ARHGAP32 Cryptic Analysis

Introduction

TDP43-dependent cryptic ARHGAP32 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 ARHGAP32 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 = ARHGAP32
- Snapcount coordinates (cryptic) = chr11:128992047-128998318
- Snapcount coordinates (annotated) = chr11:128988126-128998318
- 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 ARHGAP32 cryptic and annotated query tables were joined into a single dataframe (**ARHGAP32_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 **ARHGAP32_clinical_jir** dataframe.

In order to investigate cryptic STMN2 expression in these patients, a new dataframe (**ARHGAP32_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 ARHGAP32 expression (in general)

To visualise the ARHGAP32 reads data, a bar chart was plotted to compare the expression of ARHGAP32 across the different cancer types. This was plotted using all data (**ARHGAP32_clinical_jir**) rather than the cryptic filtered dataframe, and it compares the absolute read counts.

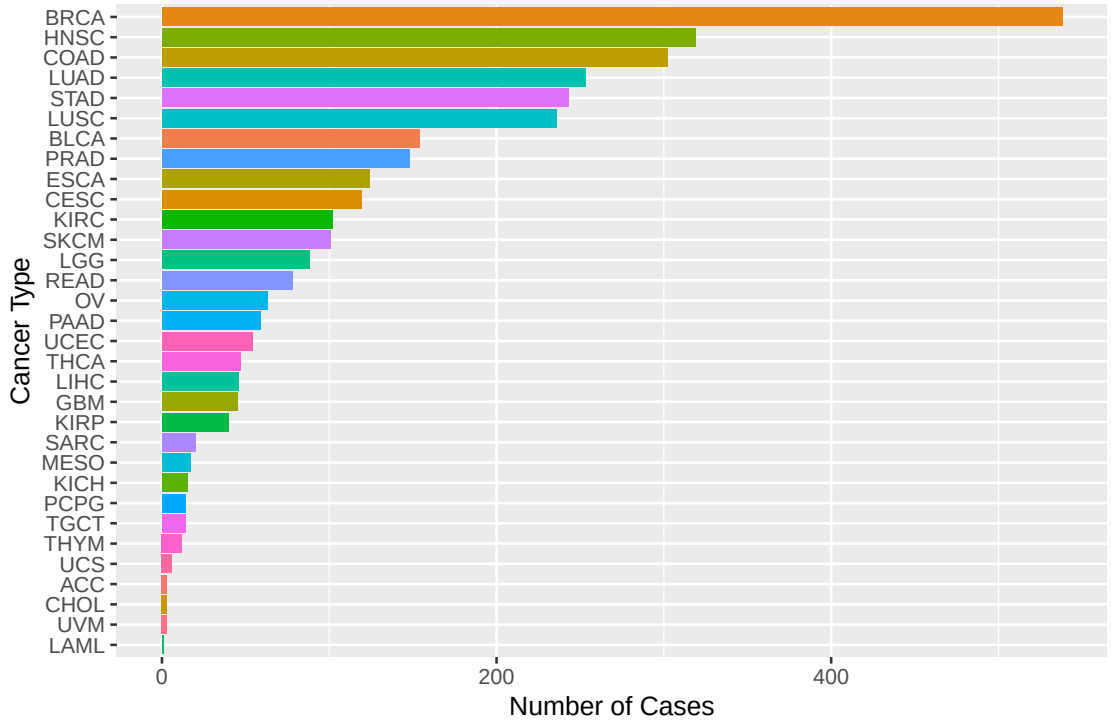

Figure 1: ARHGAP32 is expressed in mostly breast cancer patients. BRCA = breast invasive carcinoma.

Cancers with cryptic ARHGAP32 events

The same bar chart was then plotted using the cryptic filtered dataframe (**ARHGAP32_clinical_jir_cryptic**) to compare the expression of only the cryptic ARHGAP32 events across the different cancer types. Again, this plot compares the absolute read counts of the cryptic events.

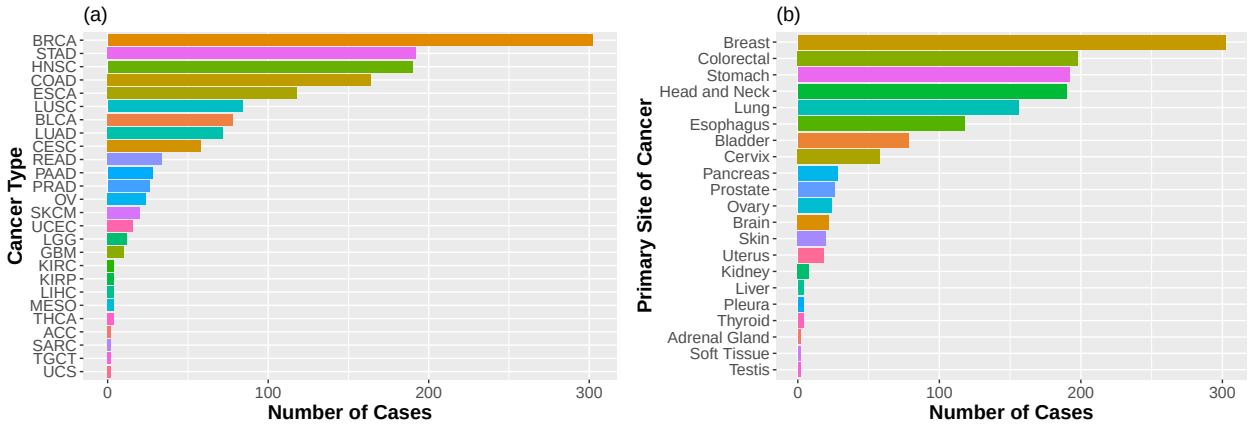

Figure 2: (a) Cryptic ARHGAP32 events are found mostly in breast cancers. BRCA = breast invasive carcinoma. (b) Cryptic ARHGAP32 events are found most abundantly in cancers of the breast.

Junction coverage

Overall ARHGAP32 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 ARHGAP32 coverage in those cancers (Figure 3a). Cryptic ARHGAP32 junction coverage was also investigated by calculating 'reads per million' as the fraction of cryptic reads of the overall junction coverage (Figure 3b).

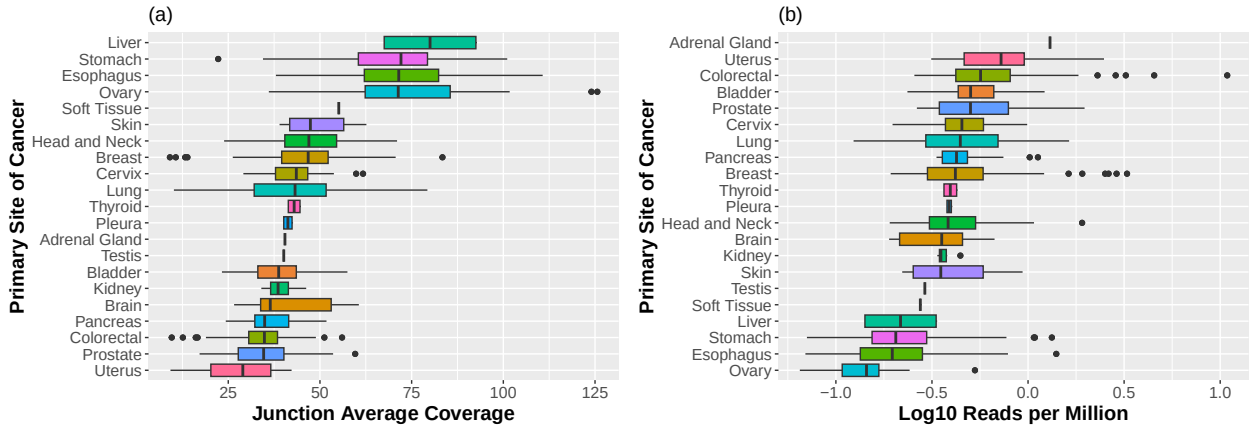

Figure 3: (a) Liver, stomach, esophageal and ovarian cancers are the most deeply sequenced. (b) Cancers with the greatest cryptic coverage.

The breast is not very deeply sequenced so the high cryptic count is likely significant in these cancers. There is only one patient with cancer of the adrenal gland who has TDP43-dependent cryptic ARHGAP32 events, hence the single data value for adrenal gland. There is an extreme data value for colorectal cancer, which indicates that it is one patient with a larger number of ARHGAP32 cryptic reads that is influencing the overall cryptic coverage for this cancer subset.

Which cancers have the most cryptic ARHGAP32 events?

To investigate where we are seeing most of the cryptic ARHGAP32 events, the number of cases of each cancer type (with cryptic ARHGAP32) was weighted against the total number of cases with cryptic ARHGAP32.

Table 1: Breast cancer has high cryptic ARHGAP32 expression.

cancer_abbrev	n	percent
BRCA	302	0.2071331
STAD	194	0.1330590
HNSC	190	0.1303155
COAD	164	0.1124829
ESCA	118	0.0809328
LUSC	84	0.0576132
BLCA	78	0.0534979
LUAD	72	0.0493827
CESC	58	0.0397805
READ	34	0.0233196
PAAD	28	0.0192044
PRAD	26	0.0178326
OV	24	0.0164609
SKCM	20	0.0137174
UCEC	16	0.0109739

cancer_abbrev	n	percent
LGG	12	0.0082305
GBM	10	0.0068587
KIRC	4	0.0027435
KIRP	4	0.0027435
LIHC	4	0.0027435
MESO	4	0.0027435
THCA	4	0.0027435
ACC	2	0.0013717
SARC	2	0.0013717
TGCT	2	0.0013717
UCS	2	0.0013717

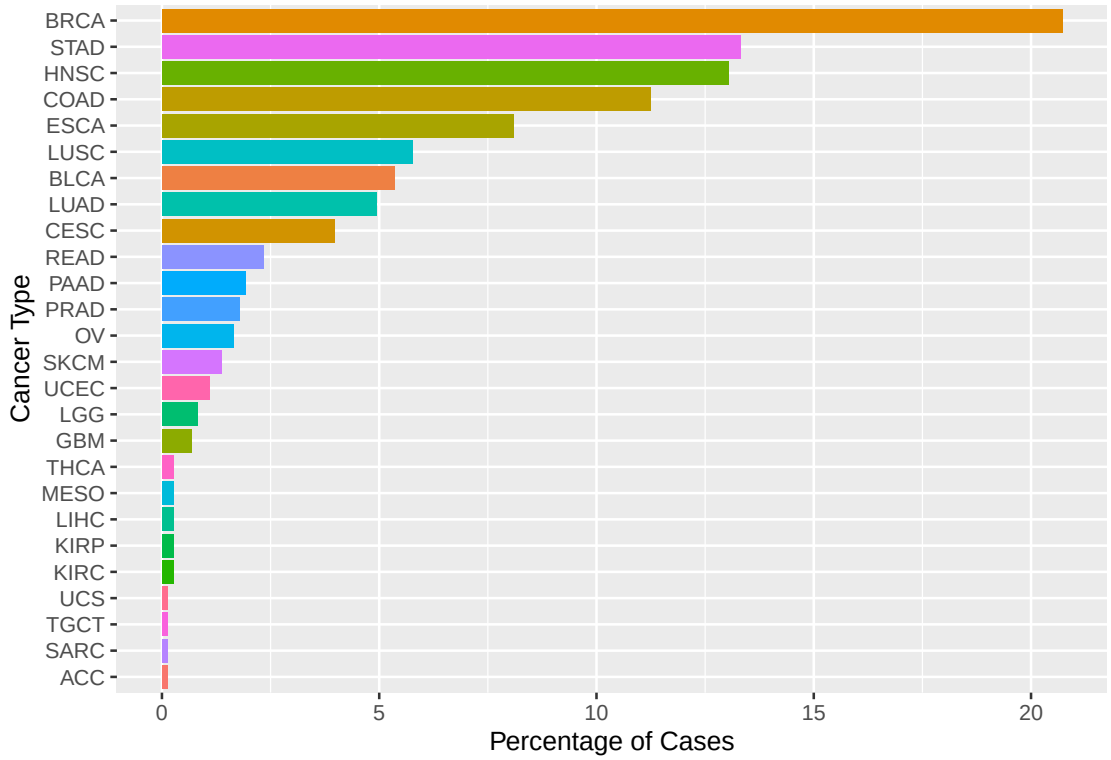

Figure 4: High proportion of cryptic ARHGAP32 events are in BRCA cancer.

The majority of the cryptic ARHGAP32 events are expressed in BRCA (20.7% of cryptic cases), consistent with previous results.

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 ARHGAP32) was weighted against the total number of cases with cryptic ARHGAP32.

gdc_cases_project_primary_site	n	percent
Breast	302	0.2074176

gdc_cases_project_primary_site	n	percent
Colorectal	198	0.1359890
Stomach	192	0.1318681
Head and Neck	190	0.1304945
Lung	156	0.1071429
Esophagus	118	0.0810440
Bladder	78	0.0535714
Cervix	58	0.0398352
Pancreas	28	0.0192308
Prostate	26	0.0178571
Ovary	24	0.0164835
Brain	22	0.0151099
Skin	20	0.0137363
Uterus	18	0.0123626
Kidney	8	0.0054945
Liver	4	0.0027473
Pleura	4	0.0027473
Thyroid	4	0.0027473
Adrenal Gland	2	0.0013736
Soft Tissue	2	0.0013736
Testis	2	0.0013736

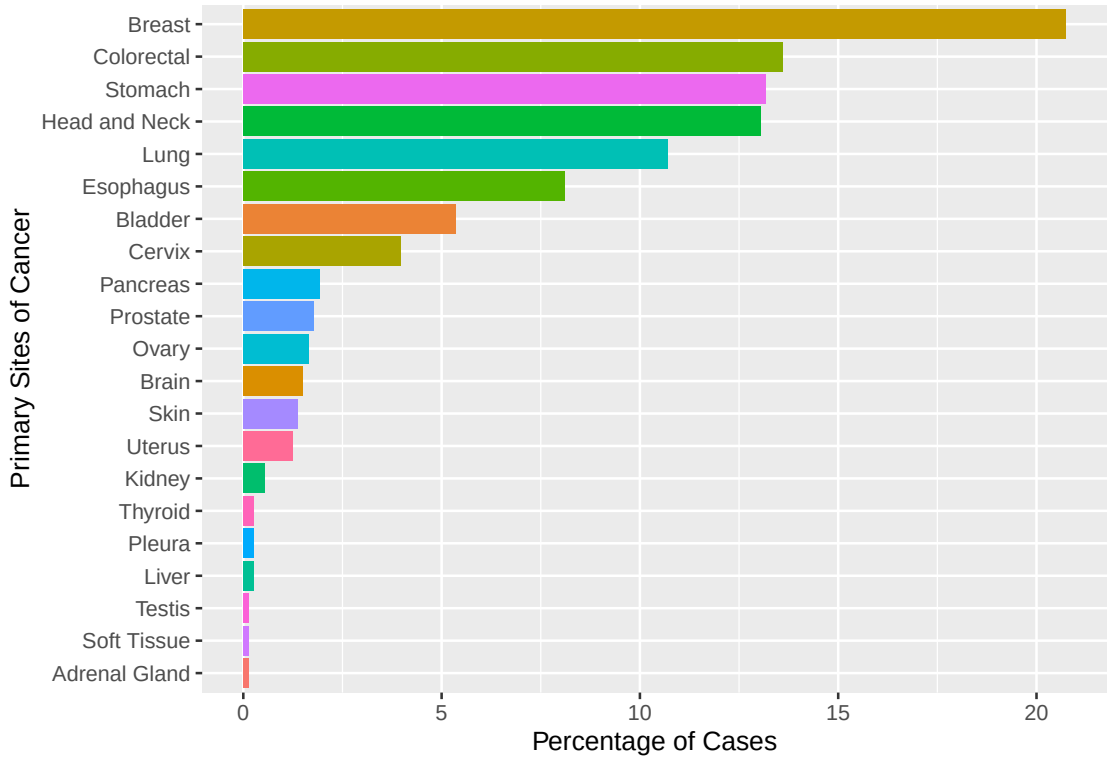

Figure 5: Cancers with cryptic ARHGAP32 events are primarily in the breast.

Consistent with all previous results so far, most of the cryptic ARHGAP32 events seen in cancer patients are in cancers of the breast (20.7% 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 ARHGAP32 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 ARHGAP32 expression (**ARHGAP32_clinical_jir_cryptic**). This join added the clinical data to only the relevant patients of interest (i.e., those with cryptic ARHGAP32) to produce the **ARHGAP32_cryptic_cBio** dataframe.

Fraction of each cancer that has cryptic ARHGAP32 events

The burden of cryptic ARHGAP32 expression was examined. This was done by looking at the fraction of each cancer type that exhibits TDP43-dependent cryptic ARHGAP32 events (Figure 6).

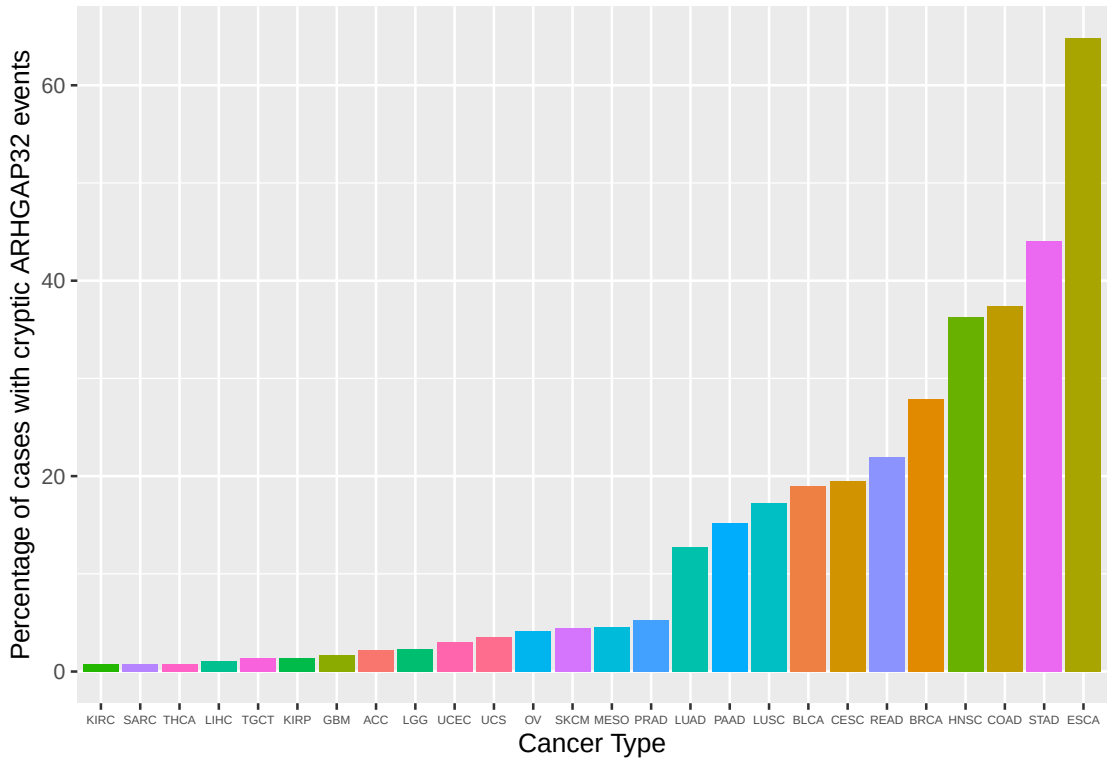

Figure 6: Fraction of each cancer type that has cryptic ARHGAP32 expression.

ESCA (esophageal carcinoma) cancer has the highest proportion of patients with cryptic ARHGAP32 events.

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 ARHGAP32, the total mutation count seen in patients with cryptic ARHGAP32 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 ARHGAP32 cases.

cancer_abbrev	total_mutations	total_mutations_cryptic	percent_with_cryptic
ESCA	25631	21044	82.1037025
STAD	147953	65978	44.5938913
HNSC	74994	28458	37.9470358
COAD	165465	62322	37.6647629
BRCA	84230	22652	26.8930310
BLCA	99709	21216	21.2779187
LUSC	126568	21526	17.0074584
LUAD	157145	21896	13.9336282
CESC	56054	7212	12.8661648
PAAD	20703	1282	6.1923393
OV	36093	1984	5.4969108
READ	43274	2324	5.3704303
SKCM	325852	16256	4.9887679
MESO	2488	100	4.0192926
TGCT	2157	58	2.6889198
PRAD	21448	576	2.6855651
ACC	7570	130	1.7173052
KIRP	20308	272	1.3393736
UCS	7303	96	1.3145283
LGG	27152	334	1.2301120
LIHC	36216	394	1.0879169
GBM	46100	396	0.8590022
KIRC	20195	122	0.6041099
SARC	17394	84	0.4829251
THCA	7622	30	0.3935975
UCEC	538960	1580	0.2931572

Table 3 shows that 82.1% of the mutations in ESCA (esophageal carcinoma) are seen in cases with cryptic ARHGAP32. This is the largest proportion out of all cancer subsets. 26.9% of all mutation in BRCA patients are in cases with cryptic ARHGAP32 events.

Mutational burden was further examined within each cancer type, using boxplots to visualise the mutation counts in non-cryptic cases and cases with cryptic ARHGAP32 expression (Figure 7). Only cancers with cryptic ARHGAP32 expression are plotted. This analysis aimed to determine whether cases with cryptic ARHGAP32 events have a greater mutational burden.

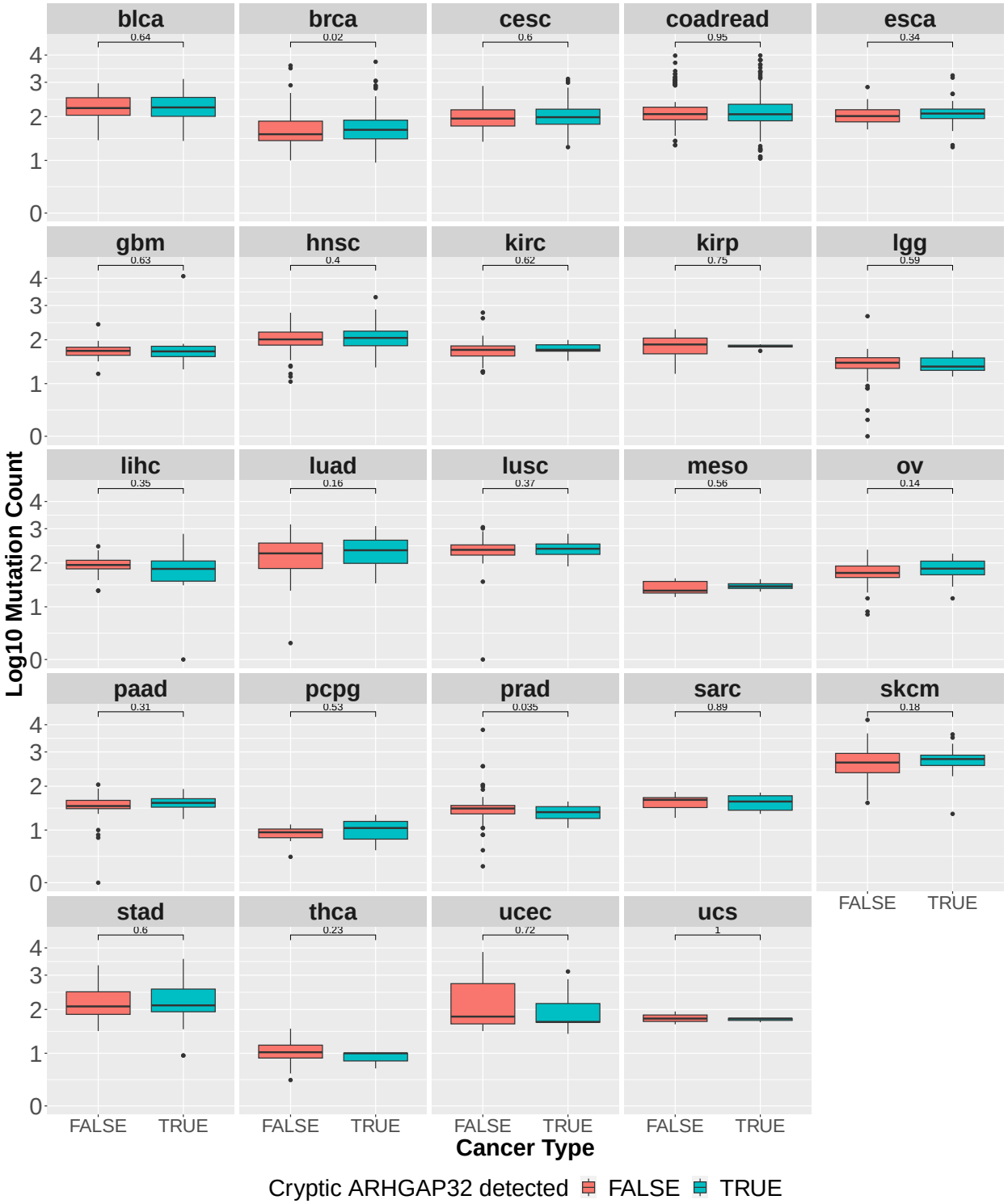

Figure 7: Mutational burden in cancer types with cryptic ARHGAP32 expression.

BRCA (breast invasive carcinoma, p = 0.02) and PRAD (prostate adenocarcinoma, p = 0.035) cancer patients with cryptic ARHGAP32 expression, on average, have a significantly higher mutation count that non-cryptic BRCA and PRAD patients. There may also be a higher mutational burden in cryptic LUAD and SKCM 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 ARHGAP32 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).

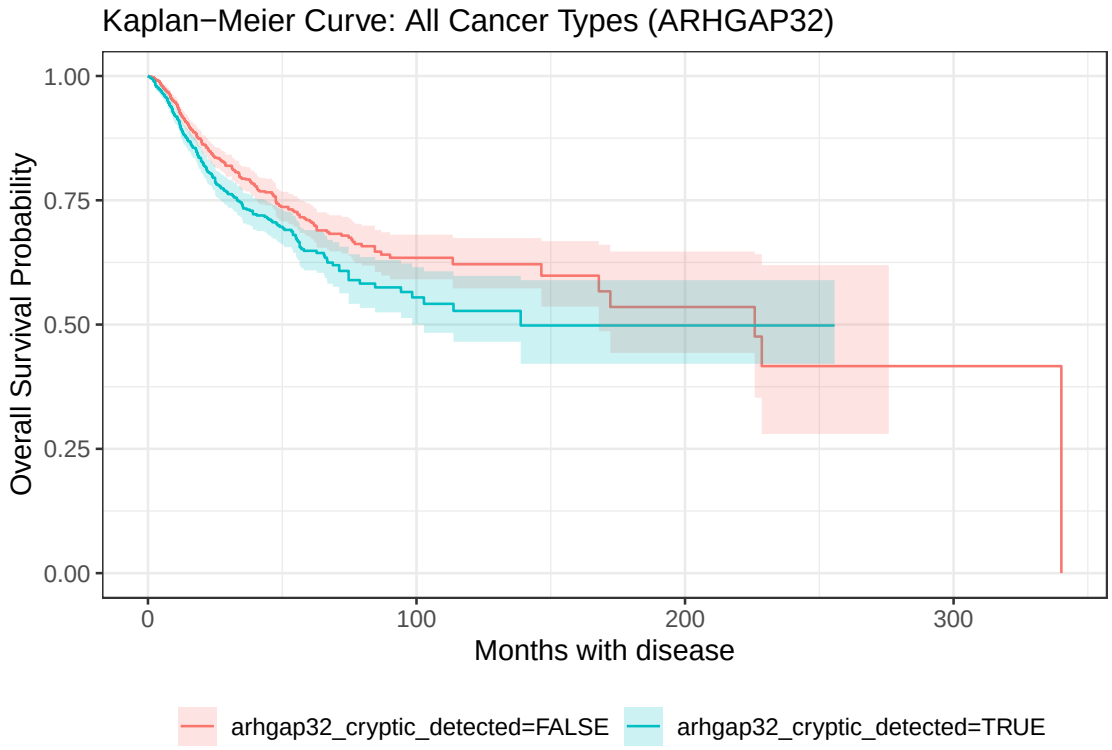

Figure 8: Kaplan-Meier survival curves for cancers with cryptic ARHGAP32 events and those without cryptic ARHGAP32 The probabilities shown are Kaplan-Meier survival probabilities.

There is no significant difference between the overall survival of cancer patients with and without cryptic ARHGAP32 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 ARHGAP32 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).

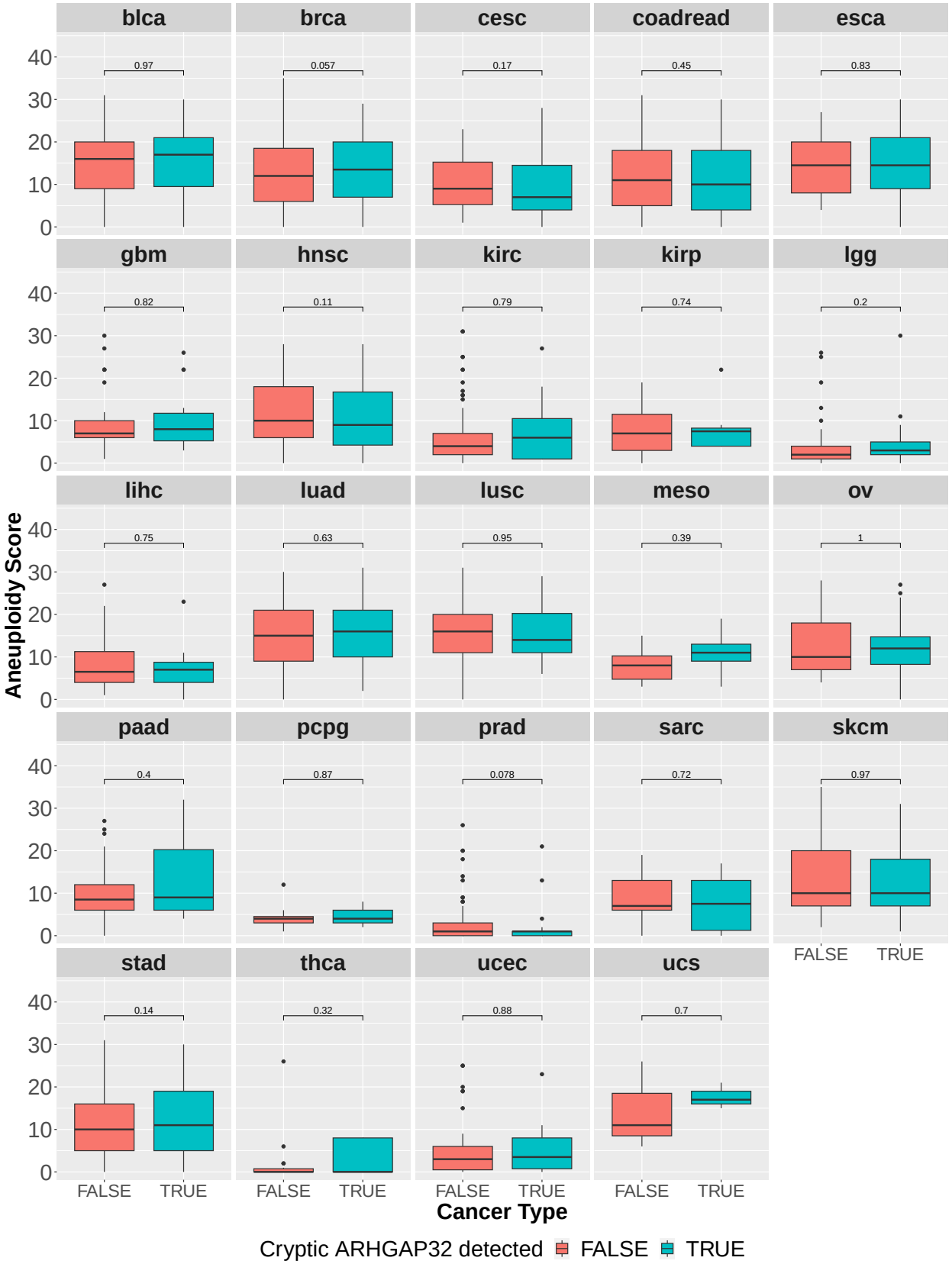

Figure 9: Aneuploidy score in cryptic ARHGAP32 versus non-cryptic cases.

There is no significant difference in aneuploidy score in cases with cryptic ARHGAP32 expression and non-cryptic cases for the majority of cancer types, indicating that the cryptic expression does not influence chromosome number. However, aneuploidy score is seemingly higher in cryptic BRCA ($p = 0.057$) but also in non-cryptic PRAD ($p = 0.078$) cases.