Chen *et al. Biology Direct* (2025) 20:11
https://doi.org/10.1186/s13062-025-00598-y

Biology Direct

RESEARCH

Open Access

Functional genomic imaging (FGI), a virtual tool for visualization of functional gene expression modules in heterogeneous tumor samples

Xinlei Chen[1], Youbing Guo[1], Xiaorong Gu[1*] and Danyi Wen[1*]

Abstract

Advances in sequencing technologies are reshaping clinical diagnostics, prompting the development of new software tools to decipher big data. To this end, we developed functional genomic imaging (FGI), a visualization tool designed to assist clinicians in interpreting RNA-Seq results from patient samples. FGI uses weighted gene co-expression network analysis (WGCNA), followed by a modified Phenograph clustering algorithm to identify co-expression gene clusters. These gene modules were annotated and projected onto a t-SNE map for visualization. Annotation of FGI gene clusters revealed three categories: tissue-specific, functional, and positional. These clusters may be used to build tumor subtypes with pre-annotated functions. At the multi-cancer cohort level, tissue-specific clusters are enriched, whereas at the single cancer level, such as in lung cancer or ovarian cancer, positional clusters can be more prominent. Moreover, FGI analysis could also reveal molecular tumor subtypes not documented in clinical records and generated a more detailed co-expression gene cluster map. Based on different levels of FGI modeling, each individual tumor sample can be customized to display various types of information such as tissue origin, molecular subtypes, immune activation status, stromal signaling pathways, cell cycle activity, and potential amplicon regions which can aid in diagnosis and guide treatment decisions. Our results highlight the potential of FGI as a robust visualization tool for personalized medicine in molecular diagnosis.

Keywords Co-expression, WGCNA, Functional genomic imaging (FGI)

Introduction

RNA sequencing (RNA-Seq) has emerged as a valuable tool in both clinical diagnostics and research endeavors [1]. By analyzing RNA-Seq data, researchers and clinicians can gain valuable insights into a multitude of sample characteristics. These include, but are not limited to, cell proliferation rates, immune activation levels, cellular differentiation processes, apoptosis dynamics, and the activity of signaling pathways. The versatility and depth of information provided by RNA-Seq have positioned it as a cornerstone technology for unraveling the complexities of biological systems and advancing our understanding of disease mechanisms.

Conventional methods used in RNA-Seq data analysis include principal component analysis (PCA), unsupervised clustering, differential gene expression (DGE), and gene set enrichment analysis (GSEA) [2]. PCA and other dimensionality reduction techniques (e.g., t-SNE [3, 4]) may provide unbiased separation of samples on 2D or 3D planes. On the other hand, DGE analysis, using tools

*Correspondence:
Xiaorong Gu
xiaorong.gu@lidebiotech.com
Danyi Wen
danyi.wen@lidebiotech.com
[1] Shanghai LIDE Biotech Co., Ltd., Shanghai 201203, China

such as DESeq2 [5] and Limma [6], aims to identify genes that exhibit statistically significant changes across different sample groups and treatment conditions. GSEA goes a step further by revealing the enrichment of functional gene sets under different phenotypic and treatment conditions, which provide pathway activation status for samples. All these methods help identify genes and pathways of interest through defining sample groups and finding inter-group differences, which provide functional interpretations of the transcriptomic data and facilitate the discovery of potential biomarkers.

Although PCA and t-SNE allow samples to be visualized along principal component axes, one drawback is that the results may not be easily interpretable at the functional level. Moreover, dimensionality reduction algorithms have their own limitations. For instance, PCA is a linear method and may not capture complex, nonlinear patterns present in the data [7]. While t-SNE is computationally intensive and sensitive to parameters, researchers often use t-SNE combined with clustering methods to better interpret high-dimensional biological data [3, 8]. DEG and GSEA methods are useful for scrutinizing inter-group differences only when the groups are pre-determined. In real-world clinical samples, significant heterogeneity exits not only between sample groups but also within each sample. This complexity often renders conventional methods reliant on grouping ineffective, leading to suboptimal results in analysis.

In this study, we developed a gene expression analysis tool, functional genomic imaging (FGI), to visualize functional gene modules in a user-friendly manner. FGI utilizes weighted gene co-expression network analysis (WGCNA) for network construction [9], followed by dimensionality reduction, clustering, and functional annotation. It effectively showcases the functional module composition of co-expression networks in both pan-cancer and individual cancer types, providing a clear representation of functional pathway differences among tumor samples from clinical RNA-Seq data.

Materials and methods

Data sources

RNA-Seq data of the Cancer Genome Atlas (TCGA) were obtained from the GDC website (https://portal.gdc.cancer.gov, last accessed on January 11, 2024) [10]. We focused on 26 cancer types with more than 100 samples. Non-tumor and metastatic tumor samples were filtered out based on sample barcodes. The tumor types included in this study ordered by sample numbers are Breast invasive carcinoma (BRCA, n = 1049), Uterine Corpus Endometrial Carcinoma (UCEC, n = 554), Kidney renal clear cell carcinoma (KIRC, n = 541), Brain Lower Grade Glioma (LGG, n = 534), Thyroid carcinoma (THCA, n = 504), Head and Neck squamous cell carcinoma (HNSC, n = 502), Lung squamous cell carcinoma (LUSC, n = 502), Lung adenocarcinoma (LUAD, n = 501), Prostate adenocarcinoma (PRAD, n = 500), Bladder Urothelial Carcinoma (BLCA, n = 412), Colon adenocarcinoma (COAD, n = 397), Ovarian serous cystadenocarcinoma (OV, n = 381), Stomach adenocarcinoma (STAD, n = 375), Liver hepatocellular carcinoma (LIHC, n = 374), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, n = 304), Kidney renal papillary cell carcinoma (KIRP, n = 291), Sarcoma (SARC, n = 262), Pheochromocytoma and Paraganglioma (PCPG, n = 182), Glioblastoma multiforme (GBM, n = 169), Esophageal carcinoma (ESCA, n = 162), Testicular Germ Cell Tumors (TGCT, n = 156), Rectum adenocarcinoma (READ, n = 152), Acute Myeloid Leukemia (LAML, n = 150), Pancreatic adenocarcinoma (PAAD, n = 143), Thymoma (THYM, n = 120), and Skin Cutaneous Melanoma (SKCM, n = 103).

RNA-Seq data and proteomic data of the CPTAC dataset were obtained from the CPTAC Pan-Cancer Data website (https://pdc.cancer.gov/pdc/cptac-pancancer, last accessed on October 10, 2024).

Microarray dataset GSE63885 was downloaded from GEO website (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63885).

FGI procedures

FGI begins by constructing a co-expression network using weighted gene co-expression network analysis (WGCNA). Co-expressed gene modules are then identified using a modified Phenograph clustering algorithm [11]. These gene modules are then annotated using metascape programs [12], after which they are projected onto a gene t-SNE map and visualized with their functions clearly labeled (Fig. 1A).

(A) Data filtering

Before FGI analysis, genes underwent several filtering steps. TPM (transcripts per million) [13] values were log-transformed as $\log_2(\text{TPM}+1)$. Genes with a maximum $\log_2$TPM less than 2 across all samples were excluded. The Top 20,000 genes with the highest median absolute deviation (MAD) values were selected.

(B) Dimensionality reduction of genes

We first used the WGCNA R package (version 1.72.1) to obtain the topological overlap matrix (TOM) [14] from filtered RNA-Seq data and transformed TOM to a distance matrix as $-\log_2(1\text{-TOM})$.

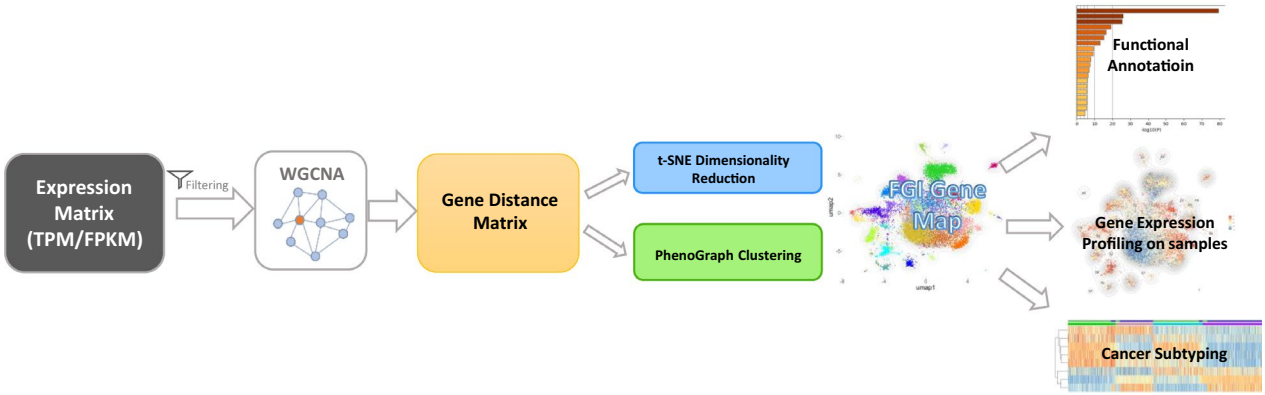

Fig. 1 Schematic workflow of the Functional Genomic Imaging (FGI) analysis. **A** The TPM/FPKM data matrix was filtered and processed using WGCNA, producing a gene distance matrix. Phenograph was applied to identify coexpression gene clusters, which were then visualized using t-SNE. Based on this, functional annotation, gene expression profiling, and cancer subtyping were performed to gain more detailed insights into gene expression patterns within different cancer types and their clinical relevance

Next, We used the Rtsne package (version 0.16) [15] to reduce the dimensionality of the gene distance matrix, with the following parameter settings: perplexity = 20, theta = 0.2. The "*is_distance*" parameter was set to TRUE to treat the input matrix as a distance matrix rather than a coordinate matrix.

(C) Gene clustering, annotation, and visualization

To normalize gene expression levels of each FGI cluster across samples, we calculated the enrichment score for each cluster using the GSVA R package (version 1.42.0) [16]. By default, WGCNA identifies gene modules through hierarchical clustering, we employed the Rphenograph package (version 0.99.1) for gene clustering, achieving more balanced clustering results which facilitated subsequent annotation steps (Details of the comparison are presented in Fig. S4). The Rphenograph package was slightly modified to allow processing distance matrices instead of gene expression matrices. The resulting FGI gene clusters were submitted to Metascape (https://metascape.org, version 3.5, last accessed on March 28, 2024) for pathway enrichment and cell type analyses. The annotated gene modules were then visualized using the ggplot2 R package (version 3.4.4) to plot clusters on t-SNE dimensions, using colors to highlight gene expression levels.

FGI maps for single samples were also generated to show the Z-score of each gene in the corresponding TCGA samples.

To validate the effectiveness of the workflow, we first tested our method using RNA-Seq data from the Cancer Genome Atlas (TCGA) repositories, including samples from different tissue origins. Then we further narrowed the focus to cancer types with known internal differences and tested our method on TCGA-LUAD/LUSC. Lastly, we applied FGI to high grade ovarian cancer, which has high heterogeneity and no histological sub-classifications.

Results

FGI analysis on the TCGA Pan-Cancer dataset

Tumor samples exhibit considerable heterogeneity in tissue sources and functionalities. We initiated our FGI (functional genomic imaging) process (Fig. 1) on the top level using TCGA pan-cancer RNA-Seq dataset which includes 780 samples from 26 cancer types. After filtering out low-expression genes and pseudogenes, the top 20,000 most variably expressed genes were used to compute gene distances via WGCNA, generating co-expression gene clusters (FGI gene clusters). This resulted in 54 clusters with gene numbers ranging from 6 to 2151 (Fig. 2A). Next, we calculated GSVA score of FGI gene

(See figure on next page.)

Fig. 2 FGI clustering analysis on TCGA pan-cancer RNA-Seq data. **A** t-SNE plot of 54 FGI gene clusters generated from 26 tumor types. **B** Clustering heatmap using GSVA scores of FGI gene clusters for samples of 26 tumor types. **C** Examples of tissue specific FGI gene clusters (liver and brain), showing boxplot of GSVA scores of clusters in tumor samples. **D** Examples of functional FGI gene clusters (cell cycle and lymphocyte activation), showing boxplot of GSVA scores of clusters in tumor samples. **E** Examples of positional FGI gene clusters, showing cluster names and gene distributions on chromosomes. **F** GSVA score violin plot of cluster_1 genes in male and female samples from TCGA

clusters for 6390 samples across 26 cancer types (TCGA sample sizes vary significantly across different cancer types, so we limited the sampling to a maximum of 300 samples per cancer type to ensure balanced data representation in the results.) and generated a pan-cancer expression heatmap of FGI gene clusters (Fig. 2B). This revealed that many clusters are expressed in a tissue-specific manner, suggesting that, at this top level, FGI may resolve tissue specificity. Next, we annotated each FGI gene cluster using Metascape (supplemental Table S1). This confirmed that many clusters are indeed specifically enriched in one or several tumor types. For example, cluster_6 is exclusively highly expressed in liver cancers (LIHC, Fig. 2C) and cluster_12 is only expressed in tumors from the brain (GBM and LGG, Fig. 2C). Some clusters are less tissue-specific but share common phenotypic characteristics. For instance, cluster_9 represents a cell cycle gene signature with elevated expression in multiple cancer types, including several squamous-like tumors (CESC, HNSC, LUSC) and cluster_14 represents a lymphocyte activation signature with elevated expression across multiple cancer types, and may be associated with a higher leukocyte fraction (KIRC, LUAD) or tissues of origin related to leukocytes (THYM, LAML) (Fig. 2D) [17, 18]. Additionally, we found that some FGI gene clusters are located within specific chromosomal regions, many with unknown functions (Fig. 2E). These may be from evolutionarily formed gene clusters or tumor amplicons. One interesting cluster is cluster_1, with 7 out of 9 genes localized to Yp11.22. As a result, cluster_1 is highly expressed in prostate cancer (PRAD) and minimally expressed in breast and cervical cancers (BRCA, CESC) as shown in Fig. 2B. Further analysis of 6,390 samples revealed that cluster_1 had significantly higher expression in males than in females (Fig. 2F). Another notable observation is that several HOX gene clusters are highlighted, with cluster_35 mapped to the HOXC gene cluster at 12q13, cluster_36 to the HOXB gene cluster at 17q21, and cluster_41 to the HOXD gene cluster at 2q31. Cluster_35 and cluster_41 are highly expressed in several cancer types, particularly KIRC, while cluster_36 shows elevated expression in COAD and READ. This may suggest the critical role of the HOX gene family in these cancers [19, 20].

Annotation of TCGA Pan-Cancer FGI gene clusters

Based on cluster annotations, we constructed an FGI map at the pan-cancer level (Fig. 3A, Table S1). We categorized gene clusters into three overlapping categories: tissue-specific, functional, and positional, representing tissue-specific clusters, clusters of known functions, and clusters of known chromosomal locations but less functional annotation (Fig. 3B, Table S4). Interestingly, gene clusters on the FGI map appear to group based on their functions. This allowed us to further categorize FGI clusters into eight subcategories such as proliferation, housekeeping, immune-related, and stroma-related, in addition to tissue-specific clusters (Fig. 3A inset, Supplemental Fig. S1A). With a well annotated FGI backbone, we demonstrated the utility of FGI in personalized diagnosis using RNA-Seq, as shown in Fig. 3C. FGI maps were generated from three individual samples of different tumor types. As expected, FGI maps displayed tissue-specific clusters for LIHC, GBM, and LAML. On the other hand, they clearly showed different levels of expression for clusters related to cell cycle, oxidative phosphorylation, and stroma/angiogenesis. This enables us to appreciate the differential activation of cancer-related pathways and tumor microenvironment changes in individual tumor samples and may help tailor therapeutic strategies.

FGI analysis on TCGA lung cancer dataset

We next applied FGI analysis to tumors derived from the same tissue type but exhibiting known heterogeneity in histology. To this end, we used non-small cell lung cancers as a relatively simple model system, including 501 LUAD plus 502 LUSC samples from TCGA. The FGI map was constructed and annotated in a similar manner, yielding 23 functionally annotated clusters, 35 positional clusters, with two gene clusters of overlap (Fig. 4A–C, supplemental Tables S2, S5). The layout of the lung FGI map closely resembled that of the pan-cancer FGI map, with housekeeping gene clusters interconnected in the center, surrounded by other functional domains and scattered positional clusters. Although we did not define "tissue-specific" clusters at this level, we did observe cell type specific clusters in lung tumors. For example, cluster_8 is enriched in genes specific to squamous epithelial cells, cluster_7 is associated with alveolar epithelial cells, and cluster_18 with ciliated epithelial cells, supporting the idea that lung cancers may arise from distinct epithelial progenitors at different developmental stages [21]. Several clusters are clearly linked to LUAD and LUSC subtypes (Supplemental Fig. S2A). This includes cluster_1 (cell cycle genes), cluster 5P (enriched with genes located on chromosome 3q26-29, a region known to be amplified in several squamous cancers) [22], cluster_8 (squamous epithelial related), and cluster_7 (alveolar epithelial related). Unsupervised consensus K-means clustering using eight well-annotated clusters separated LUAD/LUSC into four major subtypes (Fig. 4D). LUAD has two subtypes: one is high in ECM, angiogenesis, and inflammatory response genes, the other is high in ribosomal genes. LUSC also has two subtypes: one with high cell cycle genes and ribosomal genes, the other with high ECM with moderately elevated signals in angiogenesis

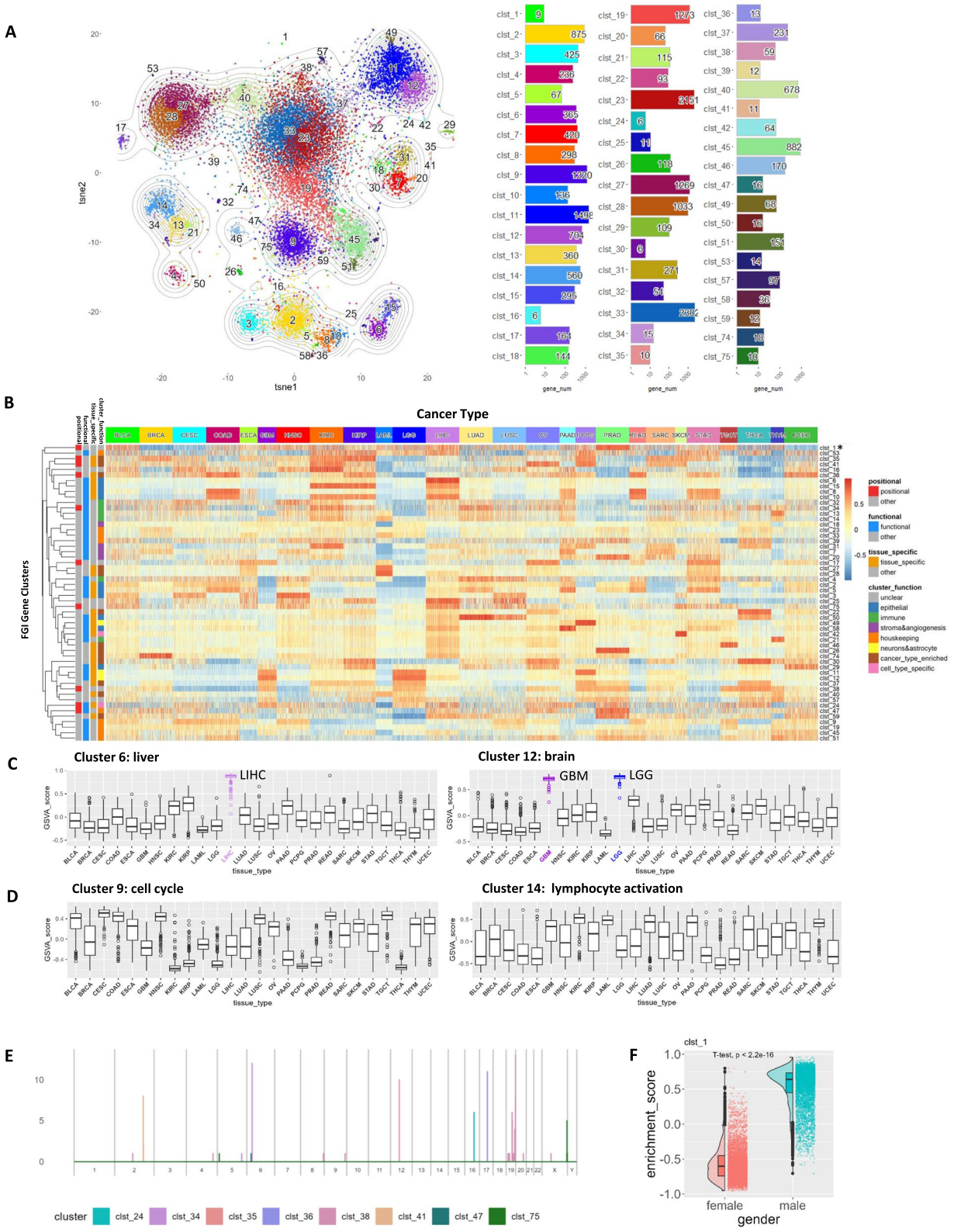

Fig. 2 (See legend on previous page.)

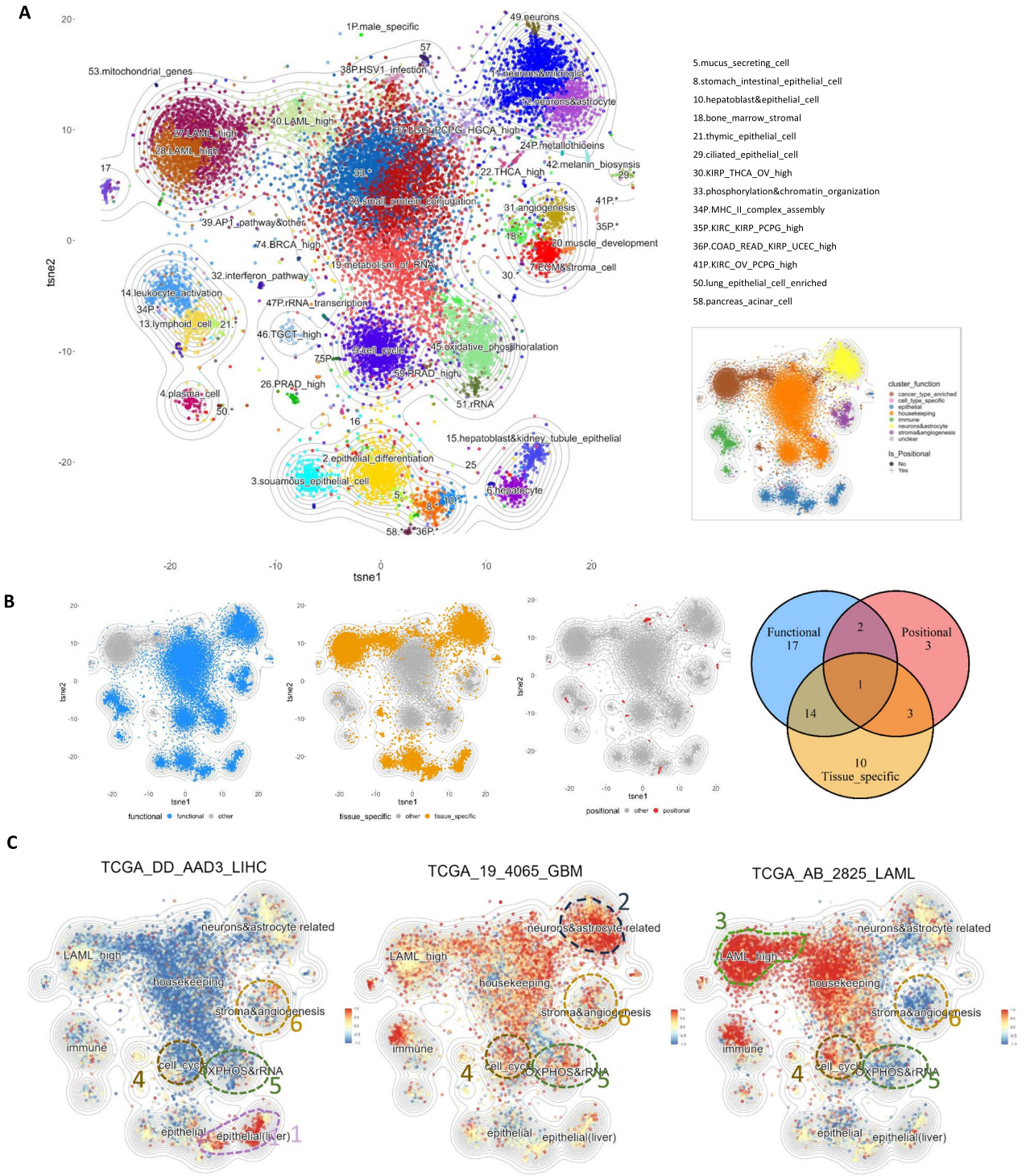

Fig. 3 FGI visualization after gene cluster annotation. **A** t-SNE plot of fully annotated FGI gene clusters from 26 tumor types. Some annotations with long names are put outside the plot. The inset is a small version of Supplemental Fig. S1A where clustered with similar functions were merged. Positional clusters are marked by the cluster number with additional letter "P". **B** Distribution of functional and positional clusters on the FGI map, with a Venn diagram showing overlaps of three types of clusters. **C** Example of FGI mapping applied to three individual cases of tumor samples from different cancer types

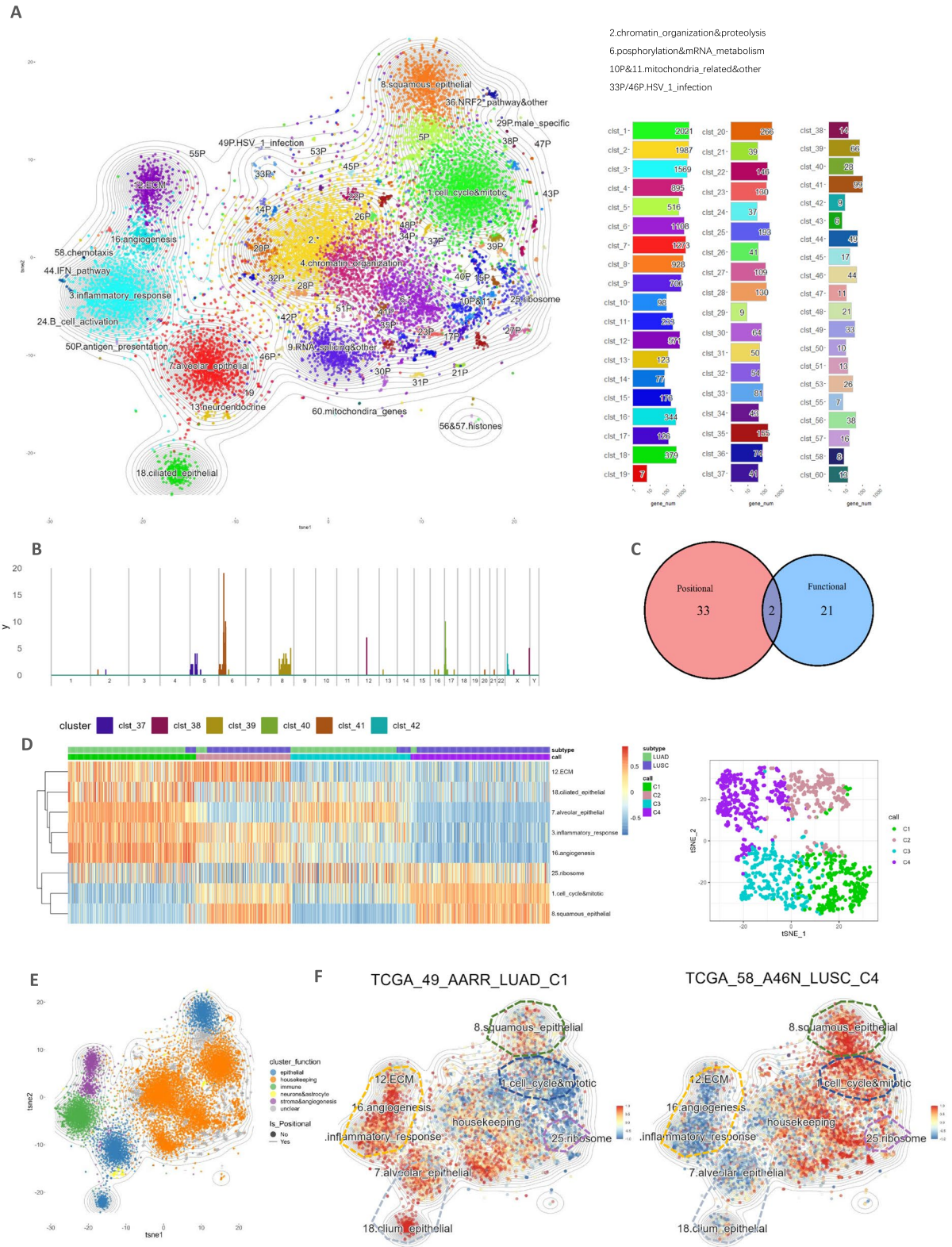

Fig. 4 FGI analysis of TCGA lung cancer RNA-Seq data. **A** t-SNE plot of annotated FGI gene clusters from TCGA LUAD and LUSC RNA-Seq data. Some annotations with long names are put outside the plot. Positional clusters are marked by the cluster number with additional letter "P". **B** Distribution of positional clusters on the chromosome. **C** Venn diagram showing overlap of functional and positional clusters. **D** Consensus k-means clustering of LUAD and LUSC samples using 8 functional gene clusters. **E** FGI map with merged functions (see full image in Supplemental Fig. S1C). **F** Example of FGI mapping applied to two individual cases of tumor samples from LUAD and LUSC

and inflammatory response genes. When FGI maps were generated for individual samples of LUAD and LUSC, we clearly observed differential activation of specific domains of FGI gene clusters corresponding to histology subtypes (Fig. 4E and F, Supplemental Fig. S1B), confirming that FGI may be used as an effective visualization approach to display tumor sources and functions.

FGI analysis on TCGA ovarian cancer dataset

Finally, we applied FGI analysis to TCGA OV dataset, which is primarily composed of high-grade serous carcinomas without refined histology classifications. Notably, among the 71 FGI gene clusters identified, 53 were positional, only 13 clusters were functional annotated (Fig. 5A–C, Tables S3, S6). When consensus K-means clustering was applied with six functional clusters and seven positional clusters, we obtained four OV subtypes (Fig. 5D). Both cluster C1 and C3 show high expression in myeloid and lymphoid cells, whereas C1 exhibits additional activation of genes in many other clusters. C1 cluster shows activation of 8q24 genes, while cluster C2 is immune "cold", with activation of 19q13 genes. Previous studies have suggested the presence of MYC/PVT1 oncogenes in the 8q24 region [23], while the 19q13 region harbors the AKT2 gene, which has been frequently reported as activated in ovarian cancer [24]. When FGI maps were generated for individual OV patient samples, we observed differential activation of specific domains of FGI gene clusters corresponding to FGI derived subtypes (Fig. 5E and F, supplemental Fig. S1C), demonstrating that FGI can be further applied to heterogeneous tumors with unknown histological subtypes.

FGI analysis reveals correlation between HOXA gene cluster expression and clinical stages in ovarian cancer

We further investigated the correlation between FGI output and clinical parameters. One important parameter for tumors is Ki-67 immunohistochemical staining. Unfortunately, this piece of information is not included in TCGA clinical records. We therefore used data from CPTAC, which provides proteomic profiles of tumor samples, including the Ki-67 protein. Pan cancer FGI analysis revealed a good correlation of cluster_9 (cell cycle cluster) with Ki-67 protein levels in most of the cancers where Ki-67 is highly expressed, except for kidney cancer and pancreatic cancer, where Ki-67 expression levels are notably low (Supplemental Fig. S5). Another important parameter is clinical stage, a key indicator of disease progression and treatment planning in OV cancer [25]. Unsupervised analysis of TCGA-OV dataset revealed potential correlation between FGI gene modules and ovarian cancer clinical stages (Fig. 6A volcano plot). To validate this finding, we analyzed an external ovarian cancer gene expression dataset (GSE63885). This confirmed that expression of cluster_24, is correlated with clinical stages (Fig. 6B and C, Supplemental Table S9). Cluster_24 is composed of HOXA family genes, which have been previously reported to be associated with lineage infidelity and cell differentiation in ovarian cancer [26]. At the single gene level, we found the observed statistical differences in cluster_24 could be attributed to HOXA4, HOXA5, and HOXA2 genes (Fig. 6D and Supplemental Fig. S6) in both TCGA-OV and GSE63885 datasets. These results suggest that FGI has the capacity to output results that align well with the medical records of patient samples.

Discussion and Conclusion

Sequencing technology is now commonly used in the clinic and provides us with a deluge of information. Amidst this abundance of 'big data', there is a need to curate and transform these complex datasets into formats that are easily grasped by human comprehension. We aim to develop FGI as a visualization tool to process RNA-Seq data and display encapsulated tumor functions that are easily understood by medical personnel and even patients.

In the field of RNA sequencing data analysis, software developed and reported in literature to assess co-expression gene modules and pathway activation for clinical purposes are rare. An example of such software with similar functions to FGI is "BrainScope" [27]. Brainscope employs a novel dual t-SNE algorithm to visualize co-expression modules significantly enriched for biological functions. In contrast, FGI's t-SNE is based on TOM matrix from WCGNA, which allows FGI to handle more complex datasets. While BrainScope is limited to the gene expression data from normal human brain, FGI can process pan-cancer data across 26 tissue types.

(See figure on next page.)

Fig. 5 FGI analysis of TCGA ovarian cancer RNA-Seq data. **A** t-SNE plot of annotated FGI gene clusters from TCGA OV RNA-Seq data. Some annotations with long names are put outside the plot. Positional clusters are marked by the cluster number with additional letter "P". **B** Distribution of positional clusters on the chromosome. **C** Venn diagram showing overlap of functional and positional clusters. **D** Consensus k-means clustering of LUAD and LUSC samples using 6 functional and 7 positional gene clusters. **E** FGI map with merged functions (see full image in Supplemental Fig. S1B). **F** Example of FGI mapping applied to two individual cases of tumor samples from OV dataset

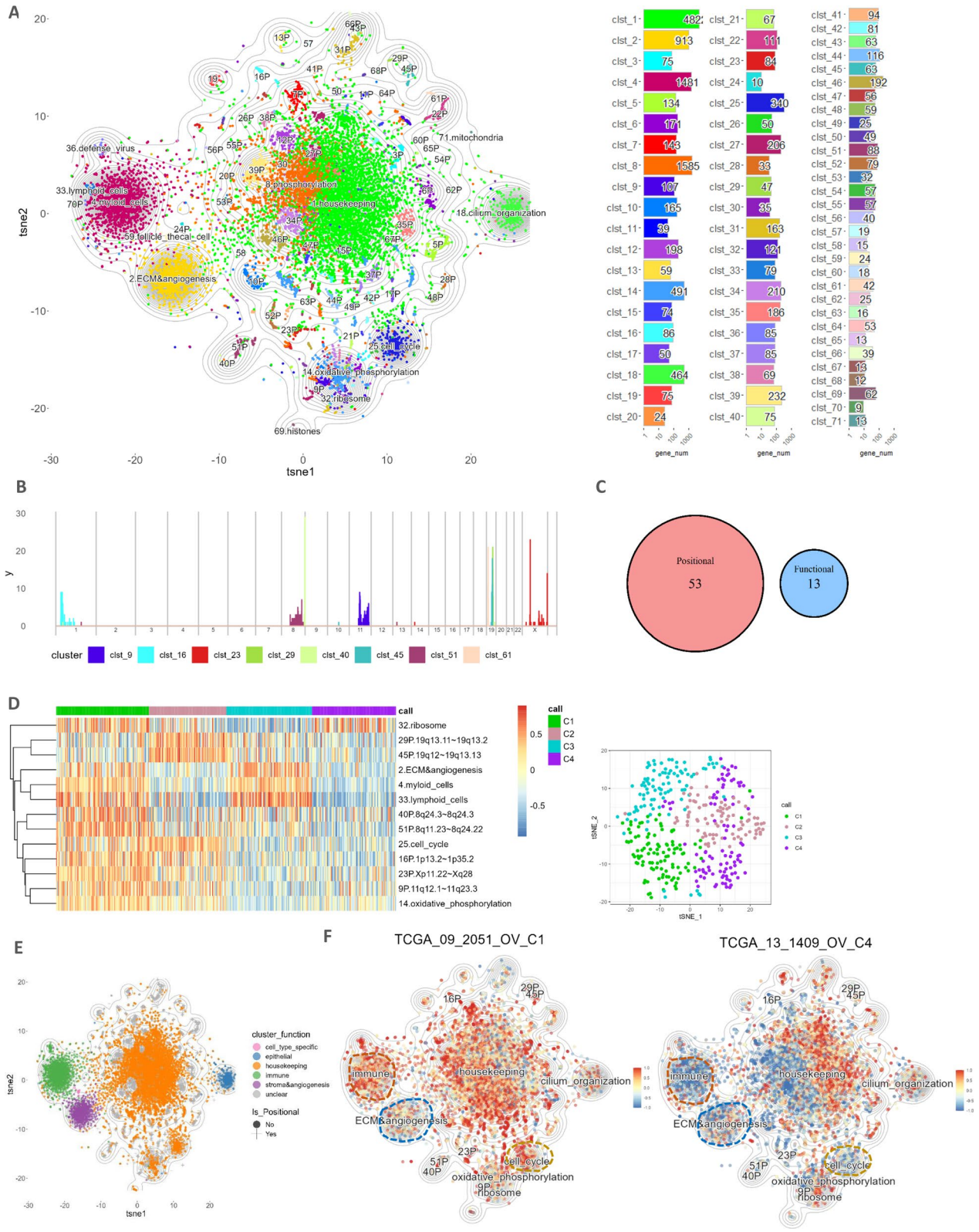

Fig. 5 (See legend on previous page.)

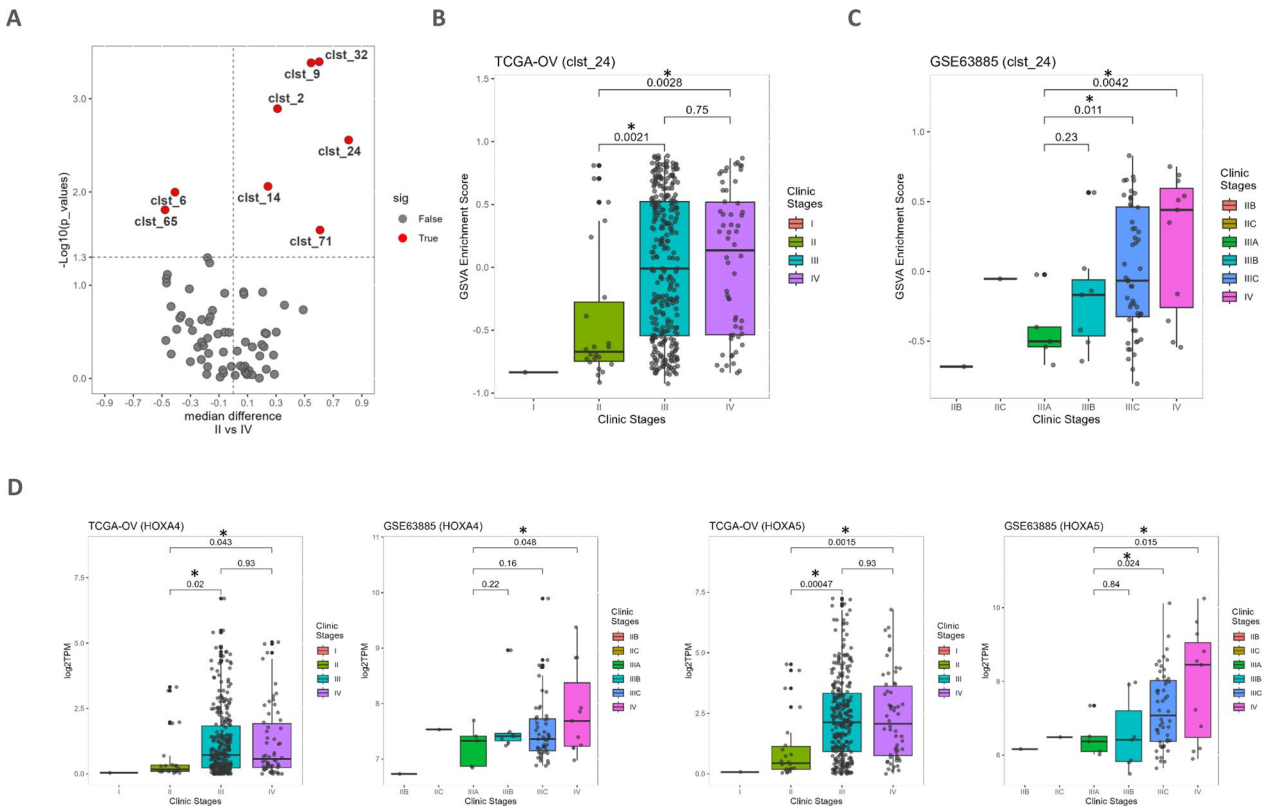

Fig. 6 Correlation between FGI Gene Clusters and Clinical Stage in Ovarian Cancer. **A** A volcano plot illustrating the differences in GSVA enrichment scores between Stage II and Stage IV in TCGA OV cancer, with the x-axis representing the median difference between the two groups. **B** & **C** The GSVA enrichment scores of cluster 24 across different stages in the TCGA and GSE63885 datasets. **D** The RNA expression levels (TPM values) of HOXA4 and HOXA5 in the TCGA and GSE63885 datasets

Our FGI analysis results from both pan-cancer and single-tissue cancer types demonstrated the effectiveness of FGI in elucidating differences among samples. FGI maps can be generated at different tiers to annotate tumor functions. For instance, a lung cancer sample can be viewed from a pan-cancer perspective, a lung cancer view (e.g., LUAD plus LUSC), and a lung cancer subtype view (e.g., LUAD only, which we did not show results here). At the highest tier, FGI can be used to identify or confirm tumor tissue source. At lower tiers, FGI can provide more information on tumor subtypes. Some FGI functional domains are found across all tiers, including pathways in proliferation, oxidative phosphorylation, immune response, angiogenesis, and stroma components. These are intricately linked to therapy decision-making processes.

In our hands, FGI can be used to generate cancer subtypes with meaningful biological functions. Using FGI clusters, we effectively separated lung cancers into LUAD and LUSC subtypes based on both distinct epithelial markers and well-known cellular functions (Fig. 4D). We also separated ovarian cancers into subtypes showing differential activation of genes in clustered chromosomal regions (Fig. 5D). This use of positional clusters may represent a novel approach for defining tumor subtypes.

Many of the positional clusters we identified consist of genes that are located close to each other in chromosomal position, yet their functional annotations remain elusive. Pengxu et al. also observed the enrichment of co-expression modules in certain chromosomal bands through their analysis of multi-cancer expression data by MEGENA pipeline [28]. Based on the distribution on the FGI map of the clusters, we observed that they tend to cluster near housekeeping genes, suggesting a potential co-expression relationship with housekeeping genes. Chromosome arm level gains and focal amplification may be the cause of co-expression and in such cases functional annotation is impossible as target genes are hiding inside by-passengers. Nevertheless, we have identified several positional clusters that can be used to identify subtypes of ovarian cancer samples (Fig. 5D). Exploring the target genes within these chromosomal regions could provide valuable insights into ovarian cancer pathway functions.

In summary, FGI, as an unsupervised method, can effectively reveal key differences within sample populations and demonstrates broad applicability. It provides visualizing co-expression networks within samples and serves as a valuable complement to traditional expression analysis methods. Moving forward, we will focus on refining the analysis techniques and explore its potential clinical applications.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13062-025-00598-y.

Additional file 1.

Additional file 2.

Additional file 3.

Author contributions

Conceptualization: XG and DW; Methodology and Validation: XC, YG, and XG; Manuscript drafting: XC; Manuscript Evaluating & Editing: XC, XG and DW. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

RNA-Seq gene expression profile data were sourced from The Cancer Genome Atlas (TCGA), the Clinical Proteomic Tumor Analysis Consortium (CPTAC), and the Gene Expression Omnibus (GEO). These datasets can be directly downloaded from https://portal.gdc.cancer.gov, https://pdc.cancer.gov/pdc/cptac-pancancer, and https://www.ncbi.nlm.nih.gov/geo. All data used in this research are publicly available. This study complies with the data use and publication guidelines.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The author(s) declared no potential Competing interests with respect to the research, authorship, and/or publication of this article.

Received: 10 September 2024 Accepted: 6 January 2025
Published online: 21 January 2025

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
2. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
3. Laurens VDM, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
4. Laurens VDM. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 2014;15(1):3221–45.
5. 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.
6. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7): e47.
7. Yao F, Coquery J, Lê Cao K-A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinform. 2012;13:1–15.
8. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. 2013;31(6):545–52.
9. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9:559.
10. Zhang Z, Hernandez K, Savage J, Li S, Miller D, Agrawal S, et al. Uniform genomic data analysis in the NCI Genomic Data Commons. Nat Commun. 2021;12(1):1226.
11. Levine JH, Simonds EF, Bendall SC, Davis KL, el Amir AD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97.
12. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.
13. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform. 2011;12:323.
14. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:17.
15. Krijthe JH. Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation. https://github.com/jkrijthe/Rtsne (2015).
16. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinform. 2013;14:7.
17. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. Immunity. 2018;48(4):812-30.e14.
18. Liu Y. A global immune gene expression signature for human cancers. Oncotarget. 2019;10(20):1993–2005.
19. Martinou E, Falgari G, Bagwan I, Angelidi AM. A systematic review on HOX genes as potential biomarkers in colorectal cancer: an emerging role of HOXB9. Int J Mol Sci. 2021;22(24):13429.
20. Wang L, Wang X, Sun H, Wang W, Cao L. A pan-cancer analysis of the role of HOXD1, HOXD3, and HOXD4 and validation in renal cell carcinoma. Aging (Albany NY). 2023;15(19):10746–66.
21. Cheung WK, Nguyen DX. Lineage factors and differentiation states in lung cancer progression. Oncogene. 2015;34(47):5771–80.
22. Voutsadakis IA. 3q26 amplifications in cervical squamous carcinomas. Curr Oncol. 2021;28(4):2868–80.
23. Onagoruwa OT, Pal G, Ochu C, Ogunwobi OO. Oncogenic role of PVT1 and therapeutic implications. Front Oncol. 2020;10:17.
24. Yuan ZQ, Sun M, Feldman RI, Wang G, Ma X, Jiang C, et al. Frequent activation of AKT2 and induction of apoptosis by inhibition of phosphoinositide-3-OH kinase/Akt pathway in human ovarian cancer. Oncogene. 2000;19(19):2324–30.
25. Burges A, Schmalfeldt B. Ovarian cancer: diagnosis and treatment. Dtsch Arztebl Int. 2011;108(38):635.
26. Cheng W, Liu J, Yoshida H, Rosen D, Naora H. Lineage infidelity of epithelial ovarian cancers is controlled by HOX genes that specify regional identity in the reproductive tract. Nat Med. 2005;11(5):531–7.
27. Huisman SM, Van Lew B, Mahfouz A, Pezzotti N, Höllt T, Michielsen L, et al. BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome. Nucleic Acids Res. 2017;45(10):e83.
28. Xu P, Zhang B. Multiscale network modeling reveals the gene regulatory landscape driving cancer prognosis in 32 cancer types. Genome Res. 2023;33(10):1806–17.

Publisher's Note