

Sequencing Informatics (Base calling, tracking, quality, vector removal)

Start your console, make sure you are in your home directory, and make a directory for this class:

```
cd ~
mkdir phreddemo
```

Ok, we are going to process some LICOR, ABI3700 and ABI3730 data, but first we need some new directories:

```
cd phreddemo
mkdir chromat_dir
mkdir phd_dir
mkdir edit_dir
ls
```

The raw sequence data is in a shared directory, please look at its contents:

```
ls /bioware/data/mcarthur/chromat_dir
ls /bioware/data/mcarthur/chromat_dir | wc -l
```

View an ABI trace file (ie. open a file in above directory with traceviewer):

```
TraceViewer
```

We are going to re-call the bases and assign quality scores using the phred program. To learn about phred, try both of the following:

```
clear
phred -h | less
phred -doc | less
```

Now we are going to run phred on the data:

```
phred -id /bioware/data/mcarthur/chromat_dir -pd phd_dir -cd chromat_dir
```

New trace files and phd files are created:

```
ls chromat_dir
ls phd_dir
```

View a new trace file (ie. in ~/phreddemo/chromat_dir):

```
TraceViewer
```

View a phd file:

```
clear
ls phd_dir
less phd_dir/filename
```

Obtain fasta format data:

```
phd2fasta -id phd_dir -os giardia.fasta -oq giardia.fasta.screen.qual
ls
less giardia.fasta
less giardia.fasta.screen.qual
```

I have some perl scripts to examine sequence data. Try them without flags:

```
summarizfasta
summarizequal
```

Examine base frequencies:

```
summarizefasta -freq giardia.fasta
```

Run the various analyses, but save the tab-delimited output:

```
summarizefasta -GC giardia.fasta > gc.tab  
summarizefasta -length giardia.fasta > length.tab  
summarizequal -curve giardia.fasta.screen.qual > qualcurve.tab  
summarizequal -histogram giardia.fasta.screen.qual > qualhist.tab
```

View the results graphically:

```
histogram gc.tab  
histogram length.tab  
plot qualcurve.tab  
histogram qualhist.tab
```

Another set of statistics:

```
seqqual giardia.fasta.screen.qual 10  
seqqual giardia.fasta.screen.qual 20  
seqqual giardia.fasta.screen.qual 30
```

What is the read length distribution of good data?

```
phred -id /bioware/data/mcarthur/chromat_dir/ -trim_fasta -trim_alt "" -sa trimmed.fasta  
summarizefasta -length trimmed.fasta > trimmed.tab  
histogram trimmed.tab
```

Before we can use this data, we need to remove vector sequences. We need a reference file of vector sequences:

```
cp /bioware/data/mcarthur/vector.fasta vector.fasta
less vector.fasta
grep ">" vector.fasta | wc -l
```

Remove the vector:

```
cross_match giardia.fasta vector.fasta -minmatch 12 -minscore 20 -screen
ls
less giardia.fasta.screen
```

Run a quick and dirty sequence assembly:

```
phrap giardia.fasta.screen -new_ace > giardia.phrap.log
mv *.ace edit_dir/
cd edit_dir
consed_linux
```

By the way, would could have done all of this automatically:

```
cd ~/phreddemo/edit_dir
phredPhrap
ls
```