Bio4342: Finishing Lab Practice Using a Mouse Contig

To do this exercise, cd to project M_BB0112A10 (it is in your home directory) and open Consed. This project is not finished and the assembly is not yet contiguous so there are multiple contigs listed. [cd M_BB0112A10/edit_dir]

1) The BAC cloning site is the junction between the project DNA and the site where it is incorporated into the BAC vector. These sites are designed to utilize restriction enzyme sites and are typically palindromes, meaning they are the same sequence both forwards and backwards (when complemented). What are the BAC cloning sites for this project and what are their base positions, including contig number in the Consed assembly? (Hint: examine the ends of each assembly, looking for the restriction sites.)

2) This project has some sequence that is repetitive. Inverted repeats are those that are complementary to each other and tandem repeats are in the same orientation. Inverted repeats often have sequence in between the two copies that is unique and is called the loop sequence. Tandem repeats are adjacent to each other. If such repeats are separated by intervening sequence, the repeated sequences are referred to as duplications (or triplications if there are 3 copies, etc.).

Where does the sequence ttgggcaagactct match and what type of repeat would is it?

3) For a project to be considered finished it must meet certain criteria. Below is an excerpt from the statement of the default standard for finished projects to be included with each finished mouse genome accession.

"This sequence was finished as follows unless otherwise noted: all regions were double stranded, or sequenced with an alternate chemistry, or covered by high quality data (i.e. phred quality >=30); an attempt was made to resolve all sequencing problems, such as compressions and repeats; all regions were covered by at least one plasmid subclone, fosmid clone or direct clone walk sequence. Sequence from the Mouse Genome Sequencing Consortium whole genome shotgun may have been used to obtain the consensus sequence. The assembly was confirmed by restriction digest."

To find the regions that do not meet these criteria so that a finisher can resolve them, the finisher navigates by low consensus quality, high quality discrepancies and single subclone regions. Single subclone regions do not require reactions if they are of good quality but they are annotated as "single subclone". The navigators are just a tool to find these regions; sometimes they will list regions that are not single subclone because there is a second subclone, typically of low quality, in the region that Consed does not consider. To find a single subclone region to be finished, open the Aligned Reads Window for contig 30. In the left hand corner, toggle the second word 'Navigate' and select on 'Regions covered by only 1 subclone (or none at all).' A box will appear, listing the single subclone regions for this contig that Consed has found.

List the single subclone regions that you found and state whether you think each one is truly a single subclone region.

Thought question: Why do you think that Consed considers some regions as *single subclone* that are not? One reason is that Consed is considering the data as if "Dim low quality" or "aligned" had been selected.

4.) When a region is annotated as being unresolved, restriction enzyme digest information is provided to estimate if there is data missing from the region. Below is the output of the Consed digest for the project that we have open for the enzyme *Hin*dIII (some irrelevant bands have been deleted for improved clarity). Consed has made an alignment between the fragments called by library core (real fragment sizes) and the *in silico* fragments that could be predicted from the location of the *Hin*dIII cut site AAGCTT in the assembly. Following this is a table with the in silico fragment sizes listed in the order in which they appear in the assembly.

Real Frag Size	*in silico* Size	Position
9419	9346	part vector/part insert Contig30 (124292-133313)
9011	8656	Contig25 (31912-36280) Contig30 (1-4644)

7974	8007	Contig30 (75231-83238)
7717	7695	Contig30 (19859-27554)
7202	7289	Contig30 (31042-38331)
7014	7077	Contig30 (39912-46989)
7014	7011	Contig25 (22942-29953)
6781	6782	Contig30 (57187-63969)
6502	6511	vector
5762	5793	Contig30 (83238-89031)
5461	5505	Contig30 (95008-100513)
5461	5496	Contig30 (66806-72302)
5461	5449	Contig30 (5714-11163)
5395	5444	Contig30 (14415-19859)
5265	5376	Contig30 (105319-110695)
5194	5248	Contig30 (113989-119237)
5102	5055	Contig30 (119237-124292)
4634	4635	Contig30 (50941-55576)
3642	3656	Contig25 (16981-20637)
3484	3488	Contig30 (27554-31042)
3268	3288	Contig30 (89031-92319)
3268	3251	Contig25 (8284-11535)
3184	3165	Contig25 (1973-5138)
3087	3082	Contig30 (46989-50071)
2936	2929	Contig30 (72302-75231)
2903	2891	Contig25 (14090-16981)
2844	2837	Contig30 (63969-66806)
2715	2718	Contig30 (110695-113413)
2691	2689	Contig30 (92319-95008)
2553	2555	Contig25 (11535-14090)
2226	2217	Contig25 (6067-8284)
2193	2164	Contig30 (12201-14365)
1956	1959	Contig25 (29953-31912)
1838	1828	Contig30 (100513-102341)
1820		
	1766	part vector/part insert Contig25 (1-762)
1592	1581	Contig30 (38331-39912)

In Silico Sorted By Position

Frag Size	Frag Position
1766	part vector/part insert Contig25 (1-762)
1211	Contig25 (762-1973)
3165	Contig25 (1973-5138)
929	Contig25 (5138-6067)
2217	Contig25 (6067-8284)
3251	Contig25 (8284-11535)

2555 Contig25 (11535-14090)
2891 Contig25 (14090-16981)
3656 Contig25 (16981-20637)
1562 Contig25 (20637-22199)
743 Contig25 (22199-22942)
7011 Contig25 (22942-29953)
1959 Contig25 (29953-31912)
8656 Contig25 (31912-36280 Contig30 (1-4644)
1070 Contig30 (4644-5714)
5449 Contig30 (5714-11163)
1038 Contig30 (11163-12201)
2164 Contig30 (12201-14365)
50 Contig30 (14365-14415)
5444 Contig30 (14415-19859)
7695 Contig30 (19859-27554)
3488 Contig30 (27554-31042)
7289 Contig30 (31042-38331)
1581 Contig30 (38331-39912)
7077 Contig30 (39912-46989)
3082 Contig30 (46989-50071)
870 Contig30 (50071-50941)
4635 Contig30 (50941-55576)
356 Contig30 (55576-55932)
1255 Contig30 (55932-57187)
6782 Contig30 (57187-63969)
2837 Contig30 (63969-66806)
5496 Contig30 (66806-72302)
2929 Contig30 (72302-75231)
8007 Contig30 (75231-83238)
5793 Contig30 (83238-89031)
3288 Contig30 (89031-92319)
2689 Contig30 (92319-95008)
5505 Contig30 (95008-100513)
1828 Contig30 (100513-102341)
879 Contig30 (102341-103220)
748 Contig30 (103220-103968)
407 Contig30 (103968-104375)
944 Contig30 (104375-105319)
5376 Contig30 (105319-110695)
2718 Contig30 (110695-113413)
576 Contig30 (113413-113989)
5248 Contig30 (113989-119237)
5055 Contig30 (119237-124292)
9346 part vector/part insert Contig30 (124292-133313)
449 vector
1527 vector

190 vector
644 vector
6511 vector

Question: There is a gap between Contig 25 and Contig30. What is the real fragment size for the region that covers this gap? What is the in silico fragment size? What is the difference between these two sizes- this is the size of the gap- the approximate amount of data that is missing?

5.) The gap region will need more work. Prefinishing has ordered oligos for PCR at the gap. The oligo sequences are tagged in the database.
What is the sequence of oligo 45 and of oligo 46 from 5' to 3'? (Hint: check the tag information.)

6.) "Assembly View" is a tool in newer versions of Consed. To open "Assembly View" click on the button 'Assembly View' in the Consed Main Window. It will bring up a visual representation of the contigs in the assembly. Contig 25 and Contig 30 are listed on the top row with purple lines connecting them. These lines symbolize forward/reverse read pairs, and indicate that the contigs should be joined. You can click on these and they will list the reads. Unfortunately, most of the reads are Whole Genome Shotgun Reads and the DNA subclones are not available to the finisher as a template for reactions. Red lines are forward/reverse pair subclones that are not consistent.

What do you think is going on with Contig29? (Hint: Contig29 consists entirely of Whole Genome Shotgun Reads.)

7.) A finisher would need to close the gap between Contig 25 and Contig 30. A finisher can either sequence an existing subclone or produce a PCR product on which to sequence in order to obtain the missing data. How many subclones span the gap between Contig 25 and Contig 30? List the names of the spanning subclones for this gap.
(Hint: click on the purple lines in Assembly View for this answer.)

Which do you think is preferable - using a subclone as a template for an additional sequencing reaction, or creating a PCR product and why?

8.) There are three main types of finishing chemistries used at the GSC:
Big Dye, dGTP and 4:1. What are the strengths and weakness of each?

9.) The sequencing reactions are named so that a user will know which subclone and chemistry was used to produce a given read. Big Dye reactions have no underscore ``_'' and have the .b if run with the standard forward primer, or the .g suffix if run with the standard reverse primer. Non-standard chemistries and oligos have an underscore following the subclonse name and a letter or number indicating the chemistry or oligo: dGTP reactions have _g incorporated into the name and 4:1 reactions have _t incorporated into the name. Oligo names are also included to the right of the g or t if appropriate. Finally, if a reaction is repeated and will create a file with the same name, the last number is incremented. Thus, if pkh30h05.g1 were repeated for some reason, the new reaction would be call pkh30h05.g2. By convention reactions run with a non-standard oligo are given the .b# suffix. Look at the read list in the Consed Main Window for examples of the naming convention.

What are some reactions that you would call for the gap? List multiple subclone options that would work and explain which ones you think would be the best options. Give the appropriate reaction names.

10.) Some reads have been done for the gap region. You need to add them to the assembly. One way is to do *Add New Reads*. An EMACS file named *reads_to_raid.fof* has been created which lists the three reads below. Add the following reads to your project (they are already in your project directory) using the *Add New Reads* function (see the Consed tutorial for additional help).

ofl68h09_46.b3
ofl68h09_t46.b3
ofl68h09_t45.b3

Where did these reads go? (Hint: use *Find reads containing* on the Consed Main Window.) Give their contig number.
[Note: you may receive an error message indicating that the reads are coming from a different libraries. Ignore this (dismiss), but this may block your further use of Assembly View. You should be able to finish the Problem Set nonetheless.]

11.) These reads do not assemble in the correct position. A finisher should always investigate where the new data goes. These reads should go into the assembly downstream and very close to the oligo sequence that created them. For instance, ofl68h09_46.b3 should start near oligo 46. Pull read ofl68h09_46.b3 into it's own contig. What is it's contig number now?
(Use a right mouse pull down after click-and-hold with the cursor on the read in the Aligned Reads window to pull it out- see the Consed tutorial for more help.)

12.) Join ofl68h09_46.b3 into the main contig near oligo 46 using *Compare Contigs* (see Consed Tutorial). When you do *Compare Contigs* which base pairs do not match ? (They are represented by an X.) Which bases could you edit in ofl68h09_46.b3? Make these edits and join the contigs.

13.) Now a finishers job would be to compare the existing contigs and see if a join can be made. Both contigs have GA's on the ends. This GA sequence is called an ssr, Simple Sequence Repeat. It is believed that this type of sequence does not contain significant data, i.e., it does not contain genes. This is a mouse BAC and the finishing standards for mouse ssrs does not require the data to be base perfect. As long as there is less than 500 bps missing in an ssr gap, it is force joined and annotated. Of course the best possible join should be made, which may require the reads to be edited so that they will align correctly. Check out ofl68h09_46.b3 again. Edit this read changing base calls and/or doing *change consensus* where necessary and make the best possible join that you can. Is the region now resolved or will it need to be annotated? Are there any bases that you are unsure of?