The Many Faces of Gene Regulation in Cancer: A Computational Oncogenomics Outlook

Enrique Hernández-Lemus [1,2,*], **Helena Reyes-Gopar [1]**, **Jesús Espinal-Enríquez [1,2]** and **Soledad Ochoa [1]**

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City 14610, Mexico; hreyes.gopar@gmail.com (H.R.-G.); jespinal@inmegen.gob.mx (J.E.-E.); mochoa@inmegen.edu.mx (S.O.)

[2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico

[*] Correspondence: ehernandez@inmegen.gob.mx; Tel.: +52-55-5350-1970

Abstract: Cancer is a complex disease at many different levels. The molecular phenomenology of cancer is also quite rich. The mutational and genomic origins of cancer and their downstream effects on processes such as the reprogramming of the gene regulatory control and the molecular pathways depending on such control have been recognized as central to the characterization of the disease. More important though is the understanding of their causes, prognosis, and therapeutics. There is a multitude of factors associated with anomalous control of gene expression in cancer. Many of these factors are now amenable to be studied comprehensively by means of experiments based on diverse omic technologies. However, characterizing each dimension of the phenomenon individually has proven to fall short in presenting a clear picture of expression regulation as a whole. In this review article, we discuss some of the more relevant factors affecting gene expression control both, under normal conditions and in tumor settings. We describe the different omic approaches that we can use as well as the computational genomic analysis needed to track down these factors. Then we present theoretical and computational frameworks developed to integrate the amount of diverse information provided by such single-omic analyses. We contextualize this within a systems biology-based multi-omic regulation setting, aimed at better understanding the complex interplay of gene expression deregulation in cancer.

Keywords: computational oncogenomics; gene expression regulation; multi-omics; integrative biology

1. Cancer as a Complex Phenotype

Cancer is a complex disease characterized by a series of highly variable and inhomogeneous phenomena. The set of individual and environmental factors associated with the onset and progression of cancer is large and diverse. These factors include several types of DNA mutations, chemical modifications to the DNA and histone proteins leading to epigenomic changes, alterations to the three-dimensional structure of chromatin and various processes that introduce biases resulting in clonal and sub-clonal selection. These changes impact the different levels of the gene regulatory programs, thus modifying gene expression in cancer. Gene regulatory processes are additionally affected by changes in the metabolic and signaling activity of the tumor cell and its surrounding microenvironment [1–4].

The rise of high-throughput omic technologies has improved our understanding of the complex regulatory landscape in both normal and tumor cells. The methodologies provided by the development of these technologies have become paramount for oncological research in the basic and clinical settings, thus paving the way to translational personalized medicine [5]. However, the high amount of data

provided by these methods needs to be supplemented with appropriate analysis and integration schemes if they are to provide insightful theories and more importantly, successful diagnostic, prognostic and therapeutic tools.

Currently, there is a need for computational implementations to handle the analysis of data created by the ever-growing arsenal of omic technologies. Such tools should allow detailed quantitative descriptions of complex and noisy data, which in turn, ought to set the foundations for integrative modeling approaches. For these reasons, an upcoming field in computational biology termed computational Oncology has attracted interest from the scientific community [6–8]. Broadly, computational oncology is considered to include two main research branches; one of them focused on data processing tasks is usually called cancer bioinformatics, and the other one, cancer systems biology or systems oncology, aims at translating a large amount of analyzed data into some form of rational model that can be used to drive focused experimental research with the potential of being useful in the clinical context [9–13].

Cancer is widely known that is a gene-based disease, mutations in specific sequences of crucial genes, such as *p53*, *MDM2*, *BRCA1*, and *BRCA2*, to mention a few, are positively correlated with cancer appearance, development, and prognosis. However, mutations are not the only mechanism behind cancer appearance. The whole gene expression pattern, which is in turn affected by other factors is responsible for the oncogenic phenotypes.

The present review discusses the current state of the computational oncology field. In the next section (Section 2) we discuss in detail both, the well-established and emerging features that have been associated with deregulated gene expression in cancer. These features and the processes from which they arise are usually measured and analyzed separately, giving rise to individual (even fragmented) results and ensuing a relatively incomplete portrait of the complex phenotype, thus highlighting the multifactorial character of genomic regulation in cancer. Section 3 introduces the omic techniques currently used to study different genomic aspects known to impact gene regulation, how they can be used to probe tumor cell behavior and which computational methods are available to analyze their data output. Finally, in Section 4 we discuss the need to develop data integration frameworks and multi-dimensional models to account, not just for the individual contributions of each omic/regulatory layer but also for their concerted interplay and its contribution to shaping altered gene regulatory patterns.

2. The Multifactorial Character of Gene Regulation in Cancer

A central issue towards the molecular understanding of tumorigenesis is the progressive increase of altered gene expression patterns as a consequence of the aberrant genetic and epigenetic programs. Many elements are known to participate in the regulatory mechanisms of gene expression (see Figure 1). Some of them have been studied for decades and their operation is substantially understood, others await to be fully characterized. As it turns out, a myriad of events affect gene regulation in cancer, from direct changes in the DNA sequence (single nucleotide variants, insertions/deletions, inversions, translocations, copy number variation, gene fusions, among others) impacting gene sequences, transcription factor (TF) genes or regulatory sequences, to post-translational and epigenomic alterations.

The synthesis of RNA by RNA polymerase enzymes is termed transcription and it is responsible for the timely expression of the required genes and therefore proteins that ensure cells perform their functions properly. Because transcription must be tightly regulated to be specific, yet flexible, it is exquisitely controlled by a set of interconnected mechanisms comprising specific DNA sequences "cis elements" and specialized proteins "trans elements" that operate in the nuclear epigenetic context to activate and repress genes in response to stimuli (i.e., signal transduction) or to manage developmental and differentiation processes.

Mutations that disrupt either cis-elements or the gene sequence of trans-elements involved in transcription can lead to dangerous deregulation of gene expression, which could eventually result

in cancer [14,15]. The latter has been evidenced by the fact that non-coding mutations in cancer can have regulatory impact [16,17] and that several oncogenes are actually TFs [18]. It has also been studied how cancer cells repurpose the transcriptional mechanisms to ensure proliferation [19] and dissemination [20].

Additionally, a number of factors known to have an influence on the regulation of gene expression on the rest of this section we will present some preliminaries on most of these factors.

2.1. The Role of Promoters and Enhancers

Promoter and enhancer cooperativity (P+E) is an established mechanism of gene expression regulation. Promoters are DNA sequences located near the transcription start site of genes comprising specific sequences known as response elements that are used as binding sites for the transcriptional machinery. On the other hand, enhancers are DNA sequences capable of binding activator proteins that recruit the transcriptional machinery to distant promoters. Enhancers are regulatory elements that can act as key regulators of tissue-specific gene expression. Promoter regions often work in conjunction with specific enhancers. A role for P+E in the expression changes associated with carcinogenesis and tumor sustenance has been established [17].

It has also been reported that mutations in cancer can lead to enhancer misregulation. For example, it has been extensively shown the crucial role that MLL3/MLL4/COMPASS-like family of histone H3 lysine 4 (H3K4) monomethyltransfereases have in cancer [21,22] and also the role that TERT promoter mutations leading to upregulation of telomerase expression observed in human cancer [23,24]. Thus, P+E activity has been recently considered a promising option in cancer therapeutics [25].

2.2. The Effect of Chromatin Structure

Structurally different regions in the chromatin fiber are able to control gene expression to a certain extent [26,27]. Traditionally and at a coarse-grained scale, genomic regions are distinguished according to whether they are in an open chromatin configuration (euchromatin) or in a closed chromatin state (heterochromatin). The former allows TFs and the transcriptional machinery to bind regulatory DNA sequences, while the latter precludes gene expression [28]. Changes in the euchromatin/heterochromatin distributions have been investigated and linked to cancer development [29].

2.3. DNA Methylation and Other Chemical Modifications

Chemical modifications to the DNA molecule, namely the covalent attachment of methyl groups to cytosine nucleotides, are able to modulate transcription at different levels: from fully preventing the process, through modulating it, to activating transcription. Differential methylation also influences spatial chromosomal configuration [30] and some DNA methylation patterns are so pervasive in some cancers [31–34], to the point where they have been used as biomarkers for classification and prognosis [35,36]. On the other hand, the distribution and intensity of chemical modification at the nucleosome level, including the covalent attachment of methyl, acetyl or phosphoryl groups and the addition of ubiquitin or SUMO proteins to histone N-terminal tails contributes to the already complex phenomenon of gene regulation in cancer.
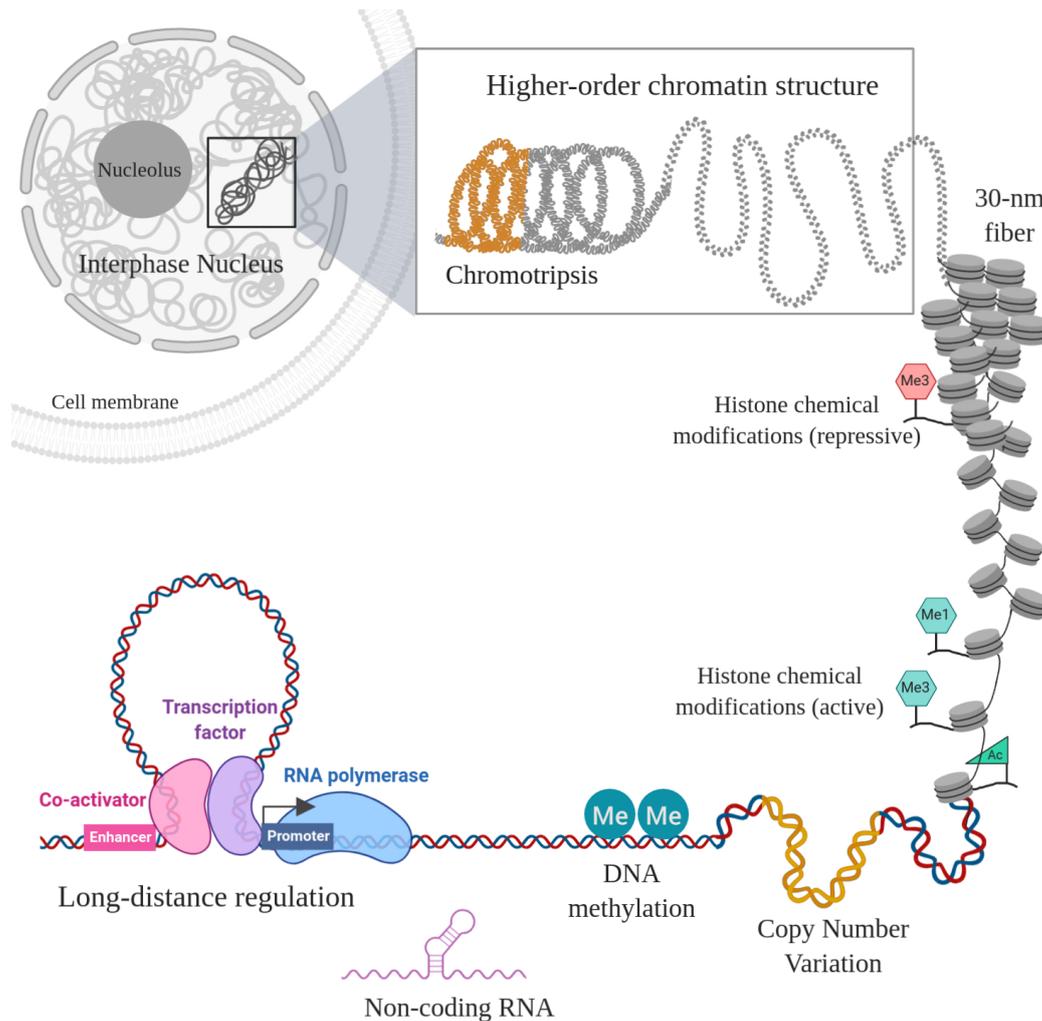
Figure 1. Gene expression regulatory mechanisms. Gene regulation in the eukaryotic cell nucleus involves different mechanisms that occur simultaneously at different spatial scales and biological contexts. These mechanisms include chromatin looping to allow contact between cis regulatory elements like enhancers and promoters, so recruitment of the transcriptional machinery can be facilitated; non coding RNA modulation of gene expression and silencing; methylation of DNA as an steric impediment to TFs binding, thus silencing transcription; and histone post transcriptional modifications that contribute to the electrostatic landscape of chromatin and encourage (e.g., H3K4me1 and H3K27ac that mark enhancer sequences or H3K4me3 that protects promoters from DNA methylation) or deter (e.g., H3K27me3 and H3K9me3 that mark heterochromatin) transcriptional processes. To date, the cooperativity and feedback between them remains to be fully characterized. (Image created with BioRender, https://biorender.com/).

2.4. Post-Transcriptional Processing

Once a primary mRNA has been synthesized by the transcriptional machinery, it is often subject to cleaving processes in which specific exons and introns are discarded or included in a final transcript in order to generate function-specific versions of mRNAs. This post-transcriptional processing is called splicing and variations in the way it is carried out are known to be cancer-related [37]. This is explained by the fact that splicing variants give rise to different proteins that display different, sometimes even antagonistic, biomolecular behaviors [38]. Therefore, splicing variants in cancer are finding their way into clinical interventions [39].

2.5. Chromosome Aberrations and Chromosomal Instability

Aneuploidy, is a chromosome instability alteration characterized by amplification or deletion of entire chromosomes or chromosomal sections and it has been found to be frequent in cancer [40–43]. Such enormous alterations to genes ratios are inevitably associated with gene expression modifications [44,45]. For instance, the region Chr8q24.3 suffers amplifications in different tumors [32,46–48] and it has also been reported in connection to oncogene overexpression and poor prognosis [49–51]. Specific regions have been described as significantly altered in different breast cancer subtypes, exemplified by the work of Smid and coworkers [51] who were able to characterize 313 primary breast tumors by their chromosomal instabilities.

2.6. Non-Coding RNAs

Non-coding RNAs (ncRNAs) are gaining increasing attention as central players behind transcriptional regulation. A decade-long research program has supported the role of ncRNAs as key drivers in cancer [52]. There are at least two relevant classes of regulatory ncRNAs and their main characteristics are their length: long non-coding RNAs are transcripts longer than 200 nucleotides, while micro-RNAs (miRs) are around 22 nucleotides long.

The phenotype-specific regulatory abilities of lncRNAs have been proven [53] and their specific role in several types of cancer has been extensively investigated. For instance, MALAT1, NEAT1, LED, HOTAIR, and MEG3, all of which are lncRNAs have been reported to target cancer-specific pathways [54–59]. miRs are smaller molecules that serve as posttranscriptional gene regulators. To date, there is a large collection of studies supporting the potentially oncogenic and tumor-suppressive c roles of miRs [60–65] regulating oncogenes or tumor suppressor genes. Given that their activity tends to be context-dependent, miRs involved in cancer gene regulation are termed oncomirs [62,63,66], regardless of whether their activity is pro- or anti-tumoral. Resorting to miR-gene regulatory networks, work from our own group [67] has shown that overexpression of members of the miR-200 family triggers a switch controlling epithelial-to-mesenchymal transition in breast cancer, via downregulation of VIM, ZEB1 and ZEB2, accompanied by aberrant TGFB signaling controlled by overexpression of the miR-199 family.

2.7. RNA Stability and Transport

Once the mRNA transcript has been synthesized and preprocessed by splicing in the cell nucleus, it must be transported through the nuclear membrane to the cytoplasm where translation can be initiated at the ribosomes. The import and export of molecules from the cell nucleus are controlled by a family of nucleoporin proteins that assemble at the nuclear membrane to form nuclear pore complexes. mRNA transcripts associate with nuclear transport proteins called exportins to achieve export from the nucleus through nuclear pore complexes. The process of mRNA export is regulated by the stability of the transcripts, both in their free and bound-to-exportins form and by the kinetics of the associated transport process [68,69]. These processes are known to be anomalous throughout cancer establishment and maintenance [70,71], and some of the involved molecules have been proposed as cancer biomarkers and therapeutic targets [72–74].

Aside from these well know features, there are other genomic factors that are now beginning to be analyzed in connection to gene expression regulatory programs. Some of these were already considered to participate in regulating gene expression, however, it was until the availability of trustworthy experimental techniques and computational methods that their unequivocal contribution to regulation was established. In this subsection, we address some of the most relevant and novel elements in gene regulation and their connection to cancer biology.

2.8. Chromothripsis and Other Catastrophic Chromosomal Events

Chromothripsis is another cancer-associated molecular phenotype in which several chromosomal rearrangements occur simultaneously in localized regions [75,76]. Particular DNA bridges (termed

micronuclei) may be related to damage of the nuclear envelope leading to these broad chromosomal rearrangements [77,78]. Chromothripsis is often concomitant with other complex chromosomal events, such as aneuploidies [79,80], localized regions of hypermutation (kataegis) [81,82] or chromoplexy (abundant DNA translocations and deletions that appear independently in multiple chromosomes) [83,84]. The co-existence of such abnormal chromosome architecture may obey a common origin. Initially, it was considered that chromothripsis appears in approximately 2–3% of the cancer cases [75], though recent studies argue that chromothripsis may be present in up to 50% of cases in some cancer types [85].

2.9. 3D Structure and DNA-Associated Complexes

Epigenomic modulation is key to regulating gene expression. Aside from chemical modifications to DNA and histones or local chromatin accessibility, higher-order chromatin structures have also resulted to be associated with transcriptional regulatory processes. Recently, it was discovered that other functional molecules involved in the epigenomic control of gene expression are SATB1, CTCF, and cohesin. These proteins are known to be relevant in tumor development and evolution [33]. For example, Lee and coworkers [86] have reported that depletion of the CTCF protein induces selective cell death of cancer cells via p53. CTCF is specifically involved in DNA spatial reconfiguration and in the formation of the so-called topological associated domains.

Histone deacetylases (HDACs), DNA-methyltransferases (DNMTs) and chromatin remodelers [87], also exert action upon the gene regulatory program by adding or removing functional chemical groups to chromatin, thus altering the spatial configuration of DNA. These dynamical processes have well-documented roles in cancer phenotypes, HDACs have been reported to participate in malignancy [88–91] and to act in concert with miR regulators [92]. HDACs have actually turned into promising epigenetic therapeutic targets [89], their inhibition has been able to potentiate immunotherapy effects on triple-negative breast tumors [90], and to slow down tumor growth when mTORC1 and *α*-estrogen are also inhibited [88]. Regarding DNMTs, their activity throughout tumorigenesis and their potential as cancer biomarkers has been discussed extensively by Zhang et al. ([93]).

3. Omic Developments to Study Cancer Genomics

During the twentieth century several features of cancer, including cell-cycle checkpoints deregulation, oncogenes, tumor suppressor genes, genetic instability, and cancer gene interactions were identified and studied with cytogenetic and classical genetic techniques [94]. As a result of this knowledge, new questions surfaced about the regulatory context of genes that are significant to cancer processes, and two things that have been crucial to advance cancer research in the last twenty years into an integrative field, the availability of the human genome sequence [95] and the development of high-throughput sequencing methods. Today it is recognized that cancer genomes, transcriptomes, and epigenomes are all key to understanding cancer biology in a detailed molecular level, which has led to several combinations of different experimental methods with high-throughput sequencing to process cancer samples and generate quantitative data of the molecules that result from the genomic mechanisms behind malignancy. In this section, we describe genomic and epigenomic methods that are currently used to investigate the cancer phenotype. Throughout, we discuss the goals of the methodologies, their general workflow, data analysis, and challenges.

3.1. Sequence-Based Methods

3.1.1. DNA Sequencing

DNA sequencing is used to determine the presence of potentially pathogenic (e.g., loss of heterozygosity) genetic variations (see [43]), including single nucleotide variations (SNVs), insertions and deletions (INDELs), copy number variations (CNVs) and genomic rearrangements. The aim of a particular DNA sequencing assay can be either detecting inherited germinal variation or characterizing

somatic variation in tumor samples. In the former case, DNA is usually extracted from buccal swabs or blood samples. In the latter, it must be extracted from tumor cells, which are habitually derived from paraffin-embedded biopsy samples [96] or fresh-frozen tumor tissue [97]. However, there have been efforts lately to achieve high-quality DNA sequencing from minimally-invasive liquid biopsies [98].

Experimental Strategies

Although whole-genome sequencing has the advantage of uncovering the whole set of variations in a given tumor, sequencing a subset of the genome such as whole-exome or targeted regions is sometimes favored due to the lower cost and input DNA requirements, particularly in clinical settings [99]. Experimental design can accommodate a broad range of objectives, from uncovering key driver mutations and their frequency in large groups of individuals presenting a specific tumor type and identifying pan-cancer significant variation [100], to detecting germline oncogenic variants or guiding therapeutic options based on surveying tumor somatic variation in an individual [101].

After DNA extraction and quantification, library preparation can follow different strategies [102], for whole-genome sequencing the DNA is fragmented and sequencing adapters are ligated to it. However, when the goal is to sequence the exome or selected regions it can be done through 'capture' where a pool of diverse oligonucleotides are used to hybridize the genetic material of interest [103] or through 'amplicons' that are primers designed to flank the target regions and amplify them through Polymerase Chain Reaction (PCR) [104]. Although the obtained data from the different platforms are equivalent (see [105] for a comprehensive review), Illumina (see sub-section 'High-Throughput DNA Sequencing' below) is currently the most used DNA sequencing technology.

Variant Discovery Analysis

Once the sequencing reads are properly mapped there are two procedures that correct for technical biases and should be applied to the data before carrying out variant calling. The first one is Indel Realignment and it is a de novo assembly of reads from regions detected to probably contain an insertion or deletion, but are mistaken for different SNVs close to each other, indel realignment algorithms include ABRA [106] and IndelRealigner from the Genome Analysis Toolkit (GATK) [107]. The second procedure is Base Quality Score Recalibration (available as a GATK tool) and it uses machine learning to analyze the sequencing read's Phred scores and corrects their values accounting for systematic errors from the sequencer machine, this avoids over- and under-estimations from the subsequently used variant caller algorithm.

Germinal Variants

Variant calling is the identification of loci in the genome where the data that is being analyzed presents differences compared to the reference genome. One of the most used variant callers for SNVs and INDELs is HaplotypeCaller [108] from GATK, which evaluates individually each site of potential variation using as expected model a De Bruijn graph [109] built from the sequence of the reference genome and comparing the sequenced reads against it to obtain the list of observed possible haplotypes and variants, the best-supported genotypes throughout the samples are obtained through a Bayesian approach. Other algorithms include MAQ [110], Freebayes [111], which is useful when analysing amplicon sequencing data, BIC-seq [112] and FermiKit [113], that can additionally handle CNVs detection.

These algorithms output files in the Variant Call Format (VCF) [114] with 'raw' variants, that should be subject to filtering and annotation (see variant filtering and annotation). Germline variants are usually investigated to assess the risk of inherited cancer susceptibility and identify affected pathways [115], however it should be noted that the frequency of pathogenic variants can be different throughout populations [116,117] and also that most cancers are 'sporadic' as opposed to familial [118].

High-Throughput DNA Sequencing

Illumina sequencing by synthesis   Illumina sequencing requires that adaptors are added through PCR to both ends of the DNA fragments, oligonucleotides complementary to the flow cell where sequencing takes place are also added.   Additionally, indices may be included too when different samples are sequenced on the same run. The flow cells are made of lanes that are coated in oligonucleotides that hybridize and physically attach the DNA to be sequenced. Once the DNA library is loaded into the flow cell, each bound fragment is amplified clonally, generating clusters of around one thousand copies of each single-stranded DNA molecule, which improves the signal. Then, all the fragments are extended in parallel by polymerase enzymes one nucleotide at a time using nucleotides (dNTPs) with an attached fluorophore which also serves as a reversible terminator. In each extension cycle, the four dNTPs are added separately and the flow cell is imaged to identify the base that was incorporated through the fluorophore emission wavelength and intensity. The base calls are made in real-time from the images by the machine's internal software. This extension cycle is repeated n times, resulting in sequencing reads of n bases. Currently, Illumina sequencers output a maximum read length of 300 base pairs.

Sequencing data processing   The first step in the analysis of sequencing data in assessing the quality of the raw reads, which are usually in FASTQ format [119], this task can be performed with readily available tools such as FASTQC [120] or MultiQC [121] that summarize the data attributes with descriptive statistics of Phred quality [122], nucleotide content and sequence length distribution, among others. If required, adapter sequences are removed and reads are trimmed to eliminate low quality and ambiguous bases, in this case too, software like Trimmomatic [123] or Cutadapt [124] can be used. The next step is the alignment of the sequencing reads to a reference genome and there is a vast offer of tools that accomplish this job [125] and output a sequence alignment map (SAM/BAM) file [126]. Afterwards, some standard data cleanup is done, including ordering the mapped reads by chromosome and position and marking the reads that are determined to be PCR duplicates by their identical start and end positions. This last step is omitted when the used library strategy is amplicon sequencing, which by definition have very similar genomic positions.

Somatic Variants

Calling somatic variants can be challenging due to a number of factors that include, heterogeneity in tumor samples [127], the high diversity among driver mutations even between cases of the same cancer type, and the fact that most genetic changes present in cancer are not the drivers of malignancy. Additionally, a high depth of sequence coverage is usually required to detect somatic mutations accurately. Several somatic variant callers with diverse algorithm strategies exist nowadays, Bohannan and Mitrofanova provide a review of these tools in the context of experimental cancer biology [128], while Zare et al. evaluate the detection performance of variant callers [129].

The analysis of somatic variations in whole-genome and whole-exome data has led to the identification of recurrent and significant pathogenic alterations, including SNVs, INDELs and CNVs [130,131], which has helped elucidate patterns of genomic anomalies in cancer phenotypes. Moreover, the work of consortia like The Cancer Genome Atlas (TCGA) [132], now integrated with Genomic Data Commons (GDC) [133], found in different cancer types the most common driver mutations, as well as genomic profiles associated with prognosis, including hypermutation, microsatellite instability, content of CNVs, mutational burden, inactivating mutations in chromatin modifiers, DNA repair pathways and immune system genes [134]. A great advantage of efforts like GDC is that the data generated by them has great quality and is publicly available to the scientific community, which can enhance discovery through reanalysis [135].

Variant Filtering and Annotation

Raw variants are filtered to remove false-positive calls. A set of criteria including alignment quality, depth, read support of the reference versus alternative alleles and strand bias is used to calculate the probability that each variant call is correct. There are also machine learning methods, such as VQSR from GATK, that build estimators using data sets of previously known true variants, however, they require thousands of raw variant calls to operate on. Importantly, SNVs and INDELs are filtered separately because their properties are different. Afterwards, variants are annotated using genetic variation databases like 1000 genomes, gnomAD, or using databases of functional categories and predicted effects just as snpEff [136], dbNSFP [137] or ClinVar [138]. Furthermore, putative predictions about variants can be made according to proximity to regulatory regions, coding regions, and splicing sites, among others.

3.1.2. RNA Sequencing

The introduction of microarrays to analyze gene expression [139] of thousands of genes simultaneously in an unbiased fashion unlocked the development of discovery-driven research and with the advent of RNA-sequencing [140,141] it emerged that studying the molecular phenotype of cells through the quantification of their whole-genome transcriptome is a very powerful tool to approximate the functional state of cells and large projects have been established to characterize not only transcriptomes from diseased phenotypes [142,143], but also from healthy ones [144].

RNA-seq experiment   RNA-seq methods are based on the conversion of extracted and fragmented RNA to complementary DNA (cDNA) by a reverse transcriptase using random primers. The obtained cDNA, which is double-stranded, can be used then to build a sequencing library. It is common practice to take steps that enrich for messenger RNA "poly(A) enrichment" or deplete ribosomal RNA, due to the overwhelming abundance of the latter in cells and usual interest in the coding transcriptome. Actually, the number of RNA-seq protocol variations nowadays is staggering (recently reviewed in [145]) because they are suited to a variety of specific research goals.

RNA-seq analysis   The main principle of RNA-seq analysis is that the number of sequencing reads for a given transcript, usually explicitly ligated to a gene, is a proportional measure of its expression level. RNA-seq is benefited from paired-end reads and depending on the objective of the assay a depth from 50 million reads, in the case of differential gene expression analysis, to 200 million reads for de novo transcriptome assembly.

The alignment of reads is usually to a reference transcriptome, however in RNA-seq from tumor samples, this step is sometimes done through de novo assembly because the alterations that are common in cancer give rise to transcriptional differences compared to the regular genome. Reads that map to multiple locations are usually filtered out since it is extremely difficult to distinguish their origin.

Assigning reads to genes or transcripts and quantifying them [146] represents a critical step for the final results and ideally should consider transcript variants, however, this is not always possible because all splice junctions are not necessarily sequenced.

Once the transcript counts are obtained, it is necessary to normalize the data to correct for biases that arise from reading depth, GC content, and intrinsic noisy differences between samples. Normalization methods have been developed since the introduction of RNA-seq and selecting one should consider the experiment attributes and goals [147], however, it is common practice to apply different normalizations to the data and compare the results to ensure the best possible results.

Finally, one of the most common goals of RNA-seq is differential gene expression (DGE) analysis, which seeks to determine the genes that are over and under-expressed in a condition compared to others. Available DGE methods and tools have been subject to evaluation [148,149] in the interest of assessing their fidelity and replicability, as well as guiding analysis decisions. Other RNA-seq analysis

applications, provided the experimental design is suitable, include detecting transcript abundance from alternative splicing events, detecting gene fusions and evaluating the expression of transcripts that contain SNVs, among others.

3.2. Epigenomic Modifications

3.2.1. DNA Methylation

DNA methylation is an epigenetic alteration with a role in transcription regulation, gene silencing, and chromatin organization [150]. DNA methyltransferases (DNMTs) add a methyl group on the fifth carbon of the cytosine ring, 5mC, avoiding TF binding. Instead, DNA methylation creates a binding site for Methyl-CpG-binding domain (MBD) proteins, which in turn recruit histone-modifying complexes [36]. Bisulfite causes the differential deamination of cytosine and 5 mC. While cytosine deamination turns the base into uracil, 5mC deamination is still detected as cytosine [151]. Then, bisulfite treatment followed by sequencing or microarray reading can effectively identify the modification.

Illumina Infinium arrays are the most common detection method [152]. The most recent one, MethylationEPIC BeadChip, can interrogate over 850,000 methylation sites across the genome. Methylation BeadChips measure trough fluorescent dyes the intensity of the methylated and unmethylated signal. Afterward, data has to go through quality control, background correction, and normalization. Illumina's GenomeStudio software cope with the whole preprocessing but can be customized through Bioconductor packages such as IMA [153], Minfi [152], and MethyLumi [154]. Biologically variant regions like SNPs and sex chromosomes should also be filtered [155].

Though GenomeStudio provides an internal control normalization, results suggest that peak-based correction and between arrays quantile normalization plus Beta-mixture quantile normalization within arrays may outperform others [156]. The `beta` values outputted by this pipeline must be transformed to `M` values and corrected for batches, but are otherwise ready to be used. `beta` values give account for hypo and hypermethylated regions, hence, a bi-modal distribution is obtained. `M`-values transformation normalize bi-modal distribution, in order to perform further analyses.

3.2.2. Chromatin Immunoprecipitation Followed by Sequencing (ChIP-seq)

Normal transcriptional programs in healthy cells are largely controlled by TFs encoded in the human genome [157], that bind specific regulatory sequences in the genome and activate transcription of their associated genes by recruiting cofactors, chromatin regulators, and the RNA polymerase II (RNAPII). Moreover, the main component of chromatin are nucleosomes that consist of histone proteins that can be modified post-translationally by covalent addition of a functional group to amino acid residues in their C- and N- terminal domains, which has a role in transcriptional control (reviewed in [158]).

Therefore, the patterns of TF binding and chromatin modifications have been investigated for their role in gene expression changes and clinical outcomes in cancer [159–164]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) [165,166] is a genome-wide method that detects in vivo binding events of TFs, positions of histone modifications or genomic presence of chromatin-associated non-histone proteins.

A validated highly specific antibody to the protein or modification of interest is necessary and depending on the abundance and stability of the target, from 10^4 to 10^7 cells. Frozen tissue may be subject to ChIP-seq, however, attention must be given to the homogeneity of the sample, because it can affect the target signal. The Encyclopedia of DNA Elements (ENCODE) consortium has developed guidelines and best practices for ChIP-seq assays [167] that can be useful in planning experiments.

ChIP-seq experiment　Cells are crosslinked using formaldehyde [168], which fixates DNA and proteins in vivo, the chromatin is sheared through sonication (although protocols using enzymatic

digestion are available) that is optimized to generate fragments ranging from 100–600 bp in length and the antibody is used to immunoprecipitate the DNA fragments bound to the target of interest. Finally, after crosslink reversal to allow for DNA purification, a sequencing library is prepared and sequencing is used to profile the events of interest across the genome. There are three types of controls for a ChIP-seq experiment, (i) input DNA, which is taken from the sample before immunoprecipitation (IP), (ii) mock IP DNA, from IP without using an antibody, and (iii) nonspecific IP DNA, from IP with an antibody against a non related protein, commonly IgG, additionally, to account for inter-sample variation a replicate experiment is recommended. Single-end sequencing is often used in ChIP-seq experiments, with the benefit of paired-end sequencing being increased mappability, while the optimal sequencing depth is in the function of the type of the experiment's target and is around 30 million and 60 million mapped reads for DNA-binding proteins and histone marks, respectively [167].

ChIP-seq data analysis　Once the raw sequencing reads have been aligned to the reference genome and appropriately filtered, peak calling is performed to identify genomic locations enriched for the targeted protein. The goal of peak calling is to find regions flanked by reads on both ends (5′ and 3′), deemed candidate peaks, and evaluate them statistically versus a background model (reads from the control experiment or expected values in the matching region) to assess their significance.

Some of the most used peak callers (for a comprehensive review see [169]) are MACS [170,171], SPP [172], and PeakSeq ; the first two use a Poisson distribution to model the data and calculate the cutoff above which a peak is determined significant, while SPP uses a binomial distribution and considers the mappability of the regions.

It is important to stress out that peak calling is sensitive to user-defined threshold values. The output of the analysis is a list of the genomic regions where significant evidence of binding/presence exists, along with *p*-values and false discovery rates (FDRs). Downstream analyses include protein binding motif discovery, differentially enriched regions analysis and integration with RNA-seq gene expression data.

3.2.3. Methods to Assess Open Chromatin

Regions where nucleosomes are sparse and physical access to the DNA sequence is enabled are identified as open chromatin. Chromatin accessibility is a dynamical and complex framework modulated by diverse elements, including nucleosome occupancy and turnover rate, histone modifications, ATP-dependent chromatin remodeling complexes, and even TF binding [26,173]. Open chromatin has emerged as indicative of transcriptional regulatory potential or activity across the human genome because most of the TFs analyzed to date bind within open regions [28].

Particularly in the context of cancer, a large survey by the TCGA analysis network [29] revealed a general increase of open regions in cancer compared to healthy phenotype, a connection between susceptibility genetic variants and accessible chromatin, as well as grouping of breast cancer and kidney renal carcinoma samples by the presence of open chromatin peaks, that turned out to be accompanied by gene overexpression and clinical implications. Moreover, studies that interrogate chromatin accessibility have helped to uncover specific TFs that play a role in the gene expression patterns of tumor samples [174]. Hence, assaying open chromatin can help researchers gain knowledge of the processes deregulated in the transition of normal cells to cancer.

Several methods exist to assess open chromatin sites, DNaseI-hypersensitive (DHS) sites derived from DNase-seq [175] coincide with nucleosome-depleted regions and Micrococcal Nuclease sequencing (MNase-seq) [176,177] experiments help determine positions where nucleosomes are present, while both techniques rely on enzymes to digest unbound, open chromatin, formaldehyde-assisted Isolation of regulatory elements sequencing (FAIRE-seq) [178] takes advantage of the different chemical properties between protein-bound DNA and nucleosome-depleted DNA. However, since the introduction of the assay for transposase-accessible chromatin using sequencing (ATAC-seq) [179], most investigations favor it, due to its simplicity and low DNA-input requirements,

and there is even a follow-up method that couples ATAC-seq with high-resolution microscopy [180]. It should be noted that the data produced by these methods present a high correlation, and it has been proposed that the differences arise mostly from sequencing biases [181].

3.2.4. Transposase-Accesible Chromatin Sequencing

ATAC-Seq experiment　　ATAC-seq is a clever method that leverages the activity of a transposase called Tn5 to simultaneously fragment and tag all the accessible DNA. The enzyme is pre-loaded with sequencing adapters, allowing for direct purification and amplification of the fragmented and tagged DNA, followed by sequencing. Usually, two replicates are used to discern biological signals from noise and while appropriate sequencing depth can depend on the target cells, the original protocol [179] suggested around 50 million aligned reads per sample.

ATAC-Seq data analysis　　After the alignment to the reference genome and corresponding filters, peak calling is carried out to identify statistically significant enrichment of reads throughout the genome. Essentially the same peak callers employed in ChIP-seq analysis can be used for this type of data. Importantly, a blacklist of ATAC-seq peaks from ENCODE [182] is available to filter the results. When multiple samples are available a custom strategy to obtain high-confidence open chromatin peaks can be developed, for example, using criteria of presence in more than n samples or normalizing peak significance score and using a threshold thereafter. With the ultimate goal of characterizing and better understanding which regulatory landscapes may underlie the studied phenotypes, downstream analyses to an ATAC-seq peak set include annotating them with data from external sources [183] to find coinciding histone marks and/or DNA-binding proteins, searching for enrichment of TFs binding motifs [184] or footprinting analysis to derive a measure of TF occupancy [185,186].

3.3. Chromosome Conformation Capture (3C Methods)

In 2002 Dekker et al. introduced an innovative technique called 3C [187] to measure at high resolution the frequency at which any two genomic loci, for example, enhancer and promoter, were found together in the nuclear space. This opened exciting avenues in the investigation of the three-dimensional conformation of the eukaryotic genome, whose structured nature had been recognized [188], but was almost exclusively studied with microscopy methods [189]. The 3C technique was followed by the development of assays to quantify chromatin interactions between all the loci within a defined region at the Megabase scale ("5C" [190]), between a viewpoint and the rest of the genome ("4C" [191]), and the genome-wide interactions ("Hi-C" [192]).

Soon thereafter, general patterns of the conformation and interactions within the chromatin framework emerged, including transcriptionally-repressed lamina-associated domains [193,194], A/B compartments that roughly correspond to euchromatin and heterochromatin [195], topologically associating domains (TADs [196]) that interact mostly within themselves and chromatin loops between regulatory sequences [197] formed by CTCF sites in convergent orientation. Proteins involved in the architecture of the 3D chromatin structure were also identified [198] and today it's well accepted that genome organization is linked to a myriad of functional processes, such as developmental regulation, gene expression or silencing throughout the cell cycle, DNA repair and deregulation in disease phenotypes.

The role of the 3D organization of the genome in genetic regulation is an ongoing and quite active research field, it has spawned variations of the C methods that are tailored to regulatory genomics questions, for example, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET [199]) to detect chromatin interactions mediated by a specific TF or protein, capture Hi-C (CHi-C [200]) to identify interactions between specific regions of interest and the rest of the genome, Hi-C methods to achieve kilobase resolution [201,202], to obtain contact maps from clinically available samples [203] and even to unmask the processes behind chromosome interactions through the quantification of their

stability [204]. In spite of the buoyant progress in the research of chromatin's functional structure, the characterization of its direct relationship to transcriptional regulation is work in progress [205–208].

Through the application of the C methods it was identified that the three-dimensional architecture of chromatin is correlated to the presence of somatic alterations in cancer [209–211], and even though Hi-C measures interaction frequency and not physical distance [212], the former can be a location predictor of chromosomal rearrangements and CNVs in cancer [213,214]. These alterations of the DNA sequence, which are typical in tumors, can lead in turn to disruption of the chromatin framework in which regulatory interactions take place [215], resulting in oncogene activation due to aberrant contacts between a foreign enhancer and their promoter [216–218].

In light of this, there have been efforts through C methods to identify non-coding alterations that impact gene expression and drive cancer progression [219] and to profile the regulatory loops that impact transcriptional programs in a clinical research context [220]. Indeed, when ChIA-PET was used to investigate the relationship between TFs mediated by hormones, namely the estrogen-receptor-alpha (ER-alpha), chromatin interactions and the transcriptome in the context of breast cancer, it was suggested that the coordinated regulation of sets of genes could be aided by their co-localization in space mediated by the RNA Polymerase II and that the perturbation of this arrangement can lead to transcriptional alterations of even secondary genes [199,221]. Later, two studies [34,222] reported that upon activation of the ER-alpha, transcriptional changes entail coordinated responses at the chromatin structure level. Finally, Hi-C experiments of genome-wide chromatin interactions have identified a switch from B to A compartments accompanied by up-regulation of their resident genes in breast cancer [223] and B-cell lymphoma [224].

3.3.1. Genome-Wide Chromosome Conformation Capture

Hi-C experiment   The Hi-C assay begins with the fixation of the DNA using formaldehyde to preserve the cellular conformation of the chromatin. Afterward, a restriction enzyme that leaves sticky ends is used to digest the DNA, it is important to note that the resolution of the data will depend on the frequency of the enzyme's restriction sites in the genome. The overhanging ends are filled with biotinylated nucleotides and religation is promoted between DNA molecules that belong to the same interaction complex; this step can be performed under dilute conditions [192] or in the cell nucleus (in situ Hi-C [201]). Finally, the crosslink is reversed, followed by sonication and biotin pull-down using streptavidin to enrich for the chimeric DNA of interest, which is amplified and sequenced.

Hi-C analysis   There are two excellent reviews [197,225] that provide guidelines to analyze data from Hi-C experiments. Although there are several options at each step of the analysis, we describe one of the most common workflows. Briefly, the sequencing reads should be cleaned, specifically, they should be scanned for ligation junctions (two restriction sites facing each other) and trimmed to improve mappability. The paired-end reads should be aligned to the reference genome separately, because the pairs do not correspond to contiguous sequences in the genome; a widely used strategy is iterative mapping in which all the reads are trimmed to n nucleotides and aligned to the reference, the reads that do not align are extended by n nucleotides in the 5' direction and mapped again. Aligned read pairs are assigned to the nearest restriction fragment in the genome and filtered to retain only informative pairs. Afterward, the genome is binned in fixed-length windows and the read pairs are assigned to them which yields a contact matrix that must be normalized before meaningful interactions can be derived. Several methods to normalize Hi-C matrices exist, explicit-factor correction [226] considers known biases as mappability and GC content to calculate the probability of contact, while matrix balancing [227] is an implicit correction method based on the Sinkhorn–Knopp balancing algorithm [228] that results in a Hi-C normalized matrix with all the rows adding up to the same quantity. It is important to note that matrix balancing requires that the raw matrix is filtered to mask bins with very few read pairs. Once a Hi-C matrix is normalized, there are different methods to obtain compartments, TADs and significant contacts from it, they are richly discussed in [197]. Afterward,

the data can be integrated with data from location-based methods, including ChIP-seq and ATAC-seq, to profile sets of interactions that are considered of interest.

3.3.2. ChIA-PET

The Chromatin Immunoprecipitation Analysis by Paired-End Tag Sequencing combines formaldehyde-crosslinking to obtain chimeric DNA molecules that are in nuclear proximity and enrichment for a subset of loci by means of a specific antibody to probe regulatory chromatin interactions mediated by a protein of interest. It should be noted that while ChIA-PET can provide de novo, unbiased short- and long-range chromatin interaction profiles, the protein involved must be previously suspected to participate in mediating these functional contacts. Usually, the assayed protein is a TF, a form of the RNA polymerase (i.e. initiation or elongation RNAPII) or a structural protein (e.g., CTCF). Some applications of ChIA-PET include identification of chromatin contacts by a TF between promoters and other regulatory sequences, evaluation of differential interactions in a myriad of phenotypes (e.g., developmental stages, response to cell signaling, disease processes), and notably, characterizing the spatial nature of miRNA genes transcriptional regulation [229].

ChIA-PET experiment The ChIA-PET experiment, as described in the original protocol [230], begins by crosslinking the DNA in the nucleus followed by sonication to lyse the cells and release the fragmented chromatin. A highly-specific and validated antibody is used to immunoprecipitate genetic material bound by the protein of interest, this increases the specificity of the library and reduces background noise. As in other IP-based protocols, a sufficient number of starting cells is required (the original protocol estimates 100 million cells) to achieve adequate library complexity. After the DNA-protein complexes are immunoprecipitated, biotinylated oligonucleotide half-linkers that contain a recognition site for the MmeI restriction enzyme are ligated to the free ends of the chromatin fragments. Two half-linker variants (A and B) with specific nucleotide barcodes to distinguish them are used; prior to ligation the ChIP chromatin is divided into two aliquots and each is ligated with half-linker A or B. Both fractions are integrated under dilute conditions to promote DNA inter-ligation, thus creating three types of junctions (heterodimer AB linkers and homodimer AA/BB linkers) that help distinguish non-specific ligation products. The crosslinking is reversed, the DNA is purified and the type IIS MmeI enzyme is used to release the tag-linker-tag constructs that are captured using streptavidin and finally used to build the sequencing library.

ChIA-PET data analysis The analysis of ChIA-PET data is complex and different methods have been proposed to extract meaningful contacts from the million reads usually obtained by an experiment, ChIA-PET tool [231] solves the problem by means of a hypergeometric model, while model-based Interaction calling from ChIA-PET data (MICC [232]) employs a hierarchical mixed probability model and ChIAPoP [233] relies on a Poisson model. However, the first step in the analysis involves classifying the read pairs in heterodimer and homodimer linkers and aligning the tags to the reference genome plus the usual post-mapping filters. Then, the overlapping read pairs are merged and peak-calling is performed to locate enrichment loci within the genome, this results in "paired" peaks, since they are located at different positions in the genome. The paired peaks correspond to pairs of loci connected by the immunoprecipitated protein, although it should be noted that ChIA-PET cannot determine if the assayed protein is directly responsible for the interactions or if it simply present e.g., as part of a protein complex.

3.4. Genome and Epigenome Editing

Studying in vivo cancer biology is a daunting task complicated by the fact that the processes governing the cancerous abnormal regulation of the genome are not yet fully characterized. For example, the onset and initial steps of tumorigenesis are hardly ever observed or probed, the contribution of most of the detected variants to different cancer types is not clear, and the tumor

microenvironment cues influence dynamically the abnormal regulatory processes [234]. Therefore, even when formidable advances in cancer research have been accomplished through the application of the high throughput methods outlined above, perturbation experimental approaches are required to dissect the genotypes underlying most cellular and molecular phenotypes of carcinogenesis.

Perturbation screens [235] act at the DNA, RNA or protein levels to gain insight about gene cellular functions and essentiality, but also to better understand the intricate regulatory mechanisms of the genome and find drivers of disease [236] as opposed to its secondary manifestations. While earlier strategies like retroviral insertion/transposon mutagenesis produced random perturbations, nowadays sequence-specific, genome-wide methods are available to perform DNA sequence [237], transcriptional [238,239] or post transcriptional [240,241] perturbation screens. Directed assays also enable multiplexed or pooled screens [242] and a higher degree of precision. Particularly, the development of clustered regularly interspaced short palindromic repeats (CRISPR) editing [243] revolutionized the entire functional genomics field and has become an obvious choice [244] to perturb gene activity due to its minimal interference with endogenous conditions, versatility and relative simplicity.

3.4.1. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas

CRISPR is the only known instance of acquired immunity in prokaryotes. It consists of a CRISPR locus in the bacterial genome that is actually foreign in its origin and usually corresponds to the genetic sequence fragments of a bacteriophage or plasmid. When a bacteria has acquired CRISPR DNA [245], it can be transcribed and processed into a mature crRNA that is used by CRISPR associated (Cas) proteins to detect and cleave pathogenic DNA from similar viruses upon reinfection. The Cas proteins are DNA endonucleases that undergo a conformational change that allows cleaving as a consequence of the base-pairing (around 20 nucleotides) between the crRNA and the foreign DNA. Additionally, The action of CRISPR-Cas depends on the presence of a protospacer adjacent motif (PAM) in the target DNA that has been theorized as a bacterial non-self recognition system.

CRISPR-Cas genome editing in cancer research Several research groups identified and characterized elements of the CRISPR-Cas system [246]. However, when a team led by Jennifer Doudna and Emmanuelle Charpentier engineered guide RNA to direct Cas9 cleavage in a sequence-specific manner [243], they rendered a programmable genome-editing tool unparalleled in power to silence or activate specific genes. Although the initial employment of the Cas nucleases was inducing double-strand breaks in specific loci that result in inactivating frameshifts, currently numerous alternatives have been developed to modulate gene expression [247]. Briefly, the nuclease domains from Cas proteins can be deleted while preserving targeting function through the crRNA. This can be exploited through protein fusions with other enzymes that are transcriptional activators or repressors, epigenetic effectors that modify chromatin and even enzymes to induce point mutations [248]. Additionally, imaging techniques that couple catalytically inactive Cas9 to fluorescent proteins have also been developed with potential applications in tracking endogenous RNA activities and post-transcriptional regulation [249].

CRISPR-Cas has generated promising strategies in translational cancer research [250,251]: from uncovering relevant mutations to establishing exemplary animal models and carrying out pooled perturbation assays (see the 'CRISPR Screening Assays' sub-section below), but it has also put forward impressive therapeutic approaches like correcting faulty DNA sequences or enhancing immune system cells to fight tumors. Importantly, methods that use CRISPR interference to probe the target promoters of enhancers [252] could help clarify the participation of non-coding mutations in cancerous gene regulatory processes, albeit different mechanisms of enhancer activation should be taken into account [253].

3.4.2. CRISPR Screening Assays

Pooled screens The goal of a pooled CRISPR screen is to assess the effects of genome-wide perturbations simultaneously in a single assay (in contrast to array screens where perturbations are assessed individually). Often, the idea is to establish a causal relationship between a genotype and phenotypes of interest or to test the functions of genes under different contexts. For example, a pooled screen can be used to evaluate phenotypes like cell death, drug resistance or transcriptome changes.

Broadly, a CRISPR screen assay is performed as follows. However, it should be noted that considerable effort needs to be put into experimental design and subsequent experiments planning in order to obtain reasonable results [244].

- A library of perturbations is built according to criteria largely determined by the assay's goal.
- The library is introduced into a population of cells using a vector that could be a lentivirus or retrovirus.
- The cell population is screened to select the subset of cells that present the phenotype of interest.
- PCR is usually employed to amplify the sequences from the genomic DNA and identify which perturbations gave rise to the phenotype.
- Finally, the relative abundance of the perturbation library can be quantified by DNA sequencing.

Technical limitations Standardizing the perturbation library introduction is important to obtain reproducible results. Additionally, the delivery system should be thought out carefully, especially in the in vivo models to avoid potential immune responses against CRISPR [254] or plain delivery failure. Finally, a highly relevant technical limitation is the off-target activity of Cas proteins that can edit wrong loci, notwithstanding CRISPR-Cas' unmatched precision [255]. It has been proposed to leverage different PAM sequences and their specificities to alleviate this problem [256,257] and a sequencing-based method has been developed to detect genome-wide double-stranded breaks by CRISPR RNA-guided nucleases [258]. More recently, anti-CRISPR molecules derived from phages have been used to block CRISPR-cas activity in mammalian cells [259], however, the authors had in mind the development of controllable synthetic gene circuits rather than CRISPR pooled assays.

4. The Need for an Integrating Framework

In the previous section, we have presented and discussed some of the most representative omic experimental approaches to characterize different facets of gene regulation patterns in cancer, as well as their main computational analysis approaches and bioinformatic tools [260]. Since each of these approaches contributes in different ways to the global phenomenon, there is a need to find theoretical frameworks and methodological techniques to integrate the knowledge derived from each of these omic technologies into a coherent, hopefully, mechanistic explanation of gene expression deregulation in cancer [11]. As is mentioned, there is a wide diversity in the types of data, dynamic ranges, sources of noise and error and other features, a fact that further complicates the development of such a holistic, integrated approach. Some preliminary proposals have been outlined in recent times [5,12]. Such proposals often combine one or more of the approaches presented in Figure 2. In this section, we will present some general methodological concepts that have resulted in relatively successful ways to integrate different classes of omic data and ultimately inform about the complex underlying cancer gene regulation patterns. Multiomics integration borrows different techniques from statistics, while multivariate analysis tackles classification and regression problems involving distinct molecular levels, probabilistic methods link entities by their chance of occurring independently of their nature creating networks that can be further explored (although it should be noted that methods like Similar network fusion [261] rely on non-probabilistic networks). Finally, statistical learning allows for feature selection amidst the broadness of omics.

4.1. Computational Approaches to Omic Integration

Multiomics integration aims to harness the interaction between the different biological levels captured by the omics. It rests on the extraction of complementary information, measured ideally on the same set of samples, in order to find co-occurring patterns [262]. Achieving such a goal will enhance the study of biological phenomena by improving models, not anymore limited to the scope of a single kind of experiment.

4.1.1. Omic Integration Stages

Late Multi-Omic Integration

The so-called late integration analyzes each omic separately and then combines the individual results ignoring inter-omics effects [263]. Though simplistic, this approach excels in classification and prediction. An example is the clustering of gynecological tumors made by Berger et al. They classified tumors in a clinically relevant way, based on previously chosen cancer biomarkers ranging from gene mutation and CNV load to receptor protein expression and immune infiltration. By simply dichotomizing biomarkers into a matrix of samples per features, where zero indicated absent or low and one meant present or high, they flattened features to uniform units that can be clustered. This way, by diminishing differences between omics type of data and range, they manage to cluster major cancer subtypes from ovary, uterus, cervix, and breast in groups with significantly different survival that do not overlap with the histological classification [264]. Though this does not strictly qualify as multiomics integration since specific elements were given to the clustering algorithm, it demonstrates the richness hidden across molecular levels, which could even inform cross-tumor-type therapies. Since some cases of uterine cancer and Luminal A breast cancer with good survival share high CNV load, low immune infiltration, and high AR, PR and ERα protein expression, Luminal A therapies could work for uterine cancer cases with such characteristics. However, now that we know these features are linked, we have to ask how are these characteristics related.

Early Multi-Omic Integration

To answer the how sort of questions, an early integration that acknowledges interomic effects with no pre-established direction [263] is needed. However, integrating the measurement of the different molecules taken up by the omics is not easy. Each omic has its own output that does not have to fit in dimension, variance, scale or noise to any other. Fortunately, methods that deal with every one of these issues exist, or are being developed [265].

The forthcoming section tackles multivariate techniques for omics integration due to its capacity to model simultaneously different omics [266] and select data-relevant features automatically. Nonetheless, these are not the only techniques available. For alternatives and ready to use tools check Bersanelli and Huang reviews [267,268].

4.2. In the Beginning There Was Statistical Learning

Omics measure thousands of molecules in tens to hundreds of samples. As a consequence, we the omics users, are by default suffering from the curse of dimensionality. When the number of subjects, n, is lower than the number of probes measured, p, we are subsampling the possible combinations of values and can hardly cover the actual space of study [269]. But this is no new problem, statisticians have been struggling with $p > n$ for longer than omics have been around. Then, the severe $p >> n$ hit when integrating different omics is not as intractable as first sight may suggest. Statistical learning, the area of statistics dedicated to model and understand complex datasets -exactly what multiomics want-, deals with $p >> n$ situations applying regularization and reduction of dimension.

Regularization shrinks the coefficients on a model towards zero. Such shrinkage comes from a penalty added to the model fitting function. When shrinkage forces some of the coefficient estimates to zero, variable selection is attained [269].
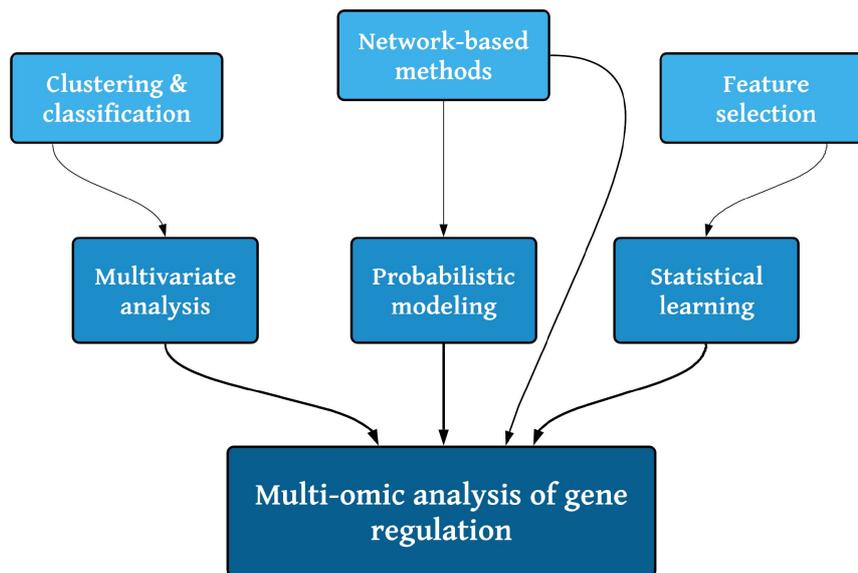
Figure 2. Several modeling approaches to integrate multiomic data in gene regulation.

4.2.1. Least Absolute Shrinkage and Selection Operator (LASSO) Methods

The regularization method least absolute shrinkage and selection operator (LASSO), accomplishes variable selection through scaling of the l1 norm of the vector of coefficients with a tuning parameter λ. As a consequence of the penalization, predictors with the highest correlation to the modeled response get selected [269]. Further, even though their value does not necessarily translate to original measurements[270], coefficients weigh the importance and direction of the predictor's effect [271], allowing deeper focus on selected variables. These properties have proven useful to find gene expression direct regulators, phenotype-specific molecules and interactions among distinct molecular levels [35,272,273].

Sohn and coworkers applied a LASSO multivariate linear regression to model ovarian cancer gene expression based on methylation, miRNA expression, and CNA. Results suggest there is a disparate impact of the regulators on the grade of expression. While highly expressed genes tend to associate with CNA; variable expressed genes are better explained by methylation features. When checking top coefficients, methylation appears as the omic with the strongest effect, with CNA coming lower in the rank and miRNAs dispersed all over. Then, the LASSO integrative setting, not only models simultaneously the distinct data types but allow a ranking of the omic's effect though its feature selection capacity. Additionally, the network linking the predictors selected by the integrative models shows better modularity and specific functional enrichment than the one derived from non-integrative models, supporting the need for integrative studies [273].

The problem with the LASSO is that it selects at most n variables before getting saturated [274], which makes necessary additional steps for the $p >> n$ multiomics case.

4.2.2. Dimension Reduction

Dimension reduction surpasses LASSO saturation by exploiting the matrix representation of omics to find $M < p$ linear combinations of the original predictors. The use of these linear combinations instead of the original predictors effectively lowers dimension from p to M [269].

The definition of data subspaces exploits matrix factorization techniques like the ones described in [275]. Both co-inertia analysis, CIA, and sparse partial least squares (sPLS) maximize the covariance

between eigenvectors, while Canonical Correlation Analysis, CCA, maximizes their correlation [276]. Multiple factor analysis (MFA) projects a multi-omics matrix that can include both numerical and discrete datasets into the principal components subspace [266]. Joint and individual variation explained (JIVE) decomposes each omics matrix into joint, individual and residual variation structures [277,278]. CCA and JIVE can become sparse trough penalization, at least dimension reduction has been combined with LASSO penalization for both regression and classification settings [279–282].

Trough sparse multi-block partial least squares, Li et al. found multilayer gene regulatory modules on ovarian cancer data. Starting from 799 microRNA and 15846 gene expression profiles, 31324 loci with CNVs and 14735 DNA methylation marks; they reduced the high dimensional dataset into modules with an average of 45 CNV loci, 42 methylation marks, 5 microRNAs and 44 genes. Inter-omic relations found had significant IPA p-values, demonstrating the power of the technique [281].

4.2.3. Elastic Net Regularization

LASSO saturation is paired with breakage of correlated groups, where just one of the grouped predictors was selected while the rest were discarded. The elastic net regularization overcomes this via a strictly convex optimization determined by an α parameter [274]. The α parameter determines how similar is the model to LASSO ($\alpha = 0$) or to ridge regression ($\alpha = 1$), another regularization method that is unable to select variables but gives similar coefficients to correlated variables. Thus, the elastic net is expected to recover variables than the LASSO does not, including both true and false positives [283–285].

Theoretically, multivariate methods recover synergistic effects that traditional paired correlation studies can not, due to the simultaneous analysis of the distinct omics [273]. The capture of entire groups of correlated variables suggests that the elastic net could overmatch LASSO on this task. Though simulation Neto et al. show the elastic-net outperforms LASSO predictive power when the variables are highly correlated. Examples of use can be found at [285,286]. Both LASSO and elastic net were used to explore the link between SNPs, DNA methylation and gene expression in bladder cancer in [284]. Results suggest SNPs and DNA methylation regulate cis gene expression, but each penalization method identifies distinct genes that would be in this situation.

4.2.4. Feature Selection under Heterogeneous Dynamic Ranges

Either regularization method has two more issues. Omics matrices need to be concatenated no matter their difference in scale and; the same penalization is applied to all the omics ignoring that each molecular level might affect the response differently. Even when these problems arise at separate time points, they exhibit the same question, we are still learning how to fit matrices of different size and range together.

To interpret co-occurrent measures of different omics a normalization that assures all the information is effectively taken into account is necessary. Such normalization can be as abrupt as the mentioned dichotomization made by Berger et al. or as broad as centering around zero with unit variance, which is the recommendation of most methods [266,282,287]. However, it is extremely important since it can shape the final results. Situations where the largest dataset dominates may require to scale each data type by its total variation to force them to contribute equally [277] or, to scale each omic by its first eigenvalue to rest heavier on the more informative omics [266]. In this sense, the decomposition method AJIVE, being insensitive to scale heterogeneity [278], has an advantage over multivariate methods.

Bringing data to the same range does not guarantee a balanced variable selection across omics. Applying the same shrinkage could shrink to zero all the coefficients of subtler effect omics. To solve this, Liu et al. tunned the extent of penalization per omic achieving an optimal shrinkage for each omic. The resulting model achieved better classification than a single penalty over simulations and cancer samples with gene expression and methylation data [286]. Weighted penalization had been used by

Lee et al. to deal with low and unbalanced sample numbers for breast cancer subtypes prediction [35], making the approach an appealing modification.

4.2.5. Modeling Related Issues

Multivariate techniques applied to multi-omics deliver models for the interaction between omics. Thus, they carry the same concerns on statistical power and overfitting that all models have. Such problems are largely driven by limited sample size. Even though many more datasets are available now, there is still a lack of co-measured omics. The constrained sample size is nevertheless exploited via k-fold cross-validation plus testing on unseen data.

However, the sample size is not the only factor driving model quality. Small effects require larger detection power, which can be tuned through weighted penalties [286] as explained in the previous section. Significance can not be measured intrinsically, but p-values can be assigned following permutation approaches [284,288]. The threat of overfitting is lessened by the sparsity of the models [271], but it is still necessary to assess prediction accuracy in samples not used for training. Such testing with independent data, tests model validity by measuring coherence with published results.

Even with all the drawbacks described, multivariate omics integration has the undeniable advantage of unbiased variable selection. Future work is expected to set the guidelines for the reproducible application of penalization models with omics data [265,289]. Special efforts need to be done on the explicit report of model fitting and evaluation processes. In this sense, Git tools could be helpful. The incorporation of network strategies is promising too, as explained in the next section.

4.3. Network-Based Methods

One almost paradigmatic way to comprehensively map and analyze system-level (genome-wide) interactions in contemporary biology is by using complex networks. The network view has been used so extensively to integrate information of high throughput experiments in biology that, for some people systems biology has become almost a synonym with network biology [290]. Gene expression regulation at the whole genome level in cancer has been extensively studied in the past [291–294].

Network analyses have been used in the past to integrate multiple omics experiments in relation to gene expression regulation in cancer [65,67,295] and other diseases [296,297]. Multi-omic networks have also been discussed in other instances of biomolecular regulation [298,299]. The mathematical foundations to integrate and analyze such multilayer networks are however being laid out [300–302], and strategies for their use in multiomics are currently also under development [303].

For instance, an approach called similarity network fusion was developed [261] to integrate gene expression, DNA methylation, and miR expression data coming from five different cancer data sets. The method was useful to better ascertain cancer subtypes predicting survival. Costa and co-workers [295], in turn, used a multinetwork consisting of correlations among differentially expressed and differentially methylated genes in head-and-neck squamous cell cancer (both HPV+ and HPV-) to identify a set of genes with methylation alteration patterns in their promoter. They observed co-expression modules leading to discover key regulatory elements.

Finding molecular signatures was also the strategy followed by Gibbs et al. [296] that aside from disease-specific findings, provide a platform-agnostic means to study the relation between gene and protein expression at a genome/proteome wise level. Integrative multi-omic studies relying on network-based methods have been able to even provide a theoretical framework to study, not only the interaction structure of the gene regulatory maps, but also approximations to the kinetics and dynamics of gene regulation by means of the so-called static signal flow, and dynamic signal flow analyses [299].

4.4. Hybrid Approaches

Multivariate, statistical learning and network-based methods are largely complementary and that, powerful as they are on their own merits; neither of them is able to capture the full complexity and the subtleties associated with the integrative multi-omic characterization of gene expression programs in cancer. Both areas are indeed growing very fast, with a number of new methods and improvement of existing methods constantly been added to the current literature. A good way to exploit the capabilities of both approaches is, well, integrating them.

5. Concluding Remarks

Cancer is a complex disease. The way in which its molecular (genomic) origins move forward to the cellular, tissue, and phenotypic level is most often mediated by changes in gene expression regulation programs. Here, we have discussed that there is a multitude of disparate factors contributing to the reprogramming of the gene regulatory mechanisms. Among these, we have analyzed the role of changes in the action of promoters and enhancers, variations in the chromatin structural disposition, DNA chemical modifications as well as transcript splicing, stability and transport kinetics, all of the factors that have been known for a while. Also relevant to modify gene expression patterns, are more recently discovered factors, such as large chromosomal aberrations, like chromothripsis; the larger-scale effect (at the chromosomal level) of DNA 3D structure and the action of regulatory non-coding RNAs. Just a few years ago, it was impossible to characterize these effects at the global, whole-genome level. However, recent developments in experimental omic techniques have allowed us to measure those effects. Some of the more commonly used techniques were also summarized in this review.

Acknowledging that the genomic landscape is only a part of the whole picture, genome regulators have acquired more attention. Given that methylation profiles, ncRNAs, or 3D DNA structure have a strong influence on gene expression, to develop computational techniques that accurately measure small variations in different expression levels provide us tools that allow us to identify how these non-genomic variations directly affect the cancer genome.

Letting aside the computational challenges inherent to the analysis of single-omic experiments, these techniques however, have unveiled an enormous problem for computational oncogenomics: i.e., how to build models and integrate this enormous wealth of disparate information, into coherent and predictive models that will help us to decipher how the different faces of gene regulation interact to develop the anomalous gene regulatory programs that we deem responsible for the rise, establishment, and maintenance of the tumor phenotype; the well-known hallmarks of cancer. We discussed how the combination of powerful computational models based on statistical learning, machine intelligence, and probabilistic modeling, but also in network (and multinetwork) based methods, presents as an appealing alternative to be developed by the coming generation of computational cancer scientists to tackle the enigma of gene regulation in cancer.

References

1. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
2. Tennant, D.A.; Durán, R.V.; Gottlieb, E. Targeting metabolic transformation for cancer therapy. *Nat. Rev. Cancer* **2010**, *10*, 267. [CrossRef]
3. Dancey, J.E.; Bedard, P.L.; Onetto, N.; Hudson, T.J. The genetic basis for cancer treatment decisions. *Cell* **2012**, *148*, 409–420. [CrossRef] [PubMed]
4. Futreal, P.A.; Coin, L.; Marshall, M.; Down, T.; Hubbard, T.; Wooster, R.; Rahman, N.; Stratton, M.R. A census of human cancer genes. *Nat. Rev. Cancer* **2004**, *4*, 177. [CrossRef] [PubMed]
5. Hernández-Lemus, E. Systems biology and integrative omics in breast cancer. In *Omics Approaches in Breast Cancer*, 1st ed.; Springer: New Delhi, India, 2014; pp. 333–352.
6. Barbolosi, D.; Ciccolini, J.; Lacarelle, B.; Barlesi, F.; Andre, N. Computational oncology–mathematical modelling of drug regimens for precision medicine. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 242. [CrossRef] [PubMed]
7. Gatenby, R.A.; Maini, P.K. Mathematical oncology: Cancer summed up. *Nature* **2003**, *421*, 321. [CrossRef]
8. Lefor, A.T. Computational oncology. *Jpn. J. Clin. Oncol.* **2016**, *13*, 242. [CrossRef]
9. Wang, E.; Zaman, N.; Mcgee, S.; Milanese, J.S.; Masoudi-Nejad, A.; O'Connor-McCourt, M. Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. In *Seminars in Cancer Biology*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 30, pp. 4–12.
10. Wang, E. Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Lett.* **2013**, *340*, 261–269. [CrossRef]
11. De Anda-Jáuregui, G.; Hernández-Lemus, E. The path to integration in computational oncology. *Precis. Med. Oncol.* **2018**, *1*, 9–23.
12. Hernández-Lemus, E. Further steps toward functional systems biology of cancer. *Front. Physiol.* **2013**, *4*, 256. [CrossRef]
13. Wang, E.; Zou, J.; Zaman, N.; Beitel, L.K.; Trifiro, M.; Paliouras, M. Cancer systems biology in the genome sequencing era: Part 1, dissecting and modeling of tumor clones and their networks. In *Seminars in Cancer Biology*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 23, pp. 279–285.
14. Cox, P.; Goding, C. Transcription and cancer. *Br. J. Cancer* **1991**, *63*, 651. [CrossRef] [PubMed]
15. Kim, H.; Kim, Y.M. Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. *Sci. Rep.* **2018**, *8*, 6041. [CrossRef] [PubMed]
16. Weinhold, N.; Jacobsen, A.; Schultz, N.; Sander, C.; Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **2014**, *46*, 1160. [CrossRef] [PubMed]
17. Sur, I.; Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* **2016**, *16*, 483. [CrossRef] [PubMed]
18. Kandoth, C.; McLellan, M.D.; Vandin, F.; Ye, K.; Niu, B.; Lu, C.; Xie, M.; Zhang, Q.; McMichael, J.F.; Wyczalkowski, M.A.; et al. Mutational landscape and significance across 12 major cancer types. *Nature* **2013**, *502*, 333. [CrossRef]
19. Lin, C.Y.; Lovén, J.; Rahl, P.B.; Paranal, R.M.; Burge, C.B.; Bradner, J.E.; Lee, T.I.; Young, R.A. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **2012**, *151*, 56–67. [CrossRef]
20. Yang, J.; Weinberg, R.A. Epithelial-mesenchymal transition: At the crossroads of development and tumor metastasis. *Dev. Cell* **2008**, *14*, 818–829. [CrossRef]
21. Herz, H.M.; Hu, D.; Shilatifard, A. Enhancer malfunction in cancer. *Mol. Cell* **2014**, *53*, 859–866. [CrossRef]
22. Sze, C.C.; Shilatifard, A. MLL3/MLL4/COMPASS family on epigenetic regulation of enhancer function and cancer. *Cold Spring Harb. Perspect. Med.* **2016**, *6*, a026427. [CrossRef]
23. Eldholm, V.; Haugen, A.; Zienolddiny, S. CTCF mediates the TERT enhancer–promoter interactions in lung cancer cells: Identification of a novel enhancer region involved in the regulation of TERT gene. *Int. J. Cancer* **2014**, *134*, 2305–2313. [CrossRef]
24. Bojesen, S.E.; Pooley, K.A.; Johnatty, S.E.; Beesley, J.; Michailidou, K.; Tyrer, J.P.; Edwards, S.L.; Pickett, H.A.; Shen, H.C.; Smart, C.E.; et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **2013**, *45*, 371. [CrossRef] [PubMed]

25. Chen, C.; Yue, D.; Lei, L.; Wang, H.; Lu, J.; Zhou, Y.; Liu, S.; Ding, T.; Guo, M.; Xu, L. Promoter-operating targeted expression of gene therapy in cancer: Current stage and prospect. *Mol. Ther.-Nucleic Acids* **2018**, *11*, 508–514. [CrossRef] [PubMed]

26. Klemm, S.L.; Shipony, Z.; Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **2019**, 20, 207–220. [CrossRef] [PubMed]

27. Cramer, P.; Wolberger, C. Proteins: Histones and chromatin. *Curr. Opin. Struct. Biol.* **2011**, *21*, 695. [CrossRef] [PubMed]

28. Thurman, R.E.; Rynes, E.; Humbert, R.; Vierstra, J.; Maurano, M.T.; Haugen, E.; Sheffield, N.C.; Stergachis, A.B.; Wang, H.; Vernot, B.; et al. The accessible chromatin landscape of the human genome. *Nature* **2012**, *489*, 75. [CrossRef] [PubMed]

29. Corces, M.R.; Granja, J.M.; Shams, S.; Louie, B.H.; Seoane, J.A.; Zhou, W.; Silva, T.C.; Groeneveld, C.; Wong, C.K.; Cho, S.W.; et al. The chromatin accessibility landscape of primary human cancers. *Science* **2018**, *362*, eaav1898. [CrossRef] [PubMed]

30. Taberlay, P.C.; Achinger-Kawecka, J.; Lun, A.T.; Buske, F.A.; Sabir, K.; Gould, C.M.; Zotenko, E.; Bert, S.A.; Giles, K.A.; Bauer, D.C.; et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.* **2016**, *26*, 719–731. [CrossRef]

31. Bediaga, N.G.; Acha-Sagredo, A.; Guerra, I.; Viguri, A.; Albaina, C.; Diaz, I.R.; Rezola, R.; Alberdi, M.J.; Dopazo, J.; Montaner, D.; et al. DNA methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Res.* **2010**, *12*, R77. [CrossRef]

32. Baxter, J.S.; Leavy, O.C.; Dryden, N.H.; Maguire, S.; Johnson, N.; Fedele, V.; Simigdala, N.; Martin, L.A.; Andrews, S.; Wingett, S.W.; et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat. Commun.* **2018**, *9*, 1028. [CrossRef]

33. Jia, R.; Chai, P.; Zhang, H.; Fan, X. Novel insights into chromosomal conformations in cancer. *Mol. Cancer* **2017**, *16*, 173. [CrossRef]

34. Le Dily, F.; Baù, D.; Pohl, A.; Vicent, G.P.; Serra, F.; Soronellas, D.; Castellano, G.; Wright, R.H.; Ballare, C.; Filion, G.; et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.* **2014**, *28*, 2151–2162. [CrossRef] [PubMed]

35. Lee, G.; Bang, L.; Kim, S.Y.; Kim, D.; Sohn, K.A. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. *BMC Med. Genom.* **2017**, *10*, 28. [CrossRef] [PubMed]

36. Szyf, M. DNA methylation signatures for breast cancer classification and prognosis. *Genome Med.* **2012**, *4*, 26. [CrossRef] [PubMed]

37. El Marabti, E.; Younis, I. The Cancer Spliceome: Reprograming of alternative splicing in cancer. *Front. Mol. Biosci.* **2018**, *5*, 80. [CrossRef] [PubMed]

38. Srebrow, A.; Kornblihtt, A.R. The connection between splicing and cancer. *J. Cell Sci.* **2006**, *119*, 2635–2641. [CrossRef] [PubMed]

39. Di, C.; Zhang, Q.; Chen, Y.; Wang, Y.; Zhang, X.; Liu, Y.; Sun, C.; Zhang, H.; Hoheisel, J.D.; et al. Function, clinical application, and strategies of pre-mRNA splicing in cancer. *Cell Death Differ.* **2019**, *26*, 1181–1194. [CrossRef] [PubMed]

40. Tang, Y.C.; Amon, A. Gene copy-number alterations: A cost-benefit analysis. *Cell* **2013**, *152*, 394–405. [CrossRef]

41. Nijhawan, D.; Zack, T.I.; Ren, Y.; Strickland, M.R.; Lamothe, R.; Schumacher, S.E.; Tsherniak, A.; Besche, H.C.; Rosenbluh, J.; Shehata, S.; et al. Cancer vulnerabilities unveiled by genomic loss. *Cell* **2012**, *150*, 842–854. [CrossRef]

42. Gordon, D.J.; Resio, B.; Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat. Rev. Genet.* **2012**, *13*, 189. [CrossRef]

43. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The cancer genome. *Nature* **2009**, *458*, 719. [CrossRef]

44. Inaki, K.; Menghi, F.; Woo, X.Y.; Wagner, J.P.; Jacques, P.É.; Lee, Y.F.; Shreckengast, P.T.; Soon, W.W.; Malhotra, A.; Teo, A.S.; et al. Systems consequences of amplicon formation in human breast cancer. *Genome Res.* **2014**, *24*, 1559–1571. [CrossRef] [PubMed]

45. Menghi, F.; Inaki, K.; Woo, X.; Kumar, P.A.; Grzeda, K.R.; Malhotra, A.; Yadav, V.; Kim, H.; Marquez, E.J.; Ucar, D.; et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E2373–E2382. [CrossRef] [PubMed]

46. Ahmadiyeh, N.; Pomerantz, M.M.; Grisanzio, C.; Herman, P.; Jia, L.; Almendro, V.; He, H.H.; Brown, M.; Liu, X.S.; Davis, M.; et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 9742–9746. [CrossRef] [PubMed]

47. Guan, Y.; Kuo, W.L.; Stilwell, J.L.; Takano, H.; Lapuk, A.V.; Fridlyand, J.; Mao, J.H.; Yu, M.; Miller, M.A.; Santos, J.L.; et al. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin. Cancer Res.* **2007**, *13*, 5745–5755. [CrossRef] [PubMed]

48. Naylor, T.L.; Greshock, J.; Wang, Y.; Colligon, T.; Yu, Q.; Clemmer, V.; Zaks, T.Z.; Weber, B.L. High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res.* **2005**, *7*, R1186. [CrossRef] [PubMed]

49. Yokota, T.; Yoshimoto, M.; Akiyama, F.; Sakamoto, G.; Kasumi, F.; Nakamura, Y.; Emi, M. Frequent multiplication of chromosomal region 8q24. 1 associated with aggressive histologic types of breast cancers. *Cancer Lett.* **1999**, *139*, 7–13. [CrossRef]

50. Chin, K.; DeVries, S.; Fridlyand, J.; Spellman, P.T.; Roydasgupta, R.; Kuo, W.L.; Lapuk, A.; Neve, R.M.; Qian, Z.; Ryder, T.; et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **2006**, *10*, 529–41. [CrossRef]

51. Smid, M.; Hoes, M.; Sieuwerts, A.M.; Sleijfer, S.; Zhang, Y.; Wang, Y.; Foekens, J.A.; Martens, J.W. Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes. *Breast Cancer Res. Treat.* **2011**, *128*, 23–30. [CrossRef]

52. Klinge, C. Non-coding RNAs in breast cancer: Intracellular and intercellular communication. *Non-Coding RNA* **2018**, *4*, 40. [CrossRef]

53. Evans, J.R.; Feng, F.Y.; Chinnaiyan, A.M. The bright side of dark matter: lncRNAs in cancer. *J. Clin. Investig.* **2016**, *126*, 2775–2782. [CrossRef]

54. Léveillé, N.; Melo, C.A.; Rooijers, K.; Díaz-Lagares, A.; Melo, S.A.; Korkmaz, G.; Lopes, R.; Moqadam, F.A.; Maia, A.R.; Wijchers, P.J.; et al. Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. *Nat. Commun.* **2015**, *6*, 6520. [CrossRef] [PubMed]

55. Zhao, H.; Kim, Y.; Wang, P.; Lapointe, J.; Tibshirani, R.; Pollack, J.R.; Brooks, J.D. Genome-wide characterization of gene expression variations and DNA copy number changes in prostate cancer cell lines. *Prostate* **2005**, *63*, 187–197. [CrossRef] [PubMed]

56. Zhu, J.; Liu, S.; Ye, F.; Shen, Y.; Tie, Y.; Zhu, J.; Wei, L.; Jin, Y.; Fu, H.; Wu, Y.; et al. Long noncoding RNA MEG3 interacts with p53 protein and regulates partial p53 target genes in hepatoma cells. *PloS ONE* **2015**, *10*, e0139790. [CrossRef] [PubMed]

57. Gutschner, T.; Hämmerle, M.; Eißmann, M.; Hsu, J.; Kim, Y.; Hung, G.; Revenko, A.; Arun, G.; Stentrup, M.; Groß, M.; et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* **2013**, *73*, 1180–1189. [CrossRef]

58. Kim, J.; Piao, H.L.; Kim, B.J.; Yao, F.; Han, Z.; Wang, Y.; Xiao, Z.; Siverly, A.N.; Lawhon, S.E.; Ton, B.N.; et al. Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nat. Genet.* **2018**, *50*, 1705. [CrossRef]

59. Sun, J.; Li, W.; Sun, Y.; Yu, D.; Wen, X.; Wang, H.; Cui, J.; Wang, G.; Hoffman, A.R.; Hu, J.F. A novel antisense long noncoding RNA within the *IGF1R* gene locus is imprinted in hematopoietic malignancies. *Nucleic Acids Res.* **2014**, *42*, 9588–9601. [CrossRef]

60. Imani, S.; Wu, R.C.; Fu, J. MicroRNA-34 family in breast cancer: From research to therapeutic potential. *J. Cancer* **2018**, *9*, 3765. [CrossRef]

61. Yang, S.J.; Yang, S.Y.; Wang, D.D.; Chen, X.; Shen, H.Y.; Zhang, X.H.; Zhong, S.L.; Tang, J.H.; Zhao, J.H. The miR-30 family: Versatile players in breast cancer. *Tumor Biol.* **2017**, *39*, 1010428317692204. [CrossRef]

62. Cho, W.C. OncomiRs: The discovery and progress of microRNAs in cancers. *Mol. Cancer* **2007**, *6*, 60. [CrossRef]

63. Esquela-Kerscher, A.; Slack, F.J. Oncomirs, microRNAs with a role in cancer. *Nat. Rev. Cancer* **2006**, *6*, 259. [CrossRef]

64. Jiang, S.; Zhang, H.W.; Lu, M.H.; He, X.H.; Li, Y.; Gu, H.; Liu, M.F.; Wang, E.D. MicroRNA-155 functions as an OncomiR in breast cancer by targeting the suppressor of cytokine signaling 1 gene. *Cancer Res.* **2010**, *70*, 3119–3127. [CrossRef] [PubMed]

65. de Anda-Jáuregui, G.; Espinal-Enríquez, J.; Drago-García, D.; Hernández-Lemus, E. Nonredundant, highly connected microRNAs control functionality in breast cancer networks. *Int. J. Genomics* **2018**, *2018*, 9585383 [CrossRef] [PubMed]

66. Svoronos, A.A.; Engelman, D.M.; Slack, F.J. OncomiR or tumor suppressor? The duplicity of microRNAs in cancer. *Cancer Res.* **2016**, *76*, 3666–3670. [CrossRef] [PubMed]

67. Drago-García, D.; Espinal-Enríquez, J.; Hernández-Lemus, E. Network analysis of EMT and MET micro-RNA regulation in breast cancer. *Sci. Rep.* **2017**, *7*, 13534. [CrossRef]

68. Vargas, D.Y.; Raj, A.; Marras, S.A.; Kramer, F.R.; Tyagi, S. Mechanism of mRNA transport in the nucleus. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17008–17013. [CrossRef]

69. Di Liegro, C.M.; Schiera, G.; Di Liegro, I. Regulation of mRNA transport, localization and translation in the nervous system of mammals. *Int. J. Mol. Med.* **2014**, *33*, 747–762. [CrossRef]

70. Siddiqui, N.; Borden, K.L. mRNA export and cancer. *RNA* **2012**, *3*, 13–25. [CrossRef]

71. Culjkovic-Kraljacic, B.; Borden, K.L. Aiding and abetting cancer: mRNA export and the nuclear pore. *Trends Cell Biol.* **2013**, *23*, 328–335. [CrossRef]

72. Cruz-Ramos, E.; Sandoval-Hernández, A.; Tecalco-Cruz, A.C. Differential expression and molecular interactions of chromosome region maintenance 1 and calreticulin exportins in breast cancer cells. *J. Steroid Biochem. Mol. Biol.* **2019**, *185*, 7–16. [CrossRef]

73. Mahipal, A.; Malafa, M. Importins and exportins as therapeutic targets in cancer. *Pharmacol. Ther.* **2016**, *164*, 135–143. [CrossRef]

74. Dickmanns, A.; Monecke, T.; Ficner, R. Structural basis of targeting the exportin CRM1 in cancer. *Cells* **2015**, *4*, 538–568. [CrossRef] [PubMed]

75. Stephens, P.J.; Greenman, C.D.; Fu, B.; Yang, F.; Bignell, G.R.; Mudie, L.J.; Pleasance, E.D.; Lau, K.W.; Beare, D.; Stebbings, L.A.; et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **2011**, *144*, 27–40. [CrossRef] [PubMed]

76. Koltsova, A.S.; Pendina, A.A.; Efimova, O.A.; Chiryaeva, O.G.; Kuznetzova, T.V.; Baranov, V.S. On the complexity of mechanisms and consequences of chromothripsis: An update. *Front. Genet.* **2019**, *10*, 393. [CrossRef] [PubMed]

77. Sansregret, L.; Vanhaesebroeck, B.; Swanton, C. Determinants and clinical implications of chromosomal instability in cancer. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 139. [CrossRef]

78. Crasta, K.; Ganem, N.J.; Dagher, R.; Lantermann, A.B.; Ivanova, E.V.; Pan, Y.; Nezi, L.; Protopopov, A.; Chowdhury, D.; Pellman, D. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **2012**, *482*, 53. [CrossRef]

79. Ly, P.; Cleveland, D.W. Rebuilding chromosomes after catastrophe: Emerging mechanisms of chromothripsis. *Trends Cell Biol.* **2017**, *27*, 917–930. [CrossRef]

80. Luijten, M.N.H.; Lee, J.X.T.; Crasta, K.C. Mutational game changer: Chromothripsis and its emerging relevance to cancer. *Mutation Res.* **2018**, *777*, 29–51. [CrossRef]

81. Nik-Zainal, S.; Alexandrov, L.B.; Wedge, D.C.; Van Loo, P.; Greenman, C.D.; Raine, K.; Jones, D.; Hinton, J.; Marshall, J.; Stebbings, L.A.; et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **2012**, *149*, 979–993. [CrossRef]

82. Alexandrov, L.B.; Nik-Zainal, S.; Wedge, D.C.; Aparicio, S.A.; Behjati, S.; Biankin, A.V.; Bignell, G.R.; Bolli, N.; Borg, A.; Børresen-Dale, A.L.; et al. Signatures of mutational processes in human cancer. *Nature* **2013**, *500*, 415. [CrossRef]

83. Baca, S.C.; Prandi, D.; Lawrence, M.S.; Mosquera, J.M.; Romanel, A.; Drier, Y.; Park, K.; Kitabayashi, N.; MacDonald, T.Y.; Ghandi, M.; et al. Punctuated evolution of prostate cancer genomes. *Cell* **2013**, *153*, 666–677. [CrossRef]

84. Shen, M.M. Chromoplexy: A new category of complex rearrangements in the cancer genome. *Cancer Cell* **2013**, *23*, 567–569. [CrossRef] [PubMed]

85. Cortes-Ciriano, I.; Lee, J.K.; Xi, R.; Jain, D.; Jung, Y.L.; Yang, L.; Gordenin, D.; Klimczak, L.J.; Zhang, C.Z.; Pellman, D.S.; et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *BioRxiv* **2018**. [CrossRef]

86. Lee, J.Y.; Mustafa, M.; Kim, C.Y.; Kim, M.H. Depletion of CTCF in breast cancer cells selectively induces cancer cell death via p53. *J. Cancer* **2017**, *8*, 2124. [CrossRef] [PubMed]

87. Audia, J.E.; Campbell, R.M. Histone modifications and cancer. *Cold Spring Harb. Perspect. Biol.* **2016**, *8*, a019521. [CrossRef]

88. Sulaiman, A.; McGarry, S.; Lam, K.M.; El-Sahli, S.; Chambers, J.; Kaczmarek, S.; Li, L.; Addison, C.; Dimitroulakos, J.; Arnaout, A.; et al. Co-inhibition of mTORC1, HDAC and ESR1α retards the growth of triple-negative breast cancer and suppresses cancer stem cells. *Cell Death Dis.* **2018**, *9*, 815. [CrossRef]

89. Damaskos, C.; Valsami, S.; Kontos, M.; Spartalis, E.; Kalampokas, T.; Kalampokas, E.; Athanasiou, A.; Moris, D.; Daskalopoulou, A.; Davakis, S.; et al. Histone deacetylase inhibitors: An attractive therapeutic strategy against breast cancer. *Anticancer Res.* **2017**, *37*, 35–46. [CrossRef]

90. Terranova-Barberio, M.; Thomas, S.; Ali, N.; Pawlowska, N.; Park, J.; Krings, G.; Rosenblum, M.D.; Budillon, A.; Munster, P.N. HDAC inhibition potentiates immunotherapy in triple negative breast cancer. *Oncotarget* **2017**, *8*, 114156. [CrossRef]

91. Halkidou, K.; Gaughan, L.; Cook, S.; Leung, H.Y.; Neal, D.E.; Robson, C.N. Upregulation and nuclear recruitment of HDAC1 in hormone refractory prostate cancer. *Prostate* **2004**, *59*, 177–189. [CrossRef]

92. Bian, X.; Liang, Z.; Feng, A.; Salgado, E.; Shim, H. HDAC inhibitor suppresses proliferation and invasion of breast cancer cells through regulation of miR-200c targeting CRKL. *Biochem. Pharmacol.* **2018**, *147*, 30–37. [CrossRef]

93. Zhang, W.; Xu, J. DNA methyltransferases and their roles in tumorigenesis. *Biomark. Res.* **2017**, *5*, 1. [CrossRef]

94. Balmain, A. Cancer genetics: From Boveri and Mendel to microarrays. *Nat. Rev. Cancer* **2001**, *1*, 77. [CrossRef] [PubMed]

95. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860.

96. Van Allen, E.M.; Wagle, N.; Stojanov, P.; Perrin, D.L.; Cibulskis, K.; Marlow, S.; Jane-Valbuena, J.; Friedrich, D.C.; Kryukov, G.; Carter, S.L.; et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **2014**, *20*, 682. [CrossRef] [PubMed]

97. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **2011**, *474*, 609.

98. Chaudhuri, A.A.; Chabon, J.J.; Lovejoy, A.F.; Newman, A.M.; Stehr, H.; Azad, T.D.; Khodadoust, M.S.; Esfahani, M.S.; Liu, C.L.; Zhou, L.; et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discov.* **2017**, *7*, 1394–1403. [CrossRef] [PubMed]

99. Frampton, G.M.; Fichtenholtz, A.; Otto, G.A.; Wang, K.; Downing, S.R.; He, J.; Schnall-Levin, M.; White, J.; Sanford, E.M.; An, P.; et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **2013**, *31*, 1023. [CrossRef] [PubMed]

100. Huang, K.l.; Mashl, R.J.; Wu, Y.; Ritter, D.I.; Wang, J.; Oh, C.; Paczkowska, M.; Reynolds, S.; Wyczalkowski, M.A.; Oak, N.; et al. Pathogenic germline variants in 10,389 adult cancers. *Cell* **2018**, *173*, 355–370. [CrossRef] [PubMed]

101. Berger, M.F.; Mardis, E.R. The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 353. [CrossRef] [PubMed]

102. Head, S.R.; Komori, H.K.; LaMere, S.A.; Whisenant, T.; Van Nieuwerburgh, F.; Salomon, D.R.; Ordoukhanian, P. Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* **2014**, *56*, 61–77. [CrossRef]

103. Hodges, E.; Xuan, Z.; Balija, V.; Kramer, M.; Molla, M.N.; Smith, S.W.; Middle, C.M.; Rodesch, M.J.; Albert, T.J.; Hannon, G.J.; et al. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **2007**, *39*, 1522. [CrossRef]

104. Forshew, T.; Murtaza, M.; Parkinson, C.; Gale, D.; Tsui, D.W.; Kaper, F.; Dawson, S.J.; Piskorz, A.M.; Jimenez-Linan, M.; Bentley, D.; et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **2012**, *4*, 136ra68. [CrossRef]

105. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333. [CrossRef]

106. Mose, L.E.; Wilkerson, M.D.; Hayes, D.N.; Perou, C.M.; Parker, J.S. ABRA: Improved coding indel detection via assembly-based realignment. *Bioinformatics* **2014**, *30*, 2813–2815. [CrossRef] [PubMed]

107. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FASTQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf.* **2013**, *43*, 1–33. [CrossRef]

108. Poplin, R.; Ruano-Rubio, V.; DePristo, M.A.; Fennell, T.J.; Carneiro, M.O.; Van der Auwera, G.A.; Kling, D.E.; Gauthier, L.D.; Levy-Moonshine, A.; Roazen, D.; et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* **2017**. [CrossRef]

109. Compeau, P.E.; Pevzner, P.A.; Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **2011**, *29*, 987. [CrossRef]

110. Li, H.; Ruan, J.; Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **2008**, *18*, 1851–1858. [CrossRef]

111. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* **2012**, arXiv:1207.3907.

112. Xi, R.; Lee, S.; Xia, Y.; Kim, T.M.; Park, P.J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* **2016**, *44*, 6274–6286. [CrossRef]

113. Li, H. FermiKit: Assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **2015**, *31*, 3694–3696. [CrossRef]

114. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]

115. Bertelsen, B.; Tuxen, I.V.; Yde, C.W.; Gabrielaite, M.; Torp, M.H.; Kinalis, S.; Oestrup, O.; Rohrberg, K.; Spangaard, I.; Santoni-Rugiu, E.; et al. High frequency of pathogenic germline variants within homologous recombination repair in patients with advanced cancer. *NPJ Genomic Med.* **2019**, *4*, 13. [CrossRef]

116. Quezada Urban, R.; Díaz Velásquez, C.; Gitler, R.; Rojo Castillo, M.; Sirota Toporek, M.; Figueroa Morales, A.; Moreno García, O.; García Esquivel, L.; Torres Mejía, G.; Dean, M.; et al. Comprehensive analysis of germline variants in mexican patients with hereditary breast and ovarian cancer susceptibility. *Cancers* **2018**, *10*, 361. [CrossRef] [PubMed]

117. Cybulski, C.; Kluźniak, W.; Huzarski, T.; Wokołorczyk, D.; Kashyap, A.; Rusak, B.; Stempa, K.; Gronwald, J.; Szymiczek, A.; Bagherzadeh, M.; et al. The spectrum of mutations predisposing to familial breast cancer in Poland. *Int. J. Cancer* **2019**. [CrossRef] [PubMed]

118. Vogelaar, I.P.; Van Der Post, R.S.; Van Krieken, J.H.J.; Spruijt, L.; van Zelst-Stams, W.A.; Kets, C.M.; Lubinski, J.; Jakubowska, A.; Teodorczyk, U.; Aalfs, C.M.; et al. Unraveling genetic predisposition to familial or early onset gastric cancer using germline whole-exome sequencing. *Eur. J. Hum. Genet.* **2017**, *25*, 1246. [CrossRef] [PubMed]

119. Cock, P.J.; Fields, C.J.; Goto, N.; Heuer, M.L.; Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **2009**, *38*, 1767–1771. [CrossRef] [PubMed]

120. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed on 1 January 2019).

121. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef] [PubMed]

122. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194. [CrossRef]

123. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]

124. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [CrossRef]

125. Canzar, S.; Salzberg, S.L. Short read mapping: An algorithmic tour. *Proc. IEEE* **2015**, *105*, 436–458. [CrossRef]

126. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Subgroup, G.P.D.P. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]

127. Shi, W.; Ng, C.K.; Lim, R.S.; Jiang, T.; Kumar, S.; Li, X.; Wali, V.B.; Piscuoglio, S.; Gerstein, M.B.; Chagpar, A.B.; et al. Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep.* **2018**, *25*, 1446–1457. [CrossRef] [PubMed]

128. Bohannan, Z.S.; Mitrofanova, A. Calling variants in the clinic: Informed variant calling decisions based on biological, clinical, and laboratory variables. *Comput. Struct. Biotechnol. J.* **2019**, 17, 561–569. [CrossRef] [PubMed]

129. Zare, F.; Dow, M.; Monteleone, N.; Hosny, A.; Nabavi, S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinform.* **2017**, *18*, 286. [CrossRef] [PubMed]

130. Macintyre, G.; Goranova, T.E.; De Silva, D.; Ennis, D.; Piskorz, A.M.; Eldridge, M.; Sie, D.; Lewsley, L.A.; Hanif, A.; Wilson, C.; et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **2018**, *50*, 1262. [CrossRef] [PubMed]

131. Ma, X.; Liu, Y.; Liu, Y.; Alexandrov, L.B.; Edmonson, M.N.; Gawad, C.; Zhou, X.; Li, Y.; Rusch, M.C.; Easton, J.; et al. Pan-cancer genome and transcriptome analyses of 1699 paediatric leukaemias and solid tumours. *Nature* **2018**, *555*, 371. [CrossRef] [PubMed]

132. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M.; Network, C.G.A.R.; et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113. [CrossRef]

133. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef]

134. Blum, A.; Wang, P.; Zenklusen, J.C. SnapShot: TCGA-analyzed tumors. *Cell* **2018**, *173*, 530. [CrossRef]

135. Spurr, L.; Li, M.; Alomran, N.; Zhang, Q.; Restrepo, P.; Movassagh, M.; Trenkov, C.; Tunnessen, N.; Apanasovich, T.; Crandall, K.A.; et al. Systematic pan-cancer analysis of somatic allele frequency. *Sci. Rep.* **2018**, *8*, 7735. [CrossRef]

136. Cingolani, P. snpEff: Variant Effect Prediction 2012. Available online: http://snpeff.sourceforge.net/ (accessed on 1 July 2019).

137. Liu, X.; Wu, C.; Li, C.; Boerwinkle, E. dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* **2016**, *37*, 235–241. [CrossRef]

138. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **2017**, *46*, D1062–D1067. [CrossRef] [PubMed]

139. Babu, M.M. Introduction to microarray data analysis. *Comput. Genomics Theory Appl.* **2004**, *225*, 249.

140. Sultan, M.; Schulz, M.H.; Richard, H.; Magen, A.; Klingenhoff, A.; Scherf, M.; Seifert, M.; Borodina, T.; Soldatov, A.; Parkhomchuk, D.; et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **2008**, *321*, 956–960. [CrossRef] [PubMed]

141. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **2008**, *5*, 621. [CrossRef] [PubMed]

142. Li, J.R.; Sun, C.H.; Li, W.; Chao, R.F.; Huang, C.C.; Zhou, X.J.; Liu, C.C. Cancer RNA-seq Nexus: A database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res.* **2015**, *44*, D944–D951. [CrossRef] [PubMed]

143. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68. [CrossRef]

144. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580. [CrossRef]

145. Stark, R.; Grzelak, M.; Hadfield, J. RNA sequencing: The teenage years. *Nat. Rev. Genet.* **2019**, doi:10.1038/s41576-019-0150-2. [CrossRef]

146. Garber, M.; Grabherr, M.G.; Guttman, M.; Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **2011**, *8*, 469. [CrossRef]

147. Evans, C.; Hardin, J.; Stoebel, D.M. Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Briefings Bioinform.* **2017**, *19*, 776–792. [CrossRef]

148. Rapaport, F.; Khanin, R.; Liang, Y.; Pirun, M.; Krek, A.; Zumbo, P.; Mason, C.E.; Socci, N.D.; Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **2013**, *14*, 3158. [CrossRef] [PubMed]

149. Costa-Silva, J.; Domingues, D.; Lopes, F.M. RNA-seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **2017**, *12*, e0190152. [CrossRef] [PubMed]

150. Selamat, S.A.; Chung, B.S.; Girard, L.; Zhang, W.; Zhang, Y.; Campan, M.; Siegmund, K.D.; Koss, M.N.; Hagen, J.A.; Lam, W.L.; et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res.* **2012**, *22*, 1197–1211. [CrossRef] [PubMed]

151. Cava, C.; Bertoli, G.; Castiglioni, I. Integrating genetics and epigenetics in breast cancer: Biological insights, experimental, computational methods and therapeutic potential. *BMC Syst. Biol.* **2015**, *9*, 62. [CrossRef]

152. Aryee, M.J.; Jaffe, A.E.; Corrada-Bravo, H.; Ladd-Acosta, C.; Feinberg, A.P.; Hansen, K.D.; Irizarry, R.A. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **2014**, *30*, 1363–1369. [CrossRef]

153. Wang, D.; Yan, L.; Hu, Q.; Sucheston, L.E.; Higgins, M.J.; Ambrosone, C.B.; Johnson, C.S.; Smiraglia, D.J.; Liu, S. IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **2012**, *28*, 729–730. [CrossRef]

154. Davis, S.; Du, P.; Bilke, S. An introduction to the methylumi package. *Biocond. Package* **2010**. [CrossRef]

155. Wilhelm-Benartzi, C.S.; Koestler, D.C.; Karagas, M.R.; Flanagan, J.M.; Christensen, B.C.; Kelsey, K.T.; Marsit, C.J.; Houseman, E.A.; Brown, R. Review of processing and analysis methods for DNA methylation array data. *Br. J. Cancer* **2013**, *109*, 1394. [CrossRef]

156. Wang, T.; Guan, W.; Lin, J.; Boutaoui, N.; Canino, G.; Luo, J.; Celedón, J.C.; Chen, W. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics* **2015**, *10*, 662–669. [CrossRef]

157. Lambert, S.A.; Jolma, A.; Campitelli, L.F.; Das, P.K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T.R.; Weirauch, M.T. The human transcription factors. *Cell* **2018**, *172*, 650–665. [CrossRef]

158. Lawrence, M.; Daujat, S.; Schneider, R. Lateral thinking: How histone modifications regulate gene expression. *Trends Genet.* **2016**, *32*, 42–56. [CrossRef] [PubMed]

159. Ross-Innes, C.S.; Stark, R.; Teschendorff, A.E.; Holmes, K.A.; Ali, H.R.; Dunning, M.J.; Brown, G.D.; Gojis, O.; Ellis, I.O.; Green, A.R.; et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **2012**, *481*, 389. [CrossRef] [PubMed]

160. Yu, F.; Shen, X.; Fan, L.; Yu, Z. Analysis of histone modifications at human ribosomal DNA in liver cancer cell. *Sci. Rep.* **2015**, *5*, 18100. [CrossRef] [PubMed]

161. Zhang, G.; Zhao, Y.; Liu, Y.; Kao, L.P.; Wang, X.; Skerry, B.; Li, Z. FOXA1 defines cancer cell specificity. *Sci. Adv.* **2016**, *2*, e1501473. [CrossRef] [PubMed]

162. Wilson, S.; Qi, J.; Filipp, F.V. Refinement of the androgen response element based on ChIP-Seq in androgen-insensitive and androgen-responsive prostate cancer cell lines. *Sci. Rep.* **2016**, *6*, 32611. [CrossRef] [PubMed]

163. Chan, H.L.; Beckedorff, F.; Zhang, Y.; Garcia-Huidobro, J.; Jiang, H.; Colaprico, A.; Bilbao, D.; Figueroa, M.E.; LaCava, J.; Shiekhattar, R.; et al. Polycomb complexes associate with enhancers and promote oncogenic transcriptional programs in cancer through multiple mechanisms. *Nat. Commun.* **2018**, *9*, 3377. [CrossRef] [PubMed]

164. Grosselin, K.; Durand, A.; Marsolier, J.; Poitou, A.; Marangoni, E.; Nemati, F.; Dahmani, A.; Lameiras, S.; Reyal, F.; Frenoy, O.; et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **2019**, *51*, 1060. [CrossRef]

165. Solomon, M.J.; Larsen, P.L.; Varshavsky, A. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell* **1988**, *53*, 937–947. [CrossRef]

166. Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*, 1497–1502. [CrossRef]

167. Landt, S.G.; Marinov, G.K.; Kundaje, A.; Kheradpour, P.; Pauli, F.; Batzoglou, S.; Bernstein, B.E.; Bickel, P.; Brown, J.B.; Cayting, P.; et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **2012**, *22*, 1813–1831. [CrossRef]

168. Hoffman, E.A.; Frey, B.L.; Smith, L.M.; Auble, D.T. Formaldehyde crosslinking: A tool for the study of chromatin complexes. *J. Biol. Chem.* **2015**, *290*, 26404–26411. [CrossRef] [PubMed]

169. Pepke, S.; Wold, B.; Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **2009**, *6*, S22. [CrossRef] [PubMed]

170. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoute, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. Model-based analysis of ChIP-Seq (MACS). *Genome Boil.* **2008**, *9*, R137. [CrossRef] [PubMed]

171. Feng, J.; Liu, T.; Zhang, Y. Using MACS to identify peaks from ChIP-seq data. *Curr. Protocols Bioinform.* **2011**, *34*, 2–14. [CrossRef]

172. Kharchenko, P.V.; Tolstorukov, M.Y.; Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **2008**, *26*, 1351. [CrossRef]

173. Bell, O.; Tiwari, V.K.; Thomä, N.H.; Schübeler, D. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.* **2011**, *12*, 554. [CrossRef]

174. Britton, E.; Rogerson, C.; Mehta, S.; Li, Y.; Li, X.; Fitzgerald, R.C.; Ang, Y.S.; Sharrocks, A.D.; et al. Open chromatin profiling identifies AP1 as a transcriptional regulator in oesophageal adenocarcinoma. *PLoS Genet.* **2017**, *13*, e1006879. [CrossRef]

175. Song, L.; Crawford, G.E. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protocols* **2010**, *2010*. [CrossRef]

176. Schones, D.E.; Cui, K.; Cuddapah, S.; Roh, T.Y.; Barski, A.; Wang, Z.; Wei, G.; Zhao, K. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **2008**, *132*, 887–898. [CrossRef]

177. Wei, G.; Hu, G.; Cui, K.; Zhao, K. Genome-wide mapping of nucleosome occupancy, histone modifications, and gene expression using next-generation sequencing technology. In *Methods in Enzymology*; Elsevier: Amsterdam, The Netherlands, 2012; Volume 513, pp. 297–313.

178. Simon, J.M.; Giresi, P.G.; Davis, I.J.; Lieb, J.D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat. Protocols* **2012**, *7*, 256. [CrossRef]

179. Buenrostro, J.D.; Wu, B.; Chang, H.Y.; Greenleaf, W.J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protocols Mol. Boil.* **2015**, *109*, 21–29.

180. Chen, X.; Shen, Y.; Draper, W.; Buenrostro, J.D.; Litzenburger, U.; Cho, S.W.; Satpathy, A.T.; Carter, A.C.; Ghosh, R.P.; East-Seletsky, A.; et al. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nature Methods* **2016**, *13*, 1013. [CrossRef] [PubMed]

181. Meyer, C.A.; Liu, X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* **2014**, *15*, 709. [CrossRef] [PubMed]

182. Amemiya, H.M.; Kundaje, A.; Boyle, A.P. The ENCODE blacklist: Identification of problematic regions of the genome. *Sci. Rep.* **2019**, *9*, 9354. [CrossRef] [PubMed]

183. Bernstein, B.E.; Stamatoyannopoulos, J.A.; Costello, J.F.; Ren, B.; Milosavljevic, A.; Meissner, A.; Kellis, M.; Marra, M.A.; Beaudet, A.L.; Ecker, J.R.; et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **2010**, *28*, 1045. [CrossRef]

184. Weirauch, M.T.; Yang, A.; Albu, M.; Cote, A.G.; Montenegro-Montero, A.; Drewe, P.; Najafabadi, H.S.; Lambert, S.A.; Mann, I.; Cook, K.; et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **2014**, *158*, 1431–1443. [CrossRef]

185. Baek, S.; Goldstein, I.; Hager, G.L. Bivariate genomic footprinting detects changes in transcription factor activity. *Cell Rep.* **2017**, *19*, 1710–1722. [CrossRef]

186. He, H.H.; Meyer, C.A.; Chen, M.W.; Zang, C.; Liu, Y.; Rao, P.K.; Fei, T.; Xu, H.; Long, H.; Liu, X.S.; et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **2014**, *11*, 73. [CrossRef]

187. Dekker, J.; Rippe, K.; Dekker, M.; Kleckner, N. Capturing chromosome conformation. *Science* **2002**, *295*, 1306–1311. [CrossRef]

188. Cremer, T.; Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2001**, *2*, 292. [CrossRef]

189. Huber, D.; von Voithenberg, L.V.; Kaigala, G. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro Nano Eng.* **2018**, *1*, 15–24. [CrossRef]

190. Dostie, J.; Richmond, T.A.; Arnaout, R.A.; Selzer, R.R.; Lee, W.L.; Honan, T.A.; Rubio, E.D.; Krumm, A.; Lamb, J.; Nusbaum, C.; et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **2006**, *16*, 1299–1309. [CrossRef] [PubMed]

191. Simonis, M.; Klous, P.; Splinter, E.; Moshkin, Y.; Willemsen, R.; De Wit, E.; Van Steensel, B.; De Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat.Genet.* **2006**, *38*, 1348. [CrossRef] [PubMed]

192. Van Berkum, N.L.; Lieberman-Aiden, E.; Williams, L.; Imakaev, M.; Gnirke, A.; Mirny, L.A.; Dekker, J.; Lander, E.S. Hi-C: A method to study the three-dimensional architecture of genomes. *JoVE* **2010**. [CrossRef] [PubMed]

193. Guelen, L.; Pagie, L.; Brasset, E.; Meuleman, W.; Faza, M.B.; Talhout, W.; Eussen, B.H.; de Klein, A.; Wessels, L.; de Laat, W.; et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **2008**, *453*, 948. [CrossRef]

194. Peric-Hupkes, D.; van Steensel, B. Role of the nuclear lamina in genome organization and gene expression. In *Cold Spring Harbor Symposia on Quantitative Biology*; Cold Spring Harbor Laboratory Press: New York, NY, USA, 2010; Volume 75, pp. 517–524.

195. Lieberman-Aiden, E.; Van Berkum, N.L.; Williams, L.; Imakaev, M.; Ragoczy, T.; Telling, A.; Amit, I.; Lajoie, B.R.; Sabo, P.J.; Dorschner, M.O.; et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **2009**, *326*, 289–293. [CrossRef]

196. Dixon, J.R.; Selvaraj, S.; Yue, F.; Kim, A.; Li, Y.; Shen, Y.; Hu, M.; Liu, J.S.; Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **2012**, *485*, 376. [CrossRef]

197. Lajoie, B.R.; Dekker, J.; Kaplan, N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* **2015**, *72*, 65–75. [CrossRef]

198. Rowley, M.J.; Corces, V.G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **2018**, *19*, 789–800. [CrossRef]

199. Fullwood, M.J.; Liu, M.H.; Pan, Y.F.; Liu, J.; Xu, H.; Mohamed, Y.B.; Orlov, Y.L.; Velkov, S.; Ho, A.; Mei, P.H.; et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **2009**, *462*, 58. [CrossRef]

200. Mifsud, B.; Tavares-Cadete, F.; Young, A.N.; Sugar, R.; Schoenfelder, S.; Ferreira, L.; Wingett, S.W.; Andrews, S.; Grey, W.; Ewels, P.A.; et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **2015**, *47*, 598. [CrossRef] [PubMed]

201. Rao, S.S.; Huntley, M.H.; Durand, N.C.; Stamenova, E.K.; Bochkov, I.D.; Robinson, J.T.; Sanborn, A.L.; Machol, I.; Omer, A.D.; Lander, E.S.; et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **2014**, *159*, 1665–1680. [CrossRef] [PubMed]

202. Belaghzal, H.; Dekker, J.; Gibcus, J.H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **2017**, *123*, 56–65. [CrossRef] [PubMed]

203. Diaz, N.; Kruse, K.; Erdmann, T.; Staiger, A.M.; Ott, G.; Lenz, G.; Vaquerizas, J.M. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat. Commun.* **2018**, *9*, 4938. [CrossRef]

204. Belaghzal, H.; Borrman, T.; Stephens, A.D.; Lafontaine, D.L.; Venev, S.V.; Marko, J.F.; Weng, Z.; Dekker, J. Compartment-dependent chromatin interaction dynamics revealed by liquid chromatin Hi-C. *bioRxiv* **2019**, 704957. [CrossRef]

205. Furlong, E. The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Boil.* **2019**, *20*, 327–337.

206. Eres, I.E.; Luo, K.; Hsiao, C.J.; Blake, L.E.; Gilad, Y. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genet.* **2019**, *15*, 1–33. doi:10.1371/journal.pgen.1008278. [CrossRef]

207. Ghavi-Helm, Y.; Jankowski, A.; Meiers, S.; Viales, R.R.; Korbel, J.O.; Furlong, E.E. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **2019**, *51*, 1272–1282. [CrossRef]

208. Rodriguez, J.; Ren, G.; Day, C.R.; Zhao, K.; Chow, C.C.; Larson, D.R. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell* **2019**, *176*, 213–226. [CrossRef]

209. Fudenberg, G.; Getz, G.; Meyerson, M.; Mirny, L.A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **2011**, *29*, 1109. [CrossRef]

210. Schuster-Böckler, B.; Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **2012**, *488*, 504. [CrossRef] [PubMed]

211. Engreitz, J.M.; Agarwala, V.; Mirny, L.A. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS ONE* **2012**, *7*, e44196. [CrossRef] [PubMed]

212. Finn, E.H.; Pegoraro, G.; Brandão, H.B.; Valton, A.L.; Oomen, M.E.; Dekker, J.; Mirny, L.; Misteli, T. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **2019**, *176*, 1502–1515. [CrossRef] [PubMed]

213. Harewood, L.; Kishore, K.; Eldridge, M.D.; Wingett, S.; Pearson, D.; Schoenfelder, S.; Collins, V.P.; Fraser, P. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Boil.* **2017**, *18*, 125. [CrossRef]

214. Chakraborty, A.; Ay, F. Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics* **2017**, *34*, 338–345. [CrossRef]

215. Katainen, R.; Dave, K.; Pitkänen, E.; Palin, K.; Kivioja, T.; Välimäki, N.; Gylfe, A.E.; Ristolainen, H.; Hänninen, U.A.; Cajuso, T.; et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **2015**, *47*, 818. [CrossRef]

216. Gröschel, S.; Sanders, M.A.; Hoogenboezem, R.; de Wit, E.; Bouwman, B.A.; Erpelinck, C.; van der Velden, V.H.; Havermans, M.; Avellino, R.; van Lom, K.; et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **2014**, *157*, 369–381. [CrossRef]

217. Flavahan, W.A.; Drier, Y.; Liau, B.B.; Gillespie, S.M.; Venteicher, A.S.; Stemmer-Rachamimov, A.O.; Suvà, M.L.; Bernstein, B.E. Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **2016**, *529*, 110. [CrossRef]

218. Hnisz, D.; Weintraub, A.S.; Day, D.S.; Valton, A.L.; Bak, R.O.; Li, C.H.; Goldmann, J.; Lajoie, B.R.; Fan, Z.P.; Sigova, A.A.; et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **2016**, *351*, 1454–1458. [CrossRef]

219. Cornish, A.J.; Hoang, P.H.; Dobbins, S.E.; Law, P.J.; Chubb, D.; Orlando, G.; Houlston, R.S. Identification of recurrent noncoding mutations in B-cell lymphoma using capture Hi-C. *Blood Adv.* **2019**, *3*, 21–32. [CrossRef]

220. Johnston, M.; Nikolic, A.; Ninkovic, N.; Guilhamon, P.; Cavalli, F.; Seaman, S.; Zemp, F.; Lee, J.; Abdelkareem, A.; Ellestad, K.; et al. High-resolution structural genomics reveals new therapeutic vulnerabilities in glioblastoma. *Genome Res.* **2019**, *29*, 1211–1222. [CrossRef] [PubMed]

221. Li, G.; Ruan, X.; Auerbach, R.K.; Sandhu, K.S.; Zheng, M.; Wang, P.; Poh, H.M.; Goh, Y.; Lim, J.; Zhang, J.; et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **2012**, *148*, 84–98. [CrossRef] [PubMed]

222. Rafique, S.; Thomas, J.S.; Sproul, D.; Bickmore, W.A. Estrogen-induced chromatin decondensation and nuclear re-organization linked to regional epigenetic regulation in breast cancer. *Genome Boil.* **2015**, *16*, 145. [CrossRef] [PubMed]

223. Barutcu, A.R.; Lajoie, B.R.; McCord, R.P.; Tye, C.E.; Hong, D.; Messier, T.L.; Browne, G.; van Wijnen, A.J.; Lian, J.B.; Stein, J.L.; et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Boil.* **2015**, *16*, 214. [CrossRef]

224. Nagai, L.A.E.; Park, S.J.; Nakai, K. Analyzing the 3D chromatin organization coordinating with gene expression regulation in B-cell lymphoma. *BMC Med. Genom.* **2019**, *11*, 127. [CrossRef]

225. Ay, F.; Noble, W.S. Analysis methods for studying the 3D architecture of the genome. *Genome Boil.* **2015**, *16*, 183. [CrossRef]

226. Yaffe, E.; Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **2011**, *43*, 1059. [CrossRef]

227. Imakaev, M.; Fudenberg, G.; McCord, R.P.; Naumova, N.; Goloborodko, A.; Lajoie, B.R.; Dekker, J.; Mirny, L.A. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **2012**, *9*, 999. [CrossRef]

228. Sinkhorn, R.; Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pac. J. Math.* **1967**, *21*, 343–348. [CrossRef]

229. Chen, D.; Fu, L.Y.; Zhang, Z.; Li, G.; Zhang, H.; Jiang, L.; Harrison, A.P.; Shanahan, H.P.; Klukas, C.; Zhang, H.Y.; et al. Dissecting the chromatin interactome of microRNA genes. *Nucleic Acids Res.* **2013**, *42*, 3028–3043. [CrossRef]

230. Goh, Y.; Fullwood, M.J.; Poh, H.M.; Peh, S.Q.; Ong, C.T.; Zhang, J.; Ruan, X.; Ruan, Y. Chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. *JoVE* **2012**, *62*, e3770. [CrossRef] [PubMed]

231. Li, G.; Sun, T.; Chang, H.; Cai, L.; Hong, P.; Zhou, Q. Chromatin interaction analysis with updated ChIA-PET Tool (V3). *Genes* **2019**, *10*, 554. [CrossRef] [PubMed]

232. He, C.; Zhang, M.Q.; Wang, X. MICC: An R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics* **2015**, *31*, 3832–3834. [CrossRef] [PubMed]

233. Huang, W.; Medvedovic, M.; Zhang, J.; Niu, L. ChIAPoP: A new tool for ChIA-PET data analysis. *Nucleic Acids Res.* **2019**, *47*, e37. [CrossRef] [PubMed]

234. Maman, S.; Witz, I.P. A history of exploring cancer in context. *Nat. Rev. Cancer* **2018**, *18*, 359. [CrossRef] [PubMed]

235. Liberali, P.; Snijder, B.; Pelkmans, L. Single-cell and multivariate approaches in genetic perturbation screens. *Nat. Rev. Genet.* **2015**, *16*, 18. [CrossRef] [PubMed]

236. Luo, B.; Cheung, H.W.; Subramanian, A.; Sharifnia, T.; Okamoto, M.; Yang, X.; Hinkle, G.; Boehm, J.S.; Beroukhim, R.; Weir, B.A.; et al. Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 20380–20385. [CrossRef]

237. Gaj, T.; Gersbach, C.A.; Barbas III, C.F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **2013**, *31*, 397–405. [CrossRef]

238. Larson, M.H.; Gilbert, L.A.; Wang, X.; Lim, W.A.; Weissman, J.S.; Qi, L.S. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protocols* **2013**, *8*, 2180. [CrossRef]

239. Tadić, V.; Josipović, G.; Zoldoš, V.; Vojta, A. CRISPR/Cas9-based epigenome editing: An overview of dCas9-based tools with special emphasis on off-target activity. *Methods* **2019**, *164*, 109–119. [CrossRef]

240. Hannon, G.J. RNA interference. *Nature* **2002**, *418*, 244. [CrossRef]

241. Abudayyeh, O.O.; Gootenberg, J.S.; Konermann, S.; Joung, J.; Slaymaker, I.M.; Cox, D.B.; Shmakov, S.; Makarova, K.S.; Semenova, E.; Minakhin, L.; et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **2016**, *353*, aaf5573. [CrossRef]

242. Dixit, A.; Parnas, O.; Li, B.; Chen, J.; Fulco, C.P.; Jerby-Arnon, L.; Marjanovic, N.D.; Dionne, D.; Burks, T.; Raychowdhury, R.; et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **2016**, *167*, 1853–1866. [CrossRef]

243. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337*, 816–821. [CrossRef]

244. Doench, J.G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* **2018**, *19*, 67. [CrossRef]

245. McGinn, J.; Marraffini, L.A. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat. Rev. Microbiol.* **2019**, *17*, 7–12. [CrossRef]

246. Sternberg, S.H.; Doudna, J.A. Expanding the biologist?s toolkit with CRISPR-Cas9. *Mol. Cell* **2015**, *58*, 568–574. [CrossRef]

247. Knott, G.J.; Doudna, J.A. CRISPR-Cas guides the future of genetic engineering. *Science* **2018**, *361*, 866–869. [CrossRef]

248. Gaudelli, N.M.; Komor, A.C.; Rees, H.A.; Packer, M.S.; Badran, A.H.; Bryson, D.I.; Liu, D.R. Programmable base editing of A? T to G? C in genomic DNA without DNA cleavage. *Nature* **2017**, *551*, 464. [CrossRef]

249. Nelles, D.A.; Fang, M.Y.; O?Connell, M.R.; Xu, J.L.; Markmiller, S.J.; Doudna, J.A.; Yeo, G.W. Programmable RNA tracking in live cells with CRISPR/Cas9. *Cell* **2016**, *165*, 488–496. [CrossRef]

250. Huang, C.H.; Lee, K.C.; Doudna, J.A. Applications of CRISPR-Cas enzymes in cancer therapeutics and detection. *Trends Cancer* **2018**, *4*, 499–512. [CrossRef]

251. O'Loughlin, T.A.; Gilbert, L.A. Functional genomics for cancer research: Applications in vivo and in vitro. *Annu. Rev. Cancer Biol.* **2019**, *3*, 345–363. [CrossRef]

252. Fulco, C.P.; Munschauer, M.; Anyoha, R.; Munson, G.; Grossman, S.R.; Perez, E.M.; Kane, M.; Cleary, B.; Lander, E.S.; Engreitz, J.M. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **2016**, *354*, 769–773. [CrossRef]

253. Benabdallah, N.S.; Williamson, I.; Illingworth, R.S.; Kane, L.; Boyle, S.; Sengupta, D.; Grimes, G.R.; Therizols, P.; Bickmore, W.A. Decreased enhancer-promoter proximity accompanying enhancer activation. *Mol. Cell* **2019**. [CrossRef]

254. Charlesworth, C.T.; Deshpande, P.S.; Dever, D.P.; Camarena, J.; Lemgart, V.T.; Cromer, M.K.; Vakulskas, C.A.; Collingwood, M.A.; Zhang, L.; Bode, N.M.; et al. Identification of preexisting adaptive immunity to Cas9 proteins in humans. *Nat. Med.* **2019**, *25*, 249. [CrossRef]

255. Fu, Y.; Foden, J.A.; Khayter, C.; Maeder, M.L.; Reyon, D.; Joung, J.K.; Sander, J.D. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **2013**, *31*, 822. [CrossRef]

256. Kleinstiver, B.P.; Prew, M.S.; Tsai, S.Q.; Topkar, V.V.; Nguyen, N.T.; Zheng, Z.; Gonzales, A.P.; Li, Z.; Peterson, R.T.; Yeh, J.R.J.; et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **2015**, *523*, 481. [CrossRef]

257. Li, Y.; Mendiratta, S.; Ehrhardt, K.; Kashyap, N.; White, M.A.; Bleris, L. Exploiting the CRISPR/Cas9 PAM constraint for single-nucleotide resolution interventions. *PLoS ONE* **2016**, *11*, e0144970. [CrossRef]

258. Tsai, S.Q.; Zheng, Z.; Nguyen, N.T.; Liebers, M.; Topkar, V.V.; Thapar, V.; Wyvekens, N.; Khayter, C.; Iafrate, A.J.; Le, L.P.; et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **2015**, *33*, 187. [CrossRef]

259. Nakamura, M.; Srinivasan, P.; Chavez, M.; Carter, M.A.; Dominguez, A.A.; La Russa, M.; Lau, M.B.; Abbott, T.R.; Xu, X.; Zhao, D.; et al. Anti-CRISPR-mediated control of gene editing and synthetic circuits in eukaryotic cells. *Nat. Commun.* **2019**, *10*, 194. [CrossRef]

260. Hernández-Lemus, E. A complex path(way) to cancer phenomenology. In *Cancer: A Complex Disease*; Copit Arxives: Mexico City, Mexico, 2018; pp. 19–41.

261. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **2014**, *11*, 333. [CrossRef]

262. Tini, G.; Marchetti, L.; Priami, C.; Scott-Boyer, M.P. Multi-omics integration—A comparison of unsupervised clustering methodologies. *Briefings Bioinform.* **2017**, *20*, 1269–1279. [CrossRef]

263. Kim, D.; Shin, H.; Sohn, K.A.; Verma, A.; Ritchie, M.D.; Kim, J.H. Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction. *Methods* **2014**, *67*, 344–353. [CrossRef]

264. Berger, A.C.; Korkut, A.; Kanchi, R.S.; Hegde, A.M.; Lenoir, W.; Liu, W.; Liu, Y.; Fan, H.; Shen, H.; Ravikumar, V.; et al. A comprehensive Pan-Cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **2018**. [CrossRef]

265. Kristensen, V.N.; Lingjærde, O.C.; Russnes, H.G.; Vollan, H.K.M.; Frigessi, A.; Børresen-Dale, A.L. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* **2014**, *14*, 299–313. doi:10.1038/nrc3721. [CrossRef]

266. De Tayrac, M.; Lê, S.; Aubry, M.; Mosser, J.; Husson, F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genom.* **2009**, *10*, 32. [CrossRef]

267. Bersanelli, M.; Mosca, E.; Remondini, D.; Giampieri, E.; Sala, C.; Castellani, G.; Milanesi, L. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinform.* **2016**, *17*, S15. [CrossRef]

268. Huang, S.; Chaudhary, K.; Garmire, L.X. More is better: Recent progress in multi-omics data integration methods. *Front. Genet.* **2017**, *8*, 84. [CrossRef]

269. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer Science + Business Media: New York, NY, USA, 2013; Volume 112.

270. Kirpich, A.; Ainsworth, E.A.; Wedow, J.M.; Newman, J.R.; Michailidis, G.; McIntyre, L.M. Variable selection in omics data: A practical evaluation of small sample sizes. *PLoS ONE* **2018**, *13*, e0197910. [CrossRef]

271. Huang, S.; Xu, W.; Hu, P.; Lakowski, T.M. Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer. *Cancers* **2019**, *11*, 507. [CrossRef]

272. Setty, M.; Helmy, K.; Khan, A.A.; Silber, J.; Arvey, A.; Neezen, F.; Agius, P.; Huse, J.T.; Holland, E.C.; Leslie, C.S. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol. Syst. Biol.* **2012**, *8*, 605. [CrossRef] [PubMed]

273. Sohn, K.A.; Kim, D.; Lim, J.; Kim, J.H. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst. Biol.* **2013**, *7*, S9. [CrossRef] [PubMed]

274. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]

275. Stein-O'Brien, G.L.; Arora, R.; Culhane, A.C.; Favorov, A.V.; Garmire, L.X.; Greene, C.S.; Goff, L.A.; Li, Y.; Ngom, A.; Ochs, M.F.; et al. Enter the matrix: Factorization uncovers knowledge from omics. *Trends Genet.* **2018**, *34*, 790–805. [CrossRef]

276. Meng, C.; Kuster, B.; Culhane, A.C.; Gholami, A.M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform.* **2014**, *15*, 162. [CrossRef]

277. Lock, E.F.; Hoadley, K.A.; Marron, J.S.; Nobel, A.B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **2013**, *7*, 523. [CrossRef]

278. Feng, Q.; Jiang, M.; Hannig, J.; Marron, J. Angle-based joint and individual variation explained. *J. Multivar. Anal.* **2018**, *166*, 241–265. [CrossRef]

279. Conesa, A.; Prats-Montalbán, J.M.; Tarazona, S.; Nueda, M.J.; Ferrer, A. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 101–111. [CrossRef]

280. Lê Cao, K.A.; Boitard, S.; Besse, P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform.* **2011**, *12*, 253. [CrossRef]

281. Li, W.; Zhang, S.; Liu, C.C.; Zhou, X.J. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* **2012**, *28*, 2458–2466. [CrossRef]

282. Rohart, F.; Gautier, B.; Singh, A.; Le Cao, K.A. mixOmics: An R package for omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef] [PubMed]

283. Neto, E.C.; Bare, J.C.; Margolin, A.A. Simulation studies as designed experiments: The comparison of penalized regression models in the "large p, small n" setting. *PLoS ONE* **2014**, *9*, e107957.

284. Pineda, S.; Real, F.X.; Kogevinas, M.; Carrato, A.; Chanock, S.J.; Malats, N.; Van Steen, K. Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. *PLoS Genet.* **2015**, *11*, e1005689. [CrossRef] [PubMed]

285. Bravo-Merodio, L.; Williams, J.A.; Gkoutos, G.V.; Acharjee, A. Omics biomarker identification pipeline for translational medicine. *J. Transl. Med.* **2019**, *17*, 155. [CrossRef]

286. Liu, J.; Liang, G.; Siegmund, K.D.; Lewinger, J.P. Data integration by multi-tuning parameter elastic net regression. *BMC Bioinform.* **2018**, *19*, 369. [CrossRef]

287. OConnell, M.J.; Lock, E.F. R.JIVE for exploration of multi-source molecular data. *Bioinformatics* **2016**, *32*, 2877–2879. [CrossRef]

288. Simon, R.M.; Subramanian, J.; Li, M.C.; Menezes, S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings Bioinform.* **2011**, *12*, 203–214. [CrossRef]

289. Waldron, L.; Pintilie, M.; Tsao, M.S.; Shepherd, F.A.; Huttenhower, C.; Jurisica, I. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics* **2011**, *27*, 3399–3406. [CrossRef]

290. Barabasi, A.L.; Oltvai, Z.N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101. [CrossRef]

291. Espinal-Enriquez, J.; Fresno, C.; Anda-Jauregui, G.; Hernandez-Lemus, E. RNA-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep.* **2017**, *7*, 1760. [CrossRef]

292. Mejía-Pedroza, R.A.; Espinal-Enríquez, J.; Hernández-Lemus, E. Pathway-based drug repositioning for breast cancer molecular subtypes. *Front. Pharmacol.* **2018**, *9*, 905. [CrossRef]

293. De Anda-Jáuregui, G.; Velázquez-Caldelas, T.E.; Espinal-Enríquez, J.; Hernández-Lemus, E. Transcriptional network architecture of breast cancer molecular subtypes. *Front. Physiol.* **2016**, *7*, 568. [CrossRef] [PubMed]

294. Alcalá-Corona, S.A.; de Anda-Jáuregui, G.; Espinal-Enríquez, J.; Hernández-Lemus, E. Network modularity in breast cancer molecular subtypes. *Front. Physiol.* **2017**, *8*, 915. [CrossRef] [PubMed]

295. Costa, R.L.; Boroni, M.; Soares, M.A. Distinct co-expression networks using multi-omic data reveal novel interventional targets in HPV-positive and negative head-and-neck squamous cell cancer. *Sci. Rep.* **2018**, *8*, 15254. [CrossRef] [PubMed]

296. Gibbs, D.L.; Gralinski, L.E.; Baric, R.S.; McWeeney, S.K. Multi-omic network signatures of disease. *Front. Physiol.* **2014**, *4*, 309. [CrossRef] [PubMed]

297. Tasaki, S.; Gaiteri, C.; Mostafavi, S.; Yu, L.; Wang, Y.; De Jager, P.L.; Bennett, D.A. Multi-omic directed networks describe features of gene regulation in aged brains and expand the set of genes driving cognitive decline. *Front. Physiol.* **2018**, *9*, 294. [CrossRef] [PubMed]

298. Pinu, F.R.; Beale, D.J.; Paten, A.M.; Kouremenos, K.; Swarup, S.; Schirra, H.J.; Wishart, D. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites* **2019**, *9*, 76. [CrossRef]

299. Yugi, K.; Kubota, H.; Hatano, A.; Kuroda, S. Trans-omics: How to reconstruct biochemical networks across multiple omic layers. *Trends Biotechnol.* **2016**, *34*, 276–290. [CrossRef]

300. Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer networks. *J. Complex Netw.* **2014**, *2*, 203–271. [CrossRef]

301. Boccaletti, S.; Bianconi, G.; Criado, R.; Del Genio, C.I.; Gómez-Gardenes, J.; Romance, M.; Sendina-Nadal, I.; Wang, Z.; Zanin, M. The structure and dynamics of multilayer networks. *Phys. Rep.* **2014**, *544*, 1–122. [CrossRef]

302. De Domenico, M.; Nicosia, V.; Arenas, A.; Latora, V. Structural reducibility of multilayer networks. *Nat. Commun.* **2015**, *6*, 6864. [CrossRef] [PubMed]

303. Hernández-Lemus, E.; Espinal-Enríquez, J.; de Anda-Jáuregui, G. Probabilistic multilayer networks. *arXiv* **2018**, arXiv:1808.07857.