


Integrative approaches based on genomic techniques in the functional studies on enhancers

Qilin Wang, Junyou Zhang, Zhaoshuo Liu, Yingying Duan and Chunyan Li 

Corresponding author: Chunyan Li, School of Engineering Medicine, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, China.
Tel.: +86-10-82313101; Fax: +86-10-82313101. E-mail: lichunyan@buaa.edu.cn

With the development of sequencing technology and the dramatic drop in sequencing cost, the functions of noncoding genes are being characterized in a wide variety of fields (e.g. biomedicine). Enhancers are noncoding DNA elements with vital transcription regulation functions. Tens of thousands of enhancers have been identified in the human genome; however, the location, function, target genes and regulatory mechanisms of most enhancers have not been elucidated thus far. As high-throughput sequencing techniques have leapt forwards, omics approaches have been extensively employed in enhancer research. Multidimensional genomic data integration enables the full exploration of the data and provides novel perspectives for screening, identification and characterization of the function and regulatory mechanisms of unknown enhancers. However, multidimensional genomic data are still difficult to integrate genome wide due to complex varieties, massive amounts, high rarity, etc. To facilitate the appropriate methods for studying enhancers with high efficacy, we delineate the principles, data processing modes and progress of various omics approaches to study enhancers and summarize the applications of traditional machine learning and deep learning in multi-omics integration in the enhancer field. In addition, the challenges encountered during the integration of multiple omics data are addressed. Overall, this review provides a comprehensive foundation for enhancer analysis.

Keywords: enhancer; multi-omics; high-throughput data analysis; data integration; machine learning

INTRODUCTION

The maintenance of transcriptional homeostasis is crucial to the development and growth of living things [1]. Transcriptional homeostasis is dependent on the interactions between transcription factors (TFs) and cis-regulatory elements (e.g. promoters and enhancers) [2]. In the 1980s, enhancers were first discovered in simian virus 40 (SV40) [3]. Subsequently, researchers have gradually characterized different types of enhancers, and various techniques have been developed to predict and study the function of enhancers (Figure 1). In 2004, Benjamin predicted enhancers in the *Drosophila* genome based on sequence conservation, initiating the application of bioinformatics methods in enhancer research [4]. The eRNA and Super-enhancer were discovered in 2010 and 2013, respectively, further enriching the understanding of enhancers [5, 6]. In 2013, the introduction of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas9 technology greatly accelerated the validation of enhancer functions [7]. Since 2016, with the rapid development of the machine learning field, scientists have gradually adopted neural network technology into enhancer research and have developed numerous models and softwares to study enhancers [8, 9]. The development of omics technology has brought a breakthrough in enhancer research. The dissection of the underlying transcriptional regulation of enhancers provides new insights into the complexity of transcriptional regulation.

Currently, omics approaches for enhancer research focus on four key questions (Figure 2). (i) How are enhancers identified? (ii) What induces the changes in enhancer activity? (iii) How do enhancers interact with their targets in the complicated 3D structure of the genome? (iv) What regulates the production and function of eRNA (enhancer RNA)? Since multiple biological processes are involved in the above four questions, genomic sequencing data provide opportunities to study such complicated issues by genomic sequencing data integration.

Enhancer-associated sequencing technologies can be divided into four categories: genomics, epigenomics, transcriptomics and gene-editing technology (Figure 3). Genomics focuses on gene sequences and genomic structures of enhancers. Specifically, the former explores enhancers through genome variation-phenotype correlation and gene sequence conservation, while the latter identifies potential enhancers and target genes under the 3D structure of the genome [10–13]. Epigenomics identifies enhancers from the perspectives of chromatin spatial information, DNA interaction and modification, and RNA secondary structure [14–16]. Since active enhancers are transcribed into eRNAs, the transcriptome is widely applied to characterize enhancers and enhancer-target pairs based on expression correlation [5, 17–19]. STARR-seq (self-transcribing active regulatory region sequencing) is specifically designed to evaluate enhancer activity [20, 21]. CRISPR gene editing technology has been applied prevalently

Qilin Wang is a PhD candidate at Beihang University, China. His research interests are bioinformatics, machine learning and data mining.

Junyou Zhang is a PhD candidate at Beihang University, China. His research interests focus on cancer genomics.

Zhaoshuo Liu is a master candidate at Beihang University, China. His research interests are data mining, computational biology and user modeling.

Yingying Duan is a master candidate at Beihang University, China. Her research interests focus on cancer genomics.

Chunyan Li is an associate professor at Beihang University, China. Her research interests are functional genomics on cancer and osteoporosis.

Received: August 28, 2023. Revised: October 22, 2023. Accepted: November 8, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

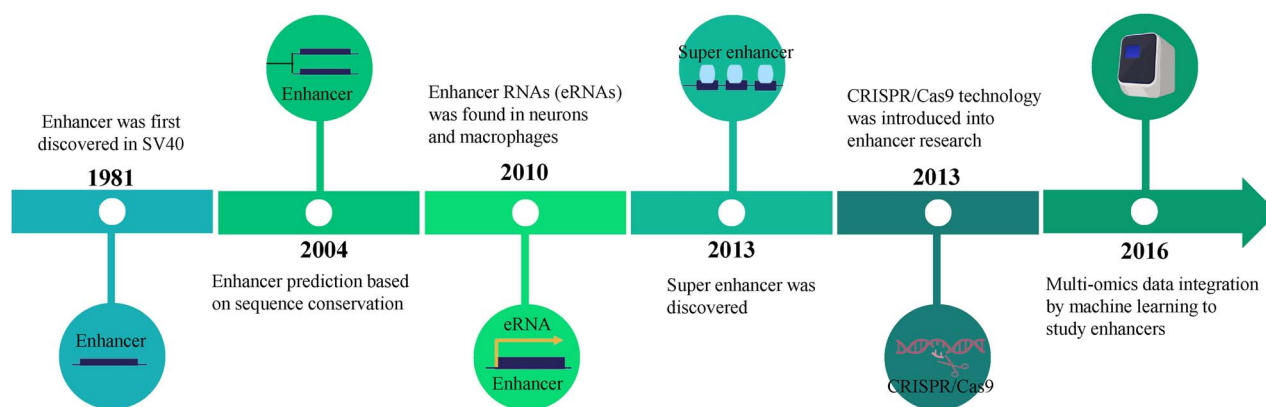


Figure 1. Timeline of enhancer research.



Figure 2. Challenges in the study of enhancers by omics methods.

to knock out/down genes or enhancers. In addition, CRISPR gene editing technology has been developed to conduct large-scale parallel screening of enhancers followed by sequencing [22].

The centermost circle represents the transcriptional regulation by enhancers on the target genes; the inner circle presents a variety of techniques, particularly sequencing; the middle layer presents molecular information acquired from each technique and the outermost displays the classification. Genomic approaches are shown in blue, epigenomic approaches are represented by pink, transcriptomic approaches are shown in yellow and gene editing technology is represented by green.

Although single omics data mining can initially screen out enhancers or genes correlated with a specific disease or phenotype, single omics analysis is subject to significant limitations, e.g. inadequate interpretation of data in a single dimension and insufficient depletion of signal noise [23]. Multi-omics analysis has the advantage of diversity and is systematic, making it more conducive to clarifying the underlying mechanisms of enhancers [24–27]. The integration of multi-omic data provides more reliable results and dramatically reduces the false-positive rate [28]. Over the past decade, various multi-omics analysis methods have been developed. However, there are still many challenges, such as the accuracy variation among different omics data, missing values, and computational and storage costs [29]. This review will discuss the data characteristics of different omics in enhancer research, the methods of multi-omics data analysis and the challenges in multi-omics research.

THE APPLICATION OF DIFFERENT OMICS DATA IN ENHANCER RESEARCH

The widespread application of sequencing technologies has provided a wealth of molecular information in the enhancer field (Table 1). To better illustrate the sources and applications of different types of molecular information, we categorized omics approaches from the perspective of the research subject: genomics, epigenomics, transcriptomics and CRISPR editing technologies (Figure 3).

Genomics

Driven by the progress of sequencing technology and the decline in sequencing cost, large-scale population genome sequencing has been initiated in many countries, and the amount of data has grown exponentially [30]. SNPs (single nucleotide polymorphisms), SVs (structural variations), CNVs (copy number variations), InDels (insertions–deletions) and other molecular information can be obtained using WGS (whole genome sequencing), WES (whole exon sequencing), WGRS (whole genome resequencing) and other genomic sequencing techniques [31]. Researchers have developed GWAS (genome-wide association study) analysis and eQTL (expression quantitative trait loci) analysis methods to study the relationship between SNPs or CNVs and phenotypes. The GWAS method can analyze millions of SNPs in the genome simultaneously, which has the advantages of high efficiency and wide coverage [32]. eQTL is used to study the relationship between gene expression level and genotype [33]. Young group summarized the results of 1675 GWAS and found 5303 SNPs associated with various diseases. The majority of SNPs are in noncoding regions (93%), and among these, 64% of the loci are enriched in enhancer regions [34]. However, the GWAS method has limitations, such as the inability to identify complex traits, the inability to assess rare genetic variants and the uncertainty of gene-phenotype associations [32]. Compared with GWAS analysis, eQTL is advantageous in exploring gene expression regulation mechanisms and gene-phenotype associations. Since eQTL information can determine genetic variants associated with gene expression levels, it can more accurately identify potential enhancer elements [35, 36]. By introducing eQTLs from the 1000 Genomes Project, Chen et al. identified 65 pairs of cancer-specific enhancer genes [36]. Chignon et al. conducted a colocalization analysis of enhancer–promoter locations with tissue eQTL locations associated with genetic coronary artery disease, evaluating the importance of genetic variability in the disease [35]. In both studies, the approach to integrate enhancers with eQTLs was a location-based approach [35, 36].

Transcription regulation is closely correlated with the 3D conformation of chromatin, which is an alternative perspective

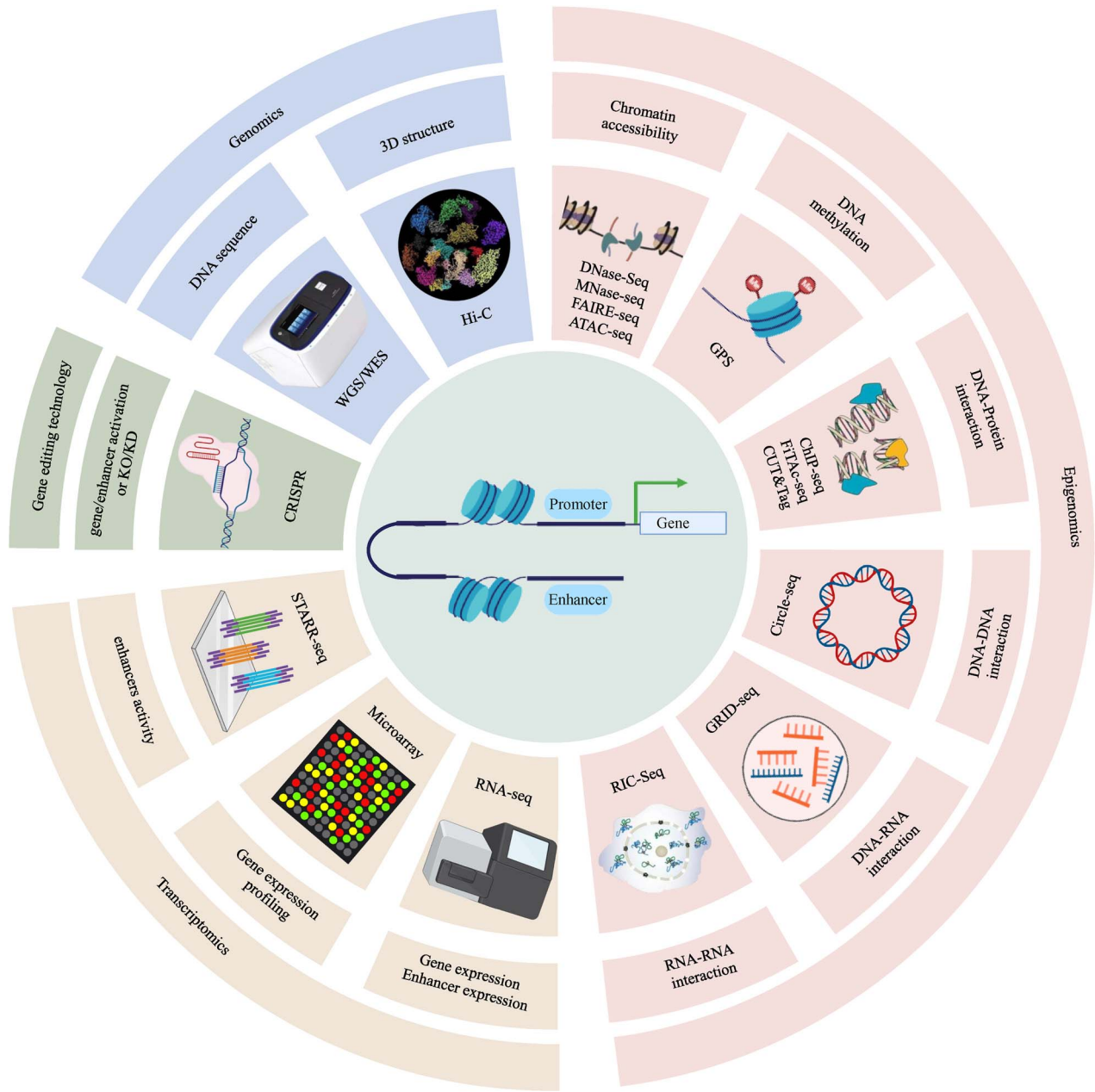


Figure 3. Applications of different omics methods and CRISPR gene editing technology in enhancer research.

Table 1: Key questions about enhancers are addressed by different molecular information

Omics methods	Molecular information	Identification	Activity	Structure	eRNA
Genomics	DNA sequence		✓	✓	
Genomics	3D structure	✓	✓	✓	
Epigenomics	Chromatin accessibility	✓	✓		
Epigenomics	DNA methylation		✓	✓	
Epigenomics	DNA-Protein interaction	✓	✓		
Epigenomics	DNA-DNA interaction	✓	✓	✓	
Epigenomics	DNA-RNA interaction	✓	✓		✓
Epigenomics	RNA secondary structure		✓	✓	✓
Transcriptomics	Gene expression		✓	✓	
Transcriptomics	Enhancer expression		✓	✓	
Transcriptomics	Enhancer activity	✓			✓
Gene editing technology	Gene/enhancer activation	✓	✓	✓	

for studying enhancers [37]. Hi-C (high-throughput chromosome conformation capture) has been the most extensively performed approach for 3D genome sequencing at the genome-wide level, with the advantages of wide coverage, high accuracy and more complete sequence positioning [38]. Since Hi-C data provide comprehensive information on chromatin interactions, they are used to determine the binding of enhancers to target genes in physical space [37]. The Hi-C derivative technologies include ChIA-PET (chromatin interaction analysis based on paired-end-tag sequencing) [39], HiChIP (in situ Hi-C followed by chromatin immunoprecipitation) [40] and PLAC-seq (proximity ligation-assisted ChIP-seq) [41]. These techniques can detect specific protein-mediated chromatin loops at high resolution, which are also used in enhancer analysis. However, for most tissues and cell lines, the Hi-C and Hi-C derivative technologies have the disadvantage of insufficient resolution [42]. The exponential growth in data volume and depth brings new analytical challenges as well [43].

Epigenomics

Epigenetics (e.g. DNA methylation, RNA modification, RNA secondary structure, histone modifications, etc.) refers to a type of regulatory mechanism on phenotypic properties by regulating gene transcription or translation processes without changing the DNA sequence [44, 45]. In enhancer studies, epigenomics methods can be divided into four categories according to the differences in research objects: chromatin accessibility, DNA modification, DNA interaction and RNA interaction (Figure 3).

Chromatin accessibility

The open state of eukaryotic chromatin is considered as a prerequisite for transcription. Four sequencing techniques have been developed to identify chromatin regions in the open state: DNase-seq (DNaseI sequencing), MNase-seq (micrococcal nuclease digestion and sequencing), FAIRE-seq (formaldehyde-assisted isolation of regulatory elements) and ATAC-seq (assay for targeting accessible chromatin with high-throughput sequencing) [46–48] (Figure 3). DNase-seq and MNase-seq are both genome sequencing techniques based on enzymatic digestion to determine chromatin accessibility. DNase-seq combines nonspecific endonuclease DNase I (Deoxyribonuclease I) to obtain DNA sequences between nucleosomes, whereas MNase-seq obtains DNA sequences wrapped around nucleosomes using micrococcal nuclease (MNase). Consistently, these techniques are used to identify active enhancers by high chromatin accessibility. However, both DNase I and MNase enzymes have sequence preferences, resulting in uneven signal distribution and false-negatives [49–52]. FAIRE-seq uses the difference in the solubility between DNA with or without nucleosome wrapping in phenol and chloroform. The DNA in nucleosome-free regions is determined by sequencing the DNA in the aqueous phase. FAIRE-seq overcomes the sequence preference of MNase and DNase I, but the low signal-to-noise ratio and the high background signal make FAIRE-seq data difficult to interpret [47, 53]. As the main sequencing technology for open chromatin so far, ATAC-seq employs the modified Tn5 transposase to randomly insert designed DNA sequences with adapter sequences into the open chromosomal regions. Fragmentation by Tn5 transposase and ligation with adapters are performed simultaneously, such that the sequencing library preparation process is notably simplified [54]. ATAC-seq has good repeatability, strong consistency and significant signals, and as few as 500 cells are needed, although mitochondrial contamination is inevitable [55]. Innovations in single-cell genomic technologies make it

possible to map regulomes in individual cells. The single-cell ATAC-seq (scATAC-seq) and single-cell DNase-seq (scDNase-seq) are two technologies for analyzing open chromatin in single cells. By adding barcode sequences to each cell, scientists are able to examine heterogeneous samples at cellular resolution [56]. In enhancer studies, chromatin accessibility analysis methods are often employed to screen potential active enhancers. For example, Chen and Liang hypothesized a negative correlation between enhancer activity and the strength of nucleosome binding. To validate the hypothesis, they integrated enhancer position information and MNase-seq data from 29 different tissues/cell types and observed a reduction in nucleosome signals on the eRNA loci compared with the flanking sequences across all 29 tissue types. Integrating these findings with RNA-seq data to determine the eRNA expression level, they identified ~200 000 new eRNA loci [57]. Through in-depth analysis of single-cell RNA sequencing (scRNA-seq) and scATAC-seq data from mouse embryonic spinal cord, an enhancer regulatory network algorithm, called eNET, successfully identified enhancers crucial to the development of spinal cord neurons [58].

DNA modification

In most cancer types, the proportion of DNA methylation in the enhancer region is negatively correlated with its activity [59]. Yu's group developed Guide Positioning Sequencing technology. By harnessing the 3' → 5' exonuclease and 5' → 3' polymerase activities of T4 DNA polymerase, methylcytosines were introduced into the 3' end of each DNA fragment. Following bisulfite treatment, the 3' read of each DNA fragment serves as a guide to determine the DNA methylation status of the paired 5' read [60]. The approach improves the efficiency and accuracy of the mapping rate, and there is no sequence preference in methylation detection [60]. By comparing the changes in DNA methylation and H3K27ac histone modification between normal liver and two liver cancer cell lines (97 L and LM3), they discovered that the DNA methylation levels were increased, the H3K27ac peaks were lost in 5 liver-specific enhancer regions, and the expression of target genes was silenced in liver cancer cells. Therefore, they concluded that aberrant DNA methylation pattern in enhancer regions may alter the activity of enhancers, resulting in alterations in the expression of target genes.

DNA interaction

The interactions between DNA and other molecules, such as proteins, DNA and RNA, modify the structure or binding affinity of DNA. These processes can also result in functional alterations in enhancers. Given the crucial role of DNA interactions in biological processes, a multitude of sequencing technologies have been devised for their study. Based on the different interaction partners, we will introduce various omics technologies from three perspectives: DNA–protein interactions, DNA–DNA interactions and DNA–RNA interactions.

DNA–protein interactions

To characterize the interaction between DNA and protein, ChIP-seq (Chromatin Immunoprecipitation sequencing), FiTac-seq (fixed-tissue ChIP-seq for H3K27ac profiling) and CUT&Tag (cleavage under targets and tagmentation) have been prevalently performed [14–16, 61–64]. ChIP-seq, first developed in 2007, has become one of the most prevalent methods for identifying the binding sites of TFs on DNA and DNA interacting with certain histone modifications to study epigenetic mechanisms [65]. However, the prolonged exposure of clinical specimens

to formalin results in excessive chemical cross-linking, which limits the isolation of soluble chromatin. Therefore, the signal intensity of ChIP-seq analysis for FFPE (formalin fixation and paraffin embedding) samples is low, and the resolution is poor [63]. Therefore, FiT-seq and FiTAc-seq were developed to obtain high-quality information on the signal distribution of H3K4me1 and H3K27ac for FFPE samples [63, 64]. Another disadvantage of ChIP-seq is the low peak signal, high background noise and sometimes uneven distribution of target DNA fragments [66]. To address the issues with ChIP-seq, in 2019, Kaya-Okur et al. developed CUT&Tag technology, which requires fewer cells, with a minimum of 60 cells, and which has library construction steps that are simplified by removing the steps of formaldehyde crosslinking and ultrasonic interruption [67]. CUT&Tag has the advantages of lower background noise, higher reading accuracy and better data repeatability [68]. In general, there are two main directions to study enhancers using sequencing data: to screen active enhancers by detecting enriched histone modifications (H3K4me1 and H3K27ac) and to characterize enhancer-target pairs by binding to certain transcription activators or coactivators [69, 70].

DNA–DNA interactions

Extrachromosomal circular DNA (eccDNA) is a circular and double-stranded molecule in the nucleus that is independent of chromosomal DNA (chrDNA). These eccDNAs can vary greatly in size, ranging from tens to millions of base pairs [71]. eccDNA interferes with the replication and expression of genes by interacting with chrDNA [71]. In addition, eccDNA functions as a mobile enhancer to increase the transcription of genome-wide target genes [72]. At present, canonical DNA sequencing can indirectly predict eccDNA through sequence information, while Circle-seq is specifically designed to detect eccDNA [73].

DNA–RNA interactions

Increasing evidence suggests that nascent RNAs mediate the chromosomal interaction between promoters and enhancers several mega-bases away in linear distance. GRID-seq (global RNA interactions with DNA by deep sequencing) is a technique for unbiased detection of DNA–RNA interactions at the genome scale [74]. GRID-seq is complementary to Hi-C in studying 3D chromatin architecture [75, 76]. However, GRID-seq requires rather deep sequencing to generate a robust contact map, which limits its application [74, 75].

RNA–RNA interactions

RNA molecules in the cell nucleus form secondary structures via intramolecular base pairing to exert their biological functions. For example, eRNA and promoter upstream antisense RNAs (also known as promoter upstream transcripts, PROMPTs) form enhancer–promoter loops to activate transcription [77]. Thus, deciphering the higher-order structure of RNA is crucial for understanding the underlying mechanisms [77, 78]. RIC-seq can accurately capture the secondary structure of RNA and identify RNA–RNA interactions through chimeric sequences. In HeLa cells, 31 genes were predicted as target genes of 7 enhancers by RIC-seq (RNA in situ conformation sequencing). Locked nucleic acid and antisense oligonucleotides were used to knock down the 7 enhancers, and the expression of 27 predicted target genes was decreased accordingly. Therefore, the prediction accuracy for target genes for enhancers was >85% based on RIC-seq. RIC-seq will be helpful to study the regulatory role of eRNA in promoter activity.

Transcriptomics

The transcriptome refers to the collection of all RNAs transcribed in a specific tissue or cell at a certain stage [79]. The most extensively employed transcriptome detection methods comprise microarray, RNA-seq, scRNA-seq, spatial transcriptome sequencing and other derivative techniques [17]. RNA-seq, the most prevalent transcriptome sequencing technology, represents low background noise, accurate quantification and higher resolution of differentially expressed genes, and has a much lower limit of detection than a standard whole genome microarray [18, 19]. However, in model organisms, microarrays are reliable and more cost effective than RNA-seq [80]. To address the different cell states within a sample, single-cell transcriptomics was developed in 2009. Nowadays, many single-cell transcriptome platforms have emerged, such as 10X Genomics, BD Rhapsody, Fluidigm C1, etc. Among these platforms, the 10X Genomics single-cell transcript platform is the most commonly used due to its high-throughput and efficiency in capturing 100–80 000 cells (per chip) [81]. The scRNA-seq and spatial transcriptome sequencing endow expression information with high accuracy and specificity at single-cell resolution, whereas the steep price and the complexities in data analysis hinder their prevalence [82, 83]. In addition to conventional RNA-seq, STARR-seq has been applied to detect enhancer activity [84]. STARR-seq is a massively parallel reporter assay that identifies transcriptional enhancers based on their activity across the genome and quantitatively assesses their activity [84].

Transcriptome sequencing (RNA-sequencing) has been performed to study the expression and genomic alterations of enhancers. First, the transcriptome provides expression information for both target genes and enhancers. Compared with mRNAs, eRNAs have the characteristics of instability and low expression level, and most eRNAs do not contain polyA tails [85, 86]. Most RNA-seq studies utilize oligo-dT enrichment to capture polyA-tailed RNAs, which results in low detection efficiency for eRNAs. scRNA-seq and spatial transcriptomics sequencing, with relatively low depth, have not yet been performed to obtain eRNA expression. In addition to RNA-seq, GRO-seq (global nuclear run-on sequencing), PRO-seq (precision nuclear run-on sequencing), CAGE-seq (cap analysis of gene expression by deep sequencing) and other RNA-seq-derived techniques have been employed to capture eRNAs [87–91].

CRISPR gene editing technology

The integration of gene editing techniques and second-generation sequencing technology implements genome-wide parallel screenings for enhancers regulating a specific phenotype. CRISPR/Cas9 technology has been applied in enhancer screening, functional verification and target gene identification [22]. Various CRISPR-derived techniques for high-throughput screening of enhancers have been developed, such as CRISPRi-FlowFISH. CRISPRi-FlowFISH integrates CRISPRi with RNA fluorescence in situ hybridization (FISH) technology. The main principle is that gRNA guides KRAB-dCas9 to bind to a specific nucleotide sequence and inhibit the transcription of the sequence 200–500 bp near the gRNA. Subsequently, RNA FISH has been used to quantitatively label single cells based on the expression level of a gene of interest. When an enhancer is targeted by gRNA, CRISPRi-FlowFISH can quantify the effect of the enhancer on the target gene(s) [92]. Furthermore, Perturb-seq (also referred to as CRISP-seq and CROP-seq) integrates multiplexed CRISPR-mediated gene inactivation with scRNA-seq to comprehensively evaluate

gene expression phenotypes for each perturbation. By designing sgRNAs (single guide RNAs) for enhancers, Perturb-seq enables simultaneous quantitative measurement of enhancer expression in many cells and a wealth of phenotypic information and greatly improves screening efficiency [93].

ENHANCER DATABASE

With the continuous growth of genomic data and experimental results, the enhancer databases have become an essential resource to study enhancers efficiently. Currently, there are more than a dozen databases that are widely used (Table 2). VISTA enhancer browser, DiseaseEnhancers and ENdb are three databases collecting enhancers experimentally validated [94–96]. So far, the VISTA enhancer browser (<https://enhancer.lbl.gov/>) has collected 1699 human or mouse noncoding elements with enhancer activity assessed in transgenic mice [94]. The DiseaseEnhancer (<https://github.com/shijianasdf/DiseaseEnhancer/tree/master>) database has collected 1059 experimentally validated disease-related enhancers from 167 human diseases based on literature [95]. And the ENdb (<https://bio.liclab.net/ENdb/index.php>) database is a manually curated enhancer database for human and mouse from 1590 published literatures, with 713 experimentally validated enhancers and their related information, including target genes, TFs, diseases and functions [96]. Cancer-specific enhancers are one of the hot topics in enhancer research. CancerEnD (<https://webs.iitd.edu.in/raghava/cancerend/>) has conducted on 18 different cancer types by RNA expression data from TCGA, providing 8599 enhancers of 8063 cancer samples [97]. CenHANCER (<http://cenhancer.chenzxlab.cn/>) has collected H3K27ac ChIP-seq data from 49 cancer types, and predicts >57 million enhancers [98]. The TCEA database (https://bioinformatics.mdanderson.org/Supplements/Super_Enhancer/TCEA_website/) has collected TCGA and GTEx RNA-seq data and provides the downloadable eRNA expression data that has been calculated [57]. In addition to cancer-specific enhancers, tissue-specific and disease-related enhancer research is another key issue in the enhancer field. The GeneHancer (<http://www.genecards.org/>) database uses an integration algorithm to eliminate redundancy and identifies >434 000 tissue-specific enhancers from multiple data sources [25]. Mutations on the DNA sequences of enhancers may cause diseases by affecting target gene expression. HACER (<http://bioinfo.vanderbilt.edu/AE/HACER/>) utilizes GWAS information on disease-related genetic variation sites to link enhancers to diseases [99]. In addition to enhancers in human, enhancer research in mouse and other mammals gains increasing attention, as well. Fantom5 (<https://fantom.gsc.riken.jp/5/>) and RAEdB (<http://www.computationalbiology.cn/RAEdB/index.php>) have predicted different enhancers in humans and mouse, respectively, through CAGE-seq and STARR-seq methods [91, 100]. EnhancerAtlas2.0 (<http://www.enhanceratlas.org/>) collected data from 12 different tissue samples and predicted enhancers for 9 different mammalian species, greatly expanding the scope of enhancer research [101]. Based on the evolutionary conservation on enhancer between species, studies on enhancers and enhancer-gene interactions were performed in other model organisms. scEnhancer (<http://enhanceratlas.net/scenhancer/>), the first database to annotate enhancers at the single-cell level, covering 14 527 776 enhancers from 1196 906 single cells in human, mouse and *Drosophila*.

MULTI-OMICS INTEGRATION METHOD

In recent years, the development of mathematics, statistics and computational science has laid the foundation for the integration of multi-omics analysis. At present, multi-omics integration methods can be divided into two categories based on whether neural networks are used: traditional machine learning models, which have the advantages of strong interpretability of algorithms and lower requirements for computing resources; and deep learning models using neural networks, which can capture complex relationships in data due to their powerful nonlinear fitting capabilities (Figure 4) [102–104]. The key factors in determining the two methods include data volume, computational resources and feature numbers. Neural networks require a large volume of data (at least thousands of samples) to avoid overfitting and abundant computational resources (hardware, software, etc.) [105]. Compared with traditional machine learning, one advantage of neural networks is that they do not require a large amount of manual labeling, and only simple data preprocessing is required for computation [106]. When choosing a method, researchers should weigh the characteristics of the issue to resolve.

Traditional machine learning models

Traditional machine learning models are algorithms that use statistics, linear algebra and optimization algorithms to extract information from existing data to build predictive models. The classical machine learning methods, such as logistic regression, random forests and naive Bayes, are used to predict and classify unknown data. Based on whether manually annotated labels are required for data, it can be divided into three types of learning: unsupervised, semi-supervised and supervised (Table 3).

Unsupervised learning

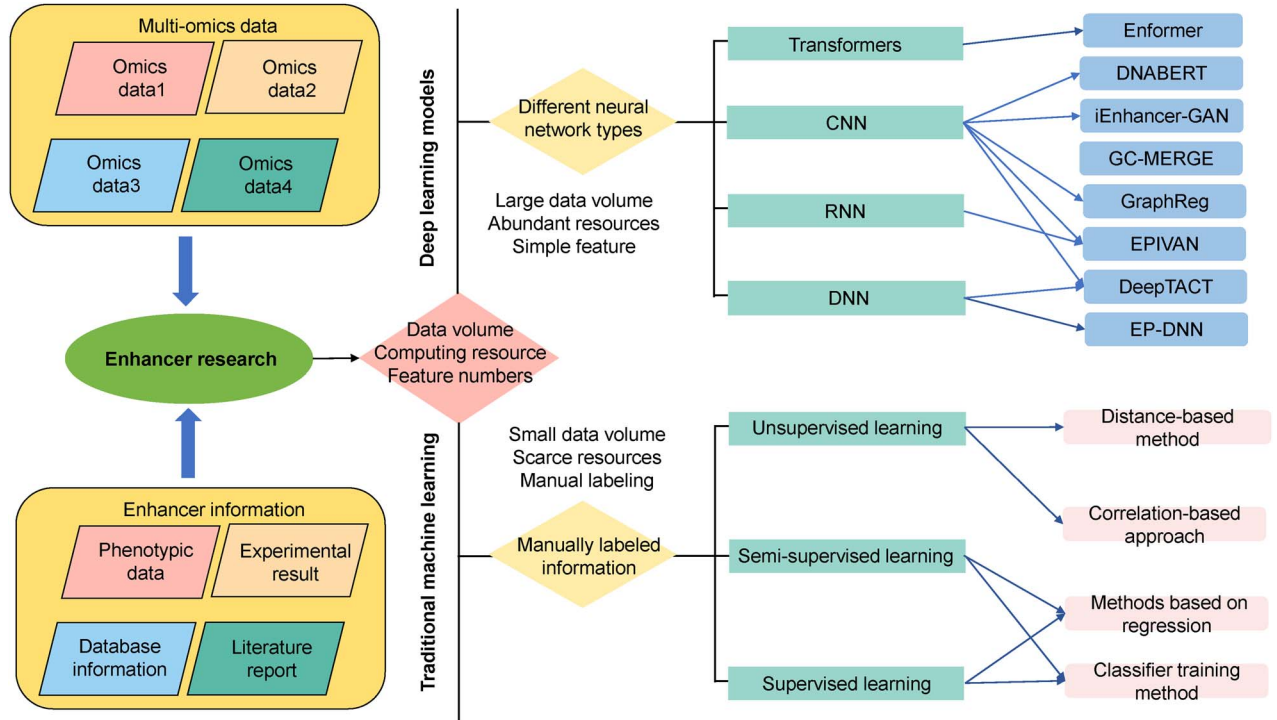
Unsupervised learning is an analytical approach that eliminates the need for prelabeled training data. The main objective of unsupervised learning is to unveil hidden patterns and establish new connections between variables within a dataset [107]. In the study of enhancers, unsupervised learning methods can be divided into distance-based methods and correlation-based methods [108].

The earliest method used to predict enhancer target genes was the distance-based method, which relies on the genomic proximity between enhancers and genes. This approach assumes that enhancers tend to regulate nearby genes in the genome [109, 110]. However, the accuracy is not high, the variation range is large and the false discovery rate (FDR) is ~40–73% [111]. Even when RNA expression data have been used to screen the enhancer's regulatory gene, its accuracy has remained low, with FDR values ranging from 53 to 77% [112]. Furthermore, the distance-based method overlooks distal regulatory interactions and the situation in which multiple enhancers target the same promoter [5]. Therefore, the distance-based method is generally used as a baseline [109]. For example, the ABC model predicts cell type-specific enhancer-target pairs based on the distance between enhancers and genes, the frequency of chromosomal contact between enhancers and promoters (by Hi-C data analysis), and enhancer activity (by DNase-seq and H3K27ac ChIP-seq) [92].

Developed from distance-based methods, correlation-based methods combine the correlation of features (e.g. histone modifications, DHS signals of enhancers and promoters, and gene transcription levels) to increase the prediction accuracy, such as ELMER and CisMapper [113–116]. ELMER identifies transcriptional targets by correlating methylation-affected enhancers with the

Table 2: Comparison of commonly used enhancer databases

Database	Species	Enhancers	eRNA	Specificity	Experimental result
CancerEnD	Human	168 464	No	Cancer	0
CenhANCER	Human	>57 000 000	No	Cancer	0
DiseaseEnhancer	Human	1059	No	Disease	1059
ENdb	Human/Mouse	713	No	Disease	713
EnhancerAtlas2.0	9 species	13 494 603	No	None	0
Fantom5	Human/Mouse	65 359	Yes	None	0
GeneHancer	Human	434 139	Yes	None	0
HACER	Human	1676 284	No	Disease	0
RAEdb	Human/Mouse	>500 000	No	None	0
scEnhancer	3 species	14 527 776	No	None	0
TCEA	Human	>300 000	Yes	Cancer	0
VISTA	Human/Mouse	3321	No	None	1699

**Figure 4.** Classification of multi-omics integration methods in enhancer research.**Table 3:** Model classification for the prediction of enhancer

Traditional machine learning	Algorithms	Tool name	Model
Unsupervised	Distance	ABC	Distance-based
Unsupervised	Correlation	ELMER	Pearson correlation
Unsupervised	Correlation	CISMAPPER	Pearson correlation
Semi-supervised	Regression	McEnhancer	Logistic regression
Semi-supervised	Classifier	DPHM	Bayesian model
Supervised	Regression	JEME	Regression-based methods
Supervised	Regression	FENRIR	Elastic net logistic regression
Supervised	Classifier	FOCS	Linear regression
Supervised	Classifier	IM-PET	Random forest
Supervised	Classifier	PETModule	Random forest
Supervised	Classifier	RIPPLE	Random forests
Supervised	Classifier	TargetFinder	Gradient tree boosting

expression of nearby genes. A nonparametric U test was used to examine the correlation degree of enhancer methylation and expression data (RNA-seq) with 10 genes upstream and 10 genes downstream of each enhancer, and all enhancer-gene pairs

with $P < 0.001$ were retained [115]. CisMapper predicts enhancer-target pairs by calculating the Pearson correlation coefficient between the log of gene expression and the log of the histone signal at the TF-binding site within 500 kb upstream of the gene

TSS [116]. CisMapper is more accurate than simple distance-based methods, with an average accuracy improvement of 2.7 times [116].

Semi-supervised and supervised learning

Semi-supervised learning uses algorithms that cover both unlabeled and labeled data for training, which is preferred when there is not enough labeled dataset available for supervised learning [117]. Compared with supervised learning, semi-supervised learning can reduce overfitting and improve the robustness of the model [117]. Supervised learning depends on high confidence positive and negative labeled training datasets (enhancers and non-enhancers, respectively). The model is usually trained to maximize the distinction between case and control sets [118]. Dependent on the algorithm applied in the model, semi-supervised and supervised learning can be divided into regression-based methods and classifier-training [109].

Regression-based methods (e.g. McEnhancer, JEME, FENRIR and FOCS) integrate enhancer and promoter features or gene expression to identify the regulatory relationship between enhancers and target genes [119–122]. McEnhancer uses a semi-supervised logistic regression model to calculate the probability of TFs binding to promoters and enhancers, and predicts the binding strength between genes and enhancers, with a prediction accuracy of 73–98% [122]. The merged regulation by multiple enhancers is considered in JEME, and sample-specific information is integrated as well to predict gene regulatory networks [121]. FENRIR integrates thousands of different epigenetic and functional genomics datasets to infer tissue-specific functional relationships between enhancers in 140 different human tissues and cell types [119]. FOCS is a statistical framework that utilizes eRNA as a marker of enhancer activity and determines enhancer–promoter interactions correlated with transcriptional activity based on information about chromatin epigenetic modifications [120].

The classifier training method uses experimentally identified enhancer–promoter interactions as the gold standard set. By learning the sequence and epigenetic modification features of the standard, a classifier can be trained to predict whether a given enhancer–promoter pair has an interaction or not [109]. DPHM, as a semi-supervised Bayesian model, predicts target genes of 47 enhancers in mice using Nkx2-5 ChIP-seq data [123]. IM-PET tool, using the random forest classifier algorithm, predicts the association between enhancers and promoters by collecting a large amount of enhancer feature data (epigenetic modification data, TF expression data, enhancer conservation data, etc.) [124]. The tool has high predictive accuracy, with an FDR reduced to ~1%, and the predicted distance between enhancers and target genes is also extended to 2 Mb [124]. PETModule, RIPPLE, TargetFinder, EAGLE and EPIP are algorithms that adopt supervised learning methods to predict enhancer-target gene interactions. Although they use different classification features, they all present good prediction performance on different datasets [112, 125–128] (Table 3).

Deep learning models

Since 2016, scientists have gradually begun to use neural network technology to study enhancers [8, 9]. Many studies have shown that neural networks have significant advantages in enhancer research, such as being able to predict across different cell types, thereby reducing computational and time costs [129, 130]. Convolutional neural networks (CNNs) have become the widely used algorithm in enhancer research, and various models such as

DNABERT [131], iEnhancer-GAN [132], GC-MERGE [133], GraphReg [134], EPIVAN [130] and DeepTACT [135] have been proposed and optimized (Figure 3). DNABERT, as a novel pre-trained bidirectional encoder representation, can reveal the potential associations between different cis-DNA by learning DNA sequence information [131]. iEnhancer-GAN integrate word embeddings and sequence generation adversarial networks to predict the binding strength of enhancer-target gene interactions [132]. DeepTACT applies a bootstrapping deep-learning model to integrate genome sequence and chromatin accessibility data to predict enhancer–promoter interactions [135]. GC-MERGE is a graph-based deep learning framework that decodes Hi-C map through graph convolutional networks to capture the potential genomic spatial structure. It models the epigenetic modification signals and DNA sequence information to predict the target genes regulated by distant enhancers [133]. GraphReg model uses CNN layers to learn 1D features of enhancer-target gene (epigenomic data, genomic DNA sequence, etc.), and then constructs different enhancer-target genes into a whole through iterative methods on 3D genomic maps (such as HiChIP, Hi-C, etc.) by using graph attention networks (GAT) [134]. Compared with linear CNN models (such as Epi-CNN, Seq-CNN, etc.), the GraphReg model requires less sample size and has higher accuracy in prediction [134]. Graph-based methods (GC-MERGE and GraphReg) have advantages in handling complex relationships, robustness and data utilization compared with traditional machine learning methods. Compared with linear CNN models, graph-based methods have the advantages of strong interpretability, fast calculation efficiency and high accuracy in predicting long-distance enhancer-target genes [133, 134]. With more and more researchers focusing on graph theory and deep learning techniques in bioinformatics, graph-based methods will provide more powerful tools for analyzing enhancer-target gene networks [136].

In addition to CNN, architectures based on deep neural networks (DNN) are used to learn enhancer features as well. For example, EP-DNN uses p300 binding sites as markers for enhancers, and TSS and random non-DHS sites as markers for non-enhancers to perform training. The prediction accuracy of EP-DNN is 91.6%, exceeding the accuracy of DEEP-ENCODE (85.3%) and RFECS (85.5%) [129]. ES-ARCNN is a computational model for predicting the enhancers strength. To train ES-ARCNN, researchers applied two data augmentation tricks (i.e. reverse complement and shift) to improve the model's predictive performance [137]. Enformer, as a developed enhancer prediction model based on the transformer, can integrate information of remote interactions in the genome (up to 100 kb away) [138]. By calculating the contribution scores of gene input gradients and attention weights, Enformer can identify the enhancer sequences that are most predictive of specific gene expression [138]. Although deep learning has outperformed many traditional computer methods in enhancer prediction applications, the problems of overparameterization and limited model performance still exist, and its interpretability lags behind traditional statistical methods. Continuous development of new deep learning methods is expected to achieve elegant applications in enhancer research.

CHALLENGES IN MULTI-OMICS APPROACHES

In recent years, there has been an increasing number of studies on enhancers by multi-omics approaches. However, there are still challenges in the application of multi-omics approaches to

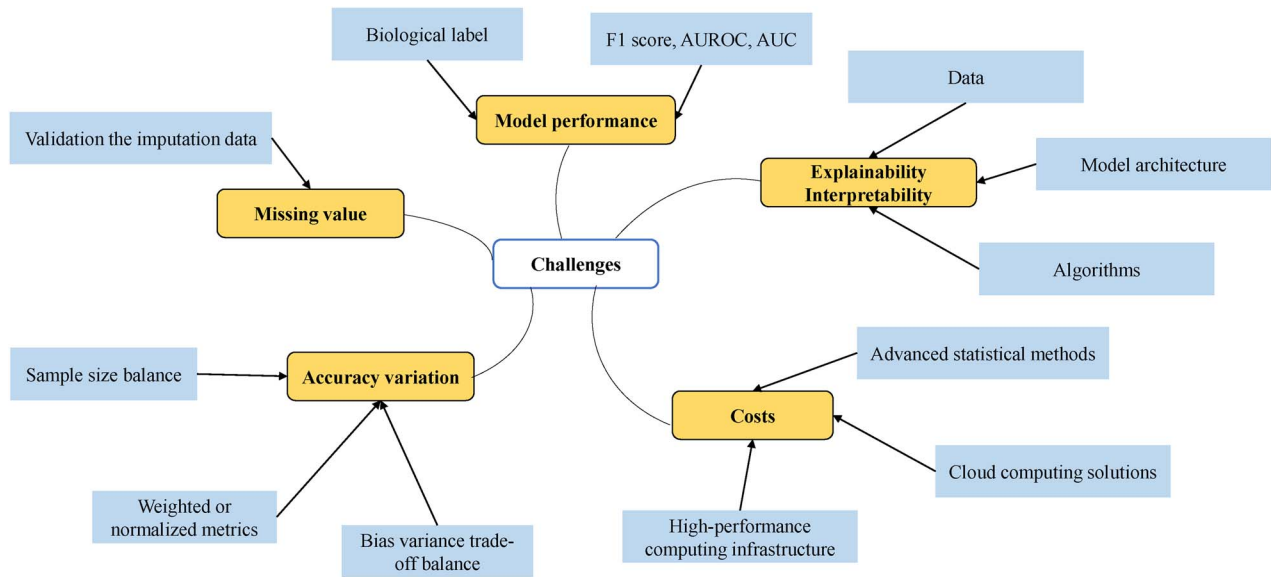


Figure 5. Challenges and resolution strategies for multi-omics integration methods in enhancer research.

study enhancers, either due to a lack of sufficient attention or limited solutions. We have summarized the five major challenges in enhancer research, and provided some possible methods to overcome these challenges (Figure 5).

The accuracy variation between different omics data

Multi-omics data from different sources are often heterogeneous, with divergence in signal-to-noise ratios and significant differences in accuracy [139]. For instance, genome sequencing has a higher coverage than RNA-seq; transcriptomics and ChIP-seq use different quantification methods (the former uses RPM or count values, while the latter quantifies based on peak areas), resulting in different data ranges and distributions [140]. Currently, increasing the number of samples and improving experimental design can improve the statistical power of different omics analyses. However, according to MultiPower software, in the estimation of sample size required to achieve specific statistical power in different omics, DNA-seq and ChIP-seq require close to more than double the sample size of RNA-seq samples to achieve the same statistical effect [141]. Therefore, it is inefficient and expensive to improve accuracy only by increasing sequencing samples. Instead, one can consider balancing sample sizes through undersampling [141]. In addition, it can also evaluate the performance of machine learning by using standardized metrics to choose the optimal sample size, or by adopting techniques (such as regularization, bagging, cross-validation) to balance bias-variance trade-offs [142, 143].

Missing value imputation in multi-omics data

Data may be missing due to experimental random errors or inherent technical defects (e.g. low coverage in repetitive regions) during sequencing [144]. Consequently, some unmatched data have to be excluded during data integration, limiting the power for detection in the genome. Surprisingly, the problem of processing missing values is often treated as a data preprocessing step, and some scientists do not believe that it will have any impact on the outcomes of subsequent statistical analyses. Instead, the distribution characteristics of the multi-omics data should be reassessed in the analysis process, and sensitivity analysis should

be performed to assess the impact of missing value inputs on the downstream analysis [29]. Imputation methods have the potential to correct missing values by leveraging the correlations within omics data and utilizing partially measured data from other omics datasets to impute missing values. MOFA analyzes the latent space across omics types to impute missing samples, and MultiBaC creates a multivariate predictive model of the incomplete omics types as a function of a shared omics modality [145, 146]. However, these two methods can create data structures that violate the assumption of independence and subsequently lead to unreliable analysis [29, 147, 148]. Therefore, the missing values across different data resources affirm the reliability and applicability of multi-omics analysis, and a better solution is urgently needed. Liew et al. compared 19 different missing value completion algorithms and found that the choice of algorithm should be assessed from an application-driven viewpoint, and validation of the imputation data is an important step in evaluating the performance of any input algorithm [149]. There is no one optimal imputation algorithm for all type of data, so it is necessary to choose an appropriate imputation algorithm according to the characteristics of the data [149].

Evaluation of model performance

Currently, the computational models for integrating multi-omics data in biology possess various characteristics (e.g. accuracy, speed, complexity and computational cost), and it is crucial to select the most suitable algorithm for multi-omics analysis [130]. Some supervised approaches are subjected to overfitting from inappropriate cross-validation policies, while certain approaches are limited by training label uncertainties [150]. In supervised learning, an incorrect label definition can lead to inaccurate prediction results. Consequently, it is essential to enlist biological expertise to properly define the labels [107]. On the other hand, models, such as the ABC model and eNet model, predict the functions of thousands of enhancers. However, most of these enhancers have not been experimentally validated, making it difficult to determine the accuracy of model prediction. Performance metrics used commonly for this purpose include the F1 score (Harmonic Mean of Precision and Recall), the area under the receiver operating characteristic curve and the area under the

precision recall curve. But due to the diversity of the principles and standard definitions of prediction, it is difficult to systematically evaluate the performance of all available computational methods [109].

Interpretability of multi-omics approaches

Interpretability is about the extent to which a cause and effect can be observed within a system [151]. Factors affect the interpretability of a model, including data, model architecture and algorithms [152, 153]. To improve the interpretability of a model integrating the multi-omics data, several approaches have been developed from different perspectives of the affecting factors mentioned above. (i) Human-labeled data can improve the interpretability of a model. For example, GenNet improves the interpretability of genotype data by constructing explainable neural networks that use prior biological knowledge to label the data [154]. (ii) Simplifying the network architecture can increase the interpretation [153]. For example, ExplainNN uses a large series of simple neural networks, each of which learns different TF binding profiles. As a result, it becomes easier to understand the prediction results of each TF, thereby improving the overall interpretability [153]. The HEAP model uses the weights of the first convolutional layer to capture important enhancer features to build a deep network model [152]. Adopting explainable artificial intelligence (XAI) architecture is another approach [155]. With this kind of architecture, researchers can have a clearer understanding of the degree of causal relationship between input signals and output results. Therefore, some teams have utilized XAI to identify enhancers based on the epigenetic feature signals of different histone groups, and discovered the connection between the enrichment of different histone modifications and the activity of enhancers [155]. (iii) From the algorithmic perspective, using interpretable algorithms (such as clustering, SHAP, etc.) can improve the interpretability of a model [152, 153]. In summary, although there are many methods available currently to improve the interpretability of models, overly pursuing interpretability may lead to a decline in model performance [156]. The existing interpretability methods often target specific model architectures or data types only [29]. Therefore, new strategies need to be continuously developed to improve the explanation ability of models.

Computational and storage costs

Multi-omics analysis incurs costs for computation and data storage [157]. Most integrated algorithms require high computational power and considerable storage capacity to store logs, results and analyses [140]. How to store multi-omics datasets to facilitate the reuse of existing research datasets is another challenge. High-performance computing infrastructure, cloud computing solutions and advanced statistical methods are all effective ways to reduce computing and storage costs. The general principle is FAIR, which stands for findable, accessible, interoperable and reusable [158]. However, many multi-omics data storage platforms (such as Figshare, Zenodo or Lifebit) do not support data retrieval and query [29]. Numerous computing models have been deployed on specialized graphics processing units and cloud computing platforms over the past few years (such as Microsoft Azure [159]), which is one of the ways to address the issues mentioned above [160, 161].

The five yellow rectangles in the figure represent the five major challenges encountered in enhancer research, while the blue squares represent different resolution strategies to address these challenges.

CONCLUSION

Enhancers are crucial regulatory elements in gene transcription, and the application of omics techniques speeds up the elucidation of the role and mechanisms of enhancers in gene regulation. In this review, we first summarized the current issues encountered in enhancer research. Next, we discussed the application and limitations of four types of omics technologies (genomics, epigenomics, transcriptomics and CRISPR gene editing). With the increasing availability of larger, high-quality datasets paralleled by the development of new omics technologies, the demand for ideas and methods for multi-omics analysis will continue to grow. Using machine learning to integrate and analyze high-dimensional and multi-omics data can effectively improve the accuracy of the enhancer prediction model. Furthermore, novel algorithms can be utilized to extract new information from existing data. However, the application of omics technology in the field of enhancer research is still challenging. Despite the widespread heterogeneity and divergent quality of multi-omics data, both data quality and quantity are being improved with the increasing application of sequencing techniques. In addition, statistical methods are continuously evolving to address the current challenges. In summary, we believe that with the development of omics technologies and statistics, multi-omics techniques will have greater value in enhancer research.

Key Points

- At present, there are four basic problems in enhancer research: identification, activity, structure and eRNA.
- Genomics, epigenomics, transcriptomics and CRISPR-gene editing technology have been widely used in enhancer research.
- Multi-omics integration methods in enhancer research are divided into traditional machine learning and deep learning methods.

ACKNOWLEDGEMENTS

We deeply thank Yong Zhang and Weiwei Zhai for their comments and helpful suggestions during the manuscript preparation.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (32270610, 31801094 and 82072499 to C.L.) and the Fundamental Research Funds for the Central Universities (YWF-21-BJ-J-T105 to C.L.).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

REFERENCES

1. Ohler U, Wassarman DA. Promoting developmental transcription. *Development* 2010;**137**:15–26.

2. Peng Y, Zhang Y. Enhancer and super-enhancer: positive regulators in gene transcription. *Animal Model Exp Med* 2018;**1**: 169–79.
3. Thomas HF, Buecker C. What is an enhancer? *Bioessays* 2023;**45**:e2300044.
4. Berman BP, Pfeiffer BD, Lavery TR, et al. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol* 2004;**5**:R61.
5. Ye R, Cao C, Xue Y. Enhancer RNA: biogenesis, function, and regulation. *Essays Biochem* 2020;**64**:883–94.
6. Agrawal P, Rao S. Super-enhancers and CTCF in early embryonic cell fate decisions. *Front Cell Dev Biol* 2021;**9**:653669.
7. Corradin O, Scacheri PC. Enhancer variants: evaluating functions in common disease. *Genome Med* 2014;**6**:85.
8. Liu F, Li H, Ren C, et al. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci Rep* 2016;**6**:28517.
9. Yang B, Liu F, Ren C, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* 2017;**33**:1930–6.
10. Bosse Y, Amos CI. A decade of GWAS results in lung cancer. *Cancer Epidemiol Biomarkers Prev* 2018;**27**:363–79.
11. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* 2019;**51**:592–9.
12. Cano-Gamez E, Trynka G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front Genet* 2020;**11**:424.
13. Dozmorov MG, Tyc KM, Sheffield NC, et al. Chromatin conformation capture (Hi-C) sequencing of patient-derived xenografts: analysis guidelines. *Gigascience* 2021;**10**:10.
14. Nakato R, Sakata T. Methods for ChIP-seq analysis: a practical workflow and advanced applications. *Methods* 2021;**187**: 44–53.
15. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010;**2010**(2):pdb prot5384.
16. Ocampo J, Cui F, Zhurkin VB, Clark DJ. The proto-chromatosome: a fundamental subunit of chromatin? *Nucleus* 2016;**7**:382–7.
17. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. *Trends Genet* 2018;**34**: 666–81.
18. Orgaz JL, Crosas-Molist E, Sadok A, et al. Myosin II reactivation and cytoskeletal remodeling as a hallmark and a vulnerability in melanoma therapy resistance. *Cancer Cell* 2020;**37**: 85–103.e9.
19. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;**17**:13.
20. Neumayr C, Pagani M, Stark A, et al. STARR-seq and UMI-STARR-seq: assessing enhancer activities for genome-wide-, high-, and low-complexity candidate libraries. *Curr Protoc Mol Biol* 2019;**128**:e105.
21. Tian W, Huang X, Ouyang X. Genome-wide prediction of activating regulatory elements in rice by combining STARR-seq with FACS. *Plant Biotechnol J* 2022;**20**:2284–97.
22. Lu T, Yang B, Wang R, et al. Xenotransplantation: current status in preclinical research. *Front Immunol* 2019;**10**:3060.
23. Subramanian I, Verma S, Kumar S, et al. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:1177932219899051.
24. Klefogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform* 2016;**17**:967–79.
25. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017;**2017**:bax028.
26. Tsai A, Alves MR, Crocker J. Multi-enhancer transcriptional hubs confer phenotypic robustness. *Elife* 2019;**8**:e45325.
27. Ribeiro DM, Rubinacci S, Ramisch A, et al. The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat Commun* 2021;**12**:4842.
28. Wörheide MA, Krumsiek J, Kastenmüller G, et al. Multi-omics integration in biomedical research – a metabolomics-centric review. *Anal Chim Acta* 2021;**1141**:144–62.
29. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat Comput Sci* 2021;**1**:395–402.
30. Investigators GPP, Smedley D, Smith KR, et al. 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. *N Engl J Med* 2021;**385**:1868–80.
31. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci* 2018;**109**: 513–22.
32. Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;**20**: 467–84.
33. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 2008;**24**:408–15.
34. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell* 2013;**155**:934–47.
35. Chignon A, Mathieu S, Rufiange A, et al. Enhancer promoter interactome and Mendelian randomization identify network of druggable vascular genes in coronary artery disease. *Hum Genomics* 2022;**16**:8.
36. Chen H, Li C, Peng X, et al. A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* 2018;**173**:386–399.e312.
37. Mohanta TK, Mishra AK, Al-Harrasi A. The 3D genome: from structure to function. *Int J Mol Sci* 2021;**22**:11585.
38. Lafontaine DL, Yang L, Dekker J, et al. Hi-C 3.0: improved protocol for genome-wide chromosome conformation capture. *Curr Protoc* 2021;**1**:e198.
39. Vardaxis I, Drablos F, Rye MB, et al. MACPET: model-based analysis for ChIA-PET. *Biostatistics* 2020;**21**:625–39.
40. Mumbach MR, Rubin AJ, Flynn RA, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* 2016;**13**:919–22.
41. Rosen JD, Yang Y, Abnoui A, et al. HPRP: quantifying reproducibility in HiChIP and PLAC-Seq datasets. *Curr Issues Mol Biol* 2021;**43**:1156–70.
42. Zhang Y, An L, Xu J, et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun* 2018;**9**:750.
43. Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80.
44. Wang KC, Chang HY. Epigenomics: technologies and applications. *Circ Res* 2018;**122**:1191–9.
45. Wilson PC, Ledru N, Humphreys BD. Epigenomics and the kidney. *Curr Opin Nephrol Hypertens* 2020;**29**:280–5.
46. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 2019;**20**:207–20.

47. Song L, Zhang Z, Grasfeder LL, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* 2011;**21**:1757–67.
48. Buenrostro JD, Wu B, Chang HY, et al. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;**109**:21.29.1–29.
49. Chen A, Chen D, Chen Y. Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals. *Gene* 2018;**667**:83–94.
50. Liu Y, Fu L, Kaufmann K, et al. A practical guide for DNase-seq data analysis: from data management to common applications. *Brief Bioinform* 2019;**20**:1865–77.
51. Kong S, Lu Y, Tan S, et al. Nucleosome-omics: a perspective on the epigenetic code and 3D genome landscape. *Genes (Basel)* 2022;**13**:1114.
52. Chereji RV, Bryson TD, Henikoff S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol* 2019;**20**:198.
53. Seuter S, Neme A, Carlberg C. Monitoring genome-wide chromatin accessibility by formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq). *Epigenetics Methods* 2020;**353**–69.
54. Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;**10**:1213–8.
55. Jia G, Preussner J, Chen X, et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat Commun* 2018;**9**:4877.
56. Ji Z, Zhou W, Hou W, et al. Single-cell ATAC-seq signal extraction and enhancement with SCATE. *Genome Biol* 2020;**21**:161.
57. Chen H, Liang H. A high-resolution map of human enhancer RNA loci characterizes super-enhancer activities in cancer. *Cancer Cell* 2020;**38**:701–715.e705.
58. Hong D, Lin H, Liu L, et al. Complexity of enhancer networks predicts cell identity and disease genes revealed by single-cell multi-omics analysis. *Brief Bioinform* 2023;**24**(1):bbac508.
59. Clermont P-L, Parolia A, Liu HH, et al. DNA methylation at enhancer regions: novel avenues for epigenetic biomarker development. *IMR Press*. 2016;**21**(2):430–46.
60. Li J, Li Y, Li W, et al. Guide positioning sequencing identifies aberrant DNA methylation patterns that alter cell identity and tumor-immune surveillance networks. *Genome Res* 2019;**29**:270–80.
61. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**:669–80.
62. Nakao K, Miyaaki H, Ichikawa T. Antitumor function of microRNA-122 against hepatocellular carcinoma. *J Gastroenterol* 2014;**49**:589–93.
63. Cejas P, Li L, O'Neill NK, et al. Chromatin immunoprecipitation from fixed clinical tissues reveals tumor-specific enhancer profiles. *Nat Med* 2016;**22**:685–91.
64. Font-Tello A, Kesten N, Xie Y, et al. FiTAC-seq: fixed-tissue ChIP-seq for H3K27ac profiling and super-enhancer analysis of FFPE tissues. *Nat Protoc* 2020;**15**:2503–18.
65. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet* 2011;**52**:413–35.
66. Mundade R, Ozer HG, Wei H, et al. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle* 2014;**13**:2847–52.
67. Kaya-Okur HS, Wu SJ, Codomo CA, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* 2019;**10**:1930.
68. Kaya-Okur HS, Janssens DH, Henikoff JG, et al. Efficient low-cost chromatin profiling with CUT&Tag. *Nat Protoc* 2020;**15**:3264–83.
69. Li QL, Lin X, Yu YL, et al. Genome-wide profiling in colorectal cancer identifies PHF19 and TBC1D16 as oncogenic super enhancers. *Nat Commun* 2021;**12**:6407.
70. Cheung K, Barter MJ, Falk J, et al. Histone ChIP-Seq identifies differential enhancer usage during chondrogenesis as critical for defining cell-type specificity. *FASEB J* 2020;**34**:5317–31.
71. Zuo S, Yi Y, Wang C, et al. Extrachromosomal circular DNA (eccDNA): from chaos to function. *Front Cell Dev Biol* 2021;**9**:792555.
72. Zhu Y, Gujar AD, Wong CH, et al. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* 2021;**39**:694–707.e697.
73. Møller HD. Circle-Seq: isolation and sequencing of chromosome-derived circular DNA elements in cells. *Methods Mol Biol* 2020;**2119**:165–81.
74. Zhou B, Li X, Luo D, et al. GRID-seq for comprehensive analysis of global RNA-chromatin interactions. *Nat Protoc* 2019;**14**:2036–68.
75. Li X, Zhou B, Chen L, et al. GRID-seq reveals the global RNA-chromatin interactome. *Nat Biotechnol* 2017;**35**:940–50.
76. Li J, Xiang Y, Zhang L, et al. Enhancer-promoter interaction maps provide insights into skeletal muscle-related traits in pig genome. *BMC Biol* 2022;**20**:136.
77. Cai Z, Cao C, Ji L, et al. RIC-seq for global in situ profiling of RNA-RNA spatial interactions. *Nature* 2020;**582**:432–7.
78. Kim TK, Shiekhata R. Architectural and functional commonalities between enhancers and promoters. *Cell* 2015;**162**:948–59.
79. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
80. Manton KJ, Kream RM, Kuzelova H, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med Sci Monit Basic Res* 2014;**20**:138–42.
81. Jovic D, Liang X, Zeng H, et al. Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med* 2022;**12**:e694.
82. Saliba AE, Westermann AJ, Gorski SA, et al. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 2014;**42**:8845–60.
83. Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods* 2022;**19**:534–46.
84. Muerdter F, Boryn LM, Arnold CD. STARR-seq - principles and applications. *Genomics* 2015;**106**:145–50.
85. Goldstein I, Hager GL. Dynamic enhancer function in the chromatin context. *Wiley Interdiscip Rev Syst Biol Med* 2018;**10**(1):10.1002.
86. Andersson R, Refsing Andersen P, Valen E, et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun* 2014;**5**:5336.
87. Lee JH, Xiong F, Li W. Enhancer RNAs in cancer: regulation, mechanisms and therapeutic potential. *RNA Biol* 2020;**17**:1550–9.
88. Hah N, Kraus WL. Hormone-regulated transcriptomes: lessons learned from estrogen signaling pathways in breast cancer cells. *Mol Cell Endocrinol* 2014;**382**:652–64.

89. Murakawa Y, Yoshihara M, Kawaji H, et al. Enhanced identification of transcriptional enhancers provides mechanistic insights into diseases. *Trends Genet* 2016;**32**:76–88.
90. Consortium F, the RP, CLST, et al. A promoter-level mammalian expression atlas. *Nature* 2014;**507**:462–70.
91. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**:455–61.
92. Fulco CP, Nasser J, Jones TR, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* 2019;**51**:1664–9.
93. Dixit A, Parnas O, Li B, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 2016;**167**:1853–1866.e1817.
94. Visel A, Minovitsky S, Dubchak I, et al. VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res* 2007;**35**:D88–92.
95. Zhang G, Shi J, Zhu S, et al. DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Res* 2017;**46**:D78–84.
96. Bai X, Shi S, Ai B, et al. ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res* 2019;**48**:D51–7.
97. Kumar R, Lathwal A, Kumar V, et al. CancerENd: a database of cancer associated enhancers. *Genomics* 2020;**112**:3696–702.
98. Luo Z-H, Shi M-W, Zhang Y, et al. CenhANCER: a comprehensive cancer enhancer database for primary tissues and cell lines. *Database* 2023;**2023**:baad022.
99. Wang J, Dai X, Berry LD, et al. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res* 2019;**47**:D106–12.
100. Cai Z, Cui Y, Tan Z, et al. RAEdB: a database of enhancers identified by high-throughput reporter assays. *Database (Oxford)* 2019;**2019**:bay140.
101. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* 2020;**48**:D58–64.
102. Tang L, Hill MC, Wang J, et al. Predicting unrecognized enhancer-mediated genome topology by an ensemble machine learning model. *Genome Res* 2020;**30**:1835–45.
103. Cai Z, Poulos RC, Liu J, et al. Machine learning for multi-omics data integration in cancer. *iScience* 2022;**25**:103798.
104. Chen Z, Zhang J, Liu J, et al. DECODE: a deep-learning framework for condensing enhancers and refining boundaries with large-scale functional assays. *Bioinformatics* 2021;**37**:i280–8.
105. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2018;**19**:1236–46.
106. Ahmad F, Mahmood A, Muhmood T. Machine learning-integrated omics for the risk and safety assessment of nanomaterials. *Biomater Sci* 2021;**9**:1598–608.
107. Correa-Aguila R, Alonso-Pupo N, Hernández-Rodríguez EW. Multi-omics data integration approaches for precision oncology. *Mol Omics* 2022;**18**:469–79.
108. Xu H, Zhang S, Yi X, et al. Exploring 3D chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer-promoter interaction. *Comput Struct Biotechnol J* 2020;**18**:558–70.
109. Tao H, Li H, Xu K, et al. Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. *Brief Bioinform* 2021;**22**(5):bbaa405.
110. Popay TM, Dixon JR. Coming full circle: on the origin and evolution of the looping model for enhancer-promoter communication. *J Biol Chem* 2022;**298**:102117.
111. Malin J, Aniba MR, Hannenhalli S. Enhancer networks revealed by correlated DNase hypersensitivity states of enhancers. *Nucleic Acids Res* 2013;**41**:6828–38.
112. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**:488–96.
113. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;**473**:43–9.
114. Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science* 2018;**362**:eaav1898.
115. Yao L, Shen H, Laird PW, et al. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol* 2015;**16**:105.
116. O'Connor T, Bodén M, Bailey TL. CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. *Nucleic Acids Res* 2017;**45**:e19.
117. Huska MR, Ramisch A, Vingron M, et al. Predicting enhancers using a small subset of high confidence examples and co-training. *PeerJ Preprints* 2016;**4**:e2407v1.
118. Greene CS, Tan J, Ung M, et al. Big data bioinformatics. *J Cell Physiol* 2014;**229**:1896–900.
119. Chen X, Zhou J, Zhang R, et al. Tissue-specific enhancer functional networks for associating distal regulatory regions to disease. *Cell Systems* 2021;**12**:353–362.e356.
120. Hait TA, Amar D, Shamir R, et al. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. *Genome Biol* 2018;**19**:56.
121. Cao Q, Anyansi C, Hu X, et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat Genet* 2017;**49**:1428–36.
122. Hafez D, Karabacak A, Krueger S, et al. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome Biol* 2017;**18**:199.
123. Mehdi TF, Singh G, Mitchell JA, et al. Variational infinite heterogeneous mixture model for semi-supervised clustering of heart enhancers. *Bioinformatics* 2019;**35**:3232–9.
124. He B, Chen C, Teng L, et al. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* 2014;**111**:E2191–9.
125. Zhao C, Li X, Hu H. PETModule: a motif module based approach for enhancer target gene prediction. *Sci Rep* 2016;**6**:30043.
126. Roy S, Siahpirani AF, Chasman D, et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res* 2015;**43**:8694–712.
127. Gao T, Qian J. EAGLE: an algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. *PLoS Comput Biol* 2019;**15**:e1007436.
128. Talukder A, Saadat S, Li X, et al. EPIP: a novel approach for condition-specific enhancer-promoter interaction prediction. *Bioinformatics* 2019;**35**:3877–83.
129. Kim SG, Harwani M, Grama A, et al. EP-DNN: a deep neural network-based global enhancer prediction algorithm. *Sci Rep* 2016;**6**:38433.
130. Hong Z, Zeng X, Wei L, et al. Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 2020;**36**:1037–43.

131. Ji Y, Zhou Z, Liu H, et al. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112–20.
132. Yang R, Wu F, Zhang C, et al. iEnhancer-GAN: a deep learning framework in combination with word embedding and sequence generative adversarial net to identify enhancers and their strength. *Int J Mol Sci* 2021;**22**(7):3589.
133. Bigness J, Loinaz X, Patel S, et al. Integrating long-range regulatory interactions to predict gene expression using graph convolutional networks. *J Comput Biol* 2022;**29**:409–24.
134. Zhao M, Ma L, Jia X, et al. GraphReg: dynamical point cloud registration with geometry-aware graph signal processing. *IEEE Trans Image Process* 2022;**31**:7449–64.
135. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* 2019;**47**:e60.
136. Xiao S, Lin H, Wang C, et al. Graph neural networks with multiple prior knowledge for multi-omics data analysis. *IEEE J Biomed Health Inform* 2023;**27**:4591–600.
137. Zhang T-H, Flores M, Huang Y. ES-ARCNN: predicting enhancer strength by using data augmentation and residual convolutional neural network. *Anal Biochem* 2021;**618**:114120.
138. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1196–203.
139. Bersanelli M, Mosca E, Remondini D, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17**(Suppl 2):15.
140. Reel PS, Reel S, Pearson E, et al. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv* 2021;**49**:107739.
141. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, et al. Harmonization of quality metrics and power calculation in multi-omic studies. *Nat Commun* 2020;**11**:3092.
142. Jeni LA, Cohn JF, Torre FDL. Facing imbalanced data-recommendations for the use of performance metrics. In: 2013 *Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, p. 245–51.
143. Siebert U, Rochau U, Claxton K. When is enough evidence enough? - Using systematic decision analysis and value-of-information analysis to determine the need for further evidence. *Z Evid Fortbild Qual Gesundheitswes* 2013;**107**:575–84.
144. Chen L, Liu P, Evans TC, Jr, et al. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 2017;**355**:752–6.
145. Argelaguet R, Velten B, Arnol D, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;**14**:e8124.
146. Ugidos M, Tarazona S, Prats-Montalbán JM, et al. MultiBaC: a strategy to remove batch effects between different omic data types. *Stat Methods Med Res* 2020;**29**:2851–64.
147. Voillet V, Besse P, Liaubet L, et al. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics* 2016;**17**:402.
148. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data* 2019;**6**:251.
149. Liew AW-C, Law N-F, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform* 2010;**12**:498–513.
150. McCabe SD, Lin DY, Love MI. Consistency and overfitting of multi-omics methods on experimental data. *Brief Bioinform* 2020;**21**:1277–84.
151. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 2018;**16**:31–57.
152. Liu Y, Wang Z, Yuan H, et al. HEAP: a task adaptive-based explainable deep learning framework for enhancer activity prediction. *Brief Bioinform* 2023;**24**(5):bbad286.
153. Smith GD, Ching WH, Cornejo-Páramo P, et al. Decoding enhancer complexity with machine learning and high-throughput discovery. *Genome Biol* 2023;**24**:116.
154. van Hilten A, Kushner SA, Kayser M, et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* 2021;**4**:1094.
155. Wolfe JC, Mikheeva LA, Hagrass H, et al. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in *Drosophila*. *Genome Biol* 2021;**22**:308.
156. McDermid JA, Jia Y, Porter Z, et al. Artificial intelligence explainability: the technical and ethical dimensions. *Philos Trans A Math Phys Eng Sci* 2021;**379**:20200363.
157. Herrmann M, Probst P, Hornung R, et al. Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief Bioinform* 2021;**22**:bbaa167.
158. Caspi R, Billington R, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 2018;**46**:D633–9.
159. Copeland M, Soh J, Puca A, et al. Microsoft Azure and cloud computing. In: Copeland M, Soh J, Puca A et al. (eds). *Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud*. Berkeley, CA: Apress, 2015, 3–26.
160. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;**61**:85–117.
161. Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. *Commun ACM* 2010;**53**:50–8.