Supplementary files

Classification of Follicular-patterned Thyroid Lesions Using a Minimal Set of Epigenetic Biomarkers (Rodriguez-Rodero et al., 2022)

Supplementary Materials and methods

**DNA isolation**

Genomic DNA was purified from fresh-frozen tissues and FNAC samples were taken using a standard phenol-chloroform extraction protocol, which was followed by quantification by Qubit® dsDNA BR Assay in the Qubit® 2.0 Fluorometer (Life Technologies).

**Mutational Profiling**

Whole exome sequencing was performed at Macrogen, Inc (Seoul) using the SureSelect V6 Target enrichment protocol and the Illumina Novaseq 6000 platform (150bp paired-end read mode). The exome-seq analyses were performed according to the GATK best practices recommendations using the GATK framework (v.4.1.8) (1) and a detailed pipeline of the exome-seq strategy is shown in Supplementary Figure 1A. Briefly, quality control and read trimming of raw FASTQ files were performed with Fastp (v.0.20.1)(2). The resulting sequencing reads were aligned to the human reference genome hg38 using Burrows-Wheeler Aligner Bwa-mem (v.0.7.17) (3). SAM files were converted to BAM files using Samtools (v.1.7) (4) and binary files were sorted with the Picard toolkit (v.2.22.8). Polymerase chain reaction duplicates were marked, and a base-quality recalibration step was performed using data from known indels.
(1000G gold standard indels) and known SNPs (dbSNP138) obtained from the GATK resource bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0). Coverage metrics, sequencing artefacts, pileup summaries and the calculation of sample contamination were performed for downstream purposes. Single nucleotide variations were identified using Mutect2 (v.4.1.8) by means of the patient BAM files mentioned above, the allele frequencies from the Genome Aggregation Database (gnomAD) and a panel of normals (PON) obtained from the 1000 genomes project. The resulting VCFs were sorted, and a read orientation model was calculated for later use in the variant filtering step. Mutect2 Calls were filtered using the results from the calculated contamination step, the read orientation model, and the list of SureSelect V6 intervals. Simple filters were applied to the resulting VCF files. Briefly, we looked at variants that passed the standard MuTect2 filters and, for downstream purposes, bcftools (v.1.9) (5) was used to select those entries with a minimum depth coverage > 35 and a minimum allele frequency > 0.15. Filtered VCF files from all the samples were merged using vcftools (v.0.1.16) (6) and the resulting variants were annotated using ANNOVAR (v.2019Oct24) (7). Pathogenic variants were defined using non-synonymous SNV, stop gains and frameshift substitutions at exonic, splicing and UTR regions with a MAF < 0.01 using annotations from the 1000 genomes project (version 2015 Aug). The functional effects of these deleterious variants were assessed with the Polyphen2 (8), MutationTaster (9) and SIFT (10) algorithms, including also variants identified in the COSMIC database (11)(v70). Additional mutations in the BRAF gene were manually curated via IGV browser as none of these mutations were initially observed due to the use of stringent filtering criteria. Frameshift insertions and deletions, as well as those variants identified as deleterious mutations in the previous analyses, were considered as potential pathogenic mutations for downstream purposes.

Determination of somatic mutations in BRAF, KRAS, NRAS and HRAS genes was performed by Real-Time qPCR with the Thyroid Cancer Mutation Detection Kit (CE-IVD, Entrogen Inc) following
the manufacturer instructions and run in a 7500 Applied Biosystems equipment. This kit detects nucleotides changes resulting in the following mutations: BRAF p.V600E; KRAS p.G12A, p.G12D, p.G12V, p.G12C, p.G12S, p.G13D; NRAS p.Q61H, p.Q61L, p.Q61K, p.Q61R and HRAS p.G12V, p.G13R and p.Q61R. The mentioned mutations are detected in five groups defined by five reactions, and no specific mutation are distinguished within each group. DNA obtained from tumor samples was quantified by fluorimetric analysis using a Quantus Fluorometer (Promega). 25 ng DNA from each sample was used per run (5 ng/reaction). Limit of detection (LOD) of the assay varies depending on each mutation (0.25-10% range).

**DNA methylation profiling**

DNA methylation profiling was performed with Illumina’s high content Infinium HumanMethylationEPIC 850K beadchip platform (17) following the Illumina Infinium HD methylation protocol. Array hybridization services were provided by Oxford Genomics Centre (University of Oxford, UK). A detailed pipeline of the exome-sequencing strategy is shown in Supplementary Figure 2A. Raw data files (IDAT) from the HumanMethylationEPIC Beadchip platform were processed using R/Bioconductor package minfi (v.1.32.0) (12). Red and green signals from the raw data were corrected using the ssNOOB algorithm using the default parameters. Probes overlapping genetic variants, probes located in sexual chromosomes, cross-reactive and multimapping probes and probes with at least one sample with a detection p-value>0.01 were discarded from downstream analyses. A Beta-mixture quantile Normalisation method (BMIQ) (13) was applied for the correction of probe bias using the R/Bioconductor package watermelon (v.1.30.0) (14). The subsequent B-values and M-values were computed and employed throughout the differential methylation pipeline. To account for potential batch effects or confounding variables, we performed a surrogate variable analysis using the sva algorithm (v.3.35.2) (15). For each comparison, coefficients of the detected surrogate variables
(SVs) were estimated using the n.sv function in the context of M-values, and n.sv -1 surrogate variables were included in the definition of the model for the detection of differentially methylated CpGs (dmCpGs). The statistical significance of the DNA methylation probes was calculated with the moderated t-test implemented in limma (v_3.38.3) (16). A linear model was fitted using methylation level (M-values) as response and sample type (cancer/normal) as the main covariate of interest. In addition, the SVs generated in the surrogate variable analysis were included in the definition of the model. The resulting p-values obtained from the different contrasts were corrected for multiple testing using the Benjamini-Hochberg method. A false discovery rate threshold of 0.05, and an absolute difference between mean DNA methylation values of cases and controls of at least 25% were used for the assessment of significant dmCpGs. However, this threshold was set to 10% in the case of blood versus thyroid gland comparisons in order to discard those probes with minimal fluctuations between different tissue types.

**Random Forest-based classification system**

The random-forest based classifier was built using the R/CRAN randomForest package (v.4.6.14) and involving a training set of 22 benign (22 NH) and 17 malignant (10 FTC, 7 FVPTC) samples, and the input of the filtered dmCpGs from the previous step. Detailed information on the machine learning approach is depicted in Supplementary Figure 3A. To focus on the most informative set of epigenetic alterations identified in the different thyroid lesion subtypes, we initially considered the set of 12,845 dmCpGs resulting from the sum/intersection of the NH, FTA, FTC and FVPTC versus normal thyroid gland comparisons, excluding those changes shared in terms of significance and direction with blood tissue. For correct model generation, we reduced the number of input variables using an additional filter of highly correlated dmCpGs (cut-off 0.9) with the R/CRAN Caret Package (v.6.0.86). A random-forest based classifier was built using the R/CRAN random Forest package (v.4.6.14) which included a training set of 22
benign (22 NH) and 17 malignant (10 FTC, 7 FVPTC) samples, and the input of the 8,214 filtered dmCpGs from the previous step. By means of a bootstrap, iterative randomization with a seed function implemented on the training set, we trained a total of 100 models using the random Forest function with its default parameters, except for the number of trees, which was set to 5,000 per model, and the cut-off value for the prediction of benign class was set to 0.7 of the votes for better sensitivity/specificity. For each model, variable importance was ranked according to the model’s Mean Decrease Gini, and the top 10 most important variables were selected for downstream filtering purposes. For final variable selection, the cross-validation method (10-fold) was used, along with the recursive feature elimination method within the R/CRAN Caret package and the rfe algorithm, using the 10 variables with the highest number of top informative observations across models. The best model performance on the training set was achieved using a minimal number of 3 variables (cg17154646, cg21915100 and cg10189462 probes). A detailed pipeline including details of the model generation is provided in Figures S3A-D. Malignancy score for a given sample was calculated based on the ratio of malignant votes over the total number of votes (5000) predicted using the model mentioned earlier, the matrix with the DNA methylation values of the three CpG sites and the predict function of the R stats package (v.4.0.2) according to the following formula: MS = M votes / M votes + B votes. where MS stands for Molecular Malignancy Score, and M and B represent the number of malignant or benign votes identified across the different trees generated during the random forest approach. This score is ranged between 0 and 1, and higher scores represent a higher molecular malignancy status of a given sample.

Statistical power of the study considering different sample sizes was calculated with the PowerTools framework(17) (Figure S2B) using the DNA methylation matrix including cg17154646, cg21915100 and cg10189462 and the aforementioned samples from the training
dataset. A range between 5 and 39 samples, a total of 100 iterations per variable and a threshold for significance of 0.001 was included for calculation purposes.

**Bisulfite Conversion and Pyrosequencing assays**

The DNA methylation pattern of selected CpGs identified by microarray analysis was evaluated by bisulfite pyrosequencing. Bisulfite modification of DNA was performed with the EZ DNA Methylation Gold kit (Zymo Research, Orange, CA, USA) according to manufacturer’s recommendations. The set of primers for PCR amplification and sequencing were designed by PyroMark Assay Design Software (v. 2.0.01.15). Briefly, 2 µl of the bisulfite-modified DNA was amplified by PCR with the specific primer set (**Supplementary Table 1**). Next, pyrosequencing was performed using Pyromark Q24 reagents (Qiagen) by immobilizing 20 µl of PCR product with 2 µl of Streptavidin Sepharose High Performance (GE Health-care Bio-Sciences, Uppsala, Sweden) for 10 min at 80ºC, and then annealing the resulting sample with a sequencing primer. Samples were processed using the PyroMark Q24 system and the results were analysed with the PyroMark analysis software.

**Supplementary Figure legends**

**Figure S1** – Additional information related to the exome sequencing experiment. **A)** Graph depicting the computational pipeline and software used for the identification of deleterious mutations in thyroid cancer samples. **B)** Bar plots reflecting the number of sequencing paired end reads mapped to the human genome (hg38, top graph) or the mean exome coverage (bottom graph) in the exome sequencing experiment. Grey and blue lines represent, respectively, the average and the median scores across all samples. **C)** Prevalence of NRAS, HRAS
and BRAF mutations in the training and the validation cohorts identified by exome-seq or RT-PCR respectively.

Figure S1

A

DNA sequencing somatic variant alignment step

DNA sequencing somatic variant calling step

FASTQ files

Stats Intervals

Q & trim Filter

Fimp (v.0.20.1)

hg38 VCF

Depth of Coverage

GATK (v.4.1.6)

sf-only gnomad

SureSelectV5 intervals

BAM files

Collect Seq artifacts

GATK (v.4.1.6)

Calculate contamination

GATK (v.4.1.6)

Filter Mutect calls

GATK (v.4.1.6)

Filter by depth

Bimber (v.1.8)

Merged VCFs

Vcftools (v.0.1.15)

Annotate VCFs

Annovar (20190924)

Annotate file

hg38 FASTA

Trimmed FASTQ files

Genome Mapping

Bowtie2 (v.2.3.4)

SAM to BAM

Samtools (v.1.7)

Sort SAM

Picard (v.2.22.6)

Mark Duplicates

GATK (v.4.1.6)

Known SNPs VCF

B

Number of Mutated 'Parent' Reads

Mean Coverage

C

Training cohort

Validation cohort

NRAS

HRAS

BRAF

None

n = 19

n = 7

n = 15

n = 22

n = 11

n = 22

n = NA
Figure S2 – Additional information related to the high-throughput DNA methylation experiment. A) Graph illustrating the computational pipeline and software used for the identification of differentially methylated CpG sites in thyroid cancer samples. B) Evaluation of the performance of the Random forest classification system using different sample sizes. Effect size and power calculations were obtained using the PowerTools framework as indicated in the supplemental material and methods section.
Figure S3 – Additional information related to the generation of the random forest classifier.

A) Graph delineating the strategy for the identification of CpG sites that could be used for discrimination. B) Barplot showing the most informative CpG sites between benign and malignant conditions. Values indicate the number of times that a CpG site was identified as being among the 10 most informative variables using an iterative random forest selection approach (100 different seeds, 5000 trees each). C) Line plot illustrating the optimal number of variables (CpGs) required for the development of an accurate model by means of a recursive feature elimination approach. Accuracy reaches its maximum level when 3 features are retained in the model. D) Histogram representing the relative importance (Mean decrease Gini) of the three CpG sites included in the final model. E) Table representing the genomic coordinates, the nearest genes and the CpG context of the previously mentioned CpG sites. F) Boxplots illustrating the DNA methylation values of the indicated CpGs in the context of benign and NIFTP samples from an independent MethylationEPIC array (GSE121377) validation dataset. NIFTP displays features of malignant tumours using the three-CpG random forest classifier. Statistical significance between conditions was calculated using a pairwise Wilcoxon rank sum test (** = p < 0.01). G) For the estimation of the molecular malignancy status, we trained a Random Forest classifier using the information obtained from the cg10705422, cg17707274 and cg26849382 probes used in the GSE121377 dataset (Park study). The Boxplots indicate the DNA methylation value of three CpG sites identified in the MethylationEPIC study (GSE121377) across different sets of samples and thyroid cancer categories. Top lane illustrates the DNA methylation values of these three CpG sites in the context of the samples from the GSE121377 study (training dataset, excluding NIFTP). Bottom lane displays the DNA methylation values of these three CpG sites using the samples from this study as a validation cohort. As expected, a statistically significant capacity to discriminate between benign and malignant thyroid lesions was independently accomplished with any one of the three CpG sites used individually in the training dataset (upper panels) but not in the validation dataset (bottom lane, our cohort). Statistical significance between
conditions was assessed using a pairwise Wilcoxon rank sum test (* = p < 0.05, ** = p < 0.01, *** = p < 0.001). H) Radar plots illustrating the metrics and performance of the three-CpG model obtained from GSE121377 in the context of the training dataset GSE121377 (top graph) and the validation dataset (this study, bottom graph).

Figure S3
Figure S4—Validation of the three-CpG epigenetic classifier using the Bisulfite Pyrosequencing technique. A) Scatter plots indicating the correlation between DNA methylation Human Methylation EPIC arrays and DNA pyrosequencing of the same set of samples using cg17154646, cg21915100 and cg10189462. Pearson’s correlation score, statistical significance and the number of cases included in the different categories are shown. B) Boxplots indicating the DNA methylation value of three CpG sites across different sets of samples and thyroid cancer categories. This minimal set of CpGs was selected using a random forest approach trained to discriminate between benign (NH) and malignant (FTC, FVPTC) thyroid lesions. Top lane displays the DNA bisulphite pyrosequencing methylation values of the indicated CpG sites in the context of the EPIC array training dataset. Middle lane contains the DNA bisulphite pyrosequencing methylation values of the previously mentioned CpG sites in the Validation dataset. For all the graphs, dots represent individual samples, boxplots were grouped according to their cancer type and the number of cases included in each category is indicated. Statistical significance between conditions was calculated using a pairwise Wilcoxon rank sum test (* = p < 0.05, ** = p < 0.01, *** = p < 0.001). C) Radar plots illustrating the metrics and performance of the proposed three-CpG model in the context of the training dataset (top graph) and the orthogonal validation dataset (middle graph).
Figure S5- Interrogation of the DNA methylation levels of FTA samples in the Training data set using Human Methylation EPIC arrays. A) Boxplot indicating the DNA methylation values of NH and FTA samples in the context of the three CpG sites used for the modelling of thyroid cancer malignancy. Statistical significance between conditions was calculated using a pairwise Wilcoxon rank sum test (** = p < 0.01). B) Dot plot displaying the inferred cancer malignancy scores of the samples used in the training dataset. The cancer type and the diagnostic prediction (benign BN or malignant M) of each sample using the three-CpG random forest classifier are indicated.
Figure S6-Malignancy scores assigned to a cohort of PTC, oncocytic FTA / FTC and FNAC samples using the three-CpG classifier. A) Bar plot indicating the molecular malignancy scores of an additional set of PTC, oncocytic FTA and oncocytic FTC samples. B) Dot plot showing the inferred cancer malignancy scores of the samples used in the FNACs dataset. The cancer type and the diagnostic prediction (benign BN or malignant M) of each sample using the three-CpG random forest classifier are indicated. C) Pie charts showing the number of samples classified as malignant or benign in the different subgroups of FNAC samples.
**Supplementary table legends**

**Supplementary Table 1** | Sets of primers used in this study.

**Supplementary Table 2** | Exome-seq statistics corresponding to human thyroid lesions analysed in this study.

**Supplementary Table 3** | Mutations with potential deleterious consequences identified in our thyroid sample cohort. Mutations were annotated with ANNOVAR.

**Supplementary Table 4** | List of dmCpG sites between human thyroid lesions from our sample cohort and thyroid gland samples from the ENCODE consortium. Each tab includes the statistical information of the dmCpGs observed in each comparison. The table also includes information on those dmCpGs identified between pooled hematopoietic cells and thyroid gland samples.

**Supplementary Table 5** | Clinico pathological information and molecular malignancy scores calculated using the prediction algorithm on the training and validation cohorts.

**References**


