OMICS APPROACHES AND INTEGRATED BIOINFORMATIC ANALYSIS FOR THE IDENTIFICATION OF BIOMARKERS

Sabrina Tait
Reference Centre for Gender Medicine, Istituto Superiore di Sanità, Rome, Italy

Introduction

In the last decades, the advent of omics technologies dramatically increased the understanding of signaling pathways and networks of co-regulated genes, proteins and metabolites, thanks to the availability of several omics applications, including genomics, transcriptomics, epigenomics, proteomics, metabolomics and, more recently, miRNAomics, which increased the possibility to cover all the aspects of cellular biology. Indeed, each omics technology provides a piece of information contributing to the understanding of the cellular processes in physiological conditions but also in different diseases or under the influence of exogenous stimuli like chemicals, physical agents, microorganisms or life style.

Biomarker discovery

One of the main contributions that omics approaches can provide is the identification of relevant biomarkers. According to the World Health Organization definition, a biomarker is "any substance, structure or process that can be measured in the body or its products and influence or predict the incidence or outcome of the disease" (UNEP, 2001). Different biomarkers could be discovered by the analysis of clinical or experimental samples, e.g., diagnostic, prognostic or predictive biomarkers that are differently able to discriminate, respectively, specific pathologies, the progress of a disease, the response to a drug. Other biomarkers may be also identified in response to environmental stimuli (biomarkers of effect) or as critical stepping points during development or in particular life stages.

Different omics are used to investigate physiological states, diseases, exposure effects, infections, and all affections of clinical relevance for human health.

Genome-Wide association studies (GWAS) are applied to the discovery of genetic risk factors related to a particular disease by comparing the genetic variants in large population groups such as Single Nucleotide Polymorphisms (SNPs) or Copy Number Variants (CNVs), thus helping in the identification of susceptible subpopulation as well as of relevant markers of disease onset/progression (Tam *et al.*, 2019).

Epigenomics is used to identify sites of gene expression regulation, such as DNA methylation sites and histone modifications, which may be altered by pathological conditions or environmental factors (Stricker *et al.*, 2016). Since these alterations are heritable when occurring in germ cells, the identification of altered epigenomic sites may not only elucidate mechanisms underlying disease regulation but also be relevant for the generations to come (Bošković & Rando, 2018).

Transcriptomics has been the first approach that provided fundamental insights on how signaling pathways proceed into cells and to which extent minimal derangements from normal functionality may provoke great imbalances in some pathways and limited in others. As an example, we demonstrated that the exposure to low and high doses of the same chemical, i.e. the plasticizer bisphenol A, exerted completely different effects on the angiogenesis of mouse placenta by triggering two distinct signaling pathways, respectively promoting and inhibiting the process (Tait *et al*., 2015).

Being started with microarray technology, transcriptomics has now moved to Next Generation Sequencing (NGS) approaches which provide different levels of information with a greater detail, such as profiles of mutation in a specific disease condition, profiles of the transcriptome or, particularly, of the exome. With Whole-Exome Sequencing (WES) the costs may be reduced due to the selective capture of only those part of the genome encoding for functional proteins (about the 3%), where genetic variants are more probably located; for this reason, WES is increasingly implied in diagnostics and prenatal screenings (Suwinski *et al.*, 2019; Best *et al.*, 2018). Otherwise, with the Whole-Genome Sequencing (WGS), RNA-Seq in particular, all the transcripts are analysed permitting to compare the gene expression levels in different conditions as well as to discover new disease-related transcripts, among which long non coding RNAs are receiving great attention due to their regulatory role in several pathologies (Stark *et al.*, 2019). In addition, by short sequencing approaches, also micro RNAs may be identified. The discovery that the so-called "junk DNA" contains small and long non coding RNA (lncRNA) sequences which play a variety of different roles in cell machinery, strongly improved the application of NGS for the identification of novel transcripts and their interaction with other coding and non-coding RNAs, as well as with proteins under both normal and pathological conditions (Qian *et al.*, 2019). We recently reported the expression of different miRNA and lncRNA patterns (including novel transcripts) in visceral adipose tissue of lean and obese patients affected by colorectal cancer (Tait *et al.*, 2020). Our results highlighted that, in each condition, different regulatory interaction networks occur between non coding and coding genes, supporting the relevance of obesity comorbidity in colorectal cancer.

By the proteomic approach, the abundances of proteins actually functioning in the cellular machinery and their post-translational modifications are identified by mass spectrometry, allowing to discover new functions as well as new interactions among co-regulated proteins (Monti *et al.*, 2019). The end products of cellular metabolism, including sugars, amino acids, lipids, etc., are identified by metabolomics using chromatographic approaches (gas or liquid) coupled with mass spectrometry or by Nuclear Magnetic Resonance (NMR) (Schrimpe-Rutledge *et al.*, 2016). Importantly, metabolites do not represent the end of the story since they trigger cellular processes then affecting signaling pathways, transcription, translation, etc. (Rinschen *et al.*, 2019). Conversely to nucleic acids, proteins and metabolites are not amplifiable, so less abundant entities cannot be quantified. However, in the last years, these two approaches rapidly increased in sensitivity due to instrument implementations, thus their ability to effectively quantify the number of peptides and metabolites improved dramatically.

Challenges

For every omics approach, a series of critical issues should be considered in order to safely and confidently use the identified biomarkers. The greatest challenge is that omics approaches are costly, although less expensive than in the past, and time consuming, both at instrumental and data analysis level. Thus, the risk is to include limited numbers of samples, reducing the statistical

power of the analysis and finally compromising the robustness of the obtained results. This ultimately limits a biomarker from being validated and adopted into clinical routine.

Other common critical steps include appropriate sample collection and processing, identification of sequences (genes or proteins) against a reference genome, annotation and quantification. The steps involving sample manipulation relies on availability of high-quality starting material, coming from tissues of patients or other human matrices, tissues of animal models from *in vivo* studies, primary cells or cell lines cultured *in vitro*. The next steps are more related to instrument performance and its supporting software.

Bioinformatic analyses have then to be performed to manage the huge amount of data obtained and to uncover the biological significance behind the disease/condition/status under study. The pipelines to obtain such results may be quite different among computational groups and the scientific community has not reached yet a consensus on general criteria to consider in a data analysis workflow, so the results may differ. This poses the question on standardization of data format, filtering and cleaning, statistical methods and software used. First of all, to maximize transparency, each step of the analysis should be documented.

In this regard, the Organization for Economic Co-operation and Development (OECD) recently launched the Omics Reporting Project to produce guidance documents and reporting templates for omics approaches in chemical testing, to be used also in a regulatory context. The aim of these guidance documents is mainly to provide researchers a platform to report omics data in a harmonized framework, but they also help regulators and stakeholders to assess the quality of omics data with practical tools. So far, both the Transcriptomics and the Metabolomic Reporting Frameworks have been drafted and made available (OECD, 2021a; OECD, 2021b).

At least for genomic, transcriptomic and proteomic data, for which public repositories are available, each dataset produced should be deposited to guarantee transparency and permit other researchers to re-analyse the data or to integrate that dataset with others for a meta-analysis. Whatever the workflow used, the aim of bioinformatics analysis of omics data is to extract relevant information from complexity identifying main affected pathways and interaction networks. The elaboration would ultimately lead to the definition of one or more candidate biomarkers which could be further validated by molecular and/or clinical assessment.

Despite with each single omics we obtain a picture on derangements occurring at a certain level of cellular organization, this may be somewhat limiting when it is necessary to obtain a complete comprehension of the biological system, especially if the final goal is precision medicine. Moreover, several regulating processes occur in cells affecting the level of expression of genes, proteins and metabolites, thus it is not assured that information obtained at gene expression level is completely "translated" at protein or metabolite levels, and vice versa. Responses are also cell-type and tissue-specific, and relative abundances of biological molecules may be affected also by inter-individual variability. Therefore, different omics should be used in combination to improve the knowledge, also because each level of cellular organization may be differently affected when considering a disease condition or when studying alterations induced by external factors. The integration of different omics data may provide a holistic overview of the underlying mechanisms that a single omics approach may fail to identify. However, this represents a seriously challenging problem which computational biologists, bioinformaticians and biostatistics are facing. So far, an increasing number of studies integrated two omics approaches, mainly transcriptomics and proteomics or proteomics and metabolomics, whereas not many implemented the integration of data from three or more omics (Misra *et al.*, 2018). Our group is increasingly involved in the implementation of multi-omics approaches and data analysis. A work is in progress with the support of the Core Facility at the Istituto Superiore di Sanità (ISS, the National Institute of Health in Italy), in which we performed transcriptomic and proteomic analyses to investigate exposure effects of some chemical contaminants on a human liver cellular

model. By our integrated bioinformatics analysis, we are identifying the affected pathways also discriminating which genes and proteins are more relevant in the regulatory network.

Data gaps

Overall, the biological relevance of omics approaches, especially if human *in vitro* models or human matrices are analysed, relies on the ability to provide relevant information on human health in a mechanism-based context, rather than being based on apical adverse effects as observed in animal studies. In particular, they give the possibility to identify which is the flow of the signal within cells, as well as the temporal profile, if time-course experiments are performed. In this frame, increasing efforts should be undertaken to cover the investigation of all possible adverse effects, avoiding underestimation of some biological systems since, at the moment, not all the organs/tissues are equally considered. Table 1 provides an illustrative example, showing a simple search in PubMed with few keywords just to compare the number of publications in this field, not filtering for review or other type of articles.

Table 1. Number of publications retrieved in PubMed (date of search 07/2021) by using as keywords the name of the organs, the Boolean operator "AND" and, alternatively, the column names

Key words	Omics	Multi-omics	Omics AND biomarker	Multi-omics AND biomarker
"AND"				
Blood	1757	410	749	157
Brain	1036	275	275	80
Liver	920	253	263	59
Lung	746	219	248	69
Breast	704	236	246	77
Kidney	439	117	184	40
Skin	250	64	61	15
Colon	181	70	53	20
Bone	227	49	57	16
Eye	217	36	37	5
Prostate	191	52	71	20
Ovaries	93	18	15	2
Bladder	88	34	43	12
Bone marrow	81	16	17	3
Thyroid	70	12	24	3
Pancreas	63	14	20	7
Stomach	61	19	28	7
Testis	42	5	8	1
Placenta	49	8	16	3
Adrenal	40	10	12	3
Lymph nodes	37	13	12	6
Uterus	28	2	11	0
Esophagus	15	8	9	6
Thymus	10	2	1	0

It is noteworthy that, by querying blood AND omics, 1757 publications were found, whereas for thymus AND omics only 10 are available. Moreover, only in a limited number of these publications the word "biomarker" matched the query. By searching for multi-omics studies, the numbers dramatically decrease and, among them, studies including "biomarker" are really scarce. This mere speculative exercise easily demonstrates how some organs are poorly investigated through omics approaches.

As a further consideration, a relevant issue often neglected in experimental studies, including omics, is the sex/gender. As stated above, in a context of a precision medicine era, gender-specific effects should be always taken into account since signaling pathways, organ functionalities, metabolism, aging, response to stimuli, immunity, etc. are different between the two genders. Repeating the same PubMed Search as shown above, but including the keywords "gender OR sex" in the query, we can see in Table 2 that in a very limited portion of studies a match with the keywords was found, indicating a scarce consideration of this fundamental aspect.

Table 2. Number of publications retrieved in PubMed (date of search 07/2021) by using as keywords the name of the organs, the Boolean operator "AND" and, alternatively, the column names

Key words	Omics AND (gender or sex)	Multi-omics AND (gender or sex)	Omics AND biomarker AND (gender or sex)	Multi-omics AND biomarker AND (gender or sex)
"AND"				
Blood	69	15	30	5
Brain	43	13	10	0
Liver	24	9	6	4
Lung	26	9	11	4
Breast	6	3	3	2
Kidney	13	6	6	1
Skin	7	2	1	1
Colon	4	3	2	1
Bone	3	0	1	0
Eye	5	2	0	0
Prostate	3	1	1	1
Ovaries	5	0	1	0
Bladder	0	0	0	0
Bone marrow	0	0	0	0
Thyroid	2	0	0	0
Pancreas	0	0	0	0
Stomach	1	0	0	0
Testis	7	0	1	0
Placenta	1	0	0	0
Adrenal	40	10	12	3
Lymph nodes	37	13	12	6
Uterus	28	2	11	0
Esophagus	15	8	9	6
Thymus	10	2	1	0

Conclusions

In conclusion, the use of omics approaches, especially multi-omics, is crucial for the identification of novel adverse effects and biomarkers, providing valuable information on system biology in general and pathological mechanisms in particular.

Harmonization of data analysis workflow and use of reporting templates are highly encouraged to allow data comparison and meta-analysis, especially in an open science perspective.

Data gaps on some tissues/organs as well as on sex/gender differences are highlighted, suggesting that more efforts should be undertaken to systematically apply omics approaches and fully exploit their potential.

As an added value, omics allow to prioritize and consequently reduce the number of studies using animal models, limiting the research to studies substantiating observed early molecular events and modes of action and to link them to an apical adverse outcome. This approach is compliant with the 3Rs framework and, if largely applied, it will progressively increase the knowledge based on mechanisms and on the discovery of new biomarkers to be used in clinical, prevention and risk assessment applications without the use of animals.

References

Best S, Wou K, Vora N, Van der Veyver IB, Wapner R, Chitty LS. Promises, pitfalls and practicalities of prenatal whole exome sequencing. *Prenat Diagn* 2018;38(1):10-9.

Bošković A, Rando OJ. Transgenerational epigenetic inheritance. *Annu Rev Genet* 2018; 52:21-41.

Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated Omics: tools, advances, and future approaches. *J Mol Endocrinol* 2018:JME-18-0055.

Monti C, Zilocchi M, Colugnat I, Alberio T. Proteomics turns functional. *J Proteomics* 2019;198:36-44.

OECD. *Transcriptomics Reporting Framework (TRF)*. Paris: Organization for Economic Co-operation and Development; 2021. Available from: https://www.oecd.org/chemicalsafety/testing/transcriptomic-reporting-framework.pdf, last visited 30/07/2021.

OECD. *Metabolomics Reporting Framework (MRF).* Paris: Organization for Economic Co-operation and Development; 2021. Available from: https://www.oecd.org/chemicalsafety/testing/metabolomics-reporting-framework.pdf, last visited 30/07/2021.

Qian X, Zhao J, Yeung PY, Zhang QC, Kwok CK. Revealing lncRNA Structures and Interactions by Sequencing-Based Approaches. *Trends Biochem Sci* 2019;44(1):33-52.

Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol* 2019;20(6):353-367.

Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies-challenges and emerging directions. *J Am Soc Mass Spectrom* 2016;27(12):1897-1905.

Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet* 2019;20(11):631-56.

Stricker SH, Köferle A, Beck S. From profiles to function in epigenomics. *Nat Rev Genet* 2017;18(1):51-66.

Suwinski P, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front Genet* 2019;10:49.

Tait S, Tassinari R, Maranghi F, Mantovani A. Bisphenol A affects placental layers morphology and angiogenesis during early pregnancy phase in mice. *J Appl Toxicol* 2015;35(11):1278-91.

Tait S, Baldassarre A, Masotti A, Calura E, Martini P, Varì R, Scazzocchio B, Gessani S, Del Cornò M. Integrated Transcriptome Analysis of human visceral adipocytes unravels dysregulated microRNA-Long Non-coding RNA-mRNA networks in obesity and colorectal cancer. *Front Oncol* 2020;10:1089.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;20(8):467-484.

UNEP (United Nations Environment Programme). *Biomarkers in risk assessment: validity and validation*. Geneva: World Health Organization, 2001. (Environmental Health Criteria 222) https://wedocs.unep.org/20.500.11822/29529.