Oak Tree Alignments

Reading in data table from Email and associated Nucmer alignment files

```
# Table comes from email Sorel Sent October 15,2018. Manually added "|" for easy split.
# Refers to chromosome and  matching scaffold names.
# "You'll need to "reverse complement" the coordinates for the ones that say "Reversed", i.e. Subtract


scaffoldNamesManuallyEdited<-">chr1 Reversed: | Scq3eQI_240;HRSCAF=820
>chr2 Reversed: | Scq3eQI_1869;HRSCAF=2693
>chr3 Reversed: | Scq3eQI_2027;HRSCAF=3423
>chr4 | Scq3eQI_2026;HRSCAF=3421
>chr5 | Scq3eQI_2018;HRSCAF=3193
>chr6 | Scq3eQI_2028;HRSCAF=3424
>chr7 Reversed: | Scq3eQI_27
>chr8 | Scq3eQI_103;HRSCAF=384
>chr9 Reversed: | Scq3eQI_316;HRSCAF=953
>chr10 | Scq3eQI_174;HRSCAF=633
>chr11 Reversed: | Scq3eQI_1982;HRSCAF=3004
>chr12 | Scq3eQI_304;HRSCAF=933
"

#Looks like 2018 reference genome uses the whole line as a name.
#Not just the chromosome part. So copying and pasting the table again without editing
scaffoldNamesNotEdited<-">chr1 Reversed: Scq3eQI_240;HRSCAF=820
>chr2 Reversed: Scq3eQI_1869;HRSCAF=2693
>chr3 Reversed: Scq3eQI_2027;HRSCAF=3423
>chr4 Scq3eQI_2026;HRSCAF=3421
>chr5 Scq3eQI_2018;HRSCAF=3193
>chr6 Scq3eQI_2028;HRSCAF=3424
>chr7 Reversed: Scq3eQI_27
>chr8 Scq3eQI_103;HRSCAF=384
>chr9 Reversed: Scq3eQI_316;HRSCAF=953
>chr10 Scq3eQI_174;HRSCAF=633
>chr11 Reversed: Scq3eQI_1982;HRSCAF=3004
>chr12 Scq3eQI_304;HRSCAF=933
"

wholeScaffoldNames<-read_table(scaffoldNamesNotEdited, col_names = F) %>%
  dplyr::rename(QLobata2018Names=X1) %>%
  mutate(QLobata2018Names=str_replace(QLobata2018Names, ">", "")) %>%
  pull(QLobata2018Names)




# Table comes from email Sorel Sent October 15,2018. Manually added "Reversed:" if needed
# for easy join with scaffoldDf. Additionnaly added "|" For easy split
```

```
# Refers to chromosome and  matching scaffold names.
chromosomesAndTheirSize<-">chr1 Reversed:   55683757
>chr2 Reversed: 104436548
>chr3 Reversed: 74551540
>chr4   97794414
>chr5   89479036
>chr6   54021141
>chr7 Reversed: 49020089
>chr8   65456833
>chr9 Reversed: 55012092
>chr10  66423012
>chr11 Reversed:    57729484
>chr12  43132062
"

scaffoldDf<-read_table(scaffoldNamesManuallyEdited, col_names=F)

scaffoldDf<-scaffoldDf %>%
  separate(X1, into = c("Chromosome", "Scaffold"), sep="\\|") %>%
  mutate(Chromosome=str_squish(Chromosome), Scaffold=str_squish(Scaffold)) %>%
  mutate(Reversed=str_detect(Chromosome, "Reversed")) %>%
  mutate(QLobata2018Names=wholeScaffoldNames)


chromoseSizeDf<-read_delim(chromosomesAndTheirSize, col_names = F, delim="\t") %>%
  dplyr::rename(Chromosome=X1, Length=X2)



scaffoldDf<-left_join(scaffoldDf,chromoseSizeDf, by=c("Chromosome"="Chromosome"))

alignment<-read_delim("../data/oak_Qr.1coords", delim="\t", col_names = F)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   X2 = col_integer(),
##   X3 = col_integer(),
##   X4 = col_integer(),
##   X5 = col_integer(),
##   X6 = col_integer(),
##   X7 = col_double(),
##   X8 = col_integer(),
##   X9 = col_integer(),
##   X10 = col_double(),
##   X11 = col_double(),
##   X12 = col_character(),
##   X13 = col_character()
## )
```

```
#https://github.com/mummer4/mummer/blob/master/MANUAL.md under "show-coords" you can
#Find a more detailed explanation of columns
#Email Chain suggests this is the headers for the oak_Qr.1coords. Basically right.
#qBeg qEnd sBeg sEnd qAligned sAligned %ID qLength sLength qSomething sSomething query subject"
```

```
alignment<-alignment %>%
  dplyr::rename(QLobataStart=X1, QLobataEnd=X2, QRoburStart=X3, QRoburEnd=X4,
         QLobataLengthAligned=X5, QRoburLengthAligned=X6, PercentIdentity=X7,
     QLobataScaffoldLength=X8, QRoburScaffoldLength=X9,
     QLobataPercentCoverage=X10, QRoburPercentCoverage=X11,
     QLobata2017Names=X12, QRobur=X13 )
```

Using the scaffoldDf to translate the 2017 alignment to 2018 reference genome alignment.

Many scaffolds from 2017 don't have a specific name translation that is recorded in the scaffoldDf. I think this is because the names haven't changed and they are the same in 2018 as they were in 2017. Therefore, after merging alignment and scaffoldDf, if an alignement has QLobata2018Names set to NA, I assume the QLobata2017Name is the correct name.

```
alignment<-left_join(alignment, scaffoldDf, by=c("QLobata2017Names"="Scaffold")) %>%
  dplyr::rename(QLobata2018Chromosome=Chromosome)

#If No 2018 Translation, Use the 2017 Name. Remove semicolons in 2017 Names because they not there in 2

removedSemicolons<-str_replace(alignment$QLobata2017Names,
                               pattern=";", replacement = " ")
alignment$QLobata2018Names<-coalesce(alignment$QLobata2018Names,removedSemicolons)

#If no indication Reversed (NA), assume not Reversed. This should only affect alignments
# with 2017 scaffolds and not 2018 translations.

alignment$Reversed<-alignment$Reversed %>% replace_na(FALSE)
```

Now I think we have an alignment dataframe that is more detailed and complete.

Still need to figure out what to do with reversed scaffolds

```
alignment %>% print(n=5)
```

```
## # A tibble: 130,618 x 17
##   QLobataStart QLobataEnd QRoburStart QRoburEnd QLobataLengthAligned
##          <int>      <int>       <int>     <int>                <int>
## 1            1       3189     5226985   5230177                 3189
## 2         3240       8429     5230739   5235930                 5190
## 3         8476      13081     5236229   5240865                 4606
## 4        12534      13832     5241723   5243026                 1299
## 5        15577      17754     5249603   5251720                 2178
## # ... with 1.306e+05 more rows, and 12 more variables:
## #   QRoburLengthAligned <int>, PercentIdentity <dbl>,
## #   QLobataScaffoldLength <int>, QRoburScaffoldLength <int>,
## #   QLobataPercentCoverage <dbl>, QRoburPercentCoverage <dbl>,
```

```
## #   QLobata2017Names <chr>, QRobur <chr>, QLobata2018Chromosome <chr>,
## #   Reversed <lgl>, QLobata2018Names <chr>, Length <int>
```

Now I need to get it to work on "reversed" chromosomes.

Create a new 2018 Start (QLobata2018Start) and End (QLobata2018End).

For all 2018 Alignments that are on a chromsome that has been reversed, got the Length of the entire chromosome, then subtracted the alignment start and then added 1. Did the same for the alignment ends.

```
alignment<-alignment %>%
  mutate(QLobata2018Start=Length-QLobataStart+1) %>%
  mutate(QLobata2018End=Length-QLobataEnd+1)



#If NA for QLobata2018Start or QLobata2018End its because there's no Length because its
# old name is used/not reversed/no length available
alignment$QLobata2018Start<-coalesce(alignment$QLobata2018Start,
                                     parse_double(alignment$QLobataStart))

alignment$QLobata2018End<-coalesce(alignment$QLobata2018End,
                                   parse_double(alignment$QLobataEnd))
```

Sanity check to make sure alignments look reasonable.

1. Check when Start > End
2. Check when "Reversed"
3. Check when "Reversed" has Start > End
4. Check normal alignment
5. Check when no new name for 2018
6. Check 5 Random Ones

```
qLobata2018Reference<-"../data/oak_1Aug2018.fa"
qRoburReference<-"../data/Qrob_PM1N.fa"
qLobata2017Reference<-"../data/oak_14Aug2017.fa"


qLobata2018Reference<-readDNAStringSet(qLobata2018Reference)
qLobata2017Reference<-readDNAStringSet(qLobata2017Reference)
qRoburReference<-readDNAStringSet(qRoburReference)


getSequenceFromReference<-function(referenceGenome, scaffoldName, start, end){
  #If start is bigger than end, I think aligns to other strand. Change start and end and
  # Reverse complement
  if (start > end){
    temp<-start
    start<-end
```

```
    end<-temp
    dna<-subseq(referenceGenome[scaffoldName], start=start, end=end)
    reversedDNA<-reverseComplement(dna)
    return(reversedDNA)
  } else {
    dna<-subseq(referenceGenome[scaffoldName], start=start, end=end)
    return(dna)
  }


}

getSequenceFromReference(referenceGenome = qLobata2018Reference,
                         scaffoldName = alignment$QLobata2018Names[7],
                         start= alignment$QLobataStart[7],
                         end=alignment$QLobataEnd[7])
```

```
##   A DNAStringSet instance of length 1
##     width seq                                               names
## [1]  1312 AAAATTTATATATGATTCTAA...TAACTTACTTTTTCAACTTT Scq3eQI_100 HRSCA...
```

```
getSequenceFromReference(referenceGenome = qRoburReference,
                         scaffoldName = alignment$QRobur[7],
                         start= alignment$QRoburStart[7],
                         end=alignment$QRoburEnd[7])
```

```
##   A DNAStringSet instance of length 1
##     width seq                                               names
## [1]  1305 AAAATTAATCTATGATCGAAA...TAACTTACCTTTTCAACTTT Qrob_Chr09
```