ISSN 1669-9106

SPECIAL ARTICLE

MEDICINA (Buenos Aires) 2019; 79 -
NÚMERO ESPECIAL 80 ANIVERSARIO: 00-00

GENOMICS AND BIOINFORMATICS AS PILLARS OF PRECISION MEDICINE IN ONCOLOGY

ROMINA CANZONERI, EZEQUIEL LACUNZA, MARTÍN C. ABBA

Centro de Investigaciones Inmunológicas Básicas y Aplicadas (CINIBA), Facultad de Ciencias Médicas, Universidad Nacional de La Plata, Buenos Aires, Argentina

Abstract The battle against cancer has advanced tremendously in the last thirty years and the survival rate has doubled. However, it is still difficult to achieve a generalized cure. The challenge is that cancer is not a unique disease; it is about dozens of different manifestations, even within the same tumor location. For systems biology, each solid tumor is a unique system characterized by its cellular heterogeneity, its interaction with the microenvironment in which it grows and develops, and its ability to adapt and modify it. Recent advances in understanding the molecular mechanisms that underlie cancer are transforming the diagnosis and treatment of the disease. In this sense, a growing set of treatments capable of attacking a specific tumor with higher efficiency has been developed, defining a new paradigm: the precision medicine in oncology. Genomics and bioinformatics are two fundamental pillars in this applied field. These technologies generate massive data (big data) that require analytical tools and trained personnel for the analysis, integration and transfer of the information to physicians. This presentation describes the concepts of personalized medicine, Big Data, the main advances in genomics and bioinformatics as well as their future perspectives and challenges.

Key words: precision medicine, genomics, bioinformatics, big data

Resumen ***Genómica y bioinformática como pilares de la medicina de precisión en oncología.*** La batalla contra el cáncer ha avanzado enormemente en los últimos treinta años y la tasa de supervivencia se ha duplicado, sin embargo aún es difícil alcanzar una cura generalizada. El desafío reside en que el cáncer no es una enfermedad única, se trata de decenas de manifestaciones diferentes incluso dentro de una misma localización tumoral. Para la biología de sistemas, cada tumor sólido es un sistema único caracterizado por su heterogeneidad celular, su interacción con el microambiente en el que crece y se desarrolla, y su capacidad de adaptarse y modificarlo. Los avances recientes en la comprensión de los mecanismos moleculares que subyacen al cáncer están transformando el diagnóstico y el tratamiento de la enfermedad. En este sentido, se ha desarrollado un conjunto creciente de tratamientos capaces de atacar con mayor eficiencia a un tumor específico dando paso a nuevo paradigma: el de la medicina de precisión. La genómica y la bioinformática son dos ejes fundamentales en el desarrollo y aplicación de la medicina personalizada. Estas tecnologías generan datos masivos (*Big Data*) que requieren de herramientas analíticas y personal capacitado para su análisis, integración y transferencia de la información hacia los médicos especialistas. En esta presentación se describen los principales avances en genómica y bioinformática aplicados a la medicina de precisión así como sus perspectivas futuras, desafíos y problemáticas.

Palabras clave: medicina de precisión, genómica, bioinformática, *big data*

Although the fight against cancer has advanced tremendously in the last thirty years and the survival rate has doubled, the search for a definitive cure remains a utopia. The challenge lies in that cancer is not a unique disease but many different manifestations, even within the tumor itself. For systems biology, each tumor constitutes a system characterized by its cellular heterogeneity, the microenvironment (surrounding tissue, immune system) in which it grows, and its ability to adapt and modify it[1]. That is, a tumor is a unique and variable entity, product of the multiple mutations and epigenetic alterations that occur in some of the thousands of cells in the early stages of malignant transformation. Since these cells can acquire new mutations and new genetic variations as the tumor grows and develops, an almost infinite amount of genomic variations can be generated[2].

For this reason, cancer treatment becomes a complex task, in which oncologists face a mobile and unpredictable target. Unfortunately, there is no single solution for all cases. While surgery, radiation, and chemotherapy remain the primary treatments against cancer, the advances in the molecular mechanisms underlying tumor growth are generating new approaches to the diagnosis and treatment of the disease.

In the last five years, many new cancer treatments have been developed to combat the most aggressive forms of

Postal address: Martín C. Abba, CINIBA, Facultad de Ciencias Médicas, UNLP, Calle 60 y 120 S/N, 1900 La Plata, Buenos Aires, Argentina
e-mail: mcabba@gmail.com

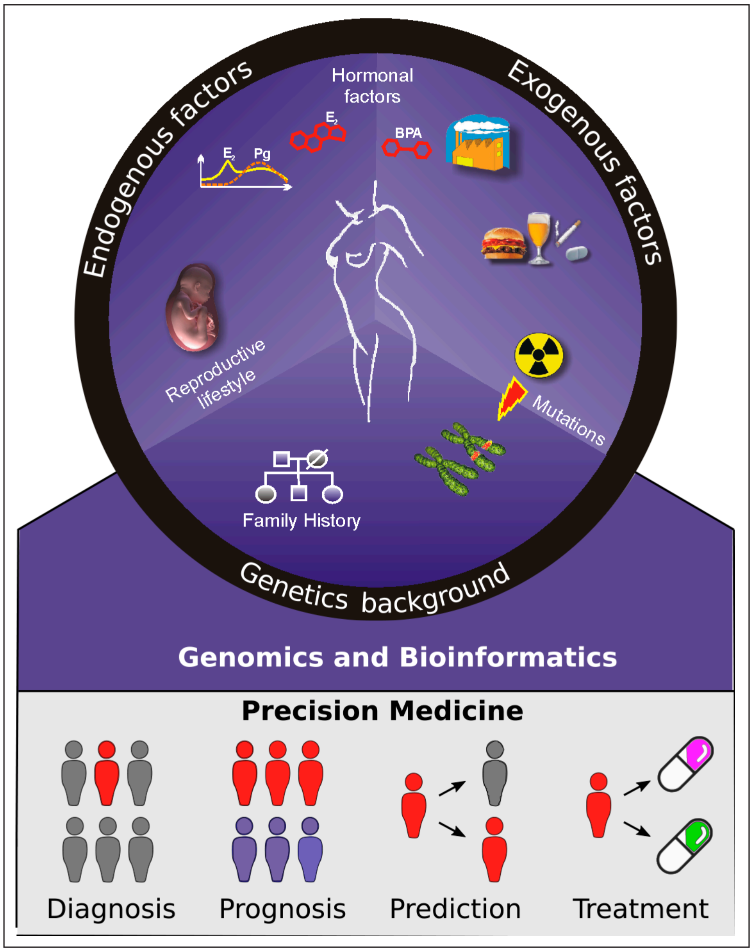
cancers. From the genetic edition to immunotherapy, the future of cancer treatment focuses on finding a specialized solution for each problem[3]. In this sense, precision medicine offers the possibility of personalized treatments to attack more accurately a specific tumor, reducing the possible side effects.

The paradigm of precision medicine is not new; recent advances in genomics and bioinformatics have helped to materialize this discipline (Fig. 1). Personalized medicine has shown benefits for patients, allowing the oncologists to prescribe the most appropriate treatment from an early stage of care, minimizing the risk of adverse reactions to the medication, or the lack of effect[4]. The personalized medicine increases the overall efficiency of health care because the molecular diagnostic profiles can rule out treatments that will not be effective –avoiding the costs involved– and identify those with the highest chance of success.

How are genetic changes identified in tumors?

Precision medicine is a medical model that proposes the customization of healthcare through medical decisions, treatments, practices, or products being tailored to the

Fig. 1.– Risk factors for cancer development (breast cancer in particular) can be classified into endogenous factors (hormonal balance, intrauterine development), exogenous factors (physical agents, chemicals, consumption habits), and the genetic background of the patient. The interaction between these factors determines that each individual responds differentially to a specific treatment. In oncological terms, patients can be stratified according to the tumor molecular profile allowing a more accurate diagnosis or the selection of the most appropriate treatment

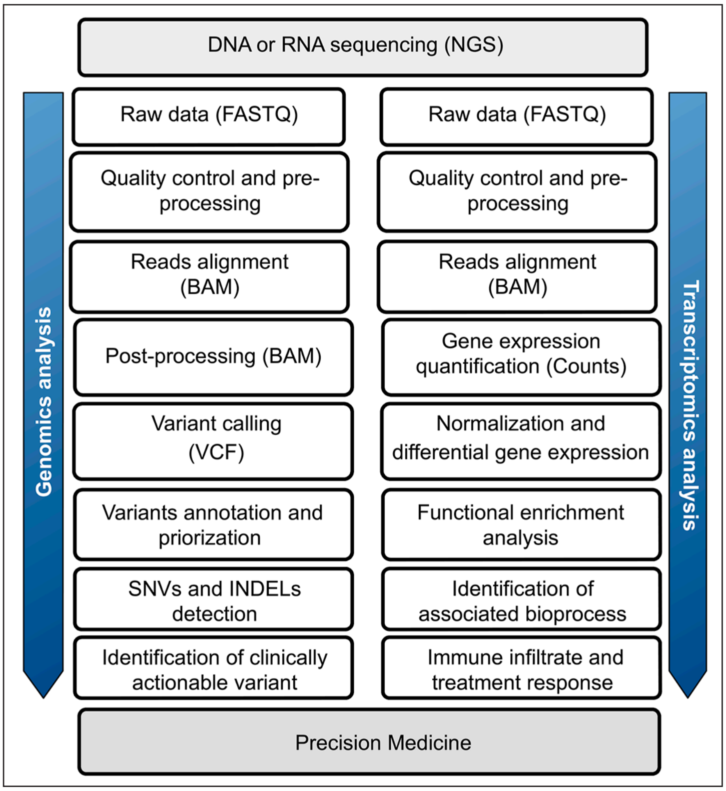


E2: estradiol; Pg: progesterone; BPA: bisphenol A individual patient[4]. This strategy requires to determine the molecular mechanisms by which a tumor proliferate, escape from the immune system control and resist to the potential applied treatment.

The technological-analytical, computational, and bioinformatics developments experienced during the last decade have allowed moving this model from the conceptual to the pragmatic scenario. For example, the current Next Generation Sequencing platforms (NGS), allow the epigenomics, genomics, and transcriptomics characterization of a specific tumor. Initially, DNA and RNA are obtained from tissue tumor biopsies, or liquid biopsies (circulating cell-free DNA or RNA derived from tumor exosomes). Quality controls are then carried out to determine the integrity and concentration of nucleic acids. DNA/RNA samples are subjets to specific library construction protocols according to the user needs [e.g., whole genome sequencing (WGSeq), whole exome sequencing (WESeq), TargetedSeq, MethylSeq, RNAseq, small RNAseq] and the sequencing platform employed (Illumina: MySeq, HiSeq, and NovaSeq series; Thermo-Fisher: PGM, Ion Proton, Ion S5 series, etc.)[4]. The raw data (FASTQ files) are pre-processed (demultiplexing, quality control, adapter trimming) and subject to a specific pipeline for the extraction of non-trivial information such as the detection of activated/deactivated signaling pathways or the identification of clinically actionable variants (Fig. 2). In this sense, the NGS platform generates a complex corpus of data that makes genomic and bioinformatics fall within the disciplines that require Big Data approaches[5]. Although technological progress has made NGS studies increasingly accessible in terms of availability and costs, the current bottleneck lies in the ability to handle the large volume of data generated.

There are more than 3 million base pairs in the human exome, distributed in 180,000 exons (25,000 coding genes)[2]. The complete exome or transcriptome sequencing for a small project (10 to 30 samples) generates terabytes of raw data (FASTQ files). These data are pre-processed and aligned against a reference genome

Fig. 2.– Bioinformatics workflows for NGS (Next Generation Sequencing) data analysis (Exome-Seq in the left and RNA-seq in the right)


DNA or RNA sequencing (NGS)
Genomics analysis
Raw data (FASTQ)
Quality control and pre-processing
Reads alignment (BAM)
Post-processing (BAM)
Variant calling (VCF)
Variants annotation and priorization
SNVs and INDELs detection
Identification of clinically actionable variant
Transcriptomics analysis
Raw data (FASTQ)
Quality control and pre-processing
Reads alignment (BAM)
Gene expression quantification (Counts)
Normalization and differential gene expression
Functional enrichment analysis
Identification of associated bioprocess
Immune infiltrate and treatment response
Precision Medicine

FASTQ: TXT file with the reads sequence and quality; BAM: binary alignment map; SNV: single nucleotide variants; INDEL: insertion or deletion variants

(BAM files for binary version of sequence alignment) that are usually on the gigabyte scale depending on the sequencing coverage (average number of unique reads that include a given nucleotide in the reconstructed sequence) and the length of reads[5].

The initial steps of pre-processing and alignment are the most demanding in terms of computational capability, and the downstream bioinformatics processes depend on the type of study to be performed, requiring less informatics resources. Various public consortiums and companies are developing cloud-based bioinformatics platforms to facilitate the implementation of the required workflows with on-demand computational capability.

Collective bioinformatics intelligence

The term "bioinformatics" is now recognized as a field that encompasses biology, medicine, computer science, mathematics, statistics, and information technology. Bioinformatics tools and databases constitute an integral component of the current research process in biomedical sciences[6]. Since the introduction of information technology in biological research, a plethora of computational tools and databases have emerged, contributing to breakthroughs in the field. The emergence of NGS technologies has inspired the development of new computational techniques, and also required the implementation of highly sophisticated pipelines.

In the past decade, the notion of 'biological data' has changed in magnitude and complexity from sets of hundreds to sets of millions of entities (e.g., genes, alternative splicing variants, protein, isoforms, CpG island). This exponential increase in the volume of biological data has stimulated the development of an ever-increasing number of bioinformatics tools[6]. As there is no just a single cure for cancer, there is no a single tool to analyze the data. As soon as new tools and techniques are developed, new data sources arise, demanding radical improvements in the algorithms for processing and analysis. No less relevant is the lack of standards regarding how genomic data is collected, stored, and processed, which raises evident problems in terms of reproducibility[5]. In this sense, several cooperative approaches have been progressively adopted with the aim to democratize the use of genomic. The Genomic Data Commons (GDC) (https://portal.gdc.cancer.gov/) is a research program of the National Cancer Institute (NCI, USA). The mission of the GDC is to provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies (TCGA, TARGET, CGCI, etc.) in support of precision medicine[7]. On the other hand, the Bioconductor (https://www.bioconductor.org/)[8], Biocontainers (https://biocontainers.pro/)[9], and Galaxy (https://usegalaxy.org/)[10] projects provide open-source bioinformatics tools allowing the reproducible implementation of data analytics tools and pipelines. In addition, the DREAM challenge is a public/private effort that appeals to the collaborative and transparent data exchange to evaluate existing analytical tools, suggest improvements, and develop new solutions (http://dreamchallenges.org/).

Challenges of implementing precision medicine

The current public and private health systems are immersed in a sea of data: medical records, results of clinical trials, biometric data monitoring, diverse diagnostic images, and genetic information of patients. But more importantly, the exploding volume and speed of unstructured data growth cannot be appropriately managed with traditional database systems. Volume, variety, and velocity are precisely the variables that characterize any Big Data environment[5]. Precision medicine is a discipline in constant evolution due to the dynamism of technological advances and the continuous development of new analytical tools. Like any technological advent, its incorporation by the scientific community implies an adaptation period, which is even longer for its implementation in the health system. Technical and human resources are two relevant components to support precision medicine implementation. Additionally, the institutional and idiosyncratic legal barriers make the clinical genomics set up a challenge for its consolidation in the health system.

Technological and human resources

The technology associated with the management of Big Data is already a consummate reality. The fundamental pillars are the systems of file distribution, the scalable databases, the mass processing software (Hadoop type), cloud computing, and the Internet[5]. This technology, however, still needs to be consolidated in the health system, being necessary to increase the public and private investments for these types of approaches. One of the fundamental aspects lies in the human resources: trained scientists and technicians for data analysis and interpretation. They constitute the link between the analysis and interpretation of the data and the transfer of the information to the physicians so that they make the appropriate decisions about their patients.

Security of genomic data

Genomic data need to be protected. Therefore, its privacy and confidentiality should be preserved similarly to other protected health information. Privacy safeguards include the utilization of data encryption, password protection, secure data transmission, auditions of data transferring methods, and the operation of institutional strategies against data breaches and mischievous abuse of the data[11]. The Fair Information Practices Principles (FIPPs) offer a framework for enabling data sharing and usage based on the guidelines adopted by the U.S. Department of Health and Human Services[12].

Current status and perspectives

Genomics and bioinformatics are two fundamental pillars in the development and implementation of personalized precision medicine in oncology. These technologies and analytical tools need to be implemented in an integrated way with the information available in the electronic medical records facilitating the physician interpretation. This must be achieved in the strictest framework of ethical and security safeguards since personal information is highly sensitive, and is under legal protection requiring the collaboration of all health system actors.

Genomics and bioinformatics-based approaches will allow the identification of the most effective treatment for each patient reducing the chances of treatment failure. For example, colorectal cancer can be considered to include many distinct molecular diseases, characterized by a partially defined pattern of molecular changes that affect various molecular pathways. This diversity of tumors has challenged the therapies developed during the past years, leading to the need to recruit patient groups with similar molecular alterations, which can be addressed with more personalized therapies. An increasing number of clinical trials have focused on the use of new drugs directed against specific pathways to be used alone or in combination. The correct stratification of patients and the appropriate choice of therapeutic agents will eventually lead to significant advances in the treatment of colorectal cancer and cancer in general[13].

In addition to inter-tumor diversity, the intra-tumor heterogeneity brings even more challenges to the field. The coexistence of multiple cellular subclones with different sets of molecular changes and different drug sensitivities implies that therapeutic strategies directed against the predominant aberrations may not be effective against the whole tumor[14]. The single-cell RNA sequencing approach is one of the recent advances in genomics that is generating a new corpus of complex data. This data will need to be analyzed to discover new entities within the same tumor and thus adjust the accuracy of the therapies. Although the first advances are already being seen, especially in the diagnosis and treatment of cancer, much remains to be done to reach a truly personalized precision medicine.

Acknowledgments: This work has been supported by PICT2015-0149 (FONCyT, ANPCyT, Argentina) and 1U54CA221208-1 (NIH-NCI, USA) grants.

Conflict of interests: None to declare

References

1. Faratian D, Bown JL, Smith VA, Langdon SP, Harrison DJ. Cancer systems biology. *Methods Mol Biol* 2010; 662: 245-63.
2. Quackenbush J. The Human Genome: Book of Essential Knowledge. Reino Unido: Imagine Publishing Inc., 2011.
3. Jaffee EM, Dang CV, Agus DB, et al. Future cancer research priorities in the USA: a Lancet Oncology Commission. *Lancet Oncol* 2017; 18:e653-706.
4. Krzyszczyk P, Acevedo A, Davidoff EJ, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology (Singap World Sci)* 2018; 6: 79-100.
5. He KY, Ge D, He MM. Big data analytics for genomic medicine. *Int J Mol Sci* 2017; 18: pii: E412.
6. Levin CL, Dynomant E, Gonzalez BJ, et al. A data-supported history of bioinformatics tools. *arXiv* 2018; arXiv:1807.06808.
7. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* 2017; 130: 453-9.
8. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 2015; 12: 115-21.
9. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* 2017; 33: 2580-2.
10. Afgan E, Baker D, Batut B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018; 46: W537-44.
11. Middleton A. Society and personal genome data. *Hum Mol Genet* 2018; 27:R8-13.
12. Baker DB, Kaye J, Terry SF. Governance through privacy, fairness, and respect for individuals. *EGEMS (Wash DC)* 2016; 4: 1207.
13. Palma S, Zwenger AO, Croce MV, Abba MC, Lacunza E. From molecular biology to clinical trials: toward personalized colorectal cancer therapy. *Clin Colorectal Cancer* 2016; 15: 104-15.
14. Yates LR, Gerstung M, Knappskog S, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* 2015; 21: 751-9.