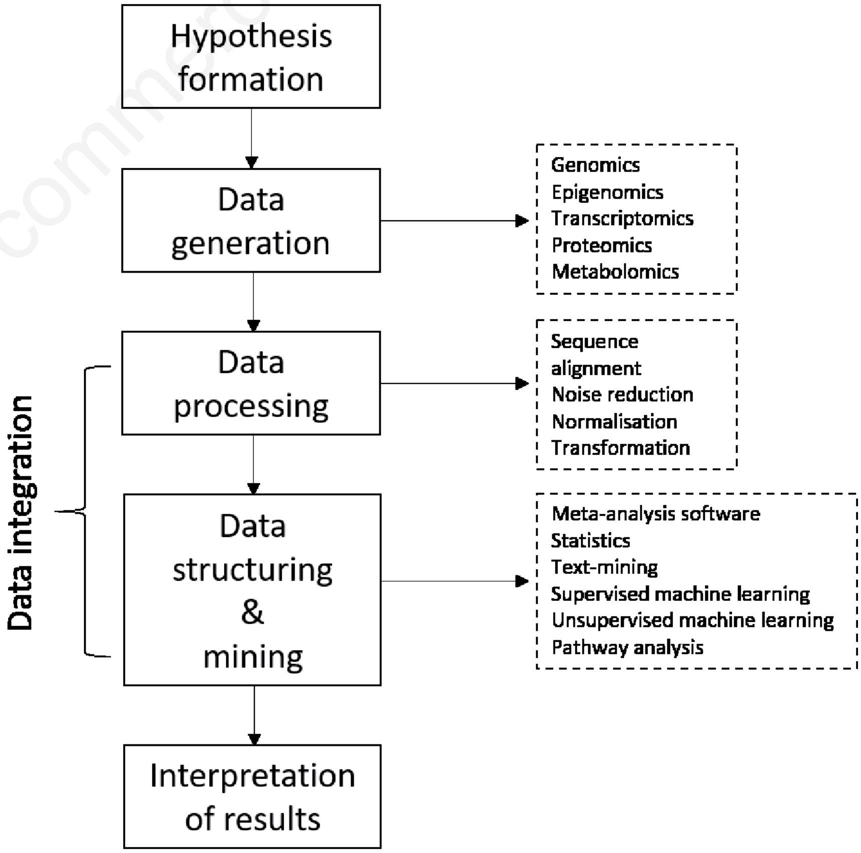

Figure 1. Schematic representation of an omics pipeline.

point of sample collection. Although the transcriptome can provide information regarding the potential proteome, there is no full correlation between the transcriptome and resulting proteome. Factors such as half-life of RNA and protein and post-translational modifications can lead to variations.[37] The proteome is a highly complex system and the concentrations of the proteins within the proteome span a wide order of magnitude. Measurement of the proteome in human material such as plasma presents significant obstacles, as concentrations of proteins can range by up to 10 orders of magnitude and the high concentration of highly abundant proteins increases the difficulty in the detection of other, low abundance proteins.[38] It is possible to focus on individual compartments of the cell by utilising additional fractionation techniques prior to proteomic analysis. Ji *et al*. utilised the analysis of secreted exosomes from primary and metastatic colorectal cancer cell lines for the discovery of differentially expressed proteins associated with metastasis.[39] In recent years, the development of data-independent mass spectrometry,[40] as is utilised in the SWATH™-based approach (ABSciex), has enabled the quantitation of large numbers of proteins from relatively small amounts of starting material. This approach enables up to 3500 routine protein identifications and quantifications within a short period of acquisition. The development of data-independent acquisition in tandem mass spectrometry approaches[40] has represented a milestone in proteome research and resulted in the discovery and validation of biomarkers such as the Apolipoprotein A-IV in ovarian cancer,[41] Carbonic Anhydrase 2 in nasopharyngeal carcinomas,[42] N-acylethanolamine Acid Amidase and Protein Tyrosine Kinase 7 in aggressive prostate cancer.[43] Multiple studies analysing the proteome of prostate cancer in comparison with healthy or benign tissue have also enabled the discovery of functional proteins altered in different disease stages.[44,45]

Metabolomics

The products of biochemical reactions, so called metabolites, comprise the metabolome. Metabolites are small molecules, normally below 1500 Da, and alterations in the metabolism of cells and therefore the metabolites produced, can be indicative of disease states and/or survival.[46] Budczies *et al.* analysed the metabolic profiles of different breast cancer stages and discovered an altered metabolite ratio of cytidine-5-monophosphate/pentadecanoic acid, which could show a strong discriminatory ability between breast cancer and healthy tissue.[47] A further study by the same group revealed that beta-alanine could distinguish between ER+ and ER- breast cancer.[48]

The most commonly applied ‘omics approaches and the latest methods/technologies used in each of these are summarised below and in Table 1.

Multi-omics datasets

Although publically-available datasets can be a good source of information, most currently available datasets are derived from genomic and transcriptomic analyses. Platforms such as ArrayExpress,[49] PRIDE[50] and NCBI Gene Expression Omnibus[51] offer free to access datasets of published work. However, matched multiple ‘omic data from a single study are rare. Matched data are important as they generate a closer insight into the relationship between transcriptional and translational activity and reduce the variability and noise within the data.

The effective use of data mining and data integration requires a clear understanding of the study aims. The integration of transcriptomic and proteomic data can improve the discovery of biomarkers[52] and identify genes/proteins that are altered in pathways. The integration of genomic and transcriptomic data is more suitable for defining categories of disease risk and has been extensively applied in the context of cancer.[53,54]

A critical factor that can define the quality of results obtained using any scientific approach is the choice of sample material to be used (Table 2). More robust results can be achieved by generating *in-house* matching datasets that are derived from the same starting material which has been subjected to identical experimental conditions. Although such an approach decreases variability within a study, it poses significant operational challenges, can be more costly and restricts sample throughput, availability and dataset size.

Defined and validated cell lines are commonly used for the discovery of potential biomarkers[11] or drug targets, as their use

Table 1. Representation of omics sources and approaches and the information they can contribute to a multi-omics study.

Information	Single nucleotide variants Insertions/deletions Copy number variations Large structural variants	Histone-modifications DNA-methylation Chromatin-assembly DNA-protein binding sequences	Gene expression Splice variants Novel transcripts Small/non-coding RNA	Protein Expression Post-translational modifications Isoforms	Small molecules (<1500 Da) Endogenous metabolome Exogenous metabolome
Omics	Genome/Epigenome		Transcriptome	Proteome	Metabolome
Methods	Whole-genome sequencing Exome sequencing DNA-microarray Sanger-sequencing NGS-sequencing Methylation arrays		NGS-Sequencing Microarray PCR-Chip nanoString analysis	Mass spectrometry Protein arrays Nanostring analysis	Mass spectrometry NRM spectrometry Chromatography

Table 2. Comparison of factors between cell line and clinical material.

	Cell line material	Clinical specimens
Availability of sample material	Good	Limited
Analysis of sample material	Easy	Difficult (complexity and biological variation)
Sample numbers for omics approches	≤15	≥100
Transferability of results into clinical application	Difficult (additional validation necessary)	Good

is reasonably cost effective and are easier to access than control and patient-derived clinical material. Such systems are also highly controlled and lower sample numbers can be sufficient for an integrative study. The main drawback of using cell lines is that these systems are often highly artificial. Commonly, cell lines can lose stability after multiple passages and this can significantly affect their phenotype and hence response to biological stimulants. The clinical relevance of cell line-based studies can therefore be questionable.

Clinical material is valuable for the identification and validation of prognostic and diagnostic biomarkers.[52] However, they have a higher complexity due to the heterogeneity of the sample material and disease state, and so their analysis and the interpretation of the data can be challenging. Major drawbacks in the use of clinical material include the limited availability, the effect of storage on the usability, the generation of highly complex data, the risk of incomplete annotation of patient clinical information (essential for correlating findings to clinical and pathological outcomes) and the high sample numbers required due to biological and disease heterogeneity.

Data processing and mining

The selection process with regards to the most appropriate sample material and analysis method is crucial for ensuring optimal outputs. It is important to consider which data processing approaches will be applied and what tools are necessary to most successfully use these approaches. Most ‘omics platforms provide analysis tools for the initial processing of the generated data, such as BaseSpace (https://basespace.illumina.com/home/index), Oneomics (https://sciex.com/applications/life-science-research/oneomics) and nSolver™ (https://www.nanostring. com/products/analysis-software/nsolver). The alignment of generated data to a reference genome or proteome enables the acquisition of comprehensive protein and gene expression and/or fold change data. These platforms also allow the confidence level of the data subsets to be adjusted in order to more stringently filter the data obtained. Despite this, additional in depth bioinformatic approaches should be considered since these offer great advantages in their application, such as literature-independent data processing and the recognition of underlying patterns in data with thousands of inputs. However, some data processing and mining approaches are demanding on computational power and/or on the number of replicates analysed and these factors must be considered at the start of the study design.

Data must be prepared and normalised prior to further analysis. In some cases, it might be useful to transform the data into identical formats (*e.g.*, $\log_2$ values) so that comparisons across datasets are feasible. It is important to be aware of the variations between gene and protein names and a uniform labelling system needs to be applied.[55,56] This is followed by a filtering process, which obtains relevant information and reduces the dimensionality of the data. Various data-structuring and data-mining approaches, which are commonly used for single omics studies are available, and these can be adjusted for integrative studies. Machine-learning approaches can be used to reduce and filter the data, thereby allowing significant factors within the data and potential underlying clusters and patterns to be identified.

Stringent statistical analysis reduces the data to a discreet selection of significant drivers and this is more easily achieved when using data with low noise. Text mining[57] can also be used, with the review of currently published research and the extraction of relevant information such as key genes within a pathway enabling further research. Commercially available reference mining tools with a curated database, such as Metacore (https://clarivate.com/products/metacore/) or iPathwayGuide (http://www.advaitabio.com/ipathwayguide) can analyse and integrate multi-omics datasets, thereby highlighting correlations and pathways between the data sets. However, these tools are based on published literature and are therefore limited to the interrogation of genes/proteins whose function has been previously described.

Various literature-independent approaches that can be supervised or unsupervised in their nature are also available. Such methods can structure and reduce the data and filter for significant changes or dominant patterns within the system. Although these approaches are commonly applied to single ‘omics data sets, an integrated approach is also theoretically possible by combining multiple ‘omics data.[58]

Unsupervised learning approaches work without predetermined grouping of the data. Here, hidden structures or clusters (co-expressed features) and potential interactions between factors and networks within the dataset are determined. Commonly used examples are Principal Component Analysis (PCA)[59-61] and hierarchical and k-means clustering.[62] However, it should be highlighted that these applications do not directly lead to a feature selection, but result in the structuring and clustering of the presented data. These applications can also function as processing steps before the application of further data-mining approaches and for quality control, *e.g.* the clustering of classed samples. PCA enables the reduction of high dimensionality by maintaining the inherent variability within the samples.[63] It can also reveal grouping within samples, be used as an internal control of biological replicates and has successfully contributed to the definition of oncogenic pathway signatures in human cancers.[64]

Hierarchical clustering groups together input factors that show a low distance/higher similarity to each other compared to other given features. Initially, all given features represent their own cluster and are incrementally clustered together by their commonalities. This continues until all clusters have been merged and the results are commonly represented in a dendrogram. Hou *et al*. generated copy number variants and transcriptomic data sets on single cell populations and applied hierarchical clustering and PCA analysis for the definition of subpopulations.[65] Hierarchical clustering was also used for the comparison of cDNA expression in prostate cancer and healthy tissue, resulting in clear clustering groups of healthy and diseased material and highlighting subclasses within the prostate tumour samples.[25] The clustering approach using k-means clustering algorithms differs in the way the clusters are created in comparison to hierarchical clustering. The k-means clustering approach is based on predefined numbers of desired clusters, termed *k*. The algorithm then creates *k* centroids to which each feature is assigned. After each assignment, the centroids are updated and the process ends after each feature is assigned to a group. Shen *et al*. applied hierarchical and k-means clustering analysis for the discovery of three distinct subclasses of colon cancers based on their genetic and epigenetic profiles.[66] These three groups were confirmed through the application of a novel dataset to k-means clustering using a subset of genetic and epigenetic markers.[67]

Supervised learning approaches can be applied to categorised data, such as *healthy* or *diseased*. These groups can be previously known through the origin of the data or generated through unsupervised learning approaches. The main aim is to accurately predict the group to which a novel signature belongs. The algorithm uses a training set taken from the data to find a solution for the given question(s) based on presented information. This method enables it to correct the predictions and to improve the output. The generated results are tested for their suitability and adjusted if needed, until either a previously defined error rate is reached or the lowest error rate within a predefined set up is selected. Commonly used examples for supervised learning approaches are Support Vector Machines (SVM)[68,69] Random Forest (RF)[69] and artificial neural networks (ANN).[70,71]

SVMs are commonly applied for classification analyses,[72] which are primarily applied for a two-group classification. The algorithm generates an optimal hyperplane to divide the two classes. Examples of preclassified input data are presented to the algorithm, which learns from this, and can in the future attribute novel samples to the given categories.[73] Carlsson *et al*. used a SVM analysis for the classification of metastatic breast cancer based on serum profiles.[74]

RF[75] use an algorithm for making classifications that creates multiple decision trees based on subsets of the given data. These *trees* build together the random forests, which together produce predictions and insight into the given data. RF shows comparable qualities and results when applied to microarray data compared to other classification approaches, such as SVM.[76] This method was used for tumour classification of renal cell carcinoma and was shown to categorise tissue samples into clear and non-clear cell tumours.[77]

ANNs are based on a machine learning approach that can deal with noisy, non-linear data and can highlight markers of interest relating to the underlying question within a given sample set.[78] This method was successfully applied for the discovery of altered miRNAs between luminal A breast cancer patients and healthy controls.[79]

Improvements through data integration?

It is undeniable that each of the previously discussed ‘omics levels and data mining approaches already offer, on their own, a large amount of crucial information. However, the data obtained represent only a proportion of the true biological complexity of a living system. For this reason, the ability to integrate multiple ‘omics levels offers great scope for the improved understanding of such complex systems. Although such approaches are varied and are normally tailored for individual studies, as mentioned previously, most data-mining approaches can be applied to both single and multi-omics studies.

Wang *et al*. integrated the tissue transcriptome with the secretomes of two cell lines. After initial data reduction based on significant changes, integration of the secretome and transcriptome datasets led to the selection of 35 key drivers, of which the importin subunit alpha 2 was defined as a potential biomarker for non-small cell lung cancer.[80] Ou *et al*. generated proteomic and transcriptomic data from cell lines using 2DE/MS and gene expression microarray analysis. A comparison between the proteomic and transcriptomic data generated a subset of concordant markers, which were then validated in tissue mRNA and tissue microarrays.[81]

A different study highlighted the ability of characterising different subtypes of diseases, in this case hyperdiploid and non-hyperdiploid multiple myeloma, using multi-omics approaches. The integrated study elucidated differences of disease characteristics of these subtypes is not only based on early promoting events, but also manifests in differentially regulated pathway activity and alterations in genetic architecture.[82]

These two examples highlight the possibilities such data integration harbours for the understanding of complex diseases such as cancer. Furthermore, it is widely accepted that certain ‘omics levels are more closely related, and consequently may be easier to integrate and understand than others. The direct integration of data does not focus solely on correlations between the datasets, but often has the purpose of filling the potential gaps within the data and highlighting interactions between the various 'omics levels[83,84] and the impact of alterations in one 'omics level on the expression of others.[82]

Conclusions

Currently, there is no *one size fits all* approach for the integration of multi-omics data. Each approach has its advantages and suitability for certain questions, but also presents limitations for others. For this reason, it is crucial to carefully select the integration methods based on their suitability for the research performed. Some unavoidable problems are faced by all researchers and the research question asked is irrelevant in this case. One such example of this are the differences in gene/protein annotations between ‘omics platforms that make the integration/combination of data difficult. Drastic improvements have been made in the scientific equipment available over the last 10 years, especially in the field of sequencing technologies. Despite these advances, obstacles which are mainly attributed to the data gap in ‘omics platforms and especially in the generation of protein datasets remain.

In conclusion, combined ‘omics studies contribute more than the sum of their individual components. This approach gives the scientist increasingly deeper insights into the complex function of biological systems, disease states and behaviour. The use of a suitable integration approach can improve the understanding of the phenotypic representation of disease states and cell behaviour and using this information, tailored studies can be designed that will lead to the development of novel drugs which target the identified pathways and/or the use of novel biomarkers in monitoring/diagnosing disease. Ultimately, this could lead to newly developed disease management approaches and improvements in patient care.

References

1. King MC, Marks JH, Mandell JB, New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 2003;302:643-6.
2. Patel KJ, Yu VP, Lee H, et al. Involvement of Brca2 in DNA repair. Mol Cell 1998;1:347-57.
3. Wang RH, Yu H, Deng CX. A requirement for breast-cancer-associated gene 1 (BRCA1) in the spindle checkpoint. Proc Natl Acad Sci USA 2004;101:17108-13.
4. Tecalco-Cruz AC, Ramirez-Jarquin JO. Mechanisms that increase stability of estrogen receptor alpha in breast cancer. Clin Breast Cancer 2017;17:1-10.
5. Deroo BJ, Korach KS. Estrogen receptors and human disease. J Clin Invest 2006;116:561-70.
6. von Minckwitz G, Procter M, de Azambuja E, et al. Adjuvant Pertuzumab and Trastuzumab in early HER2-positive breast cancer. N Engl J Med 2017;377:122-31.
7. Verma S, Miles D, Gianni L, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. N Engl J Med 2012;367:1783-91.
8. Bryant RJ, Pawlowski T, Catto JW, et al. Changes in circulating microRNA levels associated with prostate cancer. Br J Cancer 2012;106:768-74.
9. Nguyen HC, Xie W, Yang M, et al. Expression differences of circulating microRNAs in metastatic castration resistant prostate cancer and low-risk, localized prostate cancer. Prostate 2013;73:346-54.

10. Jayaram S, Gupta MK, Raju R, et al. Multi-omics data integration and mapping of altered kinases to pathways reveal gonadotropin hormone signaling in glioblastoma. OMICS 2016;20:736-46.
11. Pavlou MP, Dimitromanolakis A, Martinez-Morillo E, et al. Integrating meta-analysis of microarray data and targeted proteomics for biomarker identification: application in breast cancer. J Proteome Res 2014;13:2897-909.
12. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304-51.
13. Khurana E, Fu Y, Chakravarty D, et al. Role of non-coding sequence variants in cancer. Nat Rev Genet 2016;17:93-108.
14. Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2001;2:919-29.
15. Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet 2006;15:R17-29.
16. Wang M, Dong X, Feng Y, et al. Prognostic role of the long non-coding RNA, SPRY4 Intronic Transcript 1, in patients with cancer: a meta-analysis. Oncotarget 2017;8:33713-24.
17. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66-71.
18. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789-92.
19. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell 2007;128:669-81.
20. Tryndyak V, Kovalchuk O, Pogribny IP. Identification of differentially methylated sites within unmethylated DNA domains in normal and cancer cells. Anal Biochem 2006;356:202-7.
21. Braconi C, Huang N, Patel T. MicroRNA-dependent regulation of DNA methyltransferase-1 and tumor suppressor gene expression by interleukin-6 in human malignant cholangiocytes. Hepatology 2010;51:881-90.
22. Uhl B, Gevensleben H, Tolkach Y, et al. PITX2 DNA methylation as biomarker for individualized risk assessment of prostate cancer in core biopsies. J Mol Diagn 2017;19:107-14.
23. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res 2011;21:381-95.
24. Chen H, Lorton B, Gupta V, Shechter D. A TGFbeta-PRMT5-MEP50 axis regulates cancer cell invasion through histone H3 and H4 arginine methylation coupled transcriptional activation and repression. Oncogene 2017;36:373-86.
25. Lapointe J, Li C, Higgins JP, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA 2004;101:811-6.
26. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.
27. Ingolia NT, Brar GA, Rouskin S, et al. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 2012;7:1534-50.
28. Ramskold D, Luo S, Wang YC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 2012;30:777-82.
29. Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods 2011;8:S6-11.
30. Zhang J, Chu LF, Hou Z, et al. Functional characterization of human pluripotent stem cell-derived arterial endothelial cells. Proc Natl Acad Sci USA 2017;114:E6072-8.
31. Ren S, Peng Z, Mao J, et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res 2012;22:806.
32. Malkov VA, Serikawa KA, Balantac N, et al. Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter™ Assay System. BMC Res Note 2009;2:80.
33. Veldman-Jones MH, Brant R, Rooney C, et al. Evaluating robustness and sensitivity of the nanostring technologies ncounter platform to enable multiplexed gene expression analysis of clinical samples. Cancer Res 2015;75:2587-93.
34. Armstrong DA, Green BB, Seigne JD, et al. MicroRNA molecular profiling from matched tumor and bio-fluids in bladder cancer. Mol Canc 2015;14:194.
35. Guo G, Luc S, Marco E, et al. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. Cell Stem Cell 2013;13:492-505.
36. Cascione L, Gasparini P, Lovat F, et al. Integrated microRNA and mRNA signatures associated with survival in triple negative breast cancer. PLoS One 2013;8:e55910.
37. Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett 2009;583:3966-73.
38. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 2002;1:845-67.
39. Ji H, Greening DW, Barnes TW, et al. Proteome profiling of exosomes derived from human primary and metastatic colorectal cancer cells reveal differential expression of key metastatic factors and signal transduction components. Proteomics 2013;13:1672-86.
40. Gillet LC, Navarro P, Tate S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 2012;11:O111.016717.
41. Rauniyar N, Peng G, Lam TT. Data-Independent acquisition and parallel reaction monitoring mass spectrometry identification of serum biomarkers for ovarian cancer. Biomark Insights 2017;12:1177271917710948.
42. Luo M, Sun W, Wu C, et al. High pretreatment serum gamma-glutamyl transpeptidase predicts an inferior outcome in nasopharyngeal carcinoma. Oncotarget 2017;8:67651-62.
43. Liu Y, Chen J, Sethi A, et al. Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. Mol Cell Proteomics 2014;13:1753-68.
44. Sardana G, Dowell B, Diamandis EP. Emerging biomarkers for the diagnosis and prognosis of prostate cancer. Clin Chem 2008;54:1951-60.
45. Nakashima J, Tachibana M, Horiguchi Y, et al. Serum interleukin 6 as a prognostic factor in patients with prostate cancer. Clin Cancer Res 2000;6:2702-6.
46. Yuan C, Clish CB, Wu C, et al. Circulating metabolites and survival among patients with pancreatic cancer. J Natl Cancer Inst 2016;108:djv409.
47. Budczies J, Denkert C, Muller BM, et al. Remodeling of central metabolism in invasive breast cancer compared to normal breast tissue - a GC-TOFMS based metabolomics study. BMC Genomics 2012;13:334.
48. Budczies J, Brockmoller SF, Muller BM, et al. Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism. J Proteomics 2013;94:279-88.
49. Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress-a public repository for microarray gene expression data at the

EBI. Nucleic Acids Res 2003;31:68-71.

50. Martens L, Hermjakob H, Jones P, et al. PRIDE: the proteomics identifications database. Proteomics 2005;5:3537-45.

51. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucl Acid Res 2002;30:207-10.

52. Yu SY, Hong LC, Feng J, et al. Integrative proteomics and transcriptomics identify novel invasive-related biomarkers of non-functioning pituitary adenomas. Tumour Biol 2016;37:8923-30.

53. Roychowdhury S, Chinnaiyan AM. Translating cancer genomes and transcriptomes for precision oncology. CA Cancer J Clin 2016;66:75-88.

54. Ren S, Wei GH, Liu D, et al. Whole-genome and transcriptome sequencing of prostate cancer identify new genetic alterations driving disease progression. Eur Urol 2017;S0302-2838.

55. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44-57.

56. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucl Acid Res 2017;45:D158-69.

57. Zhu F, Patumcharoenpol P, Zhang C, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform 2013;46:200-11.

58. Li L, Tang H, Wu Z, et al. Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 2004;32:71-83.

59. Yao F, Coquery J, Le Cao KA. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinform 2012;13:24.

60. Analysis of breast cancer using data mining & statistical techniques. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on: IEEE; 2005.

61. Jiang P, Liu XS. Big data mining yields novel insights on cancer. Nat Genet 2015;47:103-4.

62. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Trans Knowled Data Eng 2004;16:1370-86.

63. Ringner M. What is principal component analysis? Nat Biotechnol 2008;26:303-4.

64. Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353-7.

65. Hou Y, Guo H, Cao C, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res 2016;26:304-19.

66. Shen L, Toyota M, Kondo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci USA 2007;104:18654-9.

67. Ahn JB, Chung WB, Maeda O, et al. DNA methylation predicts recurrence from resected stage III proximal colon cancer. Cancer 2011;117:1847-54.

68. Burges CJ. A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discovery 1998;2:121-67.

69. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinform 2008;9:319.

70. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. Neural Networks 2006;19:408-15.

71. Lisboa PJ. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks 2002;15:11-39.

72. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389-422.

73. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565-7.

74. Carlsson A, Wingren C, Ingvarsson J, et al. Serum proteome profiling of metastatic breast cancer using recombinant antibody microarrays. Eur J Canc 2008;44:472-80.

75. Boulesteix A, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisc Rev: Data Mining Knowledge Discovery 2012;2:493-507.

76. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinforma 2006;7:3.

77. Shi T, Seligson D, Belldegrun AS, et al. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 2005;18:547-57.

78. Lancashire LJ, Powe DG, Reis-Filho JS, et al. A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. Breast Canc Res Treat 2010;120:83-93.

79. McDermott AM, Miller N, Wall D, et al. Identification and validation of oncologic miRNA biomarkers for luminal A-like breast cancer. PLoS One 2014;9:e87032.

80. Wang CI, Wang CL, Wang CW, et al. Importin subunit alpha-2 is identified as a potential biomarker for non-small cell lung cancer by integration of the cancer cell secretome and tissue transcriptome. Int J Canc 2011;128:2364-72.

81. Ou K, Yu K, Kesuma D, et al. Novel breast cancer biomarkers identified by integrative proteomic and gene expression mapping. J Proteome Res 2008;7:1518-28.

82. Di Martino MT, Guzzi PH, Caracciolo D, et al. Integrated analysis of microRNAs, transcription factors and target genes expression discloses a specific molecular architecture of hyperdiploid multiple myeloma. Oncotarget 2015;6:19132-47.

83. Karagoz K, Sinha R, Arga KY. Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. Omics: J Integr Biol 2015;19:115-30.

84. Schmid A, Blank LM. Systems biology: Hypothesis-driven omics integration. Nature Chem Biol 2010;6:485-7.