List of variables for dataset:

Whole-genome sequencing of follicular thyroid carcinomas reveal recurrent mutations in microRNA processing subunit DGCR8

Table 1. Detailed characteristics of the study cohort.

Sample id	The study sample id number
Age at surgery	The patients age at the time of surgery
Sex	Gender
Final disease status	AWD = alive with disease, AWOD = alive without disease, DOD = dead of disease, DWD = dead with disease.
Site of metastasis	
WHO Subtype	MI = minimally invasive, WI = widely invasive
Onkocytic	Oncocytic/Hürthle cell histology
Size (mm)	The tumor maximum diameter
Follow-up time (Months)	The number of months between diagnosis and follow-up
T stadium	T stadium according to TNM (Tumor size, lymph node, metastasis) classification
AJCC Stage	American Joint Committee on Cancer stage
Ki-67 index (%)	Proliferation index
TERTp mutation	TERT (Telomerase reverse transcriptase) promoter mutation

Table 2. Somatic SNV's (Small nucleotide variants).

Hugo_Symbol	The official HUGO (Human genome organisation) gene symbol
Entrez_Gene_id	The unique entrez gene id
NCBI_Build	Genome Assembly
Chromosome	Chromosome number
Start_Position	Chromosome co-ordinate start
End_Position	Chromosome co-ordinate end
Strand	Indicates plus or minus strand
Variant_Classification	Translational effect of variant allele
Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP (di-nucleotide polymorphism) but for three consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of four or more (SNP, DNP, TNP, ONP, INS, DEL, or Consolidated)
Reference_Allele	The plus strand reference allele at this position.
Tumor_Seq_Allele1	Primary data genotype for tumor sequencing (discovery) allele 1
Tumor_Seq_Allele2	Tumor sequencing (discovery) allele 2
dbSNP_RS	The rs id from the dbSNP database
dbSNP_Val_Status	The dbSNP validation status is reported as a semicolon-separated list of statuses. The union of all rs-IDs is taken when there are multiple
Tumor_Sample_Barcode	Barcode for the tumor sample
Matched_Norm_Sample_Barcode	Barcode for the matched normal sample
Match_Norm_Seq_Allele1	Matched normal sequencing allele 1
Match_Norm_Seq_Allele2	Matched normal sequencing allele 2

HGVSc	Human Genome Variation Society coding DNA variant sequence
HGVSp	Human Genome Variation Society Protein variant sequence
HGVSp_Short	Same as the HGVSp column, but using 1-letter amino-acid codes
Transcript_ID	Ensembl ID of the affected transcript
Exon_Number	The exon number in the affected gene
all_effects	A semicolon delimited list of all possible variant effects, sorted by priority
Allele	The variant allele used to calculate the consequence
Gene	Ensemble ID of the affected gene
Feature	Ensembl ID of feature (transcript, regulatory, motif)
Feature_type	Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature (or blank)
Consequence	Consequence type of this variant
cDNA_position	Relative position of base pair in the cDNA sequence as a fraction
CDS_position	Relative position of base pair in coding sequence
Protein_position	Relative position of affected amino acid in protein
Amino_acids	Only given if the variation affects the protein-coding sequence
Codons	The alternative codons with the variant base in upper case
Existing_variation	Known identifier of existing variation
ALLELE_NUM	Allele number from input; 0 is reference, 1 is first alternate etc.
DISTANCE	Shortest distance from the variant to the transcript
STRAND_VEP	The DNA strand (1 or -1) on which the transcript/feature lies
SYMBOL	Gene symbol
SYMBOL_SOURCE	The source of the above gene symbol
HGNC_ID	Gene identifier from the HUGO Gene Nomenclature Committee if applicable
BIOTYPE	Biotype of transcript
CANONICAL	A flag (YES) indicating that the VEP-based canonical transcript, the longest translation, was used for this gene. If not, the value is null
CCDS	The CCDS (Cosensus coding sequence) identifier
ENSP	The ensemble protein identifier of the affected transcript
SWISSPROT	The Swissprot accession
TREMBL	UniProtKB/TrEMBL identifier of protein product
UNIPARC	Uniparc identifier of the protein product
RefSeq	Refseq identifier for the transcript
SIFT	SIFT (Sorting Intolerant from Tolerant) prediction and score
PolyPhen	PolyPhen prediction and score
EXON	The exon number
INTRON	The intron number
IMPACT	The impact modifier for the consequence type
PICK	Indicates if this block of sequence was picked by VEP's pick feature
VARIANT_CLASS	Sequence ontology class
vcf_id	Existing variation

Table 3. MutSig2CV input genes

Hugo_Symbol	The official HUGO (Human genome organisation) gene symbol
Entrez_Gene_Id	The unique entrez gene id
NCBI_Build	Genome Assembly
Chromosome	Chromosome number
Start_Position	Chromosome co-ordinate start
End_Position	Chromosome co-ordinate end
Strand	Indicates plus or minus strand
Variant_Classification	Translational effect of variant allele
Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP (di-nucleotide polymorphism) but for three consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of four or more (SNP, DNP, TNP, ONP, INS, DEL, or Consolidated)
Reference_Allele	The plus strand reference allele at this position.
Tumor_Seq_Allele1	Primary data genotype for tumor sequencing (discovery) allele 1
Tumor_Seq_Allele2	Tumor sequencing (discovery) allele 2
dbSNP_RS	The rs id from the dbSNP database
dbSNP_Val_Status	The dbSNP validation status is reported as a semicolon-separated list of statuses. The union of all rs-IDs is taken when there are multiple
Tumor_Sample_Barcode	Barcode for the tumor sample
Matched_Norm_Sample_Barcode	Barcode for the matched normal sample
Match_Norm_Seq_Allele1	Matched normal sequencing allele 1
Match_Norm_Seq_Allele2	Matched normal sequencing allele 2
HGVSc	Human Genome Variation Society coding DNA variant sequence
HGVSp	Human Genome Variation Society Protein variant sequence
HGVSp_Short	Same as the HGVSp column, but using 1-letter amino-acid codes
Transcript_ID	Ensembl ID of the affected transcript
Exon_Number	The exon number in the affected gene
cDNA_position	Relative position of base pair in the cDNA sequence as a fraction
HGNC_ID	Gene identifier from the HUGO Gene Nomenclature Committee if applicable
BIOTYPE	Biotype of transcript
CANONICAL	A flag (YES) indicating that the VEP-based canonical transcript, the longest translation, was used for this gene. If not, the value is null
CCDS	The CCDS (Cosensus coding sequence) identifier
ENSP	The ensemble protein identifier of the affected transcript
SWISSPROT	The Swissprot accession
TREMBL	UniProtKB/TrEMBL identifier of protein product
UNIPARC	Uniparc identifier of the protein product
RefSeq	Refseq identifier for the transcript

SIFT	SIFT (Sorting Intolerant from Tolerant) prediction and score
PolyPhen	PolyPhen prediction and score
EXON	The exon number
INTRON	The intron number
DOMAINS	Source and identifier of any overlapping protein domains
CLIN_SIG	Clinical significance of variant from dbSNP
SOMATIC	Somatic status of each ID reported under Existing_variation (0, 1, or null)
PUBMED	List of pubmed ID's that cite the variant
MOTIF_NAME	The source and identifier of a transcription factor binding profile aligned at this position
IMPACT	Impact modifier
VARIANT_CLASS	Sequence ontology variant class

Table 4. MutSig2CV genes ranked by p-value

MutSig2CV analyzes somatic point mutations discovered in DNA sequencing, identifying genes mutated more often than expected by chance given inferred background mutation processes. MutSig2CV consists of three independent statistical tests.

rank	Position of the gene as sorted ascending by p-/q-value.
gene	The official gene symbol
longname	The full name of the gene
codelen	Open reading frame length of the gene.
nnei	Number of neighboring genes in the bagel used to estimate background mutation rate.
nncd	Number of noncoding mutations.
nsil	Number of silent (synonymous) mutations in the gene.
nmis	Number of missense mutations in the gene.
nstp	Number of nonsense mutations in the gene.
nspl	Number of splice site mutations in the gene (defined as +/- 2 bases from the donor/acceptor site)
nind	Number of insertions or deletions in the gene.
nnon	Number of nonsilent mutations in the gene (including all indels and splice site mutations, even if the codon change is synonymous in the latter case).
npat	Number of patients with mutations in the gene.
nsite	Number of uniquely mutated sites in the gene (does not multiply count recurrently mutated positions).
pCV	Abundance p-value.
pCL	Clustering p-value.
pFN	Functional (conservation) p-value.
p	Overall p-value obtained from Fisher combination of pCV, pCL, and pFN.
q	FDR-corrected (Benjamini-Hochberg) overall p-value.

Table 5. List of genes in copy number altered minimal region of amplification.

4p11	Cytoband p11 on chromosome 4
6p21.32	Cytoband p21.32 on chromosome 6
10q11.21	Cytoband q11.21 on chromosome 10

Table 6. Aberrant cell fraction and ploidy as determined by ASCAT.

Sample	The sample study id
Aberrant Cell Fraction	The fraction of tumour cells
Ploidy	The amount of DNA per tumour cell, expressed as multiples of haploid genomes

Table 7. List of high-confidence structural variations in the FTC cohort.

Case id	The sample study id
Chromosome	Chromosome number
Position	Chromosome co-ordinate
ALT	Breakpoint
Genes	Affected genes
SV Type	The type of structural variant

Table 8. List of significant differentially expressed genes in tumor versus normal thyroid.

Gene symbol	The official gene symbol
Description	Description of the gene
Entrezid	Official Entrez ID
Ensembl Gene id	Official Ensembl gene id
baseMean	The average of the normalized count values, dividing by size factors, taken over all samples
log2FoldChange	The effect size estimate
lfcSE	The standard error estimate for the log2 fold change estimate
stat	The Wald statistic
p value	The p value of the Wald test
Adjusted p value	Benjamini-Hochberg (BH) adjustet p value