Bioinformatics Data Engineer Challenge

Prisma Erika Lopez Jimenez

9/19/2019

```
# The ACTIONABLE ALTERATIONS database was accessed from the OncoKB database:
# oncokb.org/dataAccess
# Below are R code embedded within `code chunks`. Included are previews of the code output.
# Beside each line of R code are hashtag comments describing the purpose of the code.
# Where appropriate, answers to the questions for the bioinformatics data engineer
# challenge are included.
act.alt # preview of the ACTIONABLE ALTERATIONS database data frame:
```

```
## # A tibble: 250 x 11
##    Isoform RefSeq `Entrez Gene ID` `Hugo Symbol` Alteration
##    <chr>   <chr>             <dbl> <chr>         <chr>
##  1 ENST00~ NM_00~               25 ABL1          T315I
##  2 ENST00~ NM_00~               25 ABL1          T315I
##  3 ENST00~ NM_00~             1956 EGFR          Exon 20 i~
##  4 ENST00~ NM_00~             1956 EGFR          T790M
##  5 ENST00~ NM_03~             3845 KRAS          Oncogenic~
##  6 ENST00~ NM_00~             4893 NRAS          Oncogenic~
##  7 ENST00~ NM_00~             5156 PDGFRA        D842V
##  8 ENST00~ NM_00~               25 ABL1          BCR-ABL1 ~
##  9 ENST00~ NM_00~               25 ABL1          T315I
## 10 ENST00~ NM_00~               25 ABL1          BCR-ABL1 ~
## # ... with 240 more rows, and 6 more variables: `Protein Change` <chr>,
## #   `Cancer Type` <chr>, Level <chr>, `Drugs(s)` <chr>, `PMIDs for
## #   drug` <chr>, `Abstracts for drug` <chr>
```

```
# QUESTION 1
# How many genes in total are included here?
  # logic: 250, the size of the table
  # answer: 55 unique genes with given `Entrez Gene ID`
### Code:
dim(act.alt) # data frame dimensions
```

```
## [1] 250  11
```

```
length(unique(act.alt$`Entrez Gene ID`)) # count how many unique gene IDs are in the list
```

```
## [1] 55
```

```
length(unique(act.alt$`Hugo Symbol`)) # count how many unique gene symbols are in the list
```

```
## [1] 55
```

```r
# QUESTION 2
# List all genes that are targetable by afatinib
  # logic: 1, we assume that drugs are specific to a macromolecule
  # answer: 1 targetable gene by Afatinib. Gene symbol: EGFR, gene ID: 1956
### Code:
# filtered for Afatinib targeting (inclusive of other drugs): 16 entries
act.alt.af=dplyr::filter(act.alt,grepl('Afatinib',`Drugs(s)`))
length(unique(act.alt.af$`Entrez Gene ID`)) # 1 unique targetable gene by afatinib
```

```
## [1] 1
```

```r
act.alt.af.gID=unique(act.alt.af$`Entrez Gene ID`) # entrez gene ID vector
act.alt.af.symbol=unique(act.alt.af$`Hugo Symbol`) # gene ID symbol vector
act.alt.af.symbol # gene symbols
```

```
## [1] "EGFR"
```

```r
act.alt.af.gID # gene ID
```

```
## [1] 1956
```

```r
# QUESTION 3
# What are all the cancer types that can be treated by a targeted therapy for
# any mutations at the 600th codon of BRAF?
  # logic: less than 9 since there are 9 unique cancer types with mutations in
  # the 600th location in BRAF that are targetable by a drug

  # answer: 6 cancer types:
  # Anaplastic Thyroid Cancer
  # Erdheim-Chester Disease
  # Melanoma
  # Non-Small Cell Lung Cancer
  # Colorectal Cancer
  # Hairy Cell Leukemia

  # sub-answer: 5 cancer types at V600/E/K location:
  # Anaplastic Thyroid Cancer
  # Melanoma
  # Non-Small Cell Lung Cancer
  # Colorectal Cancer
  # Hairy Cell Leukemia

  # sub-answer: 2 cancer types at V600 annotation:
  # Erdheim-Chester Disease
  # Colorectal Cancer
### Code:
braf=act.alt %>% filter(`Hugo Symbol` == 'BRAF') # filter those for BRAF gene, 19 entries
unique(braf$`Cancer Type`) # 9 unique cancer types targetable by drugs w/ mut. @ BRAF gene
```

```
## [1] "Anaplastic Thyroid Cancer"  "Erdheim-Chester Disease"
## [3] "Melanoma"                   "Non-Small Cell Lung Cancer"
## [5] "Colorectal Cancer"           "Hairy Cell Leukemia"
## [7] "Histiocytosis"               "Ovarian Cancer"
## [9] "All Solid Tumors"
```

```
braf.v6=dplyr::filter(braf,grepl('V6',`Alteration`))
braf.v6 # table of BRAF alterations at V600/E/K location
```

```
## # A tibble: 8 x 11
##   Isoform RefSeq `Entrez Gene ID` `Hugo Symbol` Alteration `Protein Change`
##   <chr>   <chr>             <dbl> <chr>         <chr>      <chr>
## 1 ENST00~ NM_00~              673 BRAF          V600E      V600E
## 2 ENST00~ NM_00~              673 BRAF          V600       V600
## 3 ENST00~ NM_00~              673 BRAF          V600E      V600E
## 4 ENST00~ NM_00~              673 BRAF          V600K      V600K
## 5 ENST00~ NM_00~              673 BRAF          V600E      V600E
## 6 ENST00~ NM_00~              673 BRAF          V600E      V600E
## 7 ENST00~ NM_00~              673 BRAF          V600E      V600E
## 8 ENST00~ NM_00~              673 BRAF          V600       V600
## # ... with 5 more variables: `Cancer Type` <chr>, Level <chr>,
## #   `Drugs(s)` <chr>, `PMIDs for drug` <chr>, `Abstracts for drug` <chr>
```

```
unique(braf.v6$`Cancer Type`) # 6 cancer types drug targetable if alteration @ BRAF V600/E/K position
```

```
## [1] "Anaplastic Thyroid Cancer"  "Erdheim-Chester Disease"
## [3] "Melanoma"                   "Non-Small Cell Lung Cancer"
## [5] "Colorectal Cancer"          "Hairy Cell Leukemia"
```

```
# QUESTION 4
# If you were annotating a patient's genome with this data, how would you match a
# patient to an EGFR Exon 19 Insertion annotation listed here? Please describe any
# assumptions you might be making.

  # ANSWER: One way to match the gene information from this ACTIONABLE ALTERATIONS
  # table is to use the RefSeq number provided in the table to link it to a patient
  # table (presumably within a larger database). If the patient information is stored
  # in a table with columns for annotations at the EGFR Exon 19 gene, than we can join
  # the tables using the EGFR Exon 19 column in both tables.

### Code:
# subset those with EGFR gene in `Hugo Symbol`
act.alt.egfr=dplyr::filter(act.alt,grepl('EGFR',`Hugo Symbol`))
# subset those with Exon 19 in the `Alteration` column
act.alt.egfr.e19=dplyr::filter(act.alt.egfr,grepl('Exon 19',`Alteration`))
act.alt.egfr.e19 # 2 alterations at the EGFR exon 19 location
```

```
## # A tibble: 2 x 11
##   Isoform RefSeq `Entrez Gene ID` `Hugo Symbol` Alteration `Protein Change`
##   <chr>   <chr>             <dbl> <chr>         <chr>      <chr>
## 1 ENST00~ NM_00~             1956 EGFR          Exon 19 d~ 729_761del
## 2 ENST00~ NM_00~             1956 EGFR          Exon 19 i~ 729_761ins
## # ... with 5 more variables: `Cancer Type` <chr>, Level <chr>,
## #   `Drugs(s)` <chr>, `PMIDs for drug` <chr>, `Abstracts for drug` <chr>
```