Antibody Responses to Different Proteins in Prostate Cancer Patients

Tun Lee Ng and Michael A. Newton

April 14, 2020

Contents

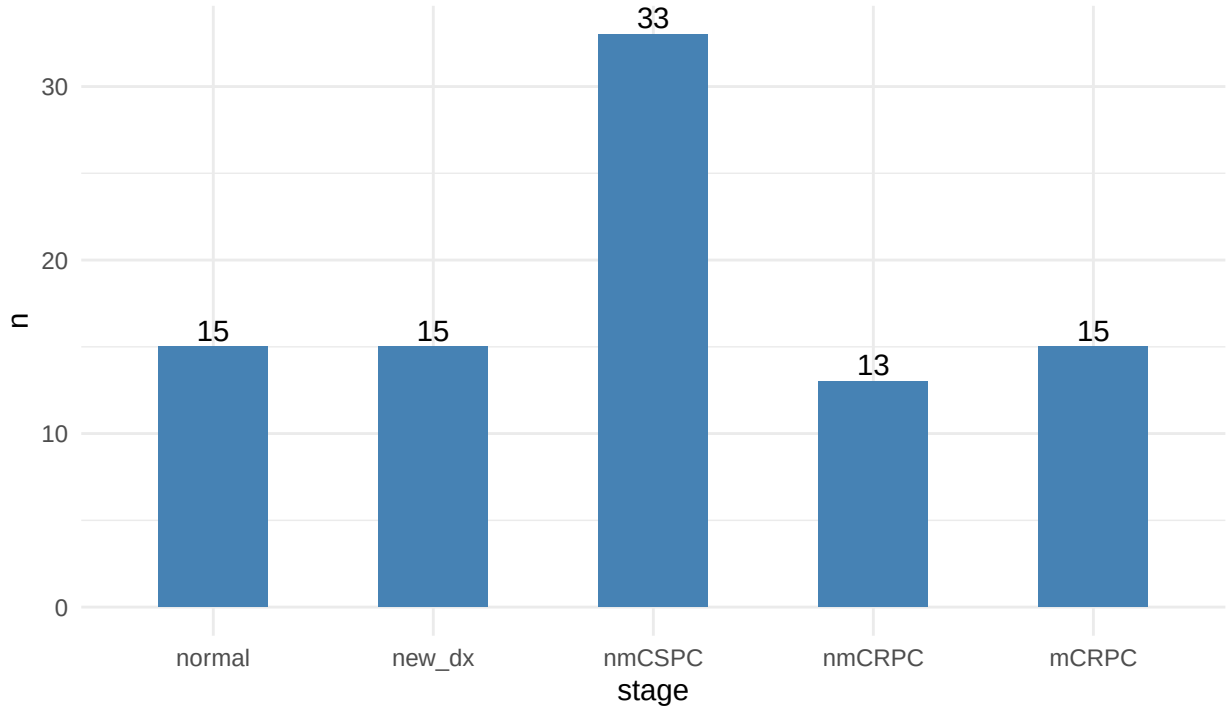
1 Introduction

This project aims to characterize antibody responses to a wide variety of proteins in prostate cancer patients at different stages of the disease. 16-mer peptides spanning the amino acid sequences of these 1611 gene products, and overlapping by 12 amino acids, were used to generate a microarray comprising 177,604 peptides. In this study, there were healthy subjects and patients with different stages of prostate cancer

- `new_dx`: newly diagnosed,
- `nmCSPC`: non-metastatic castration-sensitive,
- `mCSPC`: metastatic castration-sensitive,
- `nmCRPC`: non-metastatic castration-resistant,
- `mCRPC`: metastatic castration-resistant

stage	n
normal	17
new_dx	19
nmCSPC	52
mCSPC	16
nmCRPC	15
mCRPC	35

Note that these are not distinct patient counts, because there were 11 patients who were measured at two different stages. Number of replicates for each patient, `rep` could 1, 2, or 3. Hemanth removed patients with `rep` = 1. He also kept the latest distinct patient records only, so the distinct patient counts are:

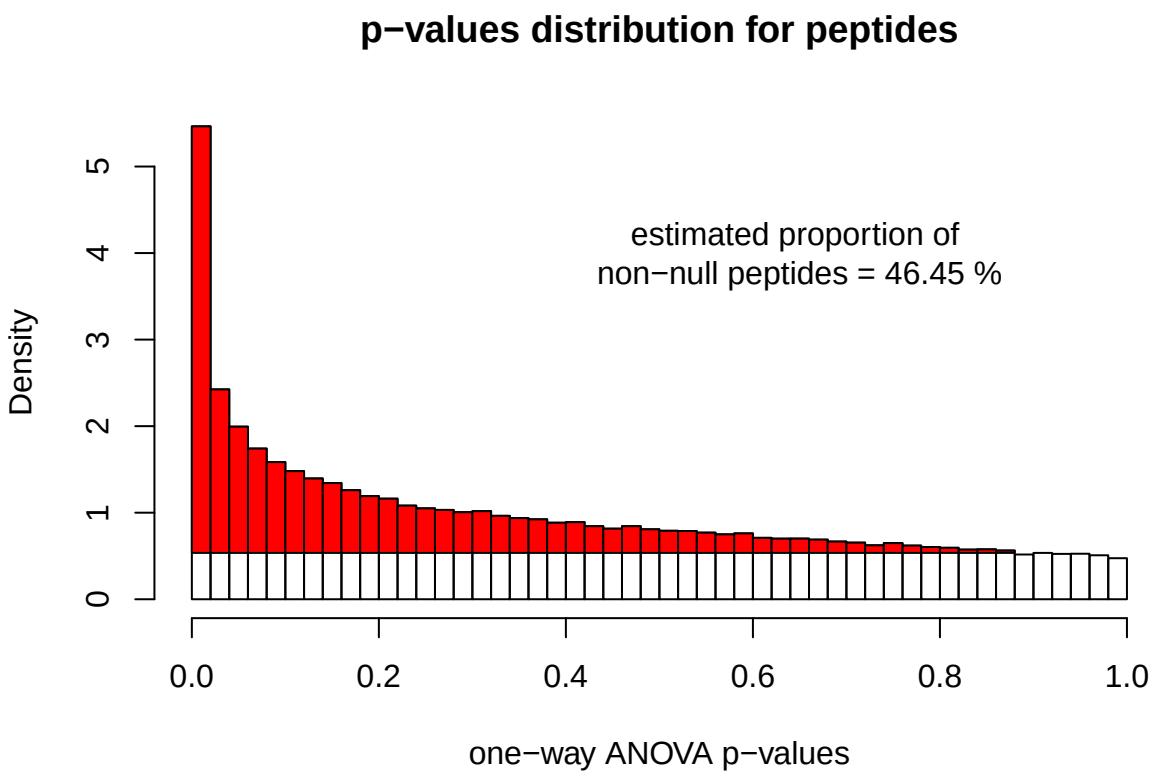


Next, we take $\log_2$ transformation of the fluorescence intensity and compute the median of the replicates of each patient.

2 One-Way ANOVA

We would like to investigate if patients at different stages of prostate cancer exhibit different antibody responses to certain peptide chains or proteins. Let μ_i be the average fluorescence level (on $\log_2$ scale) of patients, with subscript i indexing the different stages of prostate cancer as explained in Introduction section. We want to test whether
H_0: All μ_i's are the same, ie. Antibody responses are the same for patients at different stages of prostate cancer.
H_1: NOt all μ_i's are the same, ie. Antibody responses are not the same for patients at different stages of prostate cancer.

For each peptide, we perform one-way ANOVA (analysis of variance). After getting p-values for all 177k peptides, we plot the p-value histogram.



If cancer-stage effect is not present in our peptide array data, then the p-values from the ANOVA would have a uniform distribution between 0 and 1, and we expect to see a rather flat-shaped histogram of p-values.

However, the p-values histogram exhibits large counts of significant p-values (p-values close to zero), and the shape of histogram flattens off exponentially with larger p-values. Such a large count of significant p-values may not be explained by false discovery alone, and that perhaps cancer-stage effect is indeed present in some of the peptides in our profile. The red-shaded regions of the histogram represents the estimated proportion of non-null peptides in the dataset based on Storey's q-values calculation obtained via the R package `fdrtool`. The q-value is similar to the well known p-value, except that it is a measure of significance in terms of the false discovery rate rather than the false positive rate. The peptide counts at various q-values thresholds are tabulated below.

FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	181	2623	5235	7471	9510	11570	13603	15706	17900	19886

As a comparison, we also apply the Benjamini-Hochberg (BH) method on the ANOVA p-values to control for false discovery rate. The peptide counts at various FDR thresholds are tabulated below.

FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	40	194	1426	3012	4538	5716	6881	8032	9157	10236

It appears that the BH method gives a more conservative significant peptide counts. We export the list of 4538 peptides at 5% BH FDR (together with the list of their corresponding proteins) onto the *"All_Peptides"* sheet in the Excel file *"BH_FDR_5prct.xlsx"*.

3 Visualization of One-Way ANOVA

We have identified 4538 peptides at 5% Benjamini-Hochberg(BH) false discovery rate (FDR). We would like to obtain some graphical representations to illustrate how the $\log_2$ fluorescence levels differ across different cancer stages for these 4538 peptides.

For each peptide, we remove the grand mean (row mean) of the $\log_2$ fluorescence levels for all patients, before applying the following visualization techniques:

- Principal Component Analysis (PCA)
- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Heatmap

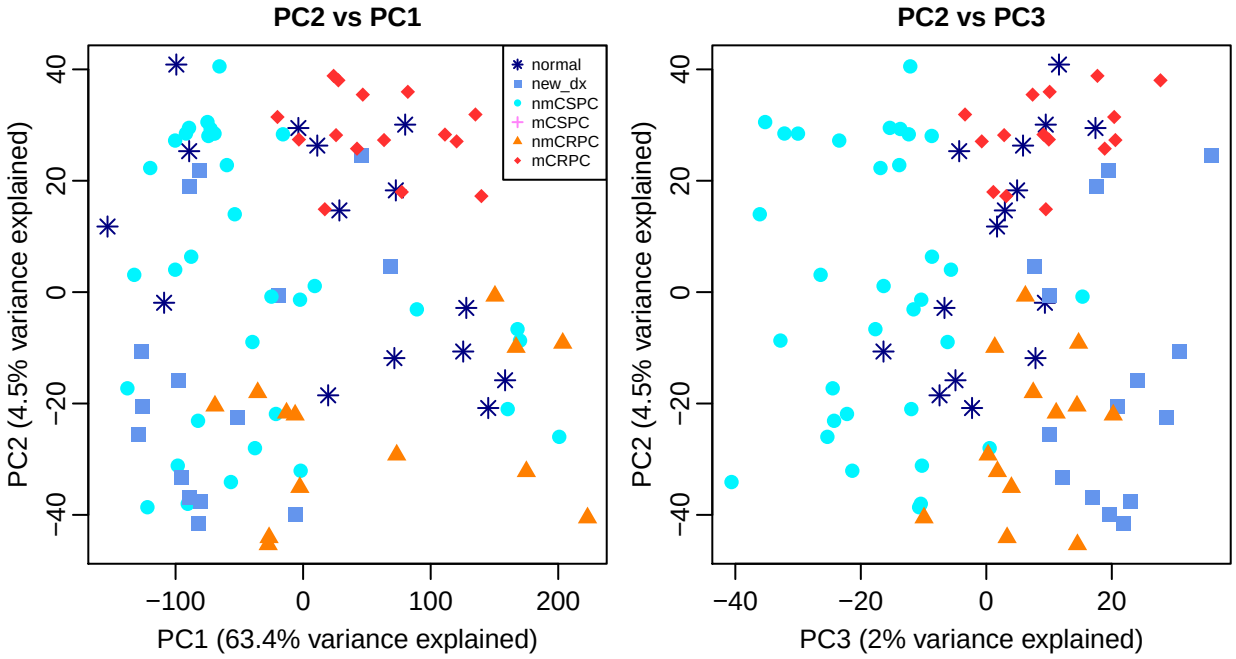
Based on our one-way-ANOVA model assumption, if there is no cancer-stage effect, we expect these residual $\log_2$ fluorescence to be random noises. Any observed (clustering) patterns among these residual data points reveal the effects of various stages of prostate cancer.

For purpose of uniformity, we also use the same color scheme to distinguish the different stages of cancer patients (notice how the spectrum of colors changes with severity of the cancer stages):

- navy for healthy subjects
- cornflower_blue for `new_dx` newly diagnosed patients
- turquoise for `nmCSPC` patients
- light pink for `mCSPC` patients – these patients have no technical replicates and are excluded from this analysis
- dark orange for `nmCSPC` patients
- dark red for `mCSPC` patients

3.1 Principal Component Analysis (PCA)

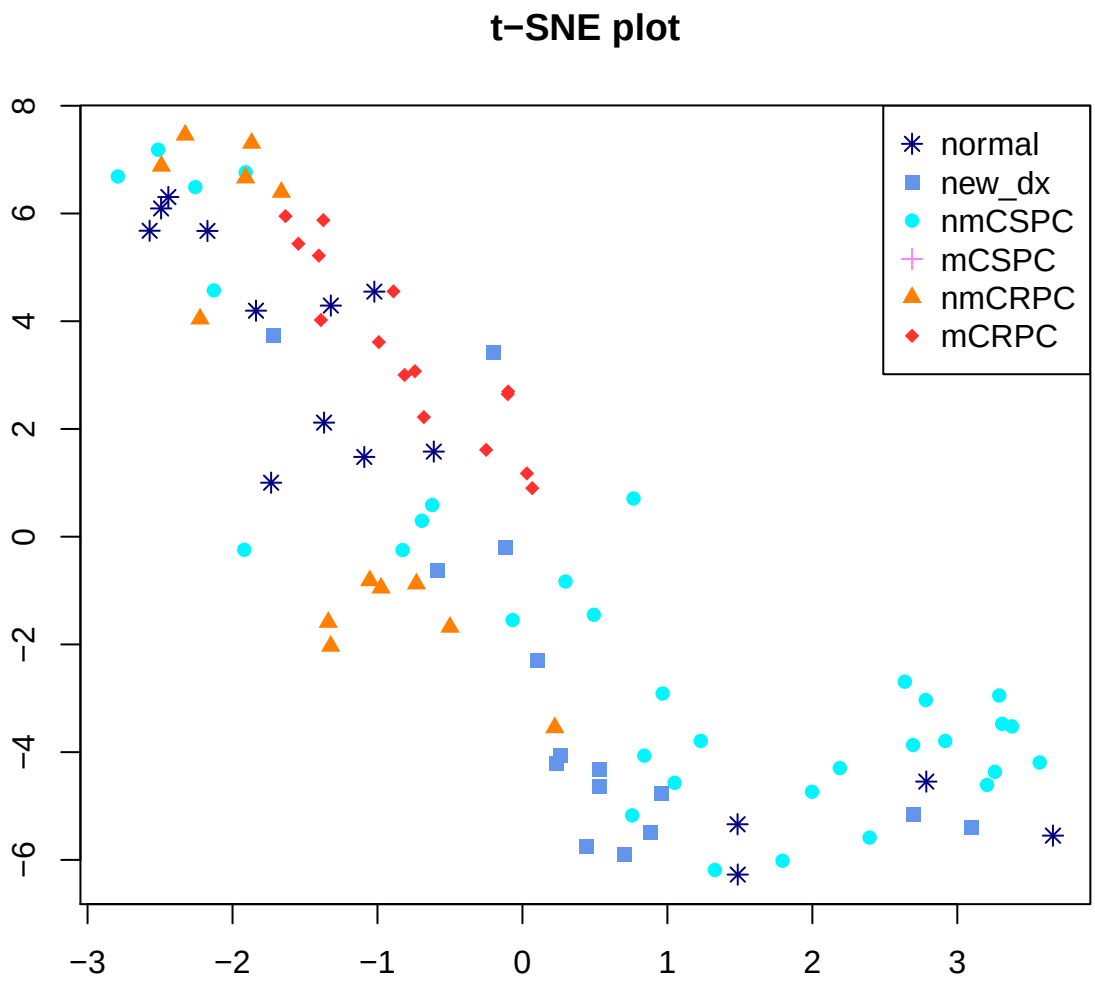
First we apply principal component analysis (PCA).



From the "*PC2 vs PC1*" plot, we observe that all mCRPC points are clustered at the top of the panel, whereas nmCRPC observations hover at the bottom of panel. Normal (healthy) and new_dx (newly diagnosed) subjects are "all over the place". In the "*PC2 vs PC3*" plot, we observe the same patterns, and most of the nmCSPC patients take up the left part of the panel. The percentage of variance explained for each principal component (PC) is shown on the axis.

3.2 t-SNE (t-distributed Stochastic Neighbor Embedding)

Just like PCA, t-SNE is a dimensionality reduction technique which was first introduced by van der Maaten and Hinton in 2008.

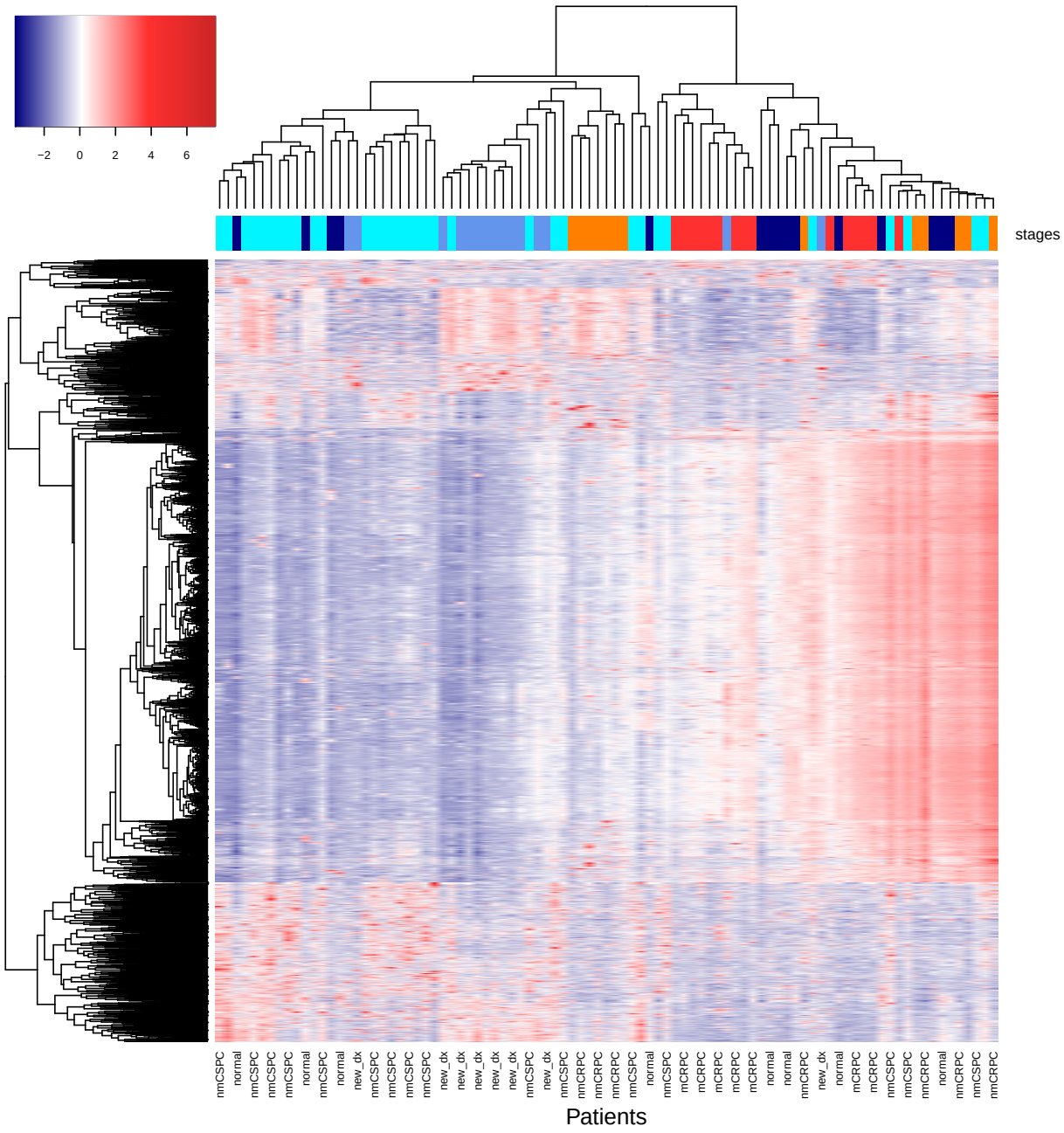


From the t-SNE plot, we observe that:

- The mCRPC points are clustered together near top-left of the plot.
- The nmCRPC patients are not too far off from the mCRPC subjects. There seems to be 2 clusters of nmCRPC patients.
- Most of the nmCSPC patients are clustered at the bottom right of the plot.
- Normal and new_dx subjects are somewhat "all over the place".

3.3 HeatMap

Finally we plot the heatmap of the $\log_2$ fluorescence (after removing row means) of the 4538 peptides at 5% BH-FDR.



Again, we observe similar patterns that nmCSPC (colored turquoise) and new_dx (colored cornflowerblue) subjects are mostly clustered to the left part of the heatmap, whereas the mCRPC (colored red) patients and most of the nmCRPC patients (colored darkorange) are clustered to the right part of the heatmap. Again, normal subjects (colored navy) are "all over the place".

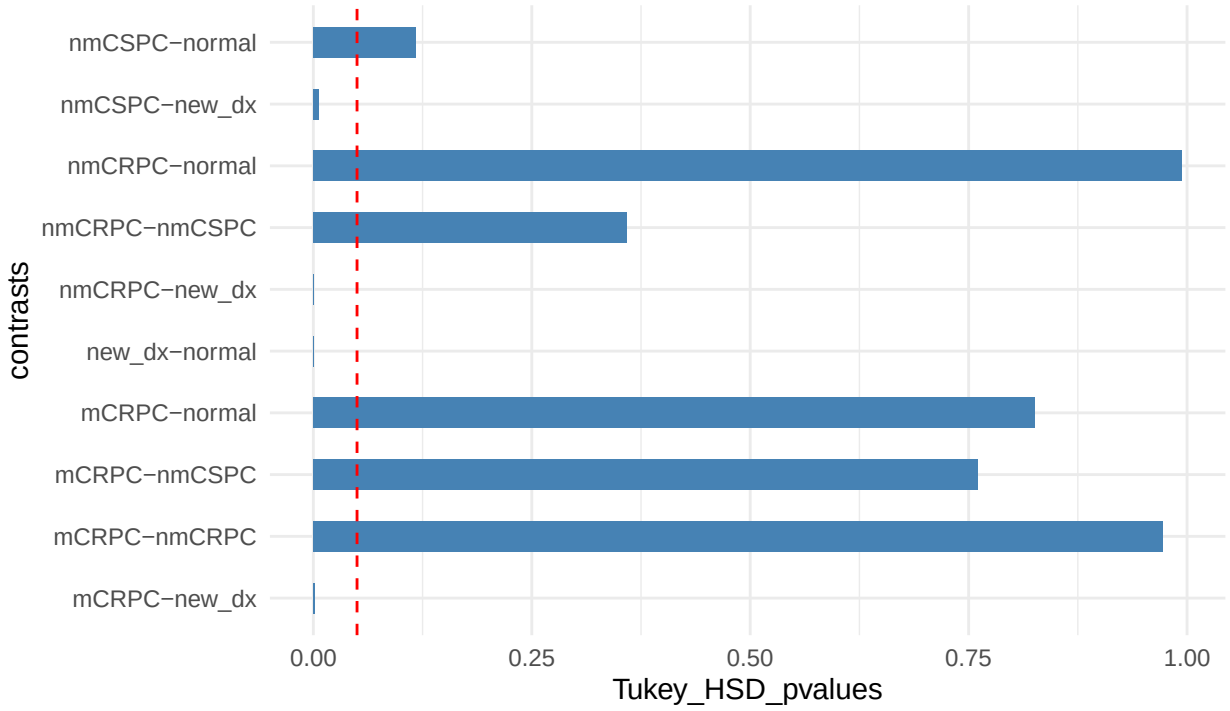
4 Tukey HSD (Honest Significant Difference)

The one-way ANOVA is helpful in revealing peptides that exhibit significant difference in the group means of $\log_2$ fluorescence levels among patients at different stages of cancer. We may be interested to find out which group or groups of patients that actually contribute to the significant difference in antibody responses.

For the 4538 peptides at 5% BH FDR, we proceed to perform the Tukey HSD analysis. Basically, it is a pairwise t-test (with standard error obtained from the previous one-way ANOVA and degrees of freedom adjusted for proper family-wise error rate) for all the possible contrasts (pairwise comparison of 2 groups of patients):

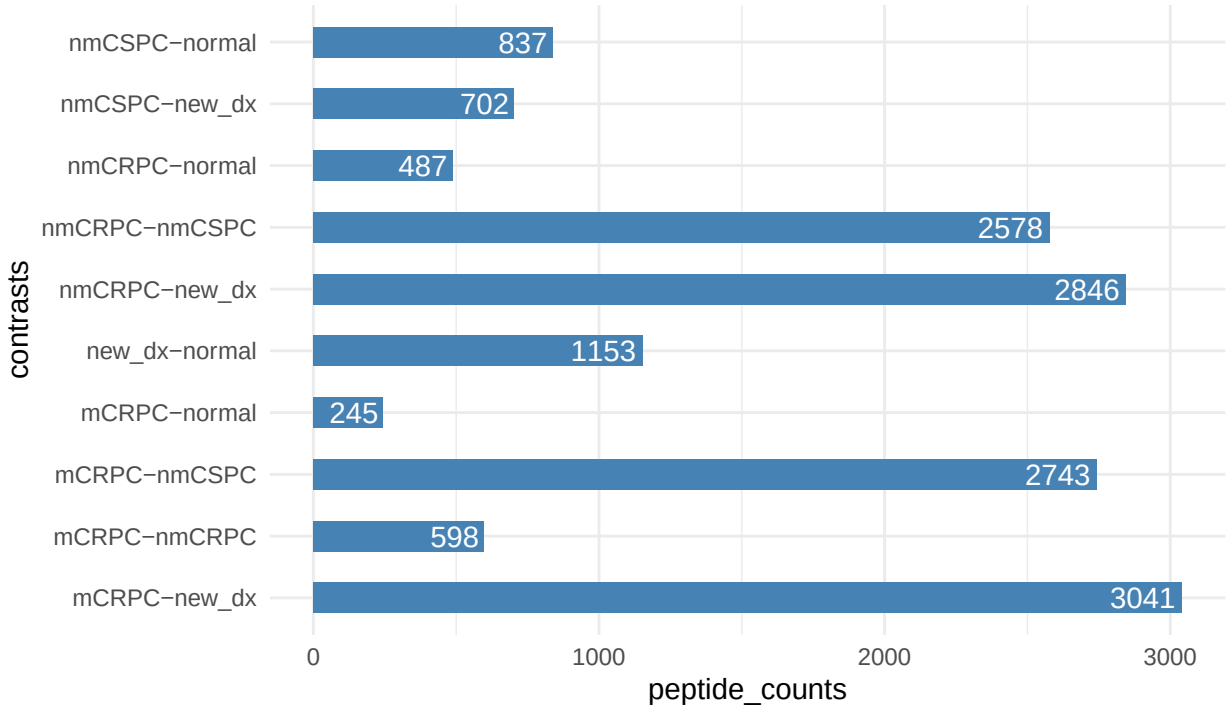
```
##  [1] "new_dx-normal" "nmCSPC-normal" "nmCRPC-normal" "mCRPC-normal"
##  [5] "nmCSPC-new_dx" "nmCRPC-new_dx" "mCRPC-new_dx"  "nmCRPC-nmCSPC"
##  [9] "mCRPC-nmCSPC"  "mCRPC-nmCRPC"
```

For example, the peptide *ADT12;129* is identified to be significant at 5% FDR. Upon closer inspection,



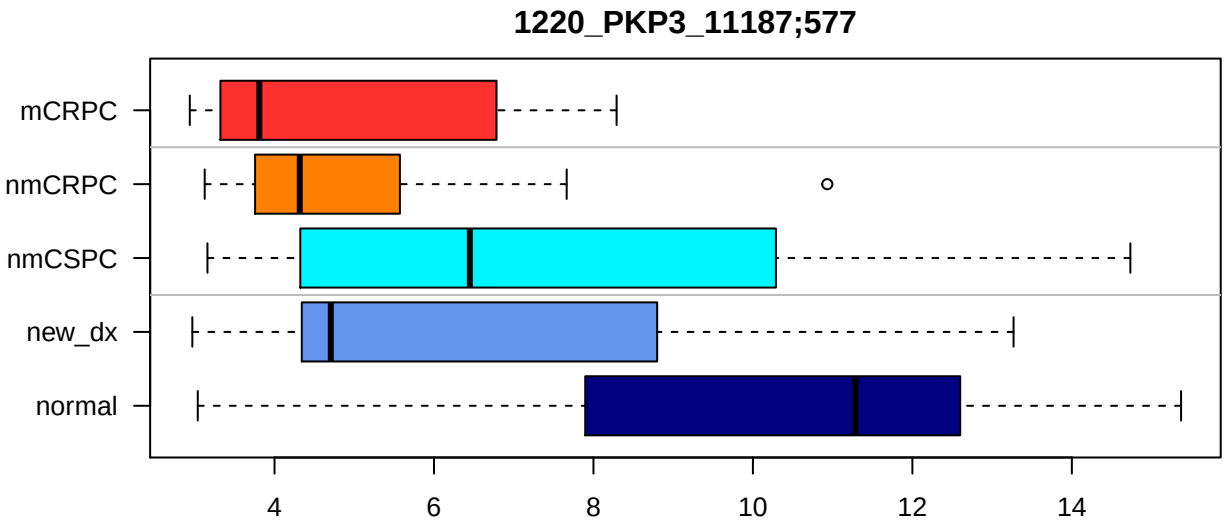
The Tukey HSD calculation reveals that for this peptide, the newly-diagnosed patients are the group of patients that are significantly different from the other groups in terms of antibody response. We export the table of Tukey HSD p-values of these 10 contrasts for the 4538 peptides to the *"Tukey_HSD"* sheet in the Excel file *"BH_FDR_5prct.xlsx"*. The table is also conditionally formatted to reveal cells that contain Tukey-HSD p-values of at most 0.05.

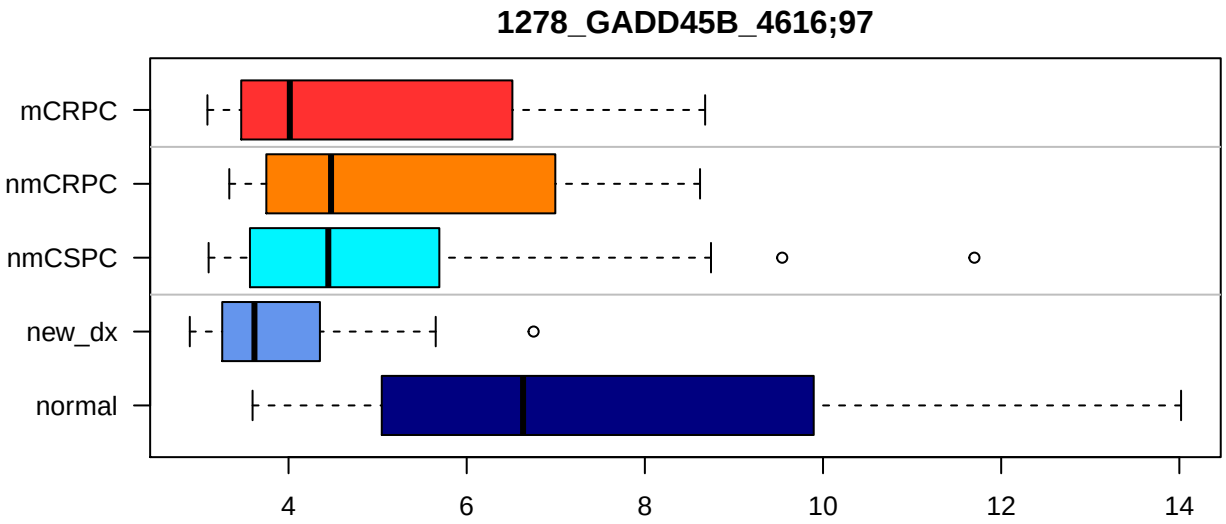
Among the 4538 peptides at 5% BH FDR, we plot the number of peptides that are found to be significant at 5% Tukey HSD under the various contrasts:

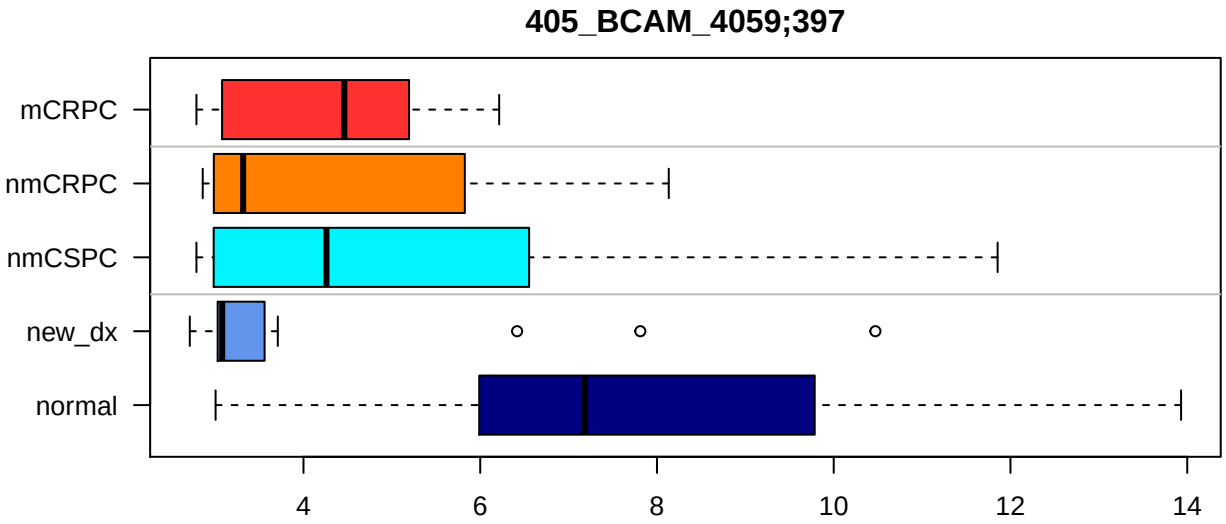


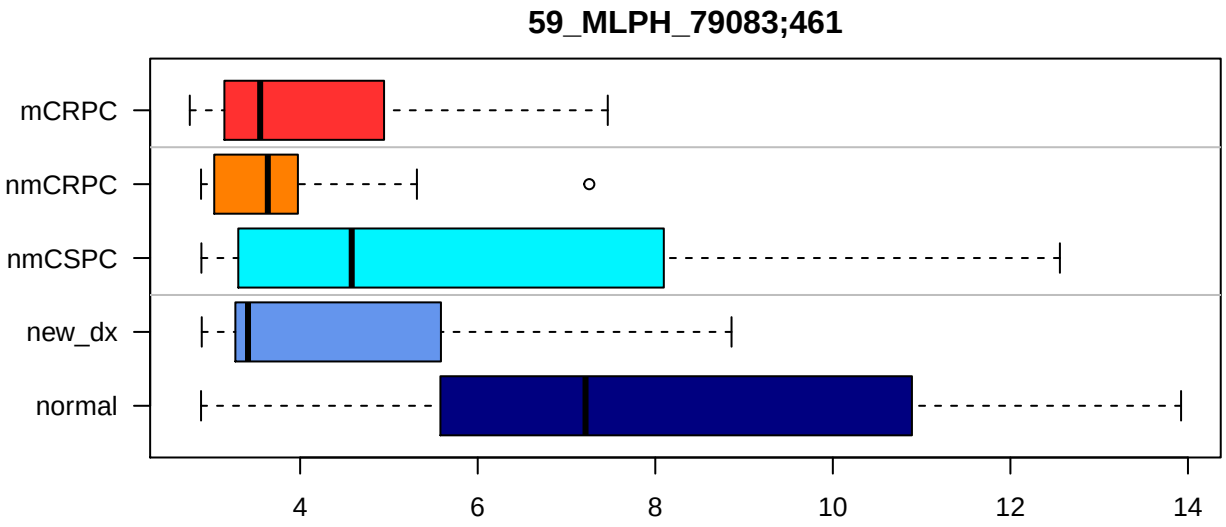
We could also plot the boxplots of a few peptides deemed significant under some interesting contrasts. The x-axis of the boxplots refers to the $\log_2$ median fluorescence level of the patients. For example, we want to find peptides that are significantly different between healthy subjects and cancer patients (but not siginificant among the different stages of patients).

```
ind <- which(
    tukey_pairwise_pattern$"new_dx-normal" == 1 &
    tukey_pairwise_pattern$"nmCSPC-normal" == 1 &
    tukey_pairwise_pattern$"nmCRPC-normal" == 1 &
    tukey_pairwise_pattern$"mCRPC-normal"  == 1 &
    tukey_pairwise_pattern$"nmCSPC-new_dx" == 0 &
    tukey_pairwise_pattern$"nmCRPC-new_dx" == 0 &
    tukey_pairwise_pattern$"mCRPC-new_dx"  == 0 &
    tukey_pairwise_pattern$"nmCRPC-nmCSPC" == 0 &
    tukey_pairwise_pattern$"mCRPC-nmCSPC"  == 0 &
    tukey_pairwise_pattern$"mCRPC-nmCRPC"  == 0 &
    rank_variance[one_way_anova_BH <= 0.05] <= 3 # just to pick out the peptides
  )
boxplot_mat <- anova_dat[ind,]
for (i in 1:nrow(boxplot_mat)){
  boxplot_func(mat = boxplot_mat, draw = i)
}
```


1220_PKP3_11187;577
mCRPC
nmCRPC
nmCSPC
new_dx
normal
4
6
8
10
12
14

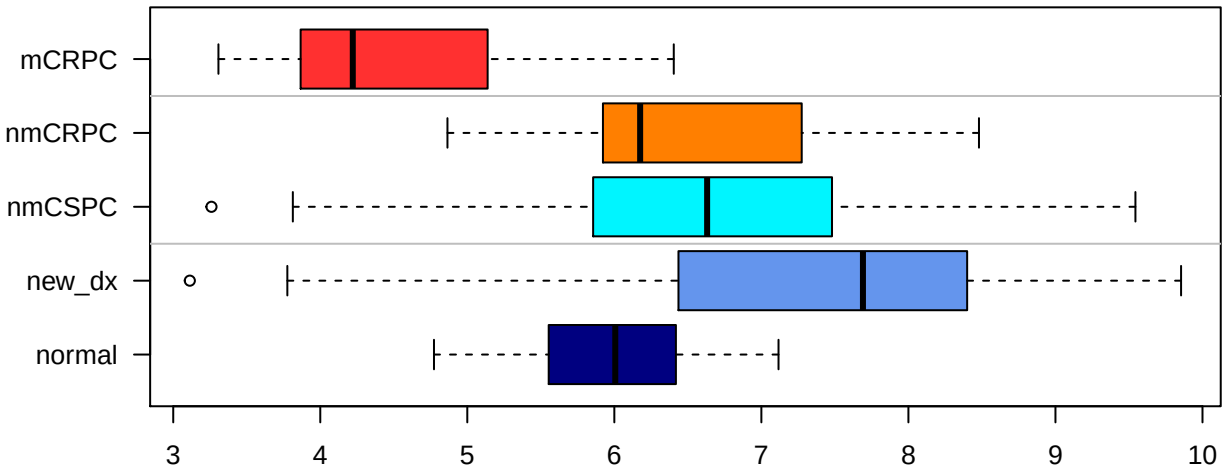

1278_GADD45B_4616;97
mCRPC
nmCRPC
nmCSPC
new_dx
normal
4
6
8
10
12
14

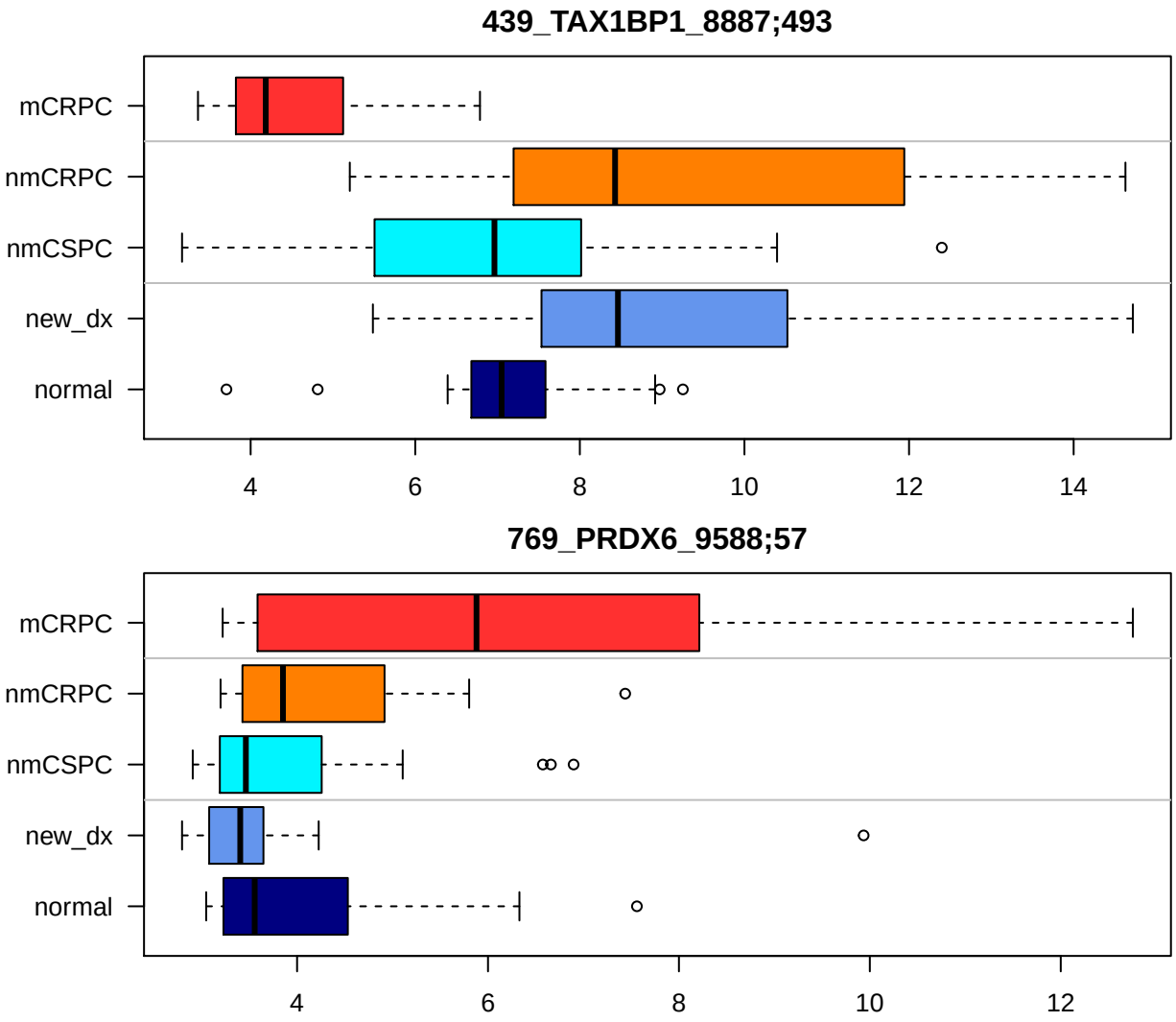

405_BCAM_4059;397
mCRPC
nmCRPC
nmCSPC
new_dx
normal
4
6
8
10
12
14



We can also look at peptides that are significantly different between mCRPC patients and the other subjects.

```
ind <- which(
    tukey_pairwise_pattern$"mCRPC-normal"  == 1 &
    tukey_pairwise_pattern$"mCRPC-new_dx"  == 1 &
    tukey_pairwise_pattern$"mCRPC-nmCSPC"  == 1 &
    tukey_pairwise_pattern$"mCRPC-nmCRPC"  == 1 &
    rank_variance[one_way_anova_BH <= 0.05] <= 13 # just to pick out some peptides with large marginal
  )
boxplot_mat <- anova_dat[ind,]
for (i in 1:nrow(boxplot_mat)){
  boxplot_func(mat = boxplot_mat, draw = i)
}
```



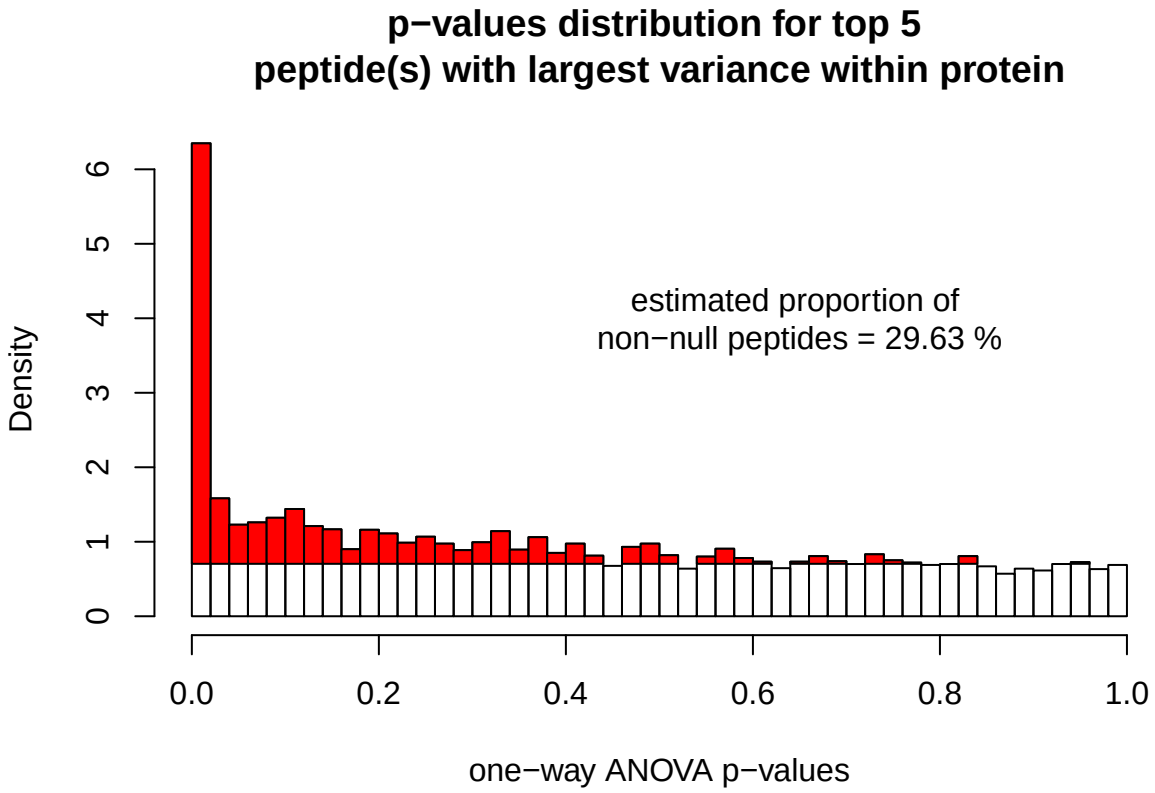


5 Marginal Variance Filtering

Next, we aim to identify proteins by first filtering for potentially "strong" peptides in terms of huge marginal variance of peptide fluorescence. Specifically, for every protein, we would like to filter the top few peptides with the largest marginal variance of $\log_2$ fluorescence before we apply FDR control.

5.1 Top 5 peptides with largest marginal variance

First, we filter top 5 peptides with largest marginal variance in each protein. Then we plot histogram of p-values.



Next, we tabulate peptide counts at different BH-adjusted p-values thresholds.

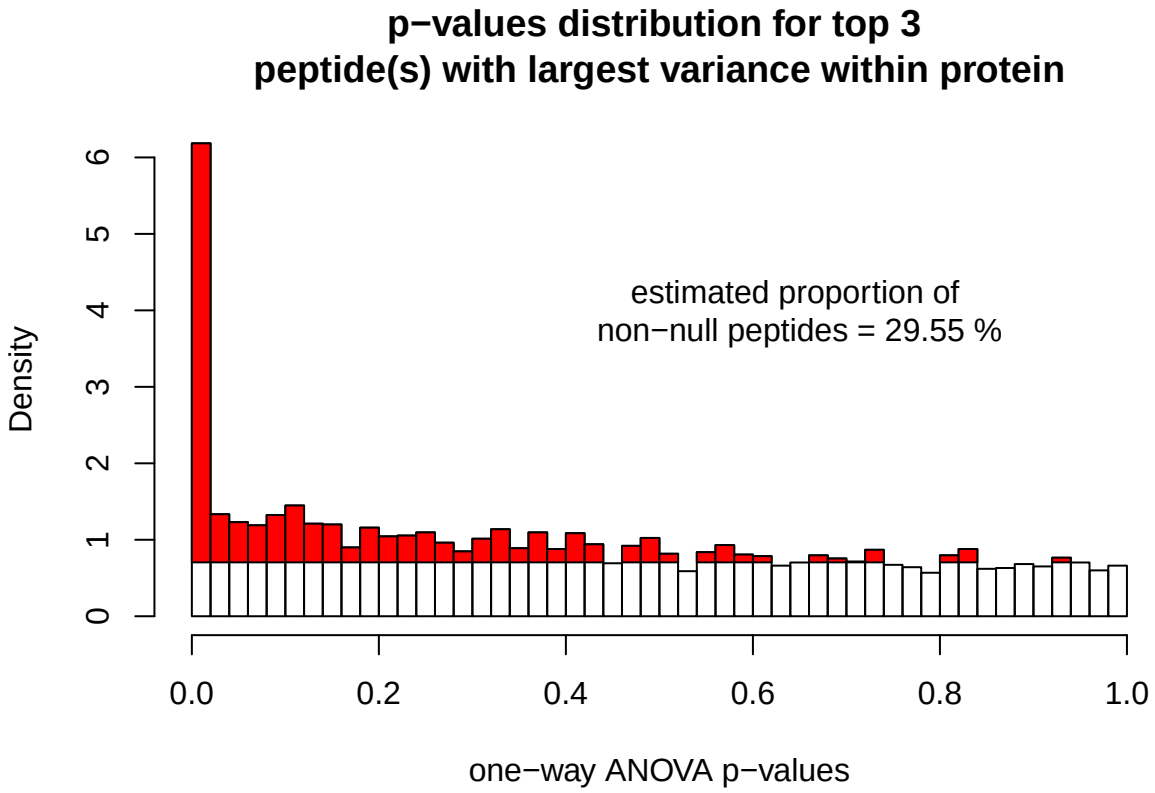
FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	3	386	522	595	671	715	762	799	833	861

The list of significant peptides (and their corresponding proteins) at 5% BH FDR could be found on the *"Top_5"* sheet in the Excel file *"BH_FDR_5prct.xlsx"*. We also tabulate peptide counts at different Storey's q-values thresholds.

FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	171	510	626	703	766	816	861	905	933	977

5.2 Top 3 peptides with largest marginal variance

Now, we filter top 3 peptides with largest marginal variance in each protein. Then we plot histogram of p-values.



Next, we tabulate peptide counts at different BH-adjusted p-values thresholds.

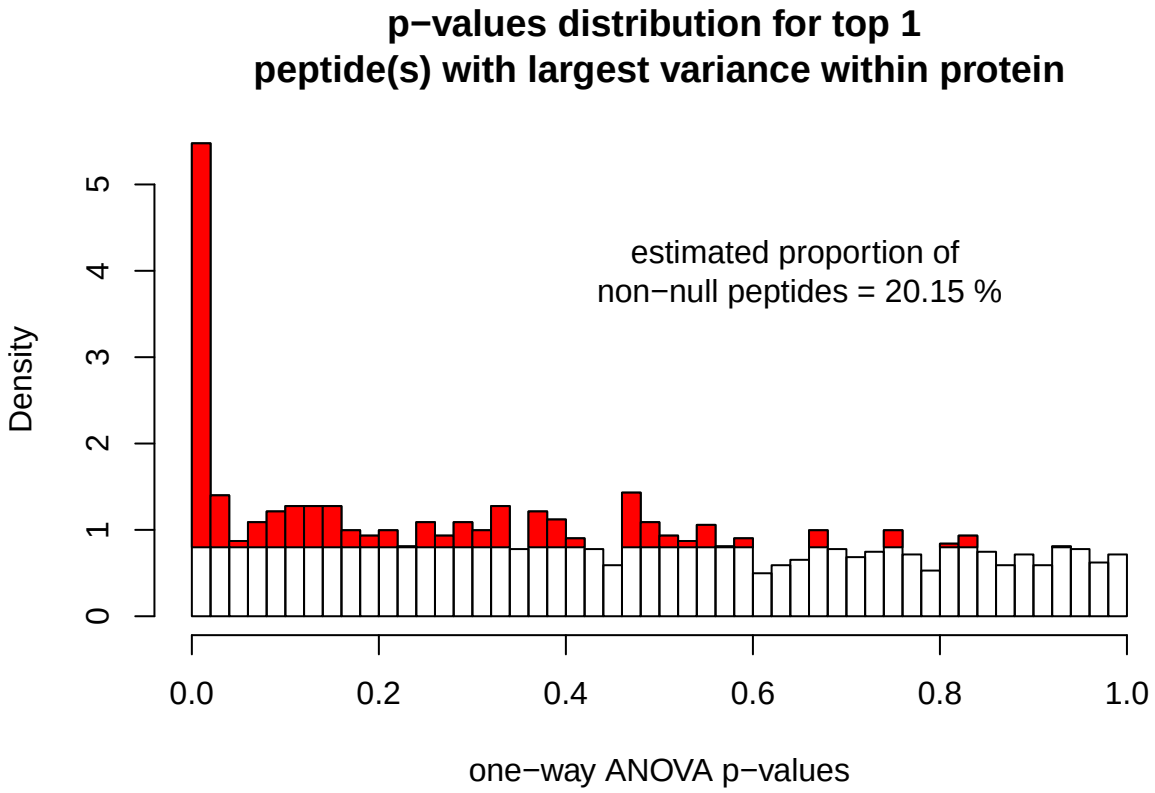
FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	49	242	320	363	402	428	446	472	491	510

The list of significant peptides (and their corresponding proteins) at 5% BH FDR could be found on the *"Top_3"* sheet in the Excel file *"BH_FDR_5prct.xlsx"*. We also tabulate peptide counts at different Storey's q-values thresholds.

FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	122	316	379	423	452	485	506	532	550	577

5.3 Top peptide with largest marginal variance

Now, we filter top peptide with largest marginal variance in each protein. This can be taken as protein-level analysis as we take the peptide with largest marginal variance a representative of its corresponding protein. Then we plot histogram of p-values.



Next, we tabulate peptide counts at different BH-adjusted p-values thresholds.

FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	21	65	88	101	106	112	121	124	131	138

The list of significant peptides (and their corresponding proteins) at 5% BH FDR could be found on the *"Top_1"* sheet in the Excel file *"BH_FDR_5prct.xlsx"*. We also tabulate peptide counts at different Storey's q-values thresholds.

FDR threshold	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Peptide counts	34	78	98	106	114	124	130	138	151	157