STMN2 Cryptic Analysis

Introduction

TDP43-dependent cryptic STMN2 events have been detected in cancer patients. These patients have been sequenced and their cryptic reads and clinical data stored in the TCGA database. We wanted to know if these STMN2 cryptic events are expressed more in particular subset(s) of cancers and if these cancers with TDP43-dependent cryptic splicing show enrichment of mutations in particular genes and pathways.

AL provided the specific genomic coordinates of TDP43-dependent STMN2 cryptic and annotated reads. A function was created to query junctions in the TCGA database for reads with the specified coordinates:

- Gene name = STMN2
- Snapcount coordinates (cryptic) = chr8:79611215-79616821
- Snapcount coordinates (annotated) = chr8:79611215-79636801
- Strand code = +

The resulting data included genomic information, sample-related clinical data and junction coverage information for cryptic and annotated reads, separately, and were read in as tables.

The STMN2 cryptic and annotated query tables were joined into a single dataframe (**STMN2_query**) and a "jir" (junction inclusion ratio) column was added to reflect the fraction of cryptic reads for each case:

A case set was made on TCGA containing all cancer patients with these TDP43-dependent STMN2 events and their clinical data downloaded and joined with the existing dataframe to make the **STMN2_clinical_jir** dataframe.

In order to investigate cryptic STMN2 expression in these patients, a new dataframe (**STMN2_clinical_jir_cryptic**) was made by filtering for patients with cryptic counts greater than 2. This is an arbitrary cut-off used for all cryptic events investigated.

Cancers with STMN2 expression (in general)

To visualise the STMN2 reads data, a bar chart was plotted to compare the expression of STMN2 across the different cancer types. This was plotted using all data (**STMN2_clinical_jir**) rather than the cryptic filtered dataframe, and it compares the absolute read counts.

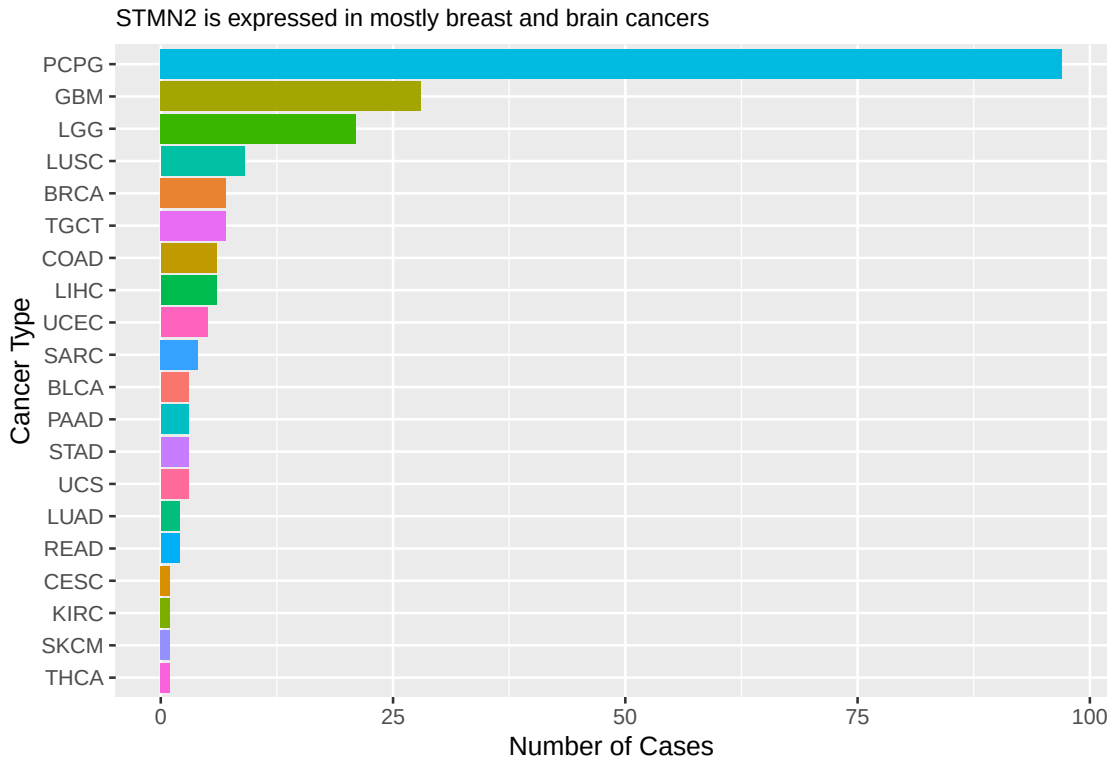

Figure 1: STMN2 is expressed in mostly breast and brain cancer patients.

Cancers with cryptic STMN2 events

The same bar chart was then plotted using the cryptic filtered dataframe (**STMN2_clinical_jir_cryptic**) to compare the expression of only the cryptic STMN2 events across the different cancer types. Again, this plot compares the absolute read counts of the cryptic events.

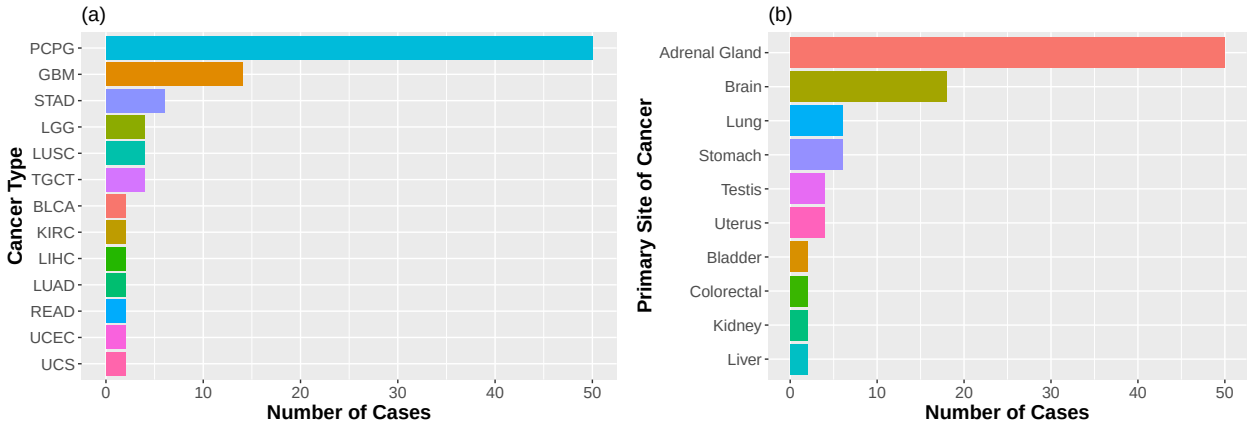

Figure 2: (a) Cryptic STMN2 events are found mostly in neuronal cancers: PCPG affects the nerve cells of the adrenal glands, and GBM affects the brain. PCPG = paraganglioma and pheochromocytoma; GBM = glioblastoma multiforme. (b) Cryptic STMN2 events are found most abundantly in cancers of the adrenal gland and brain.

Junction coverage

Overall STMN2 junction coverage was visualised in the different primary sites of cancers to see if the high cryptic read counts in adrenal gland and brain cancers were due to higher overall STMN2 coverage in those cancers (Figure 3a). Cryptic STMN2 junction coverage was also investigated by calculating 'reads per million' as the fraction of cryptic reads of the overall junction coverage (Figure 3b).

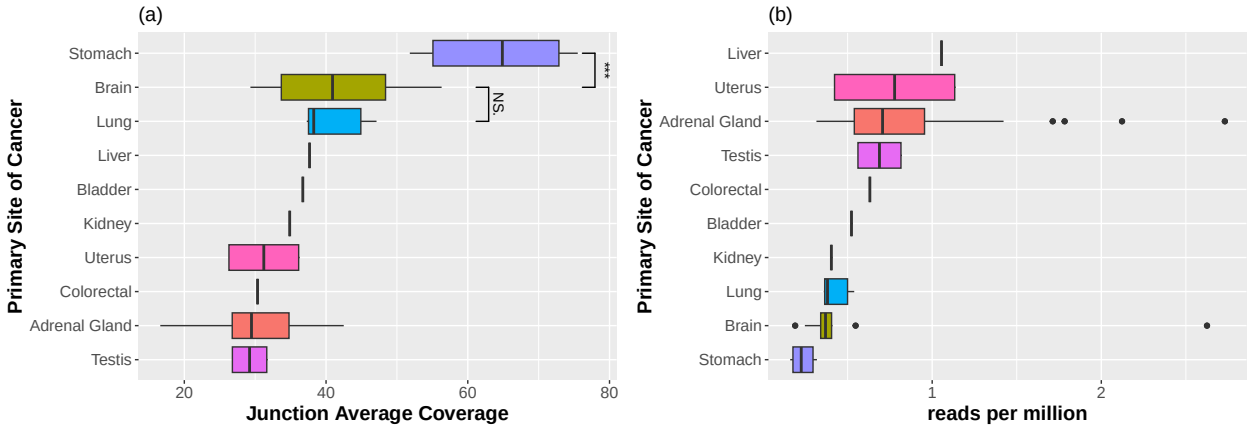

Figure 3: (a) Stomach and brain cancers are the most deeply sequenced. (b) Cancers of the uterus, adrenal gland and testis have the greatest cryptic coverage.

Indeed, the brain is deeply sequenced so it could be that the higher overall coverage resulted in a greater cryptic count for brain cancers. However, the adrenal gland is not as deeply sequenced, suggesting that the cryptic expression in adrenal gland cancers is significantly high.

There is only one liver cancer patient with TDP43-dependent cryptic STMN2 events, hence the single data value for liver. There are some extreme values for the adrenal gland and the brain, which indicates that it is few patients with a larger number of STMN2 cryptic reads that are influencing the overall cryptic coverage for these cancer subsets.

Which cancers have the most cryptic STMN2 events?

To investigate where we are seeing most of the cryptic STMN2 events, the number of cases of each cancer type (with cryptic STMN2) was weighted against the total number of cases with cryptic STMN2.

Table 1: PCPG and GBM cancers have high cryptic STMN2 expression.

cancer_abbrev	n	percent
PCPG	50	0.5208333
GBM	14	0.1458333
STAD	6	0.0625000
LGG	4	0.0416667
LUSC	4	0.0416667
TGCT	4	0.0416667
BLCA	2	0.0208333
KIRC	2	0.0208333
LIHC	2	0.0208333
LUAD	2	0.0208333
READ	2	0.0208333
UCEC	2	0.0208333

cancer_abbrev	n	percent
UCS	2	0.0208333

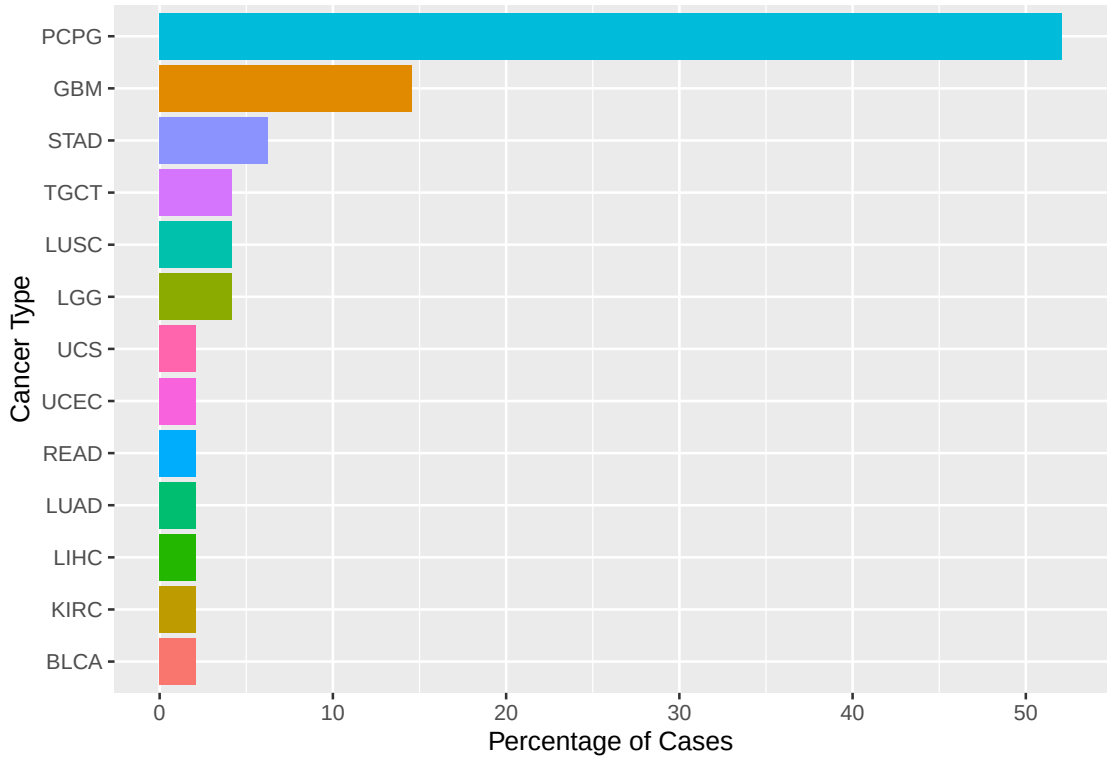

Figure 4: High proportion of cryptic STMN2 events are in PCPG and GBM cancers.

The majority of the cryptic STMN2 events are expressed in PCPG (47.2% of cryptic cases) and GBM (13.2% of cryptic cases) patients, consistent with previous results indicating high cryptic coverage in these cancers and their primary sites. It is important to note that the high coverage is influenced by patients exhibiting extreme values of cryptic coverage.

Where are the cancers with cryptic STMN2 events located?

A similar calculation was done to see where these cryptic events are occurring (i.e., where these cancers are located in the body). The number of cases of each primary cancer site (with cryptic STMN2) was weighted against the total number of cases with cryptic STMN2.

Table 2: Cancers with cryptic STMN2 events are found primarily in the adrenal gland and brain

gdc_cases_project_primary_site	n	percent
Adrenal Gland	50	0.5208333
Brain	18	0.1875000
Lung	6	0.0625000
Stomach	6	0.0625000
Testis	4	0.0416667
Uterus	4	0.0416667

gdc_cases_project_primary_site	n	percent
Bladder	2	0.0208333
Colorectal	2	0.0208333
Kidney	2	0.0208333
Liver	2	0.0208333

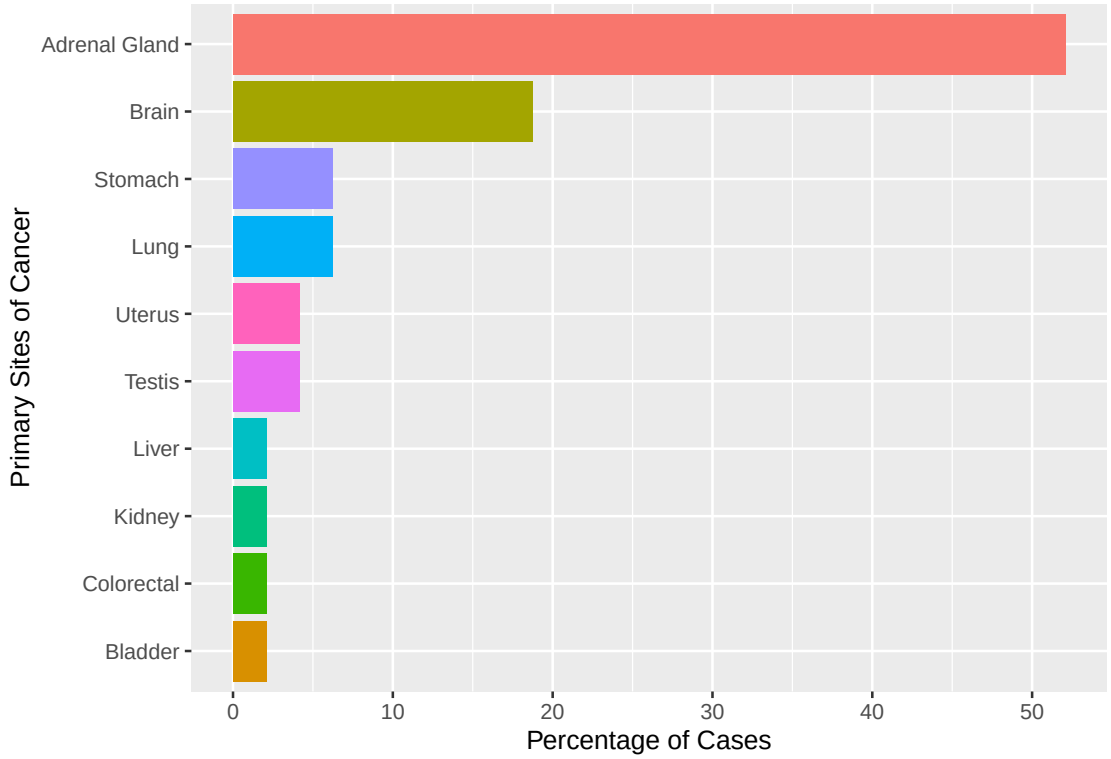

Figure 5: Cancers with cryptic STMN2 events are primarily in the adrenal gland and brain.

Consistent with all previous results so far, most of the cryptic STMN2 events seen in cancer patients are in cancers of the adrenal gland (52.1% of all cases) and brain (18.8% of all cases).

Mutational burden

All of the available clinical data for cancer patients on cBioPortal was downloaded and read in as a dataframe (**cBio_clinical**); this was **not** limited to only the patient with cryptic STMN2 expression. The resulting dataframe contains information on individual cancer patients: cancer type, survival, mutation count and aneuploidy score.

The **cBio_clinical** dataframe was joined to the existing dataframe containing the patients exhibiting cryptic STMN2 expression (**STMN2_clinical_jir_cryptic**). This join added the clinical data to only the relevant patients of interest (i.e., those with cryptic STMN2) to produce the **STMN2_cryptic_cBio** dataframe.

Fraction of each cancer that has cryptic STMN2 events

The burden of cryptic STMN2 expression was examined. This was done by looking at the fraction of each cancer type that exhibits TDP43-dependent cryptic STMN2 events (Figure 6).

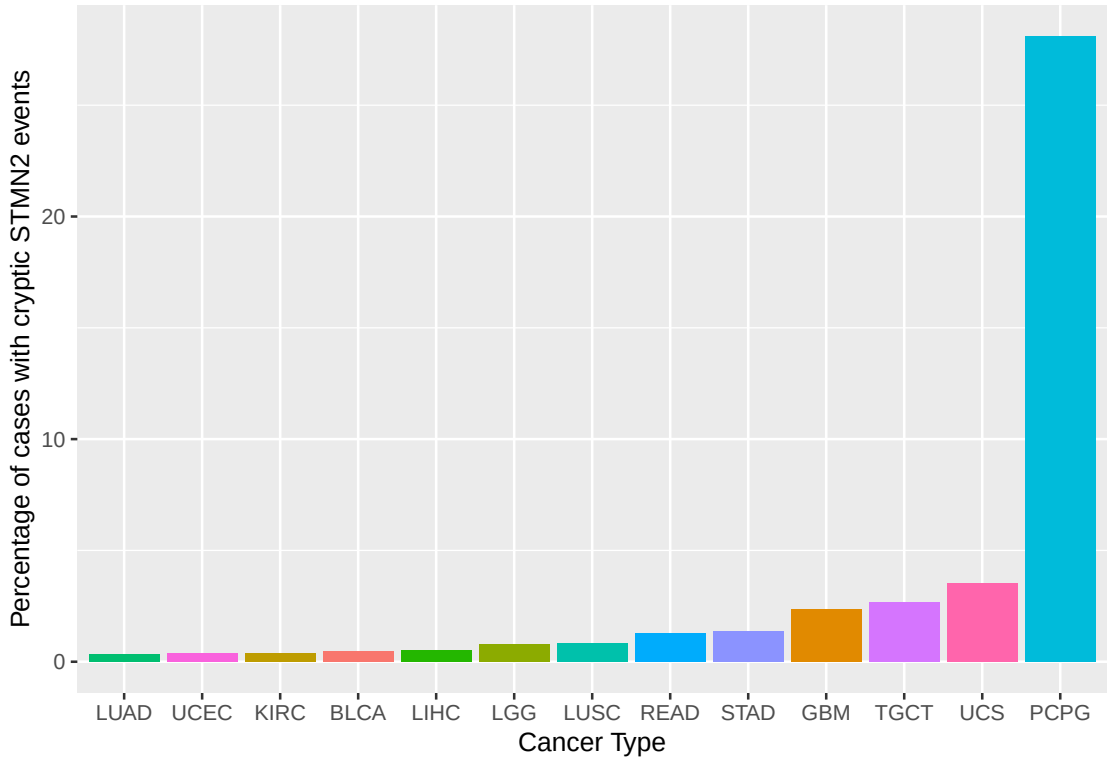

Figure 6: Fraction of each cancer type that has cryptic STMN2 expression.

The clinical data from cBioPortal provided information on the mutation counts in cancer patients of different cancer types. To investigate the cancer-by-cancer mutational burden in patients with cryptic STMN2, the total mutation count seen in patients with cryptic STMN2 expression was weighted against the total mutation count in cancer patients in general (for each cancer type). The results are displayed in Table 3.

Table 3: Mutational burden of cryptic STMN2 cases.

cancer_abbrev	total_mutations	total_mutations_cryptic	percent_with_cryptic
PCPG	1793	556	31.0094813
TGCT	2157	98	4.5433472
UCS	7303	118	1.6157743
LUSC	126568	1904	1.5043297
GBM	46100	602	1.3058568
LIHC	36216	286	0.7897062
LGG	27152	134	0.4935180
STAD	147953	676	0.4569019
BLCA	99709	356	0.3570390
LUAD	157145	374	0.2379968
UCEC	538960	120	0.0222651

Table 3 shows that 31% of the mutations in PCPG cancer patients are seen in cases with cryptic STMN2. This is the largest proportion out of all cancer subsets, with 7.3% of mutations in TGCT cancer patients being in those with cryptic STMN2 expression.

Mutational burden was further examined within each cancer type, using boxplots to visualise the mutation counts in non-cryptic cases and cases with cryptic STMN2 expression (Figure 7). Only cancers with cryptic

STMN2 expression are plotted. This analysis aimed to determine whether cases with cryptic STMN2 events have a greater mutational burden.

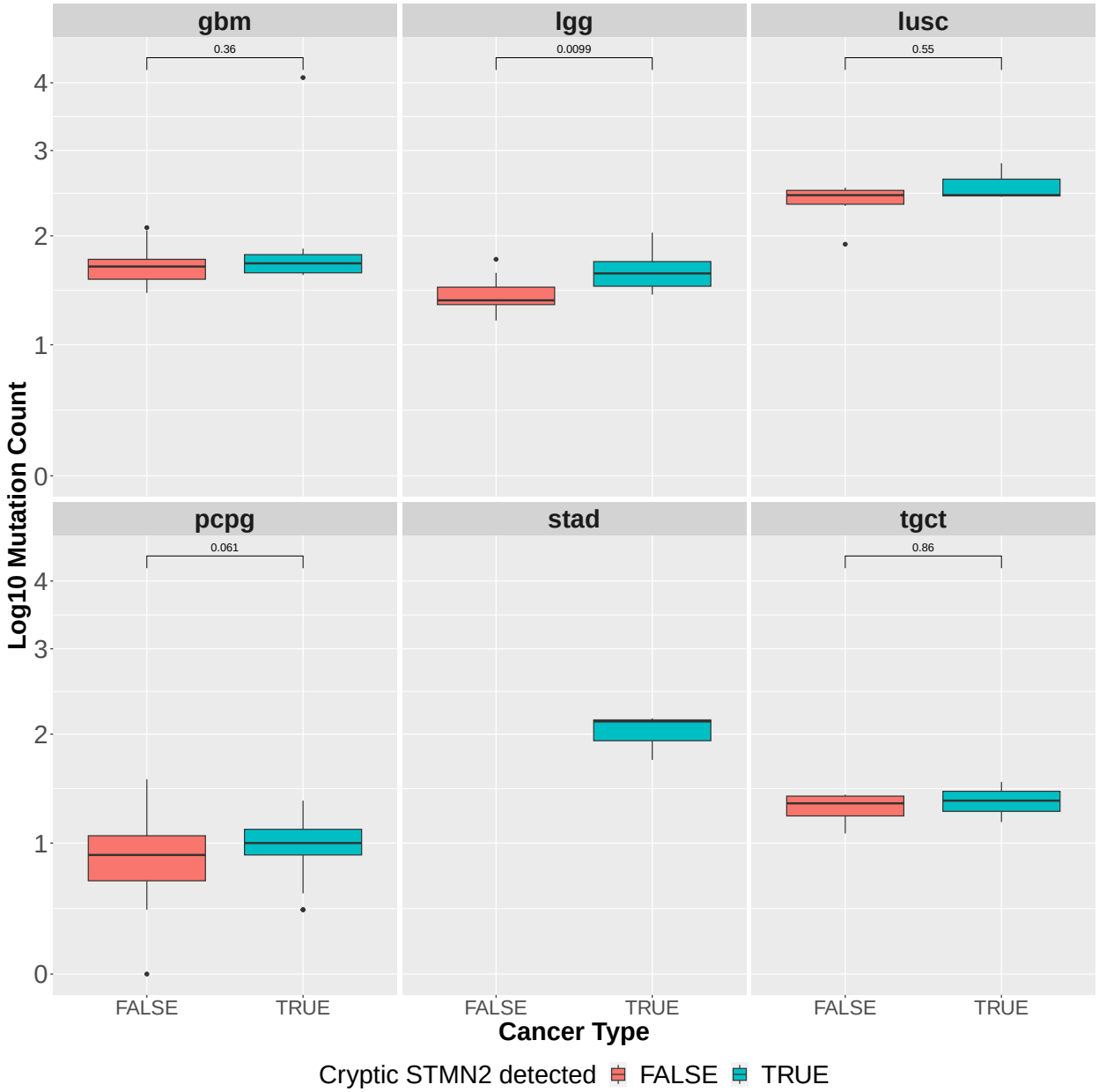

Figure 7: Mutational burden in cancer types with cryptic STMN2 expression.

LGG cancer patients with cryptic STMN2 expression, on average, have a significantly higher mutation count than non-cryptic LGG patients. There may also be a higher mutational burden in cryptic PCPG cases compared to non-cryptic PCPG cases, however the statistical significance of this is not apparent. There is no significant difference in mutation count between cryptic and non-cryptic cases in GBM, LUSC or TGCT cancers. The dataset only contained STAD cancer patients with cryptic STMN2 expression so no comparison could be made against the mutation count of non-cryptic cases.

Survival Comparisons

Survival analysis was conducted to assess any potential differences in survival in the cryptic STMN2 cases and non-cryptic cases. This was done using the survival data that had previously been pulled back in the clinical data from cBioPortal. This included disease-specific survival (months) after diagnosis and disease-specific survival status (i.e., alive or dead).

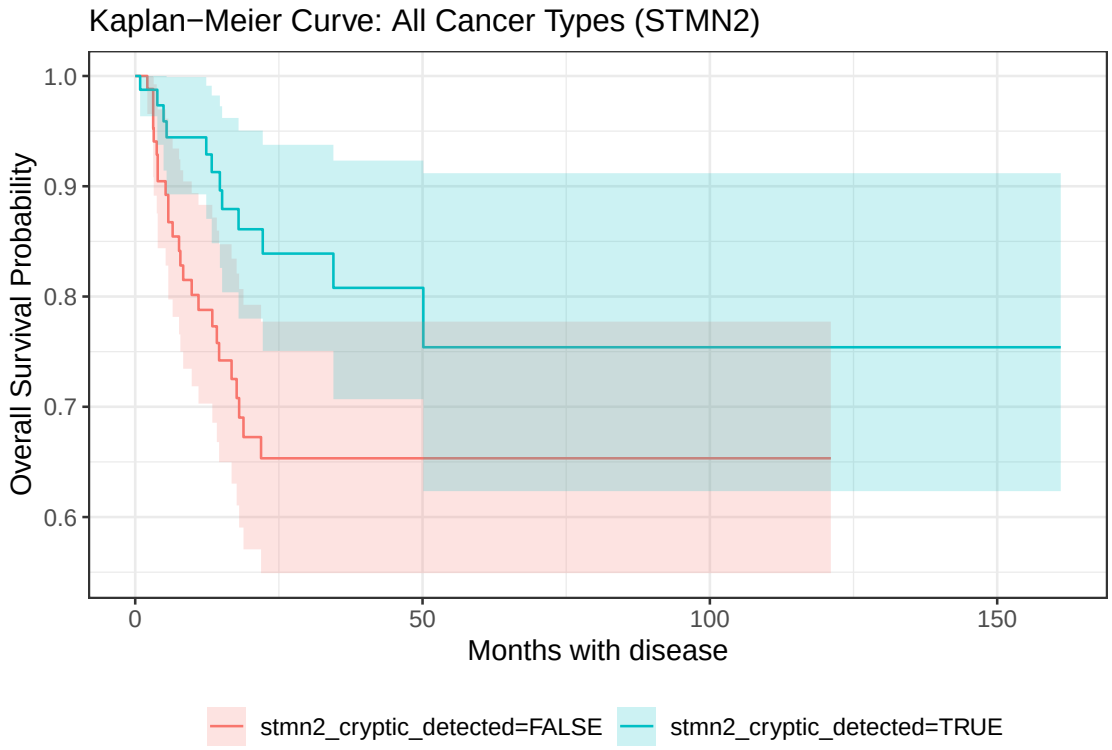

Figure 8: Kaplan-Meier survival curves for cancers with cryptic STMN2 events and those without cryptic STMN2. The probabilities shown are Kaplan-Meier survival probabilities.

There is no significant difference between the overall survival of cancer patients with and without cryptic STMN2 expression as the confidence intervals of both curves largely overlap (Figure 8).

Comparing aneuploidy cancer-by-cancer

Aneuploidy - an abnormal number of chromosomes - is associated with various genetic and developmental disorders. It is important to compare the aneuploidy score between patients with and without cryptic STMN2 events in order to explore a potential correlation between cryptic expression and abnormal chromosome number. This can also help unravel potential underlying mechanisms that the cryptic reads are involved in.

Aneuploidy score was previously pulled back in the clinical data from cBioPortal. It was plotted cancer-by-cancer to compare the scores in cryptic cases against non-cryptic cases (Figure 9).

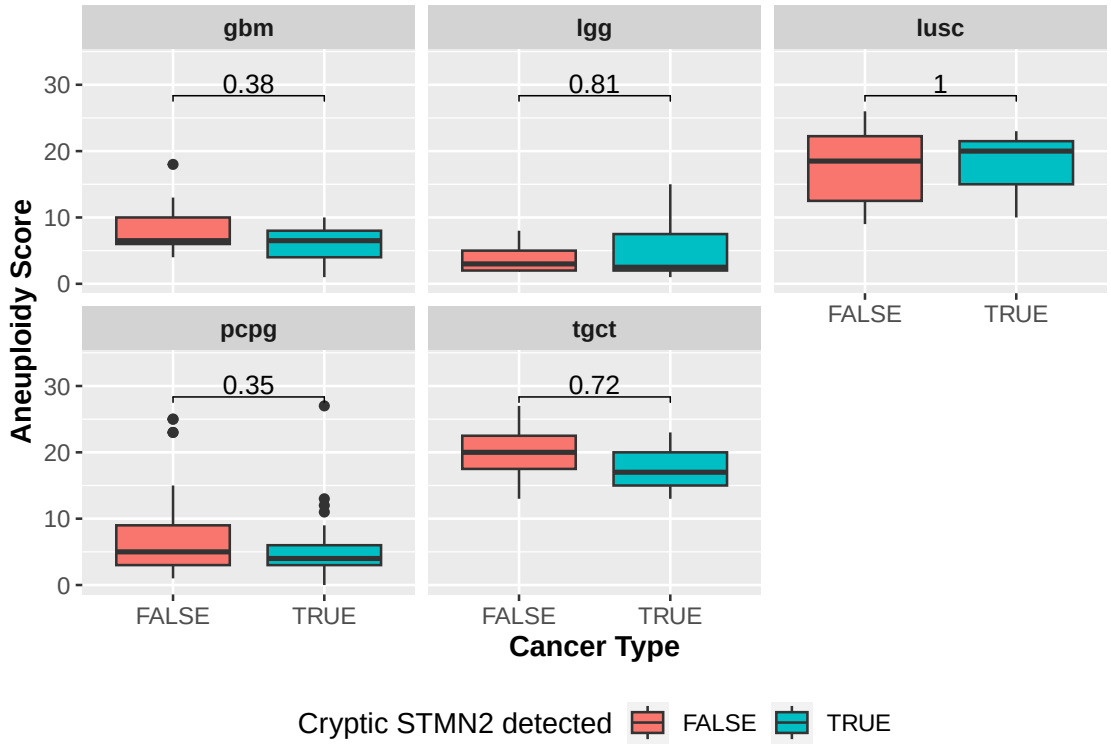

Figure 9: No significant difference in aneuploidy score in cryptic STMN2 versus non-cryptic cases.

Aneuploidy score is neither significantly higher nor lower in cases with cryptic STMN2 expression, indicating that the cryptic expression does not influence chromosome number.