Botany 2015
Introduction to Next-Generation Sequencing Workshop
Practical Exercises

The following exercises are intended to help you familiarize yourself with Illumina data and the basic kinds of analyses that are performed using short read data, including read mapping and reference-guided assembly. For the workshop you will use unpublished Illumina data we have generated for *Asclepias euphorbiifolia*. You will be performing analyses appropriate for data sets produced through target enrichment, which allows both genome skimming (see Straub et al. 2011 and Straub et al. 2012), in which the high copy fraction of the genome (e.g., nrDNA repeats, plastome, mitochondrial genome) can be well-characterized (i.e. easily assembled), and assembly of low-copy nuclear targets that have been enriched relative to untargeted regions (see Weitemier et al. 2014).

Exercise 1 – Raw Illumina Data

Illumina sequence data are typically obtained as compressed (gzip) fastq files. The files will usually be separated by index and split into subsets of 4 million reads. The paired read files are labelled R1 and R2. Some data analysis programs accept fastq files and use the quality information that they contain, while other programs will require the fastq files to be converted to fasta format.

A. The Anatomy of a fastq File

Use any text editor to open the example fastq file (example_fastq.fq). Fastq formatted files returned from the Illumina pipeline contain 4 lines of information per record. Look at the example file to make sure you understand the information contained in each line, especially the Illumina header in line 1.

Line 1: This line starts with '@' and includes text to identify the sequence. The identifiers for Illumina reads contain several pieces of information separated by colons. The information for reads that have been analyzed using Illumina-provided software will contain the following information: @instrument name:run id:flow cell id:lane number:tile number:x-coordinate for this cluster on tile:y-coordinate for this cluster on tile number of a pair(1 or 2):flag for whether this read has passed the Illumina chastity filter (N=pass, Y=fail):control bits:index (barcode) number or sequence. Some users will filter reads based on the Illumina chastity filter, but most do not.

Line2: This line contains the sequence information.

Line3: This line starts with '+' and may or may not include more information, such as a repetition of the information included in line 1.

Line4: This line contains the quality scores for each base of the sequence read in line 2.

Since the 2012 release of Casava v. 1.8, Illumina uses standard Sanger quality scores (Phred+33) encoded with ASCII characters.

$$Q_{\text{Sanger}} = -10 \log_{10} p$$

where p is the probability that the base call is incorrect. Thus a quality score of 20 corresponds to p=0.01.

ASCII characters are used to represent each Phred quality score:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ

00000000001111111111222222222233333333344

01234567890123456789012345678901234567890
```

Using this code '>' equals a Phred score of 29, '?' = 30, '@' = 31, 'A' = 32, etc.

Find examples of sequences with high and low overall quality in example_fastq.fq.

Many users filter or trim reads based on their quality scores prior to any other analyses. See http://www.usadellab.org/cms/index.php?page=trimmomatic for a versatile script for quality trimming. Note that if you have legacy Illumina data (pre-v. 1.8), be aware that quality scores may be in one of the older Illumina formats (See http://en.wikipedia.org/wiki/FASTQ_format), and will need to be converted prior to running scripts that expect v. 1.8 values.

B. Quality Score Visualization in Geneious

Click on the Asclepias_euphorbiifolia_Hyb-Seq_reads_R1_001.fastq file in your NGS_Workshop directory in Geneious. The bases of the reads will automatically be colored based on their quality scores. Light blue bases have the highest quality scores, while dark blue bases have the lowest quality scores.

1. To determine how many reads did not pass the chastity filter, search for ":Y:" in the sequence names. To do this, choose "Edit" from the Geneious menu, then select "Find in Document…," then enter :Y: in the search box before clicking "Find All."

   Why is the same number obtained for read 1 and read 2?

The sequences in the read 1 and read 2 files represent paired-end data. A fragment of DNA in an Illumina library attaches to the sequencing lane, where additional copies are synthesized. This group of DNA copies is called a cluster, and it occupies a physical XY-coordinate on the sequencing lane. Early in the sequencing process the Illumina pipeline determines whether each cluster is trustworthy (e.g. does it provide a clear signal). Those clusters that fail this filtering step are labeled with a 'Y'. Each cluster can be sequenced from either end, with one end being read 1, and the other read 2. Since both reads originate from the same cluster, and that cluster has either passed or failed the chastity filter, the number of failing reads in the two read files should be the same.

C. Set up Paired Reads

Illumina libraries can consist of DNA fragments that are longer than the length of the reads that can reliably be sequenced. However, Illumina technology allows for each end of a large DNA fragment to be sequenced separately, and the relationship between those two reads can be saved. This is known as paired-end sequencing. The DNA fragments suitable for paired-end sequencing are on the order of several hundred bases long. Fragments of much larger length can also be sequenced from each end, but are usually referred to with different names because the library preparation is different (e.g., mate pairs). Geneious can make use of the powerful information that two reads were derived from the same fragment of DNA and are therefore expected to be a certain distance apart.

1. Select the two files containing the Illumina reads, Asclepias_euphorbiifolia_Hyb-Seq_reads_R1_001.fastq and Asclepias_euphorbiifolia_Hyb-Seq_reads_R2_001.fastq.

2. Under the "Sequence" menu select "Set paired reads…" In the dialog box that appears, select the boxes "Pairs of sequence lists" and "Forward/Reverse (Illumina paired end)." Set the expected distance to 200 and click "OK."

Exercise 2 - Short read mapping

Short read mapping involves aligning reads to a reference sequence based on a set of criteria. Some reads will map to a single location, but others, such as those that originate from repetitive areas of a genome, may map to multiple locations and are called multi-maps or multi-reads. Read mapping is useful for calling single nucleotide polymorphisms (SNPs) between the sequenced sample and a reference sequence, as well as for determining sequencing depth. BWA (Li and Durbin, 2009) and Bowtie (Langmead et al., 2009), both based on Burrows-Wheeler transform, are two popular short read mapping programs. Geneious also has a read mapping function (seed and expand). See section 10.2.1 of the Geneious manual for a detailed description of the read mapping algorithm. In this exercise, you will map reads of *Asclepias euphorbiifolia* to both an annotated *Asclepias syriaca* reference sequence and an annotated *Matelea biflora* reference sequence in Geneious. We chose *Matelea* for the second example because it is placed in a different subtribe of Apocynaceae from *Asclepias*, the Gonolobinae.

A. The reference sequence

1. View the provided *Asclepias syriaca* reference plastome in Geneious by choosing the name of the file (NC_022432) in the document table pane. Take a few minutes to familiarize yourself with the types of annotations that will be present for a plastome downloaded from GenBank and how these annotations are displayed in Geneious. Zoom in to the sequence level for a better view.

B. Map the reads

1. For this exercise, use the file of paired *A. euphorbiifolia* reads that you obtained earlier. To map the reads highlight both the read file and the *Asclepias syriaca* reference file (NC_022432) by clicking on each while holding down the control key.

2. Next, select the "Align/Assemble" button from the Geneious menu and choose "Map to Reference…"

3. You should see the NC_022432 file in the "Reference Sequence" box. For "Sensitivity" choose "Low Sensitivity/Fastest" from the drop down menu and for "Fine Tuning" choose "None (fast/read mapping)." Check the boxes next to "Save assembly report" "Save in sub-folder" and "Save contigs." Click the "OK" button at the bottom of the window to begin the read mapping analysis.

4. Repeat steps 1-3 using the *Matelea* reference (KF539850).

C. Estimate the plastid DNA content of the *A. euphorbiifolia* Illumina library

1. When the read mapping is complete, in the Geneious document table pane click on the assembly report for the assembly that used *Asclepias syriaca* as the reference.

2. From the assembly report you will estimate the plastid DNA content of the *A. euphorbiifolia* library and the sequencing depth of the plastome for this individual. Assuming the sequence of the *A. euphorbiifolia* plastome is unknown, we can make estimates of each of these metrics using the number of reads that map to the reference plastome.

    a. Chloroplast content calculation

    (_119,364_____ reads mapped to cp / __1,534,468____ total reads) × 100 = ___7.78____ % cp

    b. Sequencing depth calculation – For this calculation you will also need the read length, which can be found in the document table pane under min or max sequence length, and the genome size, in this case the plastome size for the *Asclepias syriaca* reference (158,719 bp).

(___119,364_____ reads mapped × ___76___ bp read length)/___158,719_____ bp genome size = __57.2____× sequencing depth

D. Examine the read mapping results

1. Within the new folder named "Asclepias_euphorbiifolia_Hyb-Seq_reads_R_001 assembled to NC_022432" click on the document named "Contig" to open a graphical view of the read mapping results. This first view will show the consensus sequence

of the mapped reads, the coverage of each part of the plastome based on read mapping depth, and the annotated reference sequence.

2. To view some more basic information about the analysis choose the Statistics panel by clicking the tab with a % sign at the right hand part of the window.

   How well does your sequencing depth estimate from Part C match up with the Geneious base by base coverage estimate?

   The two should only differ very slightly, if at all.

3. Next, zoom in using the magnifying glass button above the Statistics panel to get a better look at the read mapping results. If you zoom in far enough you can see the actual sequence of each read. Be sure that the "Vertically compress contig" box is checked in the "Layout" section of the advanced tab (gear symbol) for the best view. Take a few minutes to explore the results, perhaps by finding your favorite plastid gene or intergenic region using the reference sequence annotations.

4. Now explore the assembly that used the *Matelea* sequence as a reference. Which parts of the plastome assembled well and which did not using read mapping? Do you notice differences between this assembly and the one for which *Asclepias syriaca* served as the reference?

   Areas with low coverage can be highlighted using options under the "Graphs" tab on the right side (bar graph icon). Areas that commonly assemble poorly using read mapping include repeats and indels between the reference and the sample. References that are more divergent from the sample will also have worse read mapping. Preferences can be set regarding how similar a read must be to the reference in order for it to be mapped. More lenient parameters can help with mapping to a divergent reference, but this also increases the risk of incorrectly mapping a read. The more similar reference from the same genus performs better than the more divergent reference sequence.

5. While still viewing the *Matelea* reference assembly, click on the display tab (screen icon) to the right side of the Statistics panel. Make sure that the box next to "Highlighting" is checked and that the next two boxes form the phrase "Disagreements to Reference." While exploring the mapped reads, you will notice

that differences in the assembled reads and reference are now highlighted in each read. How can you distinguish SNPs from sequencing errors?

Sequencing errors are disagreements to the reference that only appear in one or a few reads. SNPs are disagreements to the reference that appear in nearly all of the mapped reads.

Other highlighted features you may encounter are areas of read misalignment. Also note that apparent sequencing errors could have a biological basis and originate from plastid sequences that have been transferred to the mitochondrial or nuclear genomes. These will have been sequenced at much lower depth than the plastid genome, but still retain high enough similarity to be mapped.

6. Geneious can aid in a comparison of the *Asclepias* and *Matelea* sequences in order to identify SNPs. To find and annotate SNPs in coding regions, choose "Annotate & Predict" from the top menu bar and then select "Find Variations/SNPs." Set the "Minimum Coverage" to 25 and "Minimum Variant Frequency" to 0.8. Next, choose "Only in CDS" from the dropdown menu next to "Find Polymorphisms." Be sure that the boxes next to "Analyze effect of polymorphisms on translations" and "Calculate Variant P-values" are also selected. Then click "OK."

7. Explore the results by looking at the new track "Variations" that has been added to the "Annotations and Tracks" menu (yellow arrow tab to the right of the Contig View). Use the arrows to jump between SNPs. If you mouse over the orange SNP track markers in the Contig view you will be able to see detailed information about each SNP, such as whether it causes a synonymous or nonsynonymous amino acid change. Do you see any SNPs in the coding regions of the *A. euphorbiifolia* plastome that cause an amino acid substitution or other change in the protein (e.g., truncation) when compared to *Matelea*?

Various examples can be identified. For example A -> E in *ycf3* at 49,636 bp.

Exercise 3 – Reference-guided assembly

In reference-guided methods, a reference sequence is employed to aid genome assembly and this approach is especially amenable to the assembly of plant plastomes due

to their tractable size and conserved gene content and organization. A sequence from a closely related species is the best reference choice, but the majority of the plastome can be assembled using a reference in the same order as your species of interest (Straub et al., 2012). Some of the common assembly mistakes encountered in reference-guided assembly include incorrectly assembled or missing insertions and deletions relative to the reference and missed rearrangements relative to the reference sequence's orientation.

The Geneious read mapper has an iteration function to improve read mapping results and is thus a type of reference-guided assembler that goes beyond simple read mapping. You will use it as such for the purposes of this workshop.

1. Select the file of paired *A. euphorbiifolia* reads and the *Matelea* plastome reference and choose "Map to reference…" from the "Align/Assemble" menu just as you did in the read mapping exercise. Due to the limited time we have for the workshop, for "Fine Tuning" choose "Iterate up to 5 times" from the drop down menu. You may want to append the word "iterate" to the Assembly Name to give your sub-folder an informative name rather than a number. Then click "OK."

2. You will be able to see improvements in the numbers of reads matched as Geneious works through read mapping iterations. When mapping is complete, view the Assembly Report. How many additional reads were mapped after five iterations?

   104220-96314 = 7906 additional reads mapped

3. Now take a few minutes to explore the improved assembly. Which areas of the assembly have improved the most compared to the single pass you did in the read mapping exercise?

   intergenic regions

   What characteristics of these sequences are likely responsible for the differences?

   In this scenario most of the improvement comes in regions with small indels between the reference and sample. While this succeeds in mapping more reads, many of these regions may need special attention to ensure that Geneious has properly resolved the indel. Iterative assemblers that include a de novo assembly

component (such as Alignreads) can also make improvements by extending contigs into regions that are more divergent from the reference or have been rearranged.

4. The use of a divergent reference, the presence of indels, as well as features of individual libraries, such as low plastid to mtDNA ratios, can lead to mistakes in reference-guided assembly. A good strategy to ameliorate some of these issues is to combine the results of your reference-guided assembly with those from a de novo assembly, which will aid you in incorporating large indels or rearrangements into your final plastome sequence.

5. We will not be performing de novo assembly in today's workshop due to time constraints and the variability in the available amount of RAM on different participants' laptops. Open the provided de novo assembly of the *A. euphorbiifolia* reads. Click on "Contig 1." Note the sequence length is 129,446 bp, or essentially the whole plastome sequence with only one copy of the IR represented. Briefly explore the assembly and compare to the reference-guided assembly results.

6. If you would like to try this analysis on your own with the *A. euphorbiifolia* plastid genome do the following: Choose the file of paired reads you used for the read mapping exercise. Then choose "Align/Assemble" followed by "De Novo Assemble…" The options to "Save assembly report," "Save in sub-folder," "Save contigs," and "Save consensus sequences" will be helpful. Depending on the specifics of your computer, this may take 15-45 minutes or longer.

Exercise 4 – Hyb-Seq Exon Assembly

The Hyb-Seq process enriches a genomic library for targets of interest. In this example a list of targets representing 3385 exons from 925 genes were enriched in several samples of *Asclepias*, including *A. euphorbiifolia*. This exercise demonstrates the assembly of a consensus sequence for each exon. We will use the iterative read mapping capability of Geneious to map the reads from the sample to a reference sequence (the probe sequences used in the enrichment). This allows the assembled consensus sequence to extend beyond the edges of the reference, referred to as the "splash zone."

Examine the document named Asclepias_syriaca_single_copy_gene_exons:

What is the total length of the targeted exons? The "Statistics" tab (percent icon) on the right side of the window will indicate the number of sequences and their average length.

3385 sequences * 474 bp average length = 1,604,490 bp targeted length

If the *Asclepias syriaca* nuclear genome is assumed to be 420 Mbp (1 Mbp = 1 million bases), what proportion of the nuclear genome is being targeted by the probes?

1604490 bp targeted / 420000000 bp total = 0.4% of the nuclear genome targeted.

1. Select the file of paired *A. euphorbiifolia* reads and the list of exons targeted in this Hyb-Seq run and choose "Map to Reference…" from the "Align/Assemble" menu. We will perform a reference-guided assembly again, so for "Fine Tuning" choose "Iterate up to 5 times" from the drop down menu. Under "Reference Sequence" it should say "All 3,385 sequences in Asclepias syriaca," and the boxes for "Save assembly report," "Save in sub-folder," and "Save contigs" should be checked. This time we will also check the box for "Save consensus." Click OK.

2. When mapping is complete, view the Assembly Report.

   How many additional reads were mapped between the initial and best iteration?

   710,682-639,976 = 70,706 additional reads

   What proportion of reads mapped to the target exons?

   710,682 / 1,534,468 = 0.463

   By how much did the target enrichment process increase the proportion of targeted reads over an unenriched library?
   The targeted regions represent 0.4% of the nuclear genome, but make up 46.3% of the sequenced reads represents a nearly 116 fold increase. 46.3% / 0.4% = 115.75 (The true increase is even greater since the targeted regions are 0.4% of just the nuclear genome. The targeted regions will make up an even smaller share of the extracted DNA due to the presence of chloroplast and mitochondrial sequences.)

   How many exons (contigs) were assembled? All 3385 contigs were assembled.

3. Take a few minutes to examine some of the contigs produced by the assembly. Extra columns can be added to the Document Table, and clicking the header for a column will sort the documents based on that statistic. Click the icon on the far right side of the Document Table, just above the scroll bar (it looks like a table with an arrow on it). This will provide a list of columns that can be added to the Document Table. Check the column for Mean Coverage. Try sorting the contigs in the Document Table in different ways by clicking on the headers along the top of the Document Table.

   How can you distinguish heterozygous sites from SNPs and sequencing errors? Find examples of all three.

   SNPs and sequencing errors can be distinguished as discussed above. Heterozygous sites will be those where about half of the reads contain a variant.

   Scroll though several contigs. How well did this assembly extend contigs beyond the original reference?

   On either side of the reference sequence there is additional contig sequence i.e. the splash zone.

   What features of the wet-lab library preparation and Illumina sequencing options are most important to the size of the assembled "splash zone?" (Hint: both of these relate to specific parameters you input into Geneious during this workshop.)

   The size of the fragments in the Illumina library (created though shearing the DNA and later size selection) is directly related to the size of the splash zone.

4. Find and select the document named "Consensus Sequences." This document contains all of the consensus sequences generated by each of the contigs in the assembly. Click the "Align/Assemble" button and select "De Novo Assemble…" Check the box to "Assemble by:" 1st part of name, separated by "_ (Underscore)." Also save assembly report, save list of unused reads, save in sub-folder, save contigs, and save consensus sequences.

What did we just assemble in the step?

This step takes the contigs assembled in the previous step and attempts to find regions that overlap between them. The original contigs contain sequence that is mostly the targeted exon, but should also hold some splash zone on either end. The splash zone sequence is probably intronic, and if the splash zones of neighboring exons are large enough to overlap, they can be assembled into a combined sequence containing each exon and their shared intron.

Within each assembled contig, what part of a gene do the overlapping regions correspond to (at least in part)?

The overlapping regions will be intronic, though the intron can (and probably does) extend beyond just the area where the sequences overlap.

What sequences are in the document named "Unused Reads?"

These are the remaining assembled exons that could not be combined because their splash zone was too small relative to the intron that separates them.

5. Select the new "Consensus Sequences" document that was just created and the "Unused Reads" document. Under the "Sequence" menu select "Extract Sequences from List…" and extract to a new subfolder called "Merged and Separate Exons."

6. Navigate to the new folder and be sure the sequences are sorted by name. Under the naming scheme we use here, the number after the "m." and before the underscore represents the gene from which each exon was derived. Choose one of the "m." numbers. Select all of the documents with that number by holding the Shift key and clicking the first and the last document that share a gene name. Under the "Tools" menu select "Concatenate Sequences or Alignments…" and click OK. This produces a concatenated sequence made up of all the assembled exons for a gene. You may want to rename the document or change its description to be more informative. If we were working in this exercise with multiple samples, you could extract the sequences for the same gene from each sample, align them, and apply your favorite phylogenetic analysis. The concatenated consensus can be exported out of Geneious by selecting it, clicking the File menu, then "Export" -> "Selected Documents…" The format for the exported document can be selected under the "Files of Type:" box.

7. The assemblies for each contig in a gene can also be concatenated together and exported in a similar manner. These can be exported as a SAM file (Sequence Alignment/Map), a common file format for sequence assemblies, or a BAM file, a binary version of the same format that uses much less disk space.

Exercise 5 – Hyb-Seq Probe Development

This exercise will take you through the first steps of the Hyb-Seq pipeline to go from a genome and transcriptome sequence through to a file containing exon sequences appropriate for probe development based on the criteria used for the analysis and introduce some Linux basics. The developed probes can then be used in solution hybridization to enrich an Illumina library for the genes of interest. Sequencing of these libraries yields 50-80% on-target reads. The remainder of the reads can be used for genome skimming to assemble organellar genomes, nrDNA sequences, or other high-copy portions of the genome. The analysis you are about to conduct will utilize contigs from the unpublished common milkweed (*Asclepias syriaca* L.) genome and transcriptome assemblies. Note that in the flow of research the content of this exercise would be the first, rather than last step.

Throughout the following exercise, note that commands will appear in bold and text to enter into the command line will appear in `Courier` font. Pressing Enter at the end of the line is assumed. Be sure to carefully differentiate between **`l`** (lower case L) and **`1`**`(one)`.

1. Log into iPlant and open Atmosphere. Then click "Resume" to restart the Botany_Workshop instance. Open a shell using the "Access vis Shell" tab. Once connected, navigate into the Desktop directory.

   ```
   cd Desktop
   ```

2. Make a new directory for this workshop module and navigate into it.

   ```
   mkdir hyb-seq
   cd hyb-seq
   ```

3. Next, retrieve the scripts and sample data necessary to run the Hyb-Seq pipeline. All of the scripts are available from https://github.com/listonlab/Hyb-Seq_protocol.

For the purposes of the workshop, all of the needed scripts and sample data have been combined into a tarball (multiple files combined into a single archive). Open a browser and navigate to http://files.cgrb.oregonstate.edu/Botany/listonlab/Botany_2015_NGS_Workshop/. Once on this page, right click on "botany_ws_hybseq.tar.gz." Then click on "Copy link address" to copy the URL. Then use **wget** to transfer the file to your Atmosphere instance. Use Ctrl-Shift-V to paste for PCs or Command-Shift-V to paste for Macs.

```
wget
http://files.cgrb.oregonstate.edu/Botany/listonlab/Botany_2015_NGS_Workshop/botany_ws_hybseq.tar.gz
```

If "^M" appears when you press "Enter," delete it and press enter again. Then extract the files from the tarball.

```
tar -zxvf botany_ws_hybseq.tar.gz
```

4. The shell script Building_exon_probes.sh will call all of the scripts necessary such that given starting genome and transcriptome files, a series of sequence matching, filtering, and clustering steps will be performed, and in the end, a file of potential probe sequences for Hyb-Seq will be produced. Use the less command to view the text of this script. Note that for efficiency, you can use the Tab key to autofill any file name after typing the first few letters.

```
less Building_exon_probes.sh
```

Use the spacebar to gradually reveal more of the text of the script. Comments added to the script to help a user understand what a particular part of the script is doing are on lines beginning with "#". We intentionally added thorough comments to make this script more accessible, but unfortunately most scripts you come across will not be so well explained. You can use "q" to exit this view at any time.

5. Run the Building_exon_probes.sh shell script, which will call the various scripts you just obtained as well as several other programs. Under default settings, the script will call BLAT (Kent, 2002) to identify nuclear genome contigs that share 99% identity with transcripts. The pool of potential targets will then be narrowed to putatively single-copy genes by eliminating targets with 90% or greater sequence identity using CD-HIT-EST (Li and Godzik, 2006). The pool of potential targets is

then filtered for those genes with exons that are at least 120 bp to facilitate probe design, as well as with total exon length of at least 960 bp to increase the chances of finding phylogenetically informative variation among species for each gene. These parameters can be changed to suit your applications by modifying the shell script.

```
./Building_exon_probes.sh
```

6. Use **ls** or **ll** to view a list of the output files from the various parts of the probe design pipeline. The file named blocks_for_probe_design.fasta contains all of the exon sequences from the input genome that passed the input filtering criteria (i.e. low copy in the genome and meeting minimum length criteria). Use the **head** command to look at the top of this file.

   ```
   head blocks_for_probe_design.fasta
   ```

   The header for each sequence contains consists of the following: name of the transcript, name of the genomic contig_exon number, length of the exon.

7. Next check how many of the original genome sequences yielded probes. You can use Linux commands to quickly count the numbers of sequences in each file. There are many ways you could accomplish this task, and one example utilizing **grep** and taking advantage of the fact that the ">" symbol is only present once in the header for each sequence is given below. Essentially you are counting the number of lines in the file that contain at least one ">".

   ```
   grep -c ''>'' genome.fasta
   ```

   __106___ number of genomic contigs

   Use the same strategy to count the number of exons in the blocks_for_probe_design.fasta file.

   ```
   grep -c ''>'' blocks_for_probe_design.fasta
   ```

   __11___ number of exons for probe design

8. Counting the number of genes in the file is a bit more complicated because each exon is listed separately, genes may have different numbers of exons, and some exons are unsuitable for probe design. Use the command below to count the genes. Be sure to differentiate one and lowercase l.

```
grep ">" blocks_for_probe_design.fasta | sort | cut -f1 -d _
| uniq | wc -l
```

2 genes

Next try building up the command piece by piece by using the **head** command after each pipe (|) to view the result of each step. Pipes allow you to connect commands or programs together such that the output from the first command is the input for the one that follows it. An example appears below.

```
grep ">" blocks_for_probe_design.fasta | sort | head
```

Note that in the example exercise that you just performed, the transcriptome and genome data originated from the same species. If the genome and transcriptome are obtained from different species, you will need to adjust various parameters to account for divergence.

Concluding Encouragement

A great resource for finding answers to your questions about sample preparation, raw data handling, and basic and advanced genomics analyses is SEQanswers (http://seqanswers.com). Ripma et al. (2014) have also recently published workflows for molecular systematics studies that may be of interest. As you continue working with NGS data, gaining Linux skills and being able to operate comfortably in a command line environment are essential. Keith Bradnam and Ian Korf at UC Davis have an excellent online tutorial geared toward learning skills that you will likely find useful: Unix and Perl Primer for Biologists V.3.1.1. (http://korflab.ucdavis.edu/Unix_and_Perl/unix_and_perl_v3.1.1.html). This tutorial has also been expanded and printed in book form: *Unix and Perl to the RESUCE!* Happy sequencing and assembling!

References Cited:

KENT, W. J. 2002. BLAT—the BLAST-Like Alignment Tool. *Genome Research* 12: 656-664.

LANGMEAD, B., C. TRAPNELL, M. POP, and S. SALZBERG. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

LI, H., and R. DURBIN. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

LI, W., and A. GODZIK. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658-1659.

RIPMA, L. A., M. G. SIMPSON, and K. HASENSTAB-LEHMAN. 2014. Geneious! Simplified genome skimming methods for phylogenetic systematic studies: A case study in *Oreocarya* (Boraginaceae). *Applications in Plant Sciences* 2(10): 1400062.

STRAUB, S. C. K., M. PARKS, K. WEITEMIER, M. FISHBEIN, R. C. CRONN, and A. LISTON. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349-364.

STRAUB, S. C. K., M. FISHBEIN, T. LIVSHULTZ, Z. FOSTER, M. PARKS, K. WEITEMIER, R. C. CRONN, et al. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

WEITEMIER, K., S. C. K. STRAUB, R. CRONN, M. FISHBEIN, A. MCDONNELL, R. SCHMICKL, and A. LISTON. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.

Direct future questions to the authors of each exercise:

Exercises 1-3, 5 were prepared by:
Shannon Straub
Assistant Professor
Hobart & William Smith Colleges
straub@hws.edu

Exercise 4, 5 were prepared by:
Kevin Weitemier
Liston Laboratory
Oregon State University
weitemik@science.oregonstate.edu