Obtaining host-parasite occurence data from Genbank

Andrew Park

```
library(genbankr)
library(rentrez)
library(stringr)
library(tibble)
library(dplyr)
library(magrittr)
```

```
findHostsbySeq <- function(x) {
    # FN: finds hosts of parasite based on genbank sequences
    # ARGS: x=character parasite latin binom RTN:
    # df=tibble(para.name,host.name,uid,acc.num)
    df <- tibble(para.name = character(), host.name = character(),
        uid = character(), acc.num = character())
    para.uid <- entrez_search(db = "nucleotide", term = x)
    # for (i in para.uid$ids){
    for (j in 1:3) {
        i = para.uid$ids[j]
        print(j)
        para.acc <- entrez_fetch(db = "sequences", id = i, rettype = "acc")
        para.acc <- str_replace_all(para.acc, "\n", "")
        host <- ifelse(para.acc != "", ifelse(is.null(readGenBank(GBAccession(para.acc),
            ret.seq = F, partial = T, verbose = F)@sources@elementMetadata@listData$host),
            "", readGenBank(GBAccession(para.acc), ret.seq = F,
                partial = T, verbose = F)@sources@elementMetadata@listData$host),
            "")
        tmp <- tibble(para.name = x, host.name = host, uid = i,
            acc.num = para.acc)
        df <- bind_rows(df, tmp)
    }
    df %<>% filter(host.name != "" & host.name != x)
    return(df)
}
```

```
#q<-findHostsbySeq("Amblyomma ovale") %>% print
q<-findHostsbySeq("Yersinia pestis") %>% print
```

```
## [1] 1
## [1] 2
## [1] 3
## # A tibble: 2 x 4
##         para.name    host.name        uid        acc.num
##             <chr>        <chr>      <chr>          <chr>
## 1 Yersinia pestis Homo sapiens 1194615177  NZ_CP018770.2
## 2 Yersinia pestis Homo sapiens 1194538284 NCTN01000001.1
```