Supplementary Material Index

Supplementary Tables

Supplementary Note

The T>A mutation maps to *GTF2I* locus but not to its pseudogenes

GTF2I is a gene that spans 35 exons and is mapped on the long arm of chromosome 7 (chr7:74,072,030-74,175,022). The chr7:74146970 T>A mutation identified by exome sequencing is located in exon 15 of *GTF2I*. The T>A mutation was aligned to the same position of *GTF2I* exon 15 using either BWA (data not shown) or Novoalign algorithms (Supplementary Fig. 7A).

There are 2 known pseudogenes of *GTF2I*: *LOC100093631* and *GTF2IP1*. *GTF2I* exon 15 sequence differs by only 1 nucleotide from the sequence of the 2 pseudogenes (According to BLAT in the UCSC website: 99.5% identity, chr7:74629125-74629308 and chr7:72593127-72593310). Both these pseudogenes map to the long arm of chromosome 7 (chr7:72569012-72621336 and chr7:74601104-74653445, respectively). *GTF2IP1* maps on the negative strand of chromosome 7, whereas *LOC10093631* and *GTF2I* reside on the positive strand.

These 2 pseudogenes possess exons and their transcripts are processed into mature mRNAs, but proteins are not translated from either of them [1]. The 2 pseudogenes have a head (exon 1) that has no homology with *GTF2I* sequence and a tail (exon2 - 3'-UTR) that is very similar (99% identical; *GTF2I/GRF2IP1* 31,794bp identical on 31,929bp of sequence and *GTF2I/ LOC100093631* 31,782bp identical on 31,925bp of sequence; Supplementary Fig. 8A).

The transcripts of the two pseudogenes are almost identical since their sequences differ by only 3 nucleotides out of 3631bp. Part of the *GTF2I* RNA sequence, exon13-3'UTR, is closely related to the portion of pseudogene sequences exon2-3'UTR (Supplementary Fig. 8A). *GTF2I* exon15-3'UTR and pseudogene exon2-3'UTR RNA sequences are 99% identical. Among 3218bp of shared sequence, *GTF2I* transcript differs only by 4 and 3 nucleotides from LOC100093631 and GTF2P1, respectively. The first exon of the 2 pseudogenes is not related to *GTF2I* sequence but closely resembles the first exon of

GATS (93.9% identity) and *GATSL2* (98.9% identity) genes. *GTF2I* sequence from exon 1 to exon 12 is unique and BLAT search did not reveal close similarity to other genomic regions.

Because the chr7:74146970 T>A mutation was mapped to *GTF2I* exon 15, it resides in the region of high homology between the gene and the pseudogenes. Therefore, it was necessary to demonstrate that this T>A mutation really belongs to *GTF2I* locus rather than to the pseudogenes. At the genomic level, exon 15 of *GTF2I* is ~4500bp away from the point where *GTF2I* and the pseudogene sequences start to differ. In contrast, at mRNA level the distance is only 217bps apart in the δ-isoform. Therefore, it was possible to design specific primers able to distinguish mRNA sequences of *GTF2I* from those of the pseudogenes. The T>A mutation was observed only in *GTF2I* cDNA but not in the cDNA from the pseudogenes in the 5 samples tested. Mutations were not identified in *GTF2I* or in the pseudogenes in the negative controls (4 samples without *GTF2I* mutation). To further demonstrate that the mutation belongs to *GTF2I* locus at the genomic level we took advantage of the fact that *GTF2I* exon 15 and pseudogenes exon 4 differ by 1 nucleotide. The nucleotide chr7:74146870 is a cytidine in *GTF2I* sequence whereas the corresponding chr7:72593177 and chr7:74629258 in *LOC100093631* and *GTF2IP1* were thymidines (Supplementary Fig. 9A). Because polymorphisms have not been described in these 3 positions, according to dbSNP137, the C/T single nucleotide difference could be used as a marker of *GTF2I* and pseudogenes sequences. Therefore, if the sequenced DNA strand contains both the C/T marker and the T>A mutation, one can ascertain whether the sequences with the mutation come from the gene or from the pseudogenes. Thus, it was possible to design primers that indistinctly amplify *GTF2I* and pseudogene sequences and then to determine if the T>A mutation belongs to the gene or to the pseudogenes using the C/T marker. In order to sequence just one strand of DNA we adopted two strategies. The first was based on TopoTA cloning and the second on deep-sequencing technologies (MiSeq, Illumina). For TopoTA cloning, primers were designed in order to amplify a 218bp DNA fragment that includes the C/T (gene/pseudogenes) marker and the site of mutation (chr7:74146970 T>A). The amplicons, generated using PCR reactions, were cloned into a pCR™4-TOPO plasmid so that the expression of the toxic ccdB gene in the vector backbone, was disrupted. E. Coli DH5α bacteria were transformed and plated in a Petri dish with Ampicillin selection. Only bacteria carrying the amplicons, but not those carrying the empty vector, were able to grow. Colonies (17-40 for each tumor) that carry a single copy of DNA amplicons were picked, expanded and their DNA sequenced using specific sequencing primers. According to exome sequencing results the chr7:74146970 T>A mutation was expected to be heterozygous. Because the 2 pseudogenes were expected to be homozygous wild type, the mutated *GTF2I* amplicons should be 1:6 (~17%) of the amplicon sequenced (Supplementary Fig. 8B). Four different tumors have been studied using TopoTA cloning, in all of them approximately 1:3 of the colonies were from *GTF2I* (average 35%; range 30-40%) and included all the T>A mutations. Colonies with a copy of mutated *GTF2I* were about 12% (6-18%), slightly less than the 17% expected, which is compatible with some normal cell contamination of the samples (Supplementary Table 7).

A customized deep sequencing assay was developed in order to discriminate mutations in *GTF2I* or pseudogenes. It was based on 2 pairs of primers (P1 and P3) able to amplify a region that includes the T>A mutation site and the Gene/Pseudogene marker (C/T). An additional pair of primers (P2) was included in the deep sequencing assay exclusively for genotyping purposes and was designed in order to enrich the amplification of *GTF2I* sequences. Twelve samples were multiplexed on a MiSeq flow cell in order to obtain extremely high read counts over the region of interest (average number of total reads was 2,306,186 range 1,137,605-4,122,859; average number of informative reads was 585,714 range 167,054-1,449,423). Five samples had *GTF2I* mutation and 7 were negative controls that included 3 normal DNA, 3 tumors without *GTF2I* mutation and a thymic carcinoma cell line without *GTF2I* mutation (Supplementary Table 8). The deep sequencing assay demonstrated *GTF2I* mutation only in the 5 positive cases but not in the negative controls. The reads with the mutations belonged exclusively to *GTF2I* sequence in 3 cases, whereas 2 tumors presented 1% of the pseudogene reads with the mutation. The frequency of *GTF2I* reads with the mutation was close to what expected (average frequency 11.5%, range 8-18%, expected frequency 17%). The few pseudogene mutated reads (1% in 2 cases) did not support the presence of a pseudogene allele carrying the T>A mutations (expected 17%). These reads may be related to polymerase errors introduced in the amplification step or they can represent a real pseudogene T>A mutation present in a subclone of few tumor cells.

When all samples were evaluated, for which the *GTF2I* mutation was genotyped using MiSeq (n=250), the results matched the expectation (Supplementary Fig. 9B). Results were evaluated separately for the 12 normal samples, the WT tumors and the tumors with *GTF2I* mutations. Normal and WT tumors have an inconspicuous proportion of mutated reads either from *GTF2I* or from the pseudogenes. In the tumors with *GTF2I* mutations, the average mutated reads were 4.01% from *GTF2I* and 0.4% from the pseudogenes. This low proportion of mutated *GTF2I* reads was expected for the presence of tumors with extensive components of non-neoplastic thymocytes. Even in the tumor with highest fraction of mutated reads belonging to the pseudogenes (2.96%), the reads from *GTF2I* were significantly higher (16.13%).

The results of TopoTA cloning, MiSeq and the transcript sequences demonstrated that the chr7:74146970 T>A mutation unambiguously involves the *GTF2I* sequence.

Frequency of *GTF2I* mutation

According to the exome sequencing and transcriptome sequencing results, *GTF2I* mutation was common in the A and AB subgroups of thymomas. A larger cohort of patients was then screened for *GTF2I* mutation using standard Sanger technology and the deep sequencing approach described above. Sanger sequencing revealed *GTF2I* T>A mutation in 78 (39%) out of 199 thymic epithelial tumors. The somatic nature of the mutation was confirmed by the absence of *GTF2I* mutation in normal DNA from patients' blood. A limitation of the standard Sanger methodology is the presence of non-neoplastic thymocytes that can outnumber the epithelial tumor cells in some histotypes (in particular some AB, B1, B2). Therefore, only tumors with at least 50% cancer cells were sequenced using Sanger technology. Alternatively, a deep sequencing approach was considered for screening for the presence of the *GTF2I* mutation in thymocytes-rich tumors. However, samples with an extremely high proportion of non-neoplastic thymocytes represent a challenge for the detection of *GTF2I* mutations even using deep-sequencing. This can be the case for some B1 thymomas. According to the deep-sequencing assay, 106 tumors (42%) out of 250 had the *GTF2I* mutation. Sanger and deep sequencing technologies showed good concordance on the 172 samples assayed with both methods. The two methods detected *GTF2I* mutations in 59 cases and excluded its presence in 88 tumors. The deep sequencing approach was designed to be more sensitive than the Sanger method in lymphocyte rich tumors, where *GTF2I* mutation was identified in 20 additional cases. These cases were considered mutated. In 5 samples (3%), mutations were observed only using the Sanger method but not using the deep sequencing approach. The mutation status of these samples was considered undetermined. When discordant results were observed in cases with 3 different sequencing technologies, samples were considered mutated when 2 of these technologies detected the mutation.

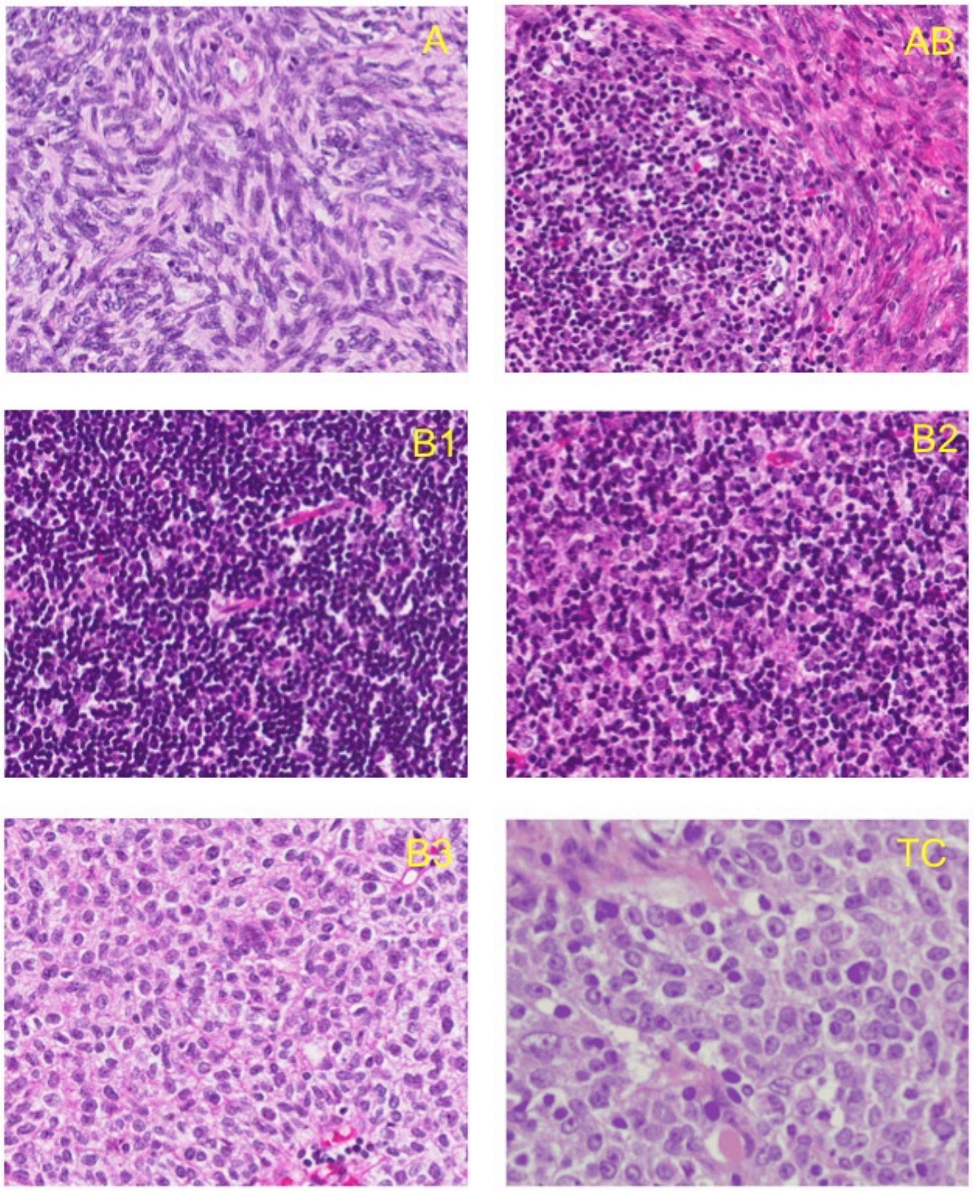
Combining exome sequencing, Sanger sequencing and *GTF2I* deep sequencing data, *GTF2I* mutation was observed in 119 tumors out of the 274 evaluated (43%). The frequency of mutation was higher in thymomas (50%) than in thymic carcinomas (8%; Fisher exact test p<0.001).

Supplementary References:

1. Perez Jurado, L.A. *et al.* A duplicated gene in the breakpoint regions of the 7q11.23 Williams-Beuren syndrome deletion encodes the initiator binding protein TFII-I and BAP-135, a phosphorylation target of BTK. *Hum Mol Genet* **7**, 325-34 (1998).

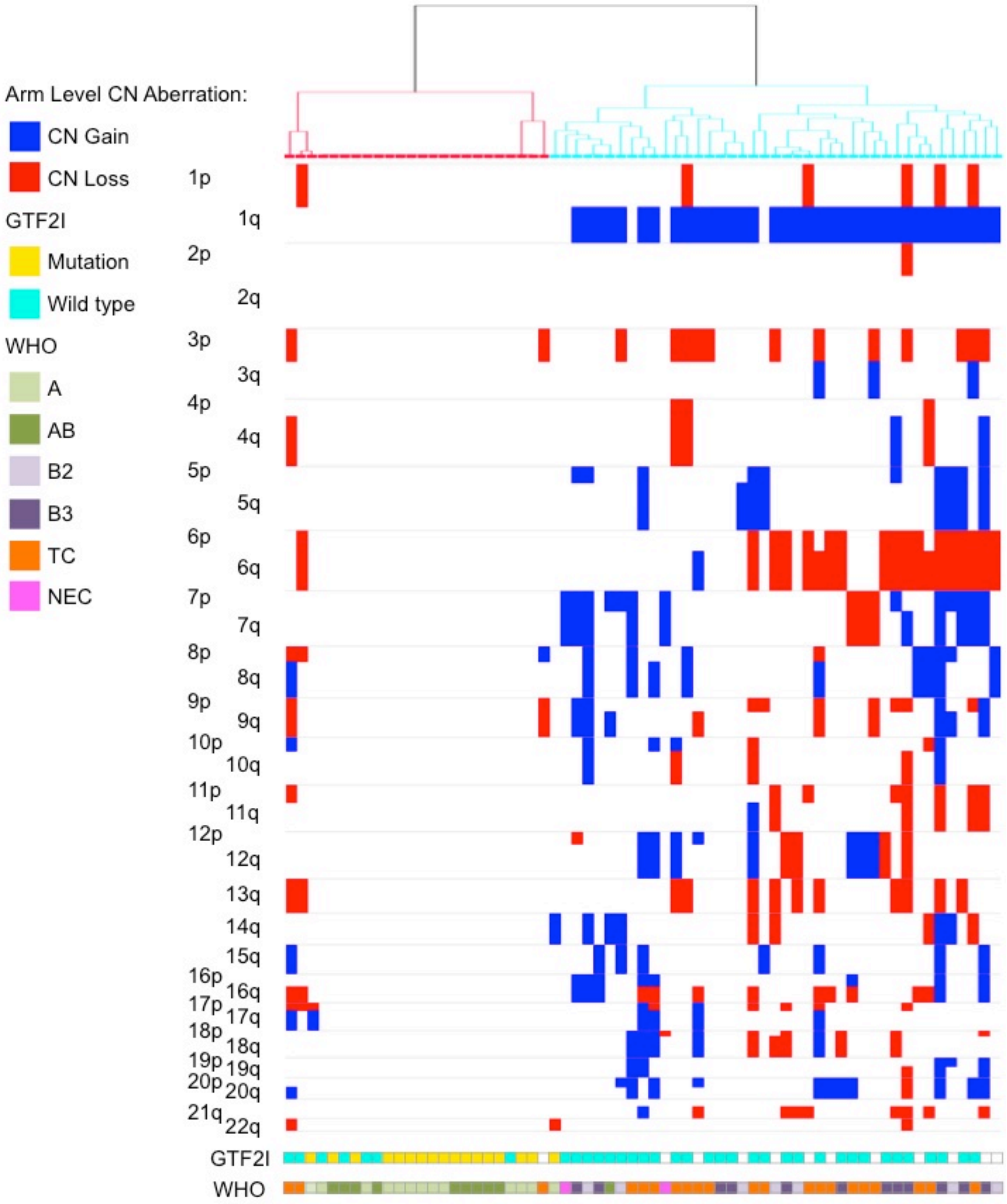
Supplementary Figures

Supplementary Figure 1



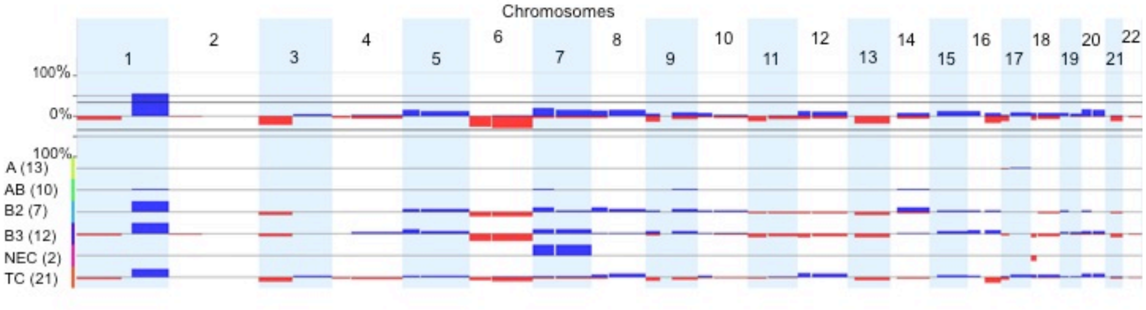
Supplementary Figure 1: Histological features of thymic epithelial tumors. According the 2004 WHO classification a clear-cut distinction has been defined between thymomas, organotypic tumors that mimic the structure of normal thymus, and thymic carcinomas (TC), more aggressive neoplasms that do not resemble the structure of normal thymus but that of carcinomas originating in other organs. **A** type thymomas present bland spindle/oval epithelial tumor cells with few or no lymphocytes. Grossly, they are usually encapsulated and easily separable from the surrounding organs even in case of tumors of conspicuous dimension. Type B thymomas show epithelial cells with a predominantly round or polygonal appearance. Type **B1** thymomas display tumor epithelial cells with very little atypia, scattered in a prominent population of immature non-neoplastic thymocytes that resemble the structure of normal thymus cortex. Type **B2** thymomas are characterized by large polygonal epithelial tumor cells arranged in a loose network containing numerous immature T lymphocytes. **B3** thymomas are composed of medium size round or polygonal epithelial tumor cells with slight atypia; these cells are mixed with a minor component of intraepithelial thymocytes. **AB** thymomas are composed of a lymphocyte-poor type A and a more lymphocyte-rich type B component. Thymic carcinomas (**TC)** are named according to their histological appearance being the squamous cell carcinoma and the undifferentiated carcinoma the most common.

Supplementary Figure 2



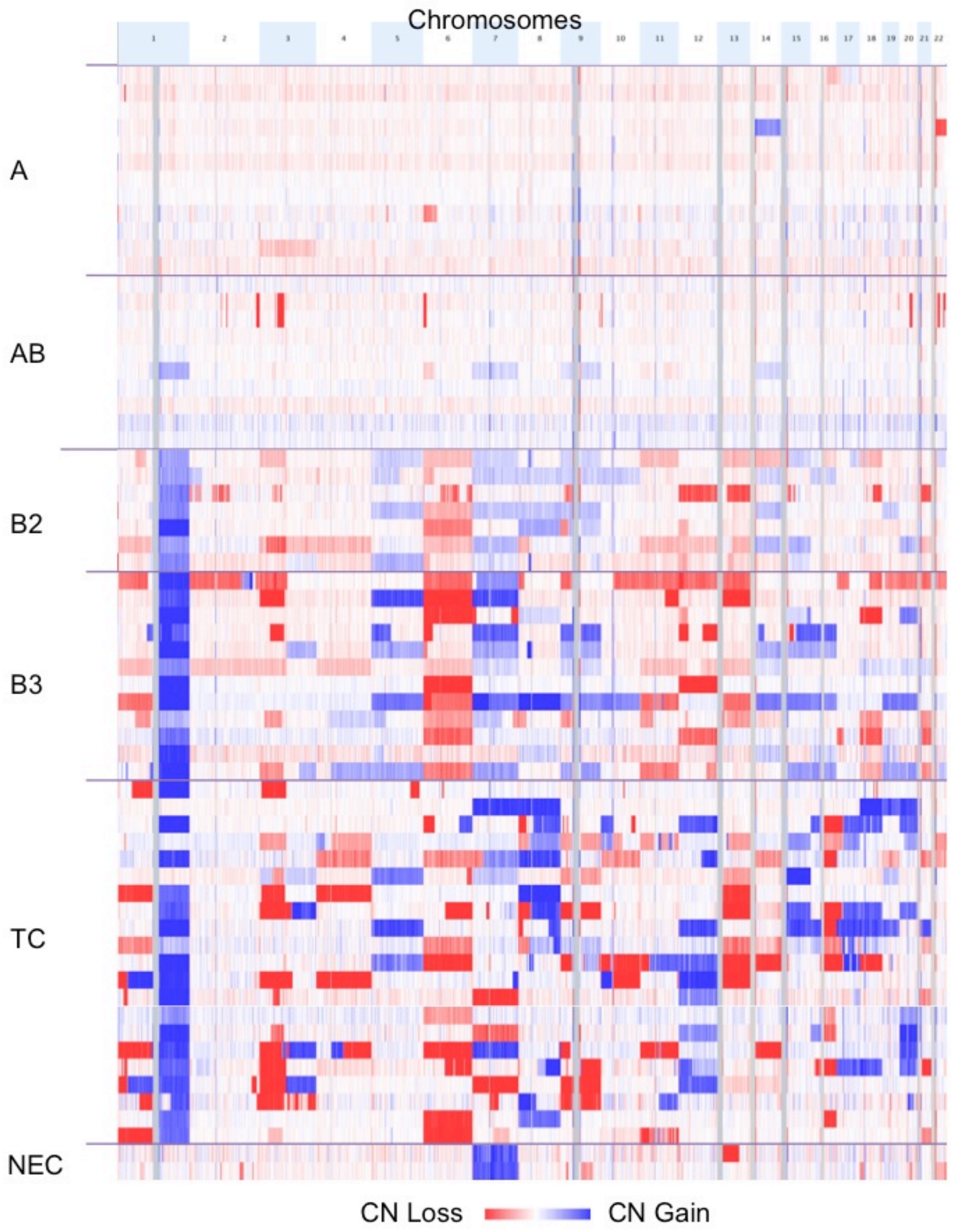
Supplementary Figure 2: Arm level copy number aberrations identify two clusters of tumors. Using array CGH, copy number aberrations were identified. An arm level copy number aberration was defined such as one event of copy number aberration that involves more than 80% of a chromosome arm (see Supplementary Methods for details). These aberrations defined two clusters of TETs: one with few arm level copy number aberrations and one rich in arm level copy number aberrations. These clusters trend to correlate with WHO histotypes and with the presence *GTF2I* mutations. There is a sub-group of samples of cluster one, to the left of the picture that presents copy number aberrations. In these samples, the copy number losses were more or equally abundant than the copy number gains, which may possibly explains the location of this subgroup in cluster 1.

Supplementary Figure S3



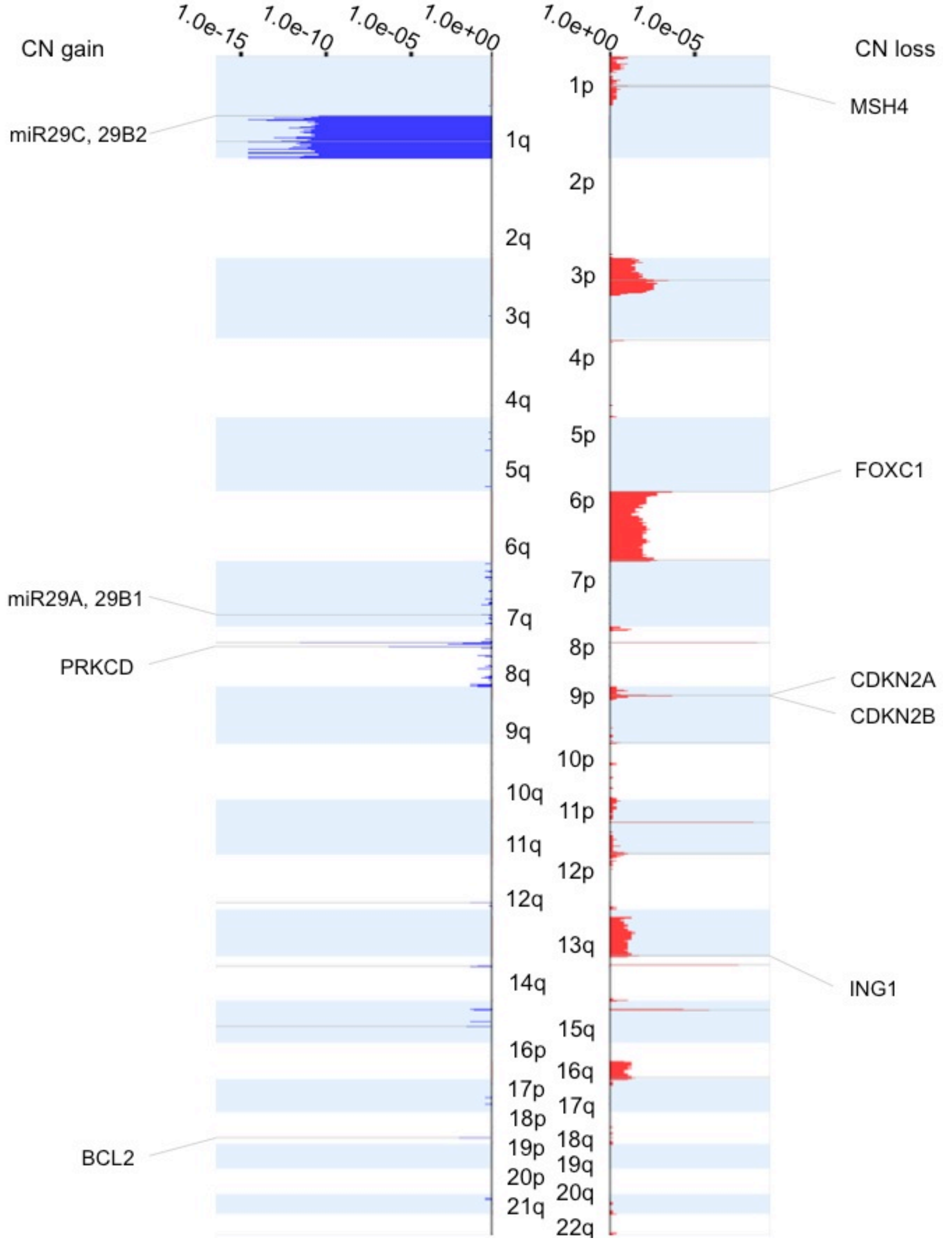
Supplementary Figure 3: Frequency of copy number aberrations is reported for all thymic epithelial tumors in the top part of the figure; the lower part of the figure reports the data by WHO histotype. copy number gain are in blue and copy number loss in red.

Supplementary Figure 4



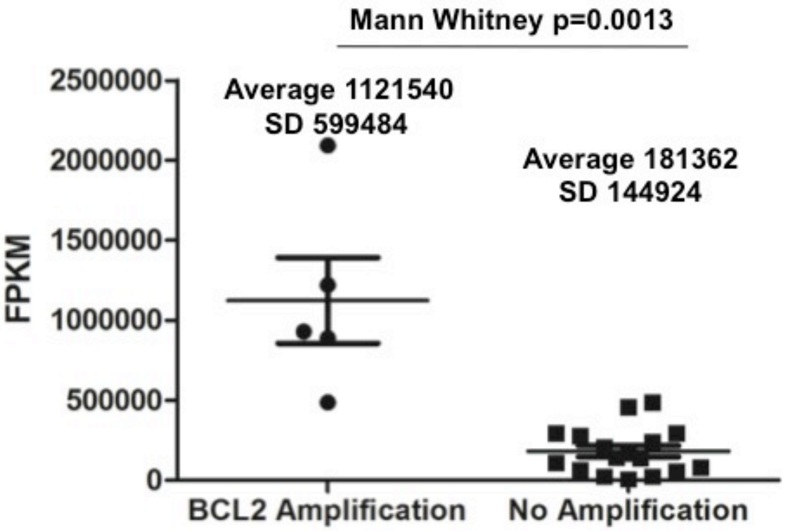
Supplementary Figure 4: Copy number aberrations at the probe set level. An estimation of copy number gain (blue) and loss (red) is reported for each probe in each tumor. The intensity of the colors represents the extent of copy number aberration predicted by the CGH arrays. TETs were grouped according to their histotypes.

Supplementary Figure 5



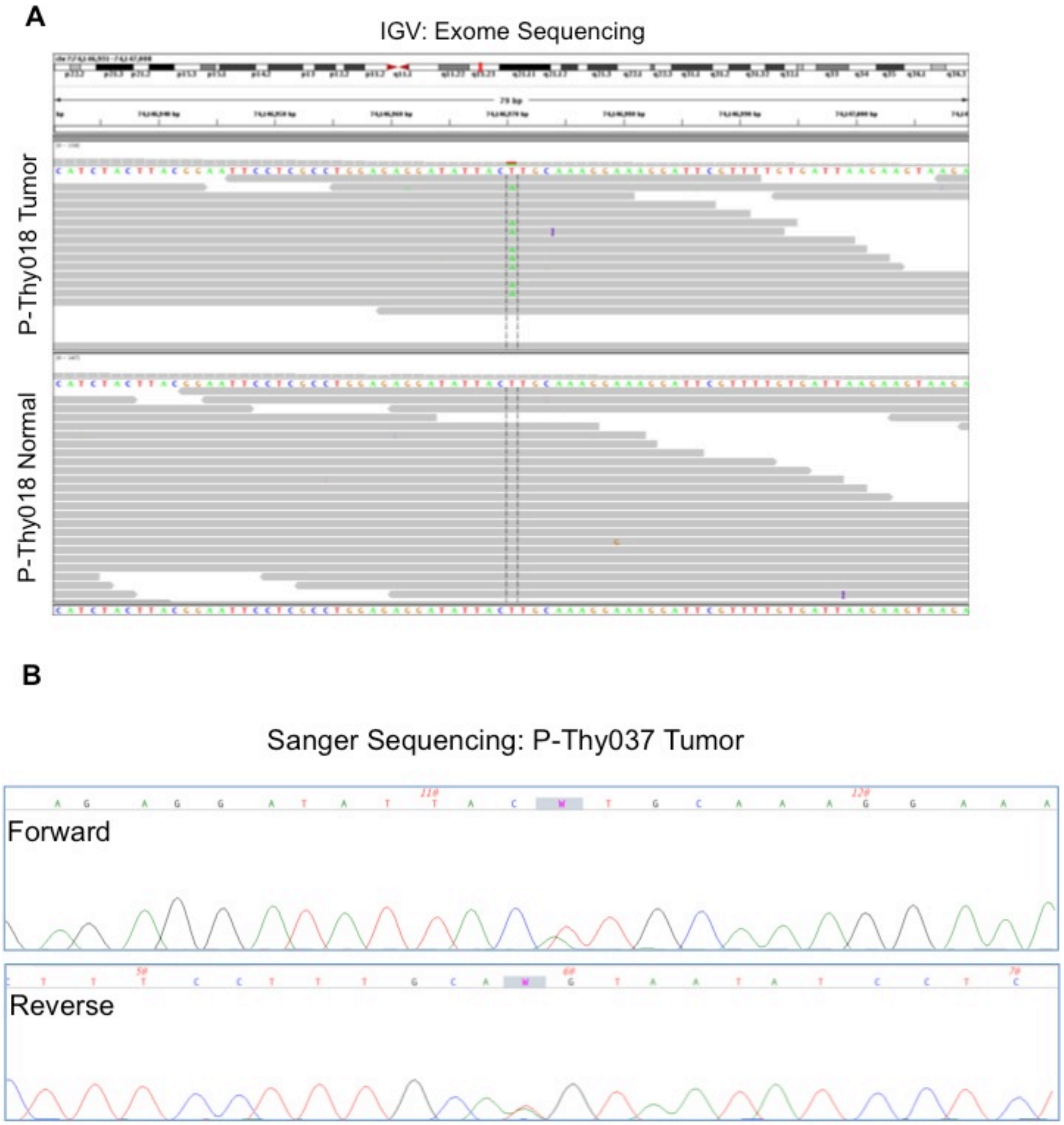
Supplementary Figure 5: GISTIC analysis identifies significant peaks of copy number gain and loss. GISTIC algorithm was applied to CGH data from 65 TETs in order to identify regions of copy number gain and loss that are candidate drivers of tumor growth. Significant peaks are labeled and their details are reported in the Supplementary Table 3.

Supplementary Figure 6



Supplementary Figure 6: Estimation of BCL2 mRNA expression using transcriptome sequencing (FPKM values) in tumors with (n=5) and without (n=17) *BCL2* amplification.

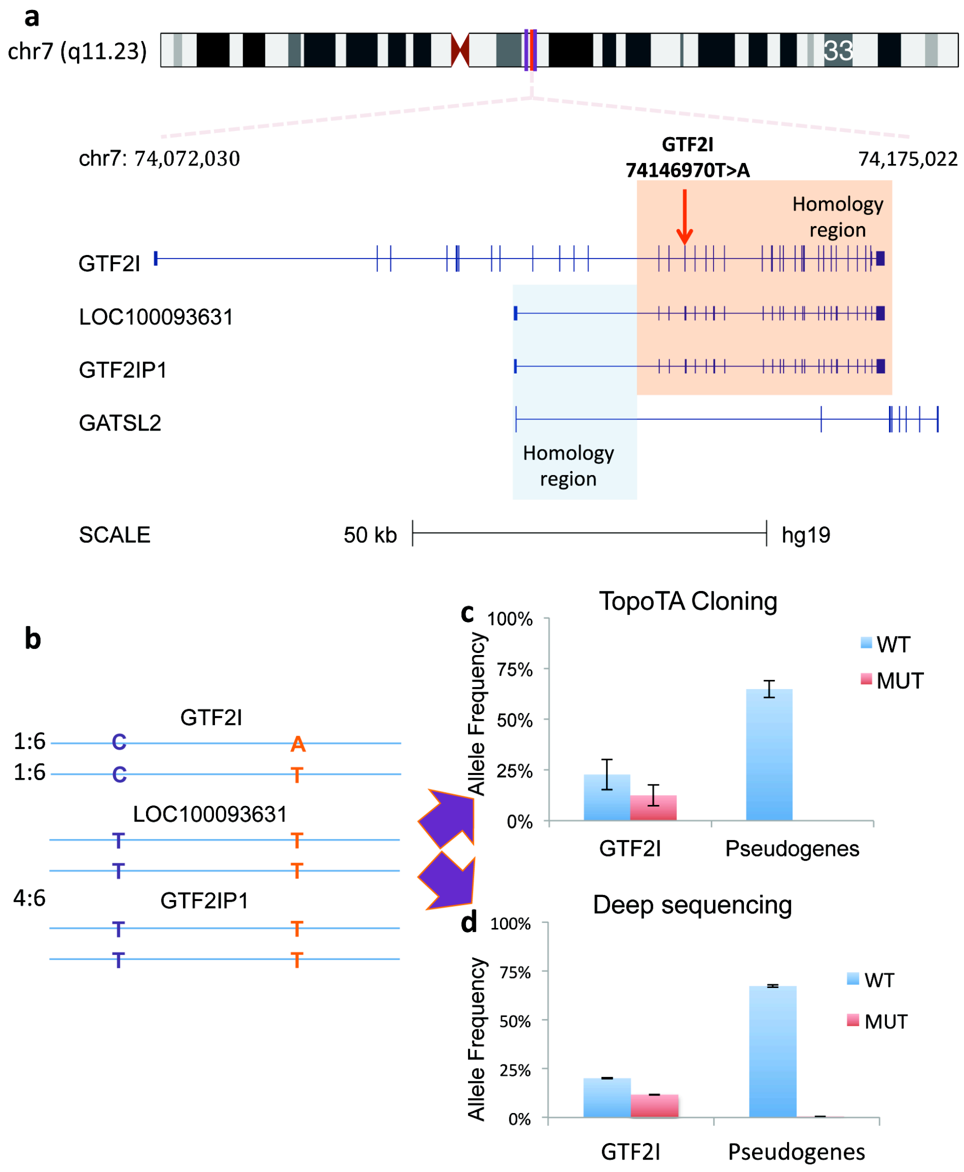
Supplementary Figure 7



Supplementary Figure 7: Representative *GTF2I* T>A mutation in whole exome and Sanger sequencing. In exome sequencing (A) representative results from tumor and normal DNA of one patient, depicting the T>A mutation. The position of the mutation is shown on chromosome 7 in the top panel; the sequence of part of *GTF2I* exon 15 is represented in both tracks, and in the tumor track there are appreciable reads (the gray bars) carrying the mutated A. Mutated reads are not present in normal genomic material. (B)

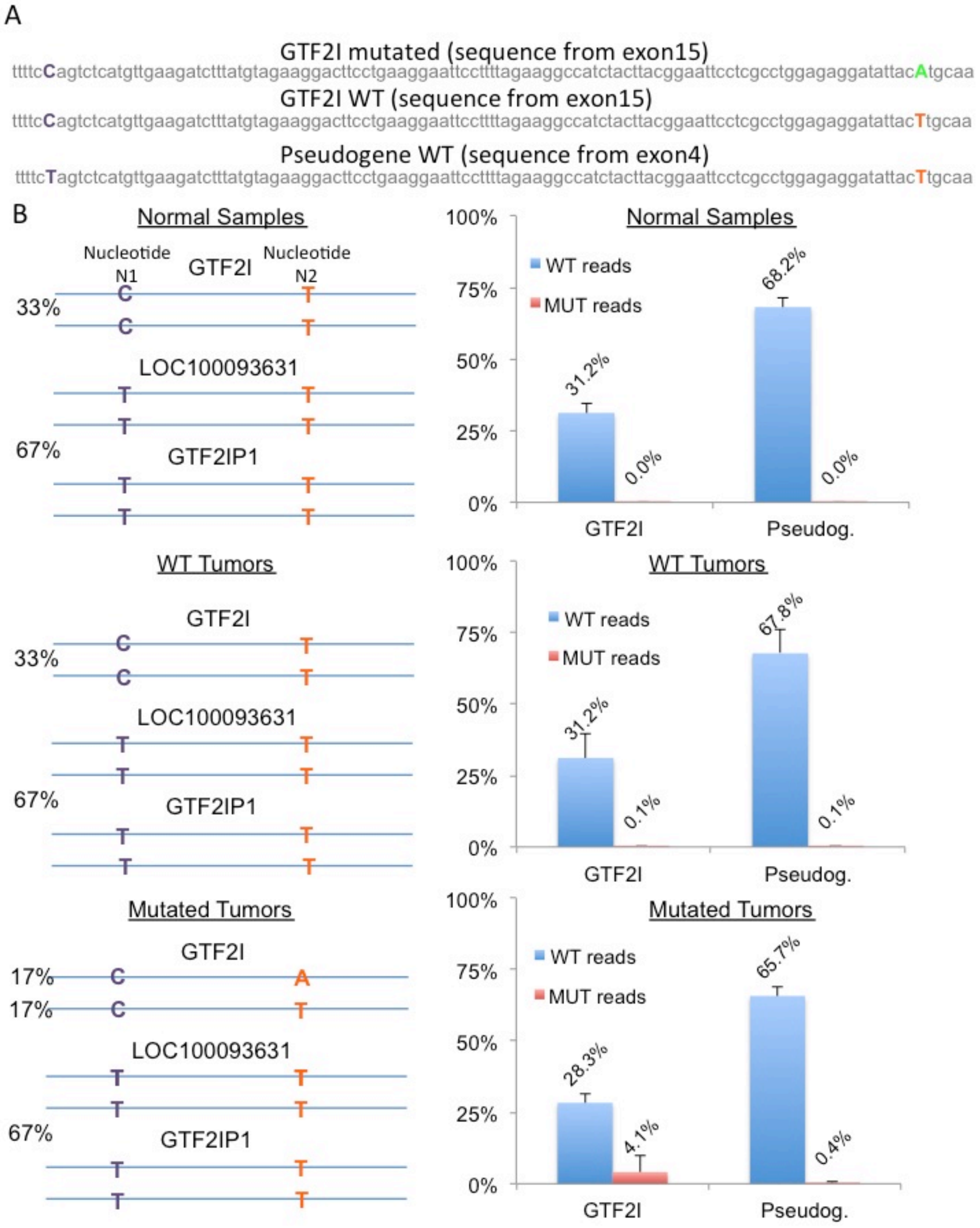
Representative *GTF2I* mutation in a Sanger pherogram of a type A thymoma: the forward and the reverse sequence are the top and bottom panels, respectively. The mutation T>A in forward and A/T in reverse is heterozygous and therefore identified by the presence of 2 peaks in that position.

Supplementary Figure 8



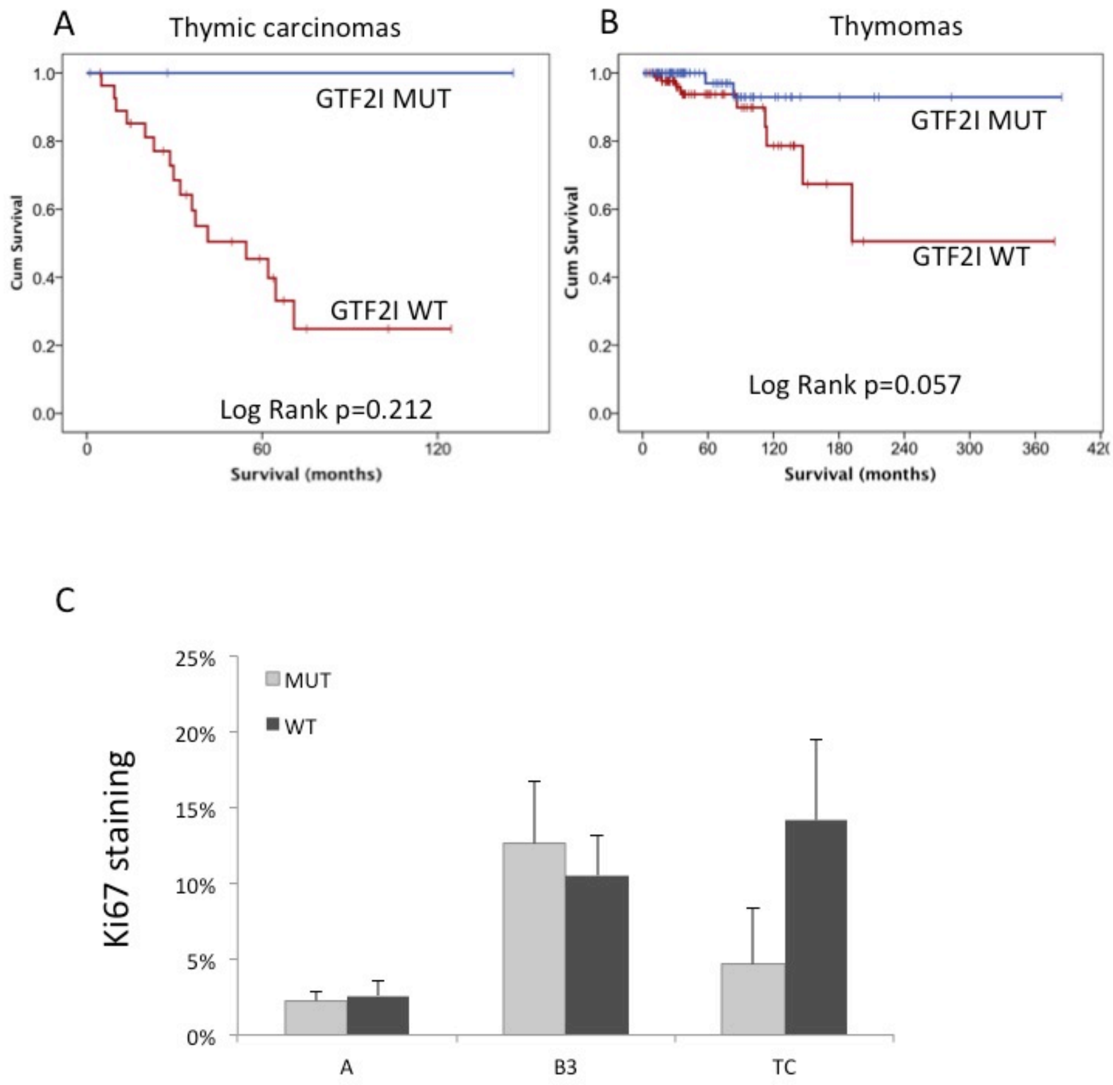
Supplementary Figure 8: Identification of mutations in the *GTF2I* gene but not in the *GTF2I* pseudogenes. (A) *GTF2I* and pseudogenes loci on chromosome 7 and their homology region. Homology regions of *GTF2i* and its pseudogenes are highlighted in orange. The first 12 exons of *GTF2I* have a unique sequence; whereas the first exons of the pseudogenes share homology sequences with *GAST* gene family (highlighted in light blue) that is composed of *GAST*, *GASTL1* and *GASTL2*: for convenience we show only *GASTL2* in the Figure. (B) *GTF2I* mutation (T>A) is mapped on exon 15. This region matches exon 4 of the pseudogenes, and differs by only 1 nucleotide: C in *GTF2I* and T in pseudogene sequences. The schema describes the allele frequencies theoretically present in a cell: one *GTF2I* mutated allele (1:6, ~17%), one GTF2I WT alleles (1:6, ~17%) and 4 pseudogenes wild type alleles (4:6, ~67%). (C) TopoTA cloning performed in 4 tumors with *GTF2I* mutation. Results are reported as average of identified allele frequencies. Sequencing of cloned amplicons identified the mutation only in *GTF2I* but not in the pseudogenes. (D) Deep sequencing performed on 5 tumors with *GTF2I* T>A mutation. The mutation was found in *GTF2I* only and not the pseudogenes, which equals to the mutation rate of ~17% or 1 out of 6 alleles (2 *GTF2I* + 4 pseudogene alleles). The mutation was not identified in the negative controls (data not shown, for details see Supplementary Table 8).

Supplementary Figure 9



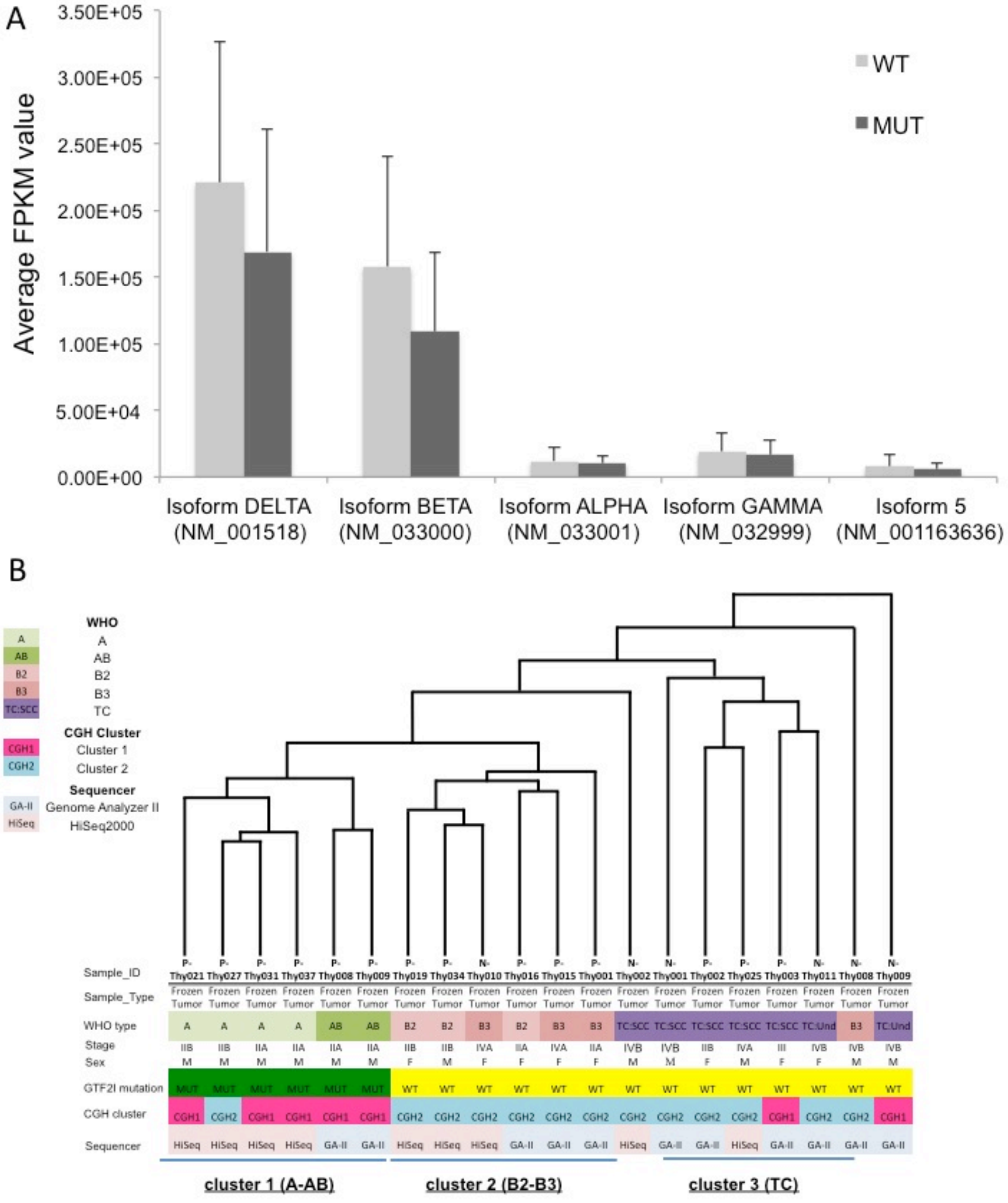
Supplementary Figure 9: *GTF2I* exon 15 and pseudogene exon 4 sequences, and their distribution in normal samples, tumors of *GTF2I* WT and *GTF2I* mutant. (A) Sequence of *GTF2I* exome 15: wild type or mutant; wild type sequence of the exon 4 of the pseudogenes. The marker that distinguishes *GTF2I* from the pseudogenes is reported in purple (nucleotide N1). The site of mutation is reported in orange when WT and in green when mutated (nucleotide N2). (B) Distribution of *GTF2I* and pseudogene reads carrying the T>A mutation in normal DNA (n=13), *GTF2I* WT (n=131) and mutated tumors (n=105) including all the samples characterized using the deep sequencing assay (n=250). In the group of *GTF2I* mutated tumors, the frequency of the mutated *GTF2I* allele was reduced compared to the expected 17% in a population of exclusively cancer cells. This result was expected, since samples rich in non-neoplastic thymocytes were included in this group.

Supplementary Figure 10



Supplementary Figure 10: Kaplan-Meier survival curves of patients with *GTF2I* mutated (in blue) and WT tumors (in red) with (A) thymic carcinomas (n=33) or (B) thymomas (n=171). The curves were compared using Log-Rank test. In thymic carcinoma the 10-year survival rate was 100% for *GTF2I* mutated cases (only 3 tumors) and 47% in WT tumors (n=30). In thymomas, the 10-year survival rate was 81% and 94% for *GTF2I* mutated and WT tumors. (C) Estimation of the fraction of proliferating cells in thymic carcinomas, A and B3 thymomas by immunohistochemistry with an anti-Ki67 antibody performed on FFPE slides. The number of cells, positive for Ki67, was similar between WT and mutated thymomas, both in A (n=5 and 11, respectively) and B3 histotypes (n=17 and 8, respectively). In thymic carcinomas Ki67 was lower in *GTF2I* mutated (n=3) tumors compared to WTs (n=7).

Supplementary Figure 11



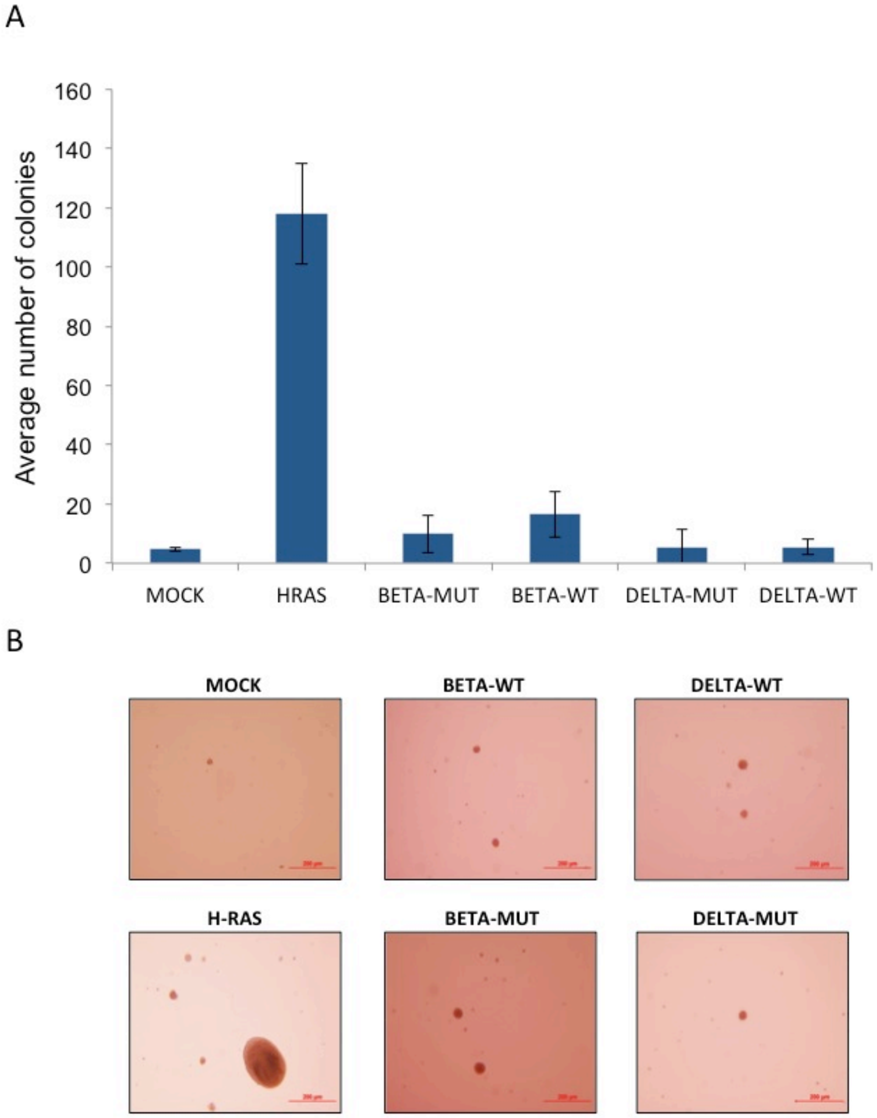
Sample_ID	P-Thy021	P-Thy027	P-Thy031	P-Thy037	P-Thy008	P-Thy009	P-Thy019	P-Thy034	N-Thy010	P-Thy016
Sample_Type	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor
WHO type	A	A	A	A	AB	AB	B2	B2	B3	B2
Stage	IIB	IIB	IIA	IIA	IIA	IIA	IIB	IIB	IVA	IIA
Sex	M	M	M	M	M	M	F	M	F	F
GTF2I mutation	MUT	MUT	MUT	MUT	MUT	MUT	WT	WT	WT	WT
CGH cluster	CGH1	CGH2	CGH1	CGH1	CGH1	CGH1	CGH2	CGH2	CGH2	CGH2
Sequencer	HiSeq	HiSeq	HiSeq	HiSeq	GA-II	GA-II	HiSeq	HiSeq	HiSeq	GA-II

Sample_ID	P-Thy015	P-Thy001	N-Thy002	N-Thy001	P-Thy002	P-Thy025	P-Thy003	N-Thy011	N-Thy008	N-Thy009
Sample_Type	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor	Frozen Tumor
WHO type	B3	B3	TC:SCC	TC:SCC	TC:SCC	TC:SCC	TC:SCC	TC:Und	B3	TC:Und
Stage	IVA	IIA	IVB	IVB	IIB	IVA	III	IVB	IVB	IVB
Sex	F	F	M	M	F	M	F	F	M	M
GTF2I mutation	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
CGH cluster	CGH2	CGH2	CGH2	CGH2	CGH2	CGH2	CGH1	CGH2	CGH2	CGH1
Sequencer	GA-II	GA-II	HiSeq	GA-II	GA-II	HiSeq	GA-II	GA-II	GA-II	GA-II

Supplementary Figure 11: Differential expression of isoforms and clustering of transcriptome data.

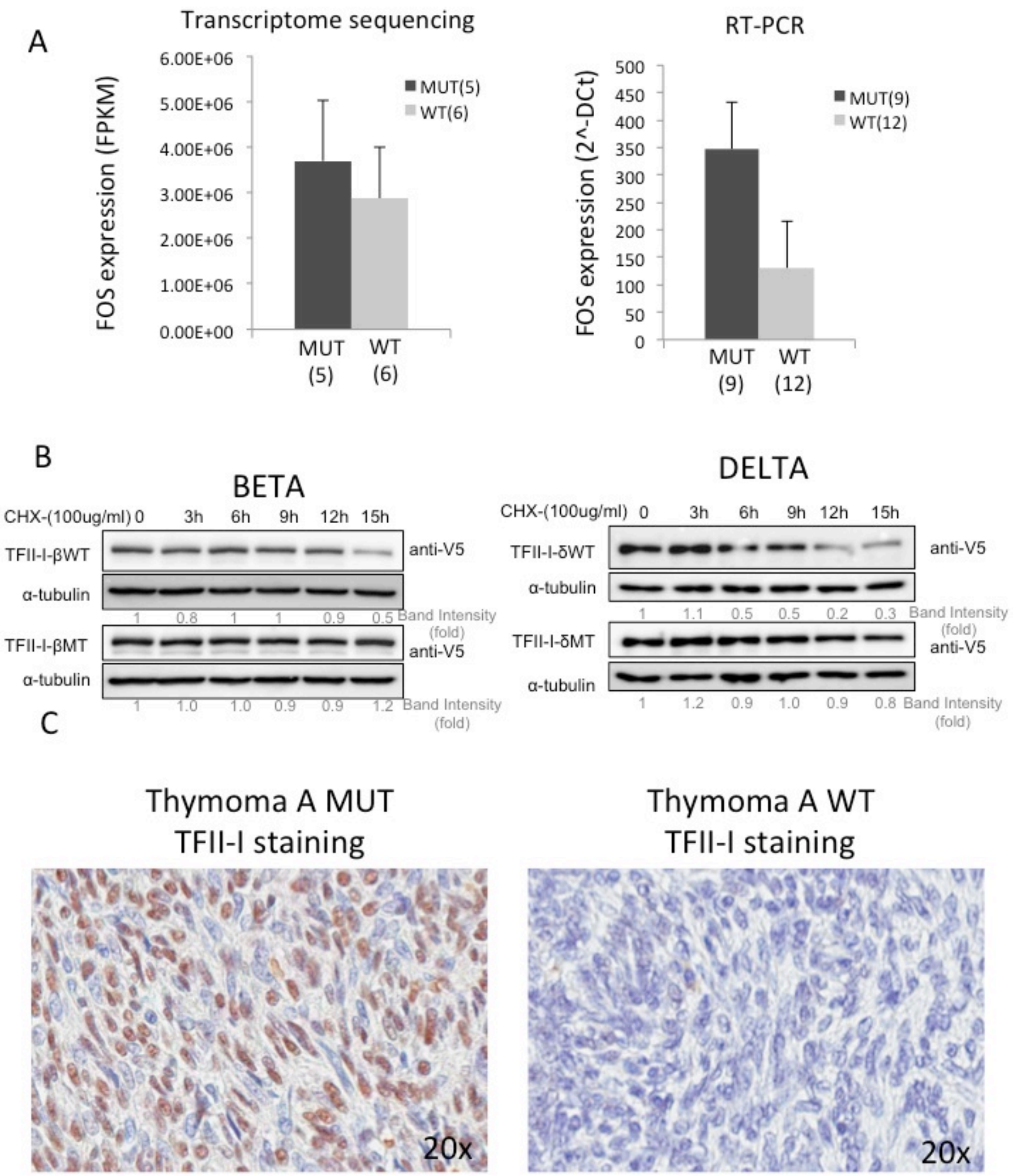
Cufflinks FPKM values were calculated for each isoform of *GTF2I* and their average value were reported for wild type (WT: 18 samples) and mutated (MUT: 7 samples) cases. The differential expression was compared using a non-parametric test (Kruskal-Wallis) and the Dunn's post hoc test that demonstrated significant differences between δ-isoforms compared with α, γ and isoform 5. Similar results were observed for β-isoform. (B) Cufflinks FPKM values were used to cluster TETs with their gene expression data. The clusters trend to segregate TETs according to their histotype. CGH cluster (CGH1 in pink and CGH2 in light blue) and *GTF2I* mutation status (MUT: Mutant in green and WT: wild type in yellow) also parallel the expression clusters. The two different platforms used to define the profile of gene expression (Genome Analyzer-II (GA-II) and HiSeq2000 (HiSeq) provided equal distribution among expression clusters.

Supplementary Figure 12



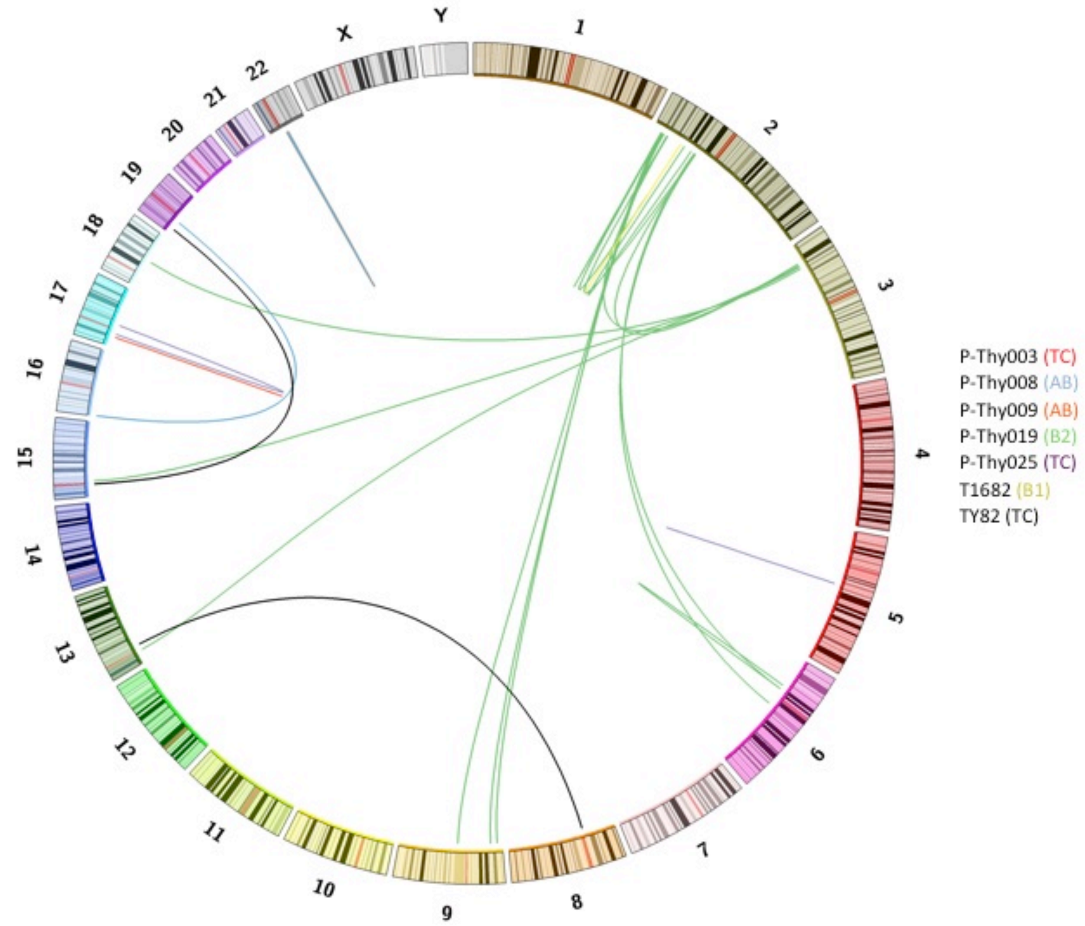
Supplementary Figure 12: Soft agar assay. (A) Average number of colonies of NIH-3T3 cells transfected with negative control (mock-construct), positive control (HRAS^{V12G}), TFII-I β-isoform WT and mutated p.(Leu404His) and δ-isoform of TFII-I WT and mutated p.(Leu383His). For positive and negative controls, results are the average of three experiments. For β- and δ- isoforms, results are the average of three experiments derived from 4 different pool transfectants. Vertical bars represent the standard deviation of triplicate experiments. (B) Pictures of soft agar colonies (5x magnification).

Supplementary Figure 13



Supplementary Figure 13: FOS expression, TFII-I protein stability and its expression in type A thymomas (A) Protein synthesis was inhibited using cycloheximide and cells were harvested at the indicated time points. Proteins were extracted and the amount of TFII-I evaluated by western blot. HELA cells transfected with mutated β-isoform had a more stable TFII-I than those transfected with the WT β-isoform. Similar results were observed with mutated and WT δ-isoforms. (B) Representative TFII-I immunohistochemistry images. Immunohistochemistry performed using anti-TFII-I antibody (not specific for β- or δ- isoform) demonstrated a higher expression in mutated type A thymomas (n=11) than WTs (n=4). The pattern of expression was predominantly nuclear.

Supplementary Figure 14



Supplementary Figure 14: Summary of fusion genes identified by transcriptome sequencing and confirmed by RT-PCR. The detected fusion genes were reported using Circos. Different colors indicate fusion genes of different TET patients. Details of the identified fusions are reported in Supplementary Table 13.

Supplementary Tables

Supplementary Table 1: Summary of Patient characteristics

		Total	*GTF2I* Sequenced	MUT *GTF2I*	WT *GTF2I*	p-value
Total in study		286	274	43%	57%	
Patients		282	270	44%	56%	
Age	median 56				range (20-86)	
Sex	Female	139	135	41%	59%	0.327
	Male	143	135	47%	53%	
	uk	4				
WHO	A	58	56	82%	18%	p<0.001*
	AB	55	54	74%	26%	
	B1	28	28	32%	68%	
	B2	33	32	22%	78%	
	B3	65	62	21%	79%	
	TC	41	36	8%	92%	
	NEC	4	4	0	100%	
	Micronodular	2	2	50%	50%	
Stage	I	41	40	58%	42%	p<0.001**
	IIA	55	53	64%	36%	
	IIB	73	71	51%	49%	
	III	29	29	35%	65%	
	IVA	21	19	16%	84%	
	IVB	34	32	6%	94%	
	uk	33	30	37%	63%	
Resection	R0	139	136	49%	51%	p=0.0267
	R1	18	18	33%	67%	
	R2	13	13	15%	85%	
	uk	116	107	41%	59%	
Paraneoplastic Syndromes	All	66	65	43%	57%	p=0.636***
	Myasthenia	63	62	45%	55%	
	No	145	136	39%	61%	
	uk	75	73	52%	48%	
CGH		65	53	32%	68%	
Whole Exome Sequencing		28	28	21%	79%	
Transcriptome Sequencing		25	25	28%	72%	
197-gene Re-sequencing		52		-	-	
Sanger Sequencing		199	199	61%	39%	
***GTF2I* Deep Sequencing**		250	250	42%	58%	

Samples Sequenced for *GTF2I*	274	43%	57%
Thymic carcinoma (TC)			

Supplementary Table 2: available in a separate file.xlsx

Supplementary Table 3: available in a separate file.xlsx

Supplementary Table 4: available in a separate file.xlsx

Supplementary Table 5: available in a separate file.xlsx

Supplementary Table 6: available in a separate file.xlsx

Supplementary Table 7: Summary of TopoTA cloning results: *GTF2I*/pseudogenes sequence

		GTF2I	*GTF2I*	Pseudogenes	Pseudogenes
	Clonies#	MUT	WT	MUT	WT
Expected ratio	-	*1:6*	*1:6*	*0*	*4:6*
Expected frequencies	-	*17%*	*17%*	*0%*	*67%*
P-Thy037	17	18%	12%	0%	70%
P-Thy021	36	11%	25%	0%	64%
P-Thy018	40	15%	25%	0%	60%
P-Thy027	17	6%	29%	0%	65%
Total	110	12%	23%	0%	65%
	average	12%	23%	0%	65%
	SD	5%	7%	0%	4%

Supplementary Table 8: available in a separate file.xlsx

Supplementary Table 9: available in a separate file.xlsx

Supplementary Table 10: Logistic regression analysis of *GTF2I* mutations in TETs

	Coeff.	**Standard Error**	**p-Value**	**Odds Ratio**	**95%CI-Lower**	**95%CI-Upper**
WHO hystotype (A,AB,B1vsB2,B3,TC)	2.433	.423	.000	11.391	4.976	26.078
Stage (I-IIvsIII-IV)	.255	.505	.614	1.290	.479	3.473
Resection R0-R1vsR2	.915	.959	.340	2.497	.381	16.370
Constant	-2.648	.888	.003	.071		
No-model prediction capacity	54.80%					
Full model prediction capacity	76.50%					
Nagelkerke R Square	0.385					
Hosmer and Lemeshow test	p=0.592					

Supplementary Table 11: Construction of the Cox Proportional hazard model

	Gropu1	Group2 (Ref)	p-value	HR	95%CI-Low	95%CI-High
Univariate Analysis						
WHO	A, AB, B1	B2, B3, TC	0.004	0.53	0.007	0.387
Stage	I-II	III-IV	<0.001	0.08	0.024	0.268
Resection	R0-R1	R2	0.002	0.17	0.057	0.51
GTF2I	Mutant	Wild type	0.002	0.1	0.024	0.42
Bivariate Analysis						
GTF2I-WHO						
GTF2I	Mutant	Wild type	0.053	0.228	0.051	1
WHO	A, AB, B1	B2, B3, TC	0.038	0.111	0.014	0.886
GTF2I-Stage						
GTF2I	Mutant	Wild type	0.026	0.1	0.013	0.755
Stage	I-II	III-IV	0.001	0.135	0.04	0.46
GTF2I-Resection						
GTF2I	Mutant	Wild type	0.023	0.094	0.012	0.725
Resection	R0-R1	R2	0.012	0.244	0.081	0.733
WHO-Stage						
WHO	A, AB, B1	B2, B3, TC	0.025	0.097	0.013	0.743
Stage	I-II	III-IV	0.002	0.145	0.042	0.495
WHO-Resection						
WHO	A, AB, B1	B2, B3, TC	0.017	0.082	0.011	0.634
Resection	R0-R1	R2	0.013	0.248	0.082	0.745
Stage-Resection						
Stage	I-II	III-IV	0.011	0.164	0.041	0.656
Resection	R0-R1	R2	0.207	0.464	0.141	1.53
Multivariate						
GTF2I-Stage-Resection						
GTF2I	Mutant	Wild type	0.209	0.227	0.023	2.291
Stage	I-II	III-IV	0.008	0.182	0.051	0.641
WHO	A, AB, B1	B2, B3, TC	0.242	0.244	0.023	2.592

Supplementary Table 12: available in a separate file.xlsx

Supplementary Table 13: available in a separate file.xlsx

Supplementary Table 14: available in a separate file.xlsx

Supplementary Table 15: available in a separate file.xlsx

Supplementary Table 16: available in a separate file.xlsx