De Novo Assembly walkthrough

PacBio

For PacBio, CANU can be used to perform the assembly. An example of the script can be found on the VSC under the Shark project. The CANU tool needs to be split into the 3 possible tasks. The walltime will be exceeded, and you have to ask for more time to the helpdesk. After assembly a QUAST report and BUSCO report should be made to asses the genome.

Illumina

For Illumina, the process is more complicated. There are several steps (some of them are optional):

- Read Quality
- Cleaning of the data [optional]
- PreQC
- Assembly (velvet/ABySS/SGA/…)
- Report
- Remapping
- Reducing of scaffolds [optional]
- Reference assisted scaffolding [optional]
- Repeat masking [optional]
- Annotation [optional]
- BUSCO report

An example of the scripts can be found on the VSC under the Octopus project. Below an overview of each task is given, together with possible tools and the duration (short is approximately a day, long means give the max walltime possible).

Read Quality

It is best to run a regular FastQC on the data to check the quality.
Tool: FastQC
Duration: short

Cleaning of the data

Optionally, if data cleaning is needed according to the previous QC. This can be removing low quality endings, or adapter removal. If adapters are present in the data, this will become a part of the final draft genome!
Tool: Trimmomatic
Duration: short

PreQC

An initial estimate of the genome can be made by using SGA. This old tool can give an estimation of the genome size, library size, heterozygosity rate, GC bias and best to use kmer.
Tool: SGA preprocess, SGA preqc
Duration: long (7 days)

- Genome size
  Figure: Est. Genome Size
  Gives an estimation of the genome size (in Mbp)
- Library size
  Figure: Estimated Fragment Size Histogram
  This should correspond to the lab traces.
- Heterozygosity rate
  Figure: 51-mer distribution
  To make this distribution, all 51-mers are taken, and there occurrence is counted. There should be a high peak around 1. These are all unique 51-mers (errors). The rest of the graph should have 1 peak. This means that the heterozygosity rate is low. When 2 peaks are observed (first peak at the half of the value (x) of the second peak), this indicates on a high heterozygosity rate. How higher the peak the higher the heterozygosity rate. The rate could be calculated by taking the area under the curve for both peaks, but dividing the area of the first peak by two times 51 (there will be 51 different kmers for 1 variation, and each variation has 2 51-mers: the 2 alleles).
- GC bias
  Figure: GC bias
  This graph should give a red circle. A smear on the y axis indicates a high coverage for a certain GC percentage, which indicates high repeat regions. A smear on the x axis indicates a high variation in GC content. This variation could indicate chromosome duplications or an evolutionary introduction of another genome (or worse contamination?).
- Kmer
  Figure: Simulated contig length vs k
  This figure shows an estimation of the contig length for kmers between 20 and 90. The ideal kmer to use, is at the peak of this graph. When multiple peaks occur, try the different kmers. Remember to only use odd kmers!

Assembly

The assembly can be done with different tools (according preference). ABySS is a fast tool, gives decent results and is regularly updated. Sometimes it could be preferable to use a different tool or method (de Brujin graph vs String graph). The adviced kmer of the PreQC report should be used here.
Tool: ABySS, SGA, Stride, Velvet
Duration: long

Report

The Quast report will give you a result of the produced draft. It is a good idea to compare the singletons, contigs and scaffolds to eachother (redo this report if you perform further steps), to see the improvements. In this report the important statistics are genome size, #scaffolds, #Ns and the continuity of the genome (combination of max, N50, L50, N75, L75 and min).
Tool: Quast
Duration: short

Remapping

Remapping the data to the created reference will give an indication of the data used for the assembly. If you map with Bowtie2, you will get a report with the mapping percentage (this should be very high, over 95%, else try a different assembler) and the number of multimappers. The number of multimappers can indicate repeat regions, but also a high number of heterozygousity that couldn't be solved by the assembler. In this case it is adviced to do a reducing of the scaffolds.
Tool: Bowtie2
Duration: short

Reducing of scaffolds

If a high heterozygousity is predicted, there is to much variation inside the genome for the assembler to create complete scaffolds. This means that for some scaffolds both alleles are still present in the draft genome. These must be reduced.
Tool: redundance
Duration: long

Reference assisted scaffolding

If a reference genome of a close species is available, this can be used to do a better scaffolding. You must also be careful, because possible errors in the used reference can be transferred to your own draft genome.
Tool: blast and chromosomer
Duration: short

Repeat masking

Repeat masking can be performed as an extra step. The found repeats in the draft will be turned to lower case letters. However, this step is only interesting when you decide to create a blast database since mappers and variant callers used to ignore the masked regions.
Tool: Windowmasker
Duration: short

Annotation

When a transcriptome is available, this can be mapped to the created reference, in order to get an annotation file
Tool: splign
Duration: short

BUSCO

Busco is a tool to check for conserved genes. You need to download the appropriate dataset for you species (there are sets for mammals, insects, plants, ...). BUSCO will check the genome for complete genes, partial genes, … How higher the number of found BUSCO genes, how better the assembly.
Tool: BUSCO
Duration: long