DNA Methylation Analysis Identifies Clinically Relevant Lung Adenocarcinoma Subgroups

Oluwadamilare I. Afolabi [1 ‡]

[1] ARL Bio, Ltd., Lagos, Nigeria
[‡] **Correspondence** dare.afolabi@outlook.com

Abstract

Lung adenocarcinoma (LUAD) is the most prevalent subtype of lung cancer and is characterized by significant molecular heterogeneity and poor prognosis, primarily due to late-stage diagnoses. Therefore, detailed molecular characterization of LUAD is crucial for developing biomarkers to accurately detect the disease in its early stages. This study investigates the role of DNA methylation in LUAD, emphasizing its potential as a biomarker for cancer detection and as a tool for understanding tumor biology. The study identified 4,925 differentially methylated sites (DMSs) and prioritized the top 200 DMSs for downstream analyses. Functional enrichment analysis revealed that site-specific hypermethylation in exon 1 and distal promoter regions are linked to critical developmental processes, including morphogenesis, pattern specification, stem cell differentiation, and synaptic transmission, suggesting that these epigenetic changes may disrupt normal cellular functions and contribute to tumorigenesis. Support vector machines demonstrated the diagnostic potential of these hypermethylated sites, achieving perfect classification of LUAD and normal adjacent tissues with as few as five features. Additionally, the strong correlation between methylation levels and feature importance scores further explained the predictive accuracy of these methylation markers. The study also identified distinct methylation subgroups within LUAD tumors, independent of traditional staging, each associated with unique transcriptional dysregulation and biological processes, such as DNA repair, immune response, and ribosome biogenesis. These findings not only enhance our understanding of LUAD pathophysiology but also underscore the clinical utility of DNA methylation as a diagnostic tool and guide for patient management.

Keywords: Lung cancer, Cancer genomics, Biomarkers, Methylation

Introduction

Lung cancer remains the most prevalent and lethal malignancy, with non-small cell lung cancer (NSCLC) constituting approximately 85% of lung cancer cases [1, 2]. Despite advances in therapeutic strategies, the five-year survival rate for lung cancer remains low, underscoring the need for improved diagnostic and therapeutic approaches. Among NSCLC subtypes, lung adenocarcinoma (LUAD) is the most prevalent and is particularly challenging to treat due to its marked heterogeneity [3]. This heterogeneity represents a broad spectrum of molecular and histopathological variations, contributing to differences in tumor behavior and responses to therapy [4, 5]. Therefore, detailed molecular characterizations of LUAD are crucial for improved diagnosis and prediction of disease progression, enabling the development of more effective, tailored treatment strategies.

DNA methylation, specifically the addition of a methyl moiety to the cytosine residues of cytosine-guanine (CpG) dinucleotides, has emerged as a critical player in cancer biology. Regulatory elements within exons, untranslated regions (UTRs), promoters, and other genomic regions are important methylation sites, directly controlling gene expression at the transcriptional level [6, 7]. Aberrant methylation of these regions often results in the silencing of tumor suppressors or the

activation of oncogenes, contributing to tumorigenesis and cancer progression. These epigenetic changes have been implicated in tumor complexity, aggressiveness, and therapeutic potential [8, 9, 10]. Thus, DNA methylation not only serves as an indicator of disease state but also represents an actionable target for therapy, making it invaluable for diagnostic and therapeutic applications.

This study leveraged array-based methylation data from The Cancer Genome Atlas (TCGA) to explore the DNA methylation landscape of LUAD. The study assessed the potential of the methylation statuses of exon 1 and distal promoter regions as biomarkers for disease detection and classification. The study findings highlight distinct methylation signatures correlating with LUAD subgroups, reflecting underlying biological differences that could be harnessed for personalized therapies. This work not only furthers our understanding of LUAD pathophysiology but also opens new avenues for developing methylation-based diagnostic tools, offering hope for improved management of this challenging disease.

Materials and Methods

Data acquisition and preprocessing

DNA methylation array data for 175 subjects diagnosed with LUAD (dbGaP accession phs000178) were retrieved from the TCGA database [11]. The dataset comprised genome-wide methylation profiles generated using the Illumina HumanMethylation450 BeadChip array, covering over 450,000 CpG sites across the human genome [12]. Matched normal adjacent tissue samples were also obtained for 31 of these subjects. The clinical characteristics and risk exposures of subjects in the TCGA LUAD cohort are summarized in Table 1. The absence of confounding factors was confirmed using T-tests for continuous variables and Chi-squared tests for discrete variables (Table 1).

Table 1. Clinical features of subjects in the TCGA LUAD cohort.

Group	Tumor (N = 175)	Normal Adjacent (N = 31)	Statistic	p-value
Age				
33.34 - 87.11	64.18 ± 10.3	64.53 ± 10.05	t = 0.179	0.858
Sex				
Female	105 (60%)	19 (61%)	$\chi^2 = 0$	1
Male	70 (40%)	12 (39%)		
Stage				
Stage I	102 (58%)	17 (55%)	$\chi^2 = 1.889$	0.756
Stage II	39 (22%)	8 (26%)		
Stage III	21 (12%)	2 (6%)		
Stage IV	10 (6%)	3 (10%)		
Unreported	3 (2%)	1 (3%)		
Cigarettes per Day				
0 - 8.44	1.72 ± 1.69	1.61 ± 1.67	t = -0.328	0.745

Differential methylation analysis

To identify CpG sites differentially methylated in LUAD, a subset comprising one-third of the LUAD cohort was randomly selected (Table S1) and analyzed using DESeq2 v1.42.1 [13]. DESeq2, widely

used for the analysis of count-based data to identify differentially expressed genes, was adapted in this study to identify differentially methylated sites (DMSs). Before differential methylation analysis, CpG sites with consistently low methylation levels, specifically those with β-values less than 0.1 in all samples of the smallest sample group, were filtered out to minimize noise. The 200 most differentially methylated sites, based on absolute fold change and a false discovery rate (FDR) adjusted p-value threshold of 0.01, were prioritized for further analyses. The overall methylation changes of these DMSs were evaluated using a volcano plot, which depicts the relationship between fold change and statistical significance. Generally, volcano plots provide a clear visualization of the extent and direction of methylation differences. Additionally, the distribution of DMSs across various genomic regions was compared using a bar plot to assess which genomic regions were more prone to differential methylation in LUAD. Furthermore, samples were clustered by methylation patterns using hierarchical clustering with Ward's method and visualized using heatmaps [14]. Ward's method minimizes the total within-cluster variance, ensuring that the resulting clusters are as homogeneous as possible [15], allowing for a clear representation of the relationships between samples based on their methylation profiles.

Functional enrichment analyses

DMSs were first linked to their associated gene. The DMS-associated genes were then stratified by the genomic location of their DMSs and subjected to Gene Ontology (GO) enrichment analysis using the clusterProfiler v4.10.1 [16]. This analysis aimed to identify biological processes significantly overrepresented among the genes associated with DMSs in specific genomic regions. A q-value threshold of 0.05 was applied to minimize the likelihood of false positives. GO enrichment analysis was performed on genes uniquely dysregulated in methylation clusters identified during the study to further explore the functional implications of methylation patterns. This analysis provided insights into the distinct biological processes that may be driving the behavior of each methylation-defined LUAD subgroup. The top 10 enriched GO processes were visualized using dot plots for all functional enrichment analyses.

Machine learning

LUAD cohort samples were split into training (60%) and test (40%) sets to build and evaluate predictive models. Support vector machine (SVM) models were trained on the training set using various feature sets, specifically the most methylated 1, 5, 10, or all exon 1 or distal promoter DMSs [17]. Several key metrics were extracted using caret v6.0-94 [18] and pROC v1.18.5 [19] to evaluate the performance of these models on the test set. Detailed breakdowns of true positives, true negatives, false positives, and false negatives were visualized in a confusion matrix. ROC curves were further generated for each model to assess their classification performance. Area Under the Curve (AUC) values were reported with 95% confidence intervals (CI) as a summary measure of model accuracy, providing a robust metric for evaluating the performance and reliability of the SVM models. Using a random forest algorithm [20], importance scores, quantifying each DMS's contribution to reducing prediction uncertainty across all decision trees, were calculated. These importance scores were then correlated with average methylation levels of the corresponding DMSs to explore potential relationships. Correlation was assessed by fitting a linear regression model, with Pearson's correlation coefficients (R) and p-values reported.

Gene expression analysis

To detect gene dysregulation within LUAD methylation subgroups, gene expression quantification data for four subjects per methylation cluster (dbGaP accession phs000178) were retrieved from the TCGA database [11]. Genes with less than ten counts in all samples of the smallest sample group were filtered out to reduce noise. Differentially expressed genes (DEGs) were then identified between the methylation clusters 2, 3, 4, and 5 and cluster 1 (normal adjacent tissue) samples at an FDR-adjusted p-value threshold of less than 0.05. To further emphasize the uniqueness of these clusters, overlaps in the sets of upregulated and downregulated genes among clusters 2, 3, 4, and 5 were visualized using upset plots [21]. This visualization approach allowed for a clear representation of shared and unique dysregulated genes across the clusters, highlighting both common pathways and cluster-specific gene expression changes that may contribute to the distinct biological behaviors of each subgroup.

Results

Aberrant hypermethylation in LUAD

The differential methylation analysis revealed that 2,875 CpG sites (24%) were hypermethylated, while 2,050 sites (17%) were hypomethylated in LUAD tumors compared to normal adjacent tissue samples (Table S2). For downstream analyses, the top 200 differentially methylated sites (DMSs) by fold change and FDR-adjusted p-values were selected, all of which exhibited above four-fold hypermethylation (Fig. 1A). Notably, the majority of these top DMSs were located in exon 1 (55 sites), closely followed by the distal promoter (TSS1500) region (46 sites). In contrast, the 3' UTR (2 sites) had the fewest top DMSs (Fig. 1B).

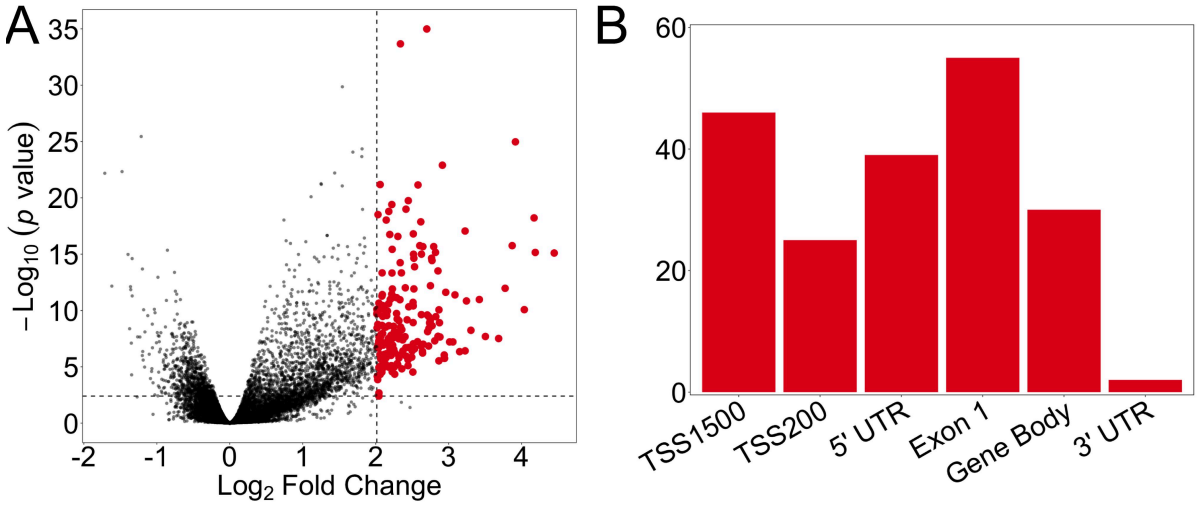

Fig. 1. Differential methylation in LUAD. (**A**) Volcano plot depicting DMSs in LUAD versus NAT samples. The top 200 DMSs, based on absolute fold change and an FDR-adjusted p-value threshold of 0.01, are highlighted in red. (**B**) Bar plot illustrating the distribution of hypermethylated CpG sites across various genomic regions.

Functional enrichment analysis of DMSs in LUAD revealed several biological processes significantly associated with hypermethylation in different genomic regions (Table S3). The analysis highlighted that genes near hypermethylated CpG sites in exon 1 regions are involved in critical developmental processes such as pattern specification process, stem cell differentiation, and regulation of neuron differentiation (Fig. 2A). These processes suggest that the hypermethylation observed may contribute to the disruption of normal development and differentiation pathways in LUAD, potentially contributing

to tumorigenesis. In 5′ UTRs, significant enrichment was also observed for processes related to morphogenesis (Fig. 2B). This could indicate a link between aberrant methylation in these regulatory regions and the altered expression of genes crucial for proper development, which might contribute to the abnormal growth in LUAD tumors. For DMSs in distal promoter regions, the enriched biological processes were similarly associated with the morphogenesis of branching structures, but also the regulation of synaptic potential and chemical synaptic transmission (Fig. 2C). The involvement of these processes underscores the potential role of epigenetic modifications in altering epithelial and neural signaling pathways, which could influence tumor behavior.

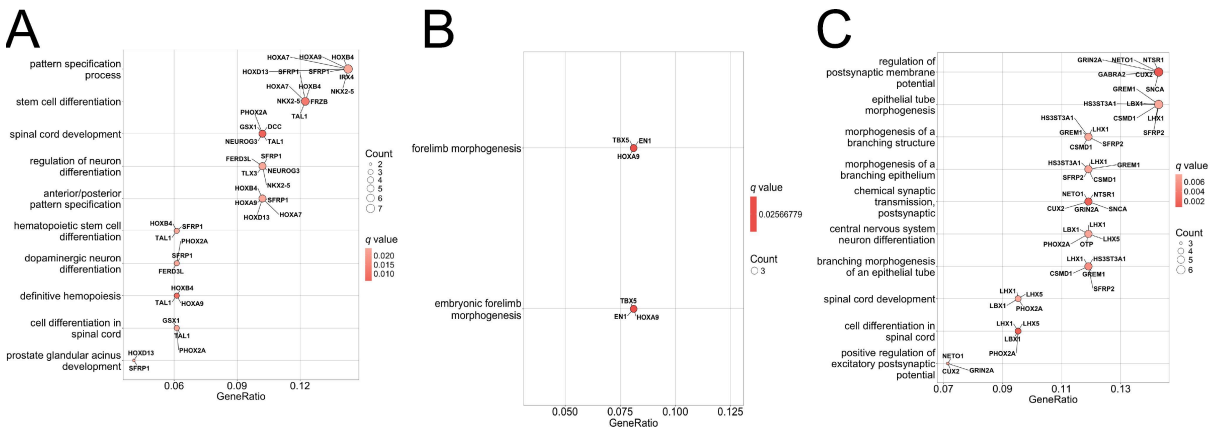

Fig. 2. Biological processes associated with hypermethylation in LUAD. Dot plots highlighting biological processes associated with differential methylation of (**A**) exon 1, (**B**) 5′ UTR, and (**C**) distal promoter regions.

Diagnostic potential of exon1 and distal promoter methylation

Since exon 1 and distal promoter region, hypermethylation appeared to be the most biologically relevant, these regions were prioritized for further analyses. Heatmaps of the methylation levels of these DMSs, with samples sorted vertically by average CpG site methylation and CpG sites sorted horizontally by average sample methylation, were generated. These heat maps confirmed that the selected CpG sites were hypermethylated exclusively in tumor samples, suggesting their potential as biomarkers for LUAD (Fig. 3A-B). Some CpG sites exhibited more consistent hypermethylation within tumor samples than others, indicating a complex yet patient-specific methylation pattern.

The efficacy of these DMSs as classifiers was validated using support vector machine (SVM) models, which were trained on varying numbers of DMSs from exon 1 (Fig. 3C) and distal promoter (Fig. 3D) regions. Remarkably, just five DMSs were sufficient to achieve perfect classification of samples into tumor and normal adjacent tissue. A model trained on a single exon 1 DMS with the highest average methylation (cg26521404 in the HOXA9 gene) achieved an AUC of 0.99 (95% CI, 0.97–1.00) (Fig. S1A). However, the model trained on the distal promoter DMS with the highest average methylation (cg22660578 in the LHX1 gene) yielded a slightly lower AUC of 0.94 (95% CI, 0.89–0.99) (Fig. S1B). Furthermore, random forest feature importance scores correlated significantly with average methylation levels for both exon 1 ($R = 0.509$, $p < 0.001$) and distal promoter regions ($R = 0.418$, $p < 0.004$), underscoring the impact of consistent methylation at these sites on prediction accuracy (Fig. 3E-F).

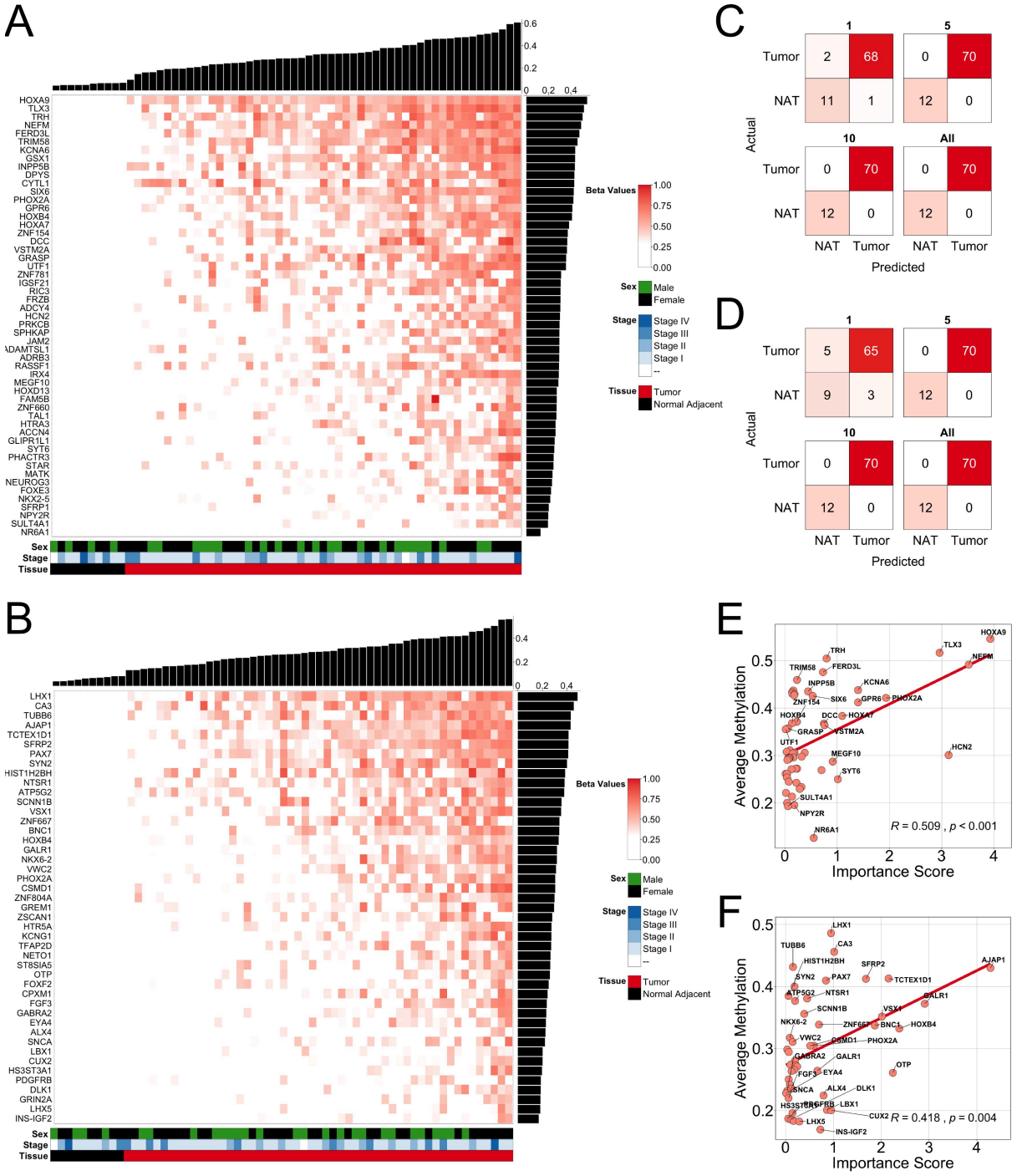

Fig. 3. Exon 1 and distal promoter methylation as classifiers for LUAD. Heatmaps showing methylation patterns in NAT and LUAD samples for (**A**) exon 1 and (**B**) distal promoter regions. Vertical bar plots represent average methylation levels across samples, while horizontal bar plots show average CpG site methylation levels. Confusion matrices depicting the performance of SVM models trained and tested on 1, 5, 10, and all DMSs from (**C**) exon 1 and (**D**) distal promoter regions. Correlation plots of average methylation and feature importance for (**E**) exon 1 and (**F**) distal promoter DMSs in random forest models.

LUAD tumors cluster into distinct methylation subgroups

Unsupervised hierarchical clustering of tumor and normal adjacent tissue samples based on methylation patterns in exon 1 and distal promoter regions identified five distinct methylation subgroups (Fig. 4). Cluster 1 exclusively comprised of normal lung tissue samples, while Clusters 2-5 were comprised entirely of tumor tissue samples. Cluster 2 had the lowest proportion of male patients

(20%), while Cluster 5 had the highest (68.75%). Cluster 3 exhibited the least smoking history among patients and showed the lowest levels of methylation, suggesting a potential link between smoking and overall methylation levels. There was no apparent clustering of samples based on cancer stages, indicating that methylation subgroups may be independent of traditional staging.

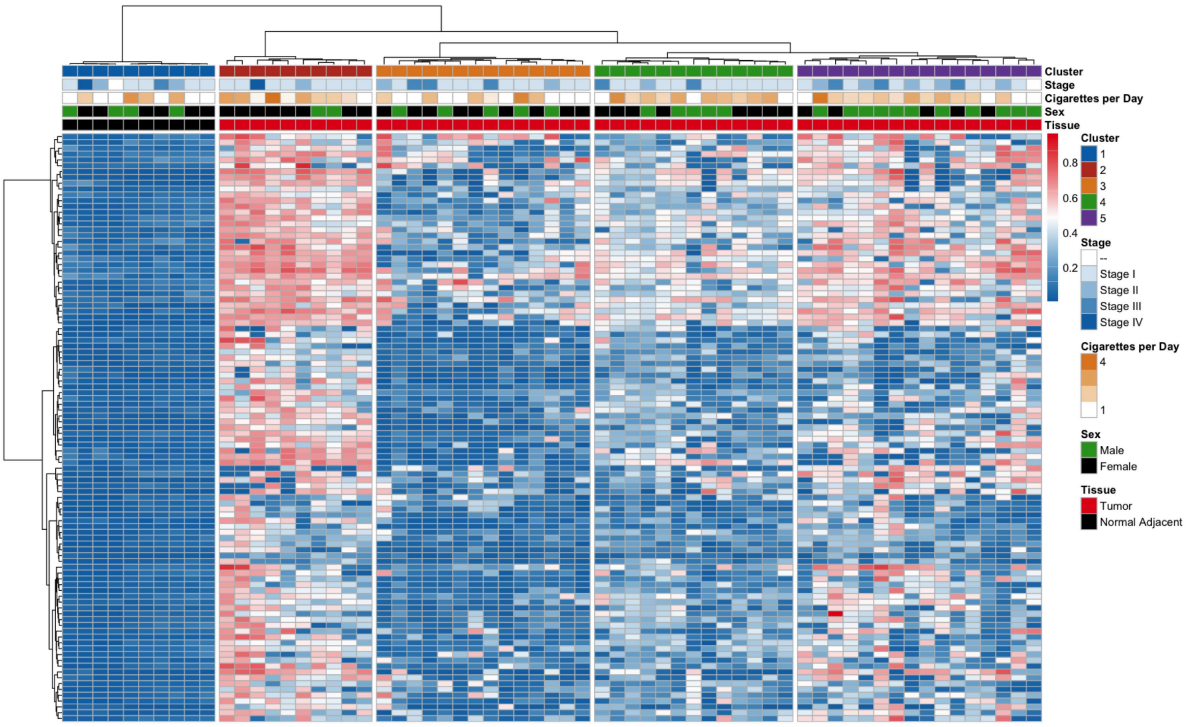

Fig. 4. Methylation subgroups identified in LUAD. Heatmap of NAT and LUAD samples clustered by methylation patterns in exon 1 and distal promoter regions.

Gene dysregulation within methylation subgroups

Gene expression analysis across the methylation-defined LUAD subgroups revealed significant dysregulation patterns specific to each cluster. Fig. 5A and 5B illustrate the overlaps of genes that are either upregulated or downregulated across Clusters 2, 3, 4, and 5, compared to normal adjacent tissues. Cluster 3 exhibited a distinct profile with the highest number of uniquely upregulated genes, which may be attributed to its relatively lower methylation levels. Enrichment analyses further identified distinct biological processes associated with the differentially expressed genes within these clusters (Fig. 5C-F). For example, Cluster 3 was strongly associated with DNA repair processes, particularly those involved in homologous recombination and replication fork processing (Fig. 5D). In contrast, Cluster 4 was enriched for immune-related processes, such as leukocyte activation and chemotaxis (Fig. 5E). Additionally, Cluster 5 displayed enrichment in processes related to ribosome biogenesis and translation, indicating high transcriptional activity (Fig. 5F). Conversely, Cluster 2 was associated with only one biological process, suggesting these tumors are less distinct in their transcriptional profile compared to the other clusters (Fig. 5C). These findings underscore the molecular heterogeneity of LUAD, with most methylation subgroups exhibiting distinct transcriptional programs, which could be leveraged for more tailored therapeutic strategies.

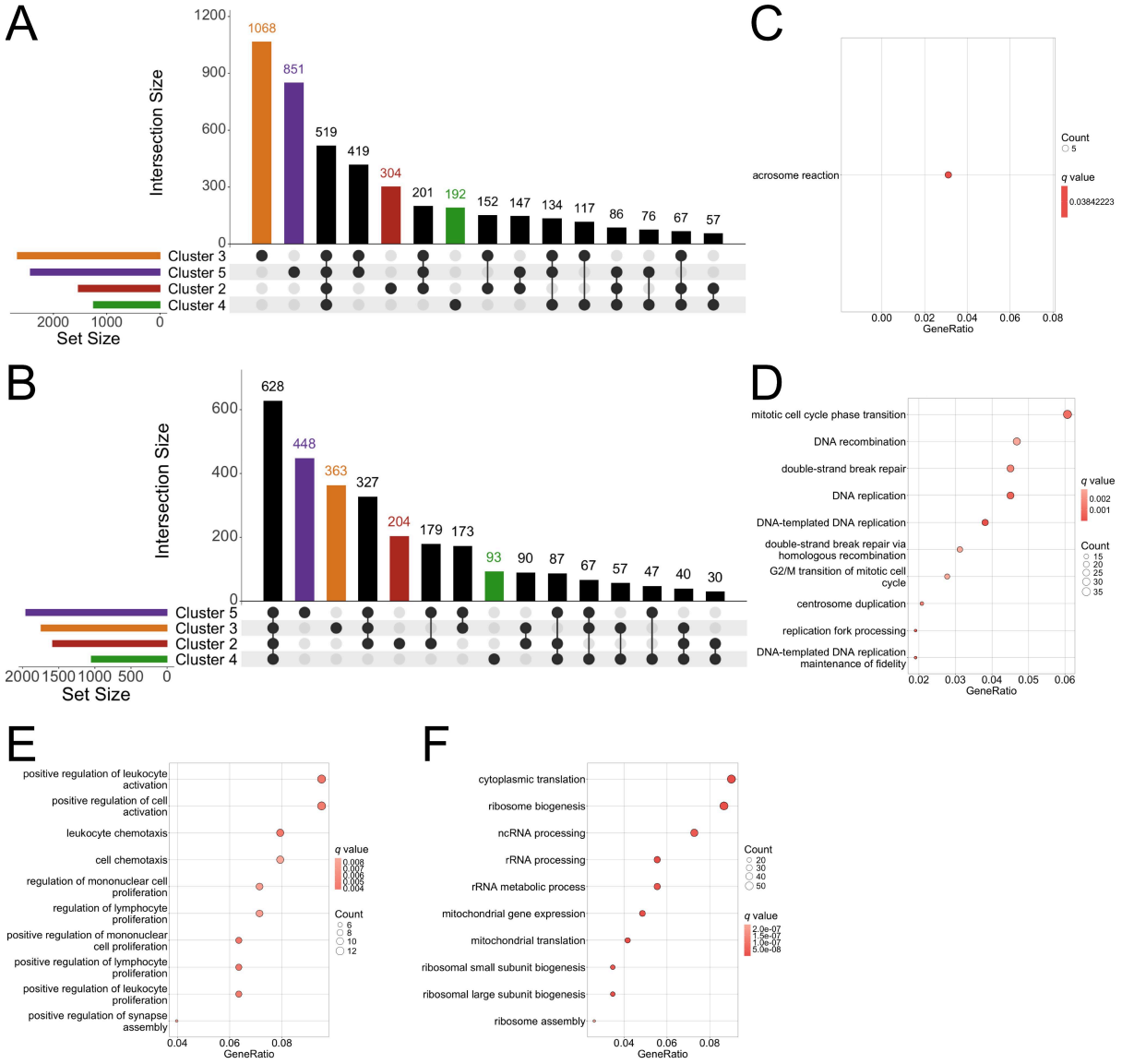

Fig. 5. Biological processes associated with LUAD subgroups. Upset plots showing overlaps of (**A**) upregulated and (**B**) downregulated genes in clusters 2, 3, 4, and 5 versus NAT samples. (**C**-**F**) Dot plots highlighting biological processes associated with clusters 2, 3, 4, and 5, respectively.

Discussion

Lung cancer remains the deadliest malignancy worldwide, primarily due to late-stage diagnoses [1, 22]. The challenge of early detection is exacerbated by the heterogeneous nature of lung cancer, which often leads to diagnostic delays and suboptimal treatment outcomes [23]. Despite therapeutic advances, the prognosis for lung cancer patients remains poor, underscoring the need for biomarkers capable of accurately detecting the disease in its early stages. Several studies have highlighted DNA methylation's promise as a cancer detection biomarker [24, 25, 26]. This study's findings emphasize the critical role of DNA methylation in LUAD and demonstrate the potential of high-resolution DNA methylation profiling for accurate cancer detection and characterization. This study identified many CpG sites with aberrant hypermethylation in LUAD tumors, particularly in exon 1 and distal promoter regions. The overrepresentation of hypermethylated CpG sites in these regions suggests they may play a crucial role in modulating gene expression, thereby contributing to LUAD pathogenesis.

Functional enrichment analysis of genes associated with these CpG sites further elucidated the biological implications of their hypermethylation. The association of exon 1 hypermethylation with developmental processes such as pattern specification, stem cell differentiation, and neuron differentiation highlights a potential mechanism by which epigenetic alterations disrupt normal cellular differentiation and contribute to tumorigenesis. These findings are consistent with previous research linking exon 1 and promoter hypermethylation to the silencing of tumor suppressor genes and activating oncogenic pathways [27, 28, 29]. Similarly, the hypermethylation observed in 5' UTRs and distal promoter regions, linked to processes like morphogenesis and synaptic transmission, suggests a broader impact of DNA methylation on epithelial and neural signaling pathways in LUAD. These findings align with the growing recognition of epigenetic modifications as key cancer progression drivers and potential therapeutic targets [30, 10].

The diagnostic potential of exon 1 and distal promoter hypermethylation was further validated by SVM models, which achieved perfect classification of LUAD and normal adjacent tissues using as few as five features. Moreover, the strong correlation between feature importance scores and average CpG site methylation levels underscores the link between the accurate LUAD prediction observed and the consistent methylation of these sites in tumor samples. The use of machine learning models in methylation analysis has been increasingly recognized for its ability to enhance the predictive power of epigenetic biomarkers in cancer diagnostics [31]. Additionally, the identification of distinct methylation subgroups within LUAD tumors, independent of traditional staging, highlights the molecular heterogeneity of this cancer type and suggests that methylation profiling could provide a more nuanced understanding of tumor biology and ultimately inform personalized treatment strategies.

Finally, the gene expression analysis across the methylation-defined subgroups revealed significant transcriptional dysregulation, with each cluster associating with unique biological processes. These findings align with previous studies that have demonstrated how distinct methylation patterns correlate with differential gene expression, influencing tumor behavior and treatment response [32, 33]. The distinct profiles observed, particularly the enrichment of DNA repair processes in Cluster 3, immune-related processes in Cluster 4, and ribosome biogenesis in Cluster 5, provide compelling evidence of the potential for these methylation subgroups to inform targeted therapeutic approaches. Overall, this study deepens our understanding of LUAD pathophysiology and underscores the clinical utility of DNA methylation as both a diagnostic tool and a guide for personalized patient management.

Acknowledgments

The results shown here are based on openly available human data from the TCGA Research Network: https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

Funding

Not applicable.

Competing Interests

None.

Ethics Approval and Consent

Not applicable.

Data Availability

All the data used in this study are openly available in the TCGA Genomic Data Commons at https://portal.gdc.cancer.gov, dbGaP accession phs000178.

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229-263. https://doi.org/10.3322/caac.21834

2. Padinharayil, H., Varghese, J., John, M. C., Rajanikant, G. K., Wilson, C. M., Al-Yozbaki, M., ... & George, A. (2023). Non-small cell lung carcinoma (NSCLC): Implications on molecular pathology and advances in early diagnostics and therapeutics. Genes and Diseases, 10(3), 960-989. https://doi.org/10.1016/j.gendis.2022.07.023

3. Hua, X., Zhao, W., Pesatori, A. C., Consonni, D., Caporaso, N. E., Zhang, T., ... & Landi, M. T. (2020). Genetic and epigenetic intratumor heterogeneity impacts prognosis of lung adenocarcinoma. Nature Communications, 11(1), 2459. https://doi.org/10.1038/s41467-020-16295-5

4. Seguin, L., Durandy, M., & Feral, C. C. (2022). Lung adenocarcinoma tumor origin: A guide for personalized medicine. Cancers, 14(7), 1759. https://doi.org/10.3390/cancers14071759

5. Imodoye, S. O., Adedokun, K. A., & Bello, I. O. (2024). From complexity to clarity: Unravelling tumor heterogeneity through the lens of tumor microenvironment for innovative cancer therapy. Histochemistry and Cell Biology, 161(4), 299-323. https://doi.org/10.1007/s00418-023-02258-6

6. Dhar, G. A., Saha, S., Mitra, P., & Nag Chaudhuri, R. (2021). DNA methylation and regulation of gene expression: Guardian of our health. The Nucleus, 64(3), 259-270. https://doi.org/10.1007/s13237-021-00367-y

7. Ramazi, S., Dadzadi, M., Sahafnejad, Z., & Allahverdi, A. (2023). Epigenetic regulation in lung cancer. MedComm, 4(6), e401. https://doi.org/10.1002/mco2.401

8. Sadida, H. Q., Abdulla, A., Al Marzooqi, S., Hashem, S., Macha, M. A., Akil, A. S. A. S., & Bhat, A. A. (2024). Epigenetic modifications: Key players in cancer heterogeneity and drug resistance. Translational Oncology, 39, 101821. https://doi.org/10.1016/j.tranon.2023.101821

9. Kaur, N., Saini, A., Walia, A., Saxena, R., Mathur, P., & Jha, V. (2023). The role of tumor suppressor gene epigenetic regulation in the prognosis of chemotherapy resistance and aggressiveness of breast and lung cancer. Advances in Clinical and Medical Research, 4, 1-17. https://doi.org/10.52793/ACMR.2023.4(3)-58

10. Castro-Muñoz, L. J., Vázquez Ulloa, E., Sahlgren, C., Lizano, M., De La Cruz-Hernández, E., & Contreras-Paredes, A. (2023). Modulating epigenetic modifications for cancer therapy. Oncology Reports, 49(3), 59. https://doi.org/10.3892/or.2023.8496

11. Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511(7511), 543-550. https://doi.org/10.1038/nature13385

12. Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., ... & Shen, R. (2011). High density DNA methylation array with single CpG site resolution. Genomics, 98(4), 288-295. https://doi.org/10.1016/j.ygeno.2011.07.007

13. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 1-21. https://doi.org/10.1186/s13059-014-0550-8

14. Kolde, R., & Kolde, M. R. (2015). Package 'pheatmap'. R package, 1(7), 790. https://linux.yz.yamagata-u.ac.jp/pub/cran/web/packages/pheatmap/pheatmap.pdf

15. Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

16. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., ... & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

17. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., & Leisch, M. F. (2006). The e1071 package. Misc Functions of Department of Statistics (e1071), TU Wien, 297-304. https://www.cs.upc.edu/~belanche/Docencia/mineria/Practiques/R/e1071.pdf

18. Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Team, R. C. (2020). Package 'caret'. The R Journal, 223(7), 48. https://cran.radicaldevelop.com/web/packages/caret/caret.pdf

19. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 1-8. https://doi.org/10.1186/1471-2105-12-77

20. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

21. Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. Bioinformatics, 33(18), 2938-2940. https://doi.org/10.1093/bioinformatics/btx364

22. Kratzer, T. B., Bandi, P., Freedman, N. D., Smith, R. A., Travis, W. D., Jemal, A., & Siegel, R. L. (2024). Lung cancer statistics, 2023. Cancer, 130(8), 1330-1348. https://doi.org/10.1002/cncr.35128

23. Nooreldeen, R., & Bach, H. (2021). Current and future development in lung cancer diagnosis. International Journal of Molecular Sciences, 22(16), 8661. https://doi.org/10.3390/ijms22168661

24. Ibrahim, J., Peeters, M., Van Camp, G., & de Beeck, K. O. (2023). Methylation biomarkers for early cancer detection and diagnosis: Current and future perspectives. European Journal of Cancer, 178, 91-113. https://doi.org/10.1016/j.ejca.2022.10.015

25. Papanicolau-Sengos, A., & Aldape, K. (2022). DNA methylation profiling: An emerging paradigm for cancer diagnosis. Annual Review of Pathology: Mechanisms of Disease, 17(1), 295-321. https://doi.org/10.1146/annurev-pathol-042220-022304

26. Wang, B., Wang, M., Lin, Y., Zhao, J., Gu, H., & Li, X. (2024). Circulating tumor DNA methylation: A promising clinical tool for cancer diagnosis and management. Clinical Chemistry and Laboratory Medicine. https://doi.org/10.1515/cclm-2023-1327

27. So, J. Y., Skrypek, N., Yang, H. H., Merchant, A. S., Nelson, G. W., Chen, W. D., ... & Yang, L. (2020). Induction of DNMT3B by PGE2 and IL6 at distant metastatic sites promotes epigenetic modification and breast cancer colonization. Cancer Research, 80(12), 2612-2627. https://doi.org/10.1158/0008-5472.CAN-19-3339

28. Zhang, T., Ma, J., Nie, K., Yan, J., Liu, Y., Bacchi, C. E., ... & Tam, W. (2014). Hypermethylation of the tumor suppressor gene PRDM1/Blimp-1 supports a pathogenetic role in EBV-positive Burkitt lymphoma. Blood Cancer Journal, 4(11), e261. https://doi.org/10.1038/bcj.2014.75

29. Kee, S. K., Lee, J. Y., Kim, M. J., Lee, S. M., Jung, Y. W., Kim, Y. J., ... & Kim, D. S. (2007). Hypermethylation of the Ras association domain family 1A (RASSF1A) gene in gallbladder cancer. Molecules and Cells, 24(3), 364-371. https://doi.org/10.1016/S1016-8478(23)07352-1

30. Ilango, S., Paital, B., Jayachandran, P., Padma, P. R., & Nirmaladevi, R. (2020). Epigenetic alterations in cancer. Frontiers in Bioscience-Landmark, 25(6), 1058-1109. https://doi.org/10.2741/4847

31. Rauschert, S., Raubenheimer, K., Melton, P. E., & Huang, R. C. (2020). Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. Clinical Epigenetics, 12, 1-11. https://doi.org/10.1186/s13148-020-00842-4

32. Romero-Garcia, S., Prado-Garcia, H., & Carlos-Reyes, A. (2020). Role of DNA methylation in the resistance to therapy in solid tumors. Frontiers in Oncology, 10, 1152. https://doi.org/10.3389/fonc.2020.01152

33. Renner, M., Wolf, T., Meyer, H., Hartmann, W., Penzel, R., Ulrich, A., ... & Mechtersheimer, G. (2013). Integrative DNA methylation and gene expression analysis in high-grade soft tissue sarcomas. Genome Biology, 14, 1-26. https://doi.org/10.1186/gb-2013-14-12-r137