

## **DNA methylation patterns define subtypes of differentiated follicular cell-derived thyroid neoplasms: an unsupervised machine learning approach**

Vicente Rodrigues Marczyk, MD<sup>1,2</sup>; Mariana Recamonde-Mendoza, PhD<sup>3,4</sup>, Ana Luiza Maia, MD, PhD<sup>1,2</sup>, Iuri Martin Goemann, MD, PhD<sup>1,2,5</sup>

<sup>1</sup>Thyroid Unit, Endocrine division, Hospital de Clínicas de Porto Alegre (HCPA)

<sup>2</sup>Medical School, Universidade Federal do Rio Grande do Sul (UFRGS)

<sup>3</sup>Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA)

<sup>4</sup>Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS)

<sup>5</sup>Medical School, Universidade do Vale do Rio dos Sinos (UNISINOS)

**Competing Interest Statement:** VRM, MRM, ALM, and IMG declare no competing interests concerning the work described.

**Funding Statement:** This study was financed by grants from the government research agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

**Data Availability:** All data used in our study is publicly available at Gene expression. Omnibus (GEO). All R packages employed in our study are also freely available at CRAN or Bioconductor.

**Ethics approval statement:** The present study was approved by the institutional ethical committee.

### **Corresponding Author:**

Iuri Martin Goemann, [igoemann@hcpa.edu.br](mailto:igoemann@hcpa.edu.br)

Thyroid Section, Endocrine Division, Hospital de Clinicas de Porto Alegre, Rua Ramiro Barcelos 2350, 90035-903, Porto Alegre, RS, Brazil

## Abstract

Alterations in DNA methylation patterns are a frequent finding in cancer. Methylation aberrations can drive tumorigenic pathways and serve as potential biomarkers. The role of epigenetic alterations in thyroid cancer is still poorly understood. Here, we analyzed methylome data of a total of 810 thyroid samples (n=256 for discovery and n=554 for validation), including benign and malignant follicular cell-derived thyroid neoplasms, as well as normal thyroid tissue. In the discovery phase, we employed an unsupervised machine-learning method to search for methylation patterns. We found evidence supporting the existence of three distinct methylation subtypes: a normal-like, a hypermethylated follicular-like, and a hypomethylated papillary-like cluster. Follicular adenomas, follicular carcinomas, oncocytic adenomas, oncocytic carcinomas, and NIFTP samples were grouped within the follicular-like cluster, indicating that these pathologies shared numerous epigenetic alterations, with a predominance of hypermethylation events. Conversely, classic papillary thyroid carcinomas (PTC) and tall cell PTC formed a separate subtype characterized by the predominance of hypomethylated positions. Interestingly, follicular variant papillary thyroid carcinomas (FVPTC) were as likely to be classified as follicular-like or PTC-like during the discovery phase, indicating a heterogeneous group likely to be formed by at least two distinct diseases. In the validation phase, we found that FVPTC with follicular-like methylation patterns were enriched for RAS mutations. In contrast, FVPTC with PTC-like methylation patterns were enriched for BRAF and RET alterations. Our data provide novel insights into the epigenetic alterations of thyroid tumors. Since the classification method relies on a fully unsupervised machine learning approach for subtype discovery, our results offer a robust background to support the classification of thyroid neoplasms based on methylation patterns.

**Keywords:** Thyroid Neoplasm; Thyroid Carcinoma; Thyroid Cancer; Thyroid Adenoma; Papillary Thyroid Carcinoma; Thyroid Nodule; DNA Methylation; Unsupervised Machine Learning

## Introduction

Follicular cell-derived thyroid neoplasms are the most frequent endocrine malignancy worldwide<sup>1</sup>. These tumors can be classified into benign lesions, low-risk neoplasms, and malignant neoplasms<sup>2</sup>. However, diagnosis can be challenging, and many lesions can only be diagnosed after surgical excision. A better understanding of tumor genetics has given new perspectives on classifying thyroid neoplasms. However, studies have focused on genetic rather than epigenetic alterations such as methylation, as recently reviewed by Zafon *et al*<sup>3</sup>. Most studies evaluating the thyroid cancer DNA methylome used array-based technologies. Nevertheless, their conclusions were limited either by the restricted number of samples or by the underrepresentation of many histological subtypes. To date, no study has employed a robust framework to classify thyroid neoplasms based on their methylation patterns. Our study aims to analyze methylation patterns among various follicular cell-derived thyroid neoplasms and identify the existence of disease clusters that can be robust and reliably identified using methylation microarray data through unsupervised consensus clustering.

Unsupervised clustering is a powerful method to identify disease subtypes, and it has been extensively employed in cancer research<sup>4,5</sup>. However, clustering techniques frequently encounter evidence for clustering even in data where no real clusters exist, leading to equivocal interpretation about the existence of subtypes within a homogenous group<sup>6</sup>. Additionally, all commonly used clustering methods, such as K-means (KM), partitioning around medoids (PAM), or even hierarchical clustering (HC), require the researcher to at some point specify the number of clusters into which to split the data. This decision is complex since the number of categories is usually unknown beforehand. Monte Carlo reference-based consensus clustering (M3C) is a recently developed approach that robustly addresses the issue of identifying the true K number of clusters<sup>7</sup>. M3C is a consensus clustering approach (i.e., it resamples the data multiple times to evaluate how often a pair of samples clusters together)

built upon the original Monti algorithm<sup>8</sup> using a robust metric, the proportion of ambiguous clustering (PAC)<sup>6</sup>, and a reference distribution<sup>7</sup>.

Here, we demonstrate that consensus clustering can identify distinct methylation patterns and help us understand thyroid tumor biology. We propose a novel way to classify follicular cell-derived thyroid neoplasms based on their methylome.

## Methods

### Search strategy and eligibility of studies

We performed a systematic search until September 2022 using the Gene Expression Omnibus (GEO) for datasets with available DNA methylation data on thyroid cancer and benign/normal thyroid tissue analyzed by array methods using the following query ("thyroid neoplasms"[MeSH Terms] OR thyroid cancer[All Fields]) AND "Homo sapiens"[porgn] AND ("Methylation profiling by array"[Filter] OR "Methylation profiling by genome tiling array"[Filter] OR "Methylation profiling by SNP array"[Filter]). The description for each resulting dataset was read in detail to evaluate whether it was suitable for inclusion. Datasets originating from *in vitro* studies or generated from blood samples were omitted. Rare histological subtypes (< 5 samples in total), anaplastic thyroid carcinomas, and medullary thyroid carcinomas were not included. The nomenclature of thyroid neoplasms was maintained as originally described in each study.

### Data analyses

All analyses were conducted in R version 4.1.3. Raw data were retrieved from GEO in the form of *idat* files using the *GEOquery* package<sup>9</sup>. Methylation data preprocessing was performed using the *minfi* package (version 1.40.0)<sup>10</sup>. Data normalization was performed using single-sample Noob normalization, which is advantageous when working across array types with large datasets arriving in batches<sup>11</sup>. Probes were filtered out to exclude those of low

quality, probes mapped to the X and Y chromosomes, probes with SNPs, and cross-reactive probes. We inspected our data for the presence of batch effects through dimensionality reduction techniques, finding only minor batch effects present within our data. Since the datasets included in our analyses were highly unbalanced regarding tumor subtypes, batch correction retaining group differences may introduce a strong bias, deflating p values and leading to false discoveries<sup>12</sup>. Therefore, we chose a more conservative approach, using our data without batch correction for all downstream analyses. For completeness, we also ran all our analyses adjusting for the study batch using the empirical Bayes method from ComBat<sup>13</sup> and retaining the differences between tumor subtypes. We found no important changes in the results, reinforcing that batch effects were not responsible for most of the variation within our data.

#### Unsupervised machine learning

We employed a Monte Carlo reference-based consensus clustering (M3C)<sup>7</sup> approach using the *M3C* package (version 1.16.0) with default parameters unless otherwise specified. We conducted 100 Monte Carlo iterations and 100 inner replications. The optimal number of clusters (K) was defined based on two metrics: the Relative Cluster Stability Index (RCSI) and the Monte Carlo p value. Consensus clustering was performed using beta-values with the largest variance, ranging from 5,000 to 20,000 features included. We executed the consensus clustering algorithm using spectral clustering, which can deal with anisotropic clusters, unequal variances, and non-Gaussian shapes<sup>7</sup>. We employed the proportion of ambiguous clustering (PAC) as the objective function for M3C clustering. M3C was also used for t-SNE plots using default parameters. Hierarchical clustering with the 5,000 most variable beta-values was performed for visualization using the *dendextend* package (version 1.16.0) with Euclidean distances and Ward's linkage.

#### Validation in an independent dataset

Data from The Cancer Genome Project (TCGA)<sup>14</sup> were retrieved as *idat* files from the GDC Data Portal. Data preprocessing using *minfi* was performed in the same manner as for GEO datasets. Samples from metastatic sites or rare histological subtypes (< 5 samples in total) were not included. We employed the nearest shrunken centroids method<sup>15</sup> available in the *pamr* package (version 1.56.1) to train a classifier using the 5,000 most variable beta-values in our discovery dataset. The classifier was trained based on cluster assignment by M3C spectral clustering with the same 5,000 features and worked by assigning new samples to the class with the minimum distance between the sample and the shrunken centroid obtained in the clustering analysis. Our classifier was unaware of the sample histological subtype. We tested our classifier using TCGA methylation data to inspect the cluster assignment for each histological subtype. Mutational data were retrieved from cBioPortal<sup>16</sup>, and oncoprint visualizations were constructed using the *ComplexHeatmap* package (version 2.10.0).

#### Differential methylation analysis

We performed differential methylation analysis (DMA) using the *DMRCate* package (version 2.8.5)<sup>17</sup>. We conducted DMA to identify differentially methylated CpG loci and differentially methylated regions among clusters. We adjusted p values for multiple comparisons using the Benjamini–Hochberg method. We set our threshold for differential methylation at a minimum beta-value difference of 0.2 and an adjusted p value < 0.05.

#### Gene set and pathway analysis

We conducted gene set analysis (GSA) to better understand the biological significance of differentially methylated CpG loci employing Reactome<sup>18</sup> gene sets for signal transduction pathways (R-HSA-162582) and diseases of signal transduction (R-HSA-5663202) with 60 to 300 genes. We conducted GSA using robust rank aggregation and overrepresentation analysis using the *methy/GSA* package<sup>19</sup>.

## Results

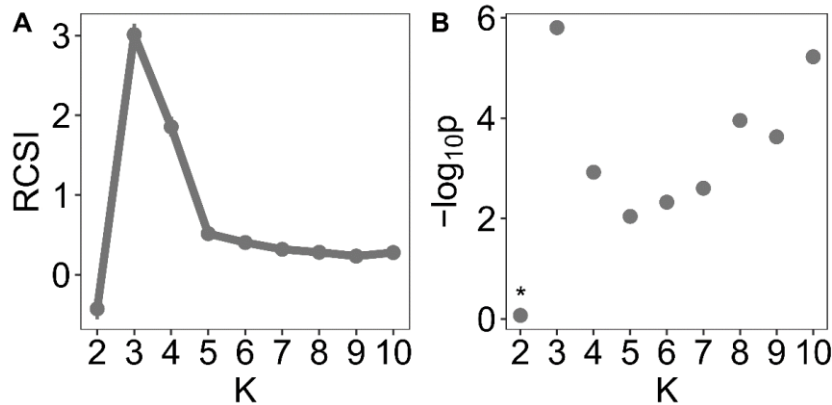
### Thyroid cancer array-based methylation studies

We identified a total of 13 studies, of which four were included (Supplementary Figure 1): GSE121377<sup>20</sup>, GSE77804<sup>21</sup>, GSE97466<sup>22</sup>, and GSE197860<sup>23</sup>. An additional study from GEO (GSE53051<sup>24</sup>) was identified from references and was included in our analysis. All five datasets employed Illumina platforms, with two studies using the EPIC array and three employing the 450K array. We included a total of 256 thyroid samples as follows: 70 classic papillary thyroid carcinomas (cPTC), 22 follicular thyroid carcinomas (FTC), 21 follicular variant papillary thyroid carcinomas (FVPTC), 16 oncocytic carcinomas (OC), 16 oncocytic adenomas (OA), 16 follicular adenomas (FA), and 6 noninvasive follicular thyroid neoplasms with papillary-like nuclear features (NIFTP) (Supplementary Table 1).

### Determining the true number of clusters using Monte Carlo consensus clustering

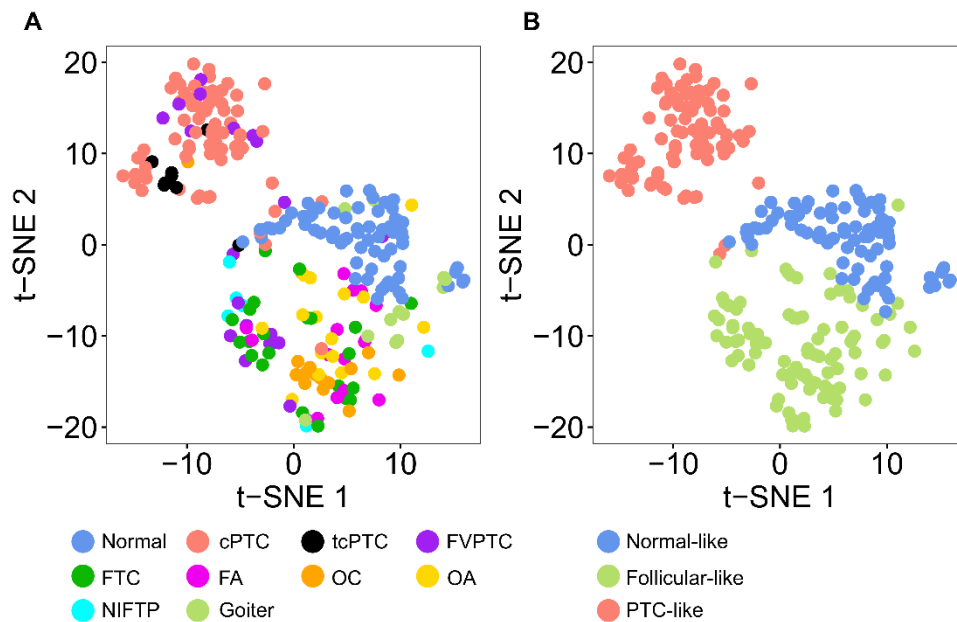
After data preprocessing, we obtained a matrix of normalized methylation beta-values for 256 samples and 332,290 genomic positions for subsequent analyses. Ideally, the optimal number of clusters should be the number  $K$  that maximizes the relative cluster stability index (RCSI) and the  $-\log_2$  Monte Carlo  $p$  value. Our results indicated the existence of three ( $K=3$ ) methylation clusters within our dataset, irrespective of whether we employed 5, 10, or 20 thousand features for clustering (Figure 1 and Supplementary Figure 2). To evaluate whether our hypothesis of  $K=3$  methylation clusters was reasonable, we performed a visual inspection to assess cluster separation using t-SNE (Figure 2), which supported  $K=3$  as the optimal number of clusters. Next, we assessed cluster assignment consistency by comparing the cluster assignment for a given sample as we varied the number of features employed. Regardless of whether we used 5, 10, or 20 thousand features for clustering, 243 out of 256 samples (94.9%) were always assigned to the same subtype, demonstrating consistent results.

**Figure 1**



**Figure 1:** Stability metrics (RCSI and Monte Carlo p value) obtained from spectral consensus clustering using the 5,000 most variable beta-values. **Abbreviations:** RCSI, Relative Cluster Stability Index. \*Not statistically significant.

**Figure 2**



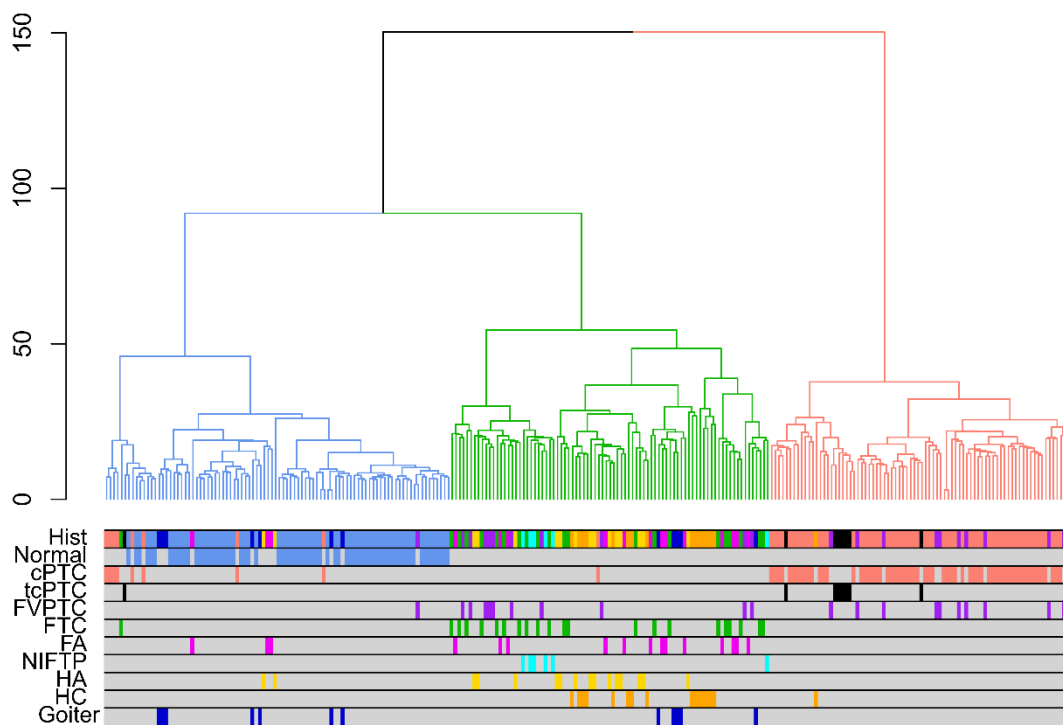
**Figure 2:** t-SNE plots were employed to inspect cluster separation. Samples were colored (A) according to histological subtype or (B) according to cluster assignment based on spectral clustering with 5,000 features. **Abbreviations:** cPTC, classic papillary thyroid carcinoma; FA, follicular adenoma; FVPTC, follicular variant papillary thyroid carcinoma; FTC, follicular thyroid carcinoma; OA, oncocytic adenoma; OC, oncocytic carcinoma; NIFTP: noninvasive follicular thyroid neoplasm with papillary-like nuclear features; tcPTC, tall cell papillary thyroid carcinoma; t-SNE, t-distributed stochastic neighbor embedding.



### Association between methylation clusters and histological subtypes

From the t-SNE plots, we observed an association between cluster assignment and histological subtypes (Figure 2). For visualization purposes, we performed conventional hierarchical clustering into K=3 clusters (Figure 3). Cluster 1 (“normal-like”) was composed primarily of normal thyroid tissue and goiter samples; cluster 2 (“follicular-like”) was composed of FA, FTC, OA, OC, and NIFTP; and cluster 3 (“PTC-like”) was composed predominantly of cPTC and tcPTC (Figure 3).

**Figure 3**



**Figure 3:** Hierarchical clustering employing the 5,000 most variable beta-values was performed to cluster samples into K=3 groups. Euclidean distances and Ward linkage were employed. **Abbreviations:** cPTC, classic papillary thyroid carcinoma; FA, follicular adenoma; FVPTC, follicular variant papillary thyroid carcinoma; FTC, follicular thyroid carcinoma; OA, oncocytic adenoma; OC, oncocytic carcinoma; NIFTP: noninvasive follicular thyroid neoplasm with papillary-like nuclear features; tcPTC, tall cell papillary thyroid carcinoma.

Consensus clustering results also demonstrated a significant association between methylation patterns and histological subtypes (Fisher's exact  $p < 0.001$ ) (Table 2). Classic PTC and tcPTC were preferentially assigned to the PTC-like cluster, with 91.4% cPTC and 100% tcPTC assigned to this group. There was also a remarkable aggregation of samples for neoplasms exhibiting follicular or oncocytic patterns. All FA and all OA were assigned to the follicular-like cluster. For their malignant counterparts, 100% FTC and 93.8% OC were also placed within this subtype. It is noteworthy that, although represented in a small number, 6 out of 6 (100%) NIFTP samples were placed within the follicular-like cluster rather than in the PTC-like cluster. Unlike all other histologic subtypes previously discussed, for which preferential clustering was evident, FVPTC did not behave as a homogeneous group, and samples were as likely to be classified as follicular-like (47.6%) as they were to be classified as PTC-like (42.9%) (Table 2). Accordingly, FVPTC samples did not cluster together on the t-SNE plot (Figure 2) or on the hierarchical clustering dendrogram (Figure 3), suggesting that this histological subtype might, in fact, represent two distinct diseases grouped as one.

**Table 2:** Cluster assignment for each histological subtype

		Normal-like	Follicular-like	PTC-like	p-value
Normal	n	69	0	0	< 0.001
	%	(100)	(0)	(0)	
cPTC	n	5	1	64	< 0.001
	%	(7.1)	(1.4)	(91.4)	
tcPTC	n	0	0	8	< 0.001
	%	(0)	(0)	(100)	
FVPTC	n	2	10	9	0.052*
	%	(9.5)	(47.6)	(42.9)	
FTC	n	0	22	0	< 0.001
	%	(0)	(100)	(0)	
FA	n	0	16	0	< 0.001
	%	(0)	(100)	(0)	
OC	n	0	15	1	< 0.001
	%	(0)	(93.8)	(6.2)	
OA	n	0	16	0	< 0.001
	%	(0)	(100)	(0)	
NIFTP	n	0	6	0	0.004
	%	(0)	(100)	(0)	
Goiter	n	6	6	0	0.031
	%	(50.0)	(50.0)	(0)	

**Legend for Table 2:** Cluster assignment for each histological subtype based on spectral consensus clustering with the 5,000 most variable beta-values. Row-wise p values were calculated using a multinomial exact test of goodness-of-fit with the null hypothesis that each tumor subtype was equally likely to be assigned to any of the clusters. **Abbreviations:** cPTC, classic papillary thyroid carcinoma; FA, follicular adenoma; FVPTC, follicular variant papillary thyroid carcinoma; FTC, follicular thyroid carcinoma; OA, oncocytic adenoma; OC, oncocytic carcinoma; NIFTP: noninvasive follicular thyroid neoplasm with papillary-like nuclear features; tcPTC, tall cell papillary thyroid carcinoma. \*Not statistically significant.

#### Validation using an independent dataset

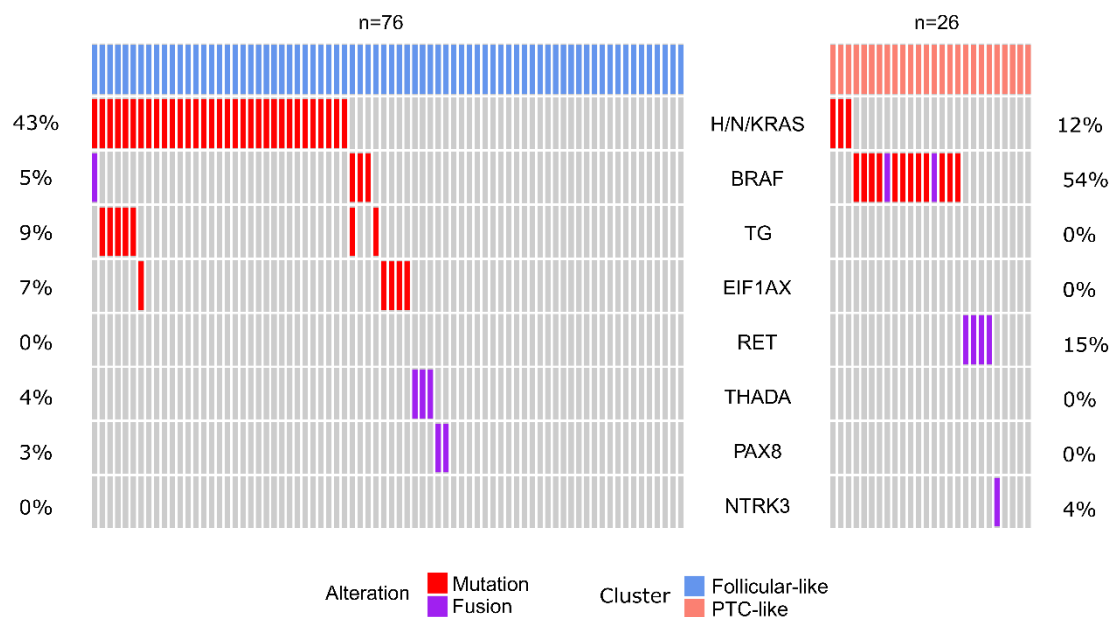
We validated our findings using TCGA samples as an independent external dataset. Samples were assigned a cluster exclusively based on their methylation pattern compared to the three previously defined clusters, without any information about histological subtypes. The results were in accordance with the discovery phase (Table 3 and Supplementary Figure 3), suggesting that true methylation clusters were correctly identified. Our model classified 94.6% of normal thyroid tissue within the normal-like cluster. All 38 (100%) tcPTC and 83.9% cPTC were assigned to the PTC-like cluster. Interestingly, FVPTC were predominantly located within the follicular-like cluster (72.4%) and less frequently in the PTC-like cluster (24.8%). For TCGA samples, we also had mutational data available, so we tested whether FVPTC assigned to the follicular-like group differed from FVPTC assigned to the PTC-like group in terms of DNA mutations (Figure 4). We observed that follicular-like FVPTC were characterized by a higher frequency of RAS mutations (43.4% vs. 11.5%, Fisher exact  $p < 0.001$ ), whereas PTC-like FVPTC were associated with alterations of BRAF (53.8% vs. 5.3%, Fisher exact  $p = 0.004$ ) and RET (15.4% vs. 0%, Fisher exact  $p = 0.004$ ). Other mutations (TG, EIF1AX) and fusions (THADA, PAX8, NTRK3) frequently associated with thyroid carcinomas were also inspected, but their low frequency did not provide sufficient power for formal testing (Figure 4).

**Table 3:** Cluster assignment for TCGA samples

	Normal-like		Follicular-like		PTC-like		p-value
	n	(%)	n	(%)	n	(%)	
Normal	53	(94.6)	3	(5.4)	0	0	<0.001
cPTC	13	(3.7)	44	(12.4)	298	(83.9)	<0.001
tcPTC	0	0	0	0	38	-100	<0.001
FVPTC	3	(2.9)	76	(72.4)	26	(24.8)	<0.001

**Legend for Table 3:** Cluster assignment for each histological subtype in the TCGA dataset. Cluster assignment was performed using a classifier based on the nearest shrunken centroids calculated from the discovery dataset. Row-wise p values were calculated using a multinomial exact test of goodness-of-fit with the null hypothesis that each tumor subtype was equally likely to be assigned to any of the clusters. **Abbreviations:** cPTC, classic papillary thyroid carcinoma; FVPTC, follicular variant papillary thyroid carcinoma; tcPTC, tall cell papillary thyroid carcinoma; TCGA, The Cancer Genome Atlas.

**Figure 4**



**Figure 4:** Oncoprint representation comparing the mutational profile of FVPTC with a follicular-like methylation pattern (left) and FVPTC with a PTC-like methylation pattern (right). Each individual column represents a single sample, and rows represent genes. **Abbreviations:** PTC, papillary thyroid carcinoma; FVPTC, follicular variant papillary thyroid carcinoma.

### Differential methylation analysis

Next, we performed differential methylation analysis to better understand the biological differences between these clusters identified during the discovery phase. Using the normal-like cluster as our reference group, we observed that follicular-like samples exhibited 1,757 differentially methylated positions, of which 89.1% were hypermethylated (Table 4). Conversely, PTC-like samples exhibited 3,241 differentially methylated sites, with a predominance of hypomethylated positions (87.5%) (Table 4). It is known, however, that most biologically significant differences in methylation affect genomic regions rather than isolated CpG loci. Therefore, we also conducted a differential methylation analysis at the genomic region level. Once again, we observed that follicular-like samples were marked mainly by hypermethylation events, whereas PTC-like samples exhibited more hypomethylated regions (Table 4). A detailed description of these genomic regions along with a complete list of differentially methylated positions is available in the Supplementary Material. We also investigated whether the differentially methylated positions mapped to genes associated with specific signaling pathways. For PTC-like samples, we found a highly significant association with the VEGF, PI3K/AKT, and MAPK signaling pathways (adjusted p value <0.001 for all three pathways) (Supplementary Table 2). For follicular-like samples, differentially methylated sites were associated with a greater variety of pathways, including VEGF, MAPK, NOTCH, and TGFB (adjusted p value <0.001) (Supplementary Table 3).

**Table 4:** Differentially methylated genomic positions and regions in comparison to the normal-like cluster

		Differentially methylated positions			Differentially methylated regions	
		Follicular-like	PTC-like		Follicular-like	PTC-like
<b>Hypermethylated</b>	n	1565	405	n	55	9
	(%)	(89.1)	(12.5)	(%)	(96.5)	(6.6)
<b>Hypomethylated</b>	n	192	2836	n	2	127
	(%)	(10.9)	(87.5)	(%)	(3.5)	(93.4)
<b>Total</b>	n	1757	3241	n	57	136

## Discussion

Our study offers the most comprehensive evaluation of the thyroid neoplasm methylome to date. Using an unsupervised machine-learning method, we identified the existence of three thyroid methylation subtypes: normal-like, hypermethylated follicular-like, and hypomethylated PTC-like.

Previous studies have shown methylome dysregulation in thyroid neoplasms, mostly using array-based methods<sup>3</sup>. Studies focused on PTC have consistently reported that these tumors are marked by hypomethylation events<sup>14,22,25,26</sup>, which is consistent with our results. FTC has been far less studied, with some studies reporting hypermethylation in this disease<sup>22,27,28</sup>, which is in accordance with our findings. To date, the most comprehensive study has been reported by Bisaro dos Reis *et al.*<sup>22</sup>, who analyzed 141 thyroid samples of various types, including normal, benign, and cancer samples. The authors proposed the existence of methylation clusters based on unsupervised hierarchical clustering, a classification that partially overlaps ours. Their study, however, had a very limited number of histological subtypes, such as FTC (n=8), OC (n=2), OA (n=3), and TCPTC (n=1). Additionally, their clustering

strategy did not formally address the issue of finding the true number of clusters. Additionally, in the TCGA study,<sup>14</sup> a methylome assessment of a large cohort of patients with thyroid cancer was also performed. Nevertheless, the data were limited to PTC and normal samples, offering a deep understanding of this histological subtype but precluding more general conclusions about follicular cell-derived thyroid neoplasms as a broader group.

Here, we show that cPTC and tcPTC have a distinct methylation pattern marked by hypomethylation events, defining the PTC-like methylation subtype. Conversely, samples exhibiting follicular or oncocytic histology have a pattern marked by hypermethylated sites, defining the follicular-like methylation subtype. It is noteworthy that follicular and oncocytic neoplasms were grouped irrespective of whether they were benign (follicular and oncocytic adenomas) or malignant tumors (follicular and oncocytic carcinomas), supporting the hypothesis that follicular adenomas and follicular carcinomas are likely to be entities within the same disease continuum. A common origin for both diseases is supported by DNA mutation analysis showing that both are characterized by RAS mutation, with follicular adenomas exhibiting these mutations in 20-40% of the cases, whereas follicular carcinomas present RAS mutations at a higher frequency (40-50%)<sup>29-31</sup>. Follicular adenomas and carcinomas also share multiple transcriptional similarities at the RNA and miRNA levels, favoring a common origin for both pathologies<sup>32</sup>.

Interestingly, our results indicate that FVPTC does not exhibit a homogenous methylation pattern, suggesting that this histological subtype might be a heterogeneous group formed by more than one disease rather than a tumor subtype itself. Most FVPTC have a follicular-like methylation pattern, suggesting that these tumors are more closely related to FTC than to PTC. These follicular-like FVPTC also show a predominance of RAS mutations, reinforcing that they are closer to FTC. Conversely, a smaller fraction of FVPTC has a PTC-like methylome, suggesting that they are truly related to PTC. These PTC-like FVPTC were enriched

for BRAF and RET alterations, strengthening this hypothesis. Our results are in accordance with previous reports describing FVPTC as a heterogeneous entity. Encapsulated FVPTC is marked by RAS mutations, resembling FTC, whereas infiltrative tumors are characterized by BRAF mutations, resembling PTC<sup>2,33</sup>. The TCGA authors also discuss the classification of FVPTC and suggest that these tumors be classified alongside FTC, since they are marked by RAS mutations and arm-level genetic alterations<sup>14</sup>. Our data offer new insights from an epigenetic perspective to help understand and classify FVPTC more precisely.

Our study has some limitations that must be recognized. Our results were derived from samples collected from different cohorts of patients, processed differently, and assessed using two distinct methylation arrays (450K and EPIC). This concern is diminished given that we could reproduce our results in a large independent cohort of patients not used during the discovery phase. Our study does not provide information about less frequent histological subtypes since they were not sufficiently represented within our dataset. Additionally, we are dependent on the reported original histological classification of samples, which might be inaccurate, especially in light of the new WHO classification of thyroid neoplasms.

In conclusion, our data provide broad insights into the major epigenetic alterations present within thyroid tumors. Using a fully unsupervised machine learning method for subtype discovery, our results offer a robust background to support the classification of thyroid neoplasms based on methylation patterns that, hinged on genomic data, could refine personalized management of thyroid neoplasms.



## Notes

**Ethics approval statement:** The present study was approved by the institutional ethical committee (Number 47605421400005327).

**Funding:** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This work received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/AWS No 032/2019, 440005/2020-5). V.R.M. received a PhD scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (88887.354162/2019-00). A.L.M. is a recipient of scholarship (No 316323/2021-7) and research grants (No 408344/2021-0) from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Programa de Apoio a Núcleos de Excelência (PRONEX/FAPERGS; nº 16/2551-0000486-2). The funder had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

**Disclosures:** VRM, MRM, ALM, and IMG declare no competing interests concerning the work described.

**CRedit author statement:** **V.R.M.:** Conceptualization, methodology, formal analysis, investigation, data curation, writing - original draft, visualization. **M.R.M.:** Methodology, Formal analysis, Data Curation. **A.L.M.:** Investigation, Writing - Review & Editing. **I.M.G.:** Conceptualization, methodology, investigation, writing - review & editing, supervision, project administration. All authors read and approved the final manuscript.

**Prior presentations:** These results have not been presented previously.

**Acknowledgments:** The results published herein are in whole or part based upon data generated in previous studies and deposited in GEO. The authors thank the patients who participated in the studies. We also thank all personnel involved in these projects for their vast

contributions.

**Data Availability:** All the data used in our study are publicly available at GEO. All R packages employed in our study are also freely available at CRAN or Bioconductor.

**Supplementary Data:**

**Supplementary Figure 1:** Flowchart illustrating the search strategy used to select studies eligible for inclusion

**Supplementary Figure 2:** Stability metrics (RCSI and Monte Carlo p value) obtained from spectral consensus clustering using the 10,000 or 20,000 most variable beta-values.

**Supplementary Figure 3:** t-SNE plots employed to inspect cluster separation in the TCGA dataset.

**Supplementary Table 1:** Description of histological subtypes within the discovery dataset.

**Supplementary Table 2:** Association between transduction pathways and differentially methylated positions in PTC-like samples

**Supplementary Table 3:** Association between transduction pathways and differentially methylated positions in follicular-like samples

**Supplementary File 1:** Genomic positions and overlapping genes for the 136 differentially methylated regions in PTC-like samples

**Supplementary File 2:** Genomic positions and overlapping genes for the 57 differentially methylated regions in follicular-like samples

## References

1. Pizzato M, Li M, Vignat J, et al. The epidemiological landscape of thyroid cancer worldwide: GLOBOCAN estimates for incidence and mortality rates in 2020. *Lancet Diabetes Endocrinol.* 2022;10(4):264-272. doi:[https://doi.org/10.1016/S2213-8587\(22\)00035-3](https://doi.org/10.1016/S2213-8587(22)00035-3)
2. Baloch ZW, Asa SL, Barletta JA, et al. Overview of the 2022 WHO Classification of Thyroid Neoplasms. *Endocr Pathol.* 2022;33(1):27-63. doi:10.1007/s12022-022-09707-3
3. Zafon C, Gil J, Pérez-González B, Jordà M. DNA methylation in thyroid cancer. *Endocr Relat Cancer.* 2019;26(7):R415-R439. doi:10.1530/ERC-19-0093
4. Bijlsma MF, Sadanandam A, Tan P, Vermeulen L. Molecular subtypes in cancers of the gastrointestinal tract. *Nat Rev Gastroenterol Hepatol.* 2017;14(6):333-342. doi:10.1038/nrgastro.2017.33
5. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415(6871):530-536. doi:10.1038/415530a
6. Şenbabaoğlu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. *Sci Rep.* 2014;4(1):6207. doi:10.1038/srep06207
7. John CR, Watson D, Russ D, et al. M3C: Monte Carlo reference-based consensus clustering. *Sci Rep.* 2020;10(1):1816. doi:10.1038/s41598-020-58766-1
8. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach Learn.* 2003;52(1):91-118. doi:10.1023/A:1023949509487
9. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007;23(14):1846-1847. doi:10.1093/bioinformatics/btm254
10. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049
11. Fortin J-P, Triche Jr TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics.* 2017;33(4):558-560. doi:10.1093/bioinformatics/btw691
12. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics.* 2016;17(1):29-39. doi:10.1093/biostatistics/kxv027
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037
14. Agrawal N, Akbani R, Aksoy BA, et al. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell.* 2014;159(3):676-690. doi:10.1016/j.cell.2014.09.050
15. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by

- shrunken centroids of gene expression. *Proc Natl Acad Sci*. 2002;99(10):6567-6572. doi:10.1073/pnas.082099299
16. Cerami E, Gao J, Dogrusoz U, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov*. 2012;2(5):401-404. doi:10.1158/2159-8290.CD-12-0095
  17. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin*. 2015;8(1):6. doi:10.1186/1756-8935-8-6
  18. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031
  19. Ren X, Kuan PF. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics*. 2019;35(11):1958-1959. doi:10.1093/bioinformatics/bty892
  20. Park J-L, Jeon S, Seo E-H, et al. Comprehensive DNA Methylation Profiling Identifies Novel Diagnostic Biomarkers for Thyroid Cancer. *Thyroid*. 2019;30(2):192-203. doi:10.1089/thy.2019.0011
  21. Boot A, Oosting J, de Miranda NFCC, et al. Imprinted survival genes preclude loss of heterozygosity of chromosome 7 in cancer cells. *J Pathol*. 2016;240(1):72-83. doi:https://doi.org/10.1002/path.4756
  22. Bizarro dos Reis M, Barros-Filho MC, Marchi FA, et al. Prognostic Classifier Based on Genome-Wide DNA Methylation Profiling in Well-Differentiated Thyroid Tumors. *J Clin Endocrinol Metab*. 2017;102(11):4089-4099. doi:10.1210/jc.2017-00881
  23. Yao Y, Xu P, Ying T, et al. Integrative Analysis of DNA Methylation and Gene Expression Identified Follicular Thyroid Cancer-Specific Diagnostic Biomarkers. *Front Endocrinol (Lausanne)*. 2022;12. <https://www.frontiersin.org/articles/10.3389/fendo.2021.736068>
  24. Timp W, Bravo HC, McDonald OG, et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med*. 2014;6(8):61. doi:10.1186/s13073-014-0061-y
  25. Beltrami CM, dos Reis MB, Barros-Filho MC, et al. Integrated data analysis reveals potential drivers and pathways disrupted by DNA methylation in papillary thyroid carcinomas. *Clin Epigenetics*. 2017;9(1):45. doi:10.1186/s13148-017-0346-2
  26. Ellis RJ, Wang Y, Stevenson HS, et al. Genome-Wide Methylation Patterns in Papillary Thyroid Cancer Are Distinct Based on Histological Subtype and Tumor Genotype. *J Clin Endocrinol Metab*. 2014;99(2):E329-E337. doi:10.1210/jc.2013-2749
  27. Mancikova V, Buj R, Castelblanco E, et al. DNA methylation profiling of well-differentiated thyroid cancer uncovers markers of recurrence free survival. *Int J Cancer*. 2014;135(3):598-610. doi:https://doi.org/10.1002/ijc.28703
  28. Affinito O, Salerno P, D'Alessio A, et al. Association between DNA methylation profile and malignancy in follicular-patterned thyroid neoplasms. *Endocr Relat Cancer*. 2019;26(4):451-462. doi:10.1530/ERC-18-0308
  29. Arora N, Scognamiglio T, Zhu B, Fahey III TJ. Do Benign Thyroid Nodules Have Malignant

- Potential? An Evidence-Based Review. *World J Surg.* 2008;32(7):1237-1246.  
doi:10.1007/s00268-008-9484-1
30. Hsiao SJ, Nikiforov YE. Molecular approaches to thyroid cancer diagnosis. *Endocr Relat Cancer.* 2014;21(5):T301-T313. doi:10.1530/ERC-14-0166
  31. Yoo S-K, Lee S, Kim S, et al. Comprehensive Analysis of the Transcriptional and Mutational Landscape of Follicular and Papillary Thyroid Cancers. *PLOS Genet.* 2016;12(8):e1006239-. <https://doi.org/10.1371/journal.pgen.1006239>
  32. Dom G, Frank S, Floor S, et al. Thyroid follicular adenomas and carcinomas: Molecular profiling provides evidence for a continuous evolution. *Oncotarget.* 2018;9(12). doi:10.18632/oncotarget.23130
  33. Rivera M, Ricarte-Filho J, Knauf J, et al. Molecular genotyping of papillary thyroid carcinoma follicular variant according to its histological subtypes (encapsulated vs infiltrative) reveals distinct BRAF and RAS mutation patterns. *Mod Pathol.* 2010;23(9):1191-1200. doi:10.1038/modpathol.2010.112