Botany 2013
Introduction to Next-Generation Sequencing Workshop
Practical Exercises version 3

The following exercises are intended to help you familiarize yourself with Illumina data and the three basic kinds of analyses that are performed using short read data: read mapping, reference-guided assembly, and de novo assembly. For the workshop you will use Illumina data generated in the Liston lab for two species of Apocynaceae (Asclepiadoideae): a milkweed vine, *Matelea biflora*, and a milkweed, *Asclepias boliviensis.* You will be performing analyses appropriate for data sets produced through genome skimming (see Straub et al. 2011 and Straub et al. 2012) for *M. biflora,* 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 some information obtained about the low-copy and single-copy portion of the genome. You will use the *A. boliviensis* data for a de novo genome assembly.

Exercise 1 – Raw Illumina Data

Illumina data analyzed using the Casava v. 1.8 pipeline are returned as fastq files. Depending on the number of reads obtained and the standards at your sequencing center, the read pool may be broken down into several smaller files of more manageable size (e.g., 4 M reads per file at Oregon State). Some data analysis programs accept fastq files and utilize 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) provided for the workshop in the "example_files" directory. These reads are a subset of the *Matelea* read pool. 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's Casava v. 1.8 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 Illumina chastity and purity filter (N=pass, Y=fail):control bits:index(barcode) sequence. Many users will filter reads based on the Illumina quality flag.

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. Illumina's Casava v. 1.8 encodes standard Sanger quality scores (Phred+33) using 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 and corresponding Phred quality scores:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
000000000011111111112222222222333333333344
012345678901234567890123456789012345678901
```

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 that quality scores may be in one of the older Illumina formats (http://en.wikipedia.org/wiki/FASTQ_format), and will need to be converted prior to running scripts that expect v1.8 values.

B. Quality Score Visualization in Geneious

Click on the Matelea_biflora.fastq file in the read_mapping_and_reference_guided directory 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.

In general, which parts of the reads seem to have the highest quality? ____________________

C. Duplicate read removal

1. The Illumina library preparation protocol includes a PCR enrichment step that can lead to the presence of PCR duplicates (identical sequences) in the sequenced read pool, although it is also possible for there to be true biological duplicates due to the shearing of DNA in the protocol. It is often desirable to remove duplicate reads before calculating the sequencing depth of your target or proceeding with downstream analyses. Use the "Find Duplicates" option in the Geneious Edit menu to find sequences with identical residues in the Matelea_biflora.fastq read file and extract the unique sequences to a separate file.

2. Look at the text view to quickly see the number of reads in each file.

How many duplicate reads were present in the file? __________

How many times was the sequence with the most duplicates present in your file? __________

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 from the genome skim of *Matelea biflora* to an annotated *Asclepias syriaca* reference sequence in Geneious.

A. The reference sequence

1. View the provided *Asclepias syriaca* reference plastome in Geneious by clicking on the "read_mapping_and_reference_guided" directory and choosing the name of the file (JF433943) 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 unique *M. biflora* reads that you obtained in duplicate read removal exercise above. To map the reads highlight both the read file and the reference file (JF433943) 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 JF433943 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 *M. biflora* 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 *M. biflora* library and the sequencing depth of the plastome for this individual. Assuming the sequence of the *M. biflora* plastome is unknown, we can make estimates of each of these metrics using the number of reads that map to the *A. syriaca* 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 *A. syriaca* reference (158,798 bp).

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

D. Examine the read mapping results

1. Click on the file named "Unique sequences from Matelea_biflora assembled to JF433943" in the document table pane 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 chloroplast 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.

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 *M. biflora* 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 well and performs 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 on our website: www.milkweedgenome.org. ARC (Assembly by Reduced Complexity) is another promising new pipeline that was introduced at Evolution 2013 and we think is one to keep your eye on in the near future for plastome and targeted sequencing data assemblies. See https://github.com/ibest/ARC for more information.

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

1. Select the file of unique *M. biflora* reads and *A. syriaca* 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.

Exercise 4 – De novo assembly

De novo assemblies of genomes use only the sequences of the short reads themselves to create longer stretches of sequences (contigs). The two main approaches employed are overlap-layout-consensus and de Bruijn graph algorithms. Some of the recently popular programs available for de novo assembly include ABySS (Simpson et al., 2009), ALLPATHS-LG (Gnerre et al., 2011), SOAPdenovo (Li et al., 2010), and Velvet (Zerbino and Birney, 2008). See section 4.7.1 of the manual to learn more about the overlap-layout-consensus assembly algorithm employed by Geneious.

A. Starting a de novo assembly in Geneious

For the de novo assembly portion of the workshop you will use data from a genome skim of *Ascelpias boliviensis.* This data set is much smaller than those normally used for whole genome sequencing, but will give you a start on developing genome assembly skills in the time allotted for the workshop. A de novo assembly of the *A. boliviensis* read pool would take approximately 30 min. to run in Geneious using the lowest sensitivity setting and assuming an allocation of 2 GB of RAM to the process and a ~2.8 GHz processor. Due to the variation in participants laptops, the steps to start a de novo analysis in Geneious are described below, but we will explore contigs from a set produced by running an assembly prior to the workshop. The raw data were included with the workshop files you downloaded, so you can re-run this analysis on your own time or explore de novo assembly with the *Matelea* data set.

1. To begin a de novo assembly, you would click to highlight the name of the *Asclepias boliviensis* read file (Asclepias_boliviensis_reads.fsa in the de_novo_assembly directory) in the top document table pane of Geneious. You would then see a graph showing the lengths of the reads in the file appear in the bottom pane. In this case, they are all 74 bp.

2. Next, you would choose the “Align/Assemble” button from the Geneious menu, select “De Novo Assemble…”, choose options and then click “OK.”

B. Exploring the results of the de novo assembly for *Asclepias boliviensis*

Start by clicking on the directory that contains the de novo assembly results you imported into Geneious prior to the workshop to view all of the result files in the document table pane of Geneious.

1. Click on the Assembly Report in the document table pane. You can use this report to explore the assembly statistics associated with this run. We will look at this information in more detail in the next section.

2. Next, click on the Sequence Length tab in the document table pane and sort the contigs from largest to smallest.

3. Click on Contig6, the longest contig recovered in the de novo assembly. Use the information reported about this contig to begin filling in Table 1 on p 9.

4. Use the same formula you used to determine the sequencing depth of the plastome to calculate a sequencing depth for this contig.

5. Now we will try to determine the genomic origin of the contig and some genes that might be contained in this contig using a Custom Blast in Geneious. To begin, click the Sequence Search Button.

6. Within the Sequence Search panel choose “selected_plant_genomes” as the database. This is the custom library you imported prior to the workshop. Then for Program, choose “Discontiguous Megablast.” The remainder of the default options should be fine, so then click the Search button to start your BLAST search.

7. When the BLAST hits are returned, explore the results and use them to complete the rest of the table for Contig6.

8. Repeat this process for contigs 825, 50, and 119. Randomly choose some additional contigs and explore those as well.

Table 1. Contig statistics for a de novo assembly of the *Asclepias boliviensis* genome

Contig Number	Contig Length	# Reads Assembled	Sequencing Depth	Genome (nuc., cp, mt)	Description of BLAST hit annotation that indicated the genome of origin for the contig
6					
825					
50					
119					

9. Based on the number of reads that assembled for contig 119, what can you say, in general, about its copy number in the genome? Considering the preliminary gene annotation, is this surprising?

10. Note that the chloroplast contigs in the de novo assembly could be ordered relative to one another using a reference sequence to produce a plastome sequence or combined with a reference-guided analysis to produce a finished plastome sequence.

C. Assessing assembly quality

Assembly quality is assessed in several different ways. In general the more sequence in long contigs and fewer contigs (provided most of the reads are being used) the better. The N50 is a value often reported for genome assemblies and indicates that 50% of the assembly is in contigs of that size or greater, and again, the higher the better. Next you will explore these values for the *A. boliviensis* assembly.

1. Return to the assembly report from the de novo assembly analysis to fill in values in Table 2 on p 10.

Table 2. Assembly statistics for the *Asclepias boliviensis* de novo assembly

Assembly method	Number of reads used	Number of Contigs	Number of Contigs >1kb	Length of the longest contig	N50 (all contigs)	Length of the assembly (bp) in contigs > 1kb
de novo						

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.

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

These exercises were prepared by:
Shannon Straub
Postdoctoral Research Associate
Liston Laboratory
Oregon State University
straubs@science.oregonstate.edu