Genes & Diseases (2023) 10, 2270–2273

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

RAPID COMMUNICATION

Unsupervised machine learning-based clustering identifies unique molecular signatures of colorectal cancer with distinct clinical outcomes

Colorectal cancer (CRC) is known to harbor considerable heterogeneity.[1] Consequently, it could be hypothesized that similar-appearing tumors might exhibit substantial genetic differences while diverse-appearing tumors may have a similar genetic landscape.[2] Due to these differences at the molecular level, they behave or respond differently to therapies as well. CRC progression is a multistep process and involves the accumulation of substantial genetic and epigenetic events in a stage-dependent manner. Alterations in Wnt, DNA repair, RAS-RAF-MAPK, and PIK3CA-AKT pathways have been well-established to play a role in the etiology of CRC. In the era of personalized medicine, it becomes essential to identify the molecular subtype of CRC so that the predictive and prognostic potential of CRC could be established.[3] Molecular subtyping has its importance and limitations in the clinical management of CRC and is very complex to comprehend. Thus, there is still a need to create robust and reliable clustering methods that could precisely identify unique molecular signatures and help in the prediction of the clinical response of the patients who share certain clinical as well as molecular characteristics.[4] Frequent mutations have been reported in RAS, BRAF, NRAS, and PIK3CA genes that are major regulators of the above pathways and significantly promote tumor initiation and progression in CRC. Similarly, the clinical significance of epigenetically deregulated MLH1, RASSF1, DAPK1, IFG2, SLITRK5, and IGFBP3 genes is also well-established and documented in CRC. In the present investigation, we have comprehensively analyzed the mutation status of KRAS, BRAF, NRAS, and PIK3CA, and methylation status of MLH1, RASSF1, SLITRK5, DAPK1, IGFBP3, and IGF2 genes in 70 CRC tumor samples and 40 matched normal, which have a significant role in CRC initiation and progression. Further, the patients were stratified according to their shared molecular heterogeneity and the prognostic value of each heterogeneous cluster was evaluated.

The mutation and methylation status of selected genes were analyzed by AS-PCR and COBRA analysis. The detailed methodology can be found in supplementary data files. The clinicopathological features of collected CRC samples ($n = 70$) are listed in Table S1. Briefly, the median age of the samples was 60.5 years, predominantly from the male population (66%). Stage III tumors were the highest (54%) in number and were mostly derived from the colon (86%). The majority of the samples were of adenocarcinoma histology (75.7%), MMR proficient (48%), negative lymph node status (61%), and moderate tumor grade (76%). The clinical and molecular data were reformatted for the hierarchical clustering analysis. The categorical variables such as stage, age, histology, *etc.* were represented by their specific type while methylation and mutation status of the molecular signature was represented by mutation/methylation = 1 and wild type/unmethylated = 0.

The mutational landscape of KRAS, BRAF, NRAS, and PIK3CA was analyzed for different codons and found to be 36%, 13%, 5.7%, and 17% of the total cohort respectively (Table S1). MLH1, RASSF1, SLITRK5, IGFBP3, DAPK1, and IGF2 genes were profiled for the methylation at promoter region on 70 CRC tumor samples (Fig. 1A). We found that SLITRK5 (57%) and IGFBP3 (40%) were highly methylated followed by IGF2 (34%), RASSF1 (26%), DAPK1 (21%), and MLH1 (4.2%) (Table S2).

Hierarchical clustering was performed for stratifying the patients sharing similar clinicopathological and molecular features. Tightest cluster first tree ordering was applied on each cluster. Hierarchical clustering analysis (considering all important parameters) of the tumor samples generated

Peer review under responsibility of Chongqing Medical University.

https://doi.org/10.1016/j.gendis.2023.01.023

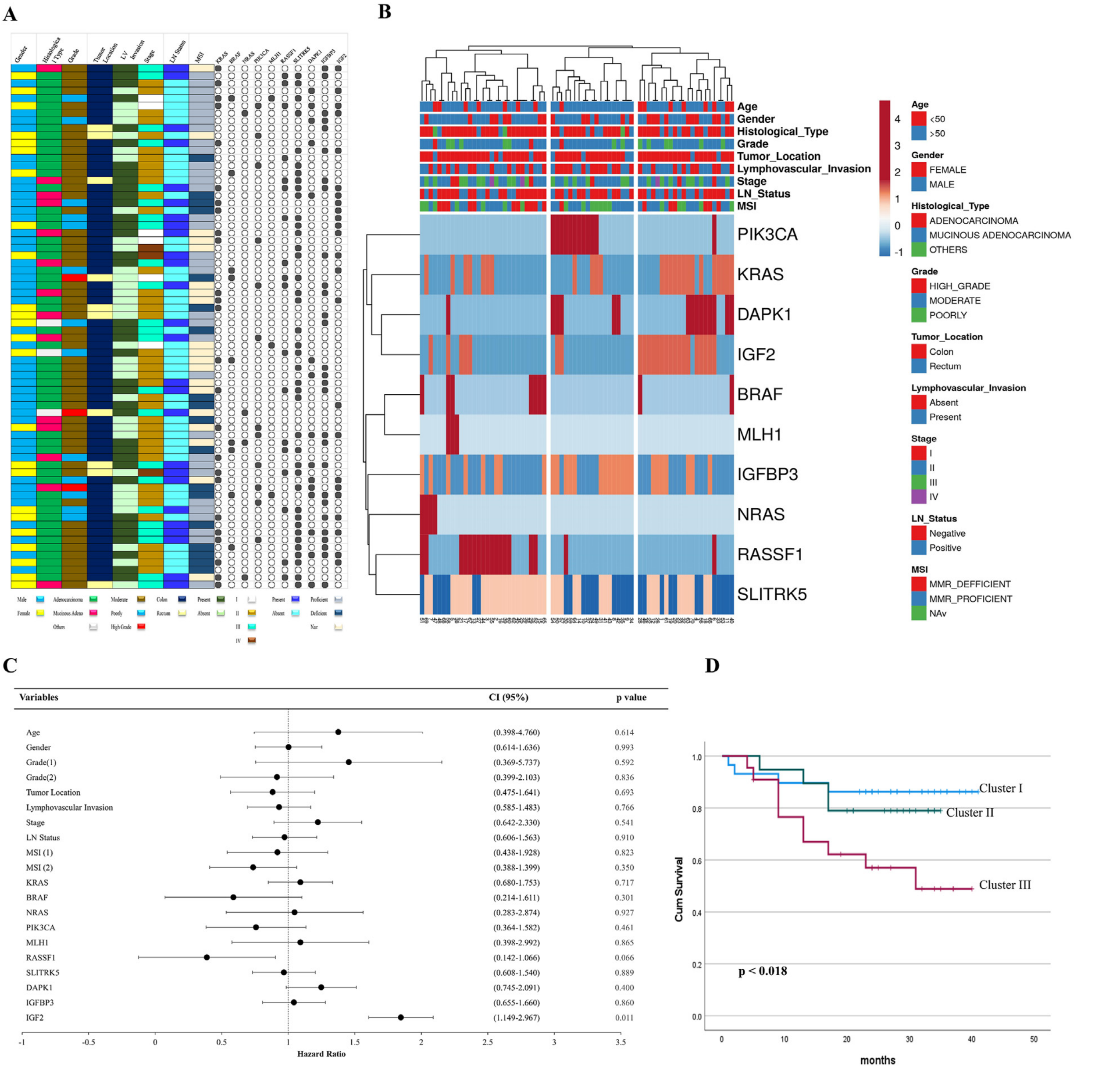

Figure 1 Molecular clustering of colorectal cancer. **(A)** Integrated genomic and clinical variable distribution among 70 colorectal cancer patients. The distribution map contains all the information including pathological features, mutation status, and methylation status of each CRC patient. The map is divided into two basic parts; the first one is for clinical variables and the second one is for genetic signatures. The clinical variables have different color codes for each pathological feature. The genetic signatures were accessed for the mutation and promoter hypermethylation, and mutated/hypermethylated samples in the map were defined by the black circle while wild-type/non-methylated samples were defined by the blank circle. **(B)** Unsupervised hierarchical clustering of 70 tumor samples. Clustering generates three major clusters based on the mutated/wild type or promoter hypermethylated/non-hypermethylated. Unit variance scaling was used for creating the clusters. Both rows and columns were clustered using correlation distance and average linkage. **(C)** Hazard ratio forest plot for the prediction of individual survival prognosis. The forest plot is considered with 95% CI to identify the prognostic role of individual clinical as well as genetic signatures ($P < 0.05$). The middle-dotted line represents the hazard ratio 1, and the hazard ratio above 1 is considered a poor patient outcome. **(D)** Kaplan-Maier survival analysis of the CRC patients. Cluster I patients have good survival compared to cluster II and Cluster III respectively ($P < 0.018$).

three major clusters of the patients with shared divergent features (Fig. 1B). Cluster I was the biggest cluster ($n = 29$, 41.4%) followed by cluster III ($n = 22$, 31.4%), and cluster II ($n = 19$, 27.1%). Demographic association studies within these clusters demonstrated that cluster I (72.4%, $n = 21$) and II (68.4%, $n = 19$) were male-dominated, while the percentage of females was 45.5% ($n = 10$) in cluster III. Age demography showed that cluster II has the highest number of older age patients (94.7%, $n = 18$) and the least below 50-year patients (5.2%, $n = 1$). Clinicopathological association studies within these clusters showed that there were no remarkable differences among different stages (I–IV) and location of the tumors (colon/rectum) of the CRC in all three clusters as they were almost equally distributed. Most of the samples were adenocarcinoma ($n = 53$) and no clear distinction was seen however cluster I had fewer tumors of mucinous histology (10.3%, $n = 10$). MMR deficient (37.9%, $n = 11$) and poor tumor grade were more specifically present in cluster I and III, respectively (Table S3).

Unique sets of genetic and epigenetic alterations were frequently distributed among all clusters and the prime aim of these clusters was to evaluate the impact of genetic and epigenetic alteration in a group of genes in disease monitoring as well as in clinical response. Cluster I, the largest cluster, shared remarkable features; for example, tumors were predominately methylated for SLITRK5 (75.8%, $n = 22$), RASSF1 (55.1%, $n = 16$), exclusively methylated for MLH1 (10.3%, $n = 3$) gene and were rich in mutated NRAS (13.7%, $n = 4$). BRAF (24.1%, $n = 7$) mutation was also present at high frequency in cluster I while KRAS, IGF2, and IGFBP3 genes were sparsely found to be methylated or mutated in this cluster. Cluster II shares three important features: exclusively mutated for PIK3CA (57.8%, $n = 11$), frequently methylated for IGFBP3 (68.4%, $n = 13$), and predominantly MMR proficient (57.8%, $n = 11$) status. Moreover, demographically older age patients >50 years (94.7%, $n = 18$) were more common in cluster II. Apart from these features, cluster II primarily had KRAS (21%, $n = 4$) mutation and DAPK1 (26.3%, $n = 5$), SLITRK5 (42.1%, $n = 8$) methylation to a significant extent. In comparison with other clusters, cluster III had a higher number of females (45.5%, $n = 10$) with no high grade of differentiation. Mutation in KRAS (63.3%, $n = 14$) and methylation in DAPK1 (40.9%, $n = 9$) and IGF2 (77.2%, $n = 17$) were frequently present in cluster III while no MLH1 methylation and NRAS mutation were seen in samples of cluster III. This cluster was, however, also mutated for PIK3CA and BRAF and methylated for IGFBP3 and SLITRK5 to some extent (Table S4).

All the clinical and molecular features were analyzed for univariate Cox regression analysis to establish the prognostic significance of these markers in clinical management. In our study, some molecular markers were found to be prognostically significant; however, no clinicopathological characteristics were reported to have any significant prognostic importance. As displayed in the forest plot, IGF2 methylation ($P < 0.011$) was associated with a poor outcome while RASSF1 methylation ($P < 0.06$) was associated with a good outcome. Similarly, some other features like age, tumor grade, tumor stage, and DAPK1 methylation were indicative of poor survival of the patients; however, their association was found to be not significant (Fig. 1C).

Survival data for 70 colorectal cancer patients were collected and compiled for the survival analysis. The mean disease free survival (DFS) for all patients was 33.26 months (95% CI = 30.1–36.4). Cumulative two-year survival for all three clusters was 86.2%, 78.9%, and 54.5% in cluster I, cluster II, and cluster III, respectively, while cumulative overall survival for all clusters was 74.3% in our cohort. Comparative analysis of all clusters shows that patients in cluster III had poorer DFS (mean DFS = 27.0 months, 95% CI = 20.9–33.1) and were dominated by KRAS mutation (63.3%, $n = 14$), and DAPK1 (40.9%, $n = 9$) and IGF2 (77.2%, $n = 17$) methylation. It was followed by cluster II which had a marginally better DFS than cluster III (mean DFS = 30.4 months, 95% CI = 26.3–34.5). Cluster II displayed IGFBP3 methylation and PIK3CA mutation with high MMR. Conspicuously patients of cluster I displayed a better DFS than those of other clusters (mean DFS = 36.3 months, 95% CI = 32.0–40.6) and shared salient molecular features such as hypermethylated genes for SLITRK5 (75.8%, $n = 22$), RASSF1 (55.1%, $n = 16$), and MLH1 (10.3%, $n = 3$) and mutated for NRAS (13.7%, $n = 4$) and BRAF (24.1%, $n = 7$) (Fig. 1D and Table S4). Cluster-specific multivariate analysis considering all variables shows the test was significant ($P < 0.001$) and the significance of the clinical and genetic variables present in the clusters is represented in Table S3 and 4.

Molecular subtype-based therapies are still in the naïve phase while several efforts have been made in the last five years to involve molecular subtyping in routine practices.[5] Our study faces certain limitations including a small number of tumor samples but provides a primary comprehensive overview of the shared genetic, epigenetic, and clinical features of sporadic CRC. Clustering-based clinical outcome is managed by the prime players (clinical and genetic factors) of the subgroups. This study suggests that KRAS mutation and DAPK1 methylation is still a major concern in the clinical management of CRC with poorer outcome. Hierarchical clustering appears to be a helpful, promising, and effective technique for further translational research, leading to the characterization of a diagnostic and prognostic signature for CRC.

Ethics declaration

Ethical approval for the present study has been granted by the Institute and hospital ethics committee (Ref. no. IEC17-18/027). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Author contributions

SS and NKS contributed to the conception/design of the study. MPS, SR, and SKG performed the experiments. All authors contributed to data acquisition and manuscript preparations.

Funding

The study was funded by the Department of Biotechnology India section order 6242-P103/RGCB/PMD/ DBT /SMSV/ 2015.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request and provide the accession no. of the data set once available.

Acknowledgements

The authors are thankful to the Director Motilal Nehru National Institute of Technology Allahabad (Prayagraj, India) for providing a research facility and Dr. Juhi Tayal (Pathologist and Coordinator Biorepository, Rajiv Gandhi Cancer Institute and Research Centre, New Delhi) for providing tumor samples and clinical data. The authors are thankful to the Ministry of Human Resource and Development, Govt. of India, New Delhi, India, for providing scholarships during this tenure.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2023.01.023.

References

1. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: molecular basis of colorectal cancer. *N Engl J Med*. 2009; 361(25):2449—2460.
2. Purcell RV, Schmeier S, Lau YC, et al. Molecular subtyping improves prognostication of Stage 2 colorectal cancer. *BMC Cancer*. 2019;19:1155.
3. Sadanandam A, Wang X, de Sousa E Melo F, et al. Reconciliation of classification systems defining molecular subtypes of colorectal cancer. *Cell Cycle*. 2014;13(3):353—357.
4. Singh MP, Rai S, Pandey A, et al. Molecular subtypes of colorectal cancer: an emerging therapeutic opportunity for personalized medicine. *Genes Dis*. 2021;8(2):133—145.
5. Alwers E, Jia M, Kloor M, et al. Associations between molecular classifications of colorectal cancer and patient survival: a systematic review. *Clin Gastroenterol Hepatol*. 2019;17(3): 402—410.

Manish Pratap Singh [a,b], Sandhya Rai [a], Sarvesh K. Gupta [a], Nand K. Singh [a], Sameer Srivastava [a,*]

[a] *Motilal Nehru National Institute of Technology Allahabad, Prayagraj 211004, India*
[b] *Deen Dayal Upadhyay Gorakhpur University, Gorakhpur 273001, India*

*Corresponding author.
E-mail addresses: manish.biophd@mnnit.ac.in (M.P. Singh), sameers@mnnit.ac.in (S. Srivastava)

4 July 2022
Available online 24 March 2023