frontiers in Oncology

REVIEW
published: 07 April 2020
doi: 10.3389/fonc.2020.00423

Computational Oncology in the Multi-Omics Era: State of the Art

Guillermo de Anda-Jáuregui [1,2]* *and Enrique Hernández-Lemus* [1,3]*

[1] *Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico,* [2] *Cátedras Conacyt Para Jóvenes Investigadores, National Council on Science and Technology, Mexico City, Mexico,* [3] *Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico*

Cancer is the quintessential complex disease. As technologies evolve faster each day, we are able to quantify the different layers of biological elements that contribute to the emergence and development of malignancies. In this multi-omics context, the use of integrative approaches is mandatory in order to gain further insights on oncological phenomena, and to move forward toward the precision medicine paradigm. In this review, we will focus on computational oncology as an integrative discipline that incorporates knowledge from the mathematical, physical, and computational fields to further the biomedical understanding of cancer. We will discuss the current roles of computation in oncology in the context of multi-omic technologies, which include: data acquisition and processing; data management in the clinical and research settings; classification, diagnosis, and prognosis; and the development of models in the research setting, including their use for therapeutic target identification. We will discuss the machine learning and network approaches as two of the most promising emerging paradigms, in computational oncology. These approaches provide a foundation on how to integrate different layers of biological description into coherent frameworks that allow advances both in the basic and clinical settings.

Keywords: multi-omics analysis, computational oncology, data integration, cancer complexity, machine learning, network science

OPEN ACCESS

Edited by:
Francesca Finotello, Innsbruck Medical University, Austria

Reviewed by:
Raoul Jean Pierre Bonnal, Istituto Nazionale Genetica Molecolare (INGM), Italy
Barbara Di Camillo, University of Padova, Italy
Dietmar Rieder, Innsbruck Medical University, Austria

***Correspondence:**
Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *01 December 2019*
Accepted: *10 March 2020*
Published: *07 April 2020*

Citation:
de Anda-Jáuregui G and Hernández-Lemus E (2020) Computational Oncology in the Multi-Omics Era: State of the Art. Front. Oncol. 10:423. doi: 10.3389/fonc.2020.00423

1. CANCER: THE COMPLEX DISEASE

Cancer is by now widely accepted to be the quintessential complex disease: a proper description of the pathological phenotype can only be achieved by properly integrating the myriad of interconnected biological elements and their relationships with their environment (1). As a complex system, cancer exhibits features, such as: emergent patterns, adaptive and collective behaviors, self-organization, non-linear dynamics, and interactions forming complex networks (2). Examples of these can be found in the *Hallmarks of Cancer* (3, 4), as seen in **Figure 1**.

On a system-wide fashion, every tumor is involved in interactions with non-cancer elements: such as gene-environment interactions (GxE) (5), micro-environmental interactions (6), and those with the immune system (7); intercellular interactions within the tumor environment (8); and intracellular interactions, such as transcriptional regulation and gene co-expression (9, 10), signaling (11, 12) and metabolic pathways (13, 14), as well as protein interactions (15). These are exemplified in **Figure 2**. It soon becomes evident that a major source of cancer complexity lies on the many layers of interacting elements involved in the phenomenon.

2. THE MULTI-OMICS PARADIGM

2.1. Multi-Omics in a Nutshell

Multiomics is the name given to the modelization approach in biology hat makes use of more than one of the current high-throughput biomlecular experimental techniques (a.k.a. omics) in order to characterize biological systems at the phenomenological level. It is understood that every omic contributes on a specific fashion to shape the actual biological phenotype under study. For this reason, it has become evident that there is a need for integrating frameworks to gather and organize the knowledge gained with each experimental approach into mechanistic or semi-mechanistic descriptions of the biological phenomenon. This issue has been deemed particularly relevant for the study of complex phenotypes, such as cancer tumors (16).

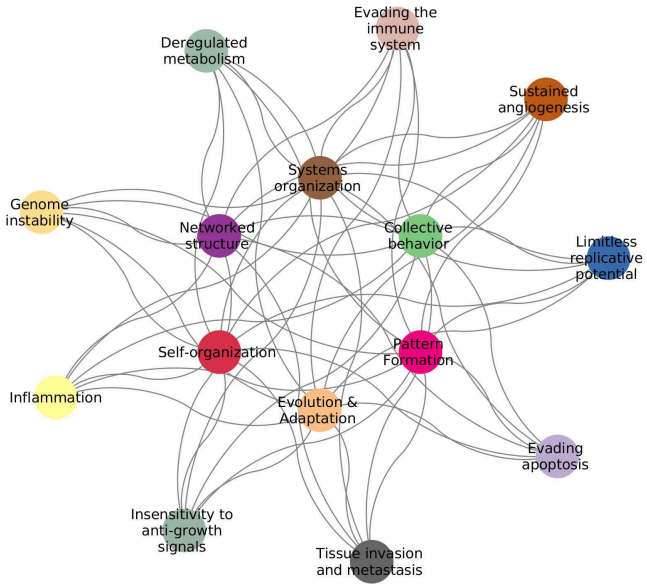

FIGURE 1 | Hallmarks of cancer complexity. The defining features of cancer (3, 4) are intrinsically connected to the defining features of complex systems (2).

The rapid development of sequencing strategies as well as genotyping and expression microarrays led to the development of gene models to account for the molecular aspects of biology at the whole cellular level (and even at the organ and organism scales). The coming of age and popularization (driven by an almost exponential lowering of the costs) of next gen sequencing techniques leads to an explosion of new approaches to understand complex phenotypes that in turn have sped up the rise of high throughput proteomics, metabolomics catching up. Single cell technologies and a number of arising sequence based approaches (ChIP-seq, ATAC-seq) are becoming usual tools of biomedical and in particular cancer research (see **Figure 3**, for an account of the fastly increasing number of PubMed publications based on these omic tools).

In spite of this, the integrative approach to multi-omic modeling is far from trivial due to the broad diversity of data types, dynamic ranges and sources of experimental and analytical errors characteristic of each omic. In spite of these facts, a number of approaches to multi-omic integration have been proposed [see, for instance, discussions in Hernández-Lemus (17, 18)]. Said approaches make use of tools from statistics, probability, machine

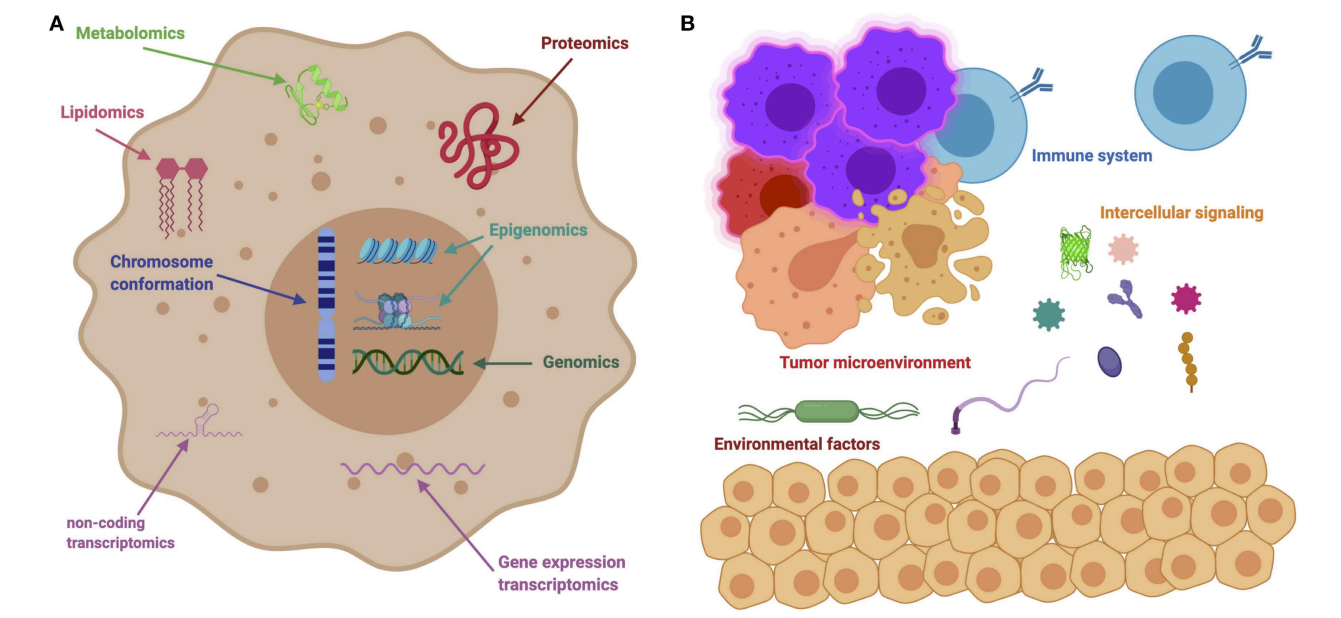

FIGURE 2 | The many levels of interactions found in a cancer system. **(A)** Depicts intracellular interactions that can be measured via the different omic technologies, such as genomics, transcriptomics, metabolomics, lipidomics, and so on. **(B)** Shows intercellular interactions, such as the ones orchestrated through immune responses, microbial interactions (metagenomics) and other instances of cell-cell interactions.

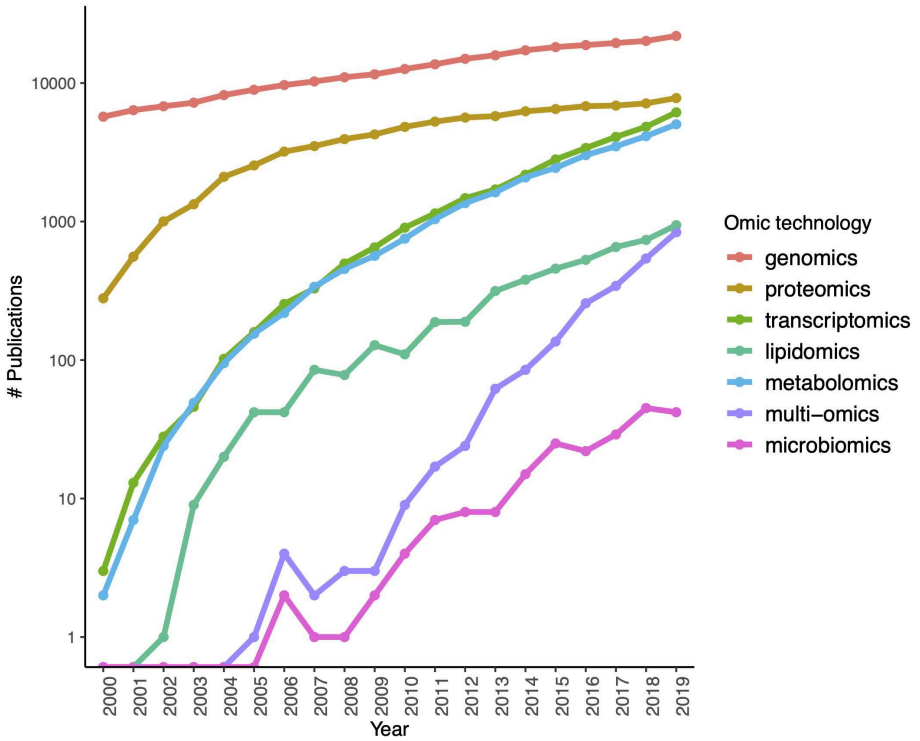

FIGURE 3 | Growth of interest in omics technologies in the twenty-first century: the number of Pubmed publications mentioning each omic technology in its title or abstract measured yearly since the year 2000.

learning and network science to classify, explore and provide guidelines for feature selection and their application is very much rooted in the tenets of systems biology.

The systematic study of cancer given by multi-omics is founded on the acknowledgment of a contribution of many different factors in the development and maintenance of the malignant state, including genetic aberrations, epigenetic alterations, changes in the response to cellular signaling, metabolic alterations, and beyond (19). Hence, by analyzing cancer as a complex pathology, the systems biology paradigm tries to gain insight into the molecular origins of the disease by looking at the diverse contributions, from DNA mutations (both germline and somatic), to deregulation of the gene expression programmes, the phenomenon of hormone disruption, that may or not be supplemented by metabolic abnormalities, and aberrant pathway signaling.

Cancer is also a multiscale pathology, aside from the biomolecular events just mentioned there is the influence of the environment and lifestyle that is known to be able to modify the onset, development, and outcome of tumors and their metastases. Multiomic analysis under a systems biology framework makes possible to use the unprecedented power of current high-throughput molecular and computational tools to draw a more complete figure of the different players in tumorigenesis and tumor establishment. At the same time, it may provide us with new instruments and strategies useful in basic and clinical research laboratories, but also in translational medicine and therapeutic endeavors.

These different levels of description have been independently studied for years. However, even if the advent of high-throughput technologies has permitted the development of systems biology, system-level models (conforming the theoretical foundations of these multiomic studies) are still under development.

2.2. The Systems Biology Framework

In essence, the foundational basis of systems biology is that of considering biological phenomena as systems, i.e., constructs formed by a large number of complex molecular and environmental components interacting at different levels to shape the functional features of said system. Tumor behavior, for instance, is determined by a combination of changes in genomic information that may (or may not) be associated with abnormal gene expression profiles; affecting protein abundance, but also modifying protein structure and folding, as well as supramolecular assembly. Changes in the regulatory patterns may also affect cell signaling mechanisms; and their responses. Hence, the complex interaction of nucleic acids and proteins in replication, transcription, metabolic, and signaling networks are considered the ultimate causes for the functioning (or misfunctioning, if preferred) of the tumor cell. We can notice that these are interdependent phenomena that cannot be treated separately, hence the need for integrative methodologies.

Another pivotal challenge in contemporary studies undertaken following a systems biology view is hence data integration. Data integration allows for the understanding of the enormous datasets generated by experimental multi-omics.

This is indeed a highly non-trivial task, since just the data management of such large amounts of information represents a challenge that has been called the big data paradigm.

3. THE ROLES OF COMPUTATION IN THE AGE OF CANCER MULTI-OMICS

We have identified four main roles that computation plays in the analysis of high-throughput data. These are the raw data acquisition from high-throughput instruments; the processing of raw data to quantitative data; the storage and management of massive omics data, for instance in remote repositories; and finally the deployment of data analysis models. These roles are illustrated in **Figure 4**. In this section, we will discuss select aspects of each of these roles.

3.1. Data Acquisition and Processing

The acquisition, processing, and manipulation of omic data generated in high throughput experiments requires, due to the very nature of these experiments (see **Figure 5**), the use of specialized bioinformatics pipelines. As the complexity of these datasets increases due to the natural evolution of these technologies, so do the associated challenges evolve (20). Bioinformatics workflow management systems can be used to develop, maintain, and foster reproducibility of a give pipeline or workflow. Examples of these systems include Galaxy (21), Snakemake (22), Nextflow (23), and the general purpose Common Workflow Language (24).

It should be noted that a large number of tools for omic data analysis are available as packages for the R language contained in the *Bioconductor* project (25), a repository of bioinformatics open source software. It is important, however, to acknowledge the existence of other software ecosystems, such as the *Biopython* project (26). Although the number of packages in Bioconductor is greater than that found in Biopython [see for instance (27)], the main takeaway should be that there is a large number of tools available to researchers that can be used in any combination suitable for their research question.

3.1.1. Genomics

The oldest of the omic technologies, genomic analyses focus on the genomic sequence and its variations: insertions, deletions (INDELs), single nucleotide variations (SNVs), copy number

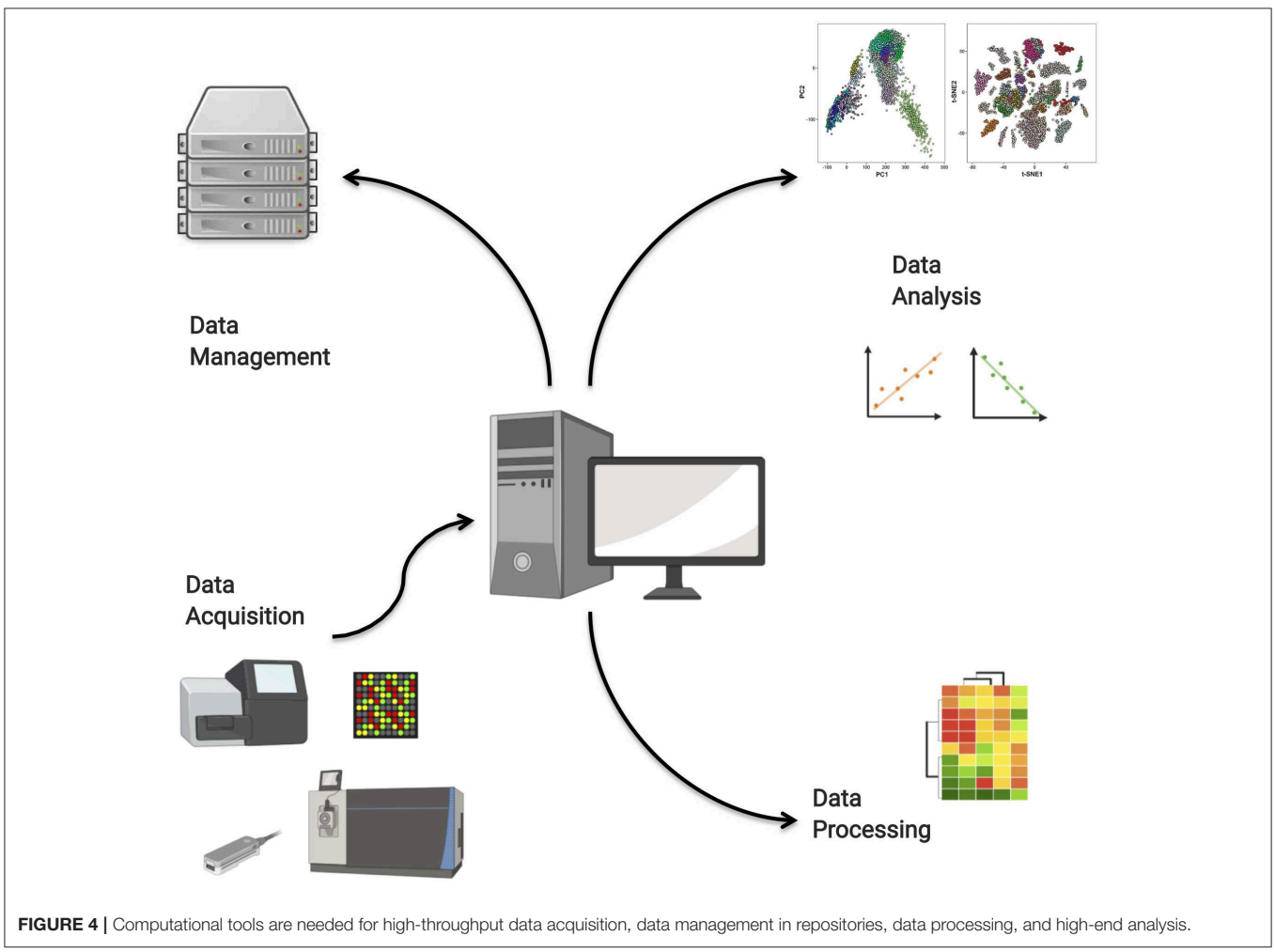

FIGURE 4 | Computational tools are needed for high-throughput data acquisition, data management in repositories, data processing, and high-end analysis.

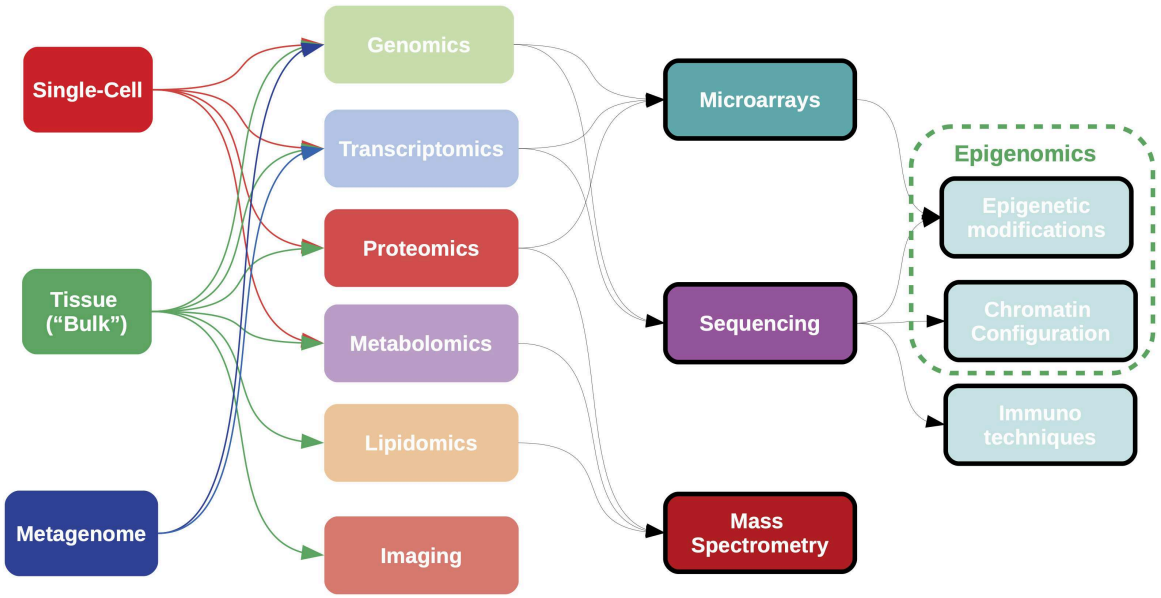

FIGURE 5 | Samples for omics analyses can be obtained from "bulk" tissue, single cell data, or heterogeneous populations, such as metagenomes. Most current omics data are generated using technologies either array-based, sequence-based, or mass spectrometry-based; although high-throughput imaging data is becoming important in the clinical setting. Complementary techniques exist for the analysis of epigenetic states. Each combination of sample type, omic measurement and analytical technology requires a specific bioinformatic pipeline for data acquisition and processing.

variations (CNVs), and so forth. The relationship between genomic alterations and cancer is well-known (28).

Microarrays have long been used for genotyping. Although specifics of microarray technology may vary across manufacturers, most modern DNA microarrays can be analyzed using well-established tools available in the *Arrays* (29). Such tools can handle arrays for different genotyping tasks, including SNP and copy number assays [for instance, copy number detection from exome sequencing using *CODEX* (30)].

Although DNA microarrays remain in use, next generation sequencing (NGS) technologies are quickly becoming commonplace. The analysis of NGS data entails a workflow that involves sequence acquisition and alignment to a reference genome, A number of downstream analysis pipelines can follow; for instance, a variant discovery workflow would involve variant calling, filtering, annotation, and prioritization (31). The first step to analyze NGS data is to use a sequence aligner tool on the sequence data (stored in FASTQ format). Some popular aligners are the stand-alone *BWA* (32), *Bowtie* (33), *Bowtie2* (34), and *SNAP* (35), with aligned sequences being stored in SAM (Sequence Alignment Map, text-based) or BAM (Binary Alignment Map) files. These aligned sequences are the input for downstream genotyping analyses (36, 37).

Such *standards* are indeed a matter of state-of-the-trade in the academic research community indeed. Regarding pipelines approved by regulatory instances, there is in fact an official FDA guideline document to this end: "Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing (NGS)—Based *in vitro* Diagnostics (IVDs) Intended to Aid in the Diagnosis of Suspected Germline Diseases" available for download at https://www.fda.gov/media/99208/download. The Guideline document (99208) actually refers to a Software Documentation Guideline: "General Principles of Software Validation; Final Guidance for Industry and FDA Staff" which is however quite outdated (last revised January, 11, 2002) (https://www.fda.gov/media/73141/download). Some NGS tools however are actually available as a web service at https://precision.fda.gov/. For a review on these guidelines and tools see (38).

3.1.2. Epigenomics

With the recent advent of high-throughput omic technologies to probe chemical modifications in the tumor genomes it has become more and more evident that such epigenomic modifications are present and likely play relevant roles in many cancers. These variations include DNA methylation and histone modifications, both in oncogenes and in other cancer-associated genes. Mutations in genes involved in epigenetic regulation have also been found in several tumor types. The computational analysis of epigenomic data may provide us new insights about cancer initiation and progression. More relevant perhaps, such studies will pave the way for a more efficient identification of genetic and epigenetic biomarkers for diagnosis, prognosis or response to therapy. These in turn, may accelerate the development of novel therapeutic approaches.

Epigenomics often presents another view of functional processes complementary to that of genomics. Sometimes epigenomic techniques even allow for a better understanding of genome-associated phenomena. Such is the case of high-throughput immunoprecipitation assays, such as ChIP-Seq.

ChIP-Seq and other experiments based on the analysis of short reads show the effects of multi-reads, i.e., reads that map to more than one genomic region. Determination of the origin of such multi-reads indeed results critical for the accurate mapping of reads to repetitive regions, such as copy number variants (39, 40). Current computational approaches have been refined to cover up for this phenomenon even at the single-cell level (41).

The epigenome contains the set of potentially inheritable chemical modifications of DNA and histone proteins that can control gene expression activity (42). There are several mechanisms which are contained within the epigenomics concept, each requiring a different high throughput molecular technique for its measurement. Each of these techniques, in turn, requires the use of a dedicated set of computational tools. These include:

- DNA methylation: The methylation state of a DNA region can alter its transcriptional activity. This state can be measured using either array-based methods or sequencing methods, such as the popular whole-genome bisulfite sequencing (WGBS) (43). Data from array based methods can be processed using the aforementioned array packages, along with dedicated packages, such as *methylationArrayAnalysis* (44). Similarly, those obtained using sequence-based methods can make use of dedicated tools, such as the *bsseq* (45) or *methyAnalysis* (46) packages.
- Chromatin remodeling: Regions where nucleosomes are sparse and physical access to the DNA sequence is enabled are identified as open chromatin. Chromatin accessibility is a dynamical and complex framework modulated by diverse elements, including nucleosome occupancy and turnover rate, histone modifications, ATP-dependent chromatin remodeling complexes and even TF binding (47, 48). Open chromatin has emerged as indicative of transcriptional regulatory potential or activity across the human genome because most of the TFs analyzed to date bind within open regions (49). Chromatin architecture is modified by changing its accessibility affecting gene expression rates. This remodeling can be controlled by histone modifications, which include acetylation, methylation, ubiquitination, and SUMOylation, among others. Overall chromatin accessibility can be also measured by techniques, such as ATAC-seq (50), a high throughput NGS technique to assess genome-wide chromatin accessibility. Due to the characteristic biochemical design of the assay ATAC-seq is a faster and more sensitive analysis of the chromatin accessibility than other alternatives, such as DNase-seq.

  ChIP-seq (51) data is used to identify genomic locations with an overabundance of proteins of interest; such identification uses the so-called *peak callers* (52, 53). These include *SICER2* (54), *PeakRanger* (55), *GEM* (56) *MUSIC* (57), *PePr* (58), *DFilter* (59), and *MACS* (60); benchmarks for these algorithms can be found at https://github.com/skchronicles/PeakCalling.

  MACS is a popular peak caller that uses dynamic Poisson distribution; its successor, *MACS2* (61), improves the algorithm to, amongst other things, make it more suitable for calling differential regions. Differential binding analysis (that is, identifying sites in which exhibit a different binding behavior between biological conditions) can be useful to identify relevant regions that may be driving cancer phenotypes, using ChIP-seq data. Tools for this task include *DiffBind* (62), a package that provides functions to handle the results of peak set callers, such as *MACS*. Another tool for this task is *csaw* (63), useful for de novo detection of differentially bound regions using a sliding window approach. In-depth comparison of differential ChIP-seq analysis tools can be found in (64).
- Chromosome conformation: The three-dimensional organization of the genome allows for interactions between regions that are distant in terms of sequence, even belonging to other chromosomes. These higher-order chromosome structures are a current area of research in oncology (65). Chromosome configuration capture techniques are able to quantify interactions between genomic loci. These *C-techs* are based on the original 3C, *Chromosome configuration capture* (66); able to quantify interactions between a single pair of loci. It was followed by: 4C (*Chromosome configuration capture-on-chip*) (67), which captures interactions between one locus and all others; 5C (chromosome conformation capture carbon copy) (68), which captures all interactions between two sets of loci; and Hi-C (high-resolution chromosome conformation capture) (69, 70) to detect interactions between all possible loci pairs. Development of computational analysis tools for chromosome conformation capture data is ongoing, although there are available packages for the detection of significant interactions for all these technologies (71–73).

It has been known for some time that higher order chromatin arrangements are associated with chromosomal alterations in cancer. For instance, it has been argued that spatial chromosome conformation and negative selection may be powerful driving forces behind somatic copy number alterations (74). More recently, chromatin conformation capture has allowed the identification of putative pharmacological targets in breast cancer (75). Genomic loci interactions may even affect the expression of biomarkers related to hallmarks of cancer, such as hypoxia (76).

Packages, such as *methylPipe* and *compEpiTools* provide an integral platform for the comprehensive and integrative analysis of the first two classes of epigenomic data (77), whereas *ATACseqQC* (78) is a package offering quality control tools for ATAC-seq data, while *esATAC* (79) offers a whole analysis pipeline and the *GenomicInteractions* package (80) offers a complete framework for the analysis of chromosome conformation data.

3.1.3. Transcriptomics

Transcriptomic analyses are used to measure the presence and abundance of RNA in a given physiological context (81). Perhaps the most common application of transcriptomic technologies is to measure gene expression. The gene expression profile of a phenotype can be used as a barcode of its biological state. Such barcodes can be compared, through differential expression analyses, to pinpoint cellular changes in cancers (82). The expression profile is the product of the gene regulatory program

encoded in the genome and the epigenome. By measuring gene expression, we are indirectly capturing the regulatory changes that are at the core of the disease.

The development of gene expression microarray technology (83) has made gene expression measurement more technically and economically viable than the measurement of protein abundance. Therefore, methods for the measurement of biological activity (i.e., pathways) have been developed with transcriptomic data in mind (84). Studying the molecular phenotype of cells via transcriptomics has become an invaluable tool providing a proxy to the functional state of cells and its regulatory interactions, both in cancer (85, 86), and in healthy phenotypes (87). Nevertheless, it should be noted that the correspondence between gene and protein abundance is far from perfect (88), which highlights the need for multi-omics.

Beyond gene expression, whole transcriptomic analyses involve the measurement of non-coding (nc) RNA, such as micro-RNA (miR), long non-coding RNAs (lnc-RNA), small nucleolar, Piwi-interacting, enhancer RNAs, among others (89, 90). The role of these transcripts, particularly in terms of their contribution to the regulatory program, remains an active area of study.

As previously mentioned, transcriptomic technologies are one of the most developed omics, second only to genomics itself. Measurement of transcript abundance can be done using either expression microarrays or RNA-sequencing (91, 92). Each methodology has technical considerations, but the general steps for their analyses are similar: acquire and preprocess data, removing technical artifacts; quality control; and data normalization. The resulting data can be represented as an expression matrix: an NxM matrix where rows represent transcripts, and columns represent samples (or observations). It should be noted that most expression pipelines are oriented toward differential expression analyses [see for instance (93)]; this should be taken into account in case that is not the intended use-case.

Starting points for RNA-seq data analysis include either alignment based methods, such as *Bowtie* (33), and *STAR* (94), or alignment-free methods, such as *kallisto* (95) and *Salmon* (96).

Cancer-related omic experiments often rely on specific, tailor-made analytics. One instance of this is provided by alignment-free RNA-Seq analysis methods, such as the ones performed by kallisto, Salmon, etc. Alignment-free methods (AFMs) are particularly well-suited to study cancer transcriptomics to look up at the role and abundance of fusion transcripts that may give rise to chimeric proteins (97, 98). Another reason behind the use of AFMs is that it is known that different RNASeq pipelines present differences that may be important when analyzing cancer genomes and transcriptomes (99, 100).

Further require different tools for quantification, quality control, and normalization of expression data. For instance, a popular pipeline is composed of the aforementioned *Bowtie* as a short read aligner, *TopHat* (101) for the identification of slice junctions, *Cufflinks* (102) for transcriptome assembly and differential expression analysis, and *CummeRbund* (103) for result exploration; it should be noted that, while this pipeline is still widely used and maintained (e.g., Bowtie2 latest release was 02/28/20), other approaches are been gradually embraced by the community (104); for instance, the *HiSat2* (105), *StringTie* (106), and *Ballgown* (107).

In the case of tools like STAR, we need to be aware that fusion detection using STAR-fusion is mainly limited by the length of single-end reads. The STAR-fusion wiki (https://github.com/STAR-Fusion/STAR-Fusion/wiki) indicates the need for at least 100 base length. In the case of other approaches, such as FusionHunter (108) the authors recommend to align to a pseudo-reference and discard junction spanning reads with <6 bp matches on either gene. *Arriba* is a relevant tool to call for gene fusions, based also in the STAR-alignment (https://github.com/suhrig/arriba/). Arriba was the winner of the DREAM SMC-RNA Challenge (https://www.synapse.org/#!Synapse:syn2813589/wiki/401435) (109).

An advantage of the modular design of these pipelines is that it is possible to combine tools from different workframes, depending on experimental and analytical needs: For instance, *Salmon* provides tools to connect with differential expression tools, such as *DESeq2* (110), *edgeR* (111), *limma* (112), or *sleuth* (113). A detailed discussion of these methods is beyond the scope of this article; please see Conesa et al. (114) for an in-depth review.

3.1.4. Proteomics

Proteomic analyses are used to identify and quantify the set of proteins present within a biological system of interest (115). The study of cancer proteomes is promising as a way of identifying biomarkers and therapeutic targets (116). This is not surprising: proteins are the molecular unit from which cellular structure and function arises.

Historically, high throughput proteomics technologies have developed at a slower pace than genomics and transcriptomics technologies. Microarray approaches to proteomics have been developed, with varied levels of success and applications (117, 118). However, the bigger breakthroughs have come through the use of mass spectrometry (119).

Various steps of proteomics analysis involve data analysis (120). During data acquisition, the detected molecular fragments must be identified. This is often done by comparing fragments to databases in real-time (121, 122). Later, the assembly of proteins from identified peptide fragments requires another set of computational methods (123). The development of such methods remains an active area of research (124, 125). The *Bioconductor* offers a streamlined set of tools for the management of proteomics data, from data processing to functional analysis (126). Another alternative for protein quantification is the *maxquant* toolset (127).

3.1.5. Metabolomics and Lipidomics

Metabolic alterations are important contributors to cancer development (128). Cancer metabolomics has become an important research topic in oncology (129), with the promise of providing novel insights on cancer development and potential therapeutic options. Lipidomics is actually a subset of metabolomics (130). The study of cancer lipidomics may lead to

the identification of biomedical important findings, such as novel biomarkers (131).

Like proteomics before, metabolomics and lipidomics studies have been possible thanks to the use of mass spectrometry. The analytical considerations for the extraction and quantification of these types of compounds have some differences to those used for proteomics. This is expected, as the chemical nature of metabolites and lipids are fundamentally different (132, 133). In turn, bioinformatic and chemoinformatic approaches to high-throughput metabolite profiling exhibit some modifications (134).

Analysis frameworks for metabolomic and lipidomic data are currently available. The *metab* package (135) provides an analysis pipeline for metabolomics derived from gas chromatography—mass spectrometry data. The *metaRbolomics* package (136) is a general toolbox that goes from data processing to functional analysis. Finally, the *lipidr* package (137) is a similar framework focused on lipidomics data.

3.1.6. Unraveling the Complexity Within Samples: Single Cell, Imaging, Microbiome

The aforementioned technologies were all developed for the detection and quantification of analytes extracted from a complex biological matrix, obtained from tissue, plasma, or a similar fluid. As such, the data from these omics is an aggregate of the different cellular contexts present in the sample. The environment within and surrounding cancer tumors is notably heterogeneous (138, 139). There is knowledge to be gained by recovering the omics diversity within samples.

Cancer is an extremely heterogeneous disease at the cellular and molecular level. Tumor heterogeneity caused by the concurrence of multiple cell lineages and differentiation stages, determined to an extent by the processes of clonal evolution. This has led to an early adoption of single cell analysis techniques. The case of single cell sequencing to study the genomic and epigenomic features of the different cell populations within a tumor by considering the characteristics of individual cells has revealed as an appealing approach to deal with said cell-to-cell variability (140–142).

Cancer cell heterogeneity also exists beyond the genome. Tumor evolution under complex environmental scenarios often leads to variability in epigenetic modifications. Single cell sequencing and imaging techniques have proven to be quite effective to characterize cellular plasticity induced by epigenomic phenomena (143). Aside from scMethSeq, and scDNAse Seq, other techniques, such as single-cell chromatin accessibility assays are starting to shed light to how epigenomic subpopulations in cancer may have the potential to impact tumor features, such as drug sensitivity and clonal dynamics (144).

Single-cell omics analyses rely on experimental techniques for the isolation of single cells from a sample, using microfluidics or fluorescence-activated cell sorting methods (145). Single-cell RNA-seq (scRNA-seq) is currently the most developed high-throughput omics technology for individual cell analysis (146).

Data from scRNA-seq experiments can be thought to be very similar to so-called "bulk" data. Data from scRNA-seq is, in fact, sparser, more variable, and with more complex expression values distributions. As such, data analyses techniques may need to account for different assumptions than their "bulk" counterparts (147). Again, the development of these novel bioinformatics tools is an active area of research (148). The *Bioconductor* ecosystem has a complete framework for the analysis of scRNA-seq from low-level (149) to functional analyses (150). *Scanpy* (151) provides a toolkit for single-cell gene expression analysis in a Python environment. Another single-cell genomics toolkit is *Seurat* (152) for R.

Integration of single-cell RNA-seq with other profiling tools is an important research area (153); as along with *single-cell*, there are other technologies that can provide a more complete picture of the cancer heterogeneity. High throughput imaging techniques (154) can be generated and computationally analyzed (155, 156). Imaging techniques can be used along with omics to recover the spatial distribution of molecules within cells and throughout tissues. Tools, such as *CellProfiler* (157) allow for a high-throughput analysis of data. Imaging techniques can be combined with single-cell methods: for instance, *MERFISH* can simultaneously measure copy number and distribution of RNA in single cells (158); *Slide-seq* (159) can measure transcriptomes at a high spatial resolution.

Space-resolved transcriptomics or spatial transcriptomics (ST) is a set of *in situ* transcript capturing methodologies aiming at quantification and visualization of gene expression patterns in individual tissue sections or regions. ST methods have indeed revealed relevant tissular phenomena linked to tumor evolution and in some cases have been able to allow the prediction of clinical outcomes in, for instance, breast cancer subtypes (160).

ST mapping of prostate tumors, on the other hand, have resulted key in the identification of gene expression gradients in stroma adjacent to tumor regions. This in turn has resulted in patient re-stratification based of tumor microenvironment features (161). A similar approach has been taken to trace tumor advance in malignant melanoma (162). A combination of ST with scRNASeq has led some researchers to propose the concept of a "tumor atlas," a roadmap to navigate tumor spatial and cellular heterogeneity (163).

Multi-omic analysis is not devoid of technical and logistic conundrums. Perhaps the most obvious is the availability of the different sample types from a single source in the same experiments. Cell cultures may provide a way out to this problem, however *in vitro* conditions are often not resembling some aspects of interest in complex phenotypes, such as cancer. In recent times, three dimensional cell culture techniques have allowed the design and development of more *realistic* models, such as the case of organoids and tumoroids. These models may represent a good compromise between cell line studies and biopsy-captured tissue experiments (164). Multi-omic approaches are starting to be applied on lab-grown organoids with relative success (165, 166). In order to analyze such data some novel computational tools are being developed and adapted (167).

The role of the immune system in cancer response is another area of active research. CITE-Seq is an RNASeq method that incorporates epitope analysis thus leading to semiquantitative information regarding surface protein abundance via antibody

assays, even at the single cell level (168). This novel technique is starting to be applied to provide the answer to fundamental questions in oncology, such is the case of tumorigenesis (169)

Finally, the role of the microbiome in cancer is being recognized (170); the integration of metagenomic, and perhaps *meta-omics* data (171), could provide key insights into cancer pathogenesis and therapeutics.

3.2. Data Management

The push for open data in the field of biomedical genomics since the gestation of the Human Genome Project has led to the emergence of a rich Genomic Commons (172). Making data available in public repositories makes for faster scientific discovery, although there are challenges to be overcome, both ethical/legal (173), and technological.

Challenges of data management include defining the type of data to be stored and how to store it; the policies for data access, sharing, and re-use; and long term archiving policies (174). Arguably, the most successful repository of cancer multiomics is NIHs Genome Data Commons (GDC) (175). The Genome Data Commons contains all data generated by the Cancer Genome Atlas (TCGA) project (176); although it should be noted that not all data is publicly accessible. The data is organized as a directed graph comprised of interconnected entities (**Figure 6**), with each entity having an associated set of properties and links. Data is publicly accessible either through the *gdc-client* command line tool, the REST API for programmatic access to the database, or through dedicated packages, such as *rtcga* (177). A recent account by *The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG)* of these resources and analyses is presented in (178). Furthermore, a larger collection of datasets can be accessed through the Broad Institute's *Firehose* (http://gdac.broadinstitute.org/); cloud computing enabled data access is provided through the Cancer Genome Collaboratory (https://cancercollaboratory.org/).

The impact of TCGA at the forefront of multiomics research is inarguable. As a publicly available resource, it provides data for method development and validation. This is used by a lot of current projects. However, there are other datasets with either single layer or multiomic datasets that can also be integrated. And wetlab researchers still carry out their projects, contributing to the cancer multiomics community. Integrating data from both, local experimental projects and large collaborative endeavors, such as TCGA is indeed a common practice in many places, such as our institution, the National Institute of Genomic Medicine in Mexico. Doing so allows to contrast specific hypothesis for the different research groups with the statistical power obtained via the much larger datasets generated by international multicentric collaborative projects.

As mentioned, it is possible to extract a lot of knowledge from the systematic re-analysis of data available in large public datasets. Perhaps, the more comprehensive of these databases is the one by the TCGA/Genome Data Commons/International Cancer Genome Consortium, TCGA. Retrieving the data via their Application Programming Interface (API) (https://gdc.cancer.gov/developers/gdc-application-programming-interface-api) demands some familiarity with command line tools and coding that may be beyond of most non-bioinformaticians. The project's data portal (https://portal.gdc.cancer.gov/) provides easy to use interfaces, but may be limited on its application to broader analyses. To date there is a number of commercially available platforms that provide a gentler access to the TCGA data. Such is the case of Qiagen's OncoLand database (https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/content-exploration-and-databases/qiagen-oncoland/) and the cloud-based analytics solution Seven Bridges

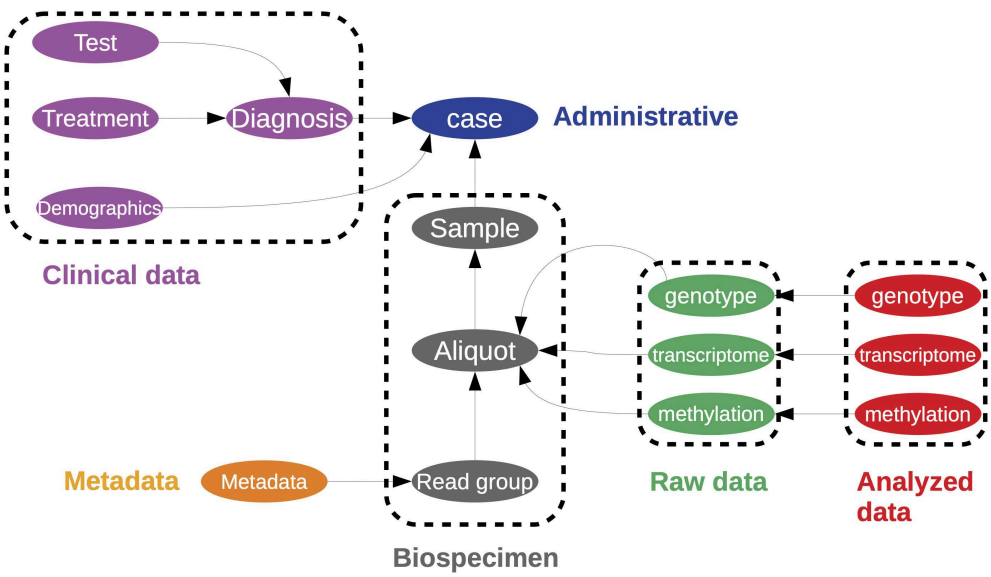

FIGURE 6 | A representation of the data structure used in to store the Cancer Genome Atlas within the Genome Data Commons. This is represented as a directed graph. This is a simplified illustration of the one found at https://gdc.cancer.gov/developers/gdc-data-model/gdc-data-model-components.

(https://docs.sevenbridges.com/docs/tcga-data). A limitation, aside from being subscription based alternatives that require a payment is that they are not customizable, which means that not all possible (nor desired) analysis may be performed.

There are, however a number of resources not only to access the data but to actually perform different levels of downstream analysis. Such is the case of imputation approaches to missing data in the TCGA database (179) (https://github.com/mrendleman/MachineLearningTCGAHNSC-BINF/).

Perhaps, the best combination of usability and versatility is present in the TCGA Workflow suite available as an R/Bioconductor package (180) (https://www.bioconductor.org/packages/release/workflows/vignettes/TCGAWorkflow/inst/doc/TCGAWorkflow.html).

4. COMPUTATIONAL TOOLS FOR MULTI-OMICS DATA INTEGRATION

An often-asked question is why try to integrate multiple omics technologies using complex models. Perhaps the simplest argument is that the biological phenomena is not comprised of independent layers of biological features: integrative models will be, due to this simple fact, closer to the system of study. As omics technologies become available, researchers have used them together to try and capture a better description of the phenomena (see **Figure 7**).

Improving our current cancer diagnostic capabilities is a major goal of biomedical research: the role of molecular technologies in the development of these tools has long been recognized (181). It is expected that multi-omic integration is able to provide better predictive tools than single molecular technologies, due to the fact that each technology is capturing just a slice of the whole complex pathological system; multi-omics data are expected to be of value for both basic and clinical research, as long as they are able to recover biological insights beyond those obtainable from the simple addition of each analysis layer (182, 183).

It may soon become evident that the formalisms that can lead to such level of description are, by necessity, complex (184). A remaining question is what multiomic combinations are able to achieve better diagnostic results. Selecting this optimal omics combination is not trivial, since there are practical constraints (such as economic and technical limitations) in the clinical setting in which such diagnostic tools are to be deployed (185). Computational tools and bioinformatic approaches play an important role in the design of such studies. A list of such tools is presented in Supplementary Materials as **Table 1**.

4.1. Multi-Omics Data Representation and Preparation

The success of a computational method could arguably be influenced by the design principles implemented in its data representation. The *MultiAssayExperiment* package (186) provides an eponymous data class to contain multi-omics experiments. Like other *Bioconductor* classes, *MultiAssayExperiment* is object-oriented. It can contain the information of different (multi-omics) experiments, linking features, patients, and experiments. Furthermore, by sharing design principles with the rest of the S4-*Bioconductor* classes, it is highly interoperable.

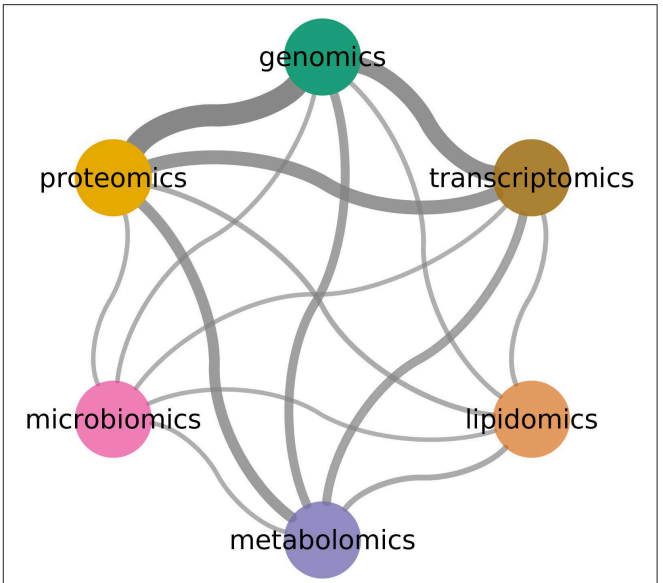

FIGURE 7 | Combinations between omics technologies. Width indicates number of co-occurrence in literature. Genomics, transcriptomics, and proteomics are the most common pairs.

An important issue with large scale multi-omics studies is the problem of missing and mislabeled samples. Whether by technical limitations or human error, the samples associated with a given patient may not have all measurements; or samples from two different patients may get mixed-up. There are packages available to handle these problems. The *missRow* package (187) can be used to handle missing data, combining multiple imputation with multiple factor analysis. The *omicsPrint* package (188), in turn, can be used to evaluate data linkage through the use of linear discriminant analysis.

The *STATegRa* (189) project provides a framework for multi-omics data analysis and integration: these are *MixOmics* (190), descended from the *integrOmics* project (191); and just like the *Bioconductor* project, the major advantage of such projects is the increased interoperability due to the sharing of design principles. For instance, within the *STATegRa* project, there is an Experiment Manager System (192); *MOSim* (193) a tool that provides methods for the generation of synthetic multi-omics datasets. These datasets can be used for the benchmarking and validating of other integration tools; and an experimental multi-omics dataset (194).

4.2. Multi-Omics Data Integration as a Data Science Problem

For this review, we approached these methods from a *data science* perspective, considering that each method is in essence solving a machine learning task (or set of tasks). In **Figure 8** we show

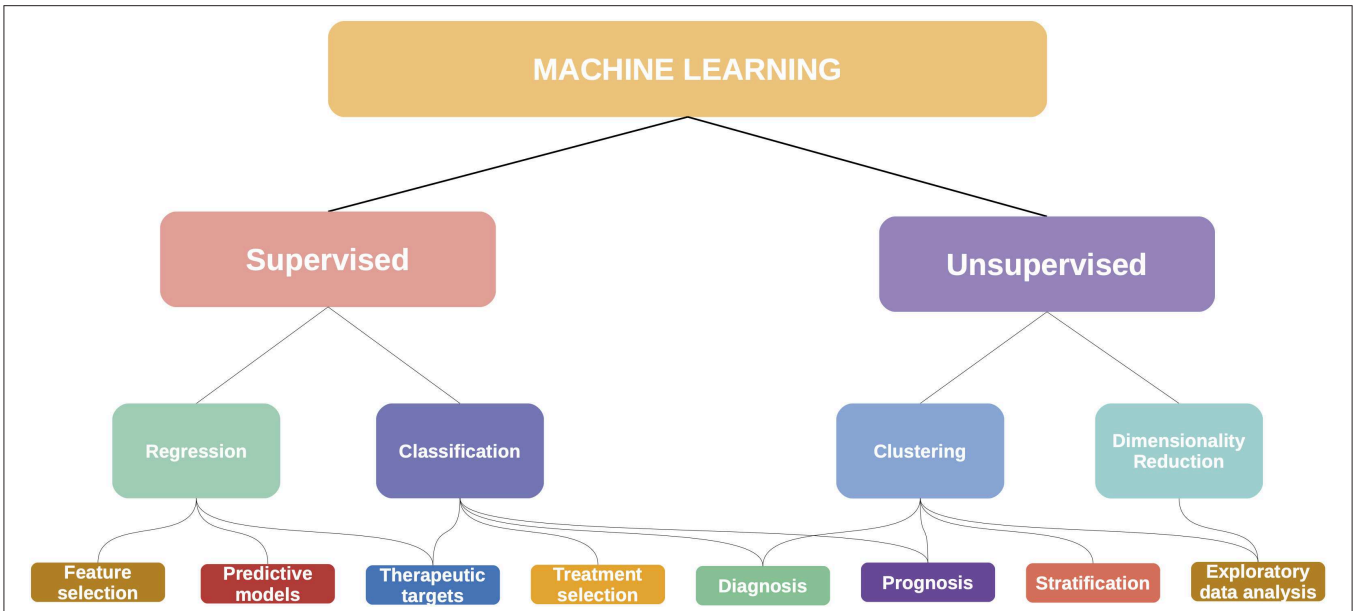

FIGURE 8 | Machine learning has many applications in cancer and multiomics.

some of these mappings, although it should be noted that these categories may be fluid: an unsupervised clustering analysis can become the basis for a supervised classifier, with diagnostic and prognostic applications. This is the story of the PAM50 algorithm for breast cancer (195).

4.3. Exploratory Data Analysis

Exploratory data analysis (EDA) is a vital first step in omics analyses (196). Through EDA the nature of the data can be understood, allowing for better decisions at a further modeling step.

Unsupervised learning approaches can provide a hypothesis-free understanding of the data behavior. This will reflect the nature of the underlying biological phenomenon. *Unsupervised clustering analyses* attempt to group samples based on the similarity of their measured features. The assumption is that this unsupervised classification will recover relevant biological differences. Multi-omics can increase the efficiency of such approaches (197).

Multi Omic data analysis is often performed with the aim of unveiling non-trivial molecular and systemic interactions that are difficult or impossible to see if one relies on a single omic approach. However, since we are tacitly assuming that the different omic levels of description may have synergistic effects that are key to develop more accurate models of tumor biology. Since multi omic approaches may generate a plethora of interdependent data it is useful to design analytical strategies for dimensionality reduction, feature selection and integration of all this information.

Aside from intelligibility, there are additional reasons to make dimensionality reduction schemes, one of these is that a multi omic study combines different information sources, hence dramatically increasing the number of features, often keeping the number of samples constant, in order to preserve statistical power we need to rely only on the most informative variables (198–200).

Computational tools to this end have been developed, such as the following: https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html https://bioconductor.org/packages/release/bioc/html/STATegRa.html For an extensive list of computational tools in the context of cancer biology, see (186).

One can make use of *dimensionality reduction techniques* in order to embed multi-omic data observations into a lower-dimensional space that can be used for either manual (i.e., visual) inspections or as the input for unsupervised clustering (or other analysis tools). Popular dimensionality reduction methods:

- Principal Component Analysis (PCA) is a classical (201) method based on an orthogonal transformation of the set of observations.
- T-distributed stochastic neighbor embedding (t-SNE) (202) is a method based on the minimization of the Kullback-Leibler divergence between the probability distribution of pairs of high-dimensional objects.
- The Uniform Manifold Approximation and Projection (UMAP) (203) is a non-linear technique in which data are projected into a Riemannian manifold.

Data visualization is an important part of EDA: the graphical representation of data can be sufficient for the identification of complex patterns (204). Visualizing high-dimensional biological data can be helpful from a purely data-driven point of view: for instance, to understand the variability within a phenomenon. Combinations of dimensionality reduction, data clustering, and visual inspection can be effective to identify subpopulations within a dataset. The most common visualization for these tasks is perhaps the scatterplot, but it is far from the only: for instance,

hexbins (205) can be used to explore sc-RNAseq data, which can be useful to overcome overplotting problems related to the order in which points are drawn in the canvas.

Visualization can also be coupled with other biological information, for instance locating the genomic regions in which epigenomic features are found. Visualizations, such as the *Circos* plot (206) can be used for the detailed representation of multi-omics data and their location in specific genomic regions; The *omicCircos* (207) implementation is compatible with the standard data classes used in*Bioconductor*. The multiOmicsViz *multiOmicsViz* package is useful to visualize the effects of one omics layer to another, visualized in within the spatial chromosome context. The *Gviz* package (208) provides a full R graphics system solution for genome browser-style visualizations. Such representation is useful to represent the behavior of different experimental layers (as tracks) in a sequence context. For ChIP-seq data visualization, tools like *PAVIS* (209) may be used. Single Cell RNA-seq data visualization suites, such as *SingleCell Signature Explorer* (210) can be useful for exploratory analysis of such datasets. In the case of chromatin capture data, visualization toolboxes, such as HiBrowse (211), the Epigenome Browser (212), and Juicebox (213). For a thorough review of Hi-C visualization consult (214).

Common exploratory data analysis tools are implemented either in base R or as packages from CRAN (since their use is not necessarily limited to biological data). However, there are packages providing integrated EDA tools for multi-omics and oncology. The *OMICsPCA* package (215) provides omics-oriented tools for PCA analysis. The *CancerSubtypes* package (216) contains several data preprocessing, quality control, and clustering methods, focused on the identification of cancer subpopulations from multi-omics data. *Biocancer* (217) provides an interactive multi-omics data exploratory toolkit. The *omicade4* package (218) provides an implementation of multiple co-inertia analysis (MCIA), another dimensionality reduction technique; these tools were used for the integration of transcriptome and proteome data from the NCI-60 cancer cell line panel. The Multi-omics Autoencoder Integration (maui) is a tool for multi-omics data analysis for Python. It allows for latent factor model coupled with artificial neural networks for multiomics data integration. *iClusterPlus* is a Bioconductor package based on the original iCluster (219) algorithm for integrative cluster analysis combining different types of genomic data.

4.4. Statistical Models: Classificators, Predictors, and Feature Selection

Exploratory methods provide a useful description of biological phenomena. Nevertheless, in the oncology context, the identification of actionable elements is most desired, to generate translational value. The generation of models and feature selection strategies can lead to such results.

In this context, *statistical models* are computational (and thus mathematical) representations of the relationships between observed variables. These models can be useful to solve a given task based on some input data (220). Examples of these tasks include the *classification* of samples and the *prediction* of the state of a feature of interest.

Classification models have important biomedical applications (185). If a classification is able to discriminate between physiological states it can have translational use: A model that discriminates between health and disease has *diagnostic* utility; A model that discriminates between different disease outcomes has *prognostic* utility, which can be used for *stratification* purposes. Molecular classifiers have been quite successful in oncology: perhaps the best example being breast cancer (221). Classification models can be developed using *supervised* methods (that is, the model is trained with class information); but *unsupervised* methods, such as the previously discussed clustering, may be able to recover groupings that capture biological and clinical differences.

Predictive models can provide insights into the molecular mechanisms driving physiological states. These can reveal the interactions between different omics, as well as between individual biomolecules. Furthermore, predictive models can have translational applications, including their use in prognostic tools (222).

Feature selection consists in the selection of a subset of measured variables that are most informative: that is, they contribute the most for the model to accomplish its task. Proper feature selection is important for biomedical models (223), as (1) removing uninformative ("irrelevant" or "redundant") features simplifies the model and increases its performance; and (2) a smaller set of features is less expensive to measure, increasing the translational potential of a given model.

Common applications of statistical models in the clinical context of cancer are the prediction of susceptibility, recurrence, and survival (223). Additionally, classification and association models are regularly used for the interpretation of molecular studies of cancer. For instance, biomarker discovery (224) is an often sought target for modeling based on biochemical and multi-omics analyses. This is an important area of study, since actionable biomarkers are not particularly common (225).

4.4.1. Implementations and Use-Cases

Novel tools for the implementation of oncology models using model data are being released constantly. Many of these packages combine exploratory, supervised, and unsupervised tools, providing a wide range of analysis tools. *mixOmics* (190) is a self-described omics data integration project; it includes an eponymous package that provides different exploratory and integrative multivariate methods, including (independent) PCA, Canonical Correlation Analysis, Partial Least Squares regression (PLS), and PLS-Discriminant Analysis (DA). Part of the larger project is the *Data Integration Analysis for Biomarker discovery using Latent Variable approaches for 'Omics studies* (DIABLO) framework, which has been used for the identification of a multi-omics signature of breast cancer molecular subtypes (226).

Other tools also follow this combined design principle. The *ropls* package (227), for instance, incorporates the tools for PCA, as well as (Orthogonal) PLS. Multi-Omics Factor Analysis (MOFA) is implemented in the eponymous package (228).

This factor analysis model has been used for the unsupervised detection of groups in a leukemia dataset, and the selection of informative multi-omic features associated with oxidative stress. *OmicsMarkeR* (229) also provides a variety of classification and feature selection tools; originally developed for metabolomics, this tool has been used for the study skin cancer progression (230). Some packages include different classifier methods to generate an ensemble model; such is the case of *Biosigner* (231) which combines PLS-DA, Random Forests, and Support Vector Machines to select discriminant features across omics.

We agree with the assumption that multi-omics specific tools can improve workflows by adhering to a single design philosophy. However, we also agree that this is convenient, but not necessary. For instance, a diagnostic panel for pancreatic cancer was recently identified with a Random Forest implementation (232) using genomics, transcriptomics, and immunohistochemistry data. In another study, biomarker candidates for pancreatic cancer are identified using a Support Vector Machine on miRNA and gene transcriptomics (233).

Predictive models can be used to identify the contribution of one omics layer to the activity of another. For instance, *epigenomix* (234) uses Bayesian mixture models to integrate ChIP-seq and gene transcription data. The *Integrative analysis of Multi-omics data for Alternative Splicing* (235) package integrates expression, sQTLs, and methylation to provide mechanistic insights behind the manifestation of alternative splicing.

Predictive methods have been used to integrate multi-omics with other sources of big data, with publicly available implementations. The packages *rexposome* and *omicRexposome* (236) have been used to study the *exposome*, defined as the set of environmental exposures. Using multi-canonical correlation analyses and multiple co-inertia analysis, exposome-wide associations have been made to multi-omic data. The *OmicsLonDA* package (237) offers a method that uses linear mixed-effect models and smoothing spline regression models to identify time periods with differential omics levels. A highlight of this package is the consideration for the use of physiological measurements from wearable sensors, which may provide applications for *nowcasting*, the prediction of near-future states.

4.4.2. Functional Aggregation

One could argue that analysis methods can be more informative if there is a way of associating the findings to the wider body of biomedical knowledge. Mapping omics data to functional features, such as pathways and functional genesets, is a strategy that can provide such readily interpretable results. *Functional enrichment* approaches, such as *over-representation analysis* (ORA) and *gene-set enrichment analysis* (GSEA), are effectively *feature extraction* methods that can be used as biologically relevant dimensionality reduction methods. The results of such methods can serve as starting points for more complex models, such as interactions among functions (238). For a detailed discussion of functional analysis, see (84).

The development of methods for effective functional enrichment based on multi-omics data is ongoing. *Multi-omics gene-set analysis* (MOGSA) (239) approaches the problem by using multivariate analysis, and using projections of data and genesets to lower dimensional spaces, to generate an enrichment score. *Massive integrative gene set analysis* (MIGSA) (240) takes a different approach, making independent functional associations for each omics layer (using ORA and Functional Class Scoring). Instead of providing an aggregated measurement, the functional associations of each layer are stored in a special data structure, allowing flexible analyses. This method has been used to functionally characterize breast cancer molecular subtypes from a multi-omics perspective.

Functional aggregation can be used as the basis for other data analysis tasks. In *pathwayPCA* (241), exploratory data analysis is done by analyzing the functional enrichment of each omics set separately, and aggregating them via consensus. This method was used to study heterogeneity in an ovarian cancer dataset. In the original work for the *Divergence analysis* (242) method for high-dimensional omics data analysis, the authors evaluate the effect of using functional aggregation for their data classification task. Functional aggregation methods are an important part of high-throughput drug initiatives, as can be seen by their prominence in the iLINCS platform (243).

4.5. The Network Paradigm

As we have stated throughout this work, biological phenomena are complex, interconnected systems. The data that we recover from high-throughput multi-omics is not isolated. Any biological system is not just the sum of its parts, but the sum of its biological elements *and their relationships*. With this in mind, the integration of high-throughput data within a network paradigm becomes appealing. Some advantages of a network approach to multi-omics integration are:

- A network representation of multi-omics data can be studied using all the foundations and tools of network science (244). Network topological parameters can be associated with important biological features; furthermore, dynamical processes can be modeled over networks.
- As previously noted, the functional level of biological description is fundamentally composed of molecular interactions. In other words, measurable functions can be thought to emerge from biological networks. Functional analyses can benefit from considering the way in which the participating molecules interact.
- The integration of interaction information can lead to more informative models (245).

A network perspective can enhance every aspect of the multi-omics analysis. For instance, mapping omics data to pathway networks can provide an opportunity to biologically contextualize the data. A classic tool for this is the *pathview* (246) package. The *Graphite* (247) package is a more flexible alternative, as it allows the visualization of pathways from different data sources, and provides proper graph objects that can be manipulated using network visualization tools. Recently, the *metaGraphite* package provided a major update to the original tool, effectively incorporating multi-omics through the addition of a metabolomics layer.

Network approaches can be used for classification and prognosis. For instance, the *micrographite* (248) package provides

a method to integrate micro-RNA and mRNA data through their association to canonical pathways. This approach has been useful in identifying key micro-RNAs in myeloma (249), primary myelofibrosis (250), and ovarian cancer (251). *Mergeomics* (252) integrates data from genomic, epigenetic, and transcriptional association studies through a functional enrichment method, the results of which are used as the basis for a network construction; however, this tool has not been used in a cancer context. *pwOmics* (253) is another tool that leverages biological network knowledge to integrate multi-omics data. In particular, this tool is well-suited for the study of time series analyses.

While mapping data to predefined networks can be useful to gain a much-needed biological context, high-throughput technologies offer the opportunity to actually *infer* networks from the data itself. With such approach, data analysis problems can be transformed into network analysis problems. For instance, feature clustering becomes network module detection, which can be then used as the basis for a functional enrichment analysis (254).

While network reconstruction from omics data can be a powerful tool, it should be stated that every network reconstructed from data has an underlying hypothesis, which defines what the links between elements represent. This hypothesis should be at the center of any interpretation of the topological or functional associations recovered from a network. Furthermore, one must remember that comparison between reconstructed networks of different biological conditions will yield information about biological differences only if the method for network reconstruction does not deviate for each condition. For a discussion on this subject, see (255). This point is particularly relevant when discussing multi-omics data integration, as many of the network reconstruction methods available were developed for gene expression data. Proper validation of a method should be conducted before using it with other types of data.

There are some recent implementations of network reconstruction methods that have been developed with multi-omics data in mind. MAGIA^2 (256) is a tool for the reconstruction of micro-RNA and transcription factor regulatory circuits; it has been used for the analysis of expression regulation in the NCI60 cell panel. The *Discordant* method (257) uses a mixture model to identify differential correlation: that is, statistical dependencies between feature pairs that are lost or gained from one biological state or another. This method has been evaluated for its use with different types of omics data. The *Netboost* (258) is a network reconstruction method infers statistical dependency based on multi-omics data, and uses a modularity approach to reduce dimensionality; the method has been used for the classification and survival analysis of acute myeloid leukemia data. *AMARETTO* (259) identifies pairwise relationships between different omic layers to select cancer driver genes. A module detection approach is used to construct a dimensionally reduced module network, which is further analyzed to identify molecular signatures.

Probabilistic network reconstruction is a powerful data analysis technique. In such a model, features are connected based on an information-theoretical similarity measure, such as mutual information, between their expression profiles. Unlike correlation metrics (260), mutual information can capture non-linear relationships between features, which makes it suitable for the analysis of transcriptomics (261). We have applied these methods for the reconstruction of micro-RNA and gene co-expression bipartite networks with minor adjustments; the analysis of such networks has yielded interesting insights on the nature of functional control by micro-RNAs (262). A current research interest the authors of this work is the extension of probabilistic network reconstruction for multi-omics reconstruction, in order to construct *probabilistic multilayer networks* (263) that can be studied using the recent tensorial formalism of multilayer networks (264).

4.6. Data Science in Biology—A Word of Warning

An important aspect of any data science project is the crucial role of both technical and domain specific expertise. The analysis of biological networks in particular can pose some complication for biological scientists not familiar with the field of network science; a network visualization may be presented as result, without an adequate evaluation of network topology or other structural and dynamic parameters. Similar behaviors can be found with other applications of data science tools.

A data-driven analysis without the participation of a domain expert risks the pursuit of non-relevant questions. On the other hand, even though a bioinformatics tool may be developed with an increased usability in mind, the level of complexity of both the computational method may require a deeper understanding of the algorithm's assumptions and limitations in order to reach valid results. With this in mind, it is evident that proper computational approaches to biological questions require a fundamental understanding of both in order to reach scientifically solid conclusions. In many cases, the key to achieve this is to strive for multidisciplinary approaches.

5. CONCLUSION

Cancer is the paradigmatic complex phenotype. We have been able to capture some of this complexity via experimental measurements with the different high throughput biomolecular technologies generically termed *omics*. Each single-technology derived data type has its own set of caveats and complexities. An additional challenge lies in the fact that each data type is able to account for a fraction of the large set of cancer aspects or features. Recent times have witnessed the development of new ways to gather and analyze these partial information layers together, under the name of multi-omics.

There are, however, multiple approaches to multi-omic computational modeling and integration, some of the most relevant have been described and discussed here. Our aim has been that of presenting the current state of the art of computational oncology tools for multiomic studies of complex cancer phenotypes. Novel developments in the multiomic computational analysis come from different fields, ranging from purely mathematical developments (263, 264), to machine learning and computational intelligence applications (179, 223), to single-cell sequencing and imaging studies (139, 145) and

more. However, in our view, the development of methods to integrate all these different analytical approaches into intelligible and statistically robust frameworks will provide the field with unprecedented advances both in our understanding of cancer biology and in our impact in the clinical settings. The field is fast-growing and currently under development, with novel algorithmic approaches being constantly released, but we believe that the present account is a good starting point.

AUTHOR CONTRIBUTIONS

GA-J and EH-L contributed to reviewing and classifying the literature, structured the review, prepared the figures, wrote, and revised the manuscript. EH-L contributed to funding and general oversight of the project.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México. EH-L was recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Laura Lucila Gómez Romero (INMEGEN) for a recent discussion on current sequence-based methods. **Figures 2**, **4** were generated using Biorender (https://biorender.com/). **Figure 4** includes images from Wikipedia, released under a Creative Commons Attribution-Share Alike 3.0.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00423/full#supplementary-material

REFERENCES

1. Knox SS. From "omics" to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int.* (2010) 10:11. doi: 10.1186/1475-2867-10-11
2. Sayama H. *Introduction to the Modeling and Analysis of Complex Systems*. Geneseo, NY: Open SUNY Textbooks (2015). Available online at: http://textbooks.opensuny.org/introduction-to-the-modeling-and-analysis-of-complex-systems/
3. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* (2000) 100:57–70. doi: 10.1016/S0092-8674(00)81683-9
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* (2011) 144:646–74. doi: 10.1016/j.cell.2011.02.013
5. McAllister K, Mechanic LE, Amos C, Aschard H, Blair IA, Chatterjee N, et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *Am J Epidemiol.* (2017) 186:753–61. doi: 10.1093/aje/kwx227
6. Barriga V, Kuol N, Nurgali K, Apostolopoulos V. The complex interaction between the tumor micro-environment and immune checkpoints in breast cancer. *Cancers.* (2019) 11:1205. doi: 10.3390/cancers11081205
7. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev.* (2018) 32:1267–84. doi: 10.1101/gad.314617.118
8. Zhou JX, Taramelli R, Pedrini E, Knijnenburg T, Huang S. Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. *Sci Rep.* (2017) 7:8815. doi: 10.1038/s41598-017-09307-w
9. Brabletz T, Jung A, Reu S, Porzner M, Hlubek F, Kunz-Schughart LA, et al. Variable β-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. *Proc Natl Acad Sci USA.* (2001) 98:10356–61. doi: 10.1073/pnas.171610498
10. Kammula US, Kuntz EJ, Francone TD, Zeng Z, Shia J, Landmann RG, et al. Molecular co-expression of the c-Met oncogene and hepatocyte growth factor in primary colon cancer predicts tumor stage and clinical outcome. *Cancer Lett.* (2007) 248:219–28. doi: 10.1016/j.canlet.2006.07.007
11. Van Gool B, Dedieu S, Emonard H, Roebroek AJ. The matricellular receptor LRP1 forms an interface for signaling and endocytosis in modulation of the extracellular tumor environment. *Front Pharmacol.* (2015) 6:271. doi: 10.3389/fphar.2015.00271
12. Terra M, Oberkampf M, Fayolle C, Rosenbaum P, Guillerey C, Dadaglio G, et al. Tumor-derived TGFβ alters the ability of plasmacytoid dendritic cells to respond to innate immune signaling. *Cancer Res.* (2018) 78:3014–26. doi: 10.1158/0008-5472.CAN-17-2719
13. Mayers JR, Vander Heiden MG. Nature and nurture: what determines tumor metabolic phenotypes? *Cancer Res.* (2017) 77:3131–4. doi: 10.1158/0008-5472.CAN-17-0165
14. Davidson SM, Papagiannakopoulos T, Olenchock BA, Heyman JE, Keibler MA, Luengo A, et al. Environment impacts the metabolic dependencies of Ras-driven non-small cell lung cancer. *Cell Metab.* (2016) 23:517–28. doi: 10.1016/j.cmet.2016.01.007
15. Serrels A, Lund T, Serrels B, Byron A, McPherson RC, von Kriegsheim A, et al. Nuclear FAK controls chemokine transcription, Tregs, and evasion of anti-tumor immunity. *Cell.* (2015) 163:160–73. doi: 10.1016/j.cell.2015.09.001
16. Hernández-Lemus E, Reyes-Gopar H, Espinal-Enríquez J, Ochoa S. The many faces of gene regulation in cancer: a computational oncogenomics outlook. *Genes.* (2019) 10:865. doi: 10.3390/genes10110865
17. Hernández-Lemus E. Systems biology and integrative omics in breast cancer. In: Barh D, editor. *Omics Approaches in Breast Cancer*. New Delhi: Springer (2014). p. 333–52. doi: 10.1007/978-81-322-0843-3_17
18. Hernández-Lemus E. Further steps toward functional systems biology of cancer. *Front Physiol.* (2013) 4:256. doi: 10.3389/fphys.2013.00256
19. Du W, Elemento O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. *Oncogene.* (2015) 34:3215–25. doi: 10.1038/onc.2014.291
20. Davis-Turak J, Courtney SM, Hazard ES, Glen WB, da Silveira WA, Wesselman T, et al. Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Rev Mol Diagn.* (2017) 17:225–37. doi: 10.1080/14737159.2017.1282822
21. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* (2010) 11:R86. doi: 10.1186/gb-2010-11-8-r86
22. Koster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics.* (2012) 28:2520–2. doi: 10.1093/bioinformatics/bts480
23. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* (2017) 35:316–9. doi: 10.1038/nbt.3820

24. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al. *Common Workflow Language, v1.0*. (2016). Available online at: https://www.commonwl.org/
25. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*. (2015) 12:115–21. doi: 10.1038/nmeth.3252
26. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. (2009) 25:1422–3. doi: 10.1093/bioinformatics/btp163
27. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: A sustainable and comprehensive software distribution for the life sciences. *Nat Methods*. (2018) 15:475–6. doi: 10.1038/s41592-018-0046-7
28. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. (2009) 458:719–24. doi: 10.1038/nature07943
29. Maintainer BP. *arrays: Using Bioconductor for Microarray Analysis*. Washington, DC (2019).
30. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*. (2015) 43:e39. doi: 10.1093/nar/gku1363
31. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. *J Mol Diagn*. (2018) 20:4–27. doi: 10.1016/j.jmoldx.2017.11.003
32. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324
33. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. (2009) 10:R25. doi: 10.1186/gb-2009-10-3-r25
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. (2012) 9:357–9. doi: 10.1038/nmeth.1923
35. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. *Faster and More Accurate Sequence Alignment with SNAP*. (2011). Available online at: http://arxiv.org/abs/1111.5572v1; http://arxiv.org/pdf/1111.5572v1
36. Magis AT, Funk CC, Price ND. SNAPR: a bioinformatics pipeline for efficient and accurate RNA-seq alignment and analysis. *IEEE Life Sci Lett*. (2015) 1:22–5. doi: 10.1109/LLS.2015.2465870
37. Arora S, Morgan M. *Sequencing: Introduction to Bioconductor for Sequence Data*. Washington, DC (2019).
38. Luh F, Yen Y. FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine. *NPJ Genom Med*. (2018) 3:1–3. doi: 10.1038/s41525-018-0067-2
39. Zhang Q, Keleş S. CNV-guided multi-read allocation for ChIP-seq. *Bioinformatics*. (2014) 30:2860–7. doi: 10.1093/bioinformatics/btu402
40. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS ONE*. (2013) 8:e65598. doi: 10.1371/journal.pone.0065598
41. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet*. (2019) 51:1060–6. doi: 10.1038/s41588-019-0424-9
42. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. *Cell*. (2007) 128:669–81. doi: 10.1016/j.cell.2007.01.033
43. Fan S, Chi W. Methods for genome-wide DNA methylation analysis in human cancer. *Brief Funct Genomics*. (2016) 15:432–42. doi: 10.1093/bfgp/elw010
44. Maksimovic J. *methylationArrayAnalysis: A Cross-Package Bioconductor Workflow for Analysing Methylation Array*. Washington, DC (2019).
45. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*. (2012) 13:R83. doi: 10.1186/gb-2012-13-10-r83
46. Du P, Bourgon R. *methyAnalysis: DNA Methylation Data Analysis and Visualization*. Washington, DC (2019).
47. Bell O, Tiwari VK, Thomä NH, Schübeler D. Determinants and dynamics of genome accessibility. *Nat Rev Genet*. (2011) 12:554. doi: 10.1038/nrg3017
48. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet*. (2019) 20:207–20. doi: 10.1038/s41576-018-0089-8
49. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature*. (2012) 489:75. doi: 10.1038/nature11232
50. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. (2015) 109:21–9. doi: 10.1002/0471142727.mb2129s109
51. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*. (2007) 316:1497–502. doi: 10.1126/science.1141319
52. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*. (2009) 6:S22–32. doi: 10.1038/nmeth.1371
53. Sarkar D, Gentleman R, Lawrence M, Yao Z. *chipseq: A Package for Analyzing Chipseq Data*. Washington, DC (2019).
54. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. (2009) 25:1952–8. doi: 10.1093/bioinformatics/btp340
55. Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*. (2011) 12:139. doi: 10.1186/1471-2105-12-139
56. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*. (2012) 8:e1002638. doi: 10.1371/journal.pcbi.1002638
57. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol*. (2014) 15:474. doi: 10.1186/s13059-014-0474-3
58. Zhang Y, Lin YH, Johnson TD, Rozek LS, Sartor MA. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*. (2014) 30:2568–75. doi: 10.1093/bioinformatics/btu372
59. Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat Biotechnol*. (2013) 31:615–22. doi: 10.1038/nbt.2596
60. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. (2008) 9:R137. doi: 10.1186/gb-2008-9-9-r137
61. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*. (2012) 7:1728–40. doi: 10.1038/nprot.2012.101
62. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. (2012) 481:389–93. doi: 10.1038/nature10730
63. Lun ATL, Smyth GK. *De novo* detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res*. (2014) 42:e95. doi: 10.1093/nar/gku351
64. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform*. (2016) 17:953–66. doi: 10.1093/bib/bbv110
65. Jia R, Chai P, Zhang H, Fan X. Novel insights into chromosomal conformations in cancer. *Mol Cancer*. (2017) 16:173. doi: 10.1186/s12943-017-0741-5
66. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science*. (2002) 295:1306–11. doi: 10.1126/science.1067799
67. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, De Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*. (2006) 38:1348. doi: 10.1038/ng1896
68. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*. (2006) 16:1299–309. doi: 10.1101/gr.5571506
69. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions

reveals folding principles of the human genome. *Science.* (2009) 326:289–93. doi: 10.1126/science.1181369
70. Van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*. (2010) 39:e1869. doi: 10.3791/1869
71. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B. r3Cseq: an R/bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res.* (2013) 41:e132. doi: 10.1093/nar/gkt373
72. Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EEM, Huber W. FourCSeq: analysis of 4C sequencing data. *Bioinformatics*. (2015) 31:3085–91. doi: 10.1093/bioinformatics/btv335
73. Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen CJ, Heard E, et al. HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics*. (2012) 28:2843–4. doi: 10.1093/bioinformatics/bts521
74. Fudenberg G, Getz G, Meyerson M, Mirny LA. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol*. (2011) 29:1109. doi: 10.1038/nbt.2049
75. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat Commun*. (2018) 9:1–13. doi: 10.1038/s41467-018-03411-9
76. Stone JK, Kim JH, Vukadin L, Richard A, Giannini HK, Lim STS, et al. Hypoxia induces cancer cell-specific chromatin interactions and increases MALAT1 expression in breast cancer cells. *J Biol Chem*. (2019) 294:11213–24. doi: 10.1074/jbc.RA118.006889
77. Kishore K, de Pretis S, Lister R, Morelli MJ, Bianchi V, Amati B, et al. methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. *BMC Bioinformatics*. (2015) 16:313. doi: 10.1186/s12859-015-0742-6
78. Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*. (2018) 19:3. doi: 10.1186/s12864-018-4559-3
79. Wei Z, Zhang W, Fang H, Li Y, Wang X. esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis. *Bioinformatics*. (2018) 34:2664–5. doi: 10.1093/bioinformatics/bty141
80. Harmston N, Ing-Simmons E, Perry M, Barešić A, Lenhard B. Genomic Interactions: an R/bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics*. (2015) 16:963. doi: 10.1186/s12864-015-2140-x
81. Zhang H, He L, Cai L. Transcriptome sequencing: RNA-seq. In: Huang T, editor. *Computational Systems Biology*. New York, NY: Humana Press (2018). p. 15–27.
82. Jeong E, Moon SU, Song M, Yoon S. Transcriptome modeling and phenotypic assays for cancer precision medicine. *Arch Pharm Res*. (2017) 40:906–14. doi: 10.1007/s12272-017-0940-z
83. Babu MM. Introduction to microarray data analysis. *Comput Genom Theory Appl*. (2004) 225:249. doi: 10.1007/0-306-47815-3_1
84. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. *Front Physiol.* (2015) 6:383. doi: 10.3389/fphys.2015.00383
85. Li JR, Sun CH, Li W, Chao RF, Huang CC, Zhou XJ, et al. Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res*. (2015) 44:D944–51. doi: 10.1093/nar/gkv1282
86. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. (2015) 19:A68. doi: 10.5114/wo.2014.47136
87. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. *Nat Genet*. (2013) 45:580. doi: 10.1038/ng.2653
88. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*. (2012) 13:227–32. doi: 10.1038/nrg3185
89. Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, et al. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell Mol Life Sci*. (2015) 72:3425–39. doi: 10.1007/s00018-015-1934-y
90. Kaikkonen MU, Adelman K. Emerging roles of non-coding RNA transcription. *Trends Biochem Sci*. (2018) 43:654–67. doi: 10.1016/j.tibs.2018.06.002
91. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. (2008) 321:956–60. doi: 10.1126/science.1160342
92. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. (2008) 5:621. doi: 10.1038/nmeth.1226
93. Slonim DK, Yanai I. Getting started in gene expression microarray analysis. *PLoS Comput Biol*. (2009) 5:e1000543. doi: 10.1371/journal.pcbi.1000543
94. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635
95. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. (2016) 34:525–7. doi: 10.1038/nbt.3519
96. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. (2017) 14:417–9. doi: 10.1038/nmeth.4197
97. Li Y, Heavican TB, Vellichirammal NN, Iqbal J, Guda C. ChimeRScope: a novel alignment-free algorithm for fusion transcript prediction using paired-end RNA-Seq data. *Nucleic Acids Res*. (2017) 45:e120. doi: 10.1093/nar/gkx315
98. Yang X, Saito Y, Rao A, Kim HJ, Singh P, Scott E, et al. Alignment-free filtering for cfNA fusion fragments. *Bioinformatics*. (2019) 35:i225–32. doi: 10.1093/bioinformatics/btz346
99. Raplee ID, Evsikov AV, Marín de Evsikova C. Aligning the aligners: comparison of RNA sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *J Personal Med*. (2019) 9:18. doi: 10.3390/jpm9020018
100. Babarinde IA, Li Y, Hutchins AP. Computational methods for mapping, assembly and quantification for coding and non-coding transcripts. *Comput Struct Biotechnol J*. (2019) 17:628–37. doi: 10.1016/j.csbj.2019.04.012
101. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. (2009) 25:1105–11. doi: 10.1093/bioinformatics/btp120
102. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. (2010) 28:511–5. doi: 10.1038/nbt.1621
103. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. (2012) 7:562–78. doi: 10.1038/nprot.2012.016
104. Baruzzo G, Hayer KE, Kim EJ, Camillo BD, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods*. (2017) 14:135–9. doi: 10.1038/nmeth.4106
105. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. (2019) 37:907–15. doi: 10.1038/s41587-019-0201-4
106. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. (2019) 20:278. doi: 10.1186/s13059-019-1910-1
107. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. (2016) 11:1650–67. doi: 10.1038/nprot.2016.095
108. Li Y, Chien J, Smith DI, Ma J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*. (2011) 27:1708–10. doi: 10.1093/bioinformatics/btr265
109. Ellrott K, Buchanan A, Creason A, Mason M, Schaffter T, Hoff B, et al. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol*. (2019) 20:1–9. doi: 10.1186/s13059-019-1794-0
110. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. (2014) 15:550. doi: 10.1186/s13059-014-0550-8

111. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. (2012) 40:4288–97. doi: 10.1093/nar/gks042
112. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. (2015) 43:e47. doi: 10.1093/nar/gkv007
113. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods*. (2017) 14:687–90. doi: 10.1038/nmeth.4324
114. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. (2016) 17:13. doi: 10.1186/s13059-016-0881-8
115. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: technologies and their applications. *J Chromatogr Sci*. (2017) 55:182–96. doi: 10.1093/chromsci/bmw167
116. Yakkioui Y, Temel Y, Chevet E, Negroni L. Integrated and quantitative proteomics of human tumors. *Methods Enzymol*. (2017) 586:229–46. doi: 10.1016/bs.mie.2016.09.034
117. Sutandy FXR, Qian J, Chen CS, Zhu H. Overview of protein microarrays. *Curr Protoc Protein Sci*. (2013) Chapter 27:Unit 27.1. doi: 10.1002/0471140864.ps2701s72
118. Atak A, Mukherjee S, Jain R, Gupta S, Singh VA, Gahoi N, et al. Protein microarray applications: autoantibody detection and posttranslational modification. *Proteomics*. (2016) 16:2557–69. doi: 10.1002/pmic.201600104
119. Cho WC. Mass spectrometry-based proteomics in cancer research. *Expert Rev Proteomics*. (2017) 14:725–7. doi: 10.1080/14789450.2017.1365604
120. Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. *BMC Syst Biol*. (2014) 8:S3. doi: 10.1186/1752-0509-8-S2-S3
121. Graumann J, Scheltema RA, Zhang Y, Cox J, Mann M. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. *Mol Cell Proteomics*. (2012) 11:M111.013185. doi: 10.1074/mcp.M111.013185
122. Hoopmann MR, Moritz RL. Current algorithmic solutions for peptide-based proteomics data generation and identification. *Curr Opin Biotechnol*. (2013) 24:31–8. doi: 10.1016/j.copbio.2012.10.013
123. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*. (2010) 73:2092–123. doi: 10.1016/j.jprot.2010.08.009
124. Kopczynski D, Sickmann A, Ahrends R. Computational proteomics tools for identification and quality control. *J Biotechnol*. (2017) 261:126–30. doi: 10.1016/j.jbiotec.2017.06.1199
125. Mihăşan M, Wormwood KL, Sokolowska I, Roy U, Woods AG, Darie CC. Mass spectrometry-and computational structural biology-based investigation of proteins and peptides. In: *Advancements of Mass Spectrometry in Biomedical Research*. Cham: Springer (2019). p. 265–287.
126. Gatto L, Christoforou A. Using R and bioconductor for proteomics data analysis. *Biochim Biophys Acta*. (2014) 1844:42–51. doi: 10.1016/j.bbapap.2013.04.032
127. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. (2008) 26:1367–72. doi: 10.1038/nbt.1511
128. Vazquez A, Kamphorst JJ, Markert EK, Schug ZT, Tardito S, Gottlieb E. Cancer metabolism at a glance. *J Cell Sci*. (2016) 129:3367–73. doi: 10.1242/jcs.181016
129. Armitage EG, Ciborowski M. Applications of metabolomics in cancer studies. *Adv Exp Med Biol*. (2017) 965:209–34. doi: 10.1007/978-3-319-47656-8_9
130. Yang K, Han X. Lipidomics: techniques, applications, and outcomes related to biomedical sciences. *Trends Biochem Sci*. (2016) 41:954–69. doi: 10.1016/j.tibs.2016.08.010
131. Perrotti F, Rosa C, Cicalini I, Sacchetta P, Del Boccio P, Genovesi D, et al. Advances in lipidomics for cancer biomarkers discovery. *Int J Mol Sci*. (2016) 17:1992. doi: 10.3390/ijms17121992
132. Zhang A, Sun H, Yan G, Wang P, Wang X. Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. *Biomed Chromatogr*. (2016) 30:7–12. doi: 10.1002/bmc.3453
133. Hu T, Zhang JL. Mass-spectrometry-based lipidomics. *J Sep Sci*. (2018) 41:351–72. doi: 10.1002/jssc.201700709
134. Meier R, Ruttkies C, Treutler H, Neumann S. Bioinformatics can boost metabolomics research. *J Biotechnol*. (2017) 261:137–41. doi: 10.1016/j.jbiotec.2017.05.018
135. Aggio R, Villas-Boas SG, Ruggiero K. Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics*. (2011) 27:2316–8. doi: 10.1093/bioinformatics/btr379
136. Stanstrup J, Broeckling CD, Helmus R, Hoffmann N, Mathé E, Naake T, et al. The metaRbolomics toolbox in bioconductor and beyond. *Metabolites*. (2019) 9:E200. doi: 10.3390/metabo9100200
137. Mohamed A, Molendijk J. *lipidr: Data Mining and Analysis of Lipidomics Datasets*. R package version 200. Washington, DC (2019).
138. Yuan Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harb Perspect Med*. (2016) 6:a026583. doi: 10.1101/cshperspect.a026583
139. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol Cancer*. (2017) 16:41. doi: 10.1186/s12943-017-0600-4
140. Sierant MC, Choi J. Single-cell sequencing in cancer: recent applications to immunogenomics and multi-omics tools. *Genomics Inform*. (2018) 16:e17. doi: 10.5808/GI.2018.16.4.e17
141. Ren X, Kang B, Zhang Z. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol*. (2018) 19:1–14. doi: 10.1186/s13059-018-1593-z
142. Navin NE. The first five years of single-cell cancer genomics and beyond. *Genome Res*. (2015) 25:1499–507. doi: 10.1101/gr.191098.115
143. Lo PK, Zhou Q. Emerging techniques in single-cell epigenomics and their applications to cancer research. *J Clin Genom*. (2018) 1:1–16. doi: 10.4172/JCG.1000103
144. Litzenburger UM, Buenrostro JD, Wu B, Shen Y, Sheffield NC, Kathiria A, et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol*. (2017) 18:15. doi: 10.1186/s13059-016-1133-7
145. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*. (2015) 16:133–45. doi: 10.1038/nrg3833
146. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. (2017) 9:75. doi: 10.1186/s13073-017-0467-4
147. Gao S. Data analysis in single-cell transcriptome sequencing. *Methods Mol Biol*. (2018) 1754:311–26. doi: 10.1007/978-1-4939-7717-8_18
148. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. (2018) 50:96. doi: 10.1038/s12276-018-0071-8
149. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*. (2016) 5:2122. doi: 10.12688/f1000research.9501.2
150. Amezquita RA, Carey VJ, Carpp LN, Geistlinger L, Lun ATL, Marini F, et al. *Orchestrating Single-Cell Analysis With Bioconductor*. Washington, DC (2019).
151. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. (2018) 19:15. doi: 10.1186/s13059-017-1382-0
152. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. (2018) 36:411–20. doi: 10.1038/nbt.4096
153. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. (2019) 20:257–72. doi: 10.1038/s41576-019-0093-7
154. Pegoraro G, Misteli T. High-throughput imaging for the discovery of cellular mechanisms of disease. *Trends Genet*. (2017) 33:604–15. doi: 10.1016/j.tig.2017.06.005
155. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. EBImage-an R package for image processing with applications to cellular phenotypes. *Bioinformatics*. (2010) 26:979–81. doi: 10.1093/bioinformatics/btq046
156. Pau G, Zhang X, Boutros M, Huber W. *imageHTS: Analysis of High-Throughput Microscopy-Based Screens*. Washington, DC (2019).
157. McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol*. (2018) 16:e2005970. doi: 10.1371/journal.pbio.2005970
158. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*. (2015) 348:aaa6090. doi: 10.1126/science.aaa6090

159. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. (2019) 363:1463–7. doi: 10.1126/science.aaw1219
160. Yoosuf N, Navarro JF, Salmén F, Ståhl PL, Daub CO. Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. *Breast Cancer Res*. (2020) 22:1–10. doi: 10.1186/s13058-019-1242-9
161. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun*. (2018) 9:1–13. doi: 10.1038/s41467-018-04724-5
162. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res*. (2018) 78:5970–9. doi: 10.1158/0008-5472.CAN-18-0747
163. Moncada R, Wagner F, Chiodin M, Devlin JC, Baron M, Hajdu CH, et al. Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. *bioRxiv*. (2018) 254375. doi: 10.1101/254375
164. Xu H, Lyu X, Yi M, Zhao W, Song Y, Wu K. Organoid technology and applications in cancer research. *J Hematol Oncol*. (2018) 11:116. doi: 10.1186/s13045-018-0662-9
165. Lindeboom RG, van Voorthuijsen L, Oost KC, Rodríguez-Colman MJ, Luna-Velez MV, Furlan C, et al. Integrative multi-omics analysis of intestinal organoid differentiation. *Mol Syst Biol*. (2018) 14:e8227. doi: 10.15252/msb.20188227
166. Finotello F, Eduati F. Multi-omics profiling of the tumor microenvironment: paving the way to precision immuno-oncology. *Front Oncol*. (2018) 8:430. doi: 10.3389/fonc.2018.00430
167. Finotello F, Rieder D, Hackl H, Trajanoski Z. Next-generation computational tools for interrogating cancer immunity. *Nat Rev Genet*. (2019) 20:724–46. doi: 10.1038/s41576-019-0166-7
168. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods*. (2019) 16:409–12. doi: 10.1038/s41592-019-0392-0
169. Praktiknjo SD, Obermayer B, Zhu Q, Fang L, Liu H, Quinn H, et al. Tracing tumorigenesis in a solid tumor model at single-cell resolution. *Nat Commun*. (2020) 11:1–12. doi: 10.1038/s41467-020-14777-0
170. Rajagopala SV, Vashee S, Oldfield LM, Suzuki Y, Venter JC, Telenti A, et al. The human microbiome and cancer. *Cancer Prev Res (Phila)*. (2017) 10:226–34. doi: 10.1158/1940-6207.CAPR-16-0249
171. Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol*. (2017) 18:228. doi: 10.1186/s13059-017-1359-z
172. Contreras JL, Knoppers BM. The genomic commons. *Annu Rev Genomics Hum Genet*. (2018) 19:429–53. doi: 10.1146/annurev-genom-083117-021552
173. Cook-Deegan R, McGuire AL. Moving beyond Bermuda: sharing data to build a medical information commons. *Genome Res.* (2017) 27:897–901. doi: 10.1101/gr.216911.116
174. Jansen P, van den Berg L, van Overveld P, Boiten JW. Research data stewardship for healthcare professionals. In: Kubben P, Dumontier M, Dekker A, editors. *Fundamentals of Clinical Data Science*. Cham: Springer (2018) p. 37–53.
175. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. (2016) 375:1109–12. doi: 10.1056/NEJMp1607591
176. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. (2013) 45:1113–20. doi: 10.1038/ng.2764
177. Kosinski M, Biecek P. *RTCGA: The Cancer Genome Atlas Data Integration*. R package version 1.14.0 (2019). Available online at: https://rtcga.github.io/RTCGA
178. Campbell PJ, Rätsch G, Kahles A, Lehmann KV, Davidson NR, Stark SG, et al. Pan-cancer analysis of whole genomes. *Nature*. (2020) 578:82–93. doi: 10.1038/s41586-020-1969-6
179. Rendleman MC, Buatti JM, Braun TA, Smith BJ, Nwakama C, Beichel RR, et al. Machine learning with the TCGA-HNSC dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality. *BMC Bioinformatics*. (2019) 20:339. doi: 10.1186/s12859-019-2929-8
180. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, et al. TCGA workflow: analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Res*. (2016) 5:1542. doi: 10.12688/f1000research.8923.1
181. Parkinson DR, Johnson BE, Sledge GW. Making personalized cancer medicine a reality: challenges and opportunities in the development of biomarkers and companion diagnostics. *Clin Cancer Res*. (2012) 18:619–24. doi: 10.1158/1078-0432.CCR-11-2017
182. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet*. (2017) 8:84. doi: 10.3389/fgene.2017.00084
183. Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-multi-OMICS approach: a new frontier in cancer research. *Biomed Res Int*. (2018) 2018:9836256. doi: 10.1155/2018/9836256
184. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*. (2016) 17:15. doi: 10.1186/s12859-015-0857-9
185. Yoo BC, Kim KH, Woo SM, Myung JK. Clinical multi-omics strategies for the effective cancer management. *J Proteomics*. (2018) 188:97–106. doi: 10.1016/j.jprot.2017.08.010
186. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, et al. Software for the integration of multiomics experiments in bioconductor. *Cancer Res*. (2017) 77:e39–42. doi: 10.1158/0008-5472.CAN-17-0344
187. Voillet V, Besse P, Liaubet L, Cristobal MS, González I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*. (2016) 17:402. doi: 10.1186/s12859-016-1273-5
188. van Iterson M, Cats D, Hop P, Heijmans BT. omicsPrint: detection of data linkage errors in multiple omics studies. *Bioinformatics*. (2018) 34:2142–3. doi: 10.1093/bioinformatics/bty062
189. Consortia S. *STATegRa: Classes and Methods for Multi-Omics Data Integration*. R package version 1.20.0 (2019).
190. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752
191. Cao KAL, González I, Déjean S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*. (2009) 25:2855–6. doi: 10.1093/bioinformatics/btp515
192. Hernández-de Diego R, Boix-Chova N, Gómez-Cabrero D, Tegner J, Abugessaisa I, Conesa A. STATegra EMS: an experiment management system for complex next-generation omics experiments. *BMC Syst Biol*. (2014) 8:S9. doi: 10.1186/1752-0509-8-S2-S9
193. Martínez-Mira C, Conesa A, Tarazona S. *MOSim: Multi-Omics Simulation in R*. Washington, DC (2018).
194. Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I, Ramirez RN, Company C, Schmidt A, et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data*. (2019) 6:256. doi: 10.1038/s41597-019-0202-7
195. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. (2009) 27:1160–7. doi: 10.1200/JCO.2008.18.1370
196. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform*. (2016) 17:628–41. doi: 10.1093/bib/bbv108
197. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration–a comparison of unsupervised clustering methodologies. *Brief Bioinform*. (2017) 20:1269–79. doi: 10.1093/bib/bbx167
198. Yugi K, Kubota H, Hatano A, Kuroda S. Trans-omics: how to reconstruct biochemical networks across multiple 'omic' layers. *Trends Biotechnol*. (2016) 34:276–90. doi: 10.1016/j.tibtech.2015.12.013
199. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*. (2019) 35:3055–62. doi: 10.1093/bioinformatics/bty1054

200. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol*. (2016) 12:878. doi: 10.15252/msb.20156651
201. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *Lond Edinb Dublin Philos Mag J Sci*. (1901) 2:559–72. doi: 10.1080/14786440109462720
202. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–605.
203. McInnes L, Healy J, Melville J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. (2018). Available online at: http://arxiv.org/abs/1802.03426v2; http://arxiv.org/pdf/1802.03426v2
204. Tufte E. *The Visual Display of Quantitative Information*. Graphics Press (1983).
205. Freytag S. *schex: Hexbin Plots for Single Cell Omics Data*. Washington, DC: R package version 1.0.0 (2019).
206. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. (2009) 19:1639–45. doi: 10.1101/gr.092759.109
207. Hu Y, Yan C, Hsu CH, Chen QR, Niu K, Komatsoulis GA, et al. OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data. *Cancer Inform*. (2014) 13:13–20. doi: 10.4137/CIN.S13495
208. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. *Stat Genom*. (2016) 1418:335–51. doi: 10.1007/978-1-4939-3578-9_16
209. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for peak annotation and visualization. *Bioinformatics*. (2013) 29:3097–9. doi: 10.1093/bioinformatics/btt520
210. Pont F, Tosolini M, Fournié JJ. Single-cell signature explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic Acids Res*. (2019) 47:e133. doi: 10.1093/nar/gkz601
211. Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, Hovig E. HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics*. (2014) 30:1620–2. doi: 10.1093/bioinformatics/btu082
212. Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU epigenome browser update 2019. *Nucleic Acids Res*. (2019) 47:W158–65. doi: 10.1093/nar/gkz348
213. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*. (2016) 3:99–101. doi: 10.1016/j.cels.2015.07.012
214. Yardımcı GG, Noble WS. Software tools for visualizing Hi-C data. *Genome Biol*. (2017) 18:26. doi: 10.1186/s13059-017-1161-y
215. Das S, Tripathy DS. *OMICsPCA: An R Package for Quantitative Integration and Analysis of Multiple Omics Assays From Heterogeneous Samples*. R package version 1.2.0 (2019).
216. Xu T, Le TD, Liu L, Su N, Wang R, Sun B, et al. CancerSubtypes: an R/bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics*. (2017) 33:3131–3. doi: 10.1093/bioinformatics/btx378
217. Mezhoud K. *bioCancer: Interactive Multi-Omics Cancers Data Visualization and Analysis*. R package version 1.12.0 (2019). Available online at: http://kmezhoud.github.io/bioCancer
218. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*. (2014) 15:162. doi: 10.1186/1471-2105-15-162
219. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. (2009) 25:2906–12. doi: 10.1093/bioinformatics/btp543
220. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Springer Series in Statistics*. New York, NY: Springer (2009).
221. Provenzano E, Ulaner GA, Chin SF. Molecular classification of breast cancer. *PET Clinics*. (2018) 13:325–38. doi: 10.1016/j.cpet.2018.02.004
222. Syed-Abdul S, Iqbal U, Li YC. Predictive analytics through machine learning in the clinical settings. *Comput Methods Prog Biomed*. (2017) 144:A1–2. doi: 10.1016/S0169-2607(17)30552-7
223. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. (2015) 13:8–17. doi: 10.1016/j.csbj.2014.11.005
224. Streeter OE, Beron PJ, Iyer PN. Precision medicine: genomic profiles to individualize therapy. *Otolaryngol Clin North Am*. (2017) 50:765–73. doi: 10.1016/j.otc.2017.03.012
225. Schwaederle M, Daniels GA, Piccioni DE, Fanta PT, Schwab RB, Shimabukuro KA, et al. On the road to precision cancer medicine: analysis of genomic biomarker actionability in 439 patients. *Mol Cancer Ther*. (2015) 14:1488–94. doi: 10.1158/1535-7163.MCT-14-1061
226. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv*. (2016). doi: 10.1101/067611
227. Thévenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J Proteome Res*. (2015) 14:3322–35. doi: 10.1021/acs.jproteome.5b00354
228. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. MultiOmics Factor Analysis–a framework for unsupervised integration of multi?omics data sets. Molecular Systems Biology. (2018) 14. doi: 10.15252/msb.20178124
229. Determan C. Optimal algorithm for metabolomics classification and feature selection varies by dataset. *Int J Biol*. (2015) 7. doi: 10.5539/ijb.v7n1p100
230. Bhalla S, Kaur H, Dhall A, Raghava GPS. Prediction and analysis of skin cancer progression using genomics profiles of patients. *Sci Rep*. (2019) 9:15790. doi: 10.1038/s41598-019-52134-4
231. Rinaudo P, Boudah S, Junot C, Thévenot EA. Biosigner: a new method for the discovery of significant molecular signatures from omics data. Front Mol Biosci. (2016) 3:26. doi: 10.3389/fmolb.2016.00026
232. Long NP, Jung KH, Anh NH, Yan HH, Nghi TD, Park S, et al. An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. *Cancers*. (2019) 11:155. doi: 10.3390/cancers11020155
233. Kwon MS, Kim Y, Lee S, Namkung J, Yun T, Yi SG, et al. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics*. (2015) 16:S4. doi: 10.1186/1471-2164-16-S9-S4
234. Klein HU, Schäfer M, Porse BT, Hasemann MS, Ickstadt K, Dugas M. Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. *Bioinformatics*. (2014) 30:1154–62. doi: 10.1093/bioinformatics/btu003
235. Han S, Lee Y. *IMAS: Integrative Analysis of Multi-Omics Data for Alternative Splicing*. R package version 1.8.0 (2019).
236. Hernandez-Ferrer C, Wellenius GA, Tamayo I, Basagaña X, Sunyer J, Vrijheid M, et al. Comprehensive study of the exposome and omic data using rexposome bioconductor packages. *Bioinformatics*. (2019) 35:5344–5. doi: 10.1093/bioinformatics/btz526
237. Metwally AA, Zhang T, Snyder M. *OmicsLonDA: Omics Longitudinal Differential Analysis*. R package version 1.0.0 (2019). Available online at: https://github.com/aametwally/OmicsLonDA
238. de Anda-Jáuregui G, Guo K, McGregor BA, Feldman EL, Hur J. Pathway crosstalk perturbation network modeling for identification of connectivity changes induced by diabetic neuropathy and pioglitazone. *BMC Syst Biol*. (2019) 13:1. doi: 10.1186/s12918-018-0674-7
239. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. *Mol Cell Proteomics*. (2019) 18:S153–68. doi: 10.1074/mcp.TIR118.001251
240. Rodriguez JC, Merino GA, Llera AS, Fernández EA. Massive integrative gene set analysis enables functional characterization of breast cancer subtypes. *J Biomed Inform*. (2019) 93:103157. doi: 10.1016/j.jbi.2019.103157
241. Odom GJ, Ban Y, Liu L, Sun X, Pico AR, Zhang B, et al. pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection. *bioRxiv*. (2019). doi: 10.1101/615435
242. Dinalankara W, Ke Q, Xu Y, Ji L, Pagane N, Lien A, et al. Digitizing omics profiles by divergence from a baseline. *Proc Natl Acad Sci USA*. (2018) 115:4545–52. doi: 10.1073/pnas.1721628115
243. Pilarczyk M, Najafabadi MF, Kouril M, Vasiliauskas J, Niu W, Shamsaei B, et al. Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINCS. *bioRxiv*. (2019). doi: 10.1101/826271

244. Albert R, Barabási AL. Statistical mechanics of complex networks. *Rev Mod Phys*. (2002) 74:47–97. doi: 10.1103/RevModPhys.74.47
245. Quesada D, Cruz-Monteagudo M, Fletcher T, Duardo-Sanchez A, González-Díaz H. Complex networks and machine learning: from molecular to social sciences. *Appl Sci*. (2019) 9:4493. doi: 10.3390/app9214493
246. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*. (2013) 29:1830–1. doi: 10.1093/bioinformatics/btt285
247. Sales G, Calura E, Cavalieri D, Romualdi C. Graphite–a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*. (2012) 13:20. doi: 10.1186/1471-2105-13-20
248. Calura E, Martini P, Sales G, Beltrame L, Chiorino G, D'Incalci M, et al. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Res*. (2014) 42:e96. doi: 10.1093/nar/gku354
249. Calura E, Bisognin A, Manzoni M, Todoerti K, Taiana E, Sales G, et al. Disentangling the microRNA regulatory milieu in multiple myeloma: integrative genomics analysis outlines mixed miRNA-TF circuits and pathway-derived networks modulated in t(4;14) patients. *Oncotarget*. (2015) 7:2367–78. doi: 10.18632/oncotarget.6151
250. Calura E, Pizzini S, Bisognin A, Coppe A, Sales G, Gaffo E, et al. A data-driven network model of primary myelofibrosis: transcriptional and post-transcriptional alterations in $CD34^{+}$ cells. Blood Cancer J. (2016) 6:e439. doi: 10.1038/bcj.2016.47
251. Calura E, Paracchini L, Fruscio R, DiFeo A, Ravaggi A, Peronne J, et al. A prognostic regulatory pathway in stage I epithelial ovarian cancer: new hints for the poor prognosis assessment. *Ann Oncol*. (2016) 27:1511–9. doi: 10.1093/annonc/mdw210
252. Shu L, Zhao Y, Kurt Z, Byars SG, Tukiainen T, Kettunen J, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics*. (2016) 17:874. doi: 10.1186/s12864-016-3198-9
253. Wachter A, Beißbarth T. pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics*. (2015) 31:3072–4. doi: 10.1093/bioinformatics/btv323
254. Alcalá-Corona SA, de Anda-Jáuregui G, Espinal-Enriquez J, Tovar H, Hernández-Lemus E. Network modularity and hierarchical structure in breast cancer molecular subtypes. In: *Springer Proceedings in Complexity*. (2018) p. 352–8.
255. de Anda-Jáuregui G. Guideline for comparing functional enrichment of biological network modular structures. *Appl Netw Sci*. (2019) 4:13. doi: 10.1007/s41109-019-0128-1
256. Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res*. (2012) 40:W13–21. doi: 10.1093/nar/gks460
257. Siska C, Bowler R, Kechris K. The discordant method: a novel approach for differential correlation. *Bioinformatics*. (2015) 32:690–6. doi: 10.1093/bioinformatics/btv633
258. Schlosser P, Knaus J, Schmutz M, Döhner K, Plass C, Bullinger L, et al. *Netboost: Boosting-Supported Network Analysis Improves High-Dimensional Omics Prediction in Acute Myeloid Leukemia and Huntington's Disease*. (2019). Available from: http://arxiv.org/abs/1909.12551v1; http://arxiv.org/pdf/1909.12551v1
259. Champion M, Brennan K, Croonenborghs T, Gentles AJ, Pochet N, Gevaert O. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine*. (2018) 27:156–66. doi: 10.1016/j.ebiom.2017.11.028
260. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. (2008) 9:559. doi: 10.1186/1471-2105-9-559
261. Khatamian A, Paull EO, Califano A, Yu J. SJARACNe: a scalable software tool for gene network reverse engineering from big data. *Bioinformatics*. (2018) 35:2165–6. doi: 10.1093/bioinformatics/bty907
262. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. Nonredundant, highly connected micrornas control functionality in breast cancer networks. *Int J Genom*. (2018) 2018:1–10. doi: 10.1155/2018/9585383
263. Hernández-Lemus E, Espinal-Enríquez J, de Anda-Jáuregui G. *Probabilistic Multilayer Networks*. (2018). Available online at: http://arxiv.org/abs/1808.07857v1; http://arxiv.org/pdf/1808.07857v1
264. De Domenico M, Solé-Ribalta A, Cozzo E, Kivelä M, Moreno Y, Porter MA, et al. Mathematical formulation of multilayer networks. *Phys Rev X*. (2013) 3:041022. doi: 10.1103/PhysRevX.3.041022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.