Predicting number of loci

An important question in planning a RAD-seq experiment is how many loci to expect from your organism's genome.

If you have a reference genome for your organism, (or a closely related one) it is easy to get a prediction.

Although you should take into consideration how much of the actual genome is represented in the reference.

From John Stanton-Geddes and seqAnswers:

```
Predicting RAD loci for ddRAD

cat pbar_scaffolds_v03.fasta | tr -d "\n" | grep -o -E "CATG.(264,336)AATT" | wc -l
```

Where pbar_scaffolds_v03.fasta is your reference genome

In "CATG.(264,336)AATT" replace CATG with the forward restriction enzyme recognition sequence and AATT with the reverse restriction enzyme recognition sequence.

This example would work if your forward restriction enzyme is NlaIII, your reverse restriction enzyme is MluCI, and your insert size is between 264 and 336 base pairs.

Does this really work?

This fake reference has 4 RAD fragments, 3 with 100-300 bp insert sizes and 1 more with a 350 bp insert size for NlaIII and MluCI.

Does the method return correct answer?

```
solution

cat test_reference.fasta | tr -d "\n" | grep -o -E "CATG.(100,300)AATT" | wc -l

cat test_reference.fasta | tr -d "\n" | grep -o -E "CATG.(100,350)AATT" | wc -l
```

Exercise:

Imagine you wanted to do ddRAD on the coral Acropora hyacinthus.

There is no reference genome for this coral, but there is one for a closely related species Acropora digitifera.

Your lab has the restriction enzymes Sphl and MspI on hand and you want to use those as your forward and reverse cutter respectively.

How would you use this to get an idea of how many loci you are likely to find.

1. Search NCBI for a reference genome
2. Download it
3. Search for your specific restriction search parameters

Solution:
#start an idev session if you have not
idev

#make a directory to do the analysis in
cds
mkdir rad_intro
cd rad_intro
mkdir in_silico_digest
cd in_silico_digest

#download the reference from NCBI
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/222/465/GCF_000222465.1_Adig_1.1/GCF_000222465.1_Adig_1.1_genomic.fna.gz

decompress it
gunzip GCF_000222465.1_Adig_1.1_genomic.fna.gz

#check the command on some different max insert sizes
cat GCF_000222465.1_Adig_1.1_genomic.fna | tr -d "\n" | grep -o -E "GCATGC.{300,400}CCGG" | wc -l
cat GCF_000222465.1_Adig_1.1_genomic.fna | tr -d "\n" | grep -o -E "GCATGC.{300,500}CCGG" | wc -l
cat GCF_000222465.1_Adig_1.1_genomic.fna | tr -d "\n" | grep -o -E "GCATGC.{300,600}CCGG" | wc -l

#use a loop to run command for a series of maximum insert sizes
seq 400 100 1000 > maxInserts.txt
echo -e "insert_range\tNsites" > prediction_results.txt
while read max; do nSites=$(cat GCF_000222465.1_Adig_1.1_genomic.fna | tr -d "\n" | grep -o -E "GCATGC.{300,${max}}CCGG" | wc -l); echo -e "$300-$(max)\t${nSites}" >> prediction_results.txt; done < maxInserts.txt

#look at the results
cat prediction_results.txt

#These values are somewhat low
#Does changing to a four-nucleotide restriction enzyme help?
#for NlaIII
cat GCF_000222465.1_Adig_1.1_genomic.fna | tr -d "\n" | grep -o -E "CATG.{300,400}CCGG" | wc -l

