MEDICAL SCIENCE • TECHNOLOGY

EDITORIAL

e-ISSN 2329-0072
© Med Sci Tech, 2018; 59: 1-3
DOI: 10.12659/MST.908220

Received: 2017.11.25
Accepted: 2017.11.30
Published: 2018.01.16

Current Challenges for Big *Omics* Data Analytics and Precision Medicine

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

AEFG 1 **Elmer Andrés Fernandez**
DEFG 2 **Federico Marcelo Casares**

1 Centre for Immunology and Infectious Diseases (CIDIE), National University of Cordoba (UNC), National Scientific and Technical Research Council (CONICET), Cordoba, Argentina
2 LISRA Institute, East Rockaway, NY, U.S.A.

Corresponding Authors: Elmer Andrés Fernandez, e-mail: efernandez@bdmg.com.ar, Federico Marcelo Casares, e-mail: fmcasares@yahoo.com
Source of support: This work was supported by grants from the following Argentine institutions: Universidad Católica de Córdoba (BOD/2016 to EAF), Ministerio de Ciencia, Tecnología e Innovación Productiva (PPL 6/2011 to EAF), Secretaría de Ciencia y Tecnología – Universidad Nacional de Córdoba (30720150101719CB to EAF), and the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

Ambitious efforts to characterize disease have been made worldwide, mainly in cancer, with initiatives such as the Cancer Genome Atlas. Many of these cost-intensive studies use cutting-edge technologies to delve deeply into the intrinsic genomic, transcriptomic, proteomic, metabolomic, etc, (i.e., omics) type of data to better explain the phenotype. But while more data is being stored, the complexity of cancer seems to challenge even more our ability to understand its nature and thus to uncover useful bio-physiological information. We strongly believe that data analytics, as well as our understanding of 'normal' cases, are still in their infancy, opening great opportunities in translational cancer research to pursue precision medicine through Big Omics Data analytics.

MeSH Keywords: **Data Interpretation, Statistical • Genomics • Microarray Analysis • Proteomics • Statistics as Topic**

Full-text PDF: https://medscitechnol.com/abstract/index/idArt/908220

Background

Precision medicine aims at the assignment of the most appropriate therapy for a particular patient. This notion entails a fundamental concept: the precise identification of the "patient-therapy" or "patient-evolution" pairs. In order to realize this concept, it is necessary, on the one hand, to have as much detailed knowledge as possible of the characteristics of the patients, how the disease manifests itself, and its molecular basis, while on the other hand, it is necessary to have analytical methods available to evaluate or compare one cohort with another, as well as algorithms for accurate identification of the patient-treatment/evolution pair (i.e., diagnosis and/or prognosis). As a first step, it is necessary to have data that allow us to obtain such a characterization by means of data-mining algorithms, allowing the extraction of knowledge and the development of accurate diagnostic methods [1,2]. In this sense, data sciences-based bioinformatics will play a crucial role in addressing '*omics*' data (e.g., genomic, transcriptomic, proteomic, and metabolomic) and clinical data all together, processing and analyzing the different sources and types of data to transform them into useful knowledge [3]. In the future, precision medicine will integrate molecular and phenotype (clinical) data from diseases through data sciences-based bioinformatics, to massively evaluate, characterize, and validate multiple molecules with clinical linkage, which were previously impossible to study with classical methods of genetics and conventional molecular biology.

The human body has now become a big data source in which different bioanalytical technologies, including biosensors and their associated algorithms [4,5], are measuring phenomena from molecules to biosignals and images, which are filling up data repositories worldwide, like the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EMBL-EBI). For instance, the Gene Expression Omnibus (GEO) [6] has more than two million biological samples of different types. However, the main drawback is that different data sources do not share their origin (i.e., the subject). In order to fill this gap, the Cancer Genome Atlas (TCGA) [7] uses many different biosensing techniques to measure a broad spectrum of the underlying biology for the same subset of patients, providing omics data at the genomic, transcriptomic, and proteomic levels, as well as clinical information, in an unprecedented manner. The availability of such an amount and diversity of data is challenging our current portfolio of bioinformatic methods, which are still in their infancy, to deal with it in a more integrative manner. In this regard, the human body prevails in the maintenance of internal homeostasis thanks to the synergy of all of its organs and molecules. Thus, diseases also need to be comprehensively approached. For this reason, the development of bioinformatic methods that can efficiently deal with all different data types simultaneously becomes crucial, allowing us to query them in an integrative manner and providing a comprehensive summary of the underlying phenomena, primarily to facilitate the interpretation by the researcher, but also to provide diagnostic support as a bedside tool for the physician in clinical practice. For this ultimate goal, we will need to include data from different populations, including those still missing from the TCGA project, like Latin American populations [8]. In this regard, it is known that certain mutations are specific to certain regions or even areas within regions and, thus, comprehensive knowledge of the subject's background, including cultural factors (e.g., dietary and environmental), which may in turn influence epigenetic factors, becomes important for accurate (and hence, cost-effective) diagnostic targeting [9–13]. For this reason, the Latin American Cancer Research Network is currently attempting to characterize Latin American breast cancer cohorts [14], and the International Cancer Genome Consortia [15] is including several different worldwide cohorts.

Discussion

Despite all these efforts to include and integrate cohorts from various geographical and cultural backgrounds, the genotypically/phenotypically 'normal' subject or tissue is currently not fully characterized. This endeavor is not simple, especially if we understand this type of normality differently from the more rigid connotation based on Johannsen's original definition of genotype and phenotype [16]. After all, which interacting factors or synergies make a genotypically normal cohort belong to a phenotypically normal group? In the answer resides the key for successful precision medicine.

Conclusions

In summary, to develop an efficient precision medicine, we will need to effectively determine both our normal and diseased molecular background. This, in turn, will enable us to achieve a better characterization of the disease by integrating several data levels from different synergistic sources (e.g., genomics, proteomics, metabolomics, phenotypic, and even cultural and social) and to develop new bioinformatic methods and tools to relate them and integrate them comprehensively.

References:

1. Fresno C, González GA, Merino GA et al: A novel non-parametric method for uncertainty evaluation of correlation-based molecular signatures: Its application on PAM50 algorithm. Bioinformatics, 2017; 33(5): 693–700
2. González-Montoro A, Prato L, Casares F et al: Appropriate sample size for standardization parameters estimation reduces misdiagnoses of molecular-based risk predictors in breast cancer. Med Sci Tech, 2017; 58: 111–18
3. Rodriguez JC, González GA, Fresno C et al: Improving information retrieval in functional analysis. Comput Biol Med, 2016; 79: 10–20
4. Stefano GB, Fernández EA: Biosensors: Enhancing the natural ability to sense and their dependence on bioinformatics. Med Sci Monit, 2017; 23: 3168–69
5. Stefano GB, Kream RM: Personalized- and one- medicine: Bioinformatics foundation in health and its economic feasibility. Med Sci Monit, 2015; 21: 201–4
6. Barrett T, Wilhite SE, Ledoux P et al: NCBI GEO: Archive for functional genomics data sets – update. Nucleic Acids Res, 2013; 41(Database issue): D991–95
7. Weinstein JH, Collisson EA, Mills GB et al: The cancer genome atlas pan-cancer analysis project. Nat Genet, 2013; 45(10): 1113–20
8. Serrano-Gómez SJ, Fejerman L, Zabaleta J: Breast cancer in Latinas: A focus on intrinsic subtypes distribution. Cancer Epidemiol Biomarkers Prev, 2017 [Epub ahead of print].
9. Teegarden D, Romieu I, Lelièvre SA: Redefining the impact of nutrition on breast cancer incidence: Is epigenetics involved? Nutr Res Rev, 2012; 25(1): 68–95
10. Porchia LM, Gonzalez-Mejia ME, Calderilla-Barbosa L et al: Common BRCA1 and BRCA2 mutations among latin american breast cancer subjects: A meta-analysis. J Carcinogene Mutagene, 2015; 6: 228
11. Villarreal-Garza C, Alvarez-Gómez RM, Pérez-Plasencia C et al: Significant clinical impact of recurrent BRCA1 and BRCA2 mutations in Mexico. Cancer, 2015; 121(3): 372–78
12. Solano AR, Cardoso FC, Romano V et al: Spectrum of BRCA1/2 variants in 940 patients from Argentina including novel, deleterious and recurrent germline mutations: Impact on healthcare and clinical practice. Oncotarget, 2016; 8(36): 60487–95
13. Ossa CA, Torres D: Founder and recurrent mutations in BRCA1 and BRCA2 genes in latin american countries: State of the art and literature review. Oncologist, 2016; 21(7): 832–39
14. Llera AS, Podhajcer OL, Breitenbach MM et al: Translational cancer research comes of age in Latin America. Sci Transl Med, 2015; 7(319): 319fs50
15. Zhang J, Baran J, Cros A et al: International Cancer Genome Consortium Data Portal - a one-stop shop for cancer genomics data. Database (Oxford), 2011; 2011: bar026
16. Johannsen W: The genotype conception of heredity. The American Naturalist, 1911; 45: 129–59