SYNJ2 Cryptic Analysis

Introduction

TDP43-dependent cryptic SYNJ2 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 SYNJ2 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 SYNJ2 cryptic and annotated reads. A function was created to query junctions in the TCGA database for reads with the specified coordinates:

- Gene name = SYNJ2
- Snapcount coordinates (cryptic) = chr6:158017291-158019983
- Snapcount coordinates (annotated) = chr6:158017291-158028755
- 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 SYNJ2 cryptic and annotated query tables were joined into a single dataframe (**SYNJ2_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 SYNJ2 events and their clinical data downloaded and joined with the existing dataframe to make the **SYNJ2_clinical_jir** dataframe.

In order to investigate cryptic SYNJ2 expression in these patients, a new dataframe (**SYNJ2_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 SYNJ2 expression (in general)

To visualise the SYNJ2 reads data, a bar chart was plotted to compare the expression of SYNJ2 across the different cancer types. This was plotted using all data (**SYNJ2_clinical_jir**) rather than the cryptic filtered dataframe, and it compares the absolute read counts.

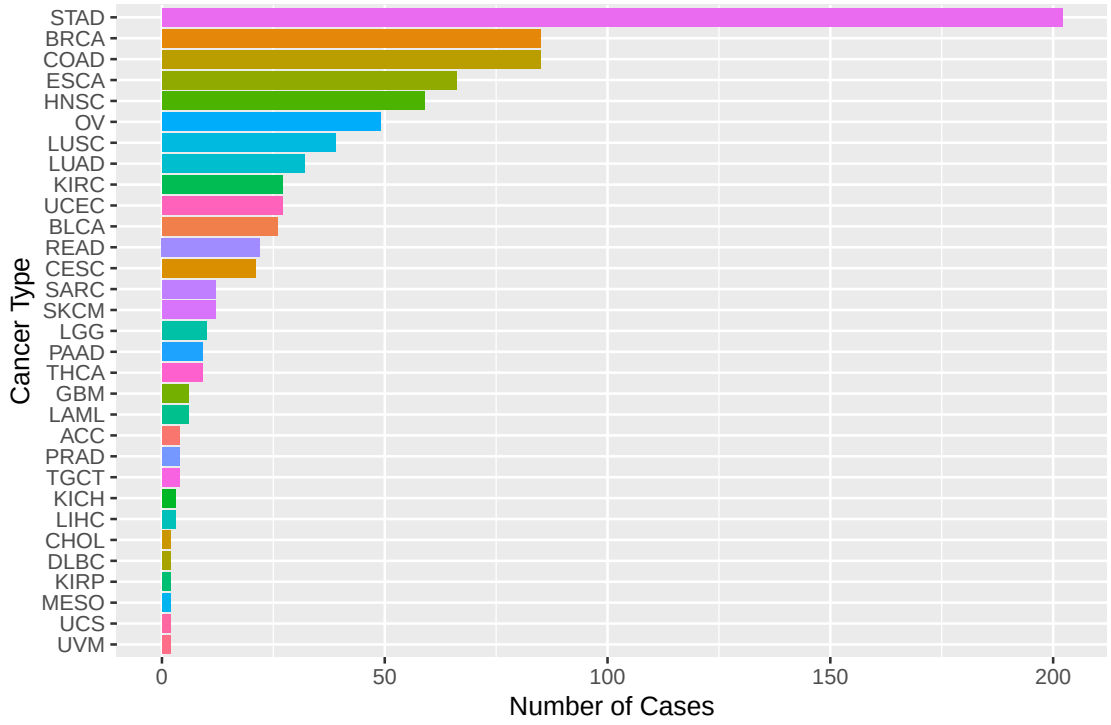

Figure 1: SYNJ2 is expressed mostly in stomach cancer patients.

Cancers with cryptic SYNJ2 events

The same bar chart was then plotted using the cryptic filtered dataframe (**SYNJ2_clinical_jir_cryptic**) to compare the expression of only the cryptic SYNJ2 events across the different cancer types. Again, this plot compares the absolute read counts of the cryptic events.

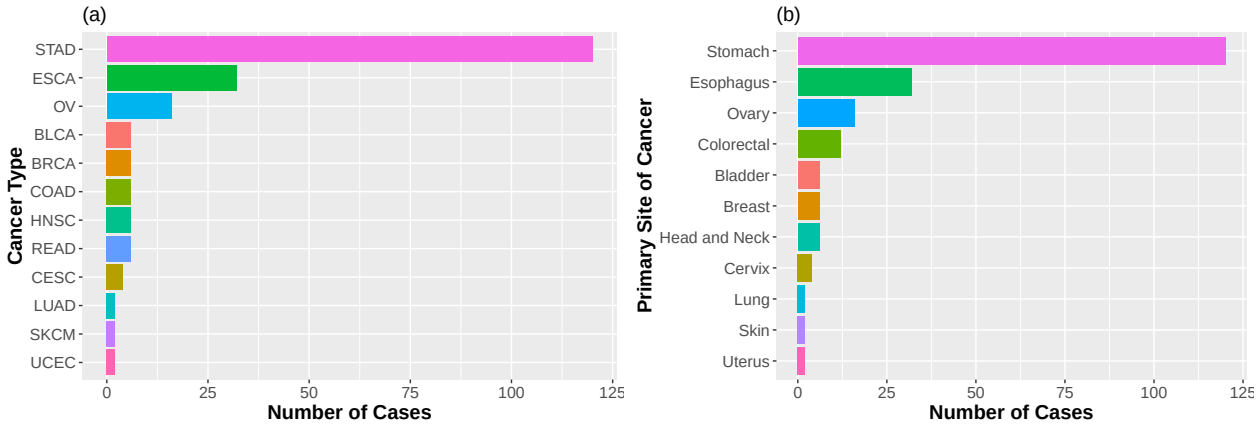

Figure 2: (a) Cryptic SYNJ2 events are found mostly in stomach cancers. STAD = stomach adenocarcinoma. (b) Cryptic SYNJ2 events are found most abundantly in cancers of the stomach.

Junction coverage

Overall SYNJ2 junction coverage was visualised in the different primary sites of cancers to see if the high cryptic read counts in breast cancers were due to higher overall SYNJ2 coverage in those cancers (Figure 3a). Cryptic SYNJ2 junction coverage was also investigated by calculating 'reads per million' as the fraction of cryptic reads of the overall junction coverage (Figure 3b).

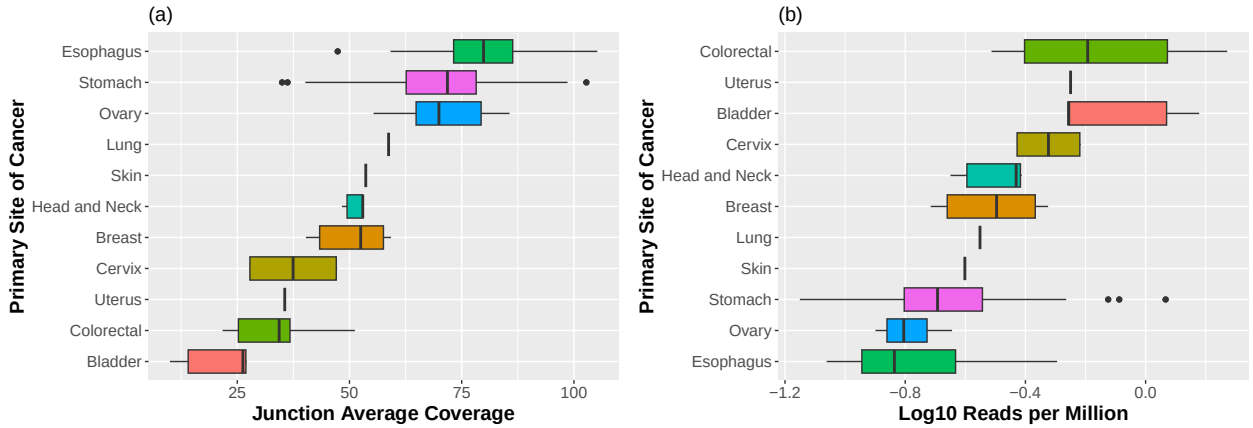

Figure 3: (a) Esophageal, stomach and ovarian cancers are the most deeply sequenced. (b) Colorectal cancers have the greatest cryptic coverage.

Indeed, the stomach is deeply sequenced (Figure 3a) so it could be that the higher overall coverage resulted in a greater cryptic count for stomach cancers. However, the cryptic coverage of stomach cancers is quite low (Figure 3b), suggesting that the cases with high cryptic SYNJ2 expression are significant.

Which cancers have the most cryptic SYNJ2 events?

To investigate where we are seeing most of the cryptic SYNJ2 events, the number of cases of each cancer type (with cryptic SYNJ2) was weighted against the total number of cases with cryptic SYNJ2.

Table 1: STAD cancer has high cryptic SYNJ2 expression

cancer_abbrev	n	percent
STAD	120	0.5769231
ESCA	32	0.1538462
OV	16	0.0769231
BLCA	6	0.0288462
BRCA	6	0.0288462
COAD	6	0.0288462
HNSC	6	0.0288462
READ	6	0.0288462
CESC	4	0.0192308
LUAD	2	0.0096154
SKCM	2	0.0096154
UCEC	2	0.0096154

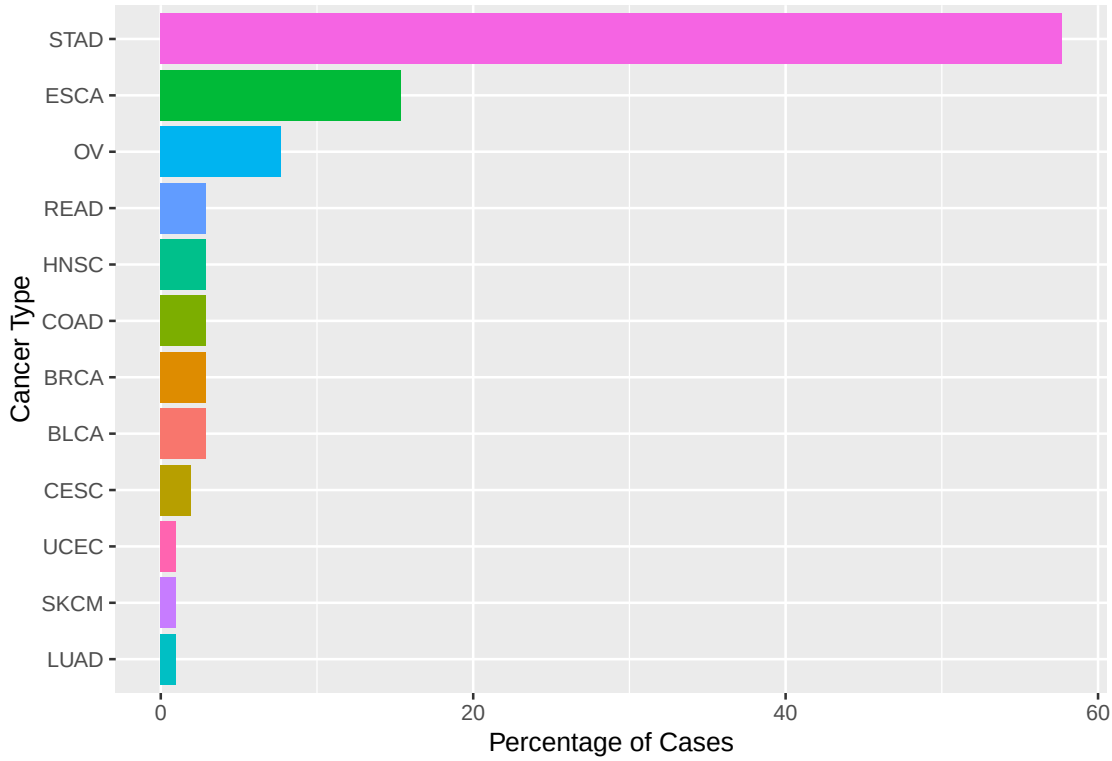

Figure 4: High proportion of cryptic SYNJ2 events are in STAD cancers.

The majority of cryptic SYNJ2 events are expressed in STAD patients (57.5% of cryptic cases), consistent with previous results. It is important to note that the high coverage is influenced by patients exhibiting extreme values of cryptic coverage (Figure 3b).

Where are the cancers with cryptic ARHGAP32 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 SYNJ2) was weighted against the total number of cases with cryptic SYNJ2.

Table 2: Cancers with cryptic SYNJ2 events are found primarily in the stomach.

gdc_cases_project_primary_site	n	percent
Stomach	120	0.5769231
Esophagus	32	0.1538462
Ovary	16	0.0769231
Colorectal	12	0.0576923
Bladder	6	0.0288462
Breast	6	0.0288462
Head and Neck	6	0.0288462
Cervix	4	0.0192308
Lung	2	0.0096154
Skin	2	0.0096154
Uterus	2	0.0096154

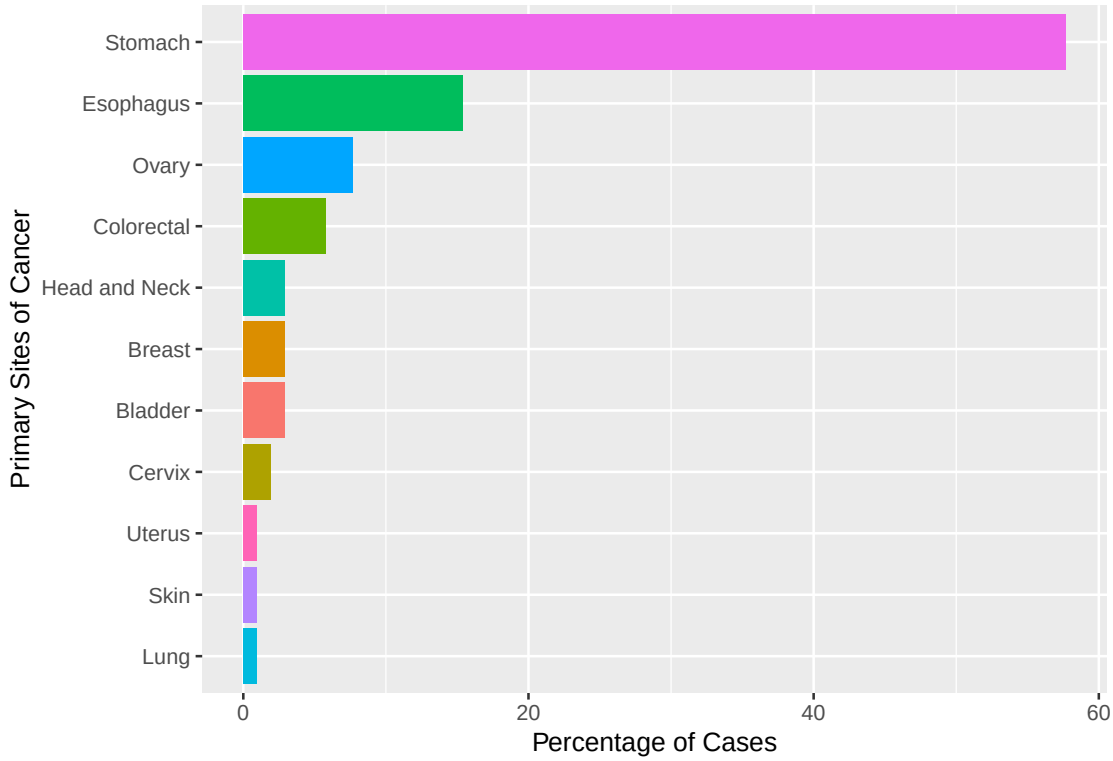

Figure 5: Cancers with cryptic SYNJ2 events are found primarily in the stomach.

Consistent with all previous results so far, most of the cryptic SYNJ2 events seen in cancer patients are in cancers of the stomach (57.5% 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 SYNJ2 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 SYNJ2 expression (**SYNJ2_clinical_jir_cryptic**). This join added the clinical data to only the relevant patients of interest (i.e., those with cryptic SYNJ2) to produce the **SYNJ2_cryptic_cBio** dataframe.

Fraction of each cancer that has cryptic SYNJ2 events

The burden of cryptic SYNJ2 expression was examined. This was done by looking at the fraction of each cancer type that exhibits TDP43-dependent cryptic SYNJ2 events (Figure 6).

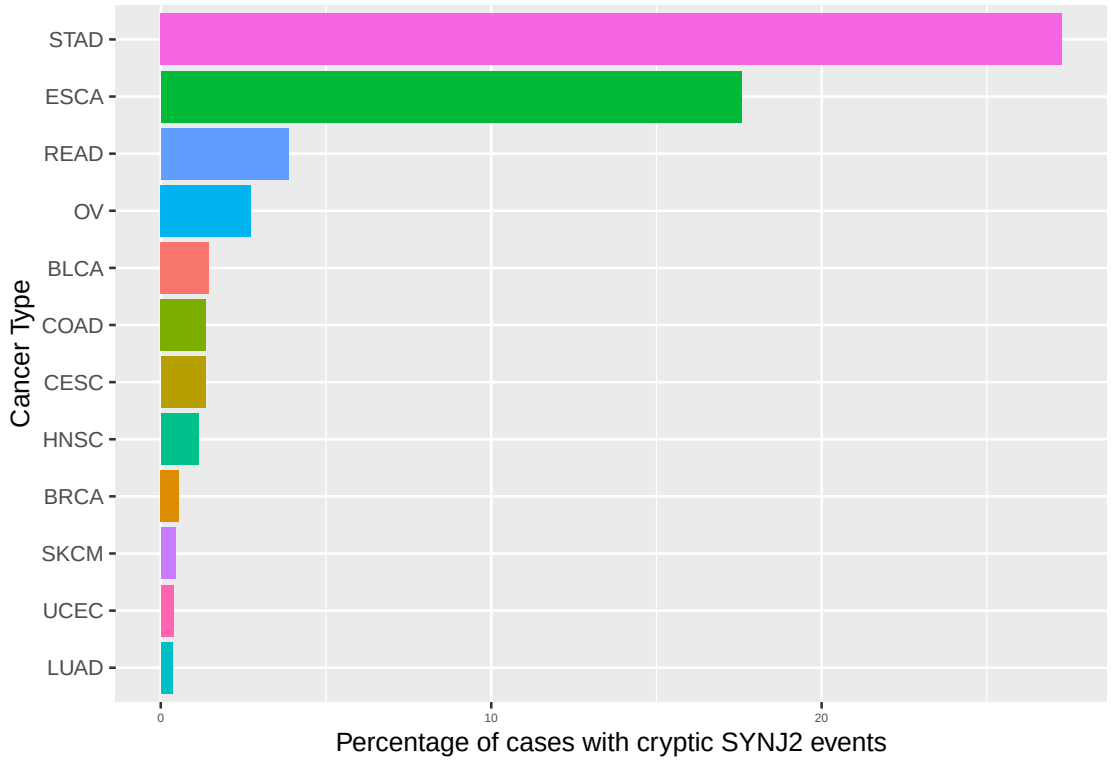

Figure 6: Fraction of each cancer type that has cryptic SYNJ2 expression.

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 SYNJ2, the total mutation count seen in patients with cryptic SYNJ2 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 SYNJ2 cases.

cancer_abbrev	total_mutations	total_mutations_cryptic	percent_with_cryptic
STAD	147953	48240	32.6049489
ESCA	25631	5248	20.4752058
OV	36093	902	2.4990995
SKCM	325852	4342	1.3325068
HNSC	74994	806	1.0747526
BLCA	99709	666	0.6679437
BRCA	84230	498	0.5912383
COAD	165465	732	0.4423896
LUAD	157145	684	0.4352668
CESC	56054	234	0.4174546
READ	43274	168	0.3882239
UCEC	538960	98	0.0181832

Table 3 shows that 32.6% of the mutations in STAD cancer patients are seen in cases with cryptic SYNJ2. This is the largest proportion out of all cancer subsets. Additionally, 20.5% of the mutations in ESCA cancer patients are seen in cases with cryptic SYNJ2.

Mutational burden was further examined within each cancer type, using boxplots to visualise the mutation counts in non-cryptic cases and cases with cryptic SYNJ2 expression (Figure 7). Only cancers with cryptic

SYNJ2 expression are plotted. This analysis aimed to determine whether cases with cryptic ARHGAP32 events have a greater mutational burden.

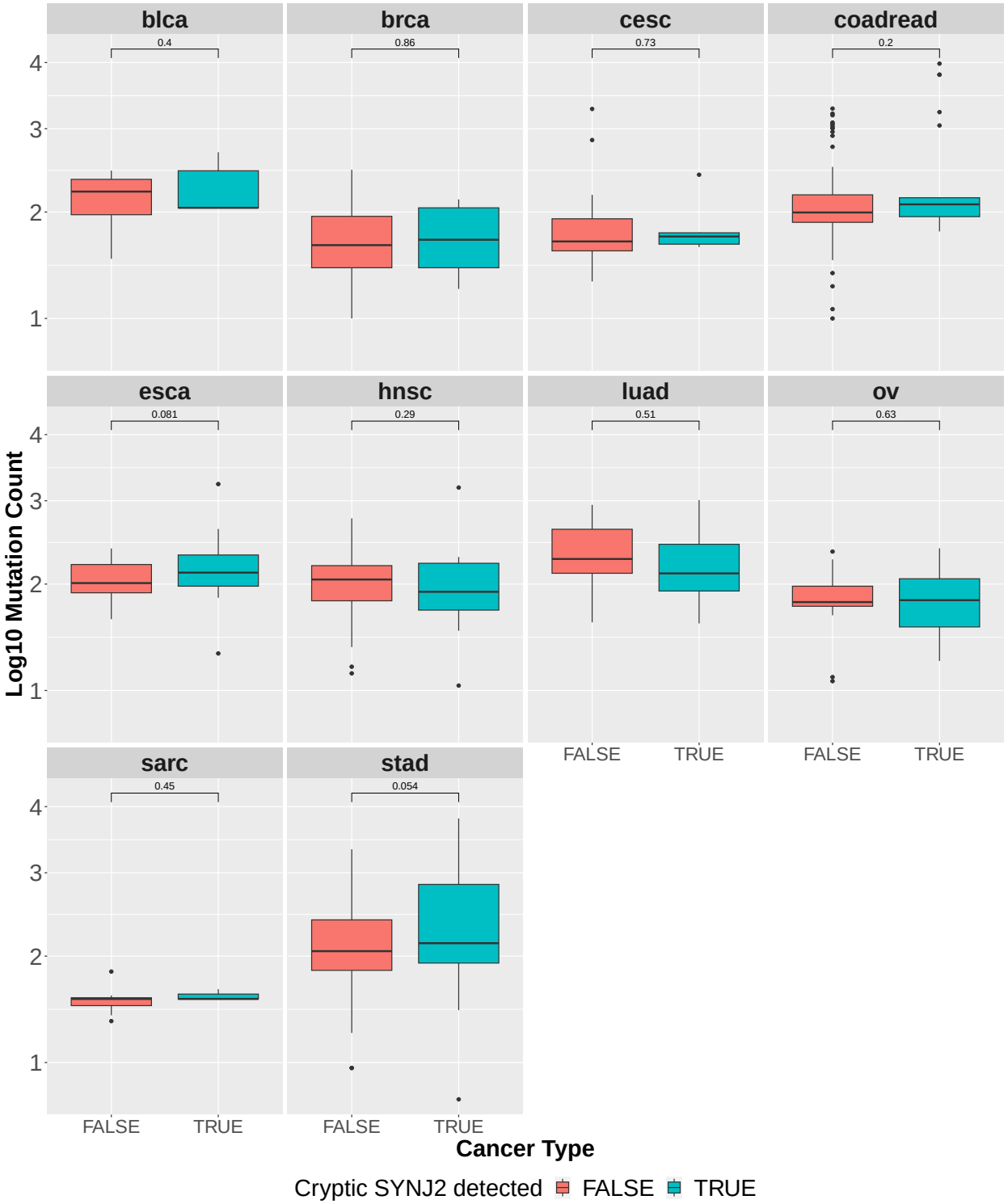

Figure 7: Mutational burden in cancer types with cryptic SYNJ2 expression.

There may be a higher mutational burden in cryptic STAD ($p = 0.054$) and cryptic ESCA ($p = 0.081$) cases compared to non-cryptic cases, however the statistical significance of this is not apparent.

Survival Comparisons

Survival analysis was conducted to assess any potential differences in survival in the cryptic SYNJ2 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).

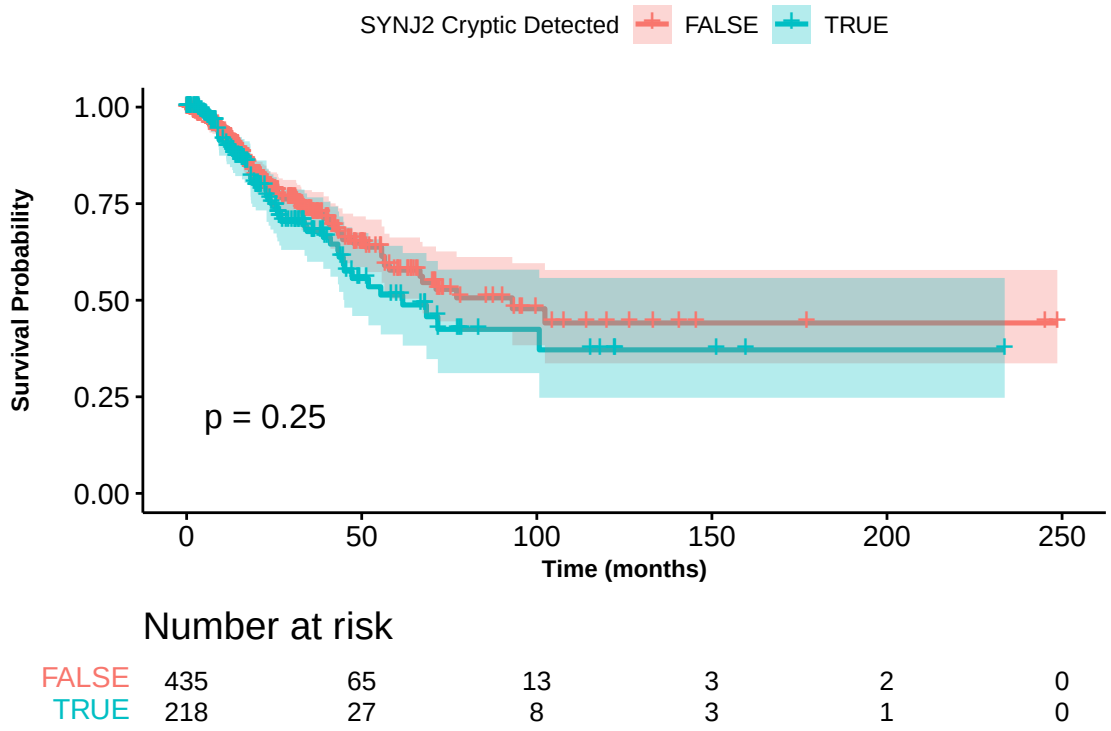

Figure 8: Kaplan-Meier survival curves for cancers with cryptic SYNJ2 events and those without cryptic SYNJ2. The probabilities shown are Kaplan-Meier survival probabilities.

There is no significant difference between the overall survival of cancer patients with and without cryptic SYNJ2 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 SYNJ2 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).

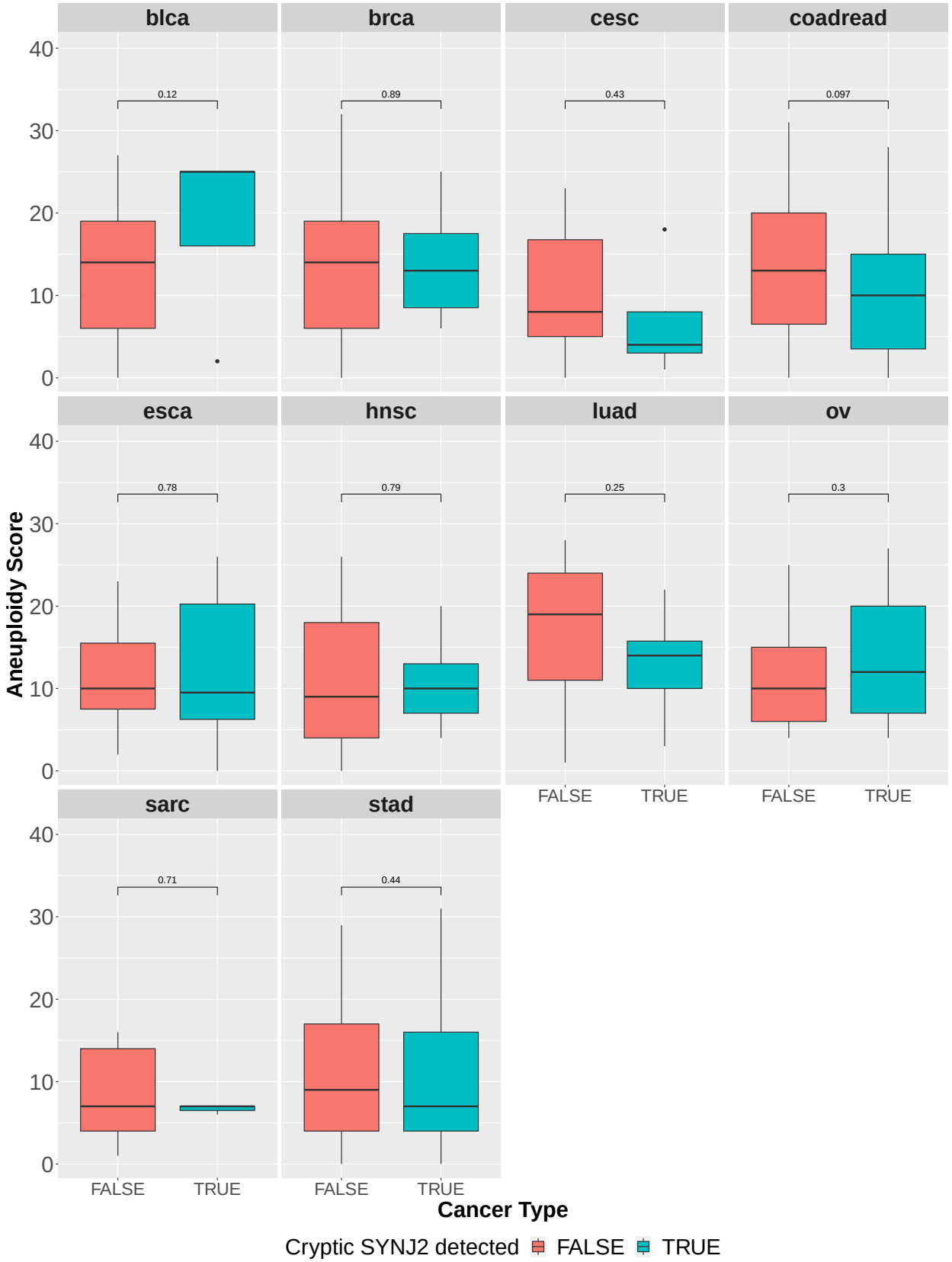

Figure 9: Aneuploidy score in cryptic SYNJ2 versus non-cryptic cases.

There is no significant difference in aneuploidy score in cases with cryptic SYNJ2 expression and non- cryptic cases for each of the cancer types, indicating that the cryptic expression does not influence chromosome number.