Capstone Question 1: What are most differentially expressed markers

David Granadier

12/10/2020

BRIEF INTRO: Objective here is to find which markers are most differentially expressed by IHC in the three most common diagnoses documented in the paper. In theory, this is something that could be helpful in differentiating clinical smaples that may not be distinguishable by standard pathology.

Download necessary packages (may not use all of them)

Read in the marker expression spreadsheet. Remove the patient identifier column and group by the clinical diagnosis then organize by man expression for each of the IHC tested surface proteins.

```
#Having some directory issues that I dn't really understand, sometimes it works and some
times it doesn't; please contact if it doesn't and I will try to fix

#dir <- "/Users/dgranadi/Desktop/TCFB_2020/Capstone/data"
#df<- read_csv("marker_expression.csv")
df<- read_csv("/Users/dgranadi/Desktop/TCFB_2020/Capstone/data/marker_expression.csv")
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   .default = col_double(),
##   clin_diag = col_character(),
##   pad = col_character()
## )
## ℹ Use `spec()` for the full column specifications.
```

```
df
```

```
## # A tibble: 439 x 29
##      ck5   ck7  ck17  ck18  ck19  ck20   vim  muc1  muc2 muc5ac  muc6 berep4
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl>
##  1 0      92.5  20    92.5  92.5     0     5  20     0      80   30      92.5
##  2 0      20     0    92.5  50       0     0   0     0       0    0       0
##  3 0      92.5   0    92.5  92.5    20     0  40     0      80    5      40
##  4 0.714  70    18.3  92.5  85       5    15  80     0      40   15      92.5
##  5 0       0     0    92.5  40       0     0   0     0       0    5       5
##  6 5      92.5   5    91.2  92.5     0    40  40     0       0   70      92.5
##  7 0      92.5  60    92.5  92.5     0     0  92.5  0      70   20      82
##  8 0       0     0    92.5  85       0     0   0     0       0    0      15
##  9 1.25   92.5  25    92.5  92.5    20     0  45.8  2.22   61.9 8.12    92.5
## 10 7.5    92.5  92.5  92.5  92.5     0     0  92.5  0      80   15      80
## # … with 429 more rows, and 17 more variables: ema <dbl>, mcea <dbl>,
## #   pcea <dbl>, ca125 <dbl>, ca19.9 <dbl>, maspin <dbl>, wt1cyt <dbl>,
## #   cdx2 <dbl>, p53 <dbl>, p63 <dbl>, ki67 <dbl>, smad4 <dbl>, chra <dbl>,
## #   cd56 <dbl>, cd10 <dbl>, clin_diag <chr>, pad <chr>
```

```
df_mean<-select(df, -pad) %>%
  group_by(clin_diag) %>%
  summarize_all(.funs = c(mean = "mean")) %>%
  print()
```

```
## # A tibble: 7 x 28
##   clin_diag ck5_mean ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean vim_mean
##   <chr>        <dbl>    <dbl>     <dbl>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 Ampullar…     1.75     59.1     10.0       90.3      92.0     32.4     11.6
## 2 Distal b…     1.13     83.1     30.6       82.8      92.5      4.38     1.88
## 3 Ductal p…     6.04     81.9     33.1       90.8      91.6     10.3     17.6
## 4 Gallblad…     5.23     76.0     26.2       92.0      90.2      7.78    15.1
## 5 Hepatoce…     1.01     10.5      5.96      86.5      16.3      7.55     6.54
## 6 Intrahep…     4.37     81.2     13.2       89.7      90.5      7.88    40.1
## 7 Perihila…     3.12     69.5     24.9       91.5      91.8      5.97    14.2
## # … with 20 more variables: muc1_mean <dbl>, muc2_mean <dbl>,
## #   muc5ac_mean <dbl>, muc6_mean <dbl>, berep4_mean <dbl>, ema_mean <dbl>,
## #   mcea_mean <dbl>, pcea_mean <dbl>, ca125_mean <dbl>, ca19.9_mean <dbl>,
## #   maspin_mean <dbl>, wt1cyt_mean <dbl>, cdx2_mean <dbl>, p53_mean <dbl>,
## #   p63_mean <dbl>, ki67_mean <dbl>, smad4_mean <dbl>, chra_mean <dbl>,
## #   cd56_mean <dbl>, cd10_mean <dbl>
```

Narrow down the above dataframe to only include the three most common diagnoses (Ductal pancreatic adenocarcinomas, Hepatocellular carcinomas, and Intrahepatic cholangiocarcinoma which will be referred to from here as DPA, HC, and IC, respectively).

```
df_only3<-df_mean %>%
  slice(3 , 5 , 6) %>%
  print()
```

```
## # A tibble: 3 x 28
##   clin_diag ck5_mean ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean vim_mean
##   <chr>        <dbl>    <dbl>     <dbl>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 Ductal p…     6.04     81.9     33.1       90.8      91.6     10.3     17.6 
## 2 Hepatoce…     1.01     10.5      5.96      86.5      16.3      7.55     6.54
## 3 Intrahep…     4.37     81.2     13.2       89.7      90.5      7.88    40.1 
## # … with 20 more variables: muc1_mean <dbl>, muc2_mean <dbl>,
## #   muc5ac_mean <dbl>, muc6_mean <dbl>, berep4_mean <dbl>, ema_mean <dbl>,
## #   mcea_mean <dbl>, pcea_mean <dbl>, ca125_mean <dbl>, ca19.9_mean <dbl>,
## #   maspin_mean <dbl>, wt1cyt_mean <dbl>, cdx2_mean <dbl>, p53_mean <dbl>,
## #   p63_mean <dbl>, ki67_mean <dbl>, smad4_mean <dbl>, chra_mean <dbl>,
## #   cd56_mean <dbl>, cd10_mean <dbl>
```

To find the difference in expression between markers calculate the ratio of the expression between diagnoses i.e. ck5 measurement for DPA divided by ck5 measurement for HC is the ratio of expression of ck5 between those diagnoses. The highest number ratio indicates the most over expressed marker for DPA relative to Diagnosis HC.

```
DPA_vs_HC <- (df_only3[1, -1] / df_only3[2, -1]) %>%
  print()
```

```
##   ck5_mean ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean vim_mean muc1_mean
## 1 5.984252  7.79969  5.549215  1.048869  5.620523  1.361432 2.692778  3.388009
##   muc2_mean muc5ac_mean muc6_mean berep4_mean ema_mean mcea_mean pcea_mean
## 1  5.506933     5.59636  1.992443    7.521266 4.734668  7.876751  1.963052
##   ca125_mean ca19.9_mean maspin_mean wt1cyt_mean cdx2_mean p53_mean p63_mean
## 1   7.940528    6.731919     4.76201    1.285514  4.703883 2.959595 2.424609
##   ki67_mean smad4_mean chra_mean cd56_mean cd10_mean
## 1  1.598131  0.5803368  5.457059 0.2607519 0.6439782
```

```
marker_DPA_vs_HC <- DPA_vs_HC[max.col(DPA_vs_HC)]
print(marker_DPA_vs_HC)
```

```
##   ca125_mean
## 1   7.940528
```

The marker most overexpressed in Ductal Pancreatic Adenocarcinoma relative to Hepatocellular Carcinoma is ca125 and it is expressed 7.94 times higher

```
#Using the inverse of the DPA vs HC dataframe, find the marker most overexpressed in HC
 relative to DPA
HC_vs_DPA <- (DPA_vs_HC ^-1 ) %>%
  print()
```

```
##    ck5_mean  ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean  vim_mean
## 1 0.1671053 0.1282102 0.1802057 0.9534076 0.1779194 0.7345205 0.3713637
##   muc1_mean muc2_mean muc5ac_mean muc6_mean berep4_mean  ema_mean mcea_mean
## 1 0.2951586 0.1815893   0.1786876 0.5018963   0.1329563 0.2112081 0.1269559
##   pcea_mean ca125_mean ca19.9_mean maspin_mean wt1cyt_mean cdx2_mean  p53_mean
## 1  0.509411  0.1259362   0.1485461   0.2099953   0.7778992 0.2125903 0.3378841
##    p63_mean ki67_mean smad4_mean chra_mean cd56_mean cd10_mean
## 1 0.4124377 0.6257311   1.723137 0.1832489  3.835063  1.552848
```

```
marker_HC_vs_DPA <- HC_vs_DPA[max.col(HC_vs_DPA)]
print(marker_HC_vs_DPA)
```

```
##   cd56_mean
## 1  3.835063
```

The marker most overexpressed in Hepatocellular Carcinoma relative to Ductal Pancreatic Adenocarcinoma is cd56 and it is expressed 3.83 times higher

```
DPA_vs_IC <- (df_only3[1, -1] / df_only3[3, -1]) %>%
  print()
```

```
##   ck5_mean ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean  vim_mean muc1_mean
## 1 1.380673 1.009237  2.505977  1.011483  1.012083  1.304828 0.4391535  1.151899
##   muc2_mean muc5ac_mean muc6_mean berep4_mean ema_mean mcea_mean pcea_mean
## 1  4.064427    3.411659 0.7698541    1.076163 1.113788  3.021853  1.171845
##   ca125_mean ca19.9_mean maspin_mean wt1cyt_mean cdx2_mean p53_mean p63_mean
## 1   2.612795    1.332964    2.125131   0.4086118  2.431299  1.79049 2.485414
##   ki67_mean smad4_mean chra_mean  cd56_mean cd10_mean
## 1  1.169172  0.5363373  1.344692 0.09595703 0.9128388
```

```
marker_DPA_vs_IC <- DPA_vs_IC[max.col(DPA_vs_IC)]
print(marker_DPA_vs_IC)
```

```
##   muc2_mean
## 1  4.064427
```

The marker most overexpressed in Ductal Pancreatic Adenocarcinoma relative to Intrahepatic Cholangiocarcinoma is muc2 and it is expressed 4.06 times higher

```
#Using the inverse of the DPA vs IC dataframe, find the marker most overexpressed in IC relative to DPA
IC_vs_DPA <- (DPA_vs_IC ^-1 ) %>%
  print()
```

```
##    ck5_mean  ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean vim_mean
## 1 0.7242845 0.9908479  0.399046 0.9886471 0.9880613 0.7663846 2.277108
##   muc1_mean muc2_mean muc5ac_mean muc6_mean berep4_mean  ema_mean mcea_mean
## 1 0.8681316 0.2460372   0.2931125  1.298947   0.9292275 0.8978371 0.3309228
##   pcea_mean ca125_mean ca19.9_mean maspin_mean wt1cyt_mean cdx2_mean  p53_mean
## 1 0.8533552   0.382732   0.7502076   0.4705592    2.447311 0.4113028 0.5585063
##    p63_mean ki67_mean smad4_mean chra_mean cd56_mean cd10_mean
## 1 0.4023475 0.8553058   1.864498 0.7436645  10.42133  1.095484
```

```
marker_IC_vs_DPA <- IC_vs_DPA[max.col(IC_vs_DPA)]
print(marker_IC_vs_DPA)
```

```
##   cd56_mean
## 1  10.42133
```

The marker most overexpressed in Intrahepatic Cholangiocarcinoma relative to Ductal Pancreatic Adenocarcinoma is cd56 and it is expressed 10.4 times higher

```
IC_vs_HC <- (df_only3[3, -1] / df_only3[2, -1]) %>%
  print()
```

```
##   ck5_mean ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean vim_mean muc1_mean
## 1 4.334301 7.728307  2.214392  1.036962  5.553421  1.043381 6.131746  2.941238
##   muc2_mean muc5ac_mean muc6_mean berep4_mean ema_mean mcea_mean pcea_mean
## 1   1.35491    1.640363  2.588079    6.988967 4.250961  2.606597   1.67518
##   ca125_mean ca19.9_mean maspin_mean wt1cyt_mean cdx2_mean p53_mean  p63_mean
## 1   3.039094    5.050337    2.240808    3.146051   1.93472 1.652952 0.9755353
##   ki67_mean smad4_mean chra_mean cd56_mean cd10_mean
## 1   1.36689   1.082037  4.058221  2.717382 0.7054676
```

```
marker_IC_vs_HC <- IC_vs_HC[max.col(IC_vs_HC)]
print(marker_IC_vs_HC)
```

```
##   ck7_mean
## 1 7.728307
```

The marker most overexpressed in Intrahepatic Cholangiocarcinoma relative to Hepatic Carcinoma is ck7 and it is expressed 7.73 times higher

```
#Using the inverse of the DPA vs IC dataframe, find the marker most overexpressed in IC
 relative to DPA
HC_vs_IC <- (IC_vs_HC ^-1 ) %>%
  print()
```

```
##    ck5_mean  ck7_mean ck17_mean ck18_mean ck19_mean ck20_mean  vim_mean
## 1 0.2307177 0.1293944 0.4515912 0.9643558 0.1800692 0.9584229 0.1630857
##   muc1_mean muc2_mean muc5ac_mean muc6_mean berep4_mean  ema_mean mcea_mean
## 1 0.3399929 0.7380564   0.6096211 0.3863869   0.1430827 0.2352409  0.383642
##   pcea_mean ca125_mean ca19.9_mean maspin_mean wt1cyt_mean cdx2_mean  p53_mean
## 1 0.5969506  0.3290455   0.1980066   0.4462676   0.3178588 0.5168706 0.6049781
##   p63_mean ki67_mean smad4_mean chra_mean cd56_mean cd10_mean
## 1 1.025078 0.7315876  0.9241828 0.2464134 0.3680013    1.4175
```

```
marker_HC_vs_IC <- HC_vs_IC[max.col(HC_vs_IC)]
print(marker_HC_vs_IC)
```

```
##   cd10_mean
## 1    1.4175
```

The marker most overexpressed in Hepatic Carcinoma relative to Intragepatic Cholangiocarcinoma is cd10 and it is expressed 1.42 times higher