R-plotting & Useful scripts

Other useful tools

```
$ module load cdbfasta
$ cdbfasta proteins.fa
$ cdbfasta reads.fastq
$ echo read_1_name | cdbyank reads.fastq.cidx > my_read1_sequence
```

Write a tool to print out sequence lengths

```
#!/usr/bin/env python3
import sys
from Bio import SeqIO
from Bio.Seq import Seq

# seqfile
if len(sys.argv) != 2:
     print("argument should be\n","sequence_length.py FILE.fasta")
     exit()

filename = sys.argv[1]
print("%s\t%s"%("Name","Length"))
for seq_record in SeqIO.parse( filename , "fasta"):
    print("%s\t%d"%(seq_record.id,len(seq_record)))
```

Generate table of sequence sizes

```
$ curl -L https://goo.gl/2gAaBt > transposon.fa.gz #SHORT URL
# ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequence/transcript

$ curl -L https://goo.gl/eUGT67 > mRNA.fa.gz # SHORT URL
# ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequence/transcript

$ gunzip *.gz
$ ./sequence_length.py transposon.fa >  transposon.lengths
$ ./sequence_length.py mRNA.fa > mRNA.lengths
```

R basics

- Cmdline R

  ```
  $ R
  or to run an existing script
  $ Rscript script.R
  ```

R basics - read in data

```
mRNA = read.csv("mRNA.lengths",sep="\t",header=T)
head(mRNA)
summary(mRNA)
mRNA2 = read.table("mRNA.lengths",sep="\t",header=T)
head(mRNA2)
mRNA3 = read.csv("mRNA.lengths",sep="\t",header=T,row.names=1)
head(mRNA3)
summary(mRNA3)
```

Summary stats and Histograms

```
mRNA = read.csv("mRNA.lengths",sep="\t",header=T,row.names=1)
TE = read.csv("transposon.lengths",sep="\t",header=T,row.names=1)
summary(mRNA)
summary(TE)
pdf("lengths_hist.pdf")
hist(TE$Length)
hist(TE$Length,100,main="TE length distribution")
hist(mRNA$Length,100,main="mRNA length distribution")

short = subset(mRNA,mRNA$Length < 10000)
summary(short)
hist(short$Length,100,main="shorter mRNA length distribution")
long = subset(mRNA,mRNA$Length > 50000)
summary(long)
```

Boxplots

```
mRNA = read.csv("mRNA.lengths",sep="\t",header=T,row.names=1)
TE = read.csv("transposon.lengths",sep="\t",header=T,row.names=1)
pdf("lengths_boxplot.pdf")

nm = c("mRNA","TE")
boxplot(mRNA$Length,TE$Length,main="Boxplot of seq type lengths",
names=c("mRNA","TE"))

boxplot(mRNA$Length,TE$Length,main="Boxplot of seq type lengths",
col=c("red","blue"),names=nm)

boxplot(subset(mRNA$Length,mRNA$Length < 10000),TE$Length,
main="Boxplot of shorter seqtype lengths",names=nm,
col=c("red","blue"),las=2)
```

Plot X-Y

Data set project on genome size: https://github.com/1KFG/genome_stats

https://github.com/1KFG/genome_stats/blob/master/fungi_genome_stats.csv