Genome Data Management Workshop
January 22nd-26th 2011
Universidad de Puerto Rico en Mayagüez

Genome assembly

I. Make a Velvet assembly (illumina data)

/media/Data_2/GDMWXY/Velvet

In the shell, type *cd /media/Data_2/GDMWXY/Velvet/*
Then, type *shuffleSequences_fastq.pl reads1.fastq reads2.fastq allreads.fastq*
This perl script will merge the .fastq files into a single interleave .fastq file, as required by Velvet.
Type *velveth KMER31 31 -shortPaired -fastq allreads.fastq*
KMER31 is the name of the folder that will be generated
31 is the length of the KMER used for the assembly
-shortPaired is the type of reads used (as specified in the Velvet manual)
-fastq is the file format
BTW saw the message? Nice little inoffensive bug… To use a bigger KMER, it must be specified when compiling the source code.
NOTE: The assembly with this dataset takes a long time (≥ 5 hours) and ≥ 24 GB of RAM per computation. Don't run it. Fabien will start it as a demonstration and kill the process after a while…
The command Fabien is typing is *velvetg KMER31 -exp_cov 50 -ins_length 340* in which:
-exp_cov is calculated as the total amount of sequence/expected genome size
-ins_length is the size of the library prepared for the sequencing

II. Make a Newbler assembly (454 data)

/media/Data_2/GDMWXY/Newbler

In the shell, type *cd /media/Data_2/GDMWXY/Newbler/*
Type *gsAssembler*
Select *New assembly project*
Enter a name for the assembly
In location, select */media/Data_2/GDMWXY/Newbler*
Click ok
Select the *Project* tab
Click on the + sign to add reads
In */media/Data_2/GDMWXY/Newbler*, select *reads1.sff* to *reads5.sff*
Hit start

III. Make a Ray assembly (illumina + 454 data)

/media/Data_2/GDMWXY/Ray

In the shell, type *cd /media/Data_2/GDMWXY/Ray/*
Type *Ray*

This will display the main command line switches
-np = number of cores for the computation
-p = paired ends reads
-s = single ends reads
-k = value of the Kmer used for the analysis
-o = the name of your output
Type *mpirun -np 2 Ray -p reads1.fastq reads2.fastq -s reads3.sff -k 21 -o NameYourOutput*
Ooops? Segmentation fault? Sometimes shit happens. Let's try again with more processing power…
Type *mpirun -np 8 Ray -p reads1.fastq reads2.fastq -s reads3.sff -k 21 -o NameYourOutput*

IV. Map reads and perform chromosome walking with Consed (454 data)

/media/Data_2/GDMWXY/CONSED/454/edit_dir

In the shell, type *cd /media/Data_2/GDMWXY/CONSED/454/edit_dir*
Type *fasta2Ace.perl Organelles.fasta*
This perl script generates a blank .ace file from a reference .fasta file
Then, type *add454Reads.perl Organelles.ace sff.fof Organelles.fasta*
This perl script will map the 454 reads on the blank .ace file
Type *consed*, then select *Organelles.ace.1*
Double click on *contig00016*
Highlight a sequence (will turn Yellow) with your mouse cursor
Click on *Search for String*
In the *Query String*, paste your sequence using your middle-mouse button, select *Exact* or *Approximate*, then click *OK*
The window popping up shows all the hits to that specific sequence. Hit ***Dismiss*** to close it.
We'll extend the contigs by doing Chromosome walking using the reads on the 3'-end of the contig. Then, we'll join a few contigs.

V. Map reads and add Sanger sequences with Consed (illumina + Sanger data)

/media/Data_2/GDMWXY/CONSED/illumina/edit_dir

In the shell, type *cd /media/Data_2/GDMWXY/CONSED/illumina/edit_dir*
Type *fasta2Ace.perl Hellem.fasta*
Then, type *addSolexaReads.perl Hellem.ace solexa.fof Hellem.fasta*
This perl script will map the illumina reads on the blank .ace file
Type *consed*, then select *Hellem.ace.1*
We could extend the contigs as in the previous exercise but we'll add a few PCR sequences to the assembly instead this time around.
Click on the *Add New reads* tab
Select *PCRs.fof* in the right field, then click OK
The text file ***PCRs.fof*** contains the names of the Sanger chromatograms (located in the chromat_dir) to add to the assembly. The ***.fof*** extension means File of Files (a CONSED convention I believe) but you could call the file anything you like (*e.g.* hot.coffee). I generated the file using the basic Linux command ***ls***.
The *New Reads In Assembly* windows that will popup contains links to the added sequences. You can then verify whether the sequence was added at the right spot or not.

Note. **After adding reads, you must save the assembly before doing anything else. In the top menu, select *File > Save assembly*.**

Genome annotation

I. Artemis (MS Windows, MacOSX or Linux)

/media/Data_2/GDMWXY/Artemis

NOTE: Artemis is a lightweight program that can run on pretty much any netbook with Java enabled. I strongly suggest that you download and install *Artemis* and run it from your laptop. You can download the chromosome to annotate with *WinSCP* (on Windows), *Cyberduck* (on OSX) or any other SFTP-capable file transfer program. It can also be run under X from our server over the net by typing *art* in the shell, but the bandwidth will be strained which will likely make it unbearably slow.

In Artemis, click *File > Open …*
Select *Chromosome_XY.fasta*, then click *OK*
From the menu, Click *Create > Mark Open Reading Frames …*
We'll use *100* as the *Minimum open reading frame size* for this exercise. Click *OK*
This will create a temporary file named ORFS_100+ in the Artemis default work folder. A good practice is to save this file, then relaunch Artemis.
From the menu, *Click File > Save An Entry As > EMBL Format > ORFS_100+*
Enter a name (e.g. ORFs.embl), then click *OK*
Close Artemis
Relaunch Artemis
In Artemis, click *File > Open …*
Select *Chromosome_XY.fasta*, then click *OK*
From the menu, *Click File > Read An Entry …*, then select the ORFs file
As you can see, a lot of these ORFs overlap and quite a few are spurious. We need to filter out the bad ones. To do so, we'll start by doing BLAST homology searches.

II. BLAST (web searches)

In a nutshell

From within Artemis
Select a feature you want to BLAST.
Click *Run > NCBI Searches > type of BLAST desired*
Once completed, the result will be displayed in your default web browser.

From a web browser
Goto *http://blast.ncbi.nlm.nih.gov/Blast.cgi*
Select the type of BLAST you want
Paste or upload your sequence and add desired options.
Click BLAST.

For this exercise
Goto *http://blast.ncbi.nlm.nih.gov/Blast.cgi*

Select *blastx*
Paste/upload the chromosome sequence you were given
In the field *Organism – Optional*, enter *Encephalitozoon intestinalis ATCC 50506 (taxid:876142)*
This will restrict the BLAST to search against this organism only. Otherwise the maximum query length would be about 12000bp. NOTE the name should appear in the scrolling menu as you type in the first few letters.
Click *BLAST*
In the results window, click *Download* on the top of the page, click *HitTable(text)* under *Alignment*, then save the file.
Open the file with any text editor.
***Remove the 7 lines starting with a #* from the file, then save it.**
Note, if you are using MS Word, make sure you save it as a text-only file.
Open the newly saved file in a spreadsheet program.
This file can be opened an edited in a spreadsheet like MS Excel or OpenOffice Calc.
***Delete columns # 4 & 5*, and then save the file.**
We need to remove these two columns before the file can be used in Artemis.
Go back to Artemis
From the menu, *Click File > Read An Entry …*, then select the BLAST file you just saved.
We'll use this file to filter most of the spurious ORFs.

III. BLAST (local searches)

/media/Data_2/GDMWXY/BLAST

NOTE: Former users of BLASTALL will notice that the command lines have change. NCBI rewrote the entire toolset from C to C++. The newest version is available *here*. A Perl wrapper (*legacy_blast.pl*) is available for the old command line syntax.

In the shell, type *cd /media/Data_2/GDMWXY/BLAST/*

To create a database:
Creating a database generates many files. It is a good idea to keep your databases in separate folders.
Type *mkdir DB*
Type *cp Chromosome_XY.fasta DB/*
Type *cd DB/*
Old syntax type ***legacy_blast.pl formatdb -i Chromosome_XY.fasta -p F -o T***
New syntax type ***makeblastdb -in Chromosome_XY.fasta -dbtype nucl -out Chromosome_XY.fasta -title "InsertSomethingFancy"***
This will create a nucleic acid database from your chromosome sequence in the current folder.

Type *cd ../* to go back to the BLAST directory

To perform a blastn:
Old syntax type ***legacy_blast.pl blastall -p blastn -i geneXY.fasta -d DB/Chromosome_XY.fasta -m 0 > blastn.out***
New syntax type ***blastn -query geneXY.fasta -db DB/Chromosome_XY.fasta -outfmt 0 -out blastn.out***
Try both commands. Look at the outputs with *less*. See anything odd?

To perform a tblastn:

Old syntax type ***legacy_blast.pl blastall -p tblastn -i proteinXY.fasta -d DB/Chromosome_XY.fasta -m 0 > tblastn.out***

New syntax type ***tblastn -query proteinXY.fasta -db DB/Chromosome_XY.fasta -outfmt 0 -out tblastn.out***

There are a lot of options hidden for local BLAST. Seek the manual for a complete description.

IV. tRNAscan-SE (web searches)

/media/Data_2/GDMWXY/tRNAscan-SE

Open up a Web browser
Goto *http://lowelab.ucsc.edu/tRNAscan-SE/*
In source, select *Mixed (general tRNA model)*
Technically, we should select Mito/Chloroplast from the Source menu as your sequence is from the plastid genome of the green alga *Pseudendoclonium akinetum* but the search would take about 2 hours due to the slow algorithm involved…
In format, select *Other* (FASTA, GenBank, EMBL, GCG, IG).
Paste or upload your sequence
Click *Run tRNAscan-SE*

Genome Submission

Sequin is the tool provided by NCBI to submit genomes to its GenBank database. Sequin can be used as a standalone tool but I do not recommend it. While it works fine for small genomes, its flaws are really problematic when using large genomes. The best way to do it is to prepare a submission table in TBL format (Artemis can save in the TBL format), and use the TBL2ASN conversion tool to generate the SQN file compatible with sequin. That way, you can modify/correct issues easily and regenerate an updated SQN version automatically.

I. Artemis > TBL2ASN > Sequin

/media/Data_2/GDMWXY/Sequin

In the shell, type *cd /media/Data_2/GDMWXY/Sequin/*
Type *art*
We'll learn to do a few more things in Artemis to prepare for a genome submission, and then we'll use TBL2ASN. NOTE: You can do this part from your laptop. Just fetch the Hel01.embl file via SFTP.
Click *File > Open …*, and then select *Hel01.embl*
This file contains all the annotations but lacks gene features. We'll create them.
From the menu, click *Select > All CDS Features*
Then, from the menu, click *Create > Gene Features*
Select a *rRNA* feature from the lower section
From the menu, click *Select > Same Key*
Type *Ctrl+D* (Command+D on a Mac) to duplicate the features selected, then click *Yes*

Artemis does not generate gene features from rRNAs and tRNAs. We must therefore create them manually. We will edit the duplicated features from the text file.

Select a *tRNA* feature from the lower section
From the menu, click *Select > Same Key*
Type *Ctrl+D* (Command+D on a Mac) to duplicate the features selected, then click *Yes*
From the menu, click *File > Save All Entries*
Close Artemis
Open the saved file with any text editor
Search for the duplicated entries and replace the *rRNA* or *tRNA* tags from the 1st duplicated entry with *gene*
Save the file
Open the file in Artemis again
Select a gene feature, then from the menu, click *Select > Same Key*
From the menu, click *Edit > Automatically Create Gene Names*
Enter *EHEL01_*, then click *…*

In the 1st window, we'll enter a descriptor for the gene names. Here we are annotating the *E. hellem* genome for which we were granted the tag EHEL for the GenBank database. We will therefore label the genes EHEL01_ to indicate the species and the chromosome number on which the genes are located.

In this window, start the counting at *10*
Click *…*
In this window, increment the number by *10*
Click *…*

As a rule of thumb, eukaryotic genes start at number 10 and are annotated by increments of 10. This allows for the addition of new genes in-between, when discovered later on, without messing up the ordering.

In this window, replace the qualifier *gene* with *locus_tag*, then click *…*
Enter *4* as the digits number, then click *…*

Because this microsporidian genome is really small and because we clearly defined on which chromosome the genes are located with the EHEL01_ tag, we only need 4 digits to annotate all the genes. On bigger genomes, we should use 5 digits or more…

In the append "c" names window, click *No*
Save the file (*File > Save All Entries*)
Then, select *File > Save An Entry As > Sequin Table Format > Hel01.embl*
Name the file *Hel01.tbl* and save it in your current directory
Click *File > Write > All Bases > Fasta Format*, then name the file *Hel01.fsa*

The TBL and FSA files will be needed by the TBL2ASN program. This program is REALLY picky about file names.

To create a template.sbt file from the web
Goto *http://www.ncbi.nlm.nih.gov/WebSub/template.cgi*
Enter the requested information
Click *create template* at the bottom once completed
Save the *template.sbt* file in your current directory

II. Artemis > TBL2ASN > Sequin

/media/Data_2/GDMWXY/Sequin

Type *mkdir TBL2ASN*

Create a folder TBL2ASN
Type *mv Hel01.fsa Hel01.tbl template.sbt TBL2ASN/*
Move the Hel01.fsa, Hel01.tbl and template.sbt files to the TBL2ASN directory
Type *cd TBL2ASN/*
Type *nano -w Hel01.tbl*
Nano is a shell-based text editor. You can also use any other text editor you like.
In the header, replace *>Feature Hel01.embl* with *>Feature Hel01.fsa*, then save the file (*Ctrl+X*, *Y*, then *Enter* in nano).
Type *nano -w Hel01.fsa*
In the header, replace *>all_bases* with *>Hel01.fsa*, then save the file
TBL2ASN is SO PICKY that you have to put the same names in and out of the files or it won't be able to match them and just do nothing even if you type in the correct command lines.
Go back in the parent folder, *cd ..*
Type *tbl2asn -t template.sbt -g -p TBL2ASN/*
The ***-t*** option loads the ***template.sbt*** you generated online. The ***-g*** option is for required for the submission of eukaryotic genomes but wasn't indicated in the manual last time I checked. The ***-p*** is the path where are your files from which the Sequin ASN file(s) will be generated.

NOTE: You will see an error message indicating the lack of transcript IDs. This is normal as the TBL file we are using is incomplete. We will not construct a complete file for this exercise. You will find an example of a complete TBL file in the REFERENCE folder.

What is missing?
The mRNA feature is not present
Each CDS feature requires an mRNA feature in addition to the gene feature.
The transcript_id is missing
Each mRNA feature require a unique transcript_id tag. This tag cannot be repeated in GenBank.
The protein_id is missing
Each CDS feature require a unique protein_id tag. This tag cannot be repeated in GenBank.
Those features are usually added to the TBL files with a bunch of Perl scripts. There are also some tricky workarounds that you can do in Artemis with text editors if you feel like being inventive.

II. Artemis > TBL2ASN > Sequin

/media/Data_2/GDMWXY/Sequin

In the shell, type *cd /media/Data_2/GDMWXY/Sequin/TBL2ASN/*
Type *sequin*
Click *Read Existing Record*, select *Hel01.sqn*, then click *OK*
You will see the file we generated. Look around the software and browse the different functions available. Most should be self-explanatory. In theory you can create an accession number straight from within Sequin but in practice, it is inefficient and error prone. Unless you are annotating small organelle genomes you should use TBL2ASN.

NOTE: While I do not recommend using Sequin to generate SQN files, it is very useful to quickly check if the files you generated are OK. Also, you can use it to generate the template.sbt file you need for TBL2ASN.

To create a template.sbt file from Sequin

Open Sequin, select *Start New Submission*
In the submission tab, click Release Date
By default the release date will be one year from now. This gives the authors the time to write the paper. The genome is usually released in the databases concurrently with the paper.
Enter a putative manuscript title
This can be changed afterwards.
Fill in all the tabs.
Return to the submission tab
Select File > *Export Submitter Info...*
In the selection box, name the file template.sbt, then click *OK*