Computational and Structural Biotechnology Journal 18 (2020) 1891–1903

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Finding new cancer epigenetic and genetic biomarkers from cell-free DNA by combining SALP-seq and machine learning

Shicai Liu [a], Jian Wu [a], Qiang Xia [a], Hongde Liu [a], Weiwei Li [b], Xinyi Xia [b,*], Jinke Wang [a,*]

[a] State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China
[b] Jinling Hospital, Nanjing University School of Medicine, Nanjing 210002, China

ARTICLE INFO

Article history:
Received 24 March 2020
Received in revised form 29 June 2020
Accepted 29 June 2020
Available online 7 July 2020

Keywords:
Cell-free DNA
Biomarkers
Esophageal cancer
Next generation sequencing
SALP-seq

ABSTRACT

The effective non-invasive diagnosis and prognosis are critical for cancer treatment. The plasma cell-free DNA (cfDNA) provides a good material for cancer liquid biopsy and its worth in this field is increasingly explored. Here we describe a new pipeline for effectively finding new cfDNA-based biomarkers for cancers by combining SALP-seq and machine learning. Using the pipeline, 30 cfDNA samples from 26 esophageal cancer (ESCA) patients and 4 healthy people were analyzed as an example. As a result, 103 epigenetic markers (including 54 genome-wide and 49 promoter markers) and 37 genetic markers were identified for this cancer. These markers provide new biomarkers for ESCA diagnosis, prognosis and therapy. Importantly, these markers, especially epigenetic markers, not only shed important new insights on the regulatory mechanisms of this cancer, but also could be used to classify the cfDNA samples. We therefore developed a new pipeline for effectively finding new cfDNA-based biomarkers for cancers by combining SALP-seq and machine learning. In this study, we also discovered new clinical worth of cfDNA distinct from other reported characters.

1. Introduction

Cancer is an important public health problem worldwide. Its morbidity and mortality are increasing year by year, and the treatment effect is poor, which seriously affects people's health and quality of life. According to GLOBOCAN data [1], there were approximately 18.08 million new cancer cases and 9.56 million deaths in the world in 2018, most of them live in low- or middle-income countries. It is estimated that by 2025, there will be about 20 million new cases of cancer every year [2]. The latest cancer statistics show that [3], 1,762,450 new cancer cases and 606,880 cancer deaths are projected to occur in the United States in 2019. Although the incidence of cancer has not changed significantly, cancer mortality has continued to decline, not only because of the development of medical standards, but also for preventive screening.

Tissue biopsy is still the gold standard for tumor dosgnosis, but due to the traumatic nature to patients, it brings a lot of interference to the dynamic treatment of patients. There are also many risks and ethical issues in tissue biopsy, which makes it have certain limitations. Liquid biopsy is a kind of cutting-edge technology to analyze a range of tumor materials in the blood or other body fluids in a minimally invasive or noninvasive manner. The tumor materials include circulating tumor cells (CTCs), circulating tumor DNA (ctDNA), cell-free DNA (cfDNA), messenger RNA (mRNA), microRNA (miRNA), and exosomes. For example, these liquid biopsies have been intensively investigated for the most dentrimental cancer of non-small cell lung cancer (NSCLC) [4–11]. Especially, the liquid biopsy for plasma ctDNA has been widely used in companion diagnosis to guide precision therapy of NSCLC. For example, the NSCLC patients diagnosed the T790M EFGR mutations with plasma had similar outcomes as those diagnosed with tissue when treating with the third-generation EGFR inhibitor Osimertinib [12]. FDA has approved the EGFR plasma mutation test as a guide to treatment of NSCLC patients with EGFR inhibitors [13]. However, ctDNA detection is still difficault

Abbreviations: CTC, circulating tumor cell; ctDNA, cell-free tumor DNA; mRNA, messenger RNA; miRNA, microRNA; cfDNA, cell-free DNA; NIPT, noninvasive prenatal testing; SNP, single nucleotide polymorphism; ESCA, esophageal cancer; SALP-seq, Single strand Adaptor Library Preparation-sequencing; NGS, next generation sequencing; PCA, principal component analysis; TSS, transcription start site; TCGA, The Cancer Genome Atlas; ATAC-seq, Assay for Transposase-Accessible Chromatin-sequencing and high-throughput sequencing; TF, transcription factor; TFBS, TF binding site; SNV, single nucleotide variant; Ti, transitions; Tv, transversion; cfMeDIP-seq, cell-free methylated DNA immunoprecipitation and high-throughput sequencing; AUC, Area Under Curve.

* Corresponding authors at: State Key Laboratory of Bioelectronics, Southeast University, Sipailou 2, Nanjing, 210096, China.

E-mail addresses: xiaxynju@163.com (X. Xia), wangjinke@seu.edu.cn (J. Wang).

https://doi.org/10.1016/j.csbj.2020.06.042

to be applied to early diagnosis of tumor because the amount of ctDNA is very low in plasma and targeted enrichment has to be done for accurate detection. Therefore, other easier liquid biopsies for early diagnosis of tumor are still in demand.

Compared to ctDNA, cfDNA is a more available material for liquid biopsy, which is all free DNA floating in blood. CfDNA was discovered in 1948 [14]. Most of the plasma cfDNA originated from the hematopoietic system in healthy subjects, but in clinical patients (e.g. pregnancy and cancer), the related cells/tissues would release additional DNA into the plasma [15,16]. The detection of this perturbation would allow us to diagnose the abnormality for people in a noninvasive way. In recent years, methods based on the analysis of plasma cfDNA have been largely explored for noninvasive prenatal testing (NIPT) and cancer liquid biopsy [17,18]. For example, the cfDNA-based fetal aneuploidy test in pregnant women was routinely deployed in many countries by 2014, and the market value is estimated to reach 3.6 billion USD in 2019 [19]. Studies on the applications of cfDNA for cancer detection and tumor origin determination have also demonstrated the high clinical potential of cfDNA [20–22]. In these studies, a variety of methods were developed for differentiating the cfDNA molecules released by the tissues of interest (e.g., ctDNA) from the background ones [23,24]. Some methods have utilized genetic biomarkers, such as the fetal-specific informative single nucleotide polymorphism (SNP) sites in pregnancies and somatic mutations in cancer patients [25,26]. However, such genetic biomarkers usually vary from case to case, which challenges the development of sensitive and generalizable approaches. Under this circumstance, the epigenetic biomarkers are more favored.

Esophageal cancer (ESCA) continues to be a leading cause of cancer death worldwide, and approximately 480,000 cases are diagnosed annually worldwide [27]. Although technologies in surgical treatment and systemic therapy advanced during the past few decades, over 400,000 cases have died from ESCA within the last 5 years [18]. The predicted 5-year survival rate of ESCA, which ranges from 15% to 20%, has barely improved in recent decades due to high recurrence rate, early metastatic tendency, and limited knowledge of biomarkers and potential therapeutic targets [28,29]. Thus, finding new biomarkers for the diagnose of ESCA is needed urgently, especially those biomarkers applicable to liquid biopsy for early discovery, diagnosis and prognosis of ESCA.

In this study, we tried to find both epigenetic and genetic biomarkers of ESCA in cfDNA with adapted Single strand Adaptor Library Preparation-sequencing (SALP-seq) in combination with machine learning. The adapted SALP-seq [30,31], developed by our laboratory, is a new single-stranded DNA library preparation technique. This technique is particularly suited to construct the next generation sequencing (NGS) libraries for highly degraded DNA samples such as cfDNA [31]. Moreover, by using the barcode T adaptors, this technique is competent to analyze many cfDNA samples in a high-throughput format [31]. In this study, the NGS libraries of 20 cfDNA samples, which were from 11 pre-operation ESCA patients, 5 post-operation ESCA patients, and 4 healthy people, were constructed by using SALP-seq. Based on bioinformatics analysis of the sequencing data, we identified 103 epigenetic markers (including 54 genome-wide and 49 promoter markers) and 37 genetic markers for ESCA, which may ultimately contribute to the development of effective diagnostic and therapeutic approaches for ESCA. Furthermore, these markers were verified by analyzing 10 new cfDNA samples from pre-operation ESCA patients.

2. Experimental procedure

2.1. Sample processing and sequencing

Plasma DNA libraries were constructed from whole blood with adapted SALP method [30,31]. CfDNA was extracted from 200 μl of plasma. The purified cfDNA was quantified with Qubit 2.0 and 4 ng of each cfDNA sample was used to prepare NGS library. Finally, the Illumina-compatible libraries were generated for 20 cfDNA samples (Table S1) by using adaptors and primers listed in Table S2. The library DNA was quantified with Qubit 2.0 and pooled at the same quality (ng) to generate a final sequencing library. The library was sequenced by two lanes of Illumina Hiseq X Ten platform (Nanjing Geneseeq). Paired-end sequencing was performed. The details of sample processing and sequencing protocols were described in the Supplementary information. In the validation experiment, we sequenced 10 new cfDNA samples using the same method.

2.2. Analysis and statistics of cfDNA sequencing data

The raw sequencing data of cfDNA were separated with the barcode by using homemade perl scripts. Then the constant sequence (19 bp) and barcode (6 bp) sequences were removed from the 5′ end of the pair-end sequencing reads 2. All sequencing reads were analyzed using the Bowtie2 tool [32], with the parameter -X 2000 to keep the long fragments. The paired-end sequencing reads were mapped to the hg19 human reference genome in a paired-end mode, allowing one mismatches for the alignment for each end. The sequence alignment map (SAM) file was converted to BAM format using samtools [33]. Only paired-end sequencing reads with both ends aligned to the same chromosome with the correct orientation were used for downstream analysis. SNV was analyzed by samtools. The annotation of SNV was performed by ANNOVAR [34] with default setting. Functional enrichment analysis through the Database DAVID [35] identified the biological significance of genes. *P*-value adjusted by Benjamini-Hochberg to <0.05 established the cut-off criteria. Reads number was calculated with bedtools [36]. The openness of 1-kb regions upstream TSS in different samples was detected with DEseq2 [37], regions with $p < 0.05$ were selected.

In the screening of ESCA-associated important regions of the whole genome, the mean decrease in accuracy (MDA) was used, which was calculated using the randomForest package in R. MDA represents the average decrease of classification accuracy on the OOB samples when the values of a particular feature are randomly permuted. Therefore, the permutation based MDA can be utilized to evaluate the contribution of each feature to the classification. After the ESCA-associated important regions were screened, machine learning was used to classify cancer and normal samples. Due to the small sample size, classifying was performed by using SVM that shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition problems. The gene expression data were downloaded from TCGA data portal, containing RNA-seq data of 23 cancers and their corresponding normal samples. The comparison of RNA-seq datasets of selected genes between cancer and normal samples were analyzed with R scripts. The H3K27ac ChIP-seq data of TE7 ESCA cell lines were downloaded from GEO database with the accession number GSE76861 [38]. The ATAC-seq data of ESCA tissues derived from donors with diverse demographic features were downloaded from the TCGA database and the hg38 coordinates were converted to hg19 using LiftOver [39]. All tracks were shown with UCSC genome browser.

3. Results

3.1. Clinical specimens and SALP-seq

All procedures used in this research were performed according to the Declaration of Helsinki. This study was approved by the

Ethics Committee of Jinling Hospital (Nanjing, China). All participants were recruited from the Jinling Hospital, Nanjing University School of Medicine (Nanjing, China), with informed consent. The detail information of 20 whole blood samples collected from the Jinling Hospital was listed in Table S1. We sequenced the cfDNAs from as many as 20 blood samples in two lanes (Lane 1 and Lane 2; Table S1) of Illumina Hiseq X Ten platform. The previously published one-lane cfDNA sequencing data [31] from the same 20 blood samples were also used in this study (Lane 3; Table S1). We merged these sequencing data of three lanes with samtools [33] for subsequent analysis (Total: Lane 1 + Lane 2 + Lane 3; Table S1), which showed that the mean value of total reads was 7.2×10^7 with the mappable ratio of 89.8% and the depth of 6.0. The data analysis in this study was, unless otherwise stated, based on the data from the merged three lanes. In the validation experiment, we sequenced 10 new cfDNA samples using the same method and the detail information of samples was listed in Table S1, which showed that the mean value of total reads was 3.4×10^7 with the mappable ratio of 89.6% and the depth of 2.9.

3.2. Finding cancer diagnostic/prognostic markers from chromatin accessibility of promoters

We showed the distribution of cfDNA in the whole genome by calculating and normalizing the reads density in each 1-Mb window. The results revealed that the distribution of reads of different samples' cfDNA was greatly different through the whole genome (Fig. 1A). In the previous study, we have confirmed that the NGS data can be used to characterize the chromatin state of different types of cfDNA [31]. In mechanism, only nucleosome-protected genomic regions can be sequenced in the NGS of cfDNA [31]. To identify normal or cancer samples by viewing the signal strength of the reads distribution around the transcription start site (TSS), we calculated the reads density of ± 5-kb region around TSSs of all human genes and the average reads density using deeptools (parameter: RPKM) for 20 cfDNA samples. The results showed that a peak was formed around the TSS in normal samples, and a valley was formed in the pre-operation cancer samples (Fig. 1B; Fig. S1). Moreover, we performed a principal component analysis (PCA) of reads density of ± 5-kb region around the TSSs of all human genes. As a result, the cfDNAs of pre-operation ESCA patients could be clearly distinguished from those of normal people by analyzing the cfDNA data with PCA (Fig. 1C). We also sequenced 10 new cfDNA samples to verify this result and got the consistent results (Fig. S2).

In the previous study we have confirmed that the SALP-seq data of cfDNA can be used to characterize the chromatin state of different types of cfDNA. In this study, we differentiated cancer and normal samples by chromatin openness. The reads density of all promoters (defined as the 1–kb region upstream TSS) was calculated for each sample. The results showed that there was a great difference of reads density of promoters between normal cfDNA samples and ESCA cfDNA samples (Fig. 2A). Notably, some promoters showed extremely low reads density in all cancer samples but high density in normal samples (Fig. 2B). There were 49 genes having such a distinct feature. Through the clustering results of these 49 genes from the heatmap, it was clearly seen that the cancer samples can be distinguished from the normal samples (Fig. 2B). We calculated the reads density of promoters of the 49 genes in the post-operation cancer samples. The results showed that the reads density of promoters of most of these genes was significantly increased in the post-operation cancer compared with pre-operation cancer (Fig. 2C), indicating the effect of surgery. Moreover, we sequenced 10 new cfDNA samples to verify this result and got the consistent results (Fig. 2D). These data suggested that the chromatin accessibility of the promoters of these 49 genes can be used as the diagnostic and prognostic markers for cancer.

Through database and literature search, we found that 15 (genes in red in Fig. 2B) of these 49 genes have been reported to be closely associated with ESCA (Table S3). Therefore, we inferred that the remaining 34 genes are newly discovered genes associated with ESCA. To validate the relationship between these 49 genes and ESCA, we downloaded the RNA-seq data (containing 163 ESCA and 11 normal samples) of ESCA from The Cancer Genome Atlas (TCGA) database for analysis. We found that the expression of most of these genes were significantly up-regulated in cancer samples (Fig. S3A). In the process of database and literature search, we found that some of these 49 genes are not only related to ESCA, but also associated with other cancers (Table S3). We downloaded RNA-seq data of 23 cancers from the TCGA database to analyze the relationship between these genes and various cancers. The results showed that most of these genes were significantly up-regulated in various cancers (Fig. S3B). The above results indicated that the chromatin accessibility of promoters of the selected 49 ESCA-associated genes is distinct between normal and cancer, and chromatin accessibility of these promoters can be also used to diagnose and prognose cancer. To further understand the functional roles of these genes in ESCA, we performed GO analysis. As a result, both chromosome organization (GO: 0051276) and chromatin organization (GO: 0006325) were significantly enriched, implying that some of these genes play critical roles in regulating the chromosome or chromatin structure (Fig. S4A). In the GO term of chromosome organization, there were nine genes, including BRPF1, NIPBL, GEN1, SUV39H1, HIST1H4E, INO80, PELO, WHSC1, and ING1. In the GO term of chromatin organization, there were seven genes including BRPF1, NIPBL, SUV39H1, HIST1H4E, INO80, WHSC1, and ING1. Among the genes enriched in GO: 0051276 and GO: 0006325, SUV39H1, INO80, WHSC1, and ING1 were known ESCA-associated genes. Other enriched GO terms were related to regulation of cell cycle (GO: 0051726), growth (GO: 0040007) and so on, which all play an important role in cancer development (Fig. S4A). Pathway annotation was used to screen out the altered biological functions arising from the 49 selected genes. The results indicated that these genes were mainly enriched in 5 pathways, including Lysine degradation, mTOR signaling pathway, mRNA Surveillance pathway, Insulin resistance, and Spliceosome (Fig. S4B).

3.3. Finding cancer diagnostic/prognostic markers from chromatin accessibility of whole genome

To find ESCA-associated important regions from the whole genome, we calculated the reads density in each 1-kb window of the whole genome, and then used MDA to screen out 88 ESCA-associated regions. Most of these regions were non-coding sequences, suggesting that these regions contain important regulatory elements (Fig. 3). Then we pinpointed the genomic location of these regions, and found that 36.36% of these regions were located in the distal intergenic regions (more than 10-kb from TSSs), and 25% of these regions were located in the Proximal regulatory region (10-kb regions upstream TSSs) (Fig. 3, inset). It can also be seen from the MDA diagram that distal elements (defined as occurring inside of distal intergenic) were much more important than promoter elements (defined as occurring inside of Proximal regulatory region) in the classification (Fig. 3), indicating that distal elements exhibited a greater specificity and wider dynamic range of activity in association with cancer, whereas promoter element accessibility was less cancer-specific. This functional specificity of distal regulatory elements was also previously observed in healthy tissues and in cancer [39,40].

We researched the ESCA-associated important regions that were located in distal intergenic and proximal regulatory regions.

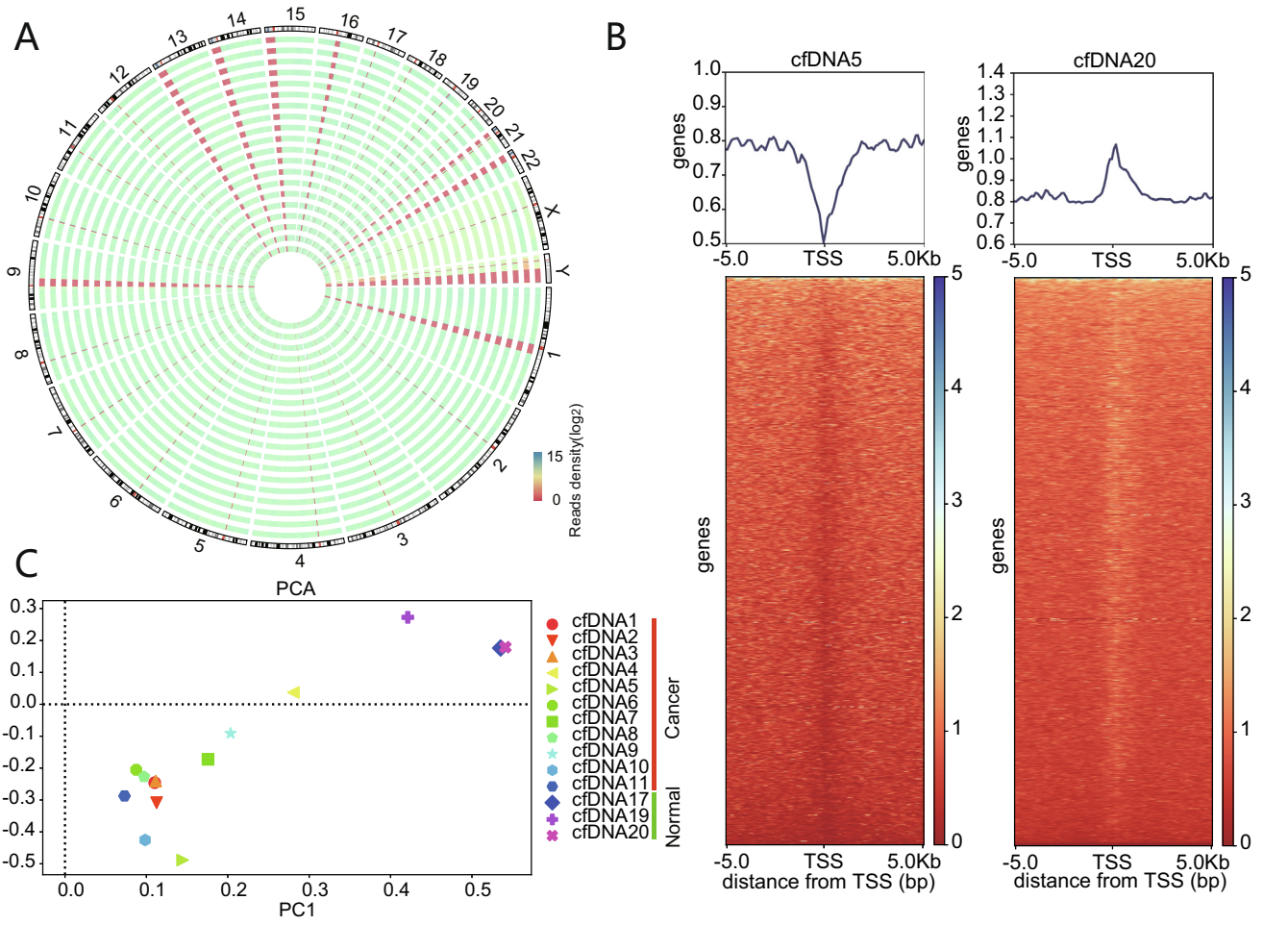

Fig. 1. Characterization of cfDNA NGS reads distribution. (A) Distribution of reads density in each 1-Mb window of whole genome in different cfDNA samples. Reads density of cfDNA 1 to 20 was shown in order from outside to inside. (B) The signal strength of reads distribution around TSSs. The line plot shows the average signal strength of all regions around TSS. The left shows the signal strength of cfDNA5 (cancer sample). The right shows the signal strength of cfDNA20 (normal sample). (C) Principal component analysis (PCA) of reads density of ± 5-kb region around the TSSs of all human genes.

The Fig. 4A showed the reads density in the ESCA-associated important regions (including 32 distal intergenic and 22 proximal regulatory regions) of each sample, which indicated that there was a great difference between normal cfDNA samples and ESCA cfDNA samples. In addition, we also calculated the reads density of the 54 regulatory regions in the post-operation cancer samples. The results showed that the reads density of most of these regions was increased in the post-operation cancer compared with pre-operation cancer (Fig. 4B), showing the effect of surgery. These data suggested that the chromatin accessibility of these genomic regions can be used as the diagnostic and prognostic markers for cancer. To further validate these markers, we sequenced 10 new cfDNA samples to verify this result. As a result, all the consistent results were obtained (Fig. 4C). These data also indicated that more new caner-associated markers could be identified using the whole-genome chromatin accessibility characterized with cfDNA.

To further verify the chromatin accessibility of these genomic regions characterized with cfDNA using SALP-seq, we downloaded the H3K27ac ChIP-seq data of TE7 ESCA cell lines from GEO with the accession number GSE76861 and the ATAC-seq data of 19 ESCA tissues derived from donors with diverse demographic features from the TCGA database [38,39]. These data together with SALP-seq data were compared by visualizing these regions (including 32 Distal Intergenic regions and 22 Proximal regulatory regions) with UCSC genome browser. The results revealed that the chromatin accessibility of these genomic regions characterized with cfDNA using SALP-seq in this study were highly consistent with those characterized with cancer cells and tissues using H3K27ac ChIP-seq and ATAC-seq (Fig. S5).

Because these regions are non-coding regions and their chromatin accessibility are significantly changed, they should play regulatory functions in tumor by providing accessible binding sites to transcription factors (TFs). We therefore searched the potential TF binding sites in these regions using FIMO with the motif matrix obtained from the HOCOMOCO (version 11) with default setting. The results showed that these regions contained a large number of TF binding sites (TFBSs). Moreover, we compared these regions with the SEdb database, a comprehensive human super-enhancer database. The results revealed that 17 of these 54 regions were well known super-enhancer elements (Fig. 3).

In order to find the target genes regulated by the selected ESCA-associated distal elements and promoter elements, we predicted target genes for these genomic regions by using EnhancerAtlas with parameter "Esophagus". The results showed that these regions were assigned to 104 genes (Fig. 3), 16 of which were also present in the 49 genes identified above with promoter chromatin accessibility, including LSM12, MAP4K1, SF3A2, SLC25A28, RCN2, YIF1B, NCBP2, EIF1AD, LPGAT1, WHSC1, SUV420H1, ATP6V1A, INO80, PPIB, MRPL45, and ING1. These data not only illustrated the reliability of our results, but also indicated that more

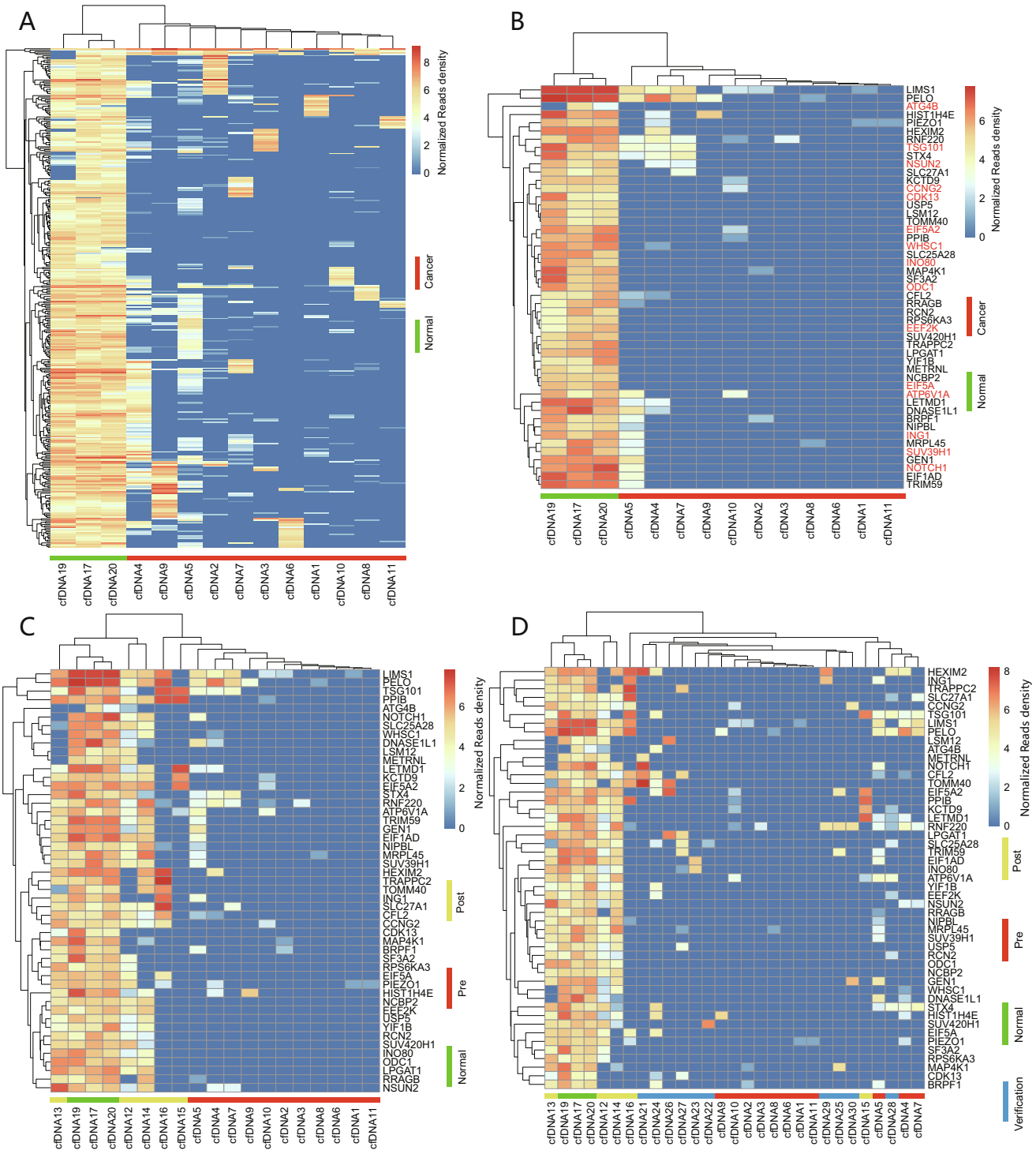

Fig. 2. Analysis of gene promoters. (A) The heat map and clustering of reads density of promoters with significant difference in chromatin accessibility in cfDNA samples. (B) The heat map and clustering of reads density of promoters of 49 selected genes. The genes in red are known ESCA-associated genes. (C) The heat map and clustering of reads density of promoters of 49 selected genes in the post-operation cancer cfDNA samples. (D) The heat map and clustering of reads density of promoters of 49 selected genes in the 10 verification cancer cfDNA samples. Pre: pre-operation cfDNA; Post: post-operation cfDNA; Normal: normal cfDNA; Verification: 10 cfDNA samples of verification. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cancer-associated genes could be identified using the whole-genome chromatin accessibility characterized with cfDNA. To further validate these potential cancer-associated genes, we downloaded the RNA-seq data of ESCA from the TCGA database for analysis. The results indicated that the expression of most of these genes were significantly up-regulated in cancer samples (Fig. S6). Through database and literature search, we found that 21 of these 104 genes have been reported to be closely associated with ESCA (Fig. 3; Table S4). Therefore, we inferred that the others are newly discovered genes associated with ESCA. The gene annotation revealed that these genes were mainly associated with the biological processes of apoptotic, metabolic, cell growth, and translational initiation, and molecular function of histone lysine N-methyltransferase activity and interferon activity (Fig. S7A).

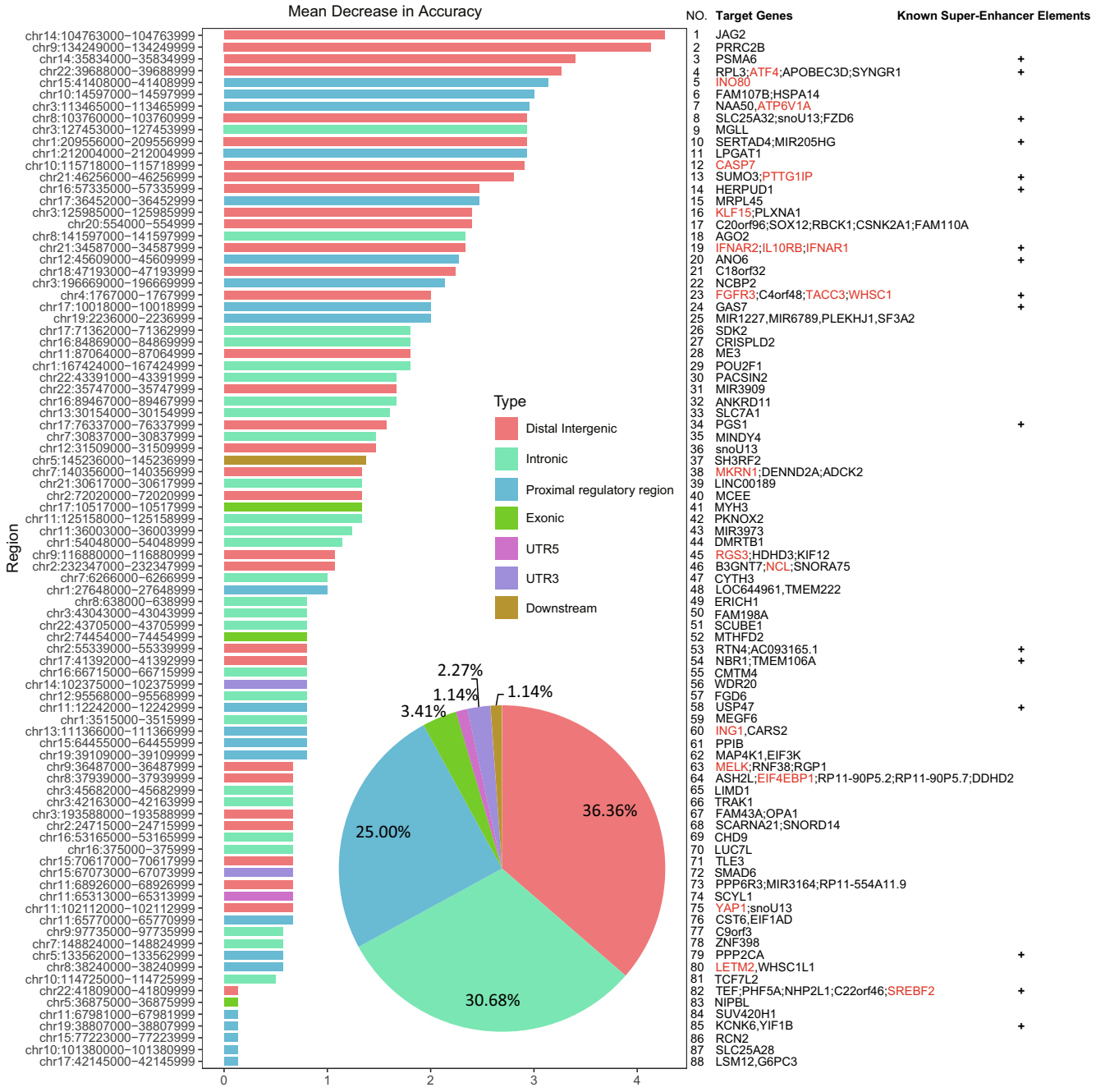

Fig. 3. Analysis of ESCA-associated important regions in genome. The 88 ESCA-associated important regions were selected with MDA. The ordinate represents the regions, and the abscissa represents the importance value. Different colors indicate the genomic location of these regions. Inset shows the distribution of genomic location of these ESCA-associated important regions in genome. Most of the regions are located in distal intergenic, intronic, and proximal regulatory region. The genes and known super-enhancer elements assigned to these regions are shown. The genes in red are the know ESCA-associated genes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The Pathway annotation showed that these genes were mainly enriched in 9 pathways, including signaling pathways of PI3K-Akt, AMPK, Hippo, and Jak-STAT (Fig. S7B). These biological processes, molecular functions and pathways all play important roles in cancers.

3.4. Establishing classification model of ESCA based on the identified ESCA-associated regions

To build a classification model for predicting ESCA, we then analyzed 88 regions with the support vector machine (SVM; e1071 package in R) algorithm [41]. The results revealed that the established SVM model could accurately distinguish cancer samples from normal samples with an Area Under Curve (AUC) value of 1.0 (Fig. 5A). To further improve the clinical applicability of the classification model, after debugging and screening, we finally selected top 24 ESCA-associated important regions to re-establish the model. As a result, the re-established model could still accurately distinguish cancer and normal samples with an AUC value of 1.0 (Fig. 5B). The validation of the model with the later sequenced cfDNA samples obtained good prediction results with accuracy 93.8% (Fig. 5C). In order to explore the effect of the reads

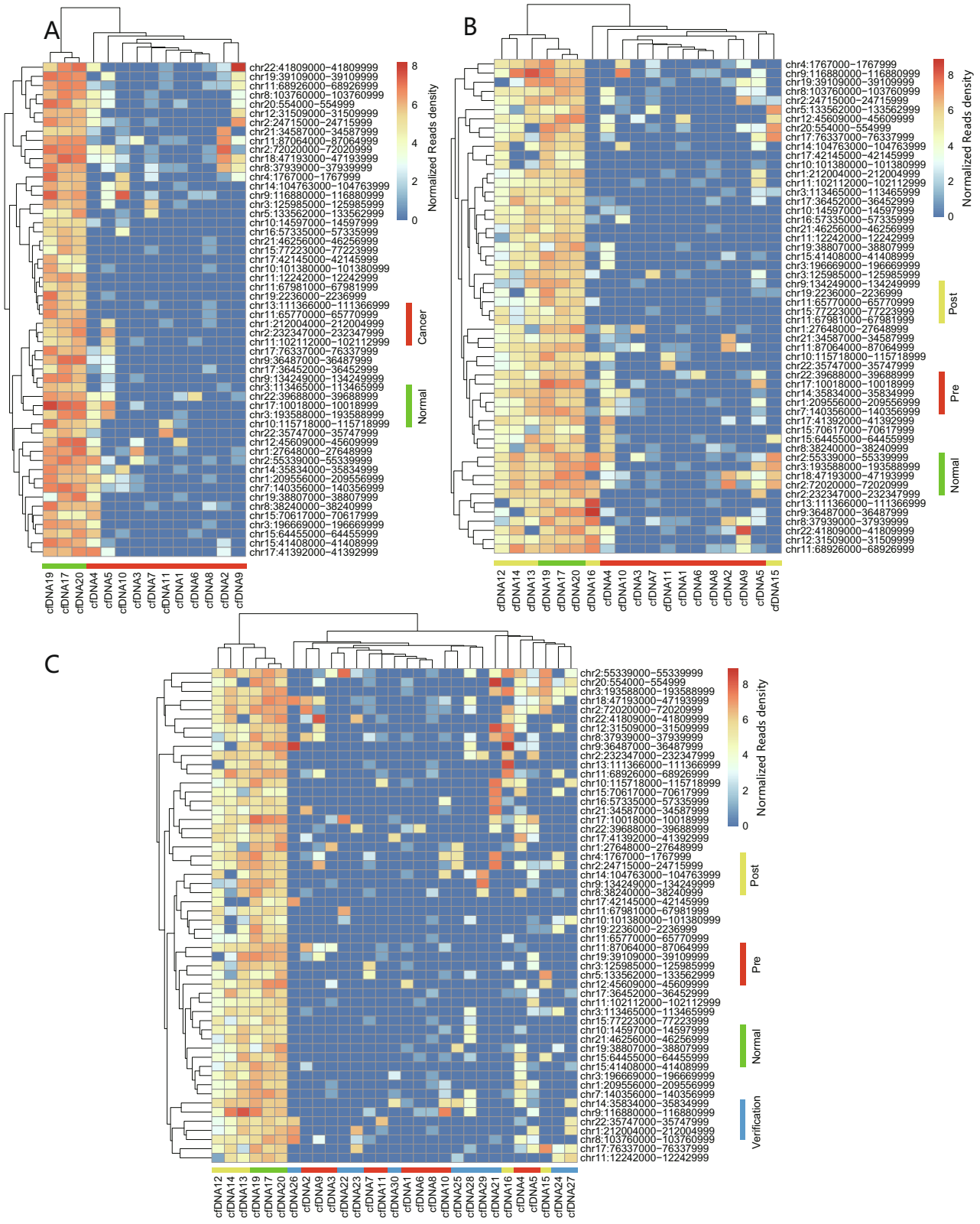

Fig. 4. Analysis of the selected 54 ESCA-associated important regions. (A) The heat map and clustering of reads density of ESCA-associated important regions (including 32 distal intergenic and 22 proximal regulatory regions) of each sample. (B) The heat map and clustering of reads density of ESCA-associated important regions in the post-operation cancer cfDNA samples. (C) The heat map and clustering of reads density of ESCA-associated important regions in the 10 verification cancer cfDNA samples.

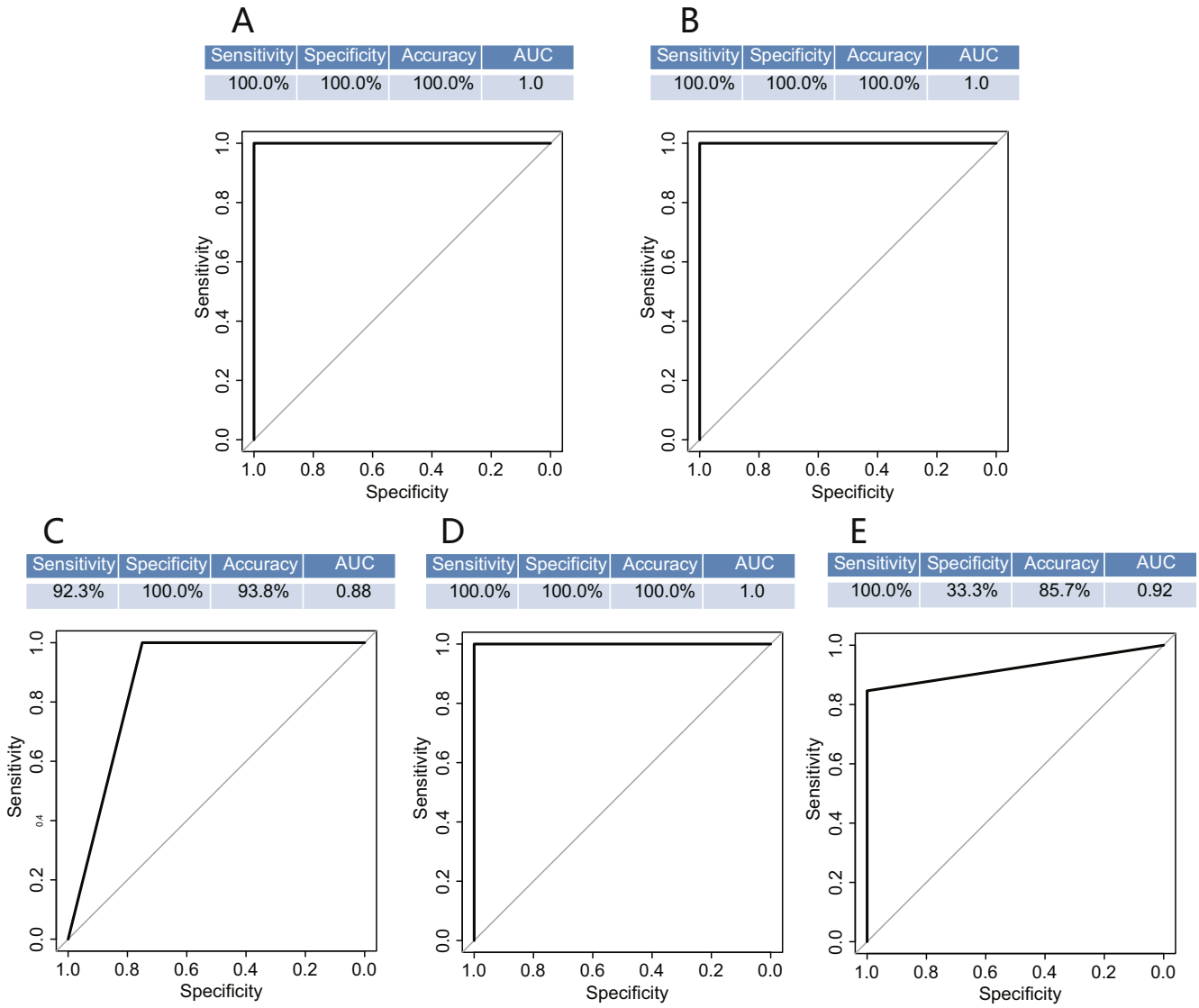

Fig. 5. Classification of ESCA and normal cfDNA based on ESCA-associated important regions. (A) Results of classification of cancer and normal cfDNA samples using SVM based on 88 ESCA-associated important regions. (B) Classification results based on top 24 ESCA-associated important regions. (C) Classification results of the verification data. (D) Classification results of 1×10^7 reads that were extracted from each sample. (E) Classification results of 1×10^6 reads that were extracted from each sample. The picture below shows the receiver operating characteristic curve (ROC).

number on the model, we selected 1×10^6 and 1×10^7 reads from 20 firstly sequenced cfDNA samples and then predicted separately. The results showed that the model still maintained a good predictive effect at 1×10^7 reads (Fig. 5D–E), suggesting that the model can be used to predict ESCA at lower cfDNA sequencing depths.

3.5. Characterizing ESCA-associated mutations with cfDNA

Mutations analysis of cfDNA with target or genome scale sequencing was widely used in NIPT or liquid biopsy. We next analyzed ESCA-associated mutations with SALP-seq reads of cfDNA samples. The results indicated that there were mutations in the whole genome (Fig. 6A). We extracted mutational signatures of 20 cfDNA samples using 6 kinds of base substitutions (C > A, C > G, C > T, T > A, T > C, and T > G). The results indicated that there were variant levels of all these base substitutions in different individuals (Fig. 6B). The results showed that the pre-operation ESCA cfDNA had lower C > T and T > C transitions than normal cfDNA, but had higher C > G and C > A transversions than normal cfDNA (Fig. 6C). There were also significant differences of transitions (Ti) and transversion (Tv) frequencies between pre-operation ESCA and normal samples (Fig. 6D). Moreover, there was a significant difference of Ti/Tv ratio between pre-operation ESCA and normal samples (Fig. 6E). Importantly, the surgical treatment evidently changed the seven mutation features (Fig. 6C–E). The seven mutation features can be developed into diagnostic markers for ESCA liquid biopsy. To pinpoint the genomic location of the single nucleotide variant (SNV), we systematically annotated the SNV using ANNOVAR [34]. The results indicated that most of the SNVs were located in distal intergenic and intronic regions (Fig. 6F).

To find the mutations in coding sequences of all genes, we analyzed mutations in each cfDNA samples. The results indicated that there were large amount of mutations in thousands of genes in pre- and post-operated ESCA cfDNAs and normal cfDNA (Table S5). To test whether the clinically relevant mutations can be detected by the cfDNA NGS, we compared these genes identified by cfDNA sequencing with the MSK-IMPACT panel genes (468 genes). MSK-IMPACT can be used to identify clinically relevant somatic mutations, novel non-coding changes, and mutational features shared between common and rare tumor types, which

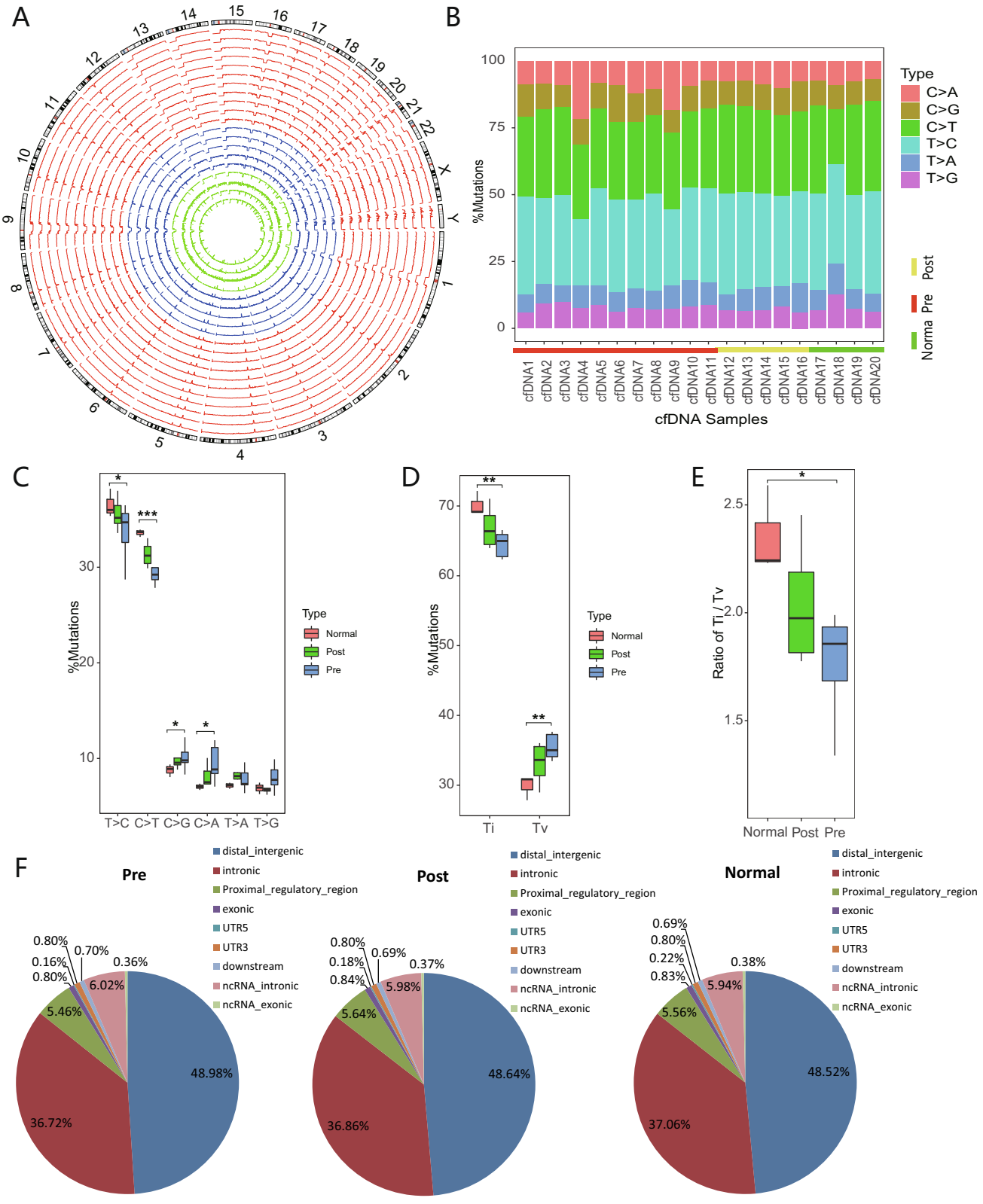

Fig. 6. Analysis of mutations. (A) Distribution of mutation density in each 1-Mb window of whole genome in different cfDNA samples. (B) Stacked bar plot shows distribution of mutation spectra for each cfDNA sample. (C) Box plot summarizing the SNV of different types of cfDNA samples. *P* values (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). (D and E) Box Plot created by dividing the SNV into Ti and Tv. Ti, transition; Tv, transversion. (F) Distribution of genomic location of the SNV in genome. Most of the SNVs were located in distal intergenic and intronic regions.

authorized by the Food and Drug Administration of USA in 2017 [42]. Finally, we found that 37 mutated genes uniquely existed in pre-operation patients (Fig. S8A; Table S6), suggesting that these genes might play a certain role in ESCA. These 37 genes contained many well-known cancer-related genes, such as PTEN, MYC, EZH1, IDH2, AKT2, and FGFR2.

We then performed functional enrichment analysis on these 37 genes. GO analysis revealed that these genes were significantly associated with cell death regulation, transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding, tissue development and so on (Fig. S8B). The KEGG pathway analysis demonstrated that these genes were significantly enriched in Pathways in cancer, Signaling pathways regulating pluripotency of stem cells, Central carbon metabolism in cancer, Transcriptional misregulation in cancer, and mTOR signaling pathway (Fig. S8C). The GAD DISEASE CLASS analysis demonstrated that these genes were significantly enriched in CANCER, REPRODUCTION, and DEVELOPMENTAL (Fig. S8D). Moreover, the GAD DISEASE analysis revealed that these genes were significantly associated with esophageal cancer (Fig. S8D). The UP KEYWORDS analysis of DAVID demonstrated that these genes were significantly enriched in Disease mutation, Transcription regulation, Tumor suppressor and Proto-oncogene (Fig. S8E). The UP SEQ FEATURE analysis of DAVID demonstrated that these genes were significantly enriched in mutagenesis site and sequence variant (Fig. S8E).

4. Discussion

In this study, we performed the NGS library construction and sequencing of cfDNA samples with SALP-seq that was developed by our laboratory [30]. As a new single-stranded DNA library preparation and sequencing technique, SALP-seq is particularly suited to construct the NGS libraries for highly degraded DNA samples such as cfDNA [30,31]. Moreover, by using the barcode T adaptors, this technique is competent to analyze many cfDNA samples in a high-throughput format [31]. In this study, we used the SALP-seq to analyze as many as 30 cfDNA samples successfully. Most of samples obtained over 80% mappable reads (Tables S1 and S2). Both ESCA-associated epigenetic and genetic biomarkers were successfully identified by analyzing the obtained sequencing data. The results demonstrated that this technique can be reliably applied to the future cfDNA NGS researches.

This study sheds important new insights on the clinical worth of cfDNA. In this study, we analyzed the cfDNA NGS reads data with machine learning algorithms. The results indicated that the main ESCA-associated epigenetic and genetic markers could be effectively identified by comparing the cancer and normal cfDNA samples. Especially, the cfDNA samples could be clearly classified by using the identified epigenetic markers (Figs. 2 and 4). Moreover, the promoter and genome-wide markers obtained highly consistent classification results (Figs. 2 and 4), indicating the reliability of these epigenetic markers in discriminating cfDNA samples. Importantly, the SVM model established with the epigenetic markers could be used to accurately distinguish cancer cfDNA samples from normal cfDNA samples with a high AUC value, even by using as few as 24 most important epigenetic markers at low cfDNA sequencing depth (1×10^7 reads/sample) (Fig. 5). These results reveal the important application of cfDNA in the systematic finding of cancer-associated markers, especially epigenetic markers associated with chromatin accessibility, at the genome-wide scale.

This study provides a new pipeline for finding new molecular markers for cancers from cfDNA by combining SALP-seq and machine learning. In recent years, as a good material for cancer liquid biopsy, plasma cfDNA has been widely analyzed by NGS for finding new molecular markers for cancer diagnosis such as fragment size [43,44], methylation [45–47], and end coordinate [48,49]. However, the size-based plasma DNA diagnostics still faced some limitations that may challenge its wide application [50]. The methylation detection has to do cell-free methylated DNA immunoprecipitation and high-throughput sequencing (cfMeDIP-seq) [45]. In comparison with these reported cfDNA-based methods, our method is much simpler and easier. Only two steps are needed. One is SALP-seq and the other is machine learning. Our method needs no pre-treatments to cfDNA such as size selection, target enrichment, chemical treatment (e.g. bisulfite conversion), and immunoprecipitation, which not only avoids the introduction of more artificial biases in cfDNA analysis, but also greatly simplifies the detection process. In the pipeline developed by this study, the cfDNA NGS reads data were analyzed with machine learning, this is different from the above mentioned previous studies. These studies did not employ machine learning to treat the NGS data. The results obtained by this study indicated that machine learning can play important roles in cfDNA NGS data analysis.

By using post-operation cfDNA samples, this study showed that the identified ESCA-associated epigenetic and genetic markers are tumor-associated. In other words, these markers should come from tumor but not from other tissues such as leukocytes, because these markers changed in response to surgical operation. It was found that most of these markers became disappeared after surgery (Figs. 2D and 4C), allowing the patient's cfDNA to be classified into or near normal cfDNA (Figs. 2D and 4C). If these markers come from other tissues such as leukocytes, they should have no such evident response to surgery. Therefore, this study revealed that these markers not only come from tumor, but also are beneficial for cancer prognosis. In other words, these markers can be used to evaluate and track the effects of cancer treatment noninvasively. In the later visiting, five patients provided the post-operation cfDNA samples still survive at present after the surgery in 2017, suggesting the good prognosis with these markers.

This study sheds important new insights on the potential regulatory and molecular mechanisms of tumorigenesis of ESCA. By analyzing and comparing the cfDNA NGS data, this study identified 49 ESCA-associated promoters and 88 ESCA-associated genome-wide regions. These ESCA-associated chromatin regions are all non-coding DNAs. Importantly, all these regions became more accessible in ESCA, suggesting that these regions play critical regulatory roles in tumorigenesis of ESCA. Especially, many of these ESCA-associated regions are distal intergenic (32, 36.36%) and proximal regulatory regions (22, 25%) (Fig. 3). Moreover, by comparing these regions with a comprehensive human super-enhancer database, the SEdb database, 17 of these 54 regions were well known super-enhancer elements. Additionally, it was found that the SVM model established with the 24 most important regions could still be used to accurately distinguish cancer samples from normal samples with a high AUC value. In these 24 regions, there are 14 distal intergenic regions (58.3%), 8 proximal regulatory regions (33.3%), and only 2 intronic regions (8.3%). By comparing with the chromatin accessibility level characterized by the H3K27ac ChIP-seq of TE7 ESCA cell lines and the ATAC-seq of ESCA tissues [38,39], it was found that the chromatin accessibility of these regions characterized by cfDNA is highly consistent with those characterized by other methods in ESCA cells and tissues (Fig. S5). Additionally, these regions contain a large number of TFBSs. Therefore, these non-coding ESCA-associated chromatin regions should play critical regulatory roles in tumorigenesis of ESCA. By assigning these regions to genes, 153 (49 plus 104) ESCA-associated genes were identified, in which 104 and 49 genes are connected with the genome-wide regions and promoters, respectively, and 16 are connected with both regions. It was found

that 15 of 49 genes and 21 of 104 genes have been reported to be closely associated with ESCA (Figs. 2B and 3). For example, studies have shown that WHSC1 has oncogenic activity and can cause protein lysine methyltransferases dysregulated in ESCA and other cancers [51]. Targeting WHSC1, specific inhibitors are currently developed for diagnosis and treatment of cancer, such as MCTP39 and LEM-06, which are in preclinical trials [51]. Furthermore, elevated expression of WHSC1 is often observed in many types of human cancers, and expression product of WHSC1 is essential for the growth of cancer cells [52,53]. The gene EIF5A2 was related to not only ESCA [54–56], but also breast cancer [57], lung cancer [58,59], bladder cancer [60,61], stomach cancer [62,63], oral cancer [64], liver cancer [65], and colorectal cancer [66]. By checking the expression of these genes detected in tumors (RNA-seq data in TCGA), most of these genes are significantly up-regulated in tumors (Figs. S3 and S6), in consistent with the increased chromatin accessibility of these regions revealed by cfDNA in this study. These results indicated that these ESCA-associated regions play regulatory roles in ESCA (Figs. S3A and S6) and other cancers (Fig. S3B). The gene annotation also revealed that these genes are also closely related to ESCA and other cancers (Figs. S4 and S7). Therefore, most of genes identified by this study are newly discovered genes associated with ESCA. For example, the newly discovered ESCA-associated gene, JAG2, plays a role in NOTCH signaling and Hedgehog signaling [67–69]. Dysregulated NOTCH signaling and Hedgehog signaling are both closely related to the development of cancer (e.g. ESCA) [68,70], further explaining the correlation between JAG2 and ESCA. The TCGA RNA-seq data revealed that JAG2 was significantly up-regulated in ESCA tissue (Fig. S6). Importantly, the JAG2-connected region identified in this study had the highest importance value of MDA (Fig. 3). Thus, targeting JAG2 may offer a promising therapeutic strategy for ESCA treatment. Other genes could also be potential targets for ESCA diagnosis and treatment.

Mutation of the cfDNA samples was analyzed in this study. The C > T and T > C transitions were the major SNVs (Fig. 6B). The C > T transitions may arise by replication of uracil generated by APOBEC cytidine deamination [71,72], while the cause of T > C currently has no clue. The C > G, C > A, and potentially additional C > T substitutions may be introduced by error-prone polymerases following uracil excision and generation of abasic sites by uracil-DNA glycosylase (UNG) [71,72]. Further analysis revealed that the differences between pre-operation ESCA samples and normal samples reached statistically significant level in the frequencies of Ti and Tv (Fig. 6D). Moreover, there was significant difference in the frequencies of Ti/Tv ratio between pre-operation ESCA samples and normal samples (Fig. 6E). These mutation characteristics of post-operation ESCA samples were closer to the normal samples, although they did not reach the same level (Fig. 6C–E). These results revealed that seven features, including the frequencies of C > T, T > C, C > G, C > A, Ti, Tv, and Ti/Tv ratio, could be developed as diagnostic markers for ESCA liquid biopsy. By analyzing mutations of cfDNA NGS data, this study finally identified 37 genetically altered ESCA-specific genes. The functional enrichment analysis showed that these genes have close relationships with the occurrence and development of cancer (Fig. S8).

In this study, to find the reliable differential characters between normal and cancer cfDNA samples, we used all mappable reads with a mean depth of 6 (Table S1). In the validation experiment, we sequenced 10 more cfDNA samples with a mean depth of 2.9 (Table S1). Clearly, the validation samples were sequenced with half depth of first 20 samples. At this seqeuncing depth, the validation samples were accurately identified (Figs. 2D and 3C). Moreover, the first 20 samples and 10 validation samples were sequences with various depth, ranging from 10.9 (cfDNA20) to 0.9 (cfDNA22) (Table S1), however, all samples were accurately differentiated (Figs. 2D and 3C), indicating that as low as mappable reads of a depth of 0.9 could be used to differentiate normal and cancer cfDNA samples. This is also in agreement with the subsequent result of the effect of the reads number on the classification model, which indicated that the model still maintained a good predictive effect at a sequencing depth as low as 1×10^7 reads (about a depth of 1.0) (Fig. 5D–E).

The study was performed with a relatively small sample size at a single institution. This study analyzed 20 cfDNA samples including 4 cfDNA samples from normal people, 5 cfDNA samples from post-operation cancer patients, and 11 cfDNA samples from pre-operation cancer patients, in which one normal cfDNA sample could not be used in the subsequent bioinformatics analysis due to too limited sequencing depth. In the verification study, only 10 new cfDNA samples from pre-operation cancer patients were used. Therefore, more normal and post-operation cfDNA samples should be included in the future study for the further validation of the current findings. It would be of value if future studies could be designed to address the clinical value of the detection of cfDNA biomarkers in larger sample cohorts. Additionally, only the cfDNA samples from one kind cancer, ESCA, were investigated. Therefore, this study only identified ESCA-associated markers, whether these markers are ESCA specific should be further investigated by analyzing cfDNA samples from various cancers. However, this more complex investigation can be effectively performed using the same pipeline, SALP-seq plus machine learning.

5. Conclusions

We have successfully analyzed many cfDNA samples from ESCA and normal participants by combining SALP-seq and machine learning, identifying both epigenetic and genetic biomarkers of ESCA. These biomarkers can be used to effectively classify cfDNAs from ESCA patients and normal participants. These biomarkers also shed important new insights on the potential regulatory and molecular mechanisms of tumorigenesis of ESCA. This study thus provides a new pipeline for finding new molecular markers for cancers from cfDNA by combining SALP-seq and machine learning. Finally, this study sheds important new insights on the clinical worth of cfDNA.

6. Ethics approval and consent to participate

All procedures used in this research were performed according to the Declaration of Helsinki. This study was approved by the Ethics Committee of Jinling Hospital (Nanjing, China). All participants were recruited from the Jinling Hospital, Nanjing University School of Medicine (Nanjing, China), with informed consent.

CRediT authorship contribution statement

Shicai Liu: Methodology, Software, Data curation, Validation, Visualization, Writing - original draft. **Jian Wu:** Investigation. **Qiang Xia:** Investigation. **Hongde Liu:** Supervision, Software, Writing - review & editing. **Weiwei Li:** Resources. **Xinyi Xia:** Resources, Supervision. **Jinke Wang:** Conceptualization, Supervision, Methodology, Project administration, Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61971122).

Appendix A. Supplementary data

Supplemental methods; Data availability; Table S1 Information of cfDNA samples and SALP-seq reads; Table S2 Oligonucleotides used as adaptors and PCR primers; Table S3 Known cancer-related genes in 49 genes; Table S4 Known cancer-related genes in target genes of 54 ESCA-related elements; Table S5 Mutation information; Table S6 Pre-operation unique genes of Supplementary Fig S8A; Fig. S1 The signal strength of reads distribution around TSSs; Fig. S2 The signal strength of reads distribution around TSSs for the verification cfDNA sample; Fig. S3 Analysis of expression of the selected 49 genes; Fig. S4 Annotation of 49 selected genes; Fig. S5 Snapshot of 54 ESCA-related important regions by using UCSC genome browser; Fig. S6 Expression of genes assigned to cancer-associated regions based on TCGA dataset of ESCA; Fig. S7 Annotation of 104 genes assigned to the ESCA-associated regions; Fig. S8 Gene annotation of 37 pre-operation sample unique genes. Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.06.042.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.
[2] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136(5):E359–86.
[3] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69 (1):7–34.
[4] Guibert N, Pradines A, Favre G, Mazieres J. Current and future applications of liquid biopsy in nonsmall cell lung cancer from early to advanced stages. Eur Respir Rev 2020;29(155):190052.
[5] Esagian SM, Grigoriadou G, Nikas IP, Boikou V, Sadow PM, Won JK, et al. Comparison of liquid-based to tissue-based biopsy analysis by targeted next generation sequencing in advanced non-small cell lung cancer: a comprehensive systematic review. J Cancer Res Clin Oncol 2020. https://doi.org/10.1007/s00432-020-03267-x.
[6] Buyuksimsek M, Togun M, Oguz KI, Bisgin A, Boga I, Tohumcuoglu M, et al. Results of liquid biopsy studies by next generation sequencing in patients with advanced stage non-small cell lung cancer: single center experience from turkey. Balkan J Med Genet 2019;22(2):17–24.
[7] de Wit S, Rossi E, Weber S, Tamminga M, Manicone M, Swennenhuis JF, et al. Single tube liquid biopsy for advanced non-small cell lung cancer. Int J Cancer 2019;144(12):3127–37.
[8] Ding PN, Becker TM, Bray VJ, Chua W, Ma YF, Lynch D, et al. The predictive and prognostic significance of liquid biopsy in advanced epidermal growth factor receptor-mutated non-small cell lung cancer: a prospective study. Lung Cancer 2019;134:187–93.
[9] Doval DC, Deshpande R, Dhabhar B, Babu KG, Prabhash K, Chopra R, et al. Liquid biopsy: A potential and promising diagnostic tool for advanced stage non-small cell lung cancer patients. Indian J Cancer 2017;54(Supplement): S25–30.
[10] Esposito Abate R, Pasquale R, Sacco A, Piccirillo MC, Morabito A, Bidoli P, et al. Liquid biopsy testing can improve selection of advanced non-small-cell lung cancer patients to rechallenge with gefitinib. Cancers (Basel) 2019;11 (10):1431.
[11] Veldore VH, Choughule A, Routhu T, Mandloi N, Noronha V, Joshi A, et al. Validation of liquid biopsy: plasma cell-free DNA testing in clinical management of advanced non-small cell lung cancer. Lung Cancer (Auckl) 2018;9:1–11.
[12] Oxnard GR, Thress KS, Alden RS, Lawrance R, Paweletz CP, Cantarini M, et al. Association between plasma genotyping and outcomes of treatment with osimertinib (AZD9291) in advanced non-small-cell lung cancer. J Clin Oncol 2016;34(28):3375–82.
[13] Kwapisz D. The first liquid biopsy test approved. Is it a new era of mutation testing for non-small cell lung cancer? Ann Transl Med 2017;5 (3):46.
[14] Mandel P, Metais P. Les acides nucléiques du plasma sanguin chez l'Homme. C R Seances Soc Biol Fil 1948;142(3–4):241–3.
[15] Celec P, Vlkova B, Laukova L, Babickova J, Boor P. Cell-free DNA: the role in pathophysiology and as a biomarker in kidney diseases. Expert Rev Mol Med 2018;20:e1.
[16] Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell 2016;164(1–2):57–68.
[17] Ferres MA, Hui L, Bianchi DW. Antenatal noninvasive DNA testing: clinical experience and impact. Am J Perinatol 2014;31(7):577–82.
[18] Sohda M, Kuwano H. Current status and future prospects for esophageal cancer treatment. Ann Thorac Cardiovasc Surg 2017;23(1):1–11.
[19] Allyse M, Minear MA, Berson E, Sridhar S, Rote M, Hung A, et al. Non-invasive prenatal testing: a review of international implementation and challenges. Int J Womens Health 2015;7:113–26.
[20] Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci USA 2015;112 (40):E5503–12.
[21] Chan KC, Jiang P, Zheng YW, Liao GJ, Sun H, Wong J, et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clin Chem 2013;59(1):211–24.
[22] Chan KCA, Woo JKS, King A, Zee BCY, Lam WKJ, Chan SL, et al. Analysis of plasma epstein-barr virus DNA to screen for nasopharyngeal cancer. N Engl J Med 2017;377(6):513–22.
[23] Sun K, Jiang P, Chan KC. The impact of digital DNA counting technologies on noninvasive prenatal testing. Expert Rev Mol Diagn 2015;15(10):1261–8.
[24] Crowley E, Di Nicolantonio F, Loupakis F, Bardelli A. Liquid biopsy: monitoring cancer-genetics in the blood. Nat Rev Clin Oncol 2013;10(8):472–84.
[25] Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2010;2 (61):61ra91.
[26] Benesova L, Belsanova B, Suchanek S, Kopeckova M, Minarikova P, Lipska L, et al. Mutation-based detection and monitoring of cell-free tumor DNA in peripheral blood of cancer patients. Anal Biochem 2013;433(2):227–34.
[27] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68 (1):7–30.
[28] Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, et al. Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin 2016;66 (4):271–89.
[29] Britton E, Rogerson C, Mehta S, Li Y, Li X, consortium O, et al. Open chromatin profiling identifies AP1 as a transcriptional regulator in oesophageal adenocarcinoma. PLoS Genet 2017;13 (8):e1006879.
[30] Wu J, Dai W, Wu L, Wang J. SALP, a new single-stranded DNA library preparation method especially useful for the high-throughput characterization of chromatin openness states. BMC Genomics 2018;19(1):143.
[31] Wu J, Dai W, Wu L, Li W, Xia X, Wang J. Decoding genetic and epigenetic information embedded in cell free DNA with adapted SALP-seq. Int J Cancer 2019;145:2395–406.
[32] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.
[33] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25 (16):2078–9.
[34] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38 (16):e164.
[35] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4 (1):44–57.
[36] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.
[37] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
[38] Jiang YY, Lin DC, Mayakonda A, Hazawa M, Ding LW, Chien WW, et al. Targeting super-enhancer-associated oncogenes in oesophageal squamous cell carcinoma. Gut 2017;66(8):1358–68.
[39] Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science 2018;362 (6413):eaav1898.
[40] Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet 2016;48 (10):1193–203.
[41] Chang C-C, Lin C-J. Libsvm. ACM Trans Intell Syst Technol 2011;2(3):1–27.
[42] Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 2017;23(6):703–13.
[43] Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med 2018;10 (466): eaat4921.
[44] Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, et al. Fragment length of circulating tumor DNA. PLoS Genet 2016;12(7):e1006162.
[45] Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 2018;563(7732):579–83.
[46] Li WY, Li QJ, Kang SL, Same M, Zhou YG, Sun C, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res 2018;46(15):e89.

[47] Sina AA, Carrascosa LG, Liang Z, Grewal YS, Wardiana A, Shiddiky MJA, et al. Epigenetically reprogrammed methylation landscape drives the DNA self-assembly and serves as a universal cancer biomarker. Nat Commun 2018;9(1):4915.
[48] Sun K, Jiang P, Wong AIC, Cheng YKY, Cheng SH, Zhang H, et al. Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. Proc Natl Acad Sci USA 2018;115(22):E5106–14.
[49] Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci USA 2018;115(46):E10925–33.
[50] Heitzer E, Speicher MR. One size does not fit all: Size-based plasma DNA diagnostics. Science Translational Medicine 2018;10 (466): eaav3873.
[51] Hamamoto R, Nakamura Y. Dysregulation of protein methyltransferases in human cancer: An emerging target class for anticancer therapy. Cancer Sci 2016;107(4):377–84.
[52] Vougiouklakis T, Hamamoto R, Nakamura Y, Saloura V. The NSD family of protein methyltransferases in human cancer. Epigenomics 2015;7(5):863–74.
[53] Saloura V, Cho HS, Kiyotani K, Alachkar H, Zuo Z, Nakakido M, et al. WHSC1 promotes oncogenesis through regulation of NIMA-related kinase-7 in squamous cell carcinoma of the head and neck. Mol Cancer Res 2015;13(2):293–304.
[54] Cai X, Yang X, Jin C, Li L, Cui Q, Guo Y, et al. Identification and verification of differentially expressed microRNAs and their target genes for the diagnosis of esophageal cancer. Oncol Lett 2018;16(3):3642–50.
[55] Yang H, Li XD, Zhou Y, Ban X, Zeng TT, Li L, et al. Stemness and chemotherapeutic drug resistance induced by EIF5A2 overexpression in esophageal squamous cell carcinoma. Oncotarget 2015;6(28):26079–89.
[56] Li Y, Fu L, Li JB, Qin Y, Zeng TT, Zhou J, et al. Increased expression of EIF5A2, via hypoxia or gene amplification, contributes to metastasis and angiogenesis of esophageal squamous cell carcinoma. Gastroenterology 2014;146 (7):1701-13 e9.
[57] Furuta S, Ren G, Mao JH, Bissell MJ. Laminin signals initiate the reciprocal loop that informs breast-specific gene expression and homeostasis by activating NO, p53 and microRNAs. Elife 2018;7:e26148.
[58] Pan Q, Sun L, Zheng D, Li N, Shi H, Song J, et al. MicroRNA-9 enhanced cisplatin sensitivity in nonsmall cell lung cancer cells by regulating eukaryotic translation initiation factor 5A2. Biomed Res Int 2018;2018:1769040.
[59] Chen C, Zhang B, Wu S, Song Y, Li J. Knockdown of EIF5A2 inhibits the malignant potential of non-small cell lung cancer cells. Oncol Lett 2018;15(4):4541–9.
[60] Chen Z, Yu T, Zhou B, Wei J, Fang Y, Lu J, et al. Mg(II)-Catechin nanoparticles delivering siRNA targeting EIF5A2 inhibit bladder cancer cell growth in vitro and in vivo. Biomaterials 2016;81:125–34.
[61] Huang Y, Wei J, Fang Y, Chen Z, Cen J, Feng Z, et al. Prognostic value of AIB1 and EIF5A2 in intravesical recurrence after surgery for upper tract urothelial carcinoma. Cancer Manag Res 2018;10:6997–7011.
[62] Zhou X, Xu M, Guo Y, Ye L, Long L, Wang H, et al. MicroRNA-588 regulates invasion, migration and epithelial-mesenchymal transition via targeting EIF5A2 pathway in gastric cancer. Cancer Manage Res 2018;10:5187–97.
[63] Wang X, Jin Y, Zhang H, Huang X, Zhang Y, Zhu J. MicroRNA-599 inhibits metastasis and epithelial-mesenchymal transition via targeting EIF5A2 in gastric cancer. Biomed Pharmacother 2018;97:473–80.
[64] Fang L, Gao L, Xie L, Xiao G. GC7 enhances cisplatin sensitivity via STAT3 signaling pathway inhibition and eIF5A2 inactivation in mesenchymal phenotype oral cancer cells. Oncol Rep 2018;39(3):1283–91.
[65] Bai HY, Liao YJ, Cai MY, Ma NF, Zhang Q, Chen JW, et al. Eukaryotic initiation factor 5A2 contributes to the maintenance of CD133(+) hepatocellular carcinoma cells via the c-Myc/microRNA-29b Axis. Stem Cells 2018;36(2):180–91.
[66] Deng B, Wang B, Fang J, Zhu X, Cao Z, Lin Q, et al. MiRNA-203 suppresses cell proliferation, migration and invasion in colorectal cancer via targeting of EIF5A2. Sci Rep 2016;6:28301.
[67] Luo B, Aster JC, Hasserjian RP, Kuo F, Sklar J. Isolation and functional analysis of a cDNA for human Jagged2, a gene encoding a ligand for the Notch1 receptor. Mol Cell Biol 1997;17(10):6057–67.
[68] Katoh Y, Katoh M. Hedgehog target genes: mechanisms of carcinogenesis induced by aberrant hedgehog signaling activation. Curr Mol Med 2009;9(7):873–86.
[69] Katoh Y, Katoh M. Integrative genomic analyses on GLI1: positive regulation of GLI1 by Hedgehog-GLI, TGFbeta-Smads, and RTK-PI3K-AKT signals, and negative regulation of GLI1 by Notch-CSL-HES/HEY, and GPCR-Gs-PKA signals. Int J Oncol 2009;35(1):187–92.
[70] Ciccarelli FD. Mutations differ in normal and cancer cells of the oesophagus. Nature 2019;565(7739):301–3.
[71] Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 2014;15(9):585–98.
[72] Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer 2014;14(12):786–800.