

PRACTICE EXERCISE

POST-ASSEMBLY ANALYSIS

From the Bambus output of our Jumpstart assembly, we can see that while some of the gaps in the sample BAC have closed, some have not.

Which gaps closed?

Which did not?

Why not?

What would be our next steps?

Which gaps closed?

For small projects, like BACs, often contig size and scaffolding information are clues enough to determine which gaps closed and which did not. For larger data sets where manual inspection is tedious, another approach is to align the new contigs/scaffolds with the previous set.

From your home directory:

```
cd AssemblyData/BAC/closure/
```

Align new Contig 4 with the old BAC contigs:

```
secretsplit final.fasta
```

```
Fasta34 asm_4.fasta bac.fasta | more
```

Which gap does the new Contig 4 span? How large was the gap?

Which other new Contigs span gaps?

Which gaps remain, and why?

Find BAC ends by aligning contigs to BAC vector sequence:

We would not expect physical ends at the edges of our sample BAC to close, so let's identify which contig ends are the ends of our BAC.

Align the new contigs to the BAC cloning vector (pBACe3.6):

```
fasta34 ../pBACe3.6 final.fasta | more
```

Which contig ends are the ends of the BAC? (Hint: the BAC is cloned into the vector at an EcoRI site)

Find repeats that might be causing misassembly:

Repeats at contig ends often frustrate gap closure. Use repeatFinder and printrepeats to find repeats in the BAC contigs.

```
repeatFinder -in final.fasta
```

View the repeatFinder output using more:

```
more final.fasta.repeats
```

Use printrepeats to look for small tandem repeats:

```
printrepeats -s 50 final.fasta
```

View the graphical output of printrepeats using ggv:

```
ggv threshold50
```

Which gaps might be having assembly problems because of repeats? Do the repeats look like large-unit repeats or small-unit, tandem repeats?

Coverage:

Using the same techniques as in the coverage exercise, analyze the sequence and clone coverage for the new set of contigs. Are there fewer low coverage regions?

Other features of note? What are the next steps?

Sequencing gaps –

The only non-repetitive sequencing gap remaining is between new Contigs 3 and 4. Perhaps the sequencing gap reads failed? Perhaps a sequencing hardstop or DNA secondary structure is present?

Repetitive sequencing gaps most often call for transposon bombing of a gap spanning subclone. Make sure the mate pairs from the spanning subclone are sufficiently anchored in unique sequence.

Physical Ends –

Note the 'extra' ~300bp on new contig 5 (previously contig 7). Check in Consed which reads created this extension. The extension is covered by POMP PCR reads from TBCNZ02 (Pools 4 and 6). Perhaps further sequence walks on this PCR product will reach the adjacent contig and close the gap?

Can we orient the remaining scaffolds?

Because the two smaller scaffolds both contain one of the BAC ends, the large scaffold must fall between these two. The only question is orientation. Since there are only two

possibilities and four possible PCR products, a combinatorial approach might be appropriate.