Genie Cancer Data

Smiti Kaul

Feb 2 - present, 2018

Load the Data

```
cancer <- read.csv("C:/Users/kauls15/Desktop/github/genie/data/derived/data_clinical_patient.txt",
    stringsAsFactors = FALSE, sep = "\t")
cancer <- cancer[-c(1, 2, 3), ]
colnames(cancer) = cancer[1, ]  # the first row will be the header
cancer <- cancer[-1, -5]
head(cancer)

sample <- read.csv("C:/Users/kauls15/Desktop/github/genie/data/derived/data_clinical_sample.txt",
    stringsAsFactors = FALSE, sep = "\t")
sample$AGE_AT_SEQ_REPORT[sample$AGE_AT_SEQ_REPORT == "<18"] <- 17
sample$AGE_AT_SEQ_REPORT[sample$AGE_AT_SEQ_REPORT == ">89"] <- 90
head(sample)
```

Map Data to Discrete Variables

```
c <- cancer
# sort(unique(df$SEX))
c$SEX <- recode(c$SEX, Female = 0, Male = 1, Unknown = 2)
c$PRIMARY_RACE <- recode(c$PRIMARY_RACE, Asian = 0, Black = 1, `Native American` = 2,
    Undefined = 3, Unknown = 4, White = 5)
c$ETHNICITY <- recode(c$ETHNICITY, `Non-Spanish/non-Hispanic` = 0, `Spanish/Hispanic` = 1,
    Unknown = 2)
head(c)

s <- sample
s <- s[, c(-3, -6)]
s$SAMPLE_TYPE <- recode(s$SAMPLE_TYPE, Metastasis = 0, Other = 1, Primary = 2,
    Unspecified = 3)
```