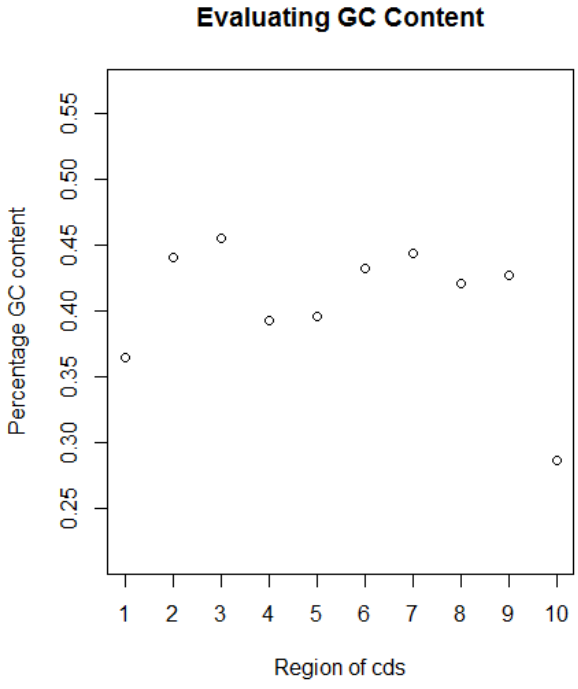
```
#GENOMICS with R: Let's use seq data to design a better primer!

#seqinR: A package used to retrieve and manipulate genomic data

#first one must download the package
install.packages("seqinr")

#one may import genomic data directly from online databases such as genbank
#where first one must specify the online database of interest:
choosebank("genbank") #specifies all data queries to genbank

#now, let's say we're interested in genes related to tomato growth
#we can use the query() function to import the names and accessions
#of ALL genbank full coding sequences for tomatoes
query("completetomatoCDS", "sp=solanum lycopersicum AND t=cds AND NOT
      k=partial")

#this command takes about half a minute to execute;
#there are around 5000 full cds to gather as of May 2014.

#How many are there now? Use the "number of elements" tool:
completetomatoCDS$nelem

#Of course, now we need to extract the coding sequence (cds)
#of interest to our analysis.
#Using the query function, the first argument names a variable,
#while the second is the genbank (etc) accession number
#(currently in R memory)
query("phot1", "N=EF063359")
#extracts the cds of phot1, a gene that aids in light sensory

#Let's make sure it's there; how long is phot1's cds?
phot1.length<-getLength(phot1)
phot1.length #prints length of sequence without printing a wall of letters

#for more information on the phot1 sequence,
#we can use the getAnnot() function,
#which prints the familiar genbank accessesion page,
#including the translated seq at bottom
getAnnot(phot1)


#Now for the cool part:

#Principal components analysis:
#GC content is an imporant feature for primer design
#It is reported as a percentage that corresponds to
#the percentage of a sequence of a DNA that is constituted by
#either guanine or cytosine, which differ from adenine and thimine
#by the number of hydrogen bonds in their secondary structure

phot1.seq<-getSequence(phot1)
phot1.char<-as.character(phot1.seq)
phot1.GC<-GC(s2c(phot1.char)) #uses the string-to-character function
phot1.GC
```

```
#However, knowing the gene's GC-content is only informative
#in the sense that one anticipates a GC-content between 45-50%
#within coding regions of genes in eukaryotes.


#For primer design, 40-60% GC-content is preferred
#in order to minimize errors in replication that are more
#common in AT-rich regions because of AT's lower temperature
#of denaturation

#Therefore, we will identify loci within the coding sequence that
#demonstrates GC-content between 0.5-0.6 to find ideal loci for
#primer design.

#To start, we'll break up the coding region into smaller
#fragments for evaluation

frag.size<-phot1.length*.1
phot1.frag1<-getFrag(phot1.seq,1,frag.size)
phot1.frag1.char<-as.character(phot1.frag1)
phot1.frag1.GC<-GC(s2c(phot1.frag1.char))
phot1.frag1.GC

phot1.frag2<-getFrag(phot1.seq,frag.size+1,2*frag.size)
phot1.frag2.char<-as.character(phot1.frag2)
phot1.frag2.GC<-GC(s2c(phot1.frag2.char))
phot1.frag2.GC

phot1.frag3<-getFrag(phot1.seq,2*frag.size+1,3*frag.size)
phot1.frag3.char<-as.character(phot1.frag3)
phot1.frag3.GC<-GC(s2c(phot1.frag3.char))
phot1.frag3.GC

phot1.frag4<-getFrag(phot1.seq,3*frag.size+1,4*frag.size)
phot1.frag4.char<-as.character(phot1.frag4)
phot1.frag4.GC<-GC(s2c(phot1.frag4.char))
phot1.frag4.GC

phot1.frag5<-getFrag(phot1.seq,4*frag.size+1,5*frag.size)
phot1.frag5.char<-as.character(phot1.frag5)
phot1.frag5.GC<-GC(s2c(phot1.frag5.char))
phot1.frag5.GC

phot1.frag6<-getFrag(phot1.seq,5*frag.size+1,6*frag.size)
phot1.frag6.char<-as.character(phot1.frag6)
phot1.frag6.GC<-GC(s2c(phot1.frag6.char))
phot1.frag6.GC

phot1.frag7<-getFrag(phot1.seq,6*frag.size+1,7*frag.size)
phot1.frag7.char<-as.character(phot1.frag7)
phot1.frag7.GC<-GC(s2c(phot1.frag7.char))
phot1.frag7.GC

phot1.frag8<-getFrag(phot1.seq,7*frag.size+1,8*frag.size)
phot1.frag8.char<-as.character(phot1.frag8)
phot1.frag8.GC<-GC(s2c(phot1.frag8.char))
phot1.frag8.GC
```

```
phot1.frag9<-getFrag(phot1.seq,8*frag.size+1,9*frag.size)
phot1.frag9.char<-as.character(phot1.frag9)
phot1.frag9.GC<-GC(s2c(phot1.frag9.char))
phot1.frag9.GC

phot1.frag10<-getFrag(phot1.seq,9*frag.size+1,10*frag.size)
phot1.frag10.char<-as.character(phot1.frag10)
phot1.frag10.GC<-GC(s2c(phot1.frag10.char))
phot1.frag10.GC

Fragmented.GC<-c(phot1.frag1.GC, phot1.frag2.GC, phot1.frag3.GC,
                 phot1.frag4.GC, phot1.frag5.GC, phot1.frag6.GC,
                 phot1.frag7.GC, phot1.frag8.GC, phot1.frag9.GC,
                 phot1.frag10.GC)


#Now we plot the fragments according to their GC-content with the intent
#to decide which areas are most satured with guanine and cytosine
X<-seq(1,10)
Y<-Fragmented.GC
GC.range.expansion<-c(min(Fragmented.GC)*.75, max(Fragmented.GC)*1.25)
plot(X,Y,type="p",ylim=GC.range.expansion, axes=FALSE, main="Evaluating GC
Content",
     xlab="Region of cds",ylab="Percentage GC content")
box()
axis(1,at=1:10)
axis(2,at=seq(0.2,0.8,0.05))
```



#References:
#Wendel, J. F. (2012). The Variation of Base Composition in Plant Genomes. Plant genome diversity. Vienna: Springer.

#SeqinR 2.07 Release literature: http://seqinr.r-forge.r-project.org/seqinr_2_0-7.pdf

#Wang, L., Stein, L., & Ware, D. (2014). The relationships among GC content, nucleosome occupancy, and exon size. arXiv preprint arXiv:1404.2487.