GeneLLM: A Large cfRNA Language Model for Cancer Screening from Raw Reads

Siwei Deng[1,2,3†], Lei Sha[2,3*†], Yongcheng Jin[3*†], Tianyao Zhou[3†], Chengen Wang[5†], Qianpu Liu[2], Hongjie Guo[7], Chengjie Xiong[4], Yangtao Xue[4], Xiaoguang Li[5], Yuanming Li[5], Yaping Gao[3], Mengyu Hong[6], Junjie Xu[4*], Shanwen Chen[1*], Pengyuan Wang[1*]

[1]Division of Gastrointestinal Surgery, Peking University First Hospital, Peking University, 8 Xishiku Street, Xicheng District, Beijing, 100034, China.

[2]Institute of Artificial Intelligence, Beihang University, No.37 Xueyuan Road, Haidian, Beijing, 100191, Beijing, China.

[3]OxTium Technology Co., Ltd, No.1 Song Ping Shan Xin Dong Road, Shenzhen, 518052, Guangdong, China.

[4]Department of General Surgery, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, 310016, China.

[5]Department of Minimally Invasive Tumor Therapies Center, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, No.1 Da Hua Road, Dong Dan, Beijing, 100730, China.

[6]Academy for Advanced Interdisciplinary Science and Technology, Key Laboratory of Advanced Materials and Devices for Post-Moore Chips Ministry of Education, State Key Laboratory for Advanced Metals and Materials, University of Science and Technology Beijing, Beijing, 100083, China.

[7]Department of Interventional Radiology and Vascular Surgery, Peking University First Hospital, Peking University, 8 Xishiku Street, Xicheng District, Beijing, 100034, China.

*Corresponding author(s). E-mail(s): shalei@buaa.edu.cn; yongcheng.jin@oxtium.com; walter235@zju.edu.cn; shanwen@pku.edu.cn; pengyuan_wang@bjmu.edu.cn;
†These authors contributed equally to this work.

Abstract

Plasma cell-free RNA (cfRNA) has recently emerged as a promising biomarker for non-invasive early cancer detection and treatment monitoring [8, 28, 31, 32, 38, 40]. Here, we introduce GeneLLM, a novel large language model designed to interpret cfRNA sequences directly, bypassing the need for genome annotations. GeneLLM significantly advances the detection accuracy of various cancer types. Our study demonstrates that this method achieves higher accuracy than traditional biomarkers and effectively handles large datasets from different centres, even with low sequencing depth. By avoiding the use of bioinformatics tools to count known genes, GeneLLM also discovered cfRNAs from previously unknown genes, referred to as 'dark matters' in the genome, as cancer detection 'pseudo-biomarkers'. Our results showcase the potential of GeneLLM to revolutionise cancer detection, making it more accessible and cost-effective. By offering a method that does not depend on bioinformatics tools to count known genes, GeneLLM opens new avenues for biomarker discovery and enhances our understanding of intercellular communication through novel RNA molecules.

Keywords: Cancer Detection, cfRNA, Artificial Intelligence, Large Language Model

1 Main

Cancer remains a significant challenge to global health, necessitating innovative approaches in early detection to enhance treatments and patient outcomes [9, 36]. Recent advances have concentrated on non-invasive, cost-effective techniques. Notably, research has highlighted the promise of cell-free RNA (cfRNA) in detecting various diseases. As RNA reflects gene expression profiles, it indicates how the body responds to diseases [8, 28, 31, 32, 38, 40]. This characteristic makes cfRNA more predictive in identifying solid tumours compared to cell-free DNA (cfDNA) [2], addressing the drawbacks of traditional methods like colonoscopy and mammography, which, although effective, are often invasive, expensive, and labour-intensive [3, 5].

Nonetheless, the field faces challenges in sensitivity and specificity due to variability in cfRNA levels and the difficulty in distinguishing cancerous cfRNA profiles from normal ones. To tackle this, studies have integrated machine learning methods, typically performing Differentially Expression (DE) analysis to discover Differentially Expressed Genes (DEGs) and identify potential biomarkers using methods such as random forest [29]. However, these methods are still reliant on known genes.

Intriguingly, research has uncovered sequences from cfRNA that do not usually reside in gene regions, such as repetitive RNA regions, which show potential as biomarkers [31]. Other RNA types, like transcribed ultra-conserved regions, also exhibit promise in cancer detection [4]. These RNAs, often termed the genome's 'dark matters', are gaining recognition in cancer detection [26]. Despite their potential, these RNAs are not annotated and thus cannot be included in traditional bioinformatics pipelines dependent on genome annotations. Consequently, the valuable insights from RNAs transcribed from the genome's 'dark matters' are frequently overlooked.

Addressing these challenges, our study introduces an innovative transformer-based approach named Gene Large Language Model (GeneLLM). By harnessing the power of large language models, this method delves into the genome's 'dark matters' by processing raw cfRNA sequencing data to identify 'pseudo-biomarkers' independently, without relying on genome annotations. Our study demonstrates the efficacy of this approach in classifying cfRNA samples from multiple cancer types across multiple centres. GeneLLM exhibits high efficiency in handling large datasets, maintaining consistent results with data sizes ranging from 6 GB to 1 GB. This reduces sequencing and computational costs. Additionally, GeneLLM successfully identified cfRNA molecules from unknown genes (RNA molecules transcribed from the 'dark genome'), which are undetectable using conventional bioinformatics tools.

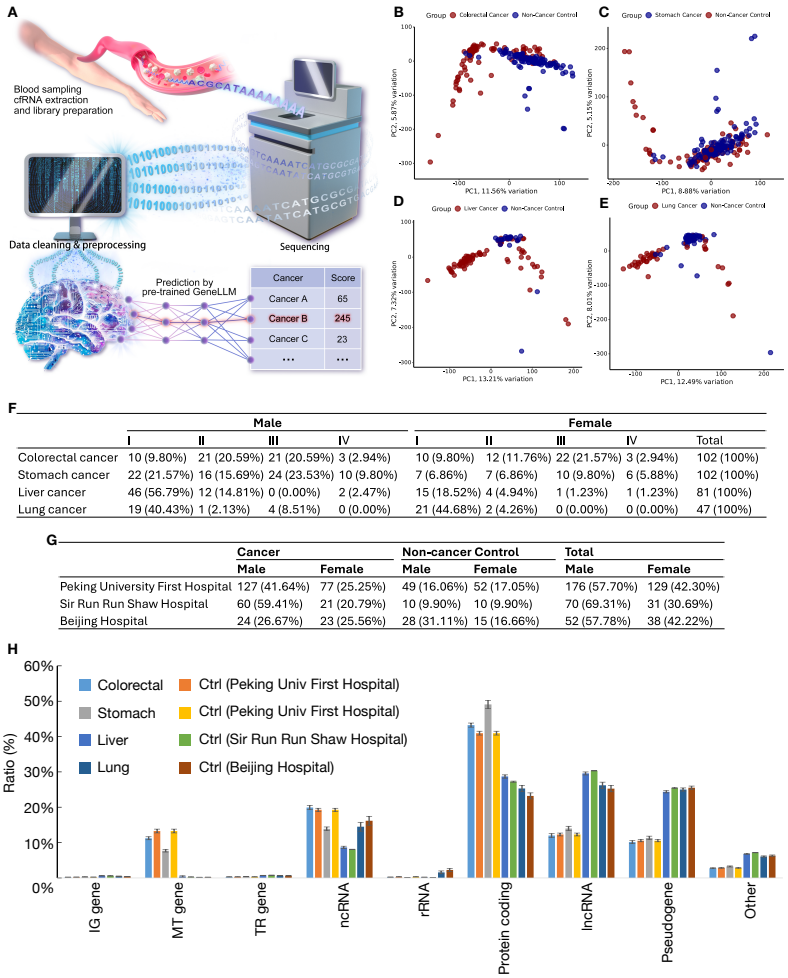
Our findings have significant implications, including improved diagnostic accuracy, reduced operational costs, and the potential for earlier cancer diagnosis, leading to

better patient outcomes. The versatile applicability of our methodology across different cancer types highlights its diagnostic significance and suggests it could revolutionise cancer screening by unlocking the potential of the 'dark genome', making it more accessible and affordable.

Additionally, this approach could reveal previously unknown genomic sequences transcribed into RNA and released from cells, thereby enhancing biomarker discovery and advancing the understanding of intercellular communication via novel RNA molecules. Finally, our framework can be applied to any sequencing-based datasets to explore 'pseudo-biomarkers,' such as tracellular bulk RNA-seq and single-cell RNA-seq, paving the way for the identification of new therapeutic targets.

2 Statistics of multi-centre cfRNA samples across multiple cancer types

CfRNA sequencing and pre-trained GeneLLM performance assessment were conducted on four types of cancer samples and their corresponding non-cancer controls, collected from patients at three centres: Sir Run-Run Shaw Hospital, Beijing Hospital, and Peking University First Hospital (Figure 1A–E). The statistics are summarised in Figure 1F & G and detailed in Table S1. Most samples were from early-stage cancers (stages I and II; Figure 1F). As we depleted ribosomal RNA (rRNA) prior to sequencing, the rRNA content was low in our samples. The cfRNA molecules in our data predominantly originated from protein-coding genes and non-coding RNA (ncRNA) (including long non-coding RNA (lncRNA) but excluding rRNA; Figure 1H).



F

	Male				Female				
	I	II	III	IV	I	II	III	IV	Total
Colorectal cancer	10 (9.80%)	21 (20.59%)	21 (20.59%)	3 (2.94%)	10 (9.80%)	12 (11.76%)	22 (21.57%)	3 (2.94%)	102 (100%)
Stomach cancer	22 (21.57%)	16 (15.69%)	24 (23.53%)	10 (9.80%)	7 (6.86%)	7 (6.86%)	10 (9.80%)	6 (5.88%)	102 (100%)
Liver cancer	46 (56.79%)	12 (14.81%)	0 (0.00%)	2 (2.47%)	15 (18.52%)	4 (4.94%)	1 (1.23%)	1 (1.23%)	81 (100%)
Lung cancer	19 (40.43%)	1 (2.13%)	4 (8.51%)	0 (0.00%)	21 (44.68%)	2 (4.26%)	0 (0.00%)	0 (0.00%)	47 (100%)

G

	Cancer		Non-cancer Control		Total	
	Male	Female	Male	Female	Male	Female
Peking University First Hospital	127 (41.64%)	77 (25.25%)	49 (16.06%)	52 (17.05%)	176 (57.70%)	129 (42.30%)
Sir Run Run Shaw Hospital	60 (59.41%)	21 (20.79%)	10 (9.90%)	10 (9.90%)	70 (69.31%)	31 (30.69%)
Beijing Hospital	24 (26.67%)	23 (25.56%)	28 (31.11%)	15 (16.66%)	52 (57.78%)	38 (42.22%)

Figure 1. Multi-centre, multi-cancer study design.
(A) Study workflow: Blood is drawn from the median cubital vein, followed by plasma centrifugation to extract cfRNA. A cfRNA library is prepared and sequenced using the Illumina NovaSeq 6000 platform. The sequencing reads are cleaned and analysed using a pre-trained large language model (GeneLLM) to determine the likelihood of various types of cancer.
(B-E) Principal Component Analysis (PCA) plots showing cancers (dark red dots) and non-cancer controls (dark blue dots). Only non-cancer controls from the same centre are shown to reduce batch effects.
(F) Statistics of cancer patients by gender and stage.
(G) Statistics of cancer patients and non-cancer controls by gender and source.
(H) RNA type statistics for different cancer samples and their matching non-cancer controls. IG gene = immunoglobulin gene; MT gene = mitochondrial gene; TR gene = T cell receptor gene; ncRNA = non-coding RNA; rRNA = ribosomal RNA; Protein coding = RNA that encodes protein-coding genes; lncRNA = long non-coding RNA; Pseudogene = RNA that encodes pseudogenes.

3 Model design and performance evaluation

3.1 GeneLLM's architecture

GeneLLM significantly enhances the accuracy of early cancer screening by leveraging the sequential data understanding capabilities of large language models. We constructed a Transformer architecture and pretrained this model using unlabelled cfRNA reads, enabling a deep understanding of the connections within cfRNA reads. Given that each patient has approximately 40 million cfRNA reads, we followed the pretraining phase by incorporating labelled data to fine-tune GeneLLM's cancer detection module. This module, acting as a complex language model head, extracts valuable information from the vast amount of cfRNA reads to provide final cancer screening results.

GeneLLM's main architecture processes each patient's extensive cfRNA reads, compares them, and identifies minor anomalies indicating potential disease (Figure 2A; see section 6 for details on the architecture, cfRNA sequence tokenisation, and the cancer detection module). The critical concepts and components are described below.

GeneLLM comprises three main stages. First, we pretrained the Transformer architecture using cfRNA reads (Figure 2A). Since disease labels are not required at this stage, we mixed and shuffled each patient's cfRNA reads, then tokenise them into 7-mer token sequences. The pretraining process follows the same methodology as that of language models [30] (see section 6 for pretraining details). Second, once GeneLLM is pretrained, each cfRNA sequence is transformed into a vector by the output hidden layer of GeneLLM's last output. We then combined all patients' cfRNA vectors to mine pseudo-biomarkers, which are the most representative cfRNA sequences. Third, we constructed a disease judgement head (Figure 2B) and conducted disease tuning based on the pseudo-biomarkers. The details of these stages are provided in the section 6.

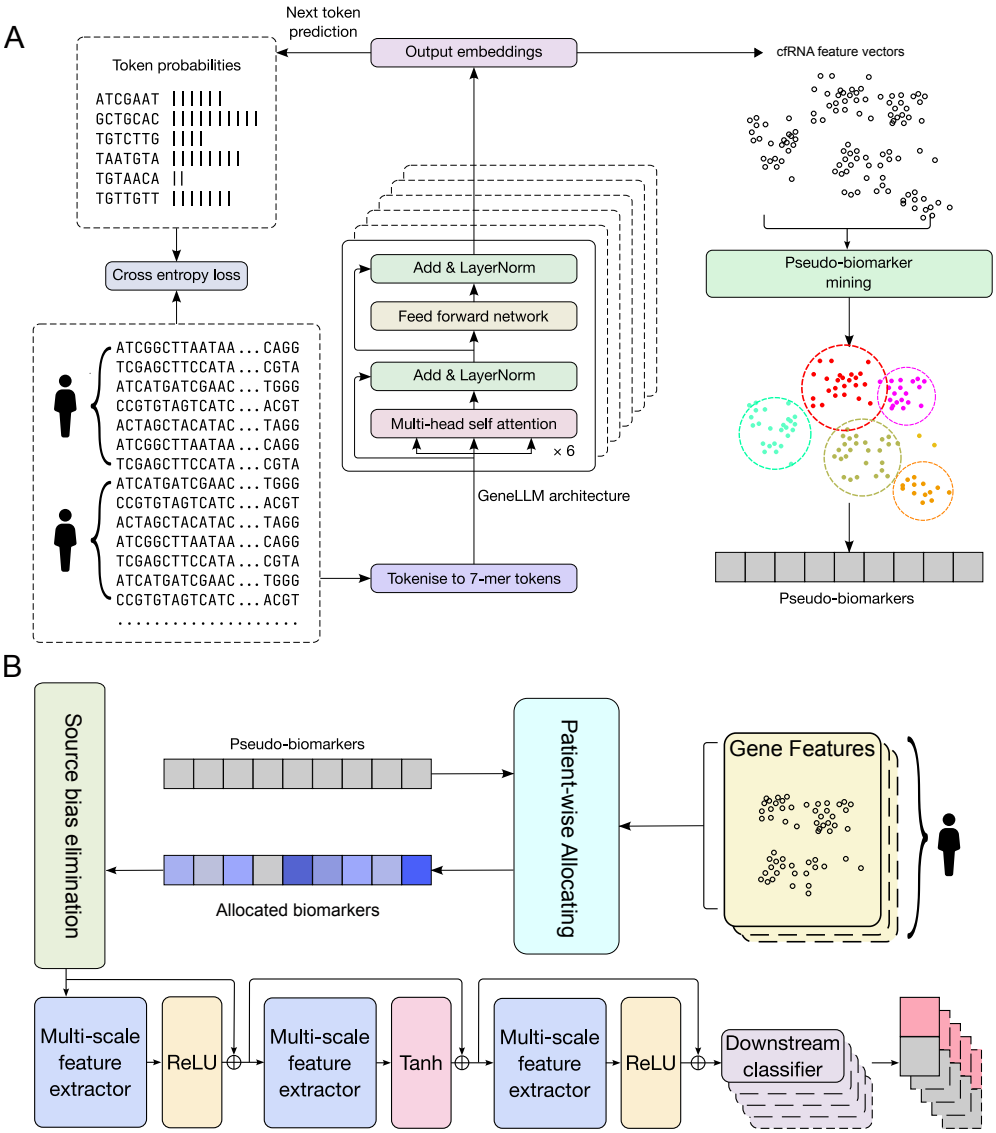


Figure 2. The architecture of GeneLLM.
(A) The pre-training and pseudo-biomarker mining phases for GeneLLM. The cfRNA reads are tokenised into 7-mer token sequences, which serve as input for pretraining the language model (LLM). Following the pretraining phase, each cfRNA read is transformed into a feature vector representation (upper right) that captures its essential characteristics. These feature vectors are then utilized for mining pseudo-biomarkers, enabling the identification of potential disease-specific signatures.
(B) The disease tuning phase for GeneLLM. Pseudo-biomarkers are used to allocate feature vectors from the given patient. Stacks of multi-scale feature extractors are employed to uncover deep, hidden information within the gene features. Finally, multiple binary classifiers, one for each disease, are utilised for decision-making. To eliminate potential bias arising from the patient's source (hospital), a source bias elimination step is conducted prior to deep information extraction.

3.2 GeneLLM's performance

We collected a comprehensive dataset consisting of 496 cfRNA cases, which included both cancer patients and non-cancer individuals. To ensure a rigorous evaluation of our proposed method, we split the dataset into three distinct subsets: training, validation, and test sets, with a ratio of 5 : 1 : 4, respectively. The GeneLLM is pretrained exclusively on the training set, and the reported results are obtained by evaluating the model's performance on the independent test set, ensuring an unbiased assessment of its generalisation ability. This stratified approach allows for a fair assessment of the model's performance and generalisation ability.

The final results of our analysis are presented in Figure 3A and B, which showcase the Receiver Operating Characteristic (ROC) plot and confusion matrix, respectively. The ROC plot demonstrates the exceptional performance of our method, with the Area Under the Curve (AUC) value for each cancer type and pan-cancer consistently above 0.9. This high AUC value indicates the model's strong ability to distinguish between cancer patients and non-cancer individuals. Furthermore, the confusion matrix reveals that the probabilities of correctly identifying each cancer type are all over 78%, emphasising the model's accuracy in classifying specific cancer types.

To further validate the effectiveness of our approach, we examined the probabilities of each case being identified as a specific cancer, as illustrated in Figures 3C–G. These figures provide a detailed view of the model's performance at the individual case level. This finding reinforces the reliability and precision of our method in correctly identifying specific diseases based on the cfRNA profiles.

These findings have significant implications for our proposed method's potential application in the early screening of cancer. The high accuracy and robustness of our approach suggest that it can be directly integrated into clinical practice, aiding in the early detection of various cancer types. By identifying cancer at an early stage, our method can contribute to improved patient outcomes and increased survival rates.

3.3 GeneLLM accurately classifies cancers with significantly lower sequencing depth

In the last experiment, we examined the standard data output for cfRNA-seq derived from plasma samples, which typically yield 6 GB of data, equivalent to approximately 40 million paired-end reads. Our primary objective was to assess the robustness and versatility of GeneLLM in handling cfRNA-seq data with significantly reduced sequencing depth, thereby mimicking real-world conditions encountered in shallow sequencing scenarios.

To thoroughly evaluate GeneLLM's performance under these constrained conditions, we employed a strategic approach. We re-sequenced the same library to generate a dataset of only 1 GB, representing a substantial reduction in sequencing depth compared to the standard 6 GB output. This step allowed us to simulate the practical limitations and challenges faced in real-world shallow sequencing experiments.

With the reduced sequencing depth dataset in hand, we proceeded to pretrain GeneLLM from scratch using the 1 GB data. After the pretraining phase, we applied the same prediction workflow as in the previous analysis, where GeneLLM was fine-tuned and tested on the standard 1 GB dataset. This consistent approach allowed for a direct comparison of the model's prediction accuracy under different sequencing depth conditions.

Remarkably, despite the substantial reduction in sequencing depth from 6 GB to 1 GB, GeneLLM demonstrated a remarkable ability to maintain its prediction accuracy. The results, as depicted in Figure 3H & I, clearly illustrate that the model's performance remained comparable to that observed with the higher sequencing depth data.

These findings hold significant implications for the applicability and feasibility of GeneLLM in real-world scenarios where shallow sequencing is employed due to cost or resource constraints. The robustness of GeneLLM in handling reduced sequencing

depth data highlights its potential to deliver accurate predictions even in challenging conditions.

Moreover, the ability of GeneLLM to maintain its performance with limited sequencing information suggests that it effectively captures and leverages the most informative features present in the cfRNA-seq data. This efficient utilisation of available data is particularly valuable in scenarios where obtaining high-depth sequencing is not feasible or practical.

3.4 GeneLLM improves the performance compared to existing datasets

To evaluate the performance of GeneLLM, we used external cfRNA datasets from the public domain with reported AUC and compared the results with the original studies. These include Yu *et al.* 2020 (NCBI BioProject PRJNA552230, comprising 284 plasma cfRNA samples in extracellular vesicles (EVs) with pancreatic ductal adenocarcinoma (PDAC), 100 patients with chronic pancreatitis, and 117 healthy controls) [40], and Chen *et al.* 2022 (NCBI BioProject PRJNA729258, containing plasma cfRNA samples from 54 colorectal cancer, 37 stomach cancer, 27 liver cancer, 35 lung cancer, 31 esophageal cancer, and 46 healthy controls) [8]. We preprocessed the raw `fastq` data using the same approach and re-trained the model using cleaned reads. The results are presented in Figure 3J & K. Our GeneLLM method substantially outperforms the previous method by Chen *et al.* [8] in pan-cancer detection by a large margin. Additionally, our approach significantly surpasses the performance of Yu *et al.* 2020 [40] in detecting pancreatic cancer.

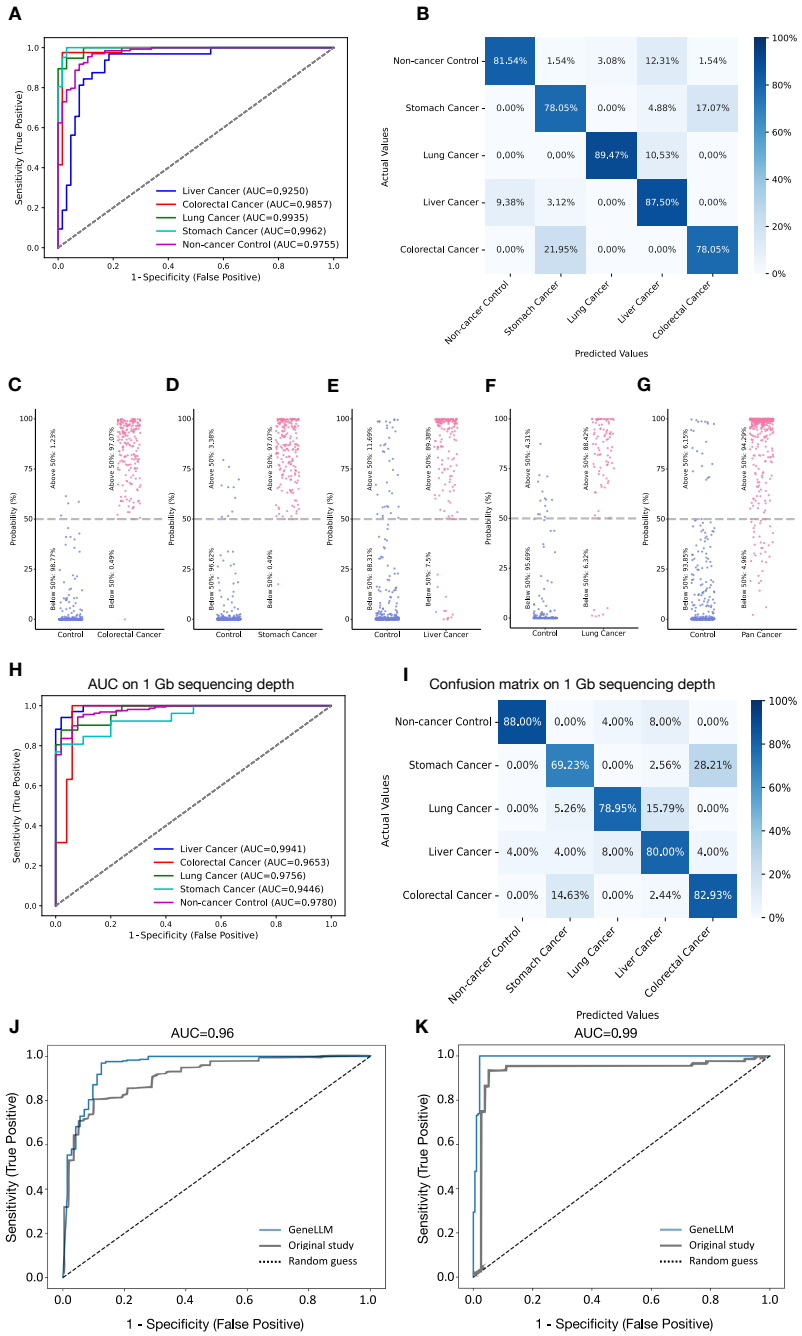


Figure 3. The performance of GeneLLM.
(A) The ROC plot for four cancers and pan-cancer, illustrating performance using AUC.
(B) The confusion matrix for four cancers and non-cancer controls.
(C–G) Probability distributions calculated by GeneLLM for each sample of (C) colorectal cancer, (D) stomach cancer, (E) liver cancer, (F) lung cancer, and (G) pan-cancers with their corresponding non-cancer controls. Each dot represents a sample.
(H) The ROC plot for four cancers and pan-cancers using 1 GB of sequencing depth.
(I) The confusion matrix for 1 GB of sequencing depth.
(J) Comparison of performance using AUC of GeneLLM and other studies using the same pan-cancer datasets [8].
(K) Comparison of performance using AUC of GeneLLM and other studies using the same PDAC datasets [40].

4 Biological characteristics of cfRNA signatures and GeneLLM-derived markers

4.1 Biological processes associated with DEGs

We performed DE analysis followed by Gene Set Enrichment Analysis (GSEA) to understand the biological changes in cfRNA between cancer and non-cancer controls. In colorectal and stomach cancers, most DEGs are downregulated, while in liver and lung cancers, most are upregulated. Specifically, the number of DEGs are 5,223 (1,043 up and 4,180 down) in colorectal cancer, 1,524 (94 up and 1,430 down) in stomach cancer, 2,137 (1,462 up and 675 down) in liver cancer, and 4,698 (2,942 up and 1,756 down) in lung cancer. The detailed DEG lists are provided in Supplementary Tables S2, S3, S4, and S5 for colorectal, stomach, liver, and lung cancers, respectively. Similarly, GSEA revealed that the enriched pathways in lung and liver cancers are similar and mostly upregulated, while those in colorectal and stomach cancers are similar and mostly downregulated (Figure 4A). Detailed GSEA results are provided in Supplementary Tables S6, S7, S8, and S9.

While GSEA offers a broad view of gene set enrichment, Pathway RespOnsive GENes for activity inference from gene expression (PROGENy) focuses on a refined set of pathway-responsive genes for a more sensitive and specific measure of pathway activities. It utilises a vast collection of perturbation experiments to identify common pathway-responsive genes [34]. From the pathway activity inference results, we observed similar patterns to those seen in GSEA, with colorectal and stomach cancers showing a resemblance, and liver and lung cancers displaying a similar profile (Figure 4B–F).

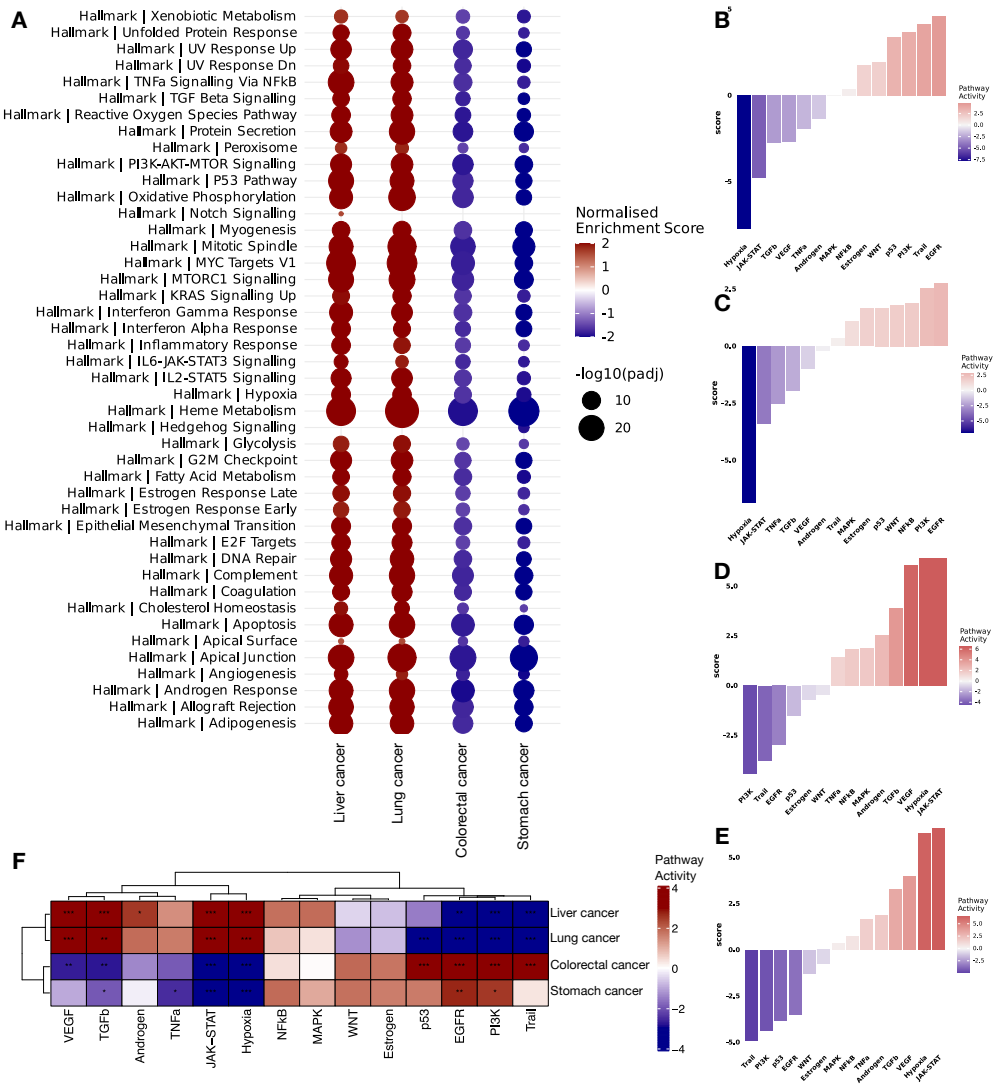


Figure 4. The pathway signatures of cfRNA in cancers
(A) Summary of GSEA significant results against Molecular Signatures Database (MSigDB) hallmark collection v2023.1.
(B–E) PROGENy pathway activity inference results for colorectal cancer, stomach cancer, liver cancer, and lung cancer.
(F) Summary of PROGENy pathway activity inference. (* = p-value < 0.05; ** = p-value < 0.01; *** = p-value < 0.001)

4.2 GeneLLM discovered 'dark matter' in the genome that aids in cancer identification

To understand which sequences GeneLLM considers most indicative of the features distinguishing cancers from non-cancer controls, we extracted the top three pseudo-biomarkers for each cancer type. We then performed Nucleotide Basic Local Alignment Search Tool (BLASTn) against the nucleotide collection (nt) database, using the database last updated on 23rd June 2024. Intriguingly, most of these sequences are not annotated as any known genes (Table 1), even though they map to the human genome with high coverage and identity. Therefore, these sequences cannot be recovered by the bioinformatics workflow as they cannot be counted in the expression matrix. This suggests that GeneLLM utilises novel sequences (known as 'dark matters' in the genome) that aid in cancer identification.

Table 1. Top three pseudo-biomarkers for each cancer types and their corresponding BLASTn top hits.

Cancer	Sequence	Top BLASTn hit	Query coverage	Percent identity
Colorectal	CGGGCTGCAACCAAGAAATTGGAAGGCTTCAGCACATGAAAGGTGAACTTTCCAAAAGCCTTGGCTTCTTTCTCAA CCGCGCAGGAAGAGGTTTATGTTTGGG	Human DNA sequence from clone RP11-300B11 on chromosome 10, complete sequence	97%	100%
	CGGCATTGCTCAGAGGATTACACCTAAGATGAGCACCATTTCTAAACATGTCCTTGTATTGGCAGAGTTGTACCTT GGCACTTTCCTCAGTCTTACTGTGGAAACAAGCTAACACACACATGTTTGCGC	Homo sapiens genomic DNA, chromosome 11q, clone:RP11-734C14, complete sequence	97%	100%
	GAGGGGCAGAAGGAGCTGGCTGTAAAATGGAGGGTGTGACCCTGAGCTGGAGAAGAACACTCTAAGGAAAACTAGT GCCTAACCTGAAATAGGGGTGGTGGATTAGATAAGAGCACAGTGGCAGTGTTGAATGTCCTATTTGGCC	Homo sapiens FOSMID clone ABC13-49758800D1 from chromosome 2, complete sequence	97%	99.30%
Stomach	GGCGGAGAGTCCCAGCTGGGCTTCAGCCTCCAAGGAAGTTGAGACACACTTTCCTTTTTCTGAAGAGGTGCTCCCC TTCATGTTTTGTAGGGGGCTCTTCTCCCTTAGGAACTAGGGTTTATATGTGTATGAGAGG	Homo sapiens glutathione synthetase (GSS), RefSeqGene (LRG_1168) on chromosome 20	100%	99.26%
	CACTATTAACCATGAACATACCTCAAACACATGTGAAAAGGTCATATTCAATGCATTTTCTAGGTCGGTGGTGCTG TTTTACCTTCCATTTGTATAGCATTTTACAAGTCCCAATGTGTTGCC	Human DNA sequence from clone RP11-686D16 on chromosome 6, complete sequence	97%	99.17%
	TGCATAACACAAAAATAGCTAACTGCTTTGCTTAAATCATATTTGGAGGGTGACTAAGCATGAGACAGTTCCCAAA CAAATCTTTGAGAAACTTATGCAGGAGTGAAAGTGAGATTTCAGCTGTAATCAGTCTCTTAATCTAGGCC	Homo sapiens methyl-CpG binding domain protein 5 (MBD5), RefSeqGene on chromosome 2	98%	100%
Liver	TAGAAATGTTATTGTTGGACACCTGTTTATTCTTCTTTTCATTCAAGTTGTTCTAACACAGGGAATTTTCTTGGGT TTAAGAGAAAATACAATGTAGCAATTCAGTTGCCTAAAGGCAATGCCC	Homo sapiens chromosome 10 clone RP11-491G3, complete sequence	95%	100%
	GGGAGCAGCTGCCTTTTCTCCCGTGTTTTGCCAAAGGGAAGACAGTTTCTGTGTCCCGTCTCTCCTTGGAGAGGGT TATCTCTTTTTGGAATTCAGTTTACATAATTGCTTTGCAGCCTCAGCTCACTGTGGGCTCAAAAATGTTATGAT	Homo sapiens BAC clone RP11-465K4 from 8, complete sequence	98%	100%
	GCGCCCCAGAGGAGCTCGTCCTGGTCCACAGCTTTATATCACCTACATGCTGAAATATTCAGGCTGCTGCTTTGAG CTCTAGGATCATACAGCCAGTGCCACTTTGAAATCTCCAAGGTAATCTTAAACCCACGCACCCCT	Homo sapiens isolate NA24385 chromosome 17	97%	100%
Lung	GGGCAAGCAGTTCATTATAATTAGAAGCATATTTACTATGAAGTTAATGAAACATAAGCTGTGAGGTCTCTCACTT GCCTTTCAAGGCCCCTAGGGGATGTGCATATGGAAAGAAGCTGGGG	Homo sapiens chromosome 4 clone C0084L10 map 4p16, complete sequence	97%	100%
	GCGGGTGCCCCAGGAGGTCAATTAGGAAAATGAGCTAAACTGAAACTAACAATCAAGACTTTATTCATTTACTGTG ACAATATAAGCAAGTAGACAAAAAGAAAGGCATCAACTCCCAGTGTTCCATTTCTTCCTGCAGGATGCC	Homo sapiens BAC clone RP11-439B18 from 4, complete sequence	96%	100%
	CCCGTGGCATATTGGGATAATCTCTATGAGGCACTTACCTCTGAGGGCTGTGATTCTTCACTGGGTCTTTTAGCTG CTGAGAGCATGGCTCAGTCTTATTGATCAAATTTGTAGAGCATAACACACAGACCC	Human DNA sequence from clone ICI-13CF11 on chromosome 1, complete sequence	97%	100%

5 Discussion

Our study presents a pioneering approach to cancer classification, called GeneLLM. This sophisticated machine learning framework is based on transformers and deeply

understands cfRNA sequences without traditional bioinformatics processing relying on genome annotations. It significantly enhances the accuracy of cancer classification across multiple cancer types. Compared to traditional biomarkers and conventional diagnostic methods, our results show that GeneLLM can reliably distinguish between cancerous and non-cancerous cfRNA samples, even in early-stage cancers where early detection is crucial for successful treatment outcomes.

Additionally, our study validated the efficacy of this methodology across diverse cancer types and patient cohorts from multiple centres. The robustness of our approach was evident in its consistent performance with shallow sequencing depth (1 GB of data). This scalability can reduce sequencing and computational costs and complexity while maintaining high diagnostic accuracy. Unlike traditional bioinformatics methods, our method discovered novel biomarkers in the unexplored genome 'dark matter' by leveraging the power of large language models to deeply understand cfRNA sequences. The potential implications for clinical practice are profound, offering a more accurate, cost-effective, and scalable solution for cancer detection.

Our study pushes early-stage cancer detection into a new era. Current challenges in cfRNA-based approaches are relying on known genes as biomarkers. They have limited scalability (specific to certain diseases) and limited overall sensitivity and specificity in real clinical settings. To overcome this pitfall, we developed GeneLLM, which does not rely on any previous knowledge and is fully data-driven to find 'pseudo-biomarkers'. We demonstrate that GeneLLM can discover the 'dark matter' in the human genome, which is not detectable using any previously reported approaches. It relies on these 'pseudo-biomarkers' for high-accuracy cancer detection. Additionally, this framework is versatile and can be adapted across multiple centres, effectively addressing high batch effects seen in different clinical settings.

Our results also demonstrated unique cfRNA signatures in cancers. Outside of the cells, RNA molecules are primarily associated with cell-cell communication and

regulatory signal transduction, which differ from their behaviours within cells. The liver and lungs, with their high metabolic and detoxification activities, have microenvironments rich in blood supply and unique immune cell populations that can promote tumour growth and metastasis. This can lead to the upregulation of pathways [6, 16, 19, 22], causing increased secretion of cfRNAs. Conversely, the gastrointestinal tract has a robust immune surveillance system. Colorectal and stomach cancers are influenced by the gut microbiome, dietary factors, and chronic inflammation. Chronic inflammation might induce a different systemic response, potentially leading to immunosuppression or metabolic down-regulation as a means to conserve energy and resources for the body's defence mechanisms [20, 27, 33, 41]. This results in reduced secretion of cfRNA molecules.

Integrating multi-omics data might be a promising future direction. It has been shown that combining cfRNA and cfDNA can enhance cancer detection [2]. Proteomics in cancer liquid biopsies are also useful [11]. Besides cfDNA fragmentomics, recent studies show that cfDNA methylation profiles are also a promising cancer detection tool [15]. Different omics provide different layers of information, and integrating multi-omics in liquid biopsies could be a significant advancement in future cancer detection.

6 Methods

6.1 Ethics approval

The ethics of this study were approved by the Ethics Committee and the Institutional Review Board of Peking University First Hospital (2022-132), Beijing Hospital (2022BJYYEC-420-02), and Sir Run-Run Shaw Hospital (2023-0300). Written informed consent was obtained from all patients prior to participation.

6.2 Cohort design

The cohort for this study comprised 496 plasma samples, including 102 colorectal cancer and 102 stomach cancer samples from Peking University First Hospital, 81 liver cancer samples from Sir Run-Run Shaw Hospital, 47 lung cancer samples from Beijing Hospital, and 164 non-cancer controls. The non-cancer controls were obtained from all three centres to calibrate potential batch effects.

6.3 Sample collection

Peripheral whole blood samples were taken from subjects before treatment using EDTA-coated tubes. The tubes were inverted between 8 and 10 times to ensure the mixing of the blood with the anticoagulant. Plasma separation was performed following a standard clinical protocol for blood centrifugation within two hours of blood collection. After separation, the plasma samples were divided into aliquots and preserved at −80°C until the extraction of cfRNA.

6.4 cfRNA extraction, library preparation, and sequencing

Total cfRNA was isolated from 200 μL of plasma utilising the miRNeasy Serum/-Plasma Advanced Kit (Qiagen), which employs Qiagen's unique resin as the separation matrix. This kit is capable of extracting all RNA molecules from approximately 18 nt upwards. DNase I (NEB) was used to eliminate remaining DNA.

To assess potential contamination in the RNA extraction and library preparation processes, two types of negative controls were utilised. These included libraries generated from RNA/DNA-free water (Invitrogen) and solutions eluted from the miRNeasy Serum/Plasma Advanced Kit with RNA/DNA-free water as input. A non-reverse transcriptase control was implemented to prevent DNA contamination. The positive control comprised human brain RNA, provided SMARTer Stranded Total

RNA-Seq Kit, and was processed using the same protocol as that for the cfRNA library preparation.

The comprehensive cfRNA library (approximately > 50 nt) was constructed using the SMARTer Stranded Total RNA-Seq Kit v2-Pico Input Mammalian (Takara), adhering to the manufacturer's recommended protocol. This kit employs a CRISPR/-DASH technique to eliminate ribosomal cDNA. Each library was sequenced over 40 million reads using Illumina NovaSeq6000 platform using a PE150 strategy.

6.5 Data quality control and preprocessing

To ensure high-quality data, we employed strict quality control processes. First, all raw data were trimmed using `fastp` v0.23.3 [7] with default parameters to automatically trim low-quality reads and adapters. The trimmed reads were then mapped to the human genome (GENCODE human release 43 [14]) using `STAR` v2.7.10b [12] with the encyclopedia of DNA elements (ENCODE) standard options [13, 17]. The mapped reads were used as input for GeneLLM. Mapping statistics were then retrieved using the `flagstat` function in `SAMtools` v1.17 [10].

6.6 Downstream bioinformatics analysis

Gene-level feature counting was performed using subread `featureCounts` v2.0.6 [24] with GENCODE human genome annotation (release 43) [14]. Downstream differential expression analysis was conducted using the `limma-voom` pipeline v3.60.2 [23], considering genes with False Discovery Rate (FDR) < 0.05 between conditions as significantly changed in expression.

GSEA was performed using `fgsea` v1.24.0 [21], employing a log2FC ordered gene vector with 1,000 permutations against gene sets from MSigDB v2023.1 [25, 37]. Terms with adjusted p-values < 0.05 were deemed significant.

Pathway activities were inferred using the `progeny` R package v1.26.0. PROGENy infers biological pathway activity from gene expression data by utilising a vast collection of perturbation experiments to identify common pathway-responsive genes [34]. This method overcomes the limitations of traditional mapping methods and provides insights into driver mutations, drug indications, and pathway distinctions [34].

6.7 AI model

6.7.1 Inputs and data sources

The input for GeneLLM is a sequence of tokens derived from the cfRNA sequencing reads of both cancer patients and individuals without cancer. Each patient contributes approximately 40 million cfRNA reads. During the pretraining phase, GeneLLM focuses on understanding the internal relationships within the cfRNA reads and storing pertinent information without using cancer labels. These cfRNA reads are then converted into 7-mers, each considered a token.

For batch processing, the lengths of different ID sequences are standardised by padding, inserting `[PAD]` tokens to ensure all token sequences in a batch are of equal length. Consistent with Natural Language Processing (NLP) tasks, specific tokens `[SOS]` and `[EOS]` are employed to mark the start and end of each sequence, respectively.

6.7.2 Pretraining

During the pretraining phase described in Figure 2A, each cfRNA sequence is treated as an independent sentence, with 7-mer sequences acting as tokens. The core model architecture is a deep Transformer decoder network. To distinguish between input and target sequences during training, a special `[SOS]` (start of sequence) token is prepended to the input tokens, and an `[EOS]` (end of sequence) token is appended to the target tokens. The model employs a standard language modelling objective,

learning to predict the next token in the target sequence based on the input sequence and preceding target tokens.

The model is composed of six Transformer blocks stacked sequentially. Each block comprises a self-attention sublayer and a feed-forward sublayer, adhering to the original Transformer architecture proposed by Vaswani *et al.* (2017) [39]. A triangular masking pattern is applied to the self-attention matrices within each block. This masking enforces the autoregressive property of the decoder, ensuring predictions for a given target token to depend only on previous tokens in the sequence and the input tokens, never on future tokens.

Pretraining on unlabelled cfRNA sequencing data enables the model to learn meaningful representations that can be fine-tuned for downstream tasks.

6.7.3 Pesudo-biomarker mining

Once the model is pretrained, we fix the parameters and use it to mine hidden information within the massive data of human cfRNA reads. Since each patient typically has approximately 40 million cfRNA reads, we first use GeneLLM to encode each read into an embedding. Denote the i-th patient's j-th cfRNA read as s_{ij}, we transformed it into a gene feature vector g_{ij} as shown in Equation (1), with an illustrative process in Figure 2A.

$$g_{ij} = \text{GeneLLM}(s_{ij}), \tag{1}$$

where g_{ij} is the last output vector.

For accurate and interpretable caner detection, we need to identify 'biomarkers', referred to as 'pseudo-biomarkers' in this study. These pseudo-biomarkers are vectors that best represent all cfRNA vectors g_{ij}. Denoting the pseudo-biomarker vectors as $v_1, \ldots, v_n$, each cfRNA representation g_{ij} should match a pseudo-biomarker that best represents it. Therefore, we aim to minimise Equation (2).

$$L_{pm} = -\sum_{k=1}^{n}\sum_{ij}\log\big(p_r(g_{ij}|v_C)p_m(C|g_{ij};\{v_1,\ldots,v_n\})\big), \tag{2}$$

where C represents an index of the pseudo-biomarkers, and p_m represents the decision probability for each cfRNA vector, defined as:

$$p_m(C|x) = \frac{\exp(-||v_C - x||_2)}{\sum_k \exp(-||v_k - x||_2)}, \tag{3}$$

where p_r represents the reconstruction probability.

Inspired by Sha *et al.* (2021) [35], each pseudo-biomarker is defined as a Gaussian distribution, with v_C as the mean and σ_C as the standard deviation. The reconstruction probability is calculated by:

$$p_r(x|v_C) = \frac{1}{\sigma_C\sqrt{2\pi}}\exp\Big(-\frac{(x-v_C)^2}{2\sigma_C^2}\Big), \tag{4}$$

where C is the index for pesudo-biomarkers, and both v_C and σ_C are trainable parameters. This pseudo-biomarker mining phase can be seen as an Large Language Model (LLM) alignment process, aligning the information in GeneLLM with the pseudo-biomarkers by minimising L_{pm} in Equation (2). After this process, the pseudo-biomarker vectors $v_1,\ldots,v_n$ are used to extract important features from the extensive cfRNA reads for each patient.

6.7.4 Disease tuning

Since each patient has approximately 40 million cfRNA reads, summarising these features is essential. This is achieved using pseudo-biomarkers, as illustrated in Figure 2B. For the i-th patient, we summarise the feature as follows:

$$f_i(k) = \sum_j p_m(k|g_{ij}), \quad for\ k = 1,\ldots,n, \tag{5}$$

$$f_i = [f_i(1), \ldots, f_i(n)], \tag{6}$$

where $f_i(k)$ represents the feature for pseudo-biomarker vector v_k.

Source bias elimination

To address biases from different sources (hospitals), we train a mapping for each patient source to align the non-cancer cases from each source to a common distribution. This process is depicted in Figure 2B as the source bias elimination module.

Formally, assuming the i-th patient has a feature vector f_i and a source s_i, we collect a set of all healthy cases from the training set, denoted as $\mathscr{H}$. We then calculate the midpoint for each source:

$$\mu_k = \frac{\sum_{i \in \mathscr{H}} \mathbb{I}_{s_i=k} f_i}{\sum_{i \in \mathscr{H}} \mathbb{I}_{s_i=k}}, \tag{7}$$

where μ_k represents the midpoint value for the k-th source, with $\mathbb{I}_{s_i=k}$ as an indicator function, which is 1 when s_i equals to k, and 0 otherwise. Next, we align all the midpoints to one standard midpoint, μ_*[1]. Each patient's feature vector is then corrected as follows:

$$f_i' = f_i + \mu_* - \mu_{s_i}, \tag{8}$$

where f_i' is the corrected feature vector without source bias.

Multi-scale feature extractor

The feature vector, derived from the vast amount of cfRNA reads, exhibits a rich set of multi-scale features that capture intricate patterns and variations at different levels of granularity. These multi-scale features are crucial for understanding the complex nature of the cfRNA sequences and their potential role in disease detection.

[1] In practice, the μ_* can be set as the average value of all midpoints.

Inspired by the Dense Convolutional Network architecture proposed by [18], we propose to extract multi-scale features by employing stacked skip-connection layers. As illustrated in Figure 2B, we utilize multiple Multi-scale Feature Extractor modules with activation functions such as the `ReLu(·)` [1] and the `Tanh(·)` to extract multi-scale features. Within each module, we iteratively apply a feedforward layer to the input feature vector and concatenate the output to the input, effectively preserving information from previous layers and enabling the network to learn multi-scale representations.

The proposed architecture leverages the power of skip connections, which allow the network to propagate information from earlier layers to later layers, mitigating the vanishing gradient problem and facilitating the learning of multi-scale features. By iteratively applying feedforward layers and concatenating the outputs, the network can capture a wide range of patterns and variations present in the cfRNA sequences at different scales. This approach enables the model to effectively learn and represent the complex characteristics of the cfRNA reads, which is essential for accurate disease detection.

By incorporating multiple Multi-scale Feature Extractor modules, the proposed architecture can extract a comprehensive set of multi-scale features that encapsulate the rich information present in the cfRNA sequences. These features serve as a robust foundation for subsequent analysis and disease detection tasks, enabling the development of more accurate and reliable diagnostic tools based on cfRNA sequencing data.

Multiple comparable downstream classifier head

In this part, we aim to use each patient's feature vector to detect any possible diseases they may have. Given the multi-scale feature vector for the i-th patient, denoted as $\tilde{f}_i$, we assume that we are going to detect N_d types of diseases. We construct a

feed-forward classifier for each disease as follows:

$$p(d_j|i) = \text{softmax}(FFN(\tilde{f}_i)), \tag{9}$$

where 'FFN' represents the feed-forward layer, and d_j represents the j-th disease. Each classifier head is trained using samples from patients with that specific disease and non-cancer controls.

However, there is still one issue: the probabilities calculated by different classifiers usually cannot be directly compared. For example, a patient may have a probability of 0.6 according to the stomach cancer classifier and 0.5 according to the lung cancer classifier. However, if the threshold for the stomach cancer classifier is 0.7 and the threshold for the lung cancer classifier is 0.4, this patient should be classified as having lung cancer. If we simply compare the absolute values of the probabilities, we will obtain an incorrect result. Therefore, it is necessary to calibrate the disease screening probabilities across all classifiers. Here, we use a simple yet effective method called 'pivot calibration'. Given the threshold $\hat{p}_j$ for the j-th disease and a general threshold p_M, we force every $\hat{p}_j$ to be equal to p_M, and every probability is tuned by ratio. The detailed calibration process is described in Eqn. (10).

$$p_{cal}(d_j|i) = \frac{p(d_j|i)}{\hat{p}_j} p_M \mathbb{I}(p(d_j|i) < \hat{p}_j) + \Big(1 - \frac{1 - p(d_j|i)}{1 - \hat{p}_j}(1 - p_M)\Big) \mathbb{I}(p(d_j|i) \geq \hat{p}_j), \tag{10}$$

where $\mathbb{I}(\cdot)$ represents the indicator function, which equals 1 if the condition inside the parentheses is true and 0 otherwise.

Finally, we compare the calibrated probability $p_{cal}(d_j|i)$ of every disease and select the disease with the highest probability as our prediction. If every probability is below p_M, then the patient is classified as non-cancer.

6.7.5 Training regimen

To train the GeneLLM, cfRNA reads were preprocessed as described in Section 6.7.1.

We trained the GeneLLM on 64 NVIDIA A100 GPUs, utilising their high-performance computing capabilities. To optimise memory usage and training efficiency, we used a batch size of 1 per GPU. The model was pre-trained using the entire training set of cfRNA sequences, allowing it to learn the underlying patterns and characteristics of the data. The pre-training process was computationally intensive and took approximately 15 days to complete.

After pre-training, we utilised the entire cfRNA dataset to extract pseudo-biomarkers. To further refine the GeneLLM for disease prediction, we performed disease tuning using labelled patient cfRNA sequential data. Each patient in the dataset has a massive amount of cfRNA sequences associated with their condition. By leveraging this labelled data, we fine-tuned the pre-trained GeneLLM to capture disease-specific patterns and characteristics. The disease tuning process adapts the model to better discriminate between different disease states and improves its performance in downstream tasks such as disease prediction.

The combination of pre-training on a large cfRNA dataset and subsequent disease tuning using labeled patient data enables the GeneLLM to learn both general and disease-specific representations of cfRNA sequences. This approach harnesses the power of transfer learning, where the knowledge gained from pre-training on a diverse dataset is transferred and specialised for specific disease applications through fine-tuning. The resulting model is expected to capture relevant biomarkers and patterns associated with various diseases, potentially enhancing the accuracy and efficiency of disease diagnosis and monitoring using cfRNA data.

6.7.6 Inference regimen

During the inference process, we sequentially input the cfRNA sequences into the GeneLLM, followed by the disease judgment head. The GeneLLM encodes each cfRNA sequence, capturing its relevant features and representations. The encoding time varies depending on the length of the cfRNA sequences, as longer sequences require more computational processing.

On average, the GeneLLM takes approximately 2.16×10^{-5} seconds to encode a single cfRNA sequence. However, it's important to note that each patient typically has a massive amount of cfRNA sequences associated with their sample. Therefore, the total encoding time for a single patient is around 14.4 minutes, considering the processing of all their cfRNA sequences.

After encoding the cfRNA sequences, the disease judgement head comes into play. This component of the model is responsible for making predictions or classifications regarding the presence or absence of specific diseases based on the encoded representations. To enhance the robustness and reliability of the disease judgement, we employ an ensemble approach. We use 10 ensemble models, each trained on different subsets or variations of the data, to make predictions. The ensemble approach helps to mitigate potential biases or variances in individual models and improves the overall accuracy of the disease judgement.

The inference time for the disease judgement head, using the 10 ensemble models, takes approximately 5.6 minutes on average for one patient. This inference time includes the processing of the encoded cfRNA sequences by each ensemble model and the aggregation of their predictions to reach a final judgement.

Compared to previous methods for cancer screening, the combination of the GeneLLM and the disease judgement head offers a remarkably quick and efficient approach. Traditional cancer screening methods often involve invasive procedures, extensive laboratory tests, and manual analysis, which can be time-consuming and

resource-intensive. In contrast, our proposed method leverages the power of machine learning and natural language processing techniques to analyze cfRNA sequences rapidly and accurately.

It's worth noting that while the inference times mentioned are impressive, they are based on the specific hardware and computational resources used in our experiments. The actual inference times may vary depending on the available computational infrastructure and the scale of the patient cohort being analyzed. Nevertheless, the efficiency and speed of our method highlight its potential for widespread adoption in clinical settings, enabling large-scale cancer screening programs and improving patient care.

6.7.7 Metrics

To evaluate the architecture's performance in detecting diseases, we calculated the confusion matrix and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The confusion matrix provides a tabular summary of the model's classification results, displaying the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. These counts offer insights into the model's ability to correctly identify positive and negative samples and the types of errors it makes. The AUC-ROC value represents the probability that a randomly selected positive sample will have a higher score assigned by the model than a randomly selected negative sample. A larger AUC value, ranging from 0 to 1, suggests better model performance in distinguishing between positive and negative samples.

Furthermore, we plotted the ROC curve to visually assess the model's performance. The ROC curve illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various classification threshold values. An ideal ROC curve would hug the top-left corner of the plot, indicating high TPR and low FPR across a wide range of thresholds. The closer the ROC curve is to the top-left corner, the better the model's performance. The AUC value summarises the overall performance of the

model across all possible thresholds, providing a single scalar value for comparison. By evaluating the architecture's performance using these metrics and visualisations, we gain a comprehensive understanding of its effectiveness in detecting diseases.

Supplementary information

Table S1. Patient information.

Table S2. Full DE analysis results for colorectal cancer.

Genes with positive LogFC and t-statistic are upregulated in cancer patients.

Table S3. Full DE analysis results for stomach cancer.

Genes with positive LogFC and t-statistic are upregulated in cancer patients.

Table S4. Full DE analysis results for liver cancer.

Genes with positive LogFC and t-statistic are upregulated in cancer patients.

Table S5. Full DE analysis results for lung cancer.

Genes with positive LogFC and t-statistic are upregulated in cancer patients.

Table S6. Full GSEA results against MSigDB hallmark collection v2023.1 for colorectal cancer.

Positive Enrichment Score (ES) and Normalised Enrichment Score (NES) values mean genes associated with the term are upregulated in cancer patients.

Table S7. Full GSEA results against MSigDB hallmark collection v2023.1 for stomach cancer.

Positive ES and NES values mean genes associated with the term are upregulated in cancer patients.

Table S8. Full GSEA results against MSigDB hallmark collection v2023.1 for liver cancer.

Positive ES and NES values mean genes associated with the term are upregulated in cancer patients.

Table S9. Full GSEA results against MSigDB hallmark collection v2023.1 for lung cancer.

Positive ES and NES values mean genes associated with the term are upregulated in cancer patients.

Declarations

Funding

This work was financially supported by Shenzhen OxTium Technology Co., Ltd (Shenzhen, China), National Natural Science Foundation of China under grant (No. KZ37117501, No. ZG216S23E8, No. 62306024), the Beijing Hospital New-star Plan of Science and Technology (BJ-2020-084), the Startup Foundation for Doctors of Beijing Hospital (BJ-2019-132), the National Key R&D Program of China (2022ZD0117700), Capital's Funds for Health Improvement and Research (2022-2-4075 and 2024-4-40711), and the National Natural Science Foundation of China (82271767).

Ethics approval

The ethics of this study were approved by the Ethics Committee and the Institutional Review Board of Peking University First Hospital (2022-132), Beijing Hospital (2022BJYYEC-420-02), and Sir Run-Run Shaw Hospital (2023-0300). Written informed consent was obtained from all patients prior to participation.

Consent to participate

Participants in this study were fully informed about the aims, procedures, and potential risks involved. All participants provided written informed consent before participation, and the study followed ethical guidelines consistent with the 1964 Helsinki declaration and its later amendments. Additionally, the study protocol was reviewed and approved by the Ethics Committee and the Institutional Review Board of Peking University First Hospital, Beijing Hospital, and Sir Run-Run Shaw Hospital, under approval number 2022-132, 2022BJYYEC-420-02, and 2023-0300, respectively.

Consent for publication

The authors affirm that all participants understand that their anonymised data will be published in a scientific journal. Consent for publication was also obtained from all participants at the time of consent for participation.

Availability of data and materials

All the raw sequencing data (including shallow sequencing) have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under the BioProject accession PRJNA1128376.

Code availability

The source code of GeneLLM, the model weights, and the inference code will be released after the paper is published.

Authors' contributions. The methodology and algorithm of GeneLLM, validated in this study, had been fully developed by L.S. with support from and exclusive authorisation granted to OxTium Technology Co., Ltd. S.D., Y.J., T.Z. and L.S. conceived current project idea. S.C. and P.W. designed clinical validation plan in this article. Y.J. designed and carried out wet-lab experiments. S.D. designed and performed all

bioinformatics. L.S. conceived, wrote and optimised artificial intelligent algorithms. T.Z., S.C., J.X., C.W. and P.W. coordinated human samples. Q.L. improved artificial intelligence. C.X., Y.X., Y.L., M.H. and Y.G. carried out various experiments. S.D., L.S., Y.J, T.Z., and S.C. wrote and edited manuscript. All authors discussed the results and commented on the manuscript.

Acknowledgments. This work was financially supported by Shenzhen OxTium Technology Co., Ltd (Shenzhen, China), National Natural Science Foundation of China under grant (No. KZ37117501, No. ZG216S23E8, No. 62306024), the Beijing Hospital New-star Plan of Science and Technology (BJ-2020-084), the Startup Foundation for Doctors of Beijing Hospital (BJ-2019-132), the National Key R&D Program of China (2022ZD0117700), Capital's Funds for Health Improvement and Research (2022-2-4075 and 2024-4-40711), and the National Natural Science Foundation of China (82271767). We also recognise the investors in OxTium Technology, whose commitment to advancing the boundaries of technology has been invaluable. We acknowledge the assistance of Jian Guo, Yelin Lai, Junhao Chen, and Yunlong Hao in this study.

Conflict of interest. The authors declare no conflict of interest.

References

[1] Agarap AF (2018) Deep learning using rectified linear units (relu). arXiv preprint arXiv:180308375

[2] Albitar M, Zhang H, Charifa A, et al (2023) Combining cell-free RNA with cell-free DNA in liquid biopsy for hematologic and solid tumors. Heliyon 9(5):e16261. https://doi.org/10.1016/j.heliyon.2023.e16261

[3] Barlow WE, Beaber EF, Geller BM, et al (2020) Evaluating Screening Participation, Follow-up, and Outcomes for Breast, Cervical, and Colorectal Cancer in the

PROSPR Consortium. Journal of the National Cancer Institute 112(3):238–246. https://doi.org/10.1093/jnci/djz137

[4] Bozgeyik I (2022) The dark matter of the human genome and its role in human cancers. Gene 811:146084. https://doi.org/10.1016/j.gene.2021.146084

[5] Bretthauer M, Løberg M, Wieszczy P, et al (2022) Effect of Colonoscopy Screening on Risks of Colorectal Cancer and Related Death. The New England Journal of Medicine 387(17):1547–1556. https://doi.org/10.1056/NEJMoa2208375

[6] Chen R, Xu X, Qian Z, et al (2019) The biological functions and clinical applications of exosomes in lung cancer. Cellular and molecular life sciences: CMLS 76(23):4613–4633. https://doi.org/10.1007/s00018-019-03233-y

[7] Chen S, Zhou Y, Chen Y, et al (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England) 34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560

[8] Chen S, Jin Y, Wang S, et al (2022) Cancer type classification using plasma cell-free RNAs derived from human and microbes. eLife 11:e75181. https://doi.org/10.7554/eLife.75181

[9] Crosby D, Bhatia S, Brindle KM, et al (2022) Early detection of cancer. Science (New York, NY) 375(6586):eaay9040. https://doi.org/10.1126/science.aay9040

[10] Danecek P, Bonfield JK, Liddle J, et al (2021) Twelve years of SAMtools and BCFtools. GigaScience 10(2):giab008. https://doi.org/10.1093/gigascience/giab008

[11] Ding Z, Wang N, Ji N, et al (2022) Proteomics technologies for cancer liquid biopsies. Molecular Cancer 21(1):53. https://doi.org/10.1186/s12943-022-01526-8

[12] Dobin A, Davis CA, Schlesinger F, et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635

[13] ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. https://doi.org/10.1038/nature11247

[14] Frankish A, Diekhans M, Jungreis I, et al (2021) GENCODE 2021. Nucleic Acids Research 49(D1):D916–D923. https://doi.org/10.1093/nar/gkaa1087

[15] Gao Q, Lin YP, Li BS, et al (2023) Unintrusive multi-cancer detection by circulating cell-free DNA methylation sequencing (THUNDER): development and independent validation studies. Annals of Oncology: Official Journal of the European Society for Medical Oncology 34(5):486–495. https://doi.org/10.1016/j.annonc.2023.02.010

[16] Gola A, Dorrington MG, Speranza E, et al (2021) Commensal-driven immune zonation of the liver promotes host defence. Nature 589(7840):131–136. https://doi.org/10.1038/s41586-020-2977-2

[17] Hitz BC, Lee JW, Jolanki O, et al (2023) The ENCODE uniform analysis pipelines. https://doi.org/10.1101/2023.04.04.535623, URL https://www.biorxiv.org/content/10.1101/2023.04.04.535623v1, pages: 2023.04.04.535623 Section: New Results

[18] Huang G, Liu Z, Van Der Maaten L, et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708

[19] Kim N, Kim HK, Lee K, et al (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nature Communications 11(1):2285. https://doi.org/10.1038/s41467-020-16164-1

[20] Koncina E, Nurmik M, Pozdeev VI, et al (2023) IL1R1+ cancer-associated fibroblasts drive tumor development and immunosuppression in colorectal cancer. Nature Communications 14(1):4251. https://doi.org/10.1038/s41467-023-39953-w

[21] Korotkevich G, Sukhov V, Budin N, et al (2016) Fast gene set enrichment analysis. bioRxiv https://doi.org/10.1101/060012

[22] Kubes P, Jenne C (2018) Immune Responses in the Liver. Annual Review of Immunology 36:247–277. https://doi.org/10.1146/annurev-immunol-051116-052415

[23] Law CW, Chen Y, Shi W, et al (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15(2):R29. https://doi.org/10.1186/gb-2014-15-2-r29, URL https://doi.org/10.1186/gb-2014-15-2-r29

[24] Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England) 30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656

[25] Liberzon A, Subramanian A, Pinchback R, et al (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics (Oxford, England) 27(12):1739–1740. https://doi.org/10.1093/bioinformatics/btr260

[26] Ling H, Girnita L, Buda O, et al (2017) Non-coding RNAs: the cancer genome dark matter that matters! Clinical Chemistry and Laboratory Medicine 55(5):705–714. https://doi.org/10.1515/cclm-2016-0740

[27] Liu Y, Li C, Lu Y, et al (2022) Tumor microenvironment-mediated immune tolerance in development and treatment of gastric cancer. Frontiers in Immunology 13:1016817. https://doi.org/10.3389/fimmu.2022.1016817

[28] Moufarrej MN, Vorperian SK, Wong RJ, et al (2022) Early prediction of preeclampsia in pregnancy with cell-free RNA. Nature 602(7898):689–694. https://doi.org/10.1038/s41586-022-04410-z

[29] Ning C, Cai P, Liu X, et al (2023) A comprehensive evaluation of full-spectrum cell-free RNAs highlights cell-free RNA fragments for early-stage hepatocellular carcinoma detection. EBioMedicine 93:104645. https://doi.org/10.1016/j.ebiom.2023.104645

[30] Radford A, Wu J, Child R, et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

[31] Reggiardo RE, Maroli SV, Peddu V, et al (2023) Profiling of repetitive RNA sequences in the blood plasma of patients with cancer. Nature Biomedical Engineering 7(12):1627–1635. https://doi.org/10.1038/s41551-023-01081-7

[32] Roskams-Hieter B, Kim HJ, Anur P, et al (2022) Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. NPJ precision oncology 6(1):28. https://doi.org/10.1038/s41698-022-00270-y

[33] Sathe A, Mason K, Grimes SM, et al (2023) Colorectal Cancer Metastases in the Liver Establish Immunosuppressive Spatial Networking between Tumor-Associated SPP1+ Macrophages and Fibroblasts. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 29(1):244–260. https://doi.org/10.1158/1078-0432.CCR-22-2041

[34] Schubert M, Klinger B, Klünemann M, et al (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. Nature Communications 9(1):20. https://doi.org/10.1038/s41467-017-02391-6, URL https://www.nature.com/articles/s41467-017-02391-6

[35] Sha L, Lukasiewicz T (2021) Multi-type disentanglement without adversarial training. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9515–9523

[36] Siegel RL, Miller KD, Wagle NS, et al (2023) Cancer statistics, 2023. CA: a cancer journal for clinicians 73(1):17–48. https://doi.org/10.3322/caac.21763

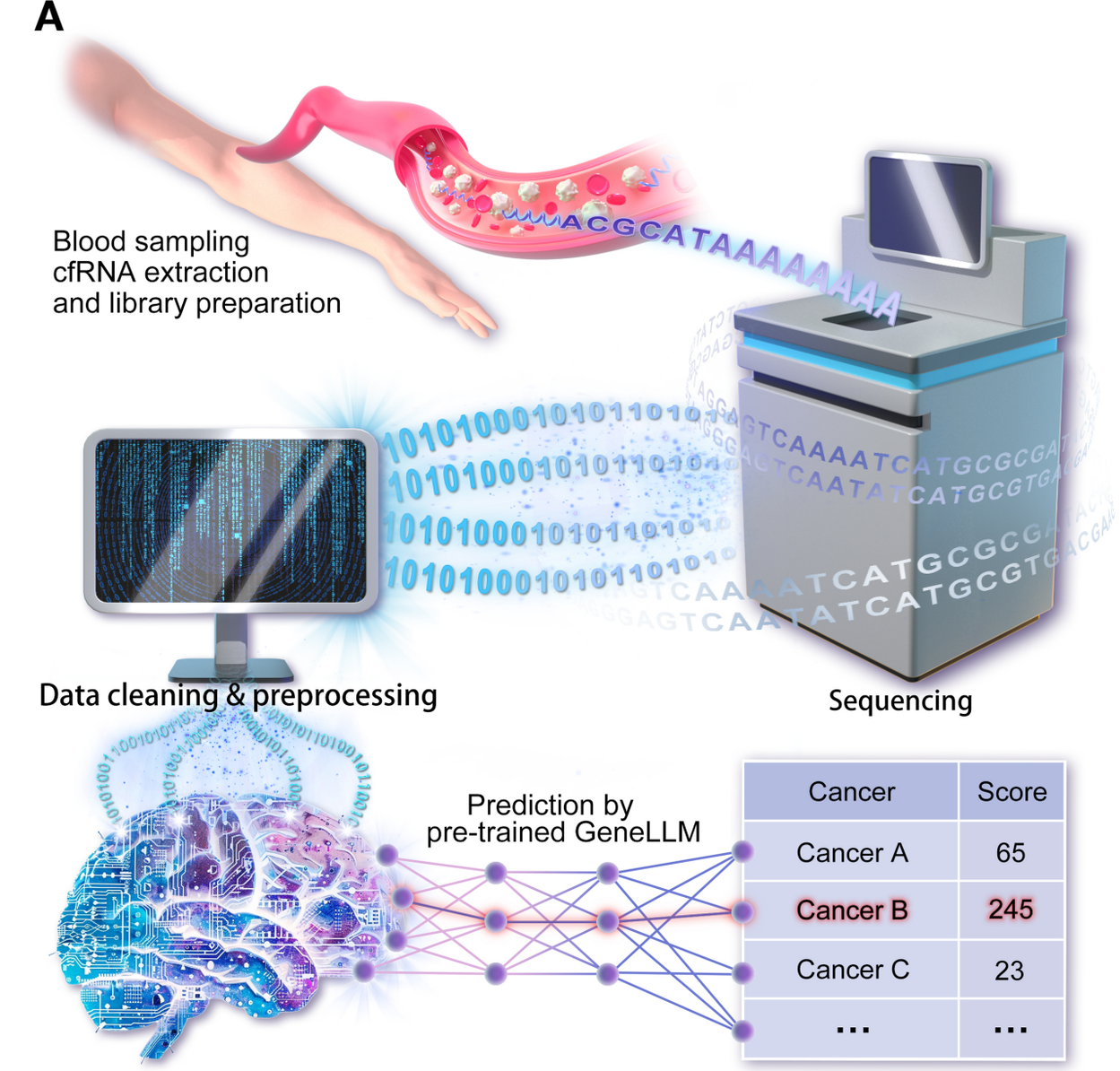
[37] Subramanian A, Tamayo P, Mootha VK, et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102

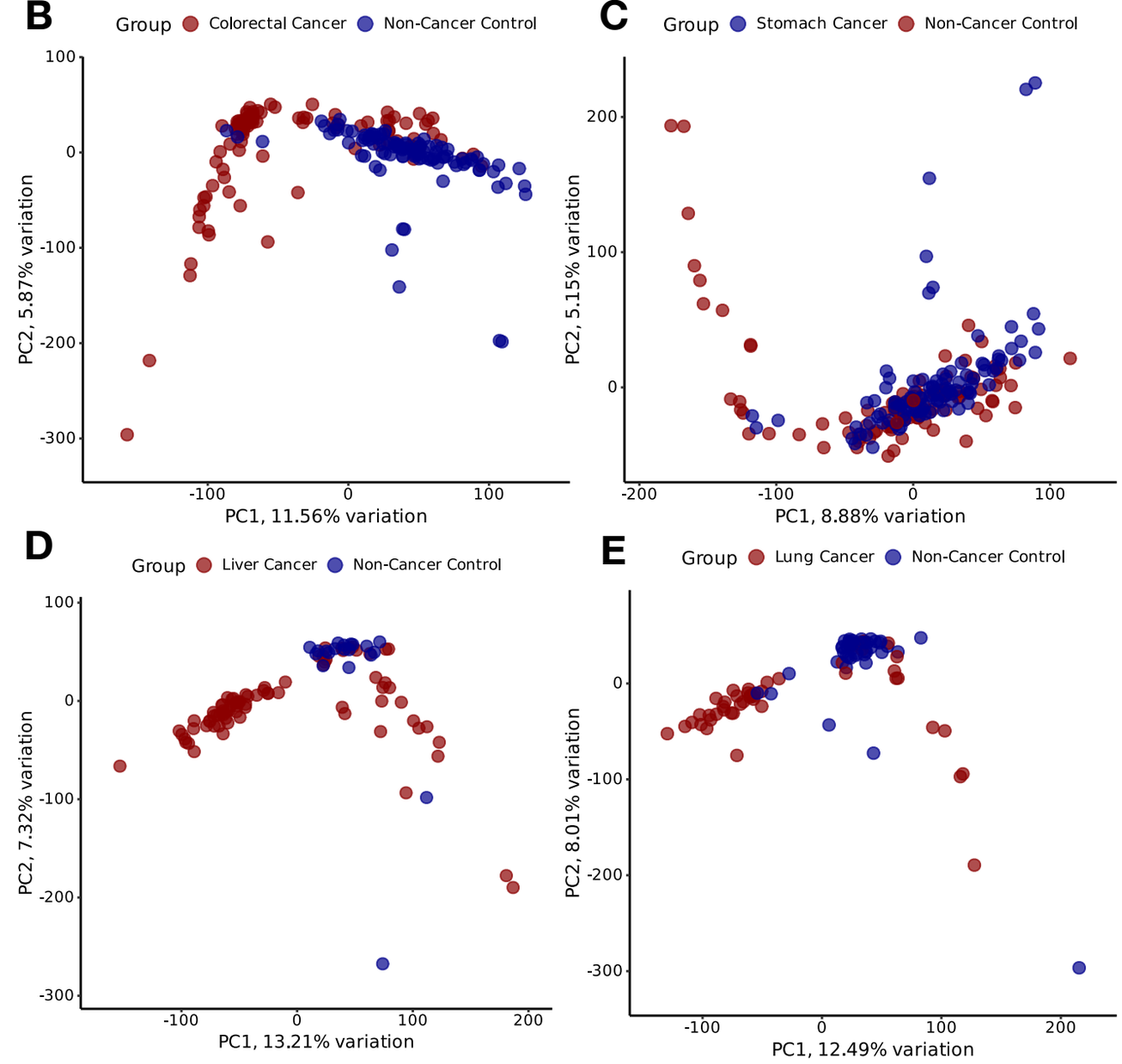
[38] Toden S, Zhuang J, Acosta AD, et al (2020) Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing. Science Advances 6(50):eabb1654. https://doi.org/10.1126/sciadv.abb1654

[39] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30

[40] Yu S, Li Y, Liao Z, et al (2020) Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut 69(3):540–550. https://doi.org/10.1136/gutjnl-2019-318860

[41] Zhu X, Liang R, Lan T, et al (2022) Tumor-associated macrophage-specific CD155 contributes to M2-phenotype transition, immunosuppression, and tumor progression in colorectal cancer. Journal for Immunotherapy of Cancer 10(9):e004219. https://doi.org/10.1136/jitc-2021-004219





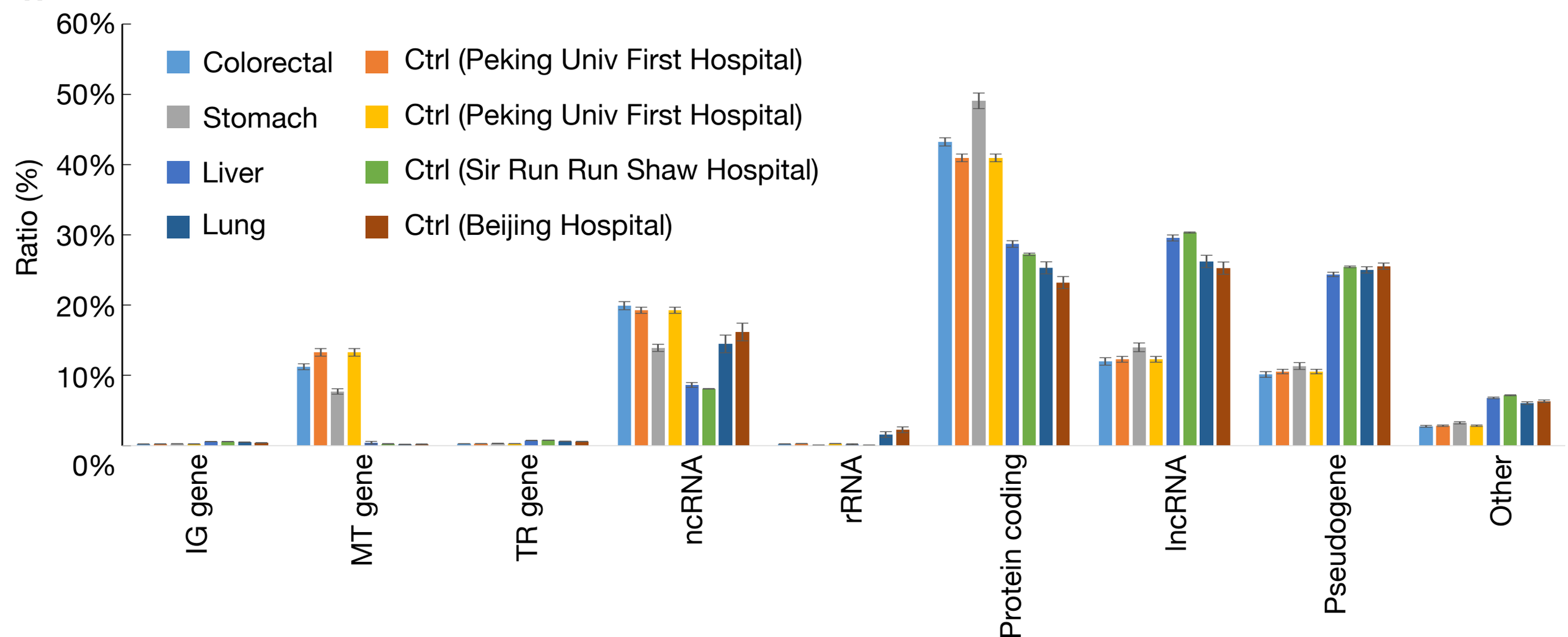
F

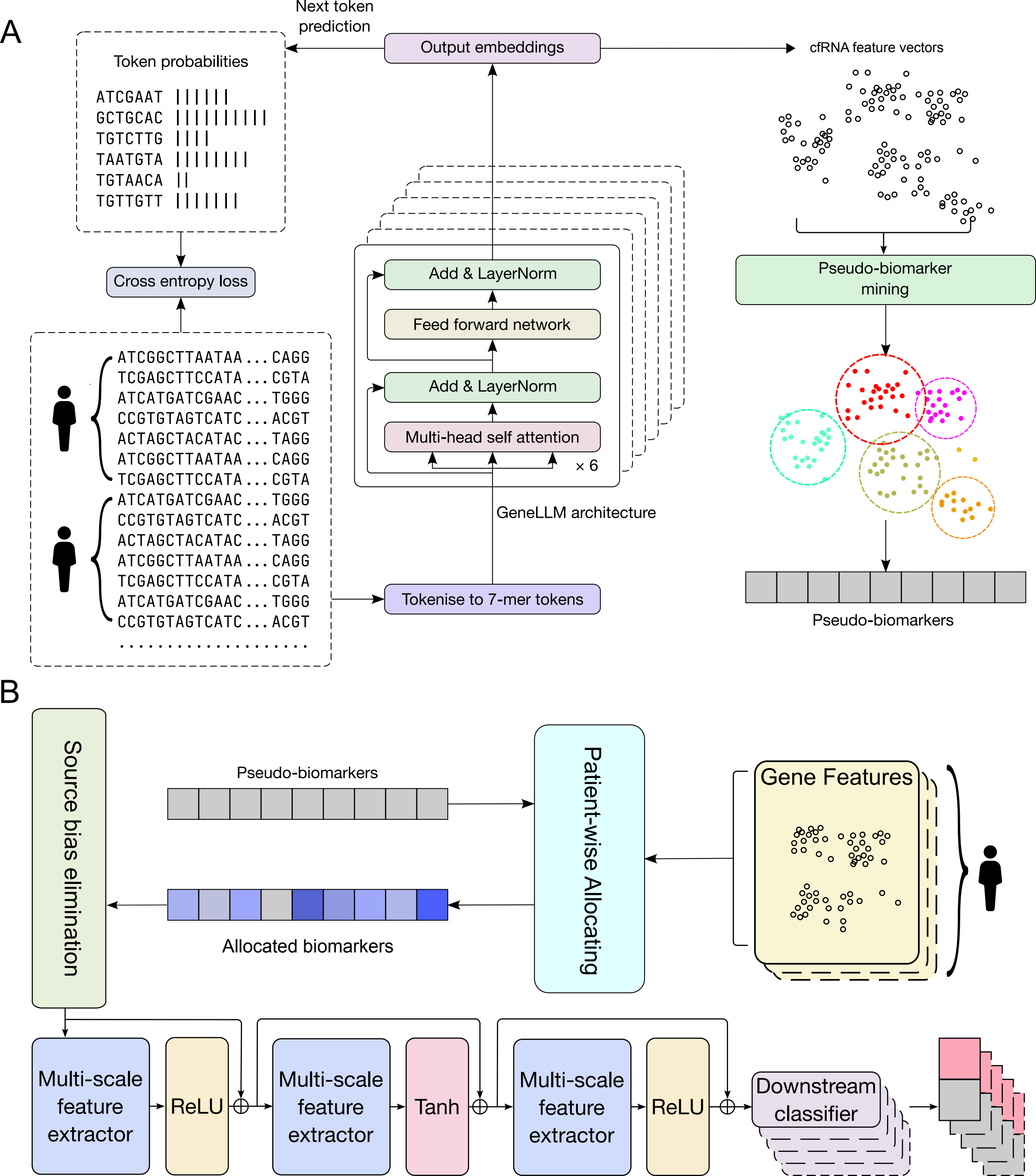
	Male				Female				
	I	II	III	IV	I	II	III	IV	Total
Colorectal cancer	10 (9.80%)	21 (20.59%)	21 (20.59%)	3 (2.94%)	10 (9.80%)	12 (11.76%)	22 (21.57%)	3 (2.94%)	102 (100%)
Stomach cancer	22 (21.57%)	16 (15.69%)	24 (23.53%)	10 (9.80%)	7 (6.86%)	7 (6.86%)	10 (9.80%)	6 (5.88%)	102 (100%)
Liver cancer	46 (56.79%)	12 (14.81%)	0 (0.00%)	2 (2.47%)	15 (18.52%)	4 (4.94%)	1 (1.23%)	1 (1.23%)	81 (100%)
Lung cancer	19 (40.43%)	1 (2.13%)	4 (8.51%)	0 (0.00%)	21 (44.68%)	2 (4.26%)	0 (0.00%)	0 (0.00%)	47 (100%)

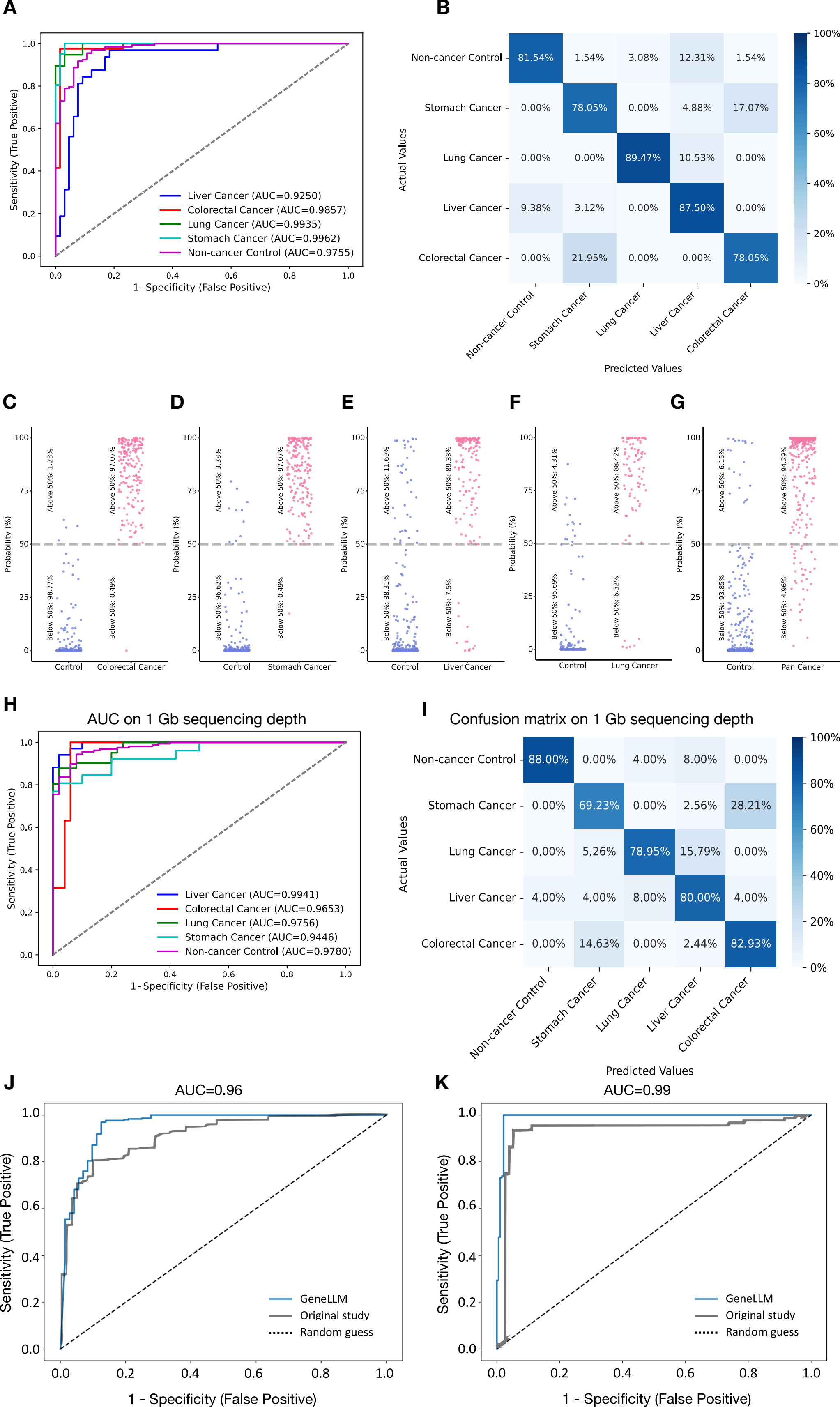
G

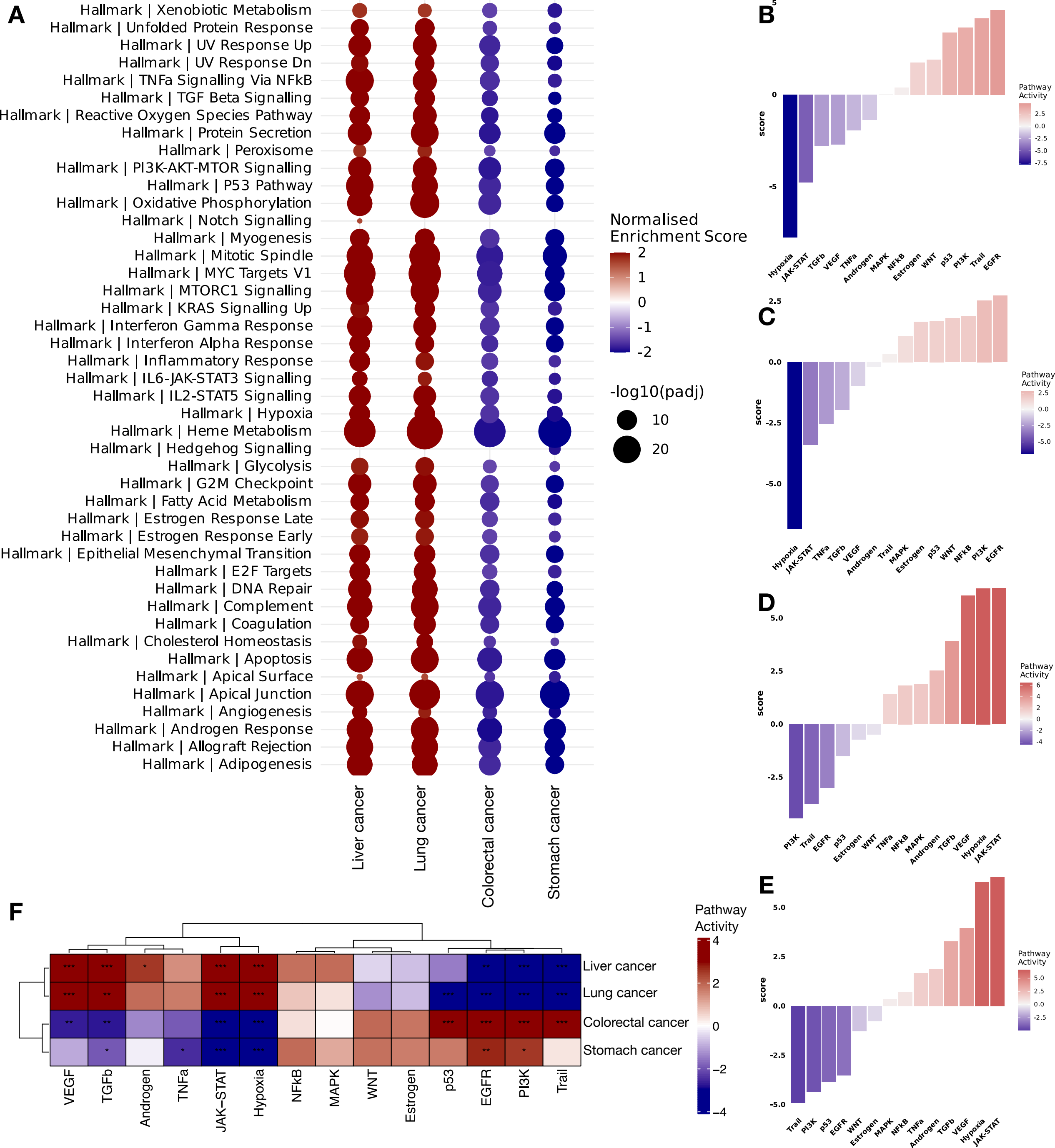
	Cancer		**Non-cancer Control**		**Total**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Peking University First Hospital	127 (41.64%)	77 (25.25%)	49 (16.06%)	52 (17.05%)	176 (57.70%)	129 (42.30%)
Sir Run Run Shaw Hospital	60 (59.41%)	21 (20.79%)	10 (9.90%)	10 (9.90%)	70 (69.31%)	31 (30.69%)
Beijing Hospital	24 (26.67%)	23 (25.56%)	28 (31.11%)	15 (16.66%)	52 (57.78%)	38 (42.22%)

H




A
Next token prediction
Token probabilities
ATCGAAT
GCTGCAC
TGTCTTG
TAATGTA
TGTAACA
TGTTGTT
Output embeddings
cfRNA feature vectors
Cross entropy loss
Add & LayerNorm
Feed forward network
Add & LayerNorm
Multi-head self attention
× 6
GeneLLM architecture
Pseudo-biomarker mining
ATCGGCTTAATAA ... CAGG
TCGAGCTTCCATA ... CGTA
ATCATGATCGAAC ... TGGG
CCGTGTAGTCATC ... ACGT
ACTAGCTACATAC ... TAGG
ATCGGCTTAATAA ... CAGG
TCGAGCTTCCATA ... CGTA
ATCATGATCGAAC ... TGGG
CCGTGTAGTCATC ... ACGT
ACTAGCTACATAC ... TAGG
ATCGGCTTAATAA ... CAGG
TCGAGCTTCCATA ... CGTA
ATCATGATCGAAC ... TGGG
CCGTGTAGTCATC ... ACGT
Tokenise to 7-mer tokens
Pseudo-biomarkers
B
Source bias elimination
Pseudo-biomarkers
Allocated biomarkers
Patient-wise Allocating
Gene Features
Multi-scale feature extractor
ReLU
Multi-scale feature extractor
Tanh
Multi-scale feature extractor
ReLU
Downstream classifier


A
Sensitivity (True Positive)
1 - Specificity (False Positive)
Liver Cancer (AUC=0.9250)
Colorectal Cancer (AUC=0.9857)
Lung Cancer (AUC=0.9935)
Stomach Cancer (AUC=0.9962)
Non-cancer Control (AUC=0.9755)
B
Actual Values
Predicted Values
Non-cancer Control
Stomach Cancer
Lung Cancer
Liver Cancer
Colorectal Cancer
81.54% 1.54% 3.08% 12.31% 1.54%
0.00% 78.05% 0.00% 4.88% 17.07%
0.00% 0.00% 89.47% 10.53% 0.00%
9.38% 3.12% 0.00% 87.50% 0.00%
0.00% 21.95% 0.00% 0.00% 78.05%
C
D
E
F
G
Probability (%)
Control
Colorectal Cancer
Stomach Cancer
Liver Cancer
Lung Cancer
Pan Cancer
Above 50%: 1.23%
Below 50%: 98.77%
Above 50%: 97.07%
Below 50%: 0.49%
Above 50%: 3.38%
Below 50%: 96.62%
Above 50%: 97.07%
Below 50%: 0.49%
Above 50%: 11.69%
Below 50%: 88.31%
Above 50%: 89.38%
Below 50%: 7.5%
Above 50%: 4.31%
Below 50%: 95.69%
Above 50%: 88.42%
Below 50%: 6.32%
Above 50%: 6.15%
Below 50%: 93.85%
Above 50%: 94.29%
Below 50%: 4.96%
H
AUC on 1 Gb sequencing depth
Sensitivity (True Positive)
1 - Specificity (False Positive)
Liver Cancer (AUC=0.9941)
Colorectal Cancer (AUC=0.9653)
Lung Cancer (AUC=0.9756)
Stomach Cancer (AUC=0.9446)
Non-cancer Control (AUC=0.9780)
I
Confusion matrix on 1 Gb sequencing depth
Actual Values
Predicted Values
Non-cancer Control
Stomach Cancer
Lung Cancer
Liver Cancer
Colorectal Cancer
88.00% 0.00% 4.00% 8.00% 0.00%
0.00% 69.23% 0.00% 2.56% 28.21%
0.00% 5.26% 78.95% 15.79% 0.00%
4.00% 4.00% 8.00% 80.00% 4.00%
0.00% 14.63% 0.00% 2.44% 82.93%
J
AUC=0.96
Sensitivity (True Positive)
1 - Specificity (False Positive)
GeneLLM
Original study
Random guess
K
AUC=0.99
Sensitivity (True Positive)
1 - Specificity (False Positive)
GeneLLM
Original study
Random guess


A
Hallmark | Xenobiotic Metabolism
Hallmark | Unfolded Protein Response
Hallmark | UV Response Up
Hallmark | UV Response Dn
Hallmark | TNFa Signalling Via NFkB
Hallmark | TGF Beta Signalling
Hallmark | Reactive Oxygen Species Pathway
Hallmark | Protein Secretion
Hallmark | Peroxisome
Hallmark | PI3K-AKT-MTOR Signalling
Hallmark | P53 Pathway
Hallmark | Oxidative Phosphorylation
Hallmark | Notch Signalling
Hallmark | Myogenesis
Hallmark | Mitotic Spindle
Hallmark | MYC Targets V1
Hallmark | MTORC1 Signalling
Hallmark | KRAS Signalling Up
Hallmark | Interferon Gamma Response
Hallmark | Interferon Alpha Response
Hallmark | Inflammatory Response
Hallmark | IL6-JAK-STAT3 Signalling
Hallmark | IL2-STAT5 Signalling
Hallmark | Hypoxia
Hallmark | Heme Metabolism
Hallmark | Hedgehog Signalling
Hallmark | Glycolysis
Hallmark | G2M Checkpoint
Hallmark | Fatty Acid Metabolism
Hallmark | Estrogen Response Late
Hallmark | Estrogen Response Early
Hallmark | Epithelial Mesenchymal Transition
Hallmark | E2F Targets
Hallmark | DNA Repair
Hallmark | Complement
Hallmark | Coagulation
Hallmark | Cholesterol Homeostasis
Hallmark | Apoptosis
Hallmark | Apical Surface
Hallmark | Apical Junction
Hallmark | Angiogenesis
Hallmark | Androgen Response
Hallmark | Allograft Rejection
Hallmark | Adipogenesis
Liver cancer
Lung cancer
Colorectal cancer
Stomach cancer
Normalised Enrichment Score
2
1
0
-1
-2
-log10(padj)
10
20
B
score
Pathway Activity
Hypoxia
JAK-STAT
TGFb
VEGF
TNFa
Androgen
MAPK
NFkB
Estrogen
WNT
p53
PI3K
Trail
EGFR
C
score
Pathway Activity
Hypoxia
JAK-STAT
TNFa
TGFb
VEGF
Androgen
Trail
MAPK
Estrogen
p53
WNT
NFkB
PI3K
EGFR
D
score
Pathway Activity
PI3K
Trail
EGFR
p53
Estrogen
WNT
TNFa
NFkB
MAPK
Androgen
TGFb
VEGF
Hypoxia
JAK-STAT
E
score
Pathway Activity
Trail
PI3K
p53
EGFR
WNT
Estrogen
MAPK
NFkB
TNFa
Androgen
TGFb
VEGF
Hypoxia
JAK-STAT
F
Liver cancer
Lung cancer
Colorectal cancer
Stomach cancer
VEGF
TGFb
Androgen
TNFa
JAK−STAT
Hypoxia
NFkB
MAPK
WNT
Estrogen
p53
EGFR
PI3K
Trail
Pathway Activity
4
2
0
−2
−4