*D40–D44* *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 20 August 2019*
*doi: 10.1093/nar/gkz715*

# CFEA: a cell-free epigenome atlas in human diseases

**Fulong Yu[1,†], Kai Li[2,†], Shuangquan Li[3,†], Jiaqi Liu[4], Yan Zhang[1], Meng Zhou[1], Hengqiang Zhao[5], Hongyan Chen[6], Nan Wu[5,7], Zhihua Liu[6,*] and Jianzhong Su[1,2,*]**

[1]School of Biomedical Engineering, School of Ophthalmology and Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou 325011, Zhejiang, China, [2]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325011, Zhejiang, China, [3]The First Affiliated Hospital of Wenzhou Medical University, Wenzhou Medical University, Wenzhou 325011, Zhejiang, China, [4]Department of Breast Surgical Oncology, National Cancer Center /National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China, [5]Department of Orthopedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100730, China, [6]State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China and [7]Beijing Key Laboratory for Genetic Research of Skeletal Deformity, Beijing 100730, China

Received June 27, 2019; Revised August 02, 2019; Editorial Decision August 02, 2019; Accepted August 14, 2019

## ABSTRACT

**Epigenetic alterations, including 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC) and nucleosome positioning (NP), in cell-free DNA (cfDNA) have been widely observed in human diseases, and many available cfDNA-based epigenome-wide profiles exhibit high sensitivity and specificity in disease detection and classification. However, due to the lack of efficient collection, standardized quality control, and analysis procedures, efficiently integrating and reusing these data remain considerable challenges. Here, we introduce CFEA (http://www.bio-data.cn/CFEA), a cell-free epigenome database dedicated to three types of widely adopted epigenetic modifications (5mC, 5hmC and NP) involved in 27 human diseases. We developed bioinformatic pipelines for quality control and standard data processing and an easy-to-use web interface to facilitate the query, visualization and download of these cell-free epigenome data. We also manually curated related biological and clinical information for each profile, allowing users to better browse and compare cfDNA epigenomes at a specific stage (such as early- or metastasis-stage) of cancer development. CFEA provides a comprehensive and timely resource to the scientific community and supports the development of liquid biopsy-based biomarkers for various human diseases.**

## INTRODUCTION

Circulating cell-free DNA (cfDNA) is defined as extracellular nucleic acid fragments that are released into the bloodstream by cell necrosis and apoptosis (1). Recent advances in cfDNA-based liquid biopsy demonstrate that the abundant genetic and epigenetic information carried in cfDNA as non-invasive molecular signatures can revolutionize the traditional screening and treatment of various human disorders (2). Epigenetics provides an important molecular link between genetic programming and environmental signals, and such changes in the cfDNA and the somatic genomic DNA (gDNA) from the tumor tissue of origin are highly consistent in many disease models (3). Based on tissue specificity and stability, epigenetic alterations have already been incorporated as valuable candidate biomarkers of human diseases (4,5).

In recent years, three types of epigenetic variations, namely, DNA methylation (5-methylcytosine, 5mC), hydroxymethylation (5-hydroxymethylcytosine, 5hmC) and nucleosome positioning (NP), in cfDNA have been successfully detected in clinical samples by a range of genome-wide approaches, demonstrating high clinical potential in disease diagnosis, prognosis and/or treatment response (6). For example, 5mC biomarkers based on plasma cfDNA are sensitive for early tumour detection in efforts for cancer interception (7). In addition to predicting cancer types, cfDNA 5hmC signatures are used to track tumour stages in human cancers (8). Genome-wide NP maps of cfDNA are also employed to infer pathological states of multiple disease types (9).

*To whom correspondence should be addressed. Tel: +86 577 8806 8270; Fax: +86 577 8806 8270; Email: sujz@wibe.ac.cn
Correspondence may also be addressed to Zhihua Liu. Email: liuzh@cicams.ac.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

At present, a large number of cfDNA-based epigenetic profiles are available for facilitating the epigenome-wide discovery of non-invasive biomarkers for different types of neoplastic, chronic inflammatory and autoimmune diseases (10), and cross-data set analysis of these epigenetic modifications for specific/multiple diseases is very useful for discovering reliable non-invasive biomarkers. However, different DNA-based epigenomic modifications have their own features and are measured by different experimental technologies. Indeed, even a single modification such as 5mC in cfDNA may be measured using five genome-wide technologies of whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), methylated DNA immunoprecipitation sequencing (MeDIP-Seq), methyl-CpG binding domain-based capture and sequencing (Methyl-Cap) and Infinium Human Methylation 450K BeadChip (450K array) (11). Moreover, various laboratories have used different bioinformatics tools and parameters to perform data analysis, further complicating the situation. Therefore, it is difficult for experimental biologists and clinicians to use these data to perform integration and comparison analyses. To obtain more reliable biomarkers, it is essential to evaluate and analyse differences in disease-specific epigenetic changes between cfDNA and corresponding gDNA data (12), yet there are no known resources for incorporating epigenome data for both cfDNA and gDNA. In addition, biological and clinical information has not been systematically or non-intuitively annotated further hindered data mining and interpretation in many disease datasets.

To address these challenges, we developed the cell-free epigenome atlas (CFEA), a comprehensive public database providing cfDNA-based epigenome profiles (5mC, 5hmC and NP) for 27 human diseases. All data in CFEA have been uniformly processed by our reproducible bioinformatics pipelines, integrating gDNA epigenome data as controls, with manually curated biological and clinical information for each sample. The user-friendly web interface provided by CFEA can facilitate easy querying, visualization and comparison of the collected data by the general scientific community.

## DATA COLLECTION AND DATABASE CONTENT

As shown in Figure 1, the raw datasets for cell-free epigenomes from public data repositories were collected, including NCBI Gene Expression Omnibus (GEO) (13), Sequence Read Archive (SRA) (14), Genome Sequence Archive (GSA) (15) and European Nucleotide Archive (ENA) (16). Consistent analysis of cfDNA and gDNA from solid tissue is important for determining valid epigenomic biomarkers for a tissue of origin and disease type. To facilitate such comparison, we also gathered tissue epigenome profiles of gDNA for patients with malignant tumours from GEO and the Cancer Genome Atlas (TCGA) (17). All CFEA data were uniformly quality controlled and processed from raw sequencing data by streamlined pipelines. In addition, we carefully reviewed the original publications in PubMed and further manually curated matched experimental and clinical information for each CFEA sample. For each sample catalogued by CFEA, we collected the metadata of 10 items including disease, gender, age, molecular level (5mC, 5hmC, NP), detection method, cancer subtype, cancer stage (benign, early, late, metastasis), American Joint Commission on Cancer (AJCC) stage of cancer, treatment before epigenomic profiling (yes, no, unknown), PubMed ID of original literature. The missing values were marked as NA in our database. These data may be useful for characterizing cfDNA-based epigenomic alterations in human diseases.

At present, CFEA contains a total of 1645 samples (1520 cfDNA-based epigenomes and 125 gDNA-based epigenomes) for three types of epigenetic information, 5mC, 5hmC and NP, in 27 types of human diseases (Supplementary Table S1). A majority (~85%) of the data available in CFEA were generated using next-generation sequencing technology. All of the reported cfDNA epigenome profiles are from blood plasma or serum. Except for several chronic inflammatory and autoimmune diseases, we found that a majority of the cfDNA-based epigenomes were closely associated with diverse malignant tumours, and nearly 40% of them were annotated with information regarding clinical subtype or cancer development.

## STANDARDIZED DATA PROCESSING

We developed the standard processing pipelines to process these raw data based on open-source and widely used bioinformatics tools (Figure 2). All the collected data (>90 billion raw sequencing reads) were reproducibly processed from raw sequencing data using our pipelines. For sequencing-based data, we applied a set of quality control (QC) metrics to evaluate experimental quality before read alignment. It is necessary to include QC steps in our pipelines because >10% of the samples would fail to pass our QC evaluation. We provided information on high-throughput sequencing samples with low sequencing quality across different disease types and technologies as Supplementary Table S2.

The bioinformatics pipeline of CFEA mainly includes several sections. First, public sequencing data were downloaded in sra or fastq format. Second, we converted sra format files to fastq format using the SRA toolkit version 2.9.2 (14). Third, we performed quality evaluation with fastqc software version 0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and any sequence in the fastq files that did not meet the criteria, such as adaptors and overrepresented sequences, were trimmed using trim_galore version 0.5.0 with default parameters (http://www.bioinformatics.babraham.ac.uk/projects/trim-galore/). Subsequently, we used different software programs based on experimental types and protocols. Data obtained using methods based on bisulfite conversion, including WGBS, RRBS, and MCTA, were processed as follows. Trimmed WGBS and RRBS reads were aligned to the reference genome using bismark version 0.18.2 with default parameters (18). Bismark_methylation_extractor program in the bismark toolset was used to extract methylated CpG from aligned bam files. Before alignment of MCTA-seq data, the primer sequences including 6 bp at the 5′-end and 12 bp at the 3′-end of reads were trimmed with custom scripts. The trimmed fastq files

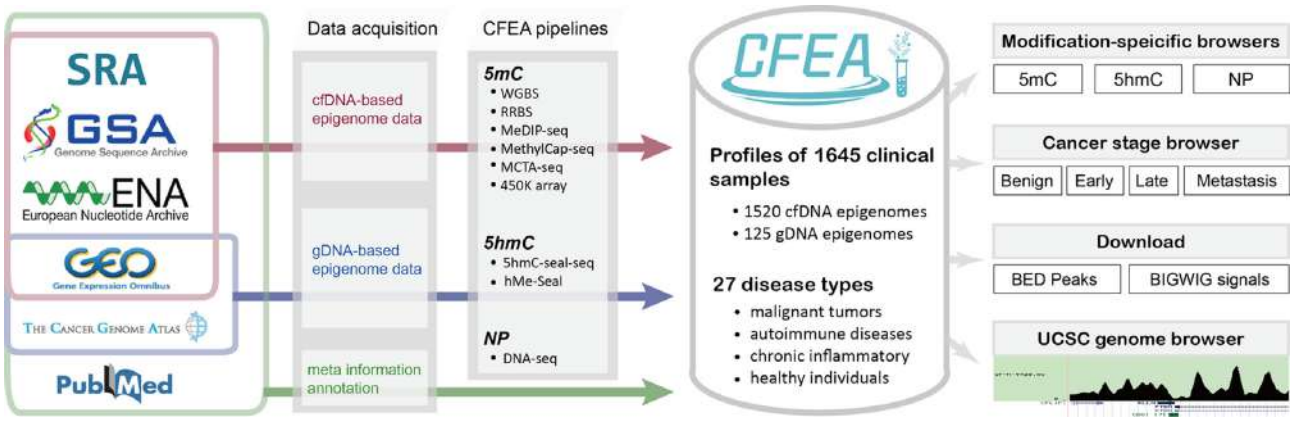


**Figure 1.** A schematic view of the CFEA database. CFEA collects publicly available 5mC, 5hmC, and NP of cfDNA epigenome data for 27 human diseases from SRA, GSA, ENA and GEO. Disease-matched epigenome data for gDNA from the corresponding solid tissue were obtained from GEO and TCGA. All CFEA data were uniformly quality controlled and processed by standard pipelines covering a broad range of experimental types. Clinical and experimental information for each CFEA sample is manually curated. Users can search for modification-specific (5mC, 5hmC and NP) browsers or cancer-stage browsers via a combination of different options. Data can be downloaded in bed or wig format for further analysis and visualized using UCSC genome browsers.

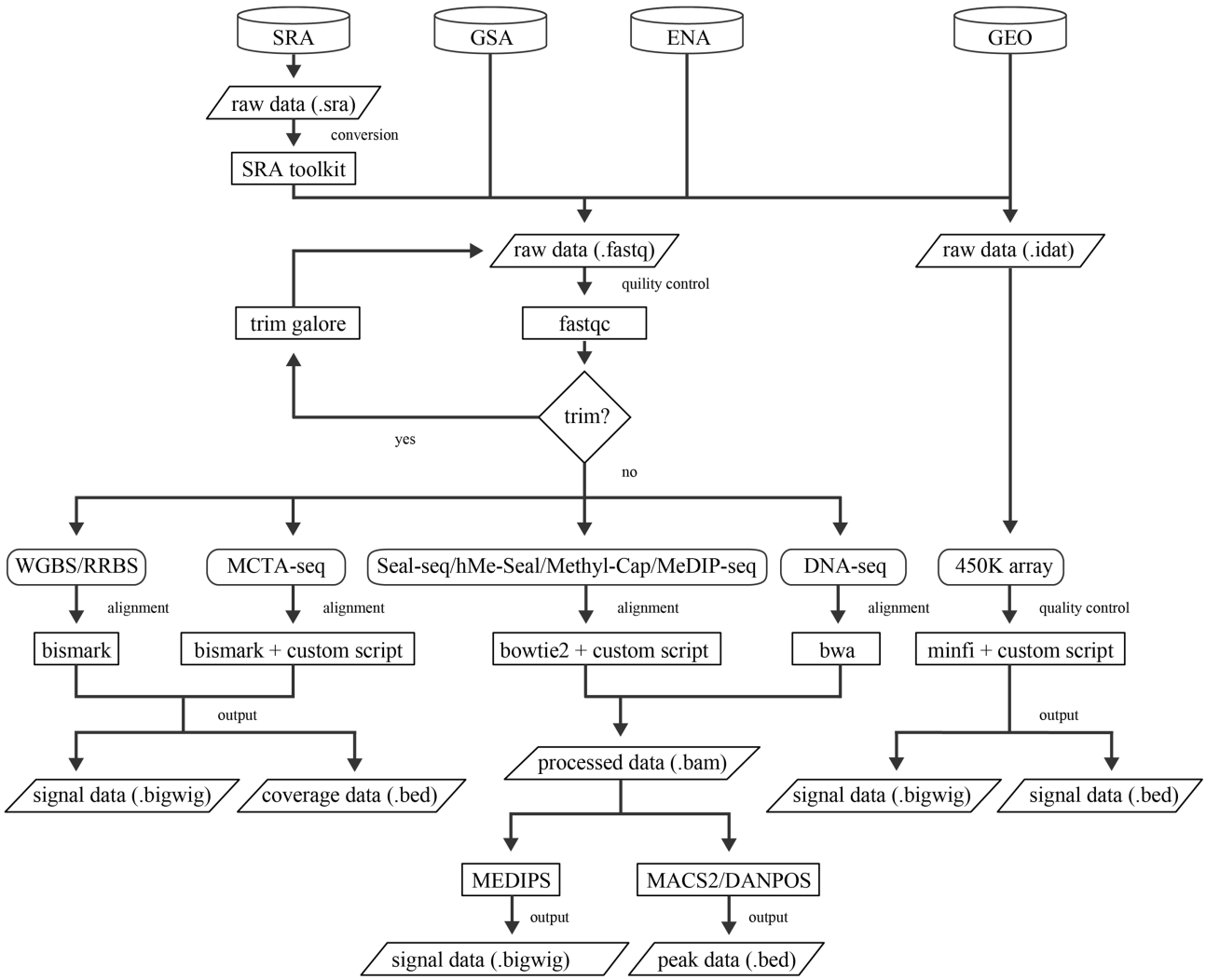


**Figure 2.** Flow diagram of CFEA processing pipelines for the collected epigenome data generated using different technologies.

were aligned to the reference genome using bismark 0.18.2 with the parameters '–bowtie1 –non_directional –fastq –phred33-quals'. Methylated CpG was extracted using bismark_methylation_extractor script. The bigwig format files for bisulfite conversion-based data were generated with custom scripts. Data from enrichment-based techniques, including Methyl-Cap, MeDIP-seq, hMe-Seal and Seal-seq, were aligned to the reference genome using bowtie2 version 2.3.4.3 with default parameters (19). MACS2 version 2.1.1 (20) was used to perform peak calling from aligned bam files with the parameters '-g hs –nomodel –extsize 200 –call-summits'. The single-end reads and the pair end reads of nucleosome positioning data were mapped to the reference genome using software bwa version 0.7.17 (21) with command aln and sampe, respectively. Peak calling for the resulting bam files was performed with DANPOS version 2.2.2 (22) with command dpos. For both enrichment-based and nucleosome positioning data, the bigwig format files were converted from the aligned bam files by MEDIPS R package version 1.36.0 (23). All data were aligned to the same reference genome (hg19) to ensure that the samples could be directly compared. For Illumina DNA methylation 450K array data, arrays were background and control normalized, and beta-values were then calculated using the raw iDAT files with the minfi package version 1.30.0 (24). A more detailed description of our pipelines and source codes are available on the CFEA website.

## DATA QUERY, VISUALIZATION AND RETRIEVAL

CFEA provides four user-friendly browsers to facilitate data browsing and searching. Because three different epigenetic molecular levels were incorporated into CFEA, three special browsers with respect to 5mC, 5hmC and NP of cfDNA on the main page are provided. Users can click icons to quickly search CFEA through corresponding browsers. These browsers can be queried by a custom combination of six dropdown menus, including diseases, detection methods, source of DNA (plasma or serum), whether to show epigenomic data from gDNA for matched cancer tissues, whether to show epigenomic data from cfDNA for healthy individual, interested gene or genomic region for visualization. To ensure non-empty results in user searching, we use jQuery-based scripts to filter the search options automatically. Additionally, we developed a cancer-stage browser to facilitate the search and analysis for epigenomes in cancer at different stages (benign, early, late and metastasis) during cancer development. Similar to the modification-specific browsers described above, the cancer stage browser can be searched using a custom combination of six menus, including cancer types, stages, detection methods, matched gDNA data, matched healthy data, and interested gene/region.

By clicking the 'search' button in the CFEA browsers, a dynamic table with sorting and filtering functions is returned. Each row in the resulting table represents an epigenome profile, and each column contains the sample-specific meta information, including a unique sample identifier assigned by our database, disease types, detection methods, gender, age, drug treatment, DNA type (cfDNA or gDNA), read mapping ratios after QC, original literature and a link to UCSC Genome Browser (25). Regarding the table resulting from the cancer-stage browser, we provide three more columns including information for clinical stages and subtypes. By default, 10 entries of the table are displayed, and the users can change the pagination at the bottom to better display the results. To allow visualization of multiple samples simultaneously or to compare the epigenome profiles of cfDNA and gDNA, users can select interesting samples and visualize them in batches by clicking the button on the top left of each data table. Users can also set up custom UCSC Genome Browser sessions or upload their own data for comparison with CFEA data.

The download page provides users with bed (peak) and bigwig (signal intensity) format data organized by categorized lists of different epigenomic molecular levels and experimental technologies. Users can also retrieve bigwig format data from the resulting table of the search browsers.

## IMPLEMENTATION

CFEA was implemented based on Struts 2 (version 1.0), a java-based web development framework. The current version of CFEA runs on an Apache Tomcat web server deployed on a CentOS Linux server (version 7.5.1804), and its web interfaces were developed based on HTML, CSS and JavaScript. All data in CFEA are stored and managed by a MySQL relational database (version 5.7.26).

## SUMMARY AND FUTURE DIRECTIONS

CFEA is an online cfDNA-based epigenome data portal that allows for the browsing and easy assessment of a large amount of high-quality data for diverse human diseases. All data collected by CFEA are processed by standard pipelines covering a broad range of experimental types. Focusing on malignant cancer, CFEA provides specialized browsers, allowing clinicians or biologists to quickly retrieve relevant clinical information and data for different stages during cancer development. To achieve good comparison and integration, CFEA also provides interfaces for conveniently visualizing and downloading the processed epigenome profiles from cfDNA and gDNA for diverse human diseases. The development of new web tools for mining disease-related cell-free epigenome profiles will advance our understanding of the nature of cfDNA and reveal useful preclinical biomarkers. The field of liquid biopsy will continue to grow rapidly, we will pay close attention to any updated cfDNA-based epigenome data and process and store them using standard pipelines once they become available.

## DATA AVAILABILITY

The CFEA database is freely accessible for non-commercial use at http://www.bio-data.cn/CFEA or http://128.1.137.15/CFEA. Users are not required to register or login to access any feature available in the database.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

National Natural Science Foundation of China [61871294, 61873193, 81822030, in part]; Science Foundation of Zhejiang Province [LR19C060001]; Beijing Natural Science Foundation [7172175]; Special Foundation for Key Basic Research of Wenzhou Institute of Biomaterials and Engineering, CAS, China [WIBEZD2017009-05]; CAMS Innovation Fund for Medical Sciences [2016-I2M-1-001, 2017-I2M-3-004]. Funding for open access charge: National Natural Science Foundation of China [61871294, 61873193, 81822030, in part]; Science Foundation of Zhejiang Province [LR19C060001]; Beijing Natural Science Foundation [7172175]; Special Foundation for Key Basic Research of Wenzhou Institute of Biomaterials and Engineering, CAS, China [WIBEZD2017009-05]; CAMS Innovation Fund for Medical Sciences [2016-I2M-1-001, 2017-I2M-3-004].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Diaz,L.A. Jr and Bardelli,A. (2014) Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.*, **32**, 579–586.
2. Corcoran,R.B. and Chabner,B.A. (2018) Application of cell-free DNA analysis to cancer treatment. *N. Engl. J. Med.*, **379**, 1754–1765.
3. Cabel,L., Proudhon,C., Romano,E., Girard,N., Lantz,O., Stern,M.H., Pierga,J.Y. and Bidard,F.C. (2018) Clinical potential of circulating tumour DNA in patients receiving anticancer immunotherapy. *Nat. Rev. Clin. Oncol.*, **15**, 639–650.
4. Moss,J., Magenheim,J., Neiman,D., Zemmour,H., Loyfer,N., Korach,A., Samet,Y., Maoz,M., Druid,H., Arner,P. *et al.* (2018) Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.*, **9**, 5068.
5. Cristiano,S., Leal,A., Phallen,J., Fiksel,J., Adleff,V., Bruhm,D.C., Jensen,S.O., Medina,J.E., Hruban,C., White,J.R. *et al.* (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, **570**, 385–389.
6. Dor,Y. and Cedar,H. (2018) Principles of DNA methylation and their implications for biology and medicine. *Lancet.*, **392**, 777–786.
7. Shen,S.Y., Singhania,R., Fehringer,G., Chakravarthy,A., Roehrl,M.H.A., Chadwick,D., Zuzarte,P.C., Borgida,A., Wang,T.T., Li,T. *et al.* (2018) Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*, **563**, 579–583.
8. Li,W., Zhang,X., Lu,X., You,L., Song,Y., Luo,Z., Zhang,J., Nie,J., Zheng,W., Xu,D. *et al.* (2017) 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res.*, **27**, 1243–1257.
9. Ulz,P., Thallinger,G.G., Auer,M., Graf,R., Kashofer,K., Jahn,S.W., Abete,L., Pristauz,G., Petru,E., Geigl,J.B. *et al.* (2016) Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.*, **48**, 1273–1278.
10. Heitzer,E., Haque,I.S., Roberts,C.E.S. and Speicher,M.R. (2019) Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.*, **20**, 71–88.
11. Feng,H., Jin,P. and Wu,H. (2019) Disease prediction by cell-free DNA methylation. *Brief. Bioinform.*, **20**, 585–597.
12. Xu,R.H., Wei,W., Krawczyk,M., Wang,W., Luo,H., Flagg,K., Yi,S., Shi,W., Quan,Q., Li,K. *et al.* (2017) Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat. Mater.*, **16**, 1155–1161.
13. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
14. Leinonen,R., Sugawara,H., Shumway,M. and International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
15. Members, B.I.G.D.C. (2019) Database resources of the BIG data center in 2019. *Nucleic Acids Res.*, **47**, D8–D14.
16. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.
17. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
18. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
19. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
20. Feng,J., Liu,T., Qin,B., Zhang,Y. and Liu,X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.*, **7**, 1728–1740.
21. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
22. Chen,K., Xi,Y., Pan,X., Li,Z., Kaestner,K., Tyler,J., Dent,S., He,X. and Li,W. (2013) DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.*, **23**, 341–351.
23. Lienhard,M., Grimm,C., Morkel,M., Herwig,R. and Chavez,L. (2014) MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, **30**, 284–286.
24. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.
25. Casper,J., Zweig,A.S., Villarreal,C., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Karolchik,D. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.