Prospects and challenges of cancer systems medicine - from genes to disease networks

Mohammad Reza Karimi[1†], Amir Hossein Karimi[1†], Shamsozoha Abolmaali[1], Mehdi Sadeghi[1*], Ulf Schmitz[2,3,4,5]

[1] Department of Cell and Molecular Biology, Faculty of Science, Semnan University, Semnan, Iran

[2] Computational BioMedicine Laboratory Centenary Institute, The University of Sydney, Camperdown 2050, Australia

[3] Gene and Stem Cell Therapy Program Centenary Institute, The University of Sydney, Camperdown 2050, Australia

[4] Faculty of Medicine and Health, The University of Sydney, Camperdown 2050, Australia

[5] College of Public Health, Medical and Veterinary Sciences, Department of Molecular and Cell Biology, James Cook University, Townsville, QLD 4811, Australia

[†] These authors have contributed equally to this work

[*] Correspondence: Mehdi Sadeghi; Email mehdisadeghi@semnan.ac.ir Telephone no +98-2331532223

Abstract

It is becoming evident that holistic perspectives towards cancer are crucial in deciphering the overwhelming complexity of tumors. Single-layer analysis of genome-wide data has greatly contributed to our understanding of cellular systems and their perturbations. However, fundamental gaps in our knowledge persist and hamper the design of effective interventions. It is becoming more apparent than ever, that cancer should not only be viewed as a disease of the genome, but as a disease of the cellular system. Integrative multi-layer approaches are emerging as vigorous assets in our endeavors to achieve systemic views on cancer biology. Herein, we provide a comprehensive review of the approaches, methods, and technologies that can serve to achieve systemic perspectives of cancer. We start with genome-wide single-layer approaches of omics analyses of cellular systems and move on to multi-layer integrative approaches in which in-depth descriptions of proteogenomics and network-based data analysis are provided. Proteogenomics is a remarkable example of how the integration of multiple levels of information can reduce our blind spots and increase the accuracy and reliability of our interpretations and network-based data analysis is a major approach for data interpretation and a robust scaffold for data integration and modeling. Overall, this review aims to increase cross-field awareness of the approaches and challenges regarding the omics-based study of cancer and to facilitate the necessary shift towards holistic approaches.

Keywords: Systems biology, Transcriptomics, Proteomics, Metabolomics, Proteogenomics, Biological networks

Key points:

- Systemic perception of cancer is essential for the design of effective interventions
- High-throughput technologies are the main arteries of systemic studies of cancer
- Emerging data integration approaches are rapidly altering current paradigms of oncology
- Vertical integration of omics data is capable of addressing multifaceted challenges
- Network-based data analysis is a major asset in data integration and interpretation

1 Introduction

According to the world health organization, an estimated number of 10 million patients worldwide succumbed to different types of cancer in 2020 alone. Despite considerable advancements in diagnostics and novel therapeutic approaches following the distilled outcomes of millions of cancer-related studies, many clinical trials do not result in major success [1–3]. This, among other reasons (e.g. implementation issues and technical limitations), can be attributed to the lack of a systemic view towards cancer and its underlying mechanisms. Indeed, the results of the recent WINTHER trial demonstrate the utility of multi-omics approaches for the improvement of cancer therapy recommendations [4]. A deeper and holistic perspective of the underlying systemic perturbations during tumor initiation and progression is a prerequisite for designing more targeted a.k.a. personalized interventions.

In cancer investigations, we are facing aberrations in extremely complex systems with enigmatic interplays between altered pathways and extensive multilevel cross-talk. The heterogeneity of subpopulations of malignant cells further contributes to the obscurity of this picture. Contrasting with conventional reductionist approaches, the field of systems biology has emerged and laid foundations for holistic investigation of biological units and mathematical modeling of molecular and cellular interplays for comprehensible exploration of biological systems [5] (refer to Figure 1 for a timeline of some of the major contributions to the field of systems biology). Fueled by genome-wide technologies and bioinformatics advancements, systems biology is establishing itself as the only reasonable approach for dissecting the complexity of tumors, identifying core components of these perturbed systems, and recognizing the vulnerabilities of specific tumors for effective patient stratification and precise interventions.

Achieving a holistic picture of cancer demands cooperation between multiple areas of research, magnification of the links between layers of information, and robust approaches for effective integration of the heterogeneous data. Hence, there is an increasing need for the research

community to move beyond single-layer omics analysis of cancer and take advantage of the value added by integrating multiple omics layers. Here, we review current approaches, methods, and technologies that can serve to achieve a systemic perspective of cancer. We start with genome-wide single-layer approaches and move on to multi-layer integrative approaches with a focus on a systems biology perspective throughout the work. In each section, an overview of the importance of each respective approach in cancer research is presented. Then, a general framework, based on the current best practices of the field or novel and promising methods, is provided. In that context, we highlight methods that require minimal computational skill and discuss outstanding challenges and future perspectives. It should be noted that while the approaches and technologies discussed in this review are presented in the context of cancer research, many of them are also applicable to fields other than oncology. The review is concluded with multiple representative examples of what these approaches have already contributed to the field of oncology. Overall, we aim to increase cross-field awareness of the approaches and challenges regarding the omics-based study of cancer for both research and medical communities in order to facilitate the necessary shift towards more holistic approaches.

2 Single-layer approaches

High-throughput technologies capable of generating comprehensive data that encompass all the molecular components at a particular level are the main arteries of systems-level studies in cancer. Genomics, transcriptomics, proteomics, and metabolomics are the four major approaches currently implemented using various technologies and comprehensive data analysis methods (Figure 2). These approaches and related technologies as well as analysis pipelines are discussed in further sections. Importantly, single-layer data analysis has greatly enhanced our understanding of cellular mechanisms and their perturbations and fundamentally contributed to our knowledge of biological systems. However, the purposive study of biological systems requires multi-level approaches that integrate the generated data from different single-layer approaches to achieve a holistic view of

cells under normal and disturbed conditions [6] (for a list of relevant researches and their contributions to the field of systems oncology refer to Supplementary Table S1).

2.1 Genomics: Elucidating the genomic landscape of tumors

The process of tumorigenesis begins (and usually progresses) with the occurrence of specific somatic driver mutations i.e. mutations that confer survival and proliferative advantages to a specific cell lineage [7]. These mutations are accompanied by a higher number of passenger mutations that do not directly contribute to tumorigenesis and cancer progression. Moreover, germline mutations can contribute to cancer predisposition [8]. The main complexity of cancer, however, arises from the lack of a consensus genomic landscape across different cancer types and even among patients stratified under certain criteria. Case-specific combinations of genomic alterations result in a wide variety of perturbations to the cellular system with the overall similar result of tumorigenesis and cancer progression. Indeed, attempts to discover mutational patterns also known as "mutational signatures" across and within tumor types have significantly contributed to our understanding of the etiology of cancer and led to the identification of cellular processes causative for specific cancer types that can serve as targets for therapeutic interventions [9–11]. Hence, it is evident that achieving an appropriate and encompassing perspective towards this complex disorder necessitates the implementation of genomics technologies.

Whole-exome sequencing (WES) is currently the most widely applied technology both in research projects [12,13] and in second-tier clinical diagnosis (implemented when gene panels are unable to pinpoint the cause of the defect) [14]. WES was developed to specifically capture and sequence all exonic regions of the genome. However, in the last decade we have learned that large parts of the human genome that were previously referred to as "junk DNA" are biologically active, i.e. translated into functional non-coding RNA [15]. Point mutations and structural variations in noncoding regions can also be cancer drivers, although less frequently compared to coding regions [16]. These findings, and the downwards trend in costs for sequencing, have already ignited the

transition from using WES to whole-genome sequencing (WGS) technologies. WGS has the advantage that it can also identify mutations in intergenic regulatory regions and mitochondrial DNA, mutations in promoters, structural variations, and viral infections, all of which are associated with different types of cancer. Moreover, the detection of copy number alterations is more effective with WGS [17]. Interestingly, WGS has been shown to be more effective than WES even when targeting coding regions [14].

Overall, current genomic technologies provide a potent vantage point for studying cancer etiology [10], biomarker discovery [18], the prediction of patients' drug response [19], and more. Recent years have witnessed the emergence of multiple international efforts such as the Pan-Cancer Analysis of Whole-Genomes (PCAWG) [16] where a considerable number of samples across different tumor types have been sequenced and analyzed. Such efforts provide unprecedented opportunities for the identification of mutational patterns across tumor types and the development of diagnostic and therapeutic approaches that are applicable to a wide range of patients.

2.1.1 Experimental workflow and data analysis pipeline

The genomics workflow generally starts with random fragmentation of the purified DNA by sonication or enzymatic digestion. Next, these fragments are enriched for target regions (genes of interest for gene panels or exonic regions when performing WES) [20]. The WGS workflow does not include this step. The acquired fragments are then ligated by oligonucleotide adapters that are complementary to the anchors on the flow cell [21]. This is commonly followed by a size selection step where ligated fragments with suitable sizes are purified [22]. Size selection can increase the sensitivity of circulating tumor DNA detection [23]. Nevertheless, selecting for specific size ranges might result in information loss and therefore, may be skipped depending on the goal of the study. Depending on the utilized method, a PCR amplification step might be required. However, considering that this step is prone to produce biased results, the utilization of a PCR-free method as a cost-efficient and more effective approach is highly recommended [24,25]. The next step is the

sequencing of the prepared library. Illumina short-read technologies are currently the dominant sequencing platforms (for a comprehensive review of different sequencing technologies refer to [26]). The NovaSeq 6000 sequencing platform is the most recent Illumina whole-genome sequencing technology. With overall results of similar quality for NovaSeq 6000 in comparison to the older Illumina whole-genome sequencing platform (HiSeq X Ten) and considering the substantial reduction in experiment costs [27], NovaSeq 6000 can be considered as the current state-of-the-art technology for whole-genome sequencing.

WGS data pre-processing begins with demultiplexing the sequencing reads using Illumina's Consensus Assessment of Sequence And Variation (CASAVA) software. Then, the raw reads are aligned against the human reference genome using an aligner tool, some of the most popular of which are BWAmem [28], Bowtie2 [29], and Novoalign (www.novocraft.com/products/novoalign/). Since duplicate reads can occur during sequence amplification and sequencing procedure, a duplicate marking step using tools such as Picard (broadinstitute.github.io/picard), Sambamba [30], or *SAMBLASTER* [31] is required.

In the next step, variant calling is performed. The most popular variant callers for somatic variant identification that have been specifically developed for the analysis of tumor samples include Mutect2 [32], VarScan [33], Strelka2 [34], and SomaticSniper [35]. A comparative study evaluating the somatic single nucleotide variant calling performance of these tools [36] reported a poor consensus among the results of variant callers. Mutect2 was identified as the best performing tool, followed closely by Strelka. Combining the high-confidence results of these methods is also a recommended approach. The study by Cai et al. [36] reported that while this approach increases the specificity of the variant calling, it results in a massive reduction of sensitivity. Thus, a combinatory approach should be opted for if higher reliability is desired while if achieving encompassing results is the goal of the study, utilizing Mutect2 or Strelka is a reasonable approach. In addition, the results of a study comparing the somatic variant calling performance of Mutect2 and Strelka2 [37] suggest that while these tools have similar overall performance, Mutect2 performs

better when dealing with lower mutation frequencies while Strelka2 is the better choice in the opposite scenario. Germline variant calling requires a different type of algorithm because the study is confined to the sequencing of normal genome [17]. This is most commonly performed using the Genome Analysis Toolkit (GATK) HaplotypeCaller (software.broadinstitute.org/gatk/). Studies indicate inconsistency among the results of different combinations of aligners and variant callers and hence, considering the intersection of the results of different pipelines is recommended to reduce false-positives [24,38]. However, a recent study suggests that some popular pipelines can produce results comparable to that of a combination of pipelines [39].

The detected variants are next subjected to annotation procedures. Annotations of previously reported alterations can be obtained from data repositories such as COSMIC [40], ClinVar [41], and OMIM [42]. The impact of novel variants with unknown significance can be predicted *in silico* using bioinformatics tools such as MutationTaster [43], SIFT [44], Polyphen [45], and VEP [46]. This is common practice in clinical diagnosis to predict the impact of novel variants before co-segregation and functional confirmation [47]. Moreover, there are algorithms such as CHASM [48] and PrimateAI [49] that are specifically developed to predict functional effects of mutations in the cancer context and distinguish driver mutations from passengers. The results of a recent comprehensive comparative study [50] that assessed 33 algorithms for their performance in predicting functional effects of mutations in cancer reported that cancer-specific algorithms significantly outperformed algorithms developed for general purposes. Furthermore, this study identified CHASM [48], CTAT-cancer [51], DEOGEN2 [52], and PrimateAI [49] as consistently well-performing algorithms. Notably, it was also proposed that incorporation of pathway and network information of the mutated genes in the prediction algorithm contributed to the outstanding performance of DEOGEN2 and thus, this should be considered in future algorithm developments. Anyhow, insignificant variants are filtered out in this step while significant variants are reported for downstream analysis and interpretation [53].

It is important to mention that there are numerous pipelines using different combinations of tools and computational approaches that attempt to address different challenges encountered in the various steps of this generalized workflow [27,54]. There are also convenient and comprehensive tools that facilitate the entire computational procedure, requiring minimal computational expertise. An example is the recently developed portable workflow named Sarek [55].

2.1.2 Challenges and perspectives

The variant allele frequency (VAF) is used to determine whether variants are heterozygous (variants with ~50% frequency) or homozygous (variants with ~100% frequency). In the cancer context, however, VAF analysis is not as precise because intratumoral heterogeneity and impurity of tumor DNA cause confusing deviations from expected VAFs [21,27,56]. The result of these ambiguities is the inability to acquire a picture of intratumoral heterogeneity that is representative of the actual biological phenomenon. Increasing the sequencing depth towards 100x coverage can ameliorate this inconvenience [24]. However, in some cases, achieving a fully representative picture of intratumoral heterogeneity requires impractical coverages of at least one order of magnitude higher than this [57]. A promising approach to tackle this problem, among others, is single-cell sequencing. Single-cell technologies provide researchers with a more accurate and less complex picture of the perturbed system both in the genomics and transcriptomics context [58]. However, single-cell technologies are still under development and a number of critical challenges both in wet lab [59] and dry lab [60] processes remain to be addressed.

The potential of tumor-specific somatic mutation profiling in guiding the administration of therapeutic interventions with precision is enormous [61]. This attracted a lot of attention towards the assessment of mutational landscapes of individuals through minimally invasive approaches such as cell-free DNA (cfDNA) sequencing. Circulating tumor DNA (ctDNA), presumably derived from necrotic and apoptotic tumor cells, comprise a portion of cfDNA in cancer patients, distinguishing them from healthy individuals [62]. Although the clinical efficacy of cfDNA monitoring in the

cancer context is yet to be validated through large-scale clinical trials, potential applications of cfDNA screening make it an attractive subject for researchers. These potential applications include postsurgical monitoring for stratification of patients for adjuvant therapy, systemic monitoring of the heterogeneity of the subclones in a metastatic tumor (as opposed to a single-site needle biopsy) for early detection of resistance to therapeutic agents, and early detection of neoplasms in asymptomatic individuals which can result in more effective interventions [63]. A major challenge for ctDNA analysis is that ctDNA VAFs are usually significantly below the detectable threshold of conventional high-throughput technologies. Ultrasensitive high-throughput technologies dedicated to ctDNA analysis such as iDES-enhanced CAPP-Seq have been introduced to ameliorate this shortcoming [64]. However, various challenges persist. These include increased risk of false-positives due to clonal hematopoiesis of indeterminate potential (CHIP) or other diseases and introduction of errors during library preparation (e.g. cfDNA degradation, contamination with normal cell lysates, etc.) and sequencing. Therefore, accurate identification of somatic mutations from cfDNA samples remains a daunting task [65]. Digital PCR approaches for ctDNA monitoring with higher sensitivities and lower costs address some of the challenges associated with high-throughput methods but require *a priori* knowledge of the targets and are particularly low in throughput [65]. Altogether, despite the remaining challenges, the analysis of ctDNA as a complement or surrogate to solid tissue specimens remains a valuable option, especially in cases where solid tumor samples are not accessible or sampling is associated with high risks.

Despite the tremendous progress made in recent years, there are still many unresolved questions in cancer genomics. The fact that no driver mutation could be identified for 5% of tumor samples [16] underscores that despite the extensive study of tumor driver genes and mutations, there are still shortcomings in our knowledge bases and/or models of cancer-initiating perturbations. Indeed, after decades of intensive research in cancer biology, the fundamentals of this complex dysfunction are still ambiguous in some areas. For example, the extent to which additional genomic/epigenomic alterations fuel the transition of a benign tumor to a malignant state is still a matter of debate [66].

Furthermore, the study of the genetic risk modifiers despite their potential to enhance our understanding of cancer is limited due to their small effect size [20]. Another important challenge is pinpointing the genomics alterations in high-complexity regions such as centromeres. Long-read sequencing technologies hold the promise of adequately addressing this problem [67]. However, certain drawbacks such as the high rate of errors in sequencing need to be tackled before these technologies would be able to effectively benefit the field.

2.2 Transcriptomics: Approaches to decipher the post-transcriptional complexity of tumors

The central dogma of biology describes the transition of information to function [68], from a semi-static genome to the highly dynamic cell. Going from genome to proteome the complexity increases as additional regulatory layers are introduced, from epigenetic [69] to post-transcriptional [70] and epi-transcriptomic regulations [71], to post-translational modifications [72]. Hence, efforts to understand the complex mechanisms of the cellular system and its perturbations exclusively from a genomic viewpoint would be futile. A widely appreciated approach to enhance our understanding of this complexity is studying the transcriptome [73].

The qualitative and quantitative analysis of transcriptomic information can yield insights into the post-transcriptional dynamics resulting from genetic events, epigenetic regulation, as well as regulation within the transcriptome, and provide means to predict the proteomics landscape. In cancer, deviations from normal transcriptomes undergo clonal evolution which in turn results in converged gene expression patterns referred to as the tumor gene signatures [74] that can be utilized in cancer subtyping [75], biomarker discovery [76], etc. The most broadly utilized functional study of the transcriptome is the comparison of expression profiles under different conditions (e.g. normal vs cancer) known as differential gene expression (DGE) analysis [77] e.g. by means of RNA-sequencing. Differential analysis of mRNA profiles can provide valuable information about perturbed signaling cascades and malfunctioning members of the cell system that gave rise to the

phenotype under investigation [78]. The study of alternative splicing and novel splicing events [79,80], variant calling [81,82], and fusion transcript detection [83] are some of the other applications of RNA-sequencing with particular importance in cancer.

mRNAs however, do not constitute the only RNA entities with relevance to cancer [84]. It is now evident that a great portion of non-coding DNA is translated to functional non-coding RNAs (ncRNAs) that are involved in almost all the aspects of cellular processes [85]. There are two general categories of ncRNA: small non-coding RNAs (sncRNAs) that are less than 200 nucleotides in length and long non-coding RNAs (lncRNAs; >200 nucleotides) [86]. sncRNAs are further categorized into a number of RNA types including microRNAs (miRNAs), small nuclear RNAs, and piwi-interacting RNAs. MiRNAs are probably the most widely studied form of ncRNAs [87,88]. With their recognized role as important regulators of many cellular processes, miRNAs are firmly established as essential players in tumorigenesis and cancer progression and have been widely studied as potential biomarkers and therapeutic targets [89–92]. The role of lncRNAs in cancer, however, is a more recent emerging view [93,94]. LncRNAs exert a variety of biological functions through interaction with a plethora of different types of macromolecules. LncRNAs roles in gene expression regulation through interactions with chromatin, protein complex assembly or disassembly, and their interplay with mRNAs have been widely studied [95]. Several lines of evidence attribute a role to lncRNAs in the regulation of virtually all of the cancer hallmarks [96,97]. The vast number of tissue- and cell-specific lncRNAs along with their importance in the regulation of cellular functions underscores their potential for annotated biomarker discovery in cancer diagnosis, prognosis, and treatment [98] as well as their potential employment as therapeutic targets [99].

2.2.1 Experimental workflow and data analysis pipeline

Illumina short-read sequencing is currently the dominant platform for transcriptomics studies [100]. The process starts with RNA extraction and target RNA enrichment to remove unwanted rRNAs or

specifically select for poly-adenylated RNAs through oligo-dT incorporation [101]. However, since other RNA types might be of interest, rRNA depletion can provide more encompassing results [102]. In any case, in the next step, the extracted RNA is subjected to fragmentation in order to become compatible with the short-read sequencing technologies. This is usually done through enzymatic digestion or by using divalent cation-containing solutions [102]. Next, reverse transcription is performed. The second strand of the synthesized cDNA is usually tagged with the incorporation of dUTPs. After the adaptor ligation, the tagged cDNAs are subjected to digestion in order to achieve a strand-specific library [103]. The remaining strands are amplified through PCR and are finally sequenced. The required sequencing depth (total number of reads) is determined by the goal of the study and the nature and condition of the sample [104]. While 15 million reads are considered a saturation point for gene expression profiling [77], a minimum of 70 million reads are required for the accurate quantification of alternative splicing events [105]. This general framework can be modified based on the experimental goals and the RNA type under investigation [106]. The use of single-end or paired-end sequencing or enriching for unique reads restricted to the 3′ end for each transcript in order to analyze DGE are examples of such modifications [102]. Another example is to take advantage of unique molecular identifiers (UMIs) to account for the misrepresentation of biological expression differences due to PCR amplification [107].

The next steps are quality control and pre-processing of the acquired reads [104]. To perform DGE analysis, the level of expression for each gene should be measured from RNA-seq reads. For that purpose, the acquired reads are mapped to an annotated genome or transcriptome using tools such as STAR [108], BWA [109], and TopHat2 [110]. Gene expression is then quantified based on the number of reads that have been aligned to each gene using tools such as HTseq-count [111]. Alternatives include methods such as Sailfish [112], Salmon [113], and Kallisto [114], which implement k-mer counts, quasi-mapping, and pseudo mapping, respectively. After batch effect correction [115,116] and data normalization [117], the last step is the actual differential gene expression analysis. While almost all of the popular methods for transcript quantification have been

shown to perform equally well [118], the utilized tool to assess differential gene/transcript expression is an influencing factor in this process. Multiple tools (e.g. NOIseq [119], limma+voom [120], and DESeq2 [121]) are known to perform a high-quality DGE analysis and are accepted as standard tools for DGE assessment [122]. Moreover, the usage of a combination of these tools has been suggested as an effective approach [118]. Quality control in multiple steps of the process (RNA quality, raw reads, alignment, and quantification) is also highly recommended [123]. Comprehensive quality control tools such as the NGS QC toolkit [124], RSeQC [125], and Qualimap2 [126] are widely applied to fulfill this purpose.

Multiple tools and web services such as IDEAMEX [127] facilitate an integrated DGE analysis for researchers with a minimal computational background. BP4RNAseq [128] is another user-friendly tool that has been recently introduced and can be utilized for a highly facilitated gene expression quantification. There are also multiple tools and pipelines that are not restricted to DGE analysis and can be implemented for a variety of RNA-seq data analysis purposes. RNACocktail [129] is a comprehensive RNA-seq analysis pipeline incorporating a variety of powerful tools for a variety of purposes including RNA variant-calling, RNA editing, and RNA fusion detection.

RNA-sequencing is at the forefront of single-cell sequencing technologies [130,131]. Sensitive full-length transcript sequencing platforms such as MATQ-seq [132] with the ability to capture and sequence ncRNAs herald the arrival of a new level of sequencing capacity. The general workflow for single-cell sequencing is similar to the bulk RNA-sequencing workflow described above [133]. It is indeed possible to perform most of the computational processing steps with the bulk RNA-sequencing methods. However, low levels of starting material coupled with additional technical requirements (such as cell-specific barcoding to be able to demultiplex the resulting data from multiplexed sequencing) and other challenges (such as the possibility of capturing damaged, dead, or multiple cells) necessitate the development of computational methods tuned for single-cell analysis [134,135] (see ***Table 1*** for a list of single-cell RNA-sequencing tools). It should be noted that large-scale comparative studies are required for the assessment of the utility of these tools in

comparison with one another and with the tools designed for bulk-RNA sequencing analysis. Indeed, bulk-RNA sequencing analysis tools have been shown to be capable of producing satisfying and in some cases, superior results compared to that of the tools specifically designed for single-cell RNA-seq [136].

2.2.2 Challenges and perspectives

A current challenge in RNA-sequencing is that the reconstruction of full-length RNA molecules from short reads is error-prone [104]. This results in incorrect assignment of reads and misrepresentation of isoform abundances and also makes isoform discovery a challenging task. Long-read technologies, as well as synthetic long-read methods, hold the promise of solving this inconvenience [100]. However, various challenges remain to be addressed. Long-read technologies are particularly low in throughput. This problem in turn would result in a reduced experiment size and low sensitivity of differential expression [100]. Hence, using a long-read technology is not currently recommended for DGE analysis, particularly when the study involves low expression levels. The high error rates and additional costs are prohibitive elements regarding long-read technologies. Moreover, the rigorous requirement to avoid RNA degradation and shearing during sample handling makes the achievement of high-quality samples laborious. However, the combination of short-read with long-read sequencing methods enhances the quality and accuracy of transcript isoform expression analysis. For instance, by combining these technologies and using algorithms for hybrid assembly of short and long reads (hybridSPAdes; [137]) enhanced results for *de novo* transcriptome assembly (e.g. with rnaSPAdes; [138]) can be achieved.

2.3 Proteomics: Studying the frontline of phenotype manifestation

Virtually all the regulatory mechanisms governing the central dogma of biology eventually serve to determine the set of expressed proteins, their expression levels, and the manner in which they function; the deviations of which from normal status can result in a malfunctioning system and give rise to various disorders such as cancer [139]. Proteins can be considered as frontline agents of

phenotype manifestation and hence, studying proteome level regulatory mechanisms, such as post-translational modifications (PTMs), the inherent properties of proteins (e.g. their 3D structures), and protein-protein interaction (PPI) networks is essential if representative views of the normal and perturbed cellular system are to be achieved. Moreover, the validity of inferring protein abundance from mRNA expression has been questioned due to the lack of consistently strong correlations between mRNA and protein abundance [140], suggesting that the direct assessment of protein abundance is a more reliable source.

All of the categorized hallmarks of cancer are either directly regulated by proteins or are highly affected by them [141]. Proteins function in protein assemblies and highly complex networks. In this context, malfunction in any member of these networks can potentially result in the disruption of the activity of other members of the same network. Therefore, an important goal of proteomics studies, in addition to assessing genome-wide protein expression under various conditions, is to achieve comprehensive and functional models of all the physical protein interactions both in normal and perturbed conditions [142]. Equally important is the study of PTMs. With more than 450 types of PTMs, these modifications regulate protein expression levels and almost all cellular processes, such as immune response, apoptosis, tumorigenesis, and cancer progression [143–146]. Exploration of these and other aspects of cell biology from omics data of other levels is either impractical or impossible. Collectively, current proteomics technologies and approaches provide researchers with powerful assets in the quest of achieving a functional view of the cellular system and addressing fundamental questions regarding the biology of cancer as well as discovering biomarkers and actionable therapeutic targets [147,148].

2.3.1 Experimental workflow and data analysis pipeline

Multiple methods have been developed to assess the proteomic landscape of cells and tissues. Targeted and top-down proteomics [149,150] are two of the established branches of such methods with dedicated software tools and platforms [151–153]. However, data-dependent bottom-up or

"shotgun" proteomics through liquid chromatography-tandem mass spectrometry (LC-MS/MS) is currently the *de facto* standard approach for genome-wide proteomics analysis [154]. The workflow for shotgun proteomics is variable and context-dependent. A general workflow based on the current best practices can be presented as follows: after the lysis of the samples, the disulfide bridges of the extracted proteins are disrupted through reduction and alkylation of the cysteine residues. Next, the proteins are subjected to enzymatic digestion through the addition of proteinases (most commonly Lys-C followed by trypsin). One- or two-dimensional chromatography is next applied, the latter is recommended to increase the dynamic range (i.e. to provide the possibility for low-abundance proteins to be identified) [155]. Currently, the most effective approach is to subject the samples to basic reversed-phase chromatography followed by acidic reversed-phase chromatography as the second dimension [156]. There is also the choice between label-free and isobaric labeling (using iTRAQ [157] or tandem mass tags (TMTs, [158])). Isobaric labeling approaches are recommended due to the provided capacity for multiplexation and the reduction of errors from manual sample handling as well as higher precision in quantification, especially when PTMs are the target of the study [155]. The wet lab procedure is concluded by the acquisition of MS spectra from MS/MS. Orbitrap-based MS/MS is the current standard. It is also possible to add a third stage (MS3) by combining Orbitrap and Ion Trap methods and it has been shown to be effective when facing highly complex samples [159]. For comprehensive and step-by-step workflows for the wet lab procedure refer to [155] and [159].

Although methods exist for *de novo* identification of peptide sequences [160], current approaches still suffer from high error rates. The preferred method is to first prepare a database of all the known protein sequences (comprehensive databases such as Uniprot [161] can be exploited for this purpose) and subject them to *in silico* digestion according to the properties of the proteinase enzymes that were utilized during sample preparation. The resulting *in silico* produced peptides are then assigned theoretical spectra and the experimentally acquired spectra are searched against this database. Each match is scored based on the similarity and the highest-scoring match reveals the

identity of each peptide with a certain false discovery rate (FDR). A stringent FDR of 1% is recommended [162]. The recommended approach to control for this FDR is the target-decoy search strategy [163]: a parallel database of incorrect peptides is constructed (usually through reversion of the peptide sequences of the main database). Matches to this database are obviously false positives and hence, can reveal the FDR based on the utilized filters. Using this method, one can tune the applied filters to achieve a suitable FDR. The identified peptides are then assigned to their respective proteins. Peptides with less than 7 residues are usually non-unique and are prone to erroneous protein assignment and thus, are recommended to be excluded [162].

Proteomics data needs to be preprocessed (including normalization, filtering, etc.) before they can be interpreted in a biological context. After preprocessing, the data can be manipulated to yield functional information through a variety of approaches. Differential expression analysis is a common approach with subsequent context-specific analyses such as expression signature discovery and co-expression network analysis.

The general workflow provided here can also be modified in order to customize the study for the analysis of PTMs [164], PPIs, and subcellular localization [142]. For the analysis of PPIs, target protein complexes should be isolated from the cell lysate. Co-immunoprecipitation (Co-IP) is a common approach for this purpose [165]. Co-IP involves the attachment of specific antibodies to bait proteins (proteins whose interacting partners are under investigation). These antibody-protein complexes are captured by agarose beads attached to A/G proteins and are “pulled-down” by means of centrifugation. Proteins in tight interaction with the bait proteins are also precipitated in this step and the unbound components of the lysate are discarded. The captured proteins can then be subjected to MS to identify PPIs. Tandem affinity purification (TAP) is a similar approach with enhanced purification that involves tagging the bait protein at its N-terminus by a TAP tag (usually a calmodulin-binding domain followed by a highly specific protease cleavage site followed by an IgG-binding fragment) prior to two steps of purification by centrifugation [166]. The major problem associated with these approaches is their restriction to identifying highly stable PPIs. For the

identification of more transient interactions in complex biological samples, another method termed crosslinking-MS (XL-MS) which also has the advantage of providing spatial information is favored [167]. This method is based on covalently binding residues in two proteins through two reactive groups (usually amine- groups due to the prevalence of lysin residues in protein structures) that are connected via a spacer with a finite distance. This limited distance confers a spatial constraint on the residues that can be linked; making the crosslinking possible only between proteins in close proximity (i.e. interacting proteins). As for the PTMs, the mass shift in the peptides due to these modifications is identifiable by LC-MS. However, an additional enrichment step for the peptides with the modification under investigation is required [168]. Various strategies for this enrichment including implementation of immunoaffinity precipitation (using antibodies highly precise for specific types of modification) and chromatography-based approaches (e.g. immobilized metal ion affinity chromatography, metal oxide affinity chromatography, etc.) have been devised. The most suitable approach, however, is dependent on the type of modification under study and the specific physical/chemical properties it confers to the peptides (refer to [169] and [168]).

MaxQuant [170] is a popular comprehensive platform that along with Perseus [171] facilitates the entire procedure of shotgun proteomics data analysis. Moreover, dedicated platforms for computational analysis of PTMs and PPIs exist [172,173]. In addition, a recently developed comprehensive toolkit named “Philosopher” [174] demonstrates a movement towards making these computationally sophisticated methods accessible to a broader community.

The prospective results of the “discovery” shotgun proteomics can be channeled into “hypothesis-driven” targeted proteomics for validation in order to extract actionable and clinically relevant directions from the plethora of information resulted from shotgun proteomics [175]. Targeted proteomics approaches are higher in sensitivity and dynamic range and tackle the problem of irreproducibility associated with shotgun proteomics which is due to the stochastic nature of precursor ion selection in shotgun approaches. Targeted proteomics is developed based on prior knowledge about the proteins of interest and the selection of signature peptides that specifically

represent those proteins. Selected reaction monitoring (SRM) is a widely-used targeted approach. A triple quadrupole instrument is used to filter the target peptides based on their predetermined mass-to-charge ratio which combined with their elution time can be sufficiently specific. The filtered peptides are subsequently fragmented using collision-induced dissociation and the resulting fragment ions are once more filtered for specific fragments based on a predetermined mass-to-charge ratio. This process is repeated for multiple different fragment ions of each filtered peptide and hence, peptides are identified and quantified utilizing MS spectra [176]. Parallel reaction monitoring (PRM) is a similar approach which through the implementation of an orbitrap or time-of-flight instrument removes the second filtering step by analyzing all the fragment ions simultaneously and provides more accurate results [176].

2.3.2 Challenges and perspectives

In spite of the remarkable progress made in proteomics methods in the last decade [147], drawbacks such as the cofragmentation problem [177] still exist and experiment design approaches, as well as computational strategies, are being constantly revised to compensate for these [178]. Overall, reduction in costs and a further increase in the sensitivity of mass spectrometers can be considered as major factors that can enhance the efficiency and accessibility of proteomics analyses [179]. Specific to targeted proteomics, a major drawback of SRM and PRM approaches is that the analysis is restricted to the preselected target proteins. Recent advances in data-independent acquisition methods (particularly SWATH-MS) circumvent the need for repeated measurements for each target protein by allowing posterior querying of the data for the desired peptides while providing multiplexing capacities comparable to shotgun proteomics [180]. However, data-independent acquisition methods lack the sensitivity of SRM and PRM and are therefore inferior to these approaches when dealing with very low-abundant proteins. In addition, SWATH-MS is still facing challenges regarding ease of data analysis [180].

From the clinical perspective, minimally invasive sample collection is critical. Body fluids (e.g. blood, saliva, urine, tears, etc.) are readily available rich sources of biomolecules (e.g. over 12,000 proteins only in plasma) with altering compositions during tumor development which can be used as tumor and/or stage-specific biomarkers [181]. Proteomics approaches were generally successful in discovering such biomarkers [182,183]. A major pitfall associated with body fluid biomarker discovery, however, is the massive dynamic range: a handful of enormously abundant proteins mask the presence of lowly abundant molecules of interest. Strategies such as immunodepletion of high-abundance proteins have been devised, which nevertheless face the caveat of information loss due to unspecific bindings to affinity ligands [184]. Nonetheless, the achievements of multiple efforts in recent years underline the possible widespread utilization of these sample types in clinical practice in the future [185,186].

Single-cell proteomics is a promising prospective approach which is still in its infancy. For single-cell technologies to become a feasible practice in proteomics, advances in both technological and computational aspects are required [187]. Considerable increase in MS sensitivity and the development of specialized tools for the analysis of such data are prerequisites of making single-cell proteomics practical. Nevertheless, various multidisciplinary efforts are already turning the dream of single-cell proteomics into reality [188].

2.4 Metabolomics: Exploring the survival strategies of cancer cells

During cancer initiation and progression, cellular systems are reprogrammed to grow and proliferate at exceptionally high rates and to acquire an enhanced capacity for survival under extreme conditions [141]. Clearly, a considerable portion of this reprogramming is dedicated to shaping an altered form of metabolism that is able to meet the massive energy needs and to provide required anabolic precursors for these highly demanding self-centered systems. Indeed, almost every aspect of cellular metabolism is affected during cancer progression [189] and since the metabolic status of a sample can be considered as the ultimate downstream manifestation of the

effects of both intrinsic (e.g. genetic) and extrinsic (i.e. environment) factors on the biological system [190], valuable insights can be gathered from the study of the metabolome.

Two core metabolites with altered metabolic pathways in cancer are glucose and glutamine [191]. Excessive glucose fermentation, overexpression of the rate-limiting enzymes of the glycolysis branch pathways, constitutive glucose influx, as well as an increased expression rate of glutamine synthesis are examples of such alterations that cancer cells exploit to provide themselves with modified sources of energy and a large collection of biosynthetic precursors [189]. In addition, cancer cells develop scavenging strategies in order to survive under the commonly encountered nutrient-poor microenvironment. These strategies include autophagy [192], consumption of extracellular proteins through macropinocytosis and subsequent lysosomal degradation of these molecules [193], entosis [194], and phagocytosis [195], as well as induction of fatty acid release from neighboring cells [196]. Cancer cells also highly influence the condition of their microenvironment. The high rate of glucose fermentation results in the accumulation of considerably high levels of extracellular lactate and H^+ which in turn contribute to angiogenesis, immune response suppression, and tumor invasiveness [189]. Since the survival of cancerous cells is highly dependent on this altered metabolic status, the metabolome is an active area of research for the discovery of cancer biomarkers as well as the identification of potential therapeutic targets [197,198].

The contribution of metabolites to the initiation of signaling cascades and their effect on the epigenetic landscape as well as PTMs are other topics of investigation. Through these investigations, the role of metabolites not only as molecules with altered behavior downstream of cancer initiation and progression but also as etiological agents (i.e. oncometabolites) that contribute to system perturbations is being rapidly established [199]. Further studies of the metabolome in this context have the potential to shed light on novel aspects of cancer biology.

2.4.1 Experimental workflow and data analysis pipeline

Due to the inherent chemical homogeneity of the polymers of genome, transcriptome, and proteome, it is possible for a single platform to capture a holistic snapshot of each respective layer. However, this does not hold in metabolomics owing to the chemical heterogeneity of different classes of metabolites [200]. Proton nuclear magnetic resonance (1H NMR) and MS-based methods are the most common approaches for metabolomics data acquisition; all of which are associated with various advantages and disadvantages [190].

NMR is highly reproducible, conveniently quantifiable, requires minimal sample preparation, and unlike MS-based approaches is nondestructive [190,201,202]. Moreover, it is considered the gold standard method for the elucidation of the metabolite structures [203]. Nevertheless, NMR suffers from low sensitivity and it is only capable of detecting 20-50 metabolites per sample which is an inadequate number for systems-level analyses [190]. MS-based approaches, on the other hand, possess the advantage of high sensitivity and are widely adopted for untargeted and system-level metabolomics analyses due to their capability to detect 100-1000 metabolites per sample [200,203]. Gas chromatography-MS (GC-MS) and LC-MS (or LC-MS/MS) are the most commonly used methods for MS-based metabolomics [204]. GC-MS is cost-effective and has the advantage of a virtually automated metabolite identification process. However, it is only applicable to volatile and thermally stable metabolites or those that can be adapted for the process with chemical derivatization [203]. This limits the versatility of GC-MS. In addition, the derivatization process can introduce artifacts and might result in erroneous quantification because of incomplete derivatization [205]. Unlike GC-MS, LC-MS does not require derivatization and with the ability to capture molecules in a wider weight range, it is highly versatile and efficient [190,203,204,206]. While these advantages make LC-MS the most widely applied method in the field, researchers are encouraged to opt for a combination of these approaches to achieve a more comprehensive representation of the metabolic status of the sample [201]. The workflows for all of the above-mentioned approaches are somewhat similar, with nuances and differences in the steps and applied

algorithms. However, due to the extensive utility of the LC-MS and LC-MS/MS, these approaches are the main focus of this section.

Unlike NMR, MS-based analysis needs a sample preparation step consisting of protein precipitation and liquid-phase extraction [207]. The higher susceptibility of the metabolome to alter under different conditions in comparison to the other omics layers [208] means that careful experimental design is a requirement to minimize confounding factors. The instruments with high mass-resolving power such as LTQ-Orbitrap and Q-TOF are instruments of choice for systems-level metabolomics. Electrospray ionization (ESI) is the most widely applied ionizing method in order to make the metabolites detectable in LC-MS metabolomics [204,209]. Of note, the validation of the results of untargeted studies through targeted approaches can increase the reliability of the acquired data [206].

The general computational workflow consists of preprocessing, peak detection or annotation, post-processing, and statistical analysis of the resulting data [210]: After the data is obtained, it should be subjected to the preprocessing procedure in order to enhance comparability and management [190]. Preprocessing usually starts with peak picking which is the process of detecting the actual informative regions of spectra and removing the background noise. For MS-derived data, a deconvolution step is required to reduce redundancy. Another requirement is the alignment of matching peaks between different samples [211,212]. A practical and popular approach for peak annotation (i.e. the assignment of the observed peaks to actual metabolites) is to search the data against the existing spectral libraries in a process similar to what has been described in the proteomics section. The desired information for metabolites is acquired by inquiring metabolome databases such as the Human Metabolome Database (HMDB) [213], METLIN [214], and MassBank [215]. It is also possible to implement a target-decoy strategy to control for the FDR. An innovative approach regarding the construction of a decoy database for metabolome studies has been proposed by Wang *et al.* [216] which is performed by violating the octet rule through the addition of extra hydrogen atoms to the molecular structures. A post-processing procedure is

performed prior to downstream analysis and interpretation of the data. Post-processing includes data filtering, imputation to account for the missing data, and normalization [210]. Data filtering is an important step in order to remove uninformative data while avoiding the loss of biologically meaningful information [217]. Recently, Schiffman et al. proposed a data-adaptive pipeline for data filtering procedure [218]. A variety of normalization methods both sample-based and metabolite-based exist. Among these, Variance Stabilization Normalization (VSN), which accounts for sample-to-sample variations and metabolite-to-metabolite variances, has proven to be a suitable and versatile method [219]. However, a recent study recognized 21 different normalization strategies based on the combination of sample-based and metabolite-based methods as consistently well-performing [220]. For an in-depth review of the computational process of the metabolomics studies we refer the readers to [221].

There are multiple robust tools for each step of the computational workflow (refer to [222] and [210] for comprehensive lists of available tools). Metabolomics researchers also enjoy the benefits of existing versatile and comprehensive workflows that cover multiple steps or even the entirety of the metabolomics computational aspects. Examples of highly popular such workflows are XCMS online [223], Galaxy-M [224], and MetaboAnalyst [225]. For a complete step-by-step guide to how to use MetaboAnalyst, we refer the readers to [226]. Moreover, novel approaches and platforms are being rapidly produced. MetaX [227] and JumpM [228] are examples of such novel and potent approaches.

2.4.2 Challenges and perspectives

The Metabolomics field is rapidly growing with the emergence of innovative technologies such as iKnife [229]. iKnife is able to perform *in situ* MS analysis with applications such as discrimination between normal and malignant tissues with 100% accuracy [230]. Single-cell metabolomics still struggles with challenges such as low throughput and sensitivity as well as computational inefficiencies. Nevertheless, efforts are being made to address such shortcomings [231]. The study

of the metabolome is not restricted to the methods discussed in this section. There are also alternative approaches such as isotope tracing fluxomics with the goal of delineation of the distribution of the metabolites in the samples of interest, and matrix-assisted laser desorption ionization-based MS imaging (MALDI-MSI) [232]. Moreover, the diverse advantages of NMR technologies attracted efforts for its synchronization for the current needs of metabolomics studies [233]. These alternative technologies, while providing the research community with improved analytical capacity, bring about their own challenges and inconveniences. Future years are expected to witness increased sensitivity of analytical platforms, improvement of interoperability among computational tools [210], as well as elevated specificity of metabolite biomarkers of cancer and enhancement of pharmacometabolomics (i.e. prediction of drug response through metabolomics) [234].

3 Multi-layer approaches

Although isolated analysis of each of the individual omics layers has substantially contributed to our understanding of a diverse range of biological phenomena, this type of analysis has an inherently limited capacity for characterizing the integrated nature of biological units. When studying the cellular system, its complexity with intertwined and highly convoluted networks of interactions and regulations necessitates a multifaceted approach where different layers of data, generated either through single-layer omics approaches or other means of data acquisition (e.g. studies of molecular interactions, imaging, etc.), are simultaneously analyzed in an integrated manner [235]. Cancer is a systemic disease and thus, achieving an accurate picture of this perturbation requires homogenization of all the different types of single-layer data through integrative approaches. This is indeed the goal of large-scale efforts such as the Cancer Genome Atlas (TCGA; [13]) which by providing publicly available multi-layer data from various tumor types, empower researchers across the globe with an unprecedented capacity for systems-level analysis of cancer (Figure 3).

Integrative approaches have three main advantages. (1) With observations validated across multiple layers of information, they allow for more reliable and representative interpretations, (2) they can substantially contribute to the delineation of the interplay among molecular levels and shed light on the hierarchy of causation, and (3) they reduce our blind spots by circumventing our limitations through combined utilization of the technological and computational power in each level.

Notably, omics data is not the only possible source of information that can be purposefully integrated in cancer studies; other types of data such as histopathological information can provide an extended panorama of tumor biology. Reportedly, the integration of histopathological features with molecular data outperforms predictions based on omics data or histopathological information in isolation in various types of cancer [236]. In one such study, an integrative, machine learning-based analysis of histopathological, molecular, and clinical data of 538 lung adenocarcinoma patients from TCGA cohorts resulted in an integrated model with more accurate prognostic power for survival outcomes of stage I lung adenocarcinoma patients [237].

The heterogeneity of the generated data across different layers is a major challenge in integrative studies [238]. However, the undeniable advantages of data integration have prompted numerous efforts to overcome its challenges. See [239] and [240] for comprehensive explorations of integrative methods, databases, and tools. In addition, Supplementary Table S2 describes some of the prominent tools and methods for the integration of multi-modal data and their comparative performance. Here, we provide an in-depth description of proteogenomics and network-based data analysis. The former is a remarkable example of how the integration of multiple levels of information can reduce our blind spots and increase the accuracy and reliability of our interpretations and the latter is a major approach for data interpretation and a robust scaffold for data integration and modeling.

3.1 Proteogenomics: vertical integration of genomics, transcriptomics, and proteomics data

Since genomic alterations are regarded as the molecular cause of tumorigenesis [7], the emergence of next-generation sequencing (NGS) technologies held the promise to greatly accelerate the identification of pathogenic alterations and thereby facilitate the design of highly effective therapeutic interventions and indeed, a variety of candidate treatments such as personalized immunotherapy, cancer vaccines, and gene therapy are being introduced [241]. However, not all of the patients stratified based on their genomic data benefit equally from the applied therapeutic interventions and the levels of response within each group of patients are diverse [242]. This has been attributed to the fact that most of the currently used treatments target specific proteins rather than genomic alterations and a great number of confounding elements are out of grasp due to the lack of proteomic information [243].

Despite recent attempts to predict specific types of PTMs [244], genomics data analysis cannot account for the numerous protein-level adaptation events in the cellular environment [243]. On the other hand, there is a considerable load of somatic mutations in cancer cells which in turn give rise to previously unidentified peptide sequences. Since proteomic analysis relies on previously identified protein sequences (to avoid false peptide sequences in *de novo* sequencing experiments), single-layer analysis of proteomic data is highly limiting in the cancer context. These and other challenges, which will be discussed here, can be addressed through vertical integration of genomics, transcriptomics, and proteomics data which is collectively termed proteogenomics (Figure 4) [245,246].

3.1.1 Experimental workflow and data analysis pipeline

The backbone of proteogenomics studies is the construction of customized protein sequence databases [245]. As previously stated in the proteomics section, the identification of peptides in samples subjected to shotgun proteomics experiments is achieved by matching the spectra against a protein sequence database [247]. However, public protein databases (e.g. UniProt and PDB) do not contain previously unidentified protein sequences such as novel altered proteins which are

696 frequently encountered in tumor-derived samples [248]. To overcome this obstacle, NGS data
697 acquired from the same sample (e.g. via WES, WGS, and RNA-seq) can be exploited to construct
698 a customized protein sequence database that contains all the hypothetical protein sequences that
699 can be inferred from the genomics or transcriptomics data and then, match the MS/MS spectra
700 against this sample-specific database [249,250].

701 The complexity of the expression system in eukaryotes makes the matching of the proteomics
702 spectra against a customized database predicted from genomics data computationally ineffective
703 and error-prone because the size of such databases will exceed any acceptable threshold [251].
704 However, customized databases from transcriptomics data are more effective and accurate since
705 they consider only expressed transcripts. To construct a customized protein database from
706 transcriptomics data, raw nucleotide sequences should be assembled into full-length transcripts.
707 There are two approaches for full-length transcript assembly: genome-guided and *de novo*
708 transcriptome assembly. Genome-guided approaches are routinely used for cancer studies.
709 However, coupling these approaches with *de novo* transcriptome assembly approaches is advised
710 [252]. *De novo* transcriptome assembly methods have the advantage of being capable of identifying
711 novel transcripts that can't be identified through reference-guided methods either due to errors in
712 the reference genome or because they are completely missing (i.e. tumor viruses) [253]. A recent
713 comparative study [252] suggested that the performance of the various existing *de novo* assembly
714 tools is dependent on the study design and the species under study. In the cancer context, where we
715 are usually dealing with human samples, Trinity [254], Trans-ABySS [255], SOAPdenovo-Trans
716 [256], and SPAdes [257] are generally well-performing tools [252]. Merging the results obtained
717 from multiple assembly tools with posterior quality control evaluation is currently considered best
718 practice. Notably, long-read sequencing technologies have the potential to circumvent challenges
719 of *de novo* transcriptome assembly. With PacBio and Nanopore technologies, read lengths of >10
720 kb are routinely achieved, capturing full-length transcripts.

Multiple tools are available for customized database construction including Galaxy-p [258], QUILTS [249], customProDB [259], and PGA [260]. Importantly, the PGA pipeline is not limited to MS/MS data searching. It incorporates database construction steps that can be done using a genome-guided approach or via a *de novo* transcriptome assembly approach and also includes post-processing steps including FDR calculation, protein inference, and spectrum annotation. In addition, the capacities of Galaxy-p for custom workflow construction prompted the development of comprehensive workflows [261] that encompass the entire computational process of proteogenomics. For a list of available tools and resources for proteogenomics studies refer to ***Table 2***.

The process of matching MS/MS spectra against a customized database is achieved by utilizing database search engines such as X!Tandem, MS-GF+ [262], and Comet [263]. Among these, the widely used X!Tandem software has been shown to have the highest false-negative rate and hence, it is not recommended to exclusively use this engine [264]. Since effective quality control methods for novel peptide identification can be utilized downstream of the matching process, a high level of false-positive can be tolerated. Hence, the best approach in this step is to combine the results of multiple search engines to gain a more comprehensive collection of putative novel peptides. Novel peptides that have been identified through the matching step can then be further validated. PepQuery [265] is a freely available tool that can be applied as an optional quality control step and can significantly reduce false positives. The definitive validation of identified novel peptides, however, can be achieved through targeted proteomics assays [243].

3.1.2 Applications

There is a variety of molecular events that can potentially give rise to a wide range of protein alterations such as chimeric proteins or single amino-acid variants in cancerous cells. However, not all of these events result in expressed proteins and even if expressed, the resulting proteins might be unstable and subjects to early degradation. Proteogenomics is an ideal approach for protein level

validation of the stable expression of these molecular events [246]. Moreover, protein-level analysis of current gene models and their somatic variations by means of proteogenomics enables the validation or correction of previous predictions of the sequence, structure, and ultimately the function of the respective proteins [246,266]. Additionally, deregulation of alternative splicing in cancer under the influence of perturbed splicing factors and altered signaling cascades is a known phenomenon [267,268]. Alternatively spliced isoforms can not only serve as tumor-specific biomarkers but can also provide stage-specific signatures and putative therapeutic targets [80]. Empowered with the capacities of both transcriptomics and proteomics, proteogenomics proves to be a competent approach for studying oncogenic splice variants and specific pipelines towards this purpose have already been developed [269].

PTMs are known to play essential roles in the biology of cancer cells [143,144]. Genomic alterations in cancer can have profound effects on protein modifications (e.g. through the addition or disruption of modification sites or alteration of PTM regulator proteins) and in turn on the signaling cascades and regulatory networks of cancer cells [251,270]. Since PTMs cannot be accurately predicted from genomics data, proteogenomics can become the tool of choice for exploring the effects of aberrations in the genome on the downstream PTM alterations [271]. In addition, it is now widely accepted that quantitative mRNA expression data is not an ideal indicator of protein expression levels and the extent to which they biologically correlate is a matter of debate [272]. Since protein expression levels are of importance both for functional inferences and therapeutic interventions, accurate measurement of protein expression levels is crucial [243]. Proteogenomics studies can not only provide us with protein expression data, but they also have the potential to deepen our understanding of the biology of this difference in expression levels.

The host immune system is known to be effective in the elimination of cancer cells [273]. For the host immune system to be able to confront cancer cells, neoantigens, which are predominantly results of the processing of altered proteins by the antigen processing pathways, should be presented as human leukocyte antigen (HLA) ligands at the cell surface and be identified by T-cell

surveillance [268,274]. the process of immune response to cancer cells is being studied with the goal of designing therapeutic interventions known as cancer vaccinations that attempt to elicit the T-cell immune response against cancer cells [275–277]. Proteogenomics can greatly accelerate the pace of neoantigen discovery and by providing candidate clonal neoantigens result in a more efficient vaccination process [278,279]. Moreover, proteogenomics studies can help delineate the underlying mechanisms of immune system evasion by cancer cells [280].

The above-mentioned applications can be used to filter more important genomic alterations, distinguish between driver and passenger mutations [281], and make for more efficient biomarker discovery [282–284]. A recent study [266] showcased the massive potential of proteogenomics studies from unraveling uncharted aspects of cancer biology to opening new avenues towards precision oncology. From PTM analysis of proteins to prioritization of somatic copy-number alterations, they exploited the full potential of current proteogenomics technologies. Importantly, they demonstrated that proteogenomics studies can result in more efficient unified multi-omics cancer subtypes that can serve to acquire an enhanced ability for prognosis, diagnosis, and precision interventions.

3.1.3 Challenges and perspectives

A long-standing challenge in the field of proteogenomics is the appropriate FDR estimation for matched peptides after database search [246]. As discussed in the proteomics section, a widely used approach is the target-decoy search strategy [163]. Since assuming the same FDR for both novel and previously identified peptide sequences is an underestimation of the FDR value for novel peptides, the efficacy of this method in proteogenomics studies has been questioned and substitute approaches such as separate FDR estimations for novel and previously identified peptides have been suggested by Nesvizhskii *et al.* [246]. Wen *et al.* [264], however, in a comparative study of FDR estimation methods utilized the prediction of retention time for peptides in comparison with the actual observed values as an evaluation metric for different quality control strategies and

identified global FDR estimation by target-decoy search (in order to attain a high level of sensitivity) with a posterior filtering step to restrict false positives (using PepQuery) as the best approach for neoantigen discovery.

Although targeted MS-based assays hold great promise for the clinical translation of the discovered biomarkers through proteogenomics studies, there are still challenges that should be addressed [243]. Targeted multiple reaction monitoring assays can be used not only to validate the results of proteogenomics analyses but can also provide clinicians with a cost-effective multiplexed platform that can analyze a high number of target proteins from a variety of sample types (e.g. urine, secretions, etc.) with satisfying sensitivity and specificity. However, there is still room for improvement since the sensitivity is not enough for dilute samples and single-cell analysis [285].

Recent advancements in proteomics technologies [286,287] and clinically valuable demonstrations such as the possibility of a micro-scaled proteogenomics study of tissues as small as 25 μg [288] are setting the stage for the emergence of a more precise and cost/time effective landscape for proteogenomics. Moreover, single-cell proteogenomics is evolving and has the potential to considerably increase our understanding of intratumoral heterogeneity [289–291]. It is expected that a greater number of researchers join this field in the years to come. However, the high number of existing tools that provide complementary results and should be utilized in combination with one another in multiple steps of the study [264,284,292] is probably a prohibitive element in attracting new researchers to the field. Other prohibitive elements are the required computational expertise and the lack of unified and comprehensive databases with user-friendly interfaces that are specifically tuned for proteogenomics studies. Although efforts have been made to provide comprehensive workflows for different study goals [293,294], international collaborations are required to overcome existing challenges and provide gold standard workflows for proteogenomics studies.

3.2 Network-based data integration

A huge amount of information regarding the interactions among molecules and biological pathways is stored in public data repositories such as STRING [295], BioGRID [296], InnateDB [297], KEGG [298], Reactome [299], VMH [300], WikiPathways [301], etc. These data are generated either from *in vivo* and *in vitro* experiments or from *in silico* predictions [302] and are essential in providing a system-based context for omics data. Biological systems in the form of interaction networks and pathways can serve as frameworks on which omics-driven data can be integrated, analyzed, and interpreted [303,304].

Combining the prior knowledge of interactions in the form of networks and pathways with genome-wide data generated through single-layer omics approaches is used to overcome issues in the interpretation of omics data by providing a larger context. On the one hand, omics data on their own are merely a representation of existing molecules and their abundances at a particular point in time. Extracting patterns and understanding the underlying mechanisms of a condition from an omics dataset in isolation is challenging [305]. On the other hand, molecular interaction networks and pathways, although highly informative, do not account for the dynamics of the cell in different states and phases. The integration of interaction networks and pathways with omics datasets facilitates pattern detection and allows the study of the dynamic nature of the cell [306]. This is of particular importance for understanding the mechanisms of complex multistage diseases such as cancer. This integrative approach has been shown to be superior to the isolated analysis of either networks or omics data [307].

An important advantage of this integrative approach is the provided capacity for topological analysis of the identified significant molecules (e.g. downstream/upstream position in a given pathway, centrality parameters [308], etc.). It is widely accepted that the upstream position of a molecule in a pathway can be considered as a predictive measure for biological significance [309]. In addition, the centrality of a node in a given network, measured by various parameters (e.g.

degree, betweenness, etc.), is a validated implication for distinct importance. Indeed, aberrations in central nodes have been shown to play vital roles in tumor development [310]. Thus, alterations in structure or function (e.g. differential expression/abundance) of a given molecule under certain conditions combined with its topological features can help prioritize candidate molecules (e.g. possible driver molecules) for further studies [306]. Identification of patterns that are unlikely to occur randomly is another important theme in network biology. These patterns include motifs and modules. Motifs are recurring small subgraphs whose interactions form the overall behavior of the complex network. Alterations in these motifs are central to cancer biology and the search for core motifs in cancer-related pathways is valuable for biomarker, therapeutic target, and subtype discovery [311]. Modules are larger subgraphs that are highly connected internally and are involved in specific processes. Modules are extensively investigated for the identification of cancer driver pathways and genes and are explained in more detail in further sections.

Guilt-by-association is another concept widely used for biological inference of topological properties of molecular networks in cancer biology [312]. This notion posits that molecules in topological proximity of each other are potentially functionally related. This is utilized in multiple ways in cancer investigations. For example, proteins of unknown significance in close topological proximity of known drivers of cancer can be investigated as candidate infrequently mutated proteins of functional importance in cancer. Alternatively, proximity as a proxy for overlap in function can be exploited to avoid utilization of redundant molecules for survival analysis, leading to higher efficacy of prognostic biomarkers [312].

Biomarkers and gene signatures identified from network-based approaches have been shown to be more reproducible [313]. In addition, network-based approaches allow the study of perturbations in specific interactions among molecules (e.g. allosteric regulations, post-translational processing, etc.) [307,314]. Deviation of these interactions from normal status is an essential factor in tumorigenesis and cancer progression [141]. Collectively, network-based analysis of cancer has been successfully implemented in cancer driver pathway identification, driver gene discovery,

somatic mutation prioritization, biomarker and therapeutic target discovery, cancer subtyping, and patient stratification [312].

The first step of the network-based analysis of omics data is to construct a context-specific subnetwork from generic data repositories of molecular interactions and pathways [311,315]. These subnetworks represent the parts of the system that are being studied and are constructed based on the experimentally acquired omics datasets. Depending on the goal of the study, different types of networks can be constructed including gene-gene and gene-protein interaction networks, signaling pathways, or a combination of these for a more comprehensive analysis. The most widely used networks are PPI networks [316] and genome-scale metabolic models [317]. The constructed subnetworks can then be amended with the results of a pathway enrichment analysis or can be mined for active module identification (Figure 5). These steps along with the visualization approaches are discussed in more detail below.

3.2.1 Subnetwork construction

Generic databases of biological interactions and pathways are still far from complete [318,319]. However, the goal of these repositories is to capture the entire repertoire of molecular and/or cellular interactions. Meanwhile, depending on the significant molecules identified in the omics dataset under analysis, only a minor subset of these interactions is relevant. Hence, the first step for network-based analysis of datasets is to construct a context-specific subnetwork. In addition to the significant molecules, identified via omics data analysis, subnetworks commonly incorporate all the known molecules that are in direct interaction with them [315]. These extra nodes provide new perspectives for a more comprehensive and accurate network interpretation.

Network-based approaches can greatly facilitate multi-omics data integration and analysis [303]. Multiple levels of omics data produced from different single-layer techniques can be layered upon a single network to achieve a more holistic view of the perturbed system [307]. Alternatively, it is possible to construct multiple networks from different levels of omics data. The comparison of

these networks can provide a deeper and more accurate view of the system under investigation and result in more reliable conclusions [320]. Several algorithms such as AMARETTO [321] and iOmicsPASS [322] facilitate network-based data integration. Interestingly, AMARETTO is able to integrate phenotypic information such as radiography data with multi-omics data. This practical approach has been shown to be effective in identifying candidate cancer driver genes [314].

3.2.2 Module identification

The constructed subnetworks are usually very complex and often referred to as hairballs. While almost impossible to manually identify functional patterns in these subnetworks, graph mining algorithms can be applied to identify functional units of the large subnetwork known as modules [323]. Modules can be regarded as sets of densely connected nodes with an overall limited connection to the rest of the network [324,325]. An important property of biological systems is that molecules with similar functions closely interact with one another and tend to cluster together in biological networks [325]. Hence, each module can be assigned a specific biological function. If a subnetwork is constructed based on differential expression/abundance of molecules under a certain condition, modules in this subnetwork are expected to represent perturbed parts of the system that gave rise to the condition under investigation.

Since the disruption of certain pathways (e.g. apoptosis, proliferation, etc.) is common to almost all cancer types [141], it is logical to consider that genes that harbor driver mutations should at least to some extent cluster together in modules [326,327]. It is expected that modules containing genes that are known to be involved in tumorigenesis and cancer progression can be utilized to predict novel cancer driver genes. Moreover, module identification can facilitate the identification of co-occurring cancer driver mutations [328].

The analysis of network modules facilitates the discovery of common disease mechanisms, disease subtypes, or the mechanics of response to drugs [329]. Interestingly, biological networks are often hierarchically organized, where for example a group of small, interconnected modules can be

clustered together to form larger modules. Researchers can use these hierarchies to adjust the magnification of the analysis for a more biologically relevant interpretation [330]. Methods such as hierarchical Hotnet are specifically developed for cancer studies to identify these module hierarchies and predict cancer driver genes [331].

Commonly used methods for module identification [332–334] first score nodes and edges based on criteria such as differential expression and experimentally validated PPIs, respectively. Then, a scoring system based on the aggregated scores of all the members of a hypothetical module is formulated. An algorithm is used to identify optimal modules (those with the highest scores). In the final step, the identified modules are queried for their statistical significance in relation to the investigated hypothesis [329].

Multiple classes of algorithms have been implemented for module identification, including diffusion-based algorithms and algorithms based on the prize-collecting Steiner tree problem [312]. Briefly, diffusion algorithms consider significant molecules as sources of a phenomenon such as heat diffusion that spreads through the edges of the network until equilibrium is achieved. Here, the goal is to find regions of the network with the most influence over them (i.e. hot regions) as these regions represent highly active modules. Prize-collecting Steiner tree algorithms seek to find modules optimized to contain the highest number of prizes (significant nodes) while minimizing the number of edges. Some algorithms [335] also exploit specific properties of tumors such as mutual exclusivity (i.e. activation/inactivation of a second driver molecule functionally related to an already perturbed molecule is obsolete and rarely observed in a single tumor).

jActiveModules [332] is a widely used plug-in for network visualization software Cytoscape [336]. It can be used for module identification and can determine whether modules are common in multiple states. jActiveModules scores all the nodes in a network based on the p-values from a differential gene expression analysis and has a scoring function to determine the statistical significance of any given module. First, it assigns an active or inactive state to each node in a

947 subnetwork (with a 0.5 probability). Then, for a defined number of iterations, it selects a random
948 node, toggles its state (active/inactive), and recomputes the module's score. If the aggregated score
949 of the module has increased, it keeps the node in its new state. Otherwise, it keeps or changes its
950 state with a defined probability. The process continues until a local optimum is achieved. The
951 identified module might not be the module with the global maximum score, but regardless it is of
952 biological interest.

953 Approaches for module identification are not limited to what has just been described. For example,
954 in [337], the authors proposed a novel module identification pipeline. In this method, gene-gene
955 correlation networks are constructed from omics data from two conditions under comparison. Then,
956 the networks are separately integrated with *a priori* knowledge of interactions to identify modules.
957 Thereafter, enriched modules (e.g. those significantly associated with upregulated genes in a certain
958 condition) can be identified and potentially be used for predictive or diagnostic purposes. A few
959 outstanding challenges regarding the existing methods and the overall approach should be
960 considered. There is a lack of a strong correlation between mRNA and protein abundance [338].
961 As a consequence, utilizing the mRNA profile on its own as the source for subnetwork construction
962 would result in an inaccurate representation of the actual system. iOmicsPASS [322] is a recently
963 developed algorithm that takes this issue into account by integrating transcriptomics and proteomics
964 data. iOmicsPass predicts phenotypic groups based on the joint expression pattern of the nodes
965 within densely connected modules. The algorithm has been shown to be effective for predictive
966 module identification especially when dealing with smaller datasets. Another major problem is that
967 it is possible for a single molecule to be shared among multiple biological modules. Current
968 methods, however, are not computationally effective in identifying overlapping modules [327].
969 Furthermore, despite a considerable rate of development of novel methods, there is a lack of
970 standard benchmarks for validation and comparison of suggested methods [329]. In addition, it
971 should be noted that the assumption that disease-related molecules cluster together in interaction
972 networks does not always hold for a complex condition such as cancer [327].

3.2.3 Pathway enrichment analysis

Pathway enrichment analysis is a common approach for identifying disrupted molecular processes and pathways underlying a certain condition [339]. The central idea is to identify common pathways that a set of molecules (e.g. differentially expressed genes) is associated with. This reduces the contextual complexity of the system and simplifies the interpretation of omics datasets by taking advantage of prior knowledge about biological processes [340].

Three generations of pathway enrichment methods have been developed [340]. The first generation was termed over-representation analysis (ORA). In this generation of methods, a list of significantly differentially expressed molecules, based on p-value and/or fold change filters, is compared against previously compiled functional lists of molecular processes to identify over-represented pathways. DAVID [341] and WebGestalt [342] are among the widely used tools that exploit ORA algorithms. A major drawback of ORA is that by defining filters, we risk the omission of important molecules [343]. Moreover, ORA algorithms treat all the molecules that passed the defined filters as equally significant [304].

The second generation of pathway enrichment methods is known as functional class scoring (FCS). Instead of using predefined filters, FCS algorithms require an input list of all the evaluated molecules, along with values corresponding to their level of differential expression (e.g. fold-change or p-value) [340]. In these methods, all the input molecules are statistically ranked and the overrepresentation of pathways is analyzed with the impact of each molecule in consideration [304]. A pitfall in this approach is that the analysis can become biased towards a few molecules that have been identified as very significant. Gene set enrichment analysis (GSEA) [344] is a widely used algorithm belonging to the second generation of pathway enrichment analysis methods. Genetrail [345] is a popular and freely accessible web service that provides users with both ORA and FCS algorithms for pathway enrichment analysis.

The most recent generation of pathway enrichment methods was developed with the goal to maximize the utilization of prior biological knowledge [346]. This generation of pathway enrichment algorithms incorporates the topological features of nodes in biological networks (e.g. upstream or downstream position in the pathway, degree and betweenness) as additional weighting factors in the enrichment process [309,315]. In addition, in topology-based methods, the analysis is not limited to input molecules but other molecules with close connections to input molecules can be incorporated to identify relevant pathways. Studies indicate that topology-based methods outperform conventional methods (ORA and FCS) both in genomics and metabolomics enrichment analyses [324,347]. This generation of algorithms provides better capacity for the analysis of molecular interactions and understanding the underlying mechanisms of a condition. In general, there is no single best-performing tool for topology-based enrichment analysis. However, a recent comparative study [324] identified DEGraph [348] as the superior method among the nine algorithms investigated.

Overall, some major challenges remain for pathway enrichment analysis. In a recent study [347] Nguyen and co-authors found that all of the tested pathway enrichment methods with the exception of GSEA are prone to report false positives. GSEA on the other hand, suffers from low sensitivity. Furthermore, the Fisher's exact test, while a highly utilized method, performed poorly in this study and produced a significant number of false-positive results. Hence, highly popular platforms such as DAVID, which use this method, should be treated with extra care.

Most comparative studies focus on gene expression data and the results of these studies are not necessarily applicable to other data types (for a list of methods and tools utilized in enrichment analyses and their comparative performance derived from comparative studies refer to Supplementary Table S3). Considering the importance of other layers of information in cancer studies, this should be considered in future developments. One tool that already supports other layers of information, including genomics, transcriptomics, proteomics, miRNAomics, epigenomics, etc. is GeneTrail [345]. In addition, although studies indicate the superiority of

topology-based enrichment methods, it is still not sufficiently recognized. It would be ideal if popular and user-friendly portals of enrichment analysis would incorporate topology-based approaches in order to make these methods accessible to a wider range of researchers.

The current lack of gold standard methods for pathway enrichment analysis coupled with the plethora of existing approaches makes the selection of a suitable method a challenging task. This is especially burdensome for researchers with limited computational expertise. With that being said, there are a number of user-friendly web-based platforms such as MetaboAnalyst [225] and Metascape [349] that offer users a comprehensive pipeline for pathway enrichment analysis. Metascape (https://metascape.org/) takes advantage of multiple databases as its resource for systems-level analysis of datasets. It provides powerful computational abilities with a simplified and user-friendly interface designed for researchers with minimal computational expertise. Since outdated data can severely impact the quality of analysis results [350], an important feature of Metascape is the monthly data synchronization with the updated information in data repositories. The workflow of Metascape can also be modified by users with more advanced computational skills to meet the requirements of individual studies. Moreover, it can be utilized for cross-omics comparisons of multiple gene lists and integrated analyses. Similar to DAVID, the resulting enriched terms in Metascape are clustered and non-redundant. The results can also be exported to Cytoscape for further analysis.

3.2.4 Network visualization

Through visualization, large amounts of data can be made more accessible for convenient pattern detection and interpretation [351]. Whether it is in the form of processed networks or categorized and functional tables, the goal of the visualization process is to reduce the overwhelming complexity of large datasets and make them more readily interpretable. Many tools such as Cytoscape [336], PaintOmics [352], and Omicsnet [353] are developed with the objective of

simplifying the visualization process and offering users a wide array of options to modify how their data is represented.

Cytoscape is a widely used freely accessible platform that provides users with an interactive interface and powerful tools for network visualization and analysis. Cytoscape's feature set can be expanded by adding plug-ins developed by the community for specific computational tasks. Omicsnet [353] is a recently developed web-based visualization tool (www.omicsnet.ca/) that provides users with a 3D structure for visualization and analysis of large networks. It can incorporate multiple heterogeneous datasets in a single subnetwork. Moreover, by taking advantage of various structural layouts such as spherical and multi-layer layouts, it facilitates network analysis and reduces the overwhelming complexity of large networks. In addition, it provides users with a variety of functional and topological analysis tools including module identification and pathway enrichment analysis.

3.2.5 Challenges and perspectives

Although there are numerous methods and tools developed to tackle the variety of problems associated with the network-based analysis of omics data, this approach to data analysis is still in its infancy. Whether it is a matter of reliability of the analysis or a matter of providing equilibrium between the amount of lost data and precision, a number of challenges remain for the community to address.

The quality of network analysis results can only be as good as the quality of the input data. Besides the quality of omics data, a major challenge in this field is incomplete or inaccurate information in network and pathway databases which has been shown to greatly affect the analysis process [350]. Hence, efforts to validate and expand the information in these databases are of essential importance. In addition, analysis tools need to regularly update their knowledge base to keep up with the expansion pace of the source databases. Moreover, limited overlap among interactome databases means that they should be used in combination for more comprehensive results [347].

A simple widespread approach for subnetwork construction is the inference of relevant nodes based on significantly differentially expressed/abundant mRNAs or proteins. However, two caveats should be considered when opting for such approaches. First, since there is evidence against a strong correlation between mRNA and protein levels [272] the accuracy of utilizing mRNA expression levels for subnetwork construction is questionable. Second, phenomena such as somatic mutations, PTMs, and alterations in cellular localization can functionally affect PPIs. These alterations might be overlooked when PPI subnetworks are constructed solely based on mRNA expression or protein abundance. When this is coupled with inaccuracies and incompleteness of current PPI databases, it becomes clear that constructed subnetworks based on differential mRNA expression or protein abundance do not necessarily provide accurate representations of the altered cellular interaction networks. Integrative approaches can ameliorate this flaw to a great extent. For instance, using integrative analysis approaches prior to subnetwork construction, one can establish a list of candidate significant molecules (e.g. genes with both somatic mutation and differential expression, overexpressed genes with hypomethylation, etc.) and subsequently create a subnetwork by mapping these molecules to the human interactome [354]. Alternatives include more sophisticated methods where a list of candidate molecules is not determined *a priori*. For example, in the very recently introduced EMOGI method specifically developed for cancer data exploration [355], novel candidate cancer genes are predicted through a machine-learning approach that uses a generic PPI network with a multi-omics feature vector for each node along with lists of high-confidence cancer/non-cancer genes as input. However, only a limited number of user-friendly tools allow for a network-based multi-omics data analysis. Moreover, current tools that provide the capacity for this type of analysis are not comprehensive with regards to the types of integration they can carry out.

Recently, efforts have been made to systematically compare the plethora of existing methods. These studies analyzed current popular methods from different perspectives, deducing different existing

challenges in the field, from the lack of a uniform distribution of p-values under the null condition for enrichment analyses to the absence of a perfect method for all the study goals [324,347].

An exciting future awaits the network-biology approaches. Single-cell multi-omics technologies provide a highly potent data source for the construction of multilayered networks providing holistic views of individual cellular systems. Moreover, it opens a great opportunity for understanding intratumoral heterogeneity [356]. From the enhanced capability to unravel the complex underlying mechanisms of cancer to drug repurposing [357] and precision medicine [358], network-based approaches facilitate the translation of raw biological data of single-layer omics experiments to practical knowledge and possible interventions.

3.3 Successful implementations of integrative approaches in cancer research

With significant growth during the last decade, high-throughput technologies prompted many studies with results of clinical relevance. The search for molecular markers predictive of the response to specific types of treatment is a hot topic in precision oncology and many studies provide encouraging results. For instance, in a study by Taber et al. [359], sequential analysis of genomics, transcriptomics, and proteomics data resulted in the identification of a subgroup of muscle-invasive bladder cancer patients with high genomic instability and non-basal/squamous expression subtype that were highly responsive to cisplatin-based chemotherapy while patients with low genomic instability and basal/squamous expression subtype showed poor response. In another study, proteogenomics analysis of HPV-negative head and neck squamous cell carcinoma shed light upon multiple clinically significant aspects of this malignancy [360]. In addition to providing insights into the underlying biology of this type of cancer, they identified multiple potentially druggable targets. Interestingly, this study proposed that amplification of EGFR does not necessarily correlate with the prevalence of EGFR ligands, suggesting that the investigation of EGFR ligand abundance is a more appropriate strategy for prediction of response to treatments with anti-EGFR monoclonal antibodies.

The interplay between molecules is best explored through network analysis. In a remarkable pan-cancer network-based integration of genomics and transcriptomics data of 9,738 samples from 20 TCGA cohorts, Paull et al. [361] identified 407 master regulator (MR) proteins responsible for channeling the functional effects of the plethora of genomic aberrations to specific gene expression signatures across tumor types. These proteins were categorized into 24 MR modules, each involved in the regulation of specific hallmarks of cancer. They proposed that based on the status of these 24 modules (activated/inactivated) in each individual, patient-tailored combinations of drugs that specifically target these modules can be administered with precision.

In addition, although in its infancy, single-cell multi-omics is an emerging mighty technology. Perhaps, the most profound contribution of single-cell technologies is that they allow us to dissect intratumoral heterogeneity at individual cell resolution and explore common cancer type- or subtype-specific patterns of heterogeneity among cellular clusters. The delineation of these patterns can enhance our understanding of how tumors with specific origins exhibit certain properties (e.g. metastasis, drug resistance, etc.), yielding insights into their assailable aspects and providing new means for patient stratification [362]. Single-cell multi-omics has the capacity to uncover intratumoral heterogeneity across layers of molecular information and provide us with a systems-level understanding of this phenomenon. Indeed, an integrative study of mRNA and protein levels at single-cell resolution evaluating the effect of BMP4 (a proposed therapeutic agent for glioblastoma [363]) on early-passage glioblastoma cultures [364] identified extensive heterogeneity in how subpopulations of cells respond to BMP4 treatment. Utilizing the mRNA and protein information in complement, they concluded that while all of the treated cells activated the BMP4 pathway, a subset of cells escapes proliferation suppressive effects of BMP4 treatment through a TNC protein-dependent mechanism. Together, such studies illustrate the massive potential of integrative approaches in deepening our understanding of tumor biology and directing clinical efforts towards precise patient stratification and treatment.

4 Conclusion

Current omics technologies and computational advancements provide unprecedented capacity to study cancer etiology and underlying mechanisms, discover clinically applicable diagnostic and predictive biomarkers, identify therapeutic targets, and develop therapeutic interventions. Despite significant progress in the field, various uncharted territories remain to be explored. The fact that no driver mutation could be identified for 5% of the cancers [365] or the unknown exact basis for metastasis [66] highlight the existence of fundamental gaps in our knowledge. Until these fundamental shortcomings in our knowledge persist, our inability to design highly effective therapeutic interventions is not surprising. With the enhancement of our knowledge during the last decades, it is becoming evident that cancer should no longer be viewed as a disease of the genome but should rather be regarded as a disease of the cellular system. Rapid advances in technologies and methodologies are paving the road for more effective study of cellular systems and their perturbations. However, the dispersion of the plethora of bioinformatics tools, the lack of benchmarked gold standard methods, and the required computational skills are major prohibitive elements. There is an ever-growing need for user-friendly workflows that have been adjusted for specific study goals. The extension of current comprehensive platforms such as Galaxy [366] that allow for designing and utilizing readymade workflows for a very wide range of omics experiments will result in further facilitation of data analysis processes.

5 Abbreviations

1H NMR: Proton nuclear magnetic resonance; CASAVA: Consensus Assessment of Sequence And Variation; cfDNA; cell-free DNA; CHIP: Clonal hematopoiesis of indeterminate potential; Co-IP: Co-immunoprecipitation; ctDNA: circulating tumor DNA; DGE: Differential gene expression; ESI: Electrospray ionization; FCS: Functional class scoring; FDR: False discovery rate; GATK: Genome Analysis Toolkit; GC-MS: Gas chromatography-mass spectrometry; GSEA: Gene set enrichment analysis; HLA: Human leukocyte antigen; HMDB: Human Metabolome Database; HPPP: Human Plasma Proteome Project; LC-MS/MS: Liquid chromatography-tandem mass spectrometry; MALDI-MSI: Matrix-assisted laser desorption ionization-based mass spectrometry imaging; MR: Master regulator; NGS: Next-generation sequencing; ORA: Over-representation analysis; PCAWG: Pan-Cancer Analysis of Whole-Genomes; PPI: Protein-protein interaction; PRM: Parallel reaction monitoring; PTM: Post-translational modification; SRM: Selected reaction monitoring; TAP: Tandem affinity purification; TCGA: The Cancer Genome Atlas; TMT: Tandem mass tag; UMI: Unique molecular identifiers; VAF: Variant allele frequency; VSN: Variance

Stabilization Normalization; WES: Whole-exome sequencing; WGS: Whole-genome sequencing; XL-MS: Crosslinking-mass spectrometry; lncRNA: long non-coding RNA; miRNA: microRNA; ncRNA: non-coding RNA; sncRNA: small non-coding RNA

6 Acknowledgments

We acknowledge Vice Chancellor for Research and Technology of the Semnan University.
Figures were created with BioRender.com.

7 Conflicts of interest

The authors declare no conflicts of interest

8 Funding

U.S. was supported via an Early Career Researcher Fellowship from Cancer Institute New South Wales and is now supported by the National Health and Medical Research Council (Fellowship #1196405). Financial support was also provided to U.S. by the Cancer Council NSW (project grant RG20-12).

9 References

[1] Le Tourneau C, Delord JP, Gonçalves A, Gavoille C, Dubot C, Isambert N, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): A multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol 2015;16:1324–34. https://doi.org/10.1016/S1470-2045(15)00188-6.

[2] Massard C, Michiels S, Ferté C, Le Deley MC, Lacroix L, Hollebecque A, et al. High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: Results of the MOSCATO 01 trial. Cancer Discov 2017;7:586–95. https://doi.org/10.1158/2159-8290.CD-16-1396.

[3] Tannock IF, Hickman JA. Limits to personalized cancer medicine. N Engl J Med 2016. https://doi.org/10.1056/NEJMsb1607705.

[4] Rodon J, Soria JC, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med 2019;25:751–8. https://doi.org/10.1038/s41591-019-0424-4.

[5] Du W, Elemento O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene 2015;34:3215–25. https://doi.org/10.1038/onc.2014.291.

[6] Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Brief Bioinform 2018;19:286–302. https://doi.org/10.1093/BIB/BBW114.

[7] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science (80-) 2013;340:1546–58. https://doi.org/10.1126/science.1235122.

[8] Panou V, Gadiraju M, Wolin A, Weipert CM, Skarda E, Husain AN, et al. Frequency of germline mutations in cancer susceptibility genes in malignant mesothelioma. J Clin Oncol 2018;36:2863–71. https://doi.org/10.1200/JCO.2018.78.5204.

[9] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21. https://doi.org/10.1038/nature12477.

[10] Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578:94–101. https://doi.org/10.1038/s41586-020-1943-3.

[11] Ma J, Setton J, Lee NY, Riaz N, Powell SN. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. Nat Commun 2018;9:1–12. https://doi.org/10.1038/s41467-018-05228-y.

[12] Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature 2010;464:993–8. https://doi.org/10.1038/nature08987.

[13] McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, et al. Comprehensive genomic characterization defines human glioblastoma genes and core

pathways. Nature 2008;455:1061–8. https://doi.org/10.1038/nature07385.

[14] Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? Hum Genet 2016;135:359–62. https://doi.org/10.1007/s00439-015-1631-9.

[15] Diamantopoulos MA, Tsiakanikas P, Scorilas A. Non-coding RNAs: the riddle of the transcriptome and their perspectives in cancer. Ann Transl Med 2018;6:241–241. https://doi.org/10.21037/atm.2018.06.10.

[16] Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature 2020. https://doi.org/10.1038/s41586-020-1969-6.

[17] Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci 2018;109:513–22. https://doi.org/10.1111/cas.13505.

[18] Itamochi H, Oishi T, Oumi N, Takeuchi S, Yoshihara K, Mikami M, et al. Whole-genome sequencing revealed novel prognostic biomarkers and promising targets for therapy of ovarian clear cell carcinoma. Br J Cancer 2017;117:717–24. https://doi.org/10.1038/bjc.2017.228.

[19] Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. Nature 2018;560:325–30. https://doi.org/10.1038/s41586-018-0409-3.

[20] Kamps R, Brandão RD, van den Bosch BJ, Paulussen ADC, Xanthoulea S, Blok MJ, et al. Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. Int J Mol Sci 2017;18. https://doi.org/10.3390/ijms18020308.

[21] Rossing M, Sørensen CS, Ejlertsen B, Nielsen FC. Whole genome sequencing of breast cancer. Apmis 2019;127:303–15. https://doi.org/10.1111/apm.12920.

[22] Van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: Tone down the bias. Exp Cell Res 2014;322:12–20. https://doi.org/10.1016/j.yexcr.2014.01.008.

[23] Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med 2018;10:eaat4921. https://doi.org/10.1126/scitranslmed.aat4921.

[24] Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 2015;6. https://doi.org/10.1038/ncomms10001.

[25] Yamaguchi I, Watanabe T, Ohara O, Hasegawa Y. PCR-free whole exome sequencing: Costeffective and efficient in detecting rare mutations. PLoS One 2019;14:1–9. https://doi.org/10.1371/journal.pone.0222562.

[26] Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.

[27] Arora K, Shah M, Johnson M, Sanghvi R, Shelton J, Nagulapalli K, et al. Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. Sci Rep 2019;9:1–13. https://doi.org/10.1038/s41598-019-55636-3.

[28] Li H. [Heng Li - Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Prepr ArXiv 2013. https://doi.org/arXiv:1303.3997 [q-bio.GN].

[29] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012. https://doi.org/10.1038/nmeth.1923.

[30] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics 2015. https://doi.org/10.1093/bioinformatics/btv098.

[31] Faust GG, Hall IM. SAMBLASTER: Fast duplicate marking and structural variant read extraction. Bioinformatics, 2014. https://doi.org/10.1093/bioinformatics/btu314.

[32] Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2 2019. https://doi.org/10.1101/861054.

[33] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76. https://doi.org/10.1101/gr.129684.111.

[34] Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 2018;15:591–4. https://doi.org/10.1038/s41592-018-0051-x.

[35] Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012;28:311–7. https://doi.org/10.1093/bioinformatics/btr665.

[36] Cai L, Yuan W, Zhang Z, He L, Chou K-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep 2016;6:36540. https://doi.org/10.1038/srep36540.

[37] Chen Z, Yuan Y, Chen X, Chen J, Lin S, Li X, et al. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. Sci Rep 2020;10:3501. https://doi.org/10.1038/s41598-020-60559-5.

[38] O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med 2013;5. https://doi.org/10.1186/gm432.

[39] Hwang KB, Lee IH, Li H, Won DG, Hernandez-Ferrer C, Negron JA, et al. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Sci Rep 2019;9:1–10. https://doi.org/10.1038/s41598-019-39108-2.

[40] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gky1015.

[41] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:980–5. https://doi.org/10.1093/nar/gkt1113.

[42] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.

Nucleic Acids Res 2005. https://doi.org/10.1093/nar/gki033.

[43] Schwarz JM, Cooper DN, Schuelke M, Seelow D. Mutationtaster2: Mutation prediction for the deep-sequencing age. Nat Methods 2014;11:361–2. https://doi.org/10.1038/nmeth.2890.

[44] Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc 2016;11:1–9. https://doi.org/10.1038/nprot.2015.123.

[45] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9. https://doi.org/10.1038/nmeth0410-248.

[46] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:1–14. https://doi.org/10.1186/s13059-016-0974-4.

[47] Karimi AH, Karimi MR, Farnia P, Parvini F, Beheshti S, Parvini F, et al. A novel homozygous truncating mutation in NALCN causing IHPRF1 : detailed clinical manifestations and a review of literature 2020:1–10.

[48] Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. Cancer Res 2009;69:6660–7. https://doi.org/10.1158/0008-5472.CAN-09-1133.

[49] Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018;50:1161–70. https://doi.org/10.1038/s41588-018-0167-z.

[50] Chen H, Li J, Wang Y, Ng PK-S, Tsang YH, Shaw KR, et al. Comprehensive assessment of computational algorithms in predicting cancer driver mutations. Genome Biol 2020;21:43. https://doi.org/10.1186/s13059-020-01954-z.

[51] Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018;173:371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060.

[52] Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res 2017;45:W201–6. https://doi.org/10.1093/nar/gkx390.

[53] Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J, et al. Next-generation sequencing informatics: Challenges and strategies for implementation in a clinical environment. Arch Pathol Lab Med 2016;140:958–75. https://doi.org/10.5858/arpa.2015-0507-RA.

[54] do Valle ÍF, Giampieri E, Simonetti G, Padella A, Manfrini M, Ferrari A, et al. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinformatics 2016;17. https://doi.org/10.1186/s12859-016-1190-7.

[55] Nystedt B, Garcia M, Juhos S, Larsson M, Olason PI, Martin M, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants.

F1000Research 2020. https://doi.org/10.12688/f1000research.16665.1.

[56] Strom SP. Current practices and guidelines for clinical next-generation sequencing oncology testing. Cancer Biol Med 2016;13:3–11. https://doi.org/10.28092/j.issn.2095-3941.2016.0004.

[57] Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer 2017;17:557–69. https://doi.org/10.1038/nrc.2017.58.

[58] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: Current state of the science. Nat Rev Genet 2016;17:175–88. https://doi.org/10.1038/nrg.2015.16.

[59] Ren X, Kang B, Zhang Z. Understanding tumor ecosystems by single-cell sequencing: Promises and limitations 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis 06 Biological Sciences 0604 Genetics. Genome Biol 2018;19:1–14. https://doi.org/10.1186/s13059-018-1593-z.

[60] Tritschler S, Büttner M, Fischer DS, Lange M, Bergen V, Lickert H, et al. Concepts and limitations for learning developmental trajectories from single cell genomics. Dev 2019;146. https://doi.org/10.1242/dev.170506.

[61] Low S-K, Zembutsu H, Nakamura Y. Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Sci 2018;109:497–506. https://doi.org/10.1111/cas.13463.

[62] Alix-Panabières C, Pantel K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. Cancer Discov 2016;6:479–91. https://doi.org/10.1158/2159-8290.CD-15-1483.

[63] Corcoran RB, Chabner BA. Application of Cell-free DNA Analysis to Cancer Treatment. N Engl J Med 2018;379:1754–65. https://doi.org/10.1056/NEJMra1706174.

[64] Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 2016;34:547–55. https://doi.org/10.1038/nbt.3520.

[65] Weng J, Atyah M, Zhou C, Ren N. Prospects and challenges of circulating tumor DNA in precision medicine of hepatocellular carcinoma. Clin Exp Med 2020;20:329–37. https://doi.org/10.1007/s10238-020-00620-9.

[66] Lambert AW, Pattabiraman DR, Weinberg RA. Emerging Biological Principles of Metastasis. Cell 2017;168:670–91. https://doi.org/10.1016/j.cell.2016.11.037.

[67] Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. J Hum Genet 2020;65:3–10. https://doi.org/10.1038/s10038-019-0658-5.

[68] Crick F. Central dogma of molecular biology. Nature 1970. https://doi.org/10.1038/227561a0.

[69] Baylin SB, Jones PA. A decade of exploring the cancer epigenome-biological and translational implications. Nat Rev Cancer 2011. https://doi.org/10.1038/nrc3130.

[70] Thomson JM, Newman M, Parker JS, Morin-Kensicki EM, Wright T, Hammond SM. Extensive post-transcriptional regulation of microRNAs and its implications for cancer.

Genes Dev 2006;20:2202–7. https://doi.org/10.1101/gad.1444406.

[71] Lian H, Wang QH, Zhu C Bin, Ma J, Jin WL. Deciphering the Epitranscriptome in Cancer. Trends in Cancer 2018;4:207–21. https://doi.org/10.1016/j.trecan.2018.01.006.

[72] Han ZJ, Feng YH, Gu BH, Li YM, Chen H. The post-Translational modification, SUMOylation, and cancer (Review). Int J Oncol 2018;52:1081–94. https://doi.org/10.3892/ijo.2018.4280.

[73] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science (80-) 2017;357. https://doi.org/10.1126/science.aan2507.

[74] Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and experiments. Mol Syst Biol 2009;5:1–8. https://doi.org/10.1038/msb.2009.66.

[75] Bartoschek M, Oskolkov N, Bocci M, Lövrot J, Larsson C, Sommarin M, et al. Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing. Nat Commun 2018;9. https://doi.org/10.1038/s41467-018-07582-3.

[76] Zhu M, Dang Y, Yang Z, Liu Y, Zhang L, Xu Y, et al. Comprehensive RNA Sequencing in Adenoma-Cancer Transition Identified Predictive Biomarkers and Therapeutic Targets of Human CRC. Mol Ther - Nucleic Acids 2020;20:25–33. https://doi.org/10.1016/j.omtn.2020.01.031.

[77] Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet 2018;19:93–109. https://doi.org/10.1038/nrg.2017.96.

[78] Davis RT, Blake K, Ma D, Gabra MBI, Hernandez GA, Phung AT, et al. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. Nat Cell Biol 2020;22:310–20. https://doi.org/10.1038/s41556-020-0477-0.

[79] Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, et al. RNA sequencing of cancer reveals novel splicing alterations. Sci Rep 2013;3. https://doi.org/10.1038/srep01689.

[80] Oltean S, Bates DO. Hallmarks of alternative splicing in cancer. Oncogene 2014;33:5311–8. https://doi.org/10.1038/onc.2013.533.

[81] Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: From SNPs to phenotypes. Trends Genet 2011;27:72–9. https://doi.org/10.1016/j.tig.2010.10.006.

[82] Brouard JS, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. J Anim Sci Biotechnol 2019;10:1–6. https://doi.org/10.1186/s40104-019-0359-0.

[83] Ozsolak F, Milos PM. RNA sequencing: Advances, challenges and opportunities. Nat Rev Genet 2011;12:87–98. https://doi.org/10.1038/nrg2934.

[84] Schmitz U, Naderi-Meshkin H, Gupta SK, Wolkenhauer O, Vera J. The RNA world in the 21st century-a systems approach to finding non-coding keys to clinical questions. Brief Bioinform 2016. https://doi.org/10.1093/bib/bbv061.

[85] Chan JJ, Tay Y. Noncoding RNA: RNA regulatory networks in cancer. Int J Mol Sci 2018;19. https://doi.org/10.3390/ijms19051310.

[86] Cech TR, Steitz JA. The noncoding RNA revolution - Trashing old rules to forge new ones. Cell 2014;157:77–94. https://doi.org/10.1016/j.cell.2014.03.008.

[87] Schmitz U, Wolkenhauer O, Julio Vera, Zinovyev A, Morozova N, Gorban AN, et al. MicroRNA Cancer Regulation. Adv Exp Med Biol 2013.

[88] Farazi TA, Spitzer JI, Morozov P, Tuschl T. MiRNAs in human cancer. J Pathol 2011. https://doi.org/10.1002/path.2806.

[89] Kosaka N, Iguchi H, Ochiya T. Circulating microRNA in body fluid: A new potential biomarker for cancer diagnosis and prognosis. Cancer Sci 2010;101:2087–92. https://doi.org/10.1111/j.1349-7006.2010.01650.x.

[90] Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: Biomarkers, functions and therapy. Trends Mol Med 2014;20:460–9. https://doi.org/10.1016/j.molmed.2014.06.005.

[91] Bhattacharya A, Schmitz U, Wolkenhauer O, Schönherr M, Raatz Y, Kunz M. Regulation of cell cycle checkpoint kinase WEE1 by miR-195 in malignant melanoma. Oncogene 2013. https://doi.org/10.1038/onc.2012.324.

[92] Amirkhah R, Schmitz U, Linnebacher M, Wolkenhauer O, Farazmand A. MicroRNA-mRNA interactions in colorectal cancer and their role in tumor progression. Genes Chromosom Cancer 2015. https://doi.org/10.1002/gcc.22231.

[93] Bhan A, Soleimani M, Mandal SS. Long noncoding RNA and cancer: A new paradigm. Cancer Res 2017;77:3965–81. https://doi.org/10.1158/0008-5472.CAN-16-2634.

[94] Naderi-Meshkin H, Lai X, Amirkhah R, Vera J, Rasko JEJ, Schmitz U. Exosomal lncRNAs and cancer: Connecting the missing links. Bioinformatics 2019. https://doi.org/10.1093/bioinformatics/bty527.

[95] Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. Cancer Cell 2016;29:452–63. https://doi.org/10.1016/j.ccell.2016.03.010.

[96] Yang G, Lu X, Yuan L. LncRNA: A link between RNA and cancer. Biochim Biophys Acta - Gene Regul Mech 2014;1839:1097–109. https://doi.org/10.1016/j.bbagrm.2014.08.012.

[97] Peng WX, Koirala P, Mo YY. LncRNA-mediated regulation of cell signaling in cancer. Oncogene 2017;36:5661–7. https://doi.org/10.1038/onc.2017.184.

[98] Yamada A, Yu P, Lin W, Okugawa Y, Boland CR, Goel A. A RNA-Sequencing approach for the identification of novel long non-coding RNA biomarkers in colorectal cancer. Sci Rep 2018;8:2–11. https://doi.org/10.1038/s41598-017-18407-6.

[99] Jiang M-C, Ni J-J, Cui W-Y, Wang B-Y, Zhuo W. Emerging roles of lncRNA in cancer and therapeutic opportunities. Am J Cancer Res 2019;9:1354–66.

[100] Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet 2019;20:631–56. https://doi.org/10.1038/s41576-019-0150-2.

[101] Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, et al. Comparison of stranded and non-

stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics 2015;16:1–14. https://doi.org/10.1186/s12864-015-1876-7.

[102] Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 2017;8. https://doi.org/10.1002/wrna.1364.

[103] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 2010;7:709–15. https://doi.org/10.1038/nmeth.1491.

[104] 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;17:1–19. https://doi.org/10.1186/s13059-016-0881-8.

[105] Vanichkina DP, Schmitz U, Wong JJL, Rasko JEJ. Challenges in defining the role of intron retention in normal biology and disease. Semin Cell Dev Biol 2018. https://doi.org/10.1016/j.semcdb.2017.07.030.

[106] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8. https://doi.org/10.1038/nmeth.1226.

[107] Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 2012;9:72–4. https://doi.org/10.1038/nmeth.1778.

[108] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013. https://doi.org/10.1093/bioinformatics/bts635.

[109] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.

[110] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013. https://doi.org/10.1186/gb-2013-14-4-r36.

[111] Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–9. https://doi.org/10.1093/bioinformatics/btu638.

[112] Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 2014;32:462–4. https://doi.org/10.1038/nbt.2862.

[113] 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. https://doi.org/10.1038/nmeth.4197.

[114] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016;34:525–7. https://doi.org/10.1038/nbt.3519.

[115] Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol 2014;32:888–95.

https://doi.org/10.1038/nbt.3000.

[116] Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods 2019;16:43–9. https://doi.org/10.1038/s41592-018-0254-1.

[117] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 2013;14:671–83. https://doi.org/10.1093/bib/bbs046.

[118] Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One 2017;12:1–18. https://doi.org/10.1371/journal.pone.0190152.

[119] Tarazona S, García F, Ferrer A, Dopazo J, Conesa A. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. EMBnetJournal 2012. https://doi.org/10.14806/ej.17.b.265.

[120] Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014;15:1–17. https://doi.org/10.1186/gb-2014-15-2-r29.

[121] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:1–21. https://doi.org/10.1186/s13059-014-0550-8.

[122] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 2013;16:59–70. https://doi.org/10.1093/bib/bbt086.

[123] Sheng Q, Vickers K, Zhao S, Wang J, Samuels DC, Koues O, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. Brief Funct Genomics 2017;16:194–204. https://doi.org/10.1093/bfgp/elw035.

[124] Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0030619.

[125] Wang L, Wang S, Li W. RSeQC: Quality control of RNA-seq experiments. Bioinformatics 2012;28:2184–5. https://doi.org/10.1093/bioinformatics/bts356.

[126] Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 2016;32:292–4. https://doi.org/10.1093/bioinformatics/btv566.

[127] Jimenez-Jacinto V, Sanchez-Flores A, Vega-Alvarado L. Integrative Differential expression analysis for multiple experiments (IDEAMEX): A web server tool for integrated RNA-seq data analysis. Front Genet 2019;10:1–16. https://doi.org/10.3389/fgene.2019.00279.

[128] Sun S, Xu L, Zou Q, Wang G. BP4RNAseq: a babysitter package for retrospective and newly generated RNA-seq data analyses using both alignment-based and alignment-free quantification method. Bioinformatics 2020. https://doi.org/10.1093/bioinformatics/btaa832.

[129] Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun 2017;8:1–14. https://doi.org/10.1038/s41467-017-00050-4.

[130] Zhu S, Qing T, Zheng Y, Jin L, Shi L. Advances in single-cell RNA sequencing and its applications in cancer research. Oncotarget 2017. https://doi.org/10.18632/oncotarget.17893.

[131] Suvà ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell 2019;75:7–12. https://doi.org/10.1016/j.molcel.2019.05.003.

[132] Sheng K, Cao W, Niu Y, Deng Q, Zong C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat Methods 2017;14:267–70. https://doi.org/10.1038/nmeth.4145.

[133] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 2019;15. https://doi.org/10.15252/msb.20188746.

[134] Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet 2019;10:1–13. https://doi.org/10.3389/fgene.2019.00317.

[135] Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods 2020. https://doi.org/10.1038/s41592-019-0654-x.

[136] Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol 2020. https://doi.org/10.1186/s13059-020-1949-z.

[137] Antipov D, Korobeynikov A, McLean JS, Pevzner PA. HybridSPAdes: An algorithm for hybrid assembly of short and long reads. Bioinformatics 2016;32:1009–15. https://doi.org/10.1093/bioinformatics/btv688.

[138] Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, et al. Extending rnaSPAdes functionality for hybrid transcriptome assembly. BMC Bioinformatics 2020;21:1–12. https://doi.org/10.1186/s12859-020-03614-2.

[139] Hanash S. Disease proteomics. Nature 2003. https://doi.org/10.1038/nature01514.

[140] Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 2016;165:535–50. https://doi.org/10.1016/j.cell.2016.03.014.

[141] Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell 2011;144:646–74. https://doi.org/10.1016/j.cell.2011.02.013.

[142] Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem 2011;80:273–99. https://doi.org/10.1146/annurev-biochem-061308-093216.

[143] Zou X, Blank M. Targeting p38 MAP kinase signaling in cancer through post-translational modifications. Cancer Lett 2017;384:19–26. https://doi.org/10.1016/j.canlet.2016.10.008.

[144] Zerdes I, Matikas A, Bergh J, Rassidakis GZ, Foukakis T. Genetic, transcriptional and post-translational regulation of the programmed death protein ligand 1 in cancer: biology and clinical correlations. Oncogene 2018;37:4639–61. https://doi.org/10.1038/s41388-018-0303-3.

[145] Glozak MA, Seto E. Histone deacetylases and cancer. Oncogene 2007;26:5420–32. https://doi.org/10.1038/sj.onc.1210610.

[146] Bode AM, Dong Z. Post-translational modification of p53 in tumorigenesis. Nat Rev Cancer 2004;4:793–805. https://doi.org/10.1038/nrc1455.

[147] Macklin A, Khan S, Kislinger T. Recent advances in mass spectrometry based clinical proteomics: Applications to cancer research. Clin Proteomics 2020;17:1–25. https://doi.org/10.1186/s12014-020-09283-w.

[148] Njoku K, Chiasserini D, Whetton AD, Crosbie EJ. Proteomic biomarkers for the detection of endometrial cancer. Cancers (Basel) 2019;11:1–25. https://doi.org/10.3390/cancers11101572.

[149] Marx V. Targeted proteomics. Nat Methods 2013;10:19–22. https://doi.org/10.1038/nmeth.2285.

[150] Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem 2016;9:499–519. https://doi.org/10.1146/annurev-anchem-071015-041550.

[151] MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, et al. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 2010;26:966–8. https://doi.org/10.1093/bioinformatics/btq054.

[152] Cai W, Guner H, Gregorich ZR, Chen AJ, Ayaz-Guner S, Peng Y, et al. MASH suite pro: A comprehensive software tool for top-down proteomics. Mol Cell Proteomics 2016;15:703–14. https://doi.org/10.1074/mcp.O115.054387.

[153] Fellers RT, Greer JB, Early BP, Yu X, Leduc RD, Kelleher NL, et al. ProSight Lite: Graphical software to analyze top-down mass spectrometry data. Proteomics 2015;15:1235–8. https://doi.org/10.1002/pmic.201400313.

[154] Gillet LC, Leitner A, Aebersold R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. Annu Rev Anal Chem 2016;9:449–72. https://doi.org/10.1146/annurev-anchem-071015-041535.

[155] Mertins P, Tang LC, Krug K, Clark DJ, Gritsenko MA, Chen L, et al. Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. Nat Protoc 2018;13:1632–61. https://doi.org/10.1038/s41596-018-0006-9.

[156] Gilar M, Olivova P, Daly AE, Gebler JC. Orthogonality of separation in two-dimensional liquid chromatography. Anal Chem 2005. https://doi.org/10.1021/ac050923i.

[157] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3:1154–69. https://doi.org/10.1074/mcp.M400129-

MCP200.

[158] Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 2003;75:1895–904. https://doi.org/10.1021/ac0262560.

[159] Zhang L, Elias JE. Relative protein quantification using tandem mass tag mass spectrometry. Methods Mol. Biol., 2017. https://doi.org/10.1007/978-1-4939-6747-6_14.

[160] Muth T, Renard BY. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief Bioinform 2018;19:954–70. https://doi.org/10.1093/bib/bbx033.

[161] Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gky1049.

[162] Sinitcyn P, Rudolph JD, Cox J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. Annu Rev Biomed Data Sci 2018;1:207–34. https://doi.org/10.1146/annurev-biodatasci-080917-013516.

[163] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4:207–14. https://doi.org/10.1038/nmeth1019.

[164] Sharma K, D'Souza RCJ, Tyanova S, Schaab C, Wiśniewski JR, Cox J, et al. Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. Cell Rep 2014;8:1583–94. https://doi.org/10.1016/j.celrep.2014.07.036.

[165] Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K, Darie CC. Protein–protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. Cell Mol Life Sci 2014;71:205–28. https://doi.org/10.1007/s00018-013-1333-1.

[166] Adelmant G, Garg BK, Tavares M, Card JD, Marto JA. Tandem Affinity Purification and Mass Spectrometry (TAP-MS) for the Analysis of Protein Complexes. Curr Protoc Protein Sci 2019;96. https://doi.org/10.1002/cpps.84.

[167] Tang X, Wippel HH, Chavez JD, Bruce JE. Crosslinking mass spectrometry: A link between structural biology and systems biology. Protein Sci 2021;30:773–84. https://doi.org/10.1002/pro.4045.

[168] Wang Y, Zhang J, Li B, He QY. Advances of Proteomics in Novel PTM Discovery: Applications in Cancer Therapy. Small Methods 2019;3:1–12. https://doi.org/10.1002/smtd.201900041.

[169] Virág D, Dalmadi-Kiss B, Vékey K, Drahos L, Klebovich I, Antal I, et al. Current Trends in the Analysis of Post-translational Modifications. Chromatographia 2020;83:1–10. https://doi.org/10.1007/s10337-019-03796-9.

[170] Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 2016;11:2301–19. https://doi.org/10.1038/nprot.2016.136.

[171] Tyanova S, Cox J. Perseus: A bioinformatics platform for integrative analysis of

proteomics data in cancer research. Methods Mol Biol 2018;1711:133–48. https://doi.org/10.1007/978-1-4939-7493-1_7.

[172] Humphrey SJ, Karayel O, James DE, Mann M. High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. Nat Protoc 2018;13:1897–916. https://doi.org/10.1038/s41596-018-0014-9.

[173] Liu F, Rijkers DTS, Post H, Heck AJR. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat Methods 2015;12:1179–84. https://doi.org/10.1038/nmeth.3603.

[174] da Veiga Leprevost F, Haynes SE, Avtonomov DM, Chang HY, Shanmugam AK, Mellacheruvu D, et al. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nat Methods 2020. https://doi.org/10.1038/s41592-020-0912-y.

[175] Faria SS, Morris CFM, Silva AR, Fonseca MP, Forget P, Castro MS, et al. A Timely shift from shotgun to targeted proteomics and how it can be groundbreaking for cancer research. Front Oncol 2017;7. https://doi.org/10.3389/fonc.2017.00013.

[176] Uzozie AC, Aebersold R. Advancing translational research and precision medicine with targeted proteomics. J Proteomics 2018;189:1–10. https://doi.org/10.1016/j.jprot.2018.02.021.

[177] McCain JSP, Bertrand EM. Prediction and Consequences of Cofragmentation in Metaproteomics. J Proteome Res 2019;18:3555–66. https://doi.org/10.1021/acs.jproteome.9b00144.

[178] McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem 2014;86:7150–8. https://doi.org/10.1021/ac502040v.

[179] Doll S, Gnad F, Mann M. The Case for Proteomics and Phospho-Proteomics in Personalized Cancer Medicine. Proteomics - Clin Appl 2019;13. https://doi.org/10.1002/prca.201800113.

[180] Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based <scp>SWATH</scp> - <scp>MS</scp> for quantitative proteomics: a tutorial. Mol Syst Biol 2018;14. https://doi.org/10.15252/msb.20178126.

[181] Huang L, Shao D, Wang Y, Cui X, Li Y, Chen Q, et al. Human body-fluid proteome: Quantitative profiling and computational prediction. Brief Bioinform 2021;22:315–33. https://doi.org/10.1093/bib/bbz160.

[182] Csősz É, Kalló G, Márkus B, Deák E, Csutak A, Tőzsér J. Quantitative body fluid proteomics in medicine — A focus on minimal invasiveness. J Proteomics 2017;153:30–43. https://doi.org/10.1016/j.jprot.2016.08.009.

[183] Hu S, Loo JA, Wong DT. Human body fluid proteome analysis. Proteomics 2006;6:6326–53. https://doi.org/10.1002/pmic.200600284.

[184] Kalogeropoulos K, Bundgaard L, auf dem Keller U. Proteomic and Degradomic Analysis of Body Fluids: Applications, Challenges and Considerations, 2020, p. 157–82. https://doi.org/10.1007/978-3-030-58330-9_8.

[185] Schwenk JM, Omenn GS, Sun Z, Campbell DS, Baker MS, Overall CM, et al. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. J Proteome Res 2017;16:4299–310. https://doi.org/10.1021/acs.jproteome.7b00467.

[186] Zhao M, Yang Y, Guo Z, Shao C, Sun H, Zhang Y, et al. A Comparative Proteomics Analysis of Five Body Fluids: Plasma, Urine, Cerebrospinal Fluid, Amniotic Fluid, and Saliva. PROTEOMICS - Clin Appl 2018;12:1800008. https://doi.org/10.1002/prca.201800008.

[187] Liu Y, Chen X, Zhang Y, Liu J. Advancing single-cell proteomics and metabolomics with microfluidic technologies. Analyst 2019;144:846–58. https://doi.org/10.1039/c8an01503a.

[188] Marx V. A dream of single-cell proteomics. Nat Methods 2019;16:809–12. https://doi.org/10.1038/s41592-019-0540-6.

[189] Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. Cell Metab 2016;23:27–47. https://doi.org/10.1016/j.cmet.2015.12.006.

[190] Ren S, Hinzman AA, Kang EL, Szczesniak RD, Lu LJ. Computational and statistical analysis of metabolomics data. Metabolomics 2015;11:1492–513. https://doi.org/10.1007/s11306-015-0823-6.

[191] Vazquez A, Kamphorst JJ, Markert EK, Schug ZT, Tardito S, Gottlieb E. Cancer metabolism at a glance. J Cell Sci 2016;129:3367–73. https://doi.org/10.1242/jcs.181016.

[192] Mulcahy Levy JM, Thorburn A. Autophagy in cancer: moving from understanding mechanism to improving therapy responses in patients. Cell Death Differ 2020. https://doi.org/10.1038/s41418-019-0474-7.

[193] Zhang Y, Commisso C. Macropinocytosis in Cancer: A Complex Signaling Network. Trends in Cancer 2019. https://doi.org/10.1016/j.trecan.2019.04.002.

[194] Hamann JC, Surcel A, Chen R, Teragawa C, Albeck JG, Robinson DN, et al. Entosis Is Induced by Glucose Starvation. Cell Rep 2017. https://doi.org/10.1016/j.celrep.2017.06.037.

[195] Krajcovic M, Krishna S, Akkari L, Joyce JA, Overholtzer M. MTOR regulates phagosome and entotic vacuole fission. Mol Biol Cell 2013. https://doi.org/10.1091/mbc.E13-07-0408.

[196] Nieman KM, Kenny HA, Penicka C V., Ladanyi A, Buell-Gutbrod R, Zillhardt MR, et al. Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth. Nat Med 2011. https://doi.org/10.1038/nm.2492.

[197] Luengo A, Gui DY, Vander Heiden MG. Targeting Metabolism for Cancer Therapy. Cell Chem Biol 2017. https://doi.org/10.1016/j.chembiol.2017.08.028.

[198] Hay N. Reprogramming glucose metabolism in cancer: Can it be exploited for cancer therapy? Nat Rev Cancer 2016. https://doi.org/10.1038/nrc.2016.77.

[199] Yong C, Stewart GD, Frezza C. Oncometabolites in renal cancer. Nat Rev Nephrol 2020;16:156–72. https://doi.org/10.1038/s41581-019-0210-z.

[200] Shulaev V. Metabolomics technology and bioinformatics. Brief Bioinform 2006;7:128–39.

https://doi.org/10.1093/bib/bbl012.

[201] Azad RK, Shulaev V. Metabolomics technology and bioinformatics for precision medicine. Brief Bioinform 2019;20:1957–71. https://doi.org/10.1093/bib/bbx170.

[202] Smolinska A, Blanchet L, Buydens LMC, Wijmenga SS. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. Anal Chim Acta 2012. https://doi.org/10.1016/j.aca.2012.05.049.

[203] Armitage EG, Ciborowski M. Applications of metabolomics in cancer studies. Adv. Exp. Med. Biol., 2017. https://doi.org/10.1007/978-3-319-47656-8_9.

[204] Nagana Gowda GA, Djukovic D. Overview of mass spectrometry-based metabolomics: Opportunities and challenges. Methods Mol Biol 2014. https://doi.org/10.1007/978-1-4939-1258-2_1.

[205] Theodoridis GA, Gika HG, Want EJ, Wilson ID. Liquid chromatography-mass spectrometry based global metabolite profiling: A review. Anal Chim Acta 2012;711:7–16. https://doi.org/10.1016/j.aca.2011.09.042.

[206] Lee MY, Hu T. Computational methods for the discovery of metabolic markers of complex traits. Metabolites 2019;9. https://doi.org/10.3390/metabo9040066.

[207] Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-Mcintyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 2011;6:1060–83. https://doi.org/10.1038/nprot.2011.335.

[208] Kell DB, Brown M, Davey HM, Dunn WB, Spasic I, Oliver SG. Metabolic footprinting and systems biology: The medium is the message. Nat Rev Microbiol 2005;3:557–65. https://doi.org/10.1038/nrmicro1177.

[209] SIUZDAK G. An introduction to mass spectrometry ionization: An excerpt from The Expanding Role of Mass Spectrometry in Biotechnology, 2nd ed.; MCC Press: San Diego, 2005. J Assoc Lab Autom 2004. https://doi.org/10.1016/j.jala.2004.01.004.

[210] Spicer R, Salek RM, Moreno P, Cañueto D, Steinbeck C. Navigating freely-available software tools for metabolomics analysis. Metabolomics 2017;13:1–16. https://doi.org/10.1007/s11306-017-1242-7.

[211] Want E, Masson P. Processing and analysis of GC/LC-MS-based metabolomics data. Methods Mol Biol 2011. https://doi.org/10.1007/978-1-61737-985-7_17.

[212] Vettukattil R. Preprocessing of raw metabonomic data. Methods Mol Biol 2015. https://doi.org/10.1007/978-1-4939-2377-9_10.

[213] Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Res 2018;46:D608–17. https://doi.org/10.1093/nar/gkx1089.

[214] Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G, et al. METLIN: A Technology Platform for Identifying Knowns and Unknowns. Anal Chem 2018. https://doi.org/10.1021/acs.analchem.7b04424.

[215] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010. https://doi.org/10.1002/jms.1777.

[216] Wang X, Jones DR, Shaw TI, Cho JH, Wang Y, Tan H, et al. Target-Decoy-Based False Discovery Rate Estimation for Large-Scale Metabolite Identification. J Proteome Res 2018;17:2328–34. https://doi.org/10.1021/acs.jproteome.8b00019.

[217] Alonso A, Marsal S, Julià A. Analytical methods in untargeted metabolomics: State of the art in 2015. Front Bioeng Biotechnol 2015;3:1–20. https://doi.org/10.3389/fbioe.2015.00023.

[218] Schiffman C, Petrick L, Perttula K, Yano Y, Carlsson H, Whitehead T, et al. Filtering procedures for untargeted lc-ms metabolomics data. BMC Bioinformatics 2019;20:1–10. https://doi.org/10.1186/s12859-019-2871-9.

[219] Li B, Tang J, Yang Q, Cui X, Li S, Chen S, et al. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. Sci Rep 2016;6:1–13. https://doi.org/10.1038/srep38881.

[220] Yang Q, Hong J, Li Y, Xue W, Li S, Yang H, et al. A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. Brief Bioinform 2019;00:1–11. https://doi.org/10.1093/bib/bbz137.

[221] Gorrochategui E, Jaumot J, Lacorte S, Tauler R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. TrAC - Trends Anal Chem 2016;82:425–42. https://doi.org/10.1016/j.trac.2016.07.004.

[222] O'Shea K, Misra BB. Software tools, databases and resources in metabolomics: updates from 2018 to 2019. Metabolomics 2020. https://doi.org/10.1007/s11306-020-01657-3.

[223] Huan T, Forsberg EM, Rinehart D, Johnson CH, Ivanisevic J, Benton HP, et al. Systems biology guided by XCMS Online metabolomics. Nat Methods 2017. https://doi.org/10.1038/nmeth.4260.

[224] Davidson RL, Weber RJM, Liu H, Sharma-Oates A, Viant MR. Galaxy-M: A Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. Gigascience 2016;5. https://doi.org/10.1186/s13742-016-0115-8.

[225] Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 2018;46:W486–94. https://doi.org/10.1093/nar/gky310.

[226] Chong J, Xia J. Using MetaboAnalyst 4.0 for Metabolomics Data Analysis, Interpretation, and Integration with Other Omics Data. Methods Mol. Biol., 2020. https://doi.org/10.1007/978-1-0716-0239-3_17.

[227] Wen B, Mei Z, Zeng C, Liu S. metaX: A flexible and comprehensive software for processing metabolomics data. BMC Bioinformatics 2017;18. https://doi.org/10.1186/s12859-017-1579-y.

[228] Wang X, Cho JH, Poudel S, Li Y, Jones DR, Shaw TI, et al. JUMPm: A tool for large-scale identification of metabolites in untargeted metabolomics. Metabolites 2020;10.

https://doi.org/10.3390/metabo10050190.

[229] Balog J, Szaniszlo T, Schaefer KC, Denes J, Lopata A, Godorhazy L, et al. Identification of biological tissues by rapid evaporative ionization mass spectrometry. Anal Chem 2010. https://doi.org/10.1021/ac101283x.

[230] Tzafetas M, Mitra A, Paraskevaidi M, Bodai Z, Kalliala I, Bowden S, et al. The intelligent knife (iKnife) and its intraoperative diagnostic advantage for the treatment of cervical disease. Proc Natl Acad Sci U S A 2020;117:7338–46. https://doi.org/10.1073/pnas.1916960117.

[231] Duncan KD, Fyrestam J, Lanekoff I. Advances in mass spectrometry based single-cell metabolomics. Analyst 2019;144:782–93. https://doi.org/10.1039/c8an01581c.

[232] Kaushik AK, DeBerardinis RJ. Applications of metabolomics to study cancer metabolism. Biochim Biophys Acta - Rev Cancer 2018;1870:2–14. https://doi.org/10.1016/j.bbcan.2018.04.009.

[233] Giraudeau P. NMR-based metabolomics and fluxomics: Developments and future prospects. Analyst 2020;145:2457–72. https://doi.org/10.1039/d0an00142b.

[234] Cheung PK, Ma MH, Tse HF, Yeung KF, Tsang HF, Chu MKM, et al. The applications of metabolomics in the molecular diagnostics of cancer. Expert Rev Mol Diagn 2019;19:785–93. https://doi.org/10.1080/14737159.2019.1656530.

[235] Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet 2018;19:299–310. https://doi.org/10.1038/nrg.2018.4.

[236] Zeng H, Chen L, Huang Y, Luo Y, Ma X. Integrative Models of Histopathological Image Features and Omics Data Predict Survival in Head and Neck Squamous Cell Carcinoma. Front Cell Dev Biol 2020;8. https://doi.org/10.3389/fcell.2020.553099.

[237] Yu K-H, Berry GJ, Rubin DL, Ré C, Altman RB, Snyder M. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. Cell Syst 2017;5:620-627.e3. https://doi.org/10.1016/j.cels.2017.10.014.

[238] Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. Metabolites 2019;9:1–31. https://doi.org/10.3390/metabo9040076.

[239] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights 2020;14:7–9. https://doi.org/10.1177/1177932219899051.

[240] Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 2014;14:299–313. https://doi.org/10.1038/nrc3721.

[241] Sahin U, Türeci Ö. Personalized vaccines for cancer immunotherapy. Science (80-) 2018;359:1355–60. https://doi.org/10.1126/science.aar7112.

[242] Prasad V. The precision-oncology illusion. Nat Outlook 2016;537:S63.

[243] Zhang B, Whiteaker JR, Hoofnagle AN, Baird GS, Rodland KD, Paulovich AG. Clinical

potential of mass spectrometry-based proteogenomics. Nat Rev Clin Oncol 2019;16:256–68. https://doi.org/10.1038/s41571-018-0135-7.

[244] Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW, et al. Quokka: A comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics 2018;34:4223–31. https://doi.org/10.1093/bioinformatics/bty522.

[245] Ruggles K V., Krug K, Wang X, Clauser KR, Wang J, Payne SH, et al. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics 2017;16:959–81. https://doi.org/10.1074/mcp.MR117.000024.

[246] Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies. Nat Methods 2014;11:1114–25. https://doi.org/10.1038/NMETH.3144.

[247] 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. https://doi.org/10.1016/j.jprot.2010.08.009.

[248] Wang X, Slebos RJC, Wang D, Halvey PJ, Tabb DL, Liebler DC, et al. Protein identification using customized protein sequence databases derived from RNA-seq data. J Proteome Res 2012. https://doi.org/10.1021/pr200766z.

[249] Ruggles K V., Tang Z, Wang X, Grover H, Askenazi M, Teubl J, et al. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. Mol Cell Proteomics 2016;15:1060–71. https://doi.org/10.1074/mcp.M115.056226.

[250] Mertins P, Mani DR, Ruggles K V., Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534:55–62. https://doi.org/10.1038/nature18003.

[251] Kumar D, Yadav AK, Jia X, Mulvenna J, Dash D. Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. Mol Cell Proteomics 2016;15:329–39. https://doi.org/10.1074/mcp.M114.047126.

[252] Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 2019;8. https://doi.org/10.1093/gigascience/giz039.

[253] Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol 2019;20:213. https://doi.org/10.1186/s13059-019-1842-9.

[254] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29:644–52. https://doi.org/10.1038/nbt.1883.

[255] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods 2010;7:909–12. https://doi.org/10.1038/nmeth.1517.

[256] Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 2014;30:1660–6. https://doi.org/10.1093/bioinformatics/btu077.

[257] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021.

[258] Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, et al. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics 2014;15:1–9. https://doi.org/10.1186/1471-2164-15-703.

[259] Wang X, Zhang B, Wren J. CustomProDB: An R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 2013;29:3235–7. https://doi.org/10.1093/bioinformatics/btt543.

[260] Wen B, Xu S, Zhou R, Zhang B, Wang X, Liu X, et al. PGA: An R/Bioconductor package for identification of novel peptides using a customized database derived from RNA-Seq. BMC Bioinformatics 2016;17:1–6. https://doi.org/10.1186/s12859-016-1133-3.

[261] Chambers MC, Jagtap PD, Johnson JE, McGowan T, Kumar P, Onsongo G, et al. An accessible proteogenomics informatics resource for cancer researchers. Cancer Res 2017. https://doi.org/10.1158/0008-5472.CAN-17-0331.

[262] Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 2014;5. https://doi.org/10.1038/ncomms6277.

[263] Eng JK, Jahan TA, Hoopmann MR. Comet: An open-source MS/MS sequence database search tool. Proteomics 2013;13:22–4. https://doi.org/10.1002/pmic.201200439.

[264] Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nat Commun 2020;11:1–14. https://doi.org/10.1038/s41467-020-15456-w.

[265] Wen B, Wang X, Zhang B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. Genome Res 2019;29:485–93. https://doi.org/10.1101/gr.235028.118.

[266] Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell 2019;177:1035-1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

[267] El Marabti E, Younis I. The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer. Front Mol Biosci 2018;5. https://doi.org/10.3389/fmolb.2018.00080.

[268] Monteuuis G, Schmitz U, Petrova V, Kearney PS, Rasko JEJ. Holding on to junk bonds: intron retention in cancer and therapy. Cancer Res 2020. https://doi.org/10.1158/0008-5472.can-20-1943.

[269] Komor MA, Pham T V., Hiemstra AC, Piersma SR, Bolijn AS, Schelfhorst T, et al. Identification of Differentially Expressed Splice Variants by the Proteogenomic Pipeline Splicify. Mol Cell Proteomics 2017;16:1850–63. https://doi.org/10.1074/mcp.TIR117.000056.

[270] Chen L, Miao Y, Liu M, Zeng Y, Gao Z, Peng D, et al. Pan-Cancer Analysis Reveals the Functional Importance of Protein Lysine Modification in Cancer Development. Front Genet 2018;9. https://doi.org/10.3389/fgene.2018.00254.

[271] Mun D-G, Bhin J, Kim S, Kim H, Jung JH, Jung Y, et al. Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell 2019;35:111-124.e10. https://doi.org/10.1016/j.ccell.2018.12.003.

[272] Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett 2009;583:3966–73. https://doi.org/10.1016/j.febslet.2009.10.036.

[273] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015. https://doi.org/10.1016/j.cell.2014.12.033.

[274] Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science (80-) 2015;348:69–74. https://doi.org/10.1126/science.aaa4971.

[275] Peng M, Mo Y, Wang Y, Wu P, Zhang Y, Xiong F, et al. Neoantigen vaccine: An emerging tumor immunotherapy. Mol Cancer 2019;18:1–14. https://doi.org/10.1186/s12943-019-1055-6.

[276] Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature 2019;565:234–9. https://doi.org/10.1038/s41586-018-0792-9.

[277] Gupta SK, Jaitly T, Schmitz U, Schuler G, Wolkenhauer O, Vera J. Personalized cancer immunotherapy using Systems Medicine approaches. Brief Bioinform 2016. https://doi.org/10.1093/bib/bbv046.

[278] Kanaseki T, Tokita S, Torigoe T. Proteogenomic discovery of cancer antigens: Neoantigens and beyond. Pathol Int 2019;69:511–8. https://doi.org/10.1111/pin.12841.

[279] McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science (80-) 2016. https://doi.org/10.1126/science.aaf1490.

[280] Vinay DS, Ryan EP, Pawelec G, Talib WH, Stagg J, Elkord E, et al. Immune evasion in cancer: Mechanistic basis and therapeutic strategies. Semin Cancer Biol 2015;35:S185–98. https://doi.org/10.1016/j.semcancer.2015.03.004.

[281] Haber DA, Settleman J. Cancer: Drivers and passengers. Nature 2007;446:145–6. https://doi.org/10.1038/446145a.

[282] Shukla HD. Comprehensive analysis of cancer-proteogenome to identify biomarkers for the early diagnosis and prognosis of cancer. Proteomes 2017;5:1–17. https://doi.org/10.3390/proteomes5040028.

[283] Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. Cell 2017;168:707–23. https://doi.org/10.1016/j.cell.2017.01.017.

[284] Nishimura T, Nakamura H, Végvári Á, Marko-Varga G, Furuya N, Saji H. Current status of clinical proteogenomics in lung cancer. Expert Rev Proteomics 2019;16:761–72. https://doi.org/10.1080/14789450.2019.1654861.

[285] Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, et al. A targeted proteomics-based pipeline for verification of biomarkers in plasma. Nat Biotechnol 2011;29:625–34.

https://doi.org/10.1038/nbt.1900.

[286] Bache N, Geyer PE, Bekker-Jensen DB, Hoerning O, Falkenby L, Treit P V., et al. A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. Mol Cell Proteomics 2018;17:2284–96. https://doi.org/10.1074/mcp.TIR118.000853.

[287] Meier F, Brunner AD, Koch S, Koch H, Lubeck M, Krause M, et al. Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. Mol Cell Proteomics 2018;17:2534–45. https://doi.org/10.1074/mcp.TIR118.000900.

[288] Satpathy S, Jaehnig EJ, Krug K, Kim BJ, Saltzman AB, Chan DW, et al. Microscaled proteogenomic methods for precision oncology. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-14381-2.

[289] Moshkovskii SA, Lobas AA, Gorshkov M V. Single Cell Proteogenomics — Immediate Prospects. Biochem 2020;85:140–6. https://doi.org/10.1134/S0006297920020029.

[290] Specht H, Emmott E, Petelski AA, Huffman RG, Perlman DH, Serra M, et al. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. Genome Biol 2021;22:50. https://doi.org/10.1186/s13059-021-02267-5.

[291] Budnik B, Levy E, Harmange G, Slavov N. SCoPE-MS: Mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation 06 Biological Sciences 0601 Biochemistry and Cell Biology 06 Biological Sciences 0604 Genetics. Genome Biol 2018;19:1–12. https://doi.org/10.1186/s13059-018-1547-5.

[292] Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics 2009. https://doi.org/10.1002/pmic.200800473.

[293] Li Y, Wang X, Cho JH, Shaw TI, Wu Z, Bai B, et al. JUMPg: An integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells. J Proteome Res 2016;15:2309–20. https://doi.org/10.1021/acs.jproteome.6b00344.

[294] Li Y, Wang G, Tan X, Ouyang J, Zhang M, Song X, et al. ProGeo-neo: A customized proteogenomic workflow for neoantigen prediction and selection. BMC Med Genomics 2020;13:1–11. https://doi.org/10.1186/s12920-020-0683-4.

[295] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362–8. https://doi.org/10.1093/nar/gkw937.

[296] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006. https://doi.org/10.1093/nar/gkj109.

[297] Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation. Nucleic Acids Res 2013. https://doi.org/10.1093/nar/gks1147.

[298] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017. https://doi.org/10.1093/nar/gkw1092.

[299] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res 2020;48:D498–503. https://doi.org/10.1093/nar/gkz1031.

[300] Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, et al. The Virtual Metabolic Human database: Integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Res 2019. https://doi.org/10.1093/nar/gky992.

[301] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 2018;46:D661–7. https://doi.org/10.1093/nar/gkx1064.

[302] Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. Int J Proteomics 2014. https://doi.org/10.1155/2014/147648.

[303] Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front Oncol 2020;10. https://doi.org/10.3389/fonc.2020.01030.

[304] Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. Brief Bioinform 2017;19:1370–81. https://doi.org/10.1093/bib/bbx066.

[305] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: A universal amplifier of genetic associations. Nat Rev Genet 2017;18:551–62. https://doi.org/10.1038/nrg.2017.38.

[306] Jalili M, Gebhardt T, Wolkenhauer O, Salehzadeh-Yazdi A. Unveiling network-based functional features through integration of gene expression into protein networks. Biochim Biophys Acta - Mol Basis Dis 2018;1864:2349–59. https://doi.org/10.1016/j.bbadis.2018.02.010.

[307] Robinson JL, Nielsen J. Integrative analysis of human omics data using biomolecular networks. Mol Biosyst 2016;12:2953–64. https://doi.org/10.1039/c6mb00476h.

[308] Jalili M, Salehzadeh-Yazdi A, Gupta S, Wolkenhauer O, Yaghmaie M, Resendis-Antonio O, et al. Evolution of centrality measurements for the detection of essential proteins in biological networks. Front Physiol 2016;7. https://doi.org/10.3389/fphys.2016.00375.

[309] Yang Q, Wang S, Dai E, Zhou S, Liu Di, Liu H, et al. Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. Brief Bioinform 2019;20:168–77. https://doi.org/10.1093/bib/bbx091.

[310] Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. Nature 2000;408:307–10. https://doi.org/10.1038/35042675.

[311] Khan FM, Marquardt S, Gupta SK, Knoll S, Schmitz U, Spitschak A, et al. Unraveling a tumor type-specific regulatory core underlying E2F1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. Nat Commun 2017. https://doi.org/10.1038/s41467-017-00268-2.

[312] Ozturk K, Dow M, Carlin DE, Bejar R, Carter H. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. J Mol Biol 2018;430:2875–99. https://doi.org/10.1016/j.jmb.2018.06.016.

[313] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007;3:1–10. https://doi.org/10.1038/msb4100180.

[314] 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. https://doi.org/10.1016/j.ebiom.2017.11.028.

[315] Zhou G, Li S, Xia J. Network-Based Approaches for Multi-omics Integration. Methods Mol. Biol., 2020. https://doi.org/10.1007/978-1-0716-0239-3_23.

[316] Koh GCKW, Porras P, Aranda B, Hermjakob H, Orchard SE. Analyzing protein-protein interaction networks. J Proteome Res 2012. https://doi.org/10.1021/pr201211w.

[317] Zhang C, Hua Q. Applications of genome-scale metabolic models in biotechnology and systems medicine. Front Physiol 2016. https://doi.org/10.3389/fphys.2015.00413.

[318] Mosca R, Pons T, Céol A, Valencia A, Aloy P. Towards a detailed atlas of protein-protein interactions. Curr Opin Struct Biol 2013;23:929–40. https://doi.org/10.1016/j.sbi.2013.07.005.

[319] Stumpf MPH, Thorne T, De Silva E, Stewart R, Hyeong JA, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A 2008;105:6959–64. https://doi.org/10.1073/pnas.0708078105.

[320] Sadeghi M, Ordway B, Rafiei I, Borad P, Fang B, Koomen JL, et al. Integrative Analysis of Breast Cancer Cells Reveals an Epithelial-Mesenchymal Transition Role in Adaptation to Acidic Microenvironment. Front Oncol 2020;10:1–14. https://doi.org/10.3389/fonc.2020.00304.

[321] Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus 2013;3. https://doi.org/10.1098/rsfs.2013.0013.

[322] Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. Npj Syst Biol Appl 2019;5. https://doi.org/10.1038/s41540-019-0099-y.

[323] Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 2007;1:1–17. https://doi.org/10.1186/1752-0509-1-8.

[324] Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. BMC Bioinformatics 2019;20:1–14. https://doi.org/10.1186/s12859-019-3146-1.

[325] Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999. https://doi.org/10.1038/35011540.

[326] Goh K Il, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A 2007;104:8685–90. https://doi.org/10.1073/pnas.0701361104.

[327] Liu C, Ma Y, Zhao J, Nussinov R, Zhang YC, Cheng F, et al. Computational network biology: Data, models, and applications. Phys Rep 2020;846:1–66.

2114 https://doi.org/10.1016/j.physrep.2019.12.004.

2115 [328] Hodzic E, Shrestha R, Zhu K, Cheng K, Collins CC, Cenk Sahinalp S. Combinatorial
2116 Detection of Conserved Alteration Patterns for Identifying Cancer Subnetworks.
2117 Gigascience 2019;8:1–13. https://doi.org/10.1093/gigascience/giz024.

2118 [329] Nguyen H, Shrestha S, Tran D, Shafi A, Draghici S, Nguyen T. A comprehensive survey
2119 of tools and software for active subnetwork identification. Front Genet 2019;10.
2120 https://doi.org/10.3389/fgene.2019.00155.

2121 [330] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of
2122 modularity in metabolic networks. Science (80-) 2002;297:1551–5.
2123 https://doi.org/10.1126/science.1073374.

2124 [331] Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: Identifying hierarchies of
2125 altered subnetworks. Bioinformatics 2018;34:i972–80.
2126 https://doi.org/10.1093/bioinformatics/bty613.

2127 [332] Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling
2128 circuits in molecular interaction networks. Bioinformatics, 2002.
2129 https://doi.org/10.1093/bioinformatics/18.suppl_1.S233.

2130 [333] Ghiassian SD, Menche J, Barabási AL. A DIseAse MOdule Detection (DIAMOnD)
2131 Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease
2132 Proteins in the Human Interactome. PLoS Comput Biol 2015.
2133 https://doi.org/10.1371/journal.pcbi.1004120.

2134 [334] Ma H, Schadt EE, Kaplan LM, Zhao H. COSINE: COndition-SpecIfic sub-NEtwork
2135 identification using a global optimization method. Bioinformatics 2011.
2136 https://doi.org/10.1093/bioinformatics/btr136.

2137 [335] Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies
2138 oncogenic network modules. Genome Res 2012;22:398–406.
2139 https://doi.org/10.1101/gr.125567.111.

2140 [336] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A
2141 software Environment for integrated models of biomolecular interaction networks. Genome
2142 Res 2003;13:2498–504. https://doi.org/10.1101/gr.1239303.

2143 [337] Kusonmano K, Halle MK, Wik E, Hoivik EA, Krakstad C, Mauland KK, et al.
2144 Identification of highly connected and differentially expressed gene subnetworks in
2145 metastasizing endometrial cancer. PLoS One 2018;13:1–21.
2146 https://doi.org/10.1371/journal.pone.0206665.

2147 [338] Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS Approach: A
2148 New Frontier in Cancer Research. Biomed Res Int 2018;2018.
2149 https://doi.org/10.1155/2018/9836256.

2150 [339] Das S, Mcclain CJ, Rai SN. Fifteen years of gene set analysis for high-throughput genomic
2151 data: A review of statistical approaches and future challenges. Entropy 2020;22:1–23.
2152 https://doi.org/10.3390/E22040427.

2153 [340] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: Current approaches and
2154 outstanding challenges. PLoS Comput Biol 2012;8.

https://doi.org/10.1371/journal.pcbi.1002375.

[341] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57. https://doi.org/10.1038/nprot.2008.211.

[342] Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res 2019;47:W199–205. https://doi.org/10.1093/nar/gkz401.

[343] Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform 2008;9:189–97. https://doi.org/10.1093/bib/bbn001.

[344] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.

[345] Gerstner N, Kehl T, Lenhof K, Müller A, Mayer C, Eckhart L, et al. GeneTrail 3: advanced high-throughput enrichment analysis. Nucleic Acids Res 2020;48:W515–20. https://doi.org/10.1093/nar/gkaa306.

[346] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics 2009;25:75–82. https://doi.org/10.1093/bioinformatics/btn577.

[347] Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: A comprehensive review and assessment. Genome Biol 2019;20:1–15. https://doi.org/10.1186/s13059-019-1790-4.

[348] Jacob L, Neuvial P, Dudoit S. More power via graph-structured tests for differential expression of gene networks. Ann Appl Stat 2012;6:561–600. https://doi.org/10.1214/11-AOAS528.

[349] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-09234-6.

[350] Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. Impact of outdated gene annotations on pathway enrichment analysis. Nat Methods 2016;13:705–6. https://doi.org/10.1038/nmeth.3963.

[351] Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. Nat Methods 2010;7:S56–68. https://doi.org/10.1038/nmeth.1436.

[352] Hernández-De-Diego R, Tarazona S, Martínez-Mira C, Balzano-Nogueira L, Furió-Tarí P, Pappas GJ, et al. PaintOmics 3: A web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res 2018;46:W503–9. https://doi.org/10.1093/nar/gky466.

[353] Zhou G, Xia J. OmicsNet: A web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Res 2018;46:W514–22. https://doi.org/10.1093/nar/gky510.

[354] Wang L, Xiao Y, Ping Y, Li J, Zhao H, Li F, et al. Integrating Multi-Omics for Uncovering the Architecture of Cross-Talking Pathways in Breast Cancer. PLoS One 2014;9:e104282. https://doi.org/10.1371/journal.pone.0104282.

[355] Schulte-Sasse R, Budach S, Hnisz D, Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat Mach Intell 2021. https://doi.org/10.1038/s42256-021-00325-y.

[356] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (80-) 2014. https://doi.org/10.1126/science.1254257.

[357] Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. Brief Bioinform 2018;19:878–92. https://doi.org/10.1093/bib/bbx017.

[358] Turanli B, Karagoz K, Gulfidan G, Sinha R, Mardinoglu A, Arga KY. A Network-Based Cancer Drug Discovery: From Integrated Multi-Omics Approaches to Precision Medicine. Curr Pharm Des 2019. https://doi.org/10.2174/1381612824666181106095959.

[359] Taber A, Christensen E, Lamy P, Nordentoft I, Prip F, Lindskrog SV, et al. Molecular correlates of cisplatin-based chemotherapy response in muscle invasive bladder cancer by integrated multi-omics analysis. Nat Commun 2020;11:4858. https://doi.org/10.1038/s41467-020-18640-0.

[360] Huang C, Chen L, Savage SR, Eguez RV, Dou Y, Li Y, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell 2021;39:361-379.e16. https://doi.org/10.1016/j.ccell.2020.12.007.

[361] Paull EO, Aytes A, Jones SJ, Subramaniam PS, Giorgi FM, Douglass EF, et al. A modular master regulator landscape controls cancer transcriptional identity. Cell 2021;184:334-351.e20. https://doi.org/10.1016/j.cell.2020.11.045.

[362] Suvà ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell 2019;75:7–12. https://doi.org/10.1016/j.molcel.2019.05.003.

[363] Piccirillo SGM, Reynolds BA, Zanetti N, Lamorte G, Binda E, Broggi G, et al. Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. Nature 2006;444:761–5. https://doi.org/10.1038/nature05349.

[364] Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, Flamourakis G, et al. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. Cell Rep 2016;14:380–9. https://doi.org/10.1016/j.celrep.2015.12.021.

[365] Analysis P, Genomes W, Cancer I, Consortium G, Cancer T, Atlas G, et al. Pan-cancer analysis of whole genomes 2020;578. https://doi.org/10.1038/s41586-020-1969-6.

[366] Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res 2020. https://doi.org/10.1093/nar/gkaa434.

[367] Van Helden P. Data-driven hypotheses. EMBO Rep 2013. https://doi.org/10.1038/embor.2012.207.

[368] Rusch M, Nakitandwe J, Shurtleff S, Newman S, Zhang Z, Edmonson MN, et al. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat Commun 2018;9. https://doi.org/10.1038/s41467-018-06485-7.

[369] Franzén O, Björkegren JLM. alona: a web server for single-cell RNA-seq analysis. Bioinformatics 2020. https://doi.org/10.1093/bioinformatics/btaa269.

[370] Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods 2017. https://doi.org/10.1038/nmeth.4150.

[371] Xu J, Cai L, Liao B, Zhu W, Yang JL. CMF-Impute: An accurate imputation tool for single-cell RNA-seq data. Bioinformatics 2020. https://doi.org/10.1093/bioinformatics/btaa109.

[372] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst 2019. https://doi.org/10.1016/j.cels.2019.03.003.

[373] Gong W, Kwak IY, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: Imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics 2018. https://doi.org/10.1186/s12859-018-2226-y.

[374] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018. https://doi.org/10.1038/nbt.4091.

[375] Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: Gene expression recovery for single-cell RNA sequencing. Nat Methods 2018. https://doi.org/10.1038/s41592-018-0033-z.

[376] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell 2019. https://doi.org/10.1016/j.cell.2019.05.031.

[377] McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 2017. https://doi.org/10.1093/bioinformatics/btw777.

[378] Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014. https://doi.org/10.1038/nmeth.2967.

[379] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: Single-cell regulatory network inference and clustering. Nat Methods 2017. https://doi.org/10.1038/nmeth.4463.

[380] Cai JJ. scGEAToolbox: A matlab toolbox for single-cell RNA sequencing data analysis. Bioinformatics 2020. https://doi.org/10.1093/bioinformatics/btz830.

[381] Kim CY, Na K, Park S, Jeong SK, Cho JY, Shin H, et al. FusionPro, a versatile proteogenomic tool for identification of novel fusion transcripts and their potential translation products in cancer cells. Mol Cell Proteomics 2019;18:1651–68. https://doi.org/10.1074/mcp.RA119.001456.

[382] Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, Mandyam RA, et al.

PGTools: A software suite for proteogenomic data analysis and visualization. J Proteome Res 2015. https://doi.org/10.1021/acs.jproteome.5b00029.

[383] Li Y, Wang G, Tan X, Ouyang J, Zhang M, Song X, et al. ProGeo-neo: A customized proteogenomic workflow for neoantigen prediction and selection. BMC Med Genomics 2020. https://doi.org/10.1186/s12920-020-0683-4.

[384] Verbruggen S, Ndah E, Van Criekinge W, Gessulat S, Kuster B, Wilhelm M, et al. PROTEOFORMER 2.0: Further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. Mol Cell Proteomics 2019. https://doi.org/10.1074/mcp.RA118.001218.

[385] Lee SY, Hwang H, Kang YM, Kim H, Kim DG, Jeong JE, et al. SAAVpedia: Identification, Functional Annotation, and Retrieval of Single Amino Acid Variants for Proteogenomic Interpretation. J Proteome Res 2019. https://doi.org/10.1021/acs.jproteome.9b00366.

[386] Cesnik AJ, Miller RM, Ibrahim K, Lu L, Millikin RJ, Shortreed MR, et al. Spritz: A Proteogenomic Database Engine. J Proteome Res 2020. https://doi.org/10.1021/acs.jproteome.0c00407.

Figure 1. A timeline of some of the major contributions to the field of systems biology

Figure 2. General workflows for different omics studies. The wet lab and computational procedures are distinguished by different background colors.

Figure 3. Integrative study of biological phenomena. The first fundamental decision for modern large-scale studies is the choice between hypothesis-driven or data-driven study design. While both types of study designs are applicable, complementary approaches are recommended since hypothesis-driven studies are vulnerable to bias while data-driven studies are highly prone to false-positives [367]. The extracted omics data can be subjected to integration through multiple approaches. The resulting functional data will improve our knowledge base and can serve as a starting point for future studies. Already emerging pipelines demonstrate the clinical utility of the integrative approaches [368]. The integration approaches provided in this figure are based on the categorization in [240]. Sequential analysis: the integration of datasets subsequent to independent analysis. Latent variable analysis: Partitioning of samples into functional groups through unsupervised clustering for example by implementation of an expectation-maximization algorithm. Penalized likelihood analysis: outcome prediction through penalized regression. Pairwise correlation analysis: association estimation for related molecule pairs across datasets. Gene set analysis: homogenization of multiple datasets by replacing every molecule with its respective gene and subsequent enrichment of the resulting datasets. Network analysis: using prior knowledge of molecular interactions to provide an environment for integration. Bayesian analysis: Utilization of the information in an omics layer as the prior information for the analysis of another through Bayesian approaches.

Figure 4. General workflow for the integration of genomics and tandem mass spectrometry data in proteogenomics. The MS/MS spectra of the sample are searched against the theoretical spectra inferred from the NGS data (most commonly RNA-seq) obtained from the same sample. The identified novel peptides should be validated (using PepQuery). The resulting data can be utilized for the study of post-translational modifications, identification of neo-antigens and biomarkers, and mutation prioritization in the downstream interpretation. Network-based analysis of these data can provide a critical vantage point for functional study of system perturbations.

Figure 5. General workflow for the network-based analysis of omics data. The constructed sub-networks from the integration of the omics-driven data and prior knowledge of molecular interactions can be subjected to module identification or enrichment analysis. The identified modules can also be enriched to yield functional information. Note that it is possible to enrich the omics data independent of the subnetwork construction process. An example of downstream interpretation is to demonstrate multi-omics data in multi-layered networks for computational and/or visual pattern detection. Going from either raw omics data or interactome databases to subnetwork modules and enriched data, the complexity decreases, and the data is constantly narrowed down to yield functional information. ORA, over-representation analysis; FCS, functional class scoring.