Botany 2014

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 Illumina data generated in the Liston lab 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 (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 2012 (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.

To determine how many reads did not pass the chastity filter, search (Edit; Find in Document; Find All) for :Y: in the sequence names. Why is the same number obtained for read 1 and read 2?

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.

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. 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 short read kit)." 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 4.7.1 of the Geneious manual for a detailed description of the read mapping algorithm. You will map reads of *Asclepias euphorbiifolia* to an annotated *Matelea biflora* reference sequence in Geneious. *Matelea* is placed in a different subtribe from *Asclepias*, the Gonolobinae.

A. The reference sequence

1. View the provided *Matelea* reference plastome in Geneious by choosing the name of the file (KF539850) 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.

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 reference file (KF539850) 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 KF539850 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.

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

1. When the read mapping is complete, click on the assembly report in the Geneious document table pane.

2. From the assembly report we 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 *Matelea* plastome.

a. Chloroplast content calculation

(_____ reads mapped to cp / _____ total reads) * 100 = _____ % 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 *Matelea* reference (127384 bp).

(_____ reads mapped * _____ bp read length)/_____ bp genome size = _____× sequencing depth

D. Examine the read mapping results

1. Within the new folder named "Asclepias_euphorbiifolia_Hyb-Seq_reads_R_001 assembled to KF539850" 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. 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? _____

2. 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.

Which parts of the plastome assembled well and which did not using read mapping?

3. 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?

   _____

   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 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.

4. Geneious can aid in the search for 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."

5. 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)?

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.

Alignreads (Straub et al., 2011) is an assembly pipeline developed for reference-guided assembly. The pipeline incorporates YASRA (Yet Another Short Read Aligner) (Ratan, 2009), which handles indels and performs well even when the reference sequence and the sequence to be assembled are divergent. YASRA maps reads to a reference genome using the LASTZ algorithm (Harris, 2007) and refines their alignment using ReAligner (Anson and Myers, 1997). Genomic regions without coverage after the initial round of alignment are then masked, and the program attempts to assemble the missing sequence by tiling the remaining reads across the masked region. The resulting assembly is used as the reference for a subsequent round of alignment and tiling, and the process iterates until no further improvement is made. Next, NUCmer and delta filter from the MUMmer 3.0 suite (Delcher et al., 2002; Kurtz et al., 2004) are used to align the assembled YASRA contigs to a reference sequence, which can be either the reference sequence used for assembly or another reference of interest. Two scripts, sumqual and qualtofa, are used to integrate the YASRA assembly information and the NUCmer alignment information and then format the results into a user-friendly fasta file, as well as apply filters based on user inputs, such as masking for sequencing depth. We are in the process of finalizing a new version of Alignreads due to the release of a new version of YASRA and hope to make it available soon (https://github.com/zachary-foster/alignreads).

The Geneious read mapper also has an interation 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? _____

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? _____

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

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.
   - We won't be covering the de novo assembly in today's workshop, but if you would like to try this on your own with the *A. euphorbiifolia* plastid genome do the following: map the reads to the *Matelea* reference as you did before, but this time select the options to "Save list of used reads" and "Include mates." This will produce the mapped reads contig as before, along with a "Used Reads" document. Select this document and choose "Align/Assemble" then "De Novo Assemble…" The options to "Save assembly report," "Save in sub-folder," "Save contigs," and "Save consensus sequences" will be helpful. (Using the Medium/Low sensitivity option, this may take 30-45 minutes).

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. _____

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?

_____

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 final iteration?

   _____

   What proportion of reads mapped to the target exons? _____

By how much did the target enrichment process increase the proportion of targeted reads over an unenriched library? _____

How many exons (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.

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

   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.)

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 this step?

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

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

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."

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.

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.

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). 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:

ANSON, E. L., and E. W. MYERS. 1997. ReAligner: a program for refining DNA sequence multi-alignments. *Journal of Computational Biology* 4: 369-383.

DELCHER, A. L., A. PHILLIPPY, J. CARLTON, and S. L. SALZBERG. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* 30: 2478-2483.

GNERRE, S., I. MACCALLUM, D. PRZYBYLSKI, F. J. RIBEIRO, J. N. BURTON, B. J. WALKER, T. SHARPE, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences, USA* 108: 1513-1518.

HARRIS, R. S. 2007. Improved pairwise alignment of genomic DNA. Ph.D. dissertation, Pennsylvania State University, University Park, Pennsylvania, USA.

KURTZ, S., A. PHILLIPPY, A. L. DELCHER, M. SMOOT, M. SHUMWAY, C. ANTONESCU, and S. L. SALZBERG. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5: R12.

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, R., W. FAN, G. TIAN, H. ZHU, L. HE, J. CAI, Q. HUANG, et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* 463: 311-317.

RATAN, A. 2009. Assembly algorithms for next-generation sequence data. Ph.D. Dissertation, The Pennsylvania State University, University Park.

SIMPSON, J. T., K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. M. JONES, and İ. BIROL. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* 19: 1117-1123.

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.

ZERBINO, D. R., and E. BIRNEY. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821-829.

Exercises 1-3 were prepared by:

Shannon Straub

Assistant Professor

Hobart & William Smith Colleges

straubs@science.oregonstate.edu

Exercise 4 was prepared by:

Kevin Weitemier

Liston Laboratory

Oregon State University

weitemik@science.oregonstate.edu