Finishing *Drosophila Ananassae* Fosmid 2728G16

Kyle Jung
March 8, 2013
Bio434W
Professor Elgin
Abstract

For my finishing project, I chose to finish fosmid 2728G16. This fosmid carries a segment of the Drosophila Ananassae D Element near the centromere of Chromosome Three. The project initially contained a number of high quality discrepancies, an incorrect overlapping of reads, a few areas of low quality, and a gap between the two main contigs. Through three rounds of Sanger Sequencing and some manipulation of Consed, I was able to resolve the high quality discrepancies and areas of low quality. I was unable to resolve the gap between the two main contigs, most likely because of a hairpin secondary structure that prevented sequencing using both normal DNA sequencing methods and sequencing of PCR products. Further work is needed to resolve the gap area between the two main contigs.

Introduction

As a control region for the study of the Muller F element of the Dot Chromosome (fourth chromosome) for Drosophila Ananassae, a small euchromatic region near the centromere of chromosome three was chosen as a control region. Fosmid 2728G16 is part of scaffold 13337 located in a euchromatic region of the Muller D element. Finishing of the sequence of the D element will provide a control to compare against the F element. These comparisons will then allow us to better understand the F element’s unusual euchromatic and heterochromatic structure.

Initial Assembly & Tagging Clone Ends

Figure 1: Initial Assembly with Crossmatch
Initial assessment of the first assembly (Figure 1) of fosmid 2728G16 showed two main contigs with a gap between them and several smaller contigs (not shown) that did not belong within either of the two main contigs. There were a number of high quality discrepancies and a few low quality regions located near the gap. To start, I searched in the two main contigs to find the clone ends. I was unsure as to the orientation of the major contigs. I searched for the beginning and end of the fosmids by looking for the characteristic string of x’s (used to mark the vector sequence) that are labeled to the left of the 5’ contig end (Figure 2) and to the right of the 3’ contig end (Figure 3). In addition to the positioning of the string of x’s, the 5’ assembly end started with a read labeled fosmidendwgc2728G16.g1 and the 3’ assembly end was marked at the 3’ end of a read labeled fosmidendwgc2728G16.b1.

Figure 2: Left end of the assembly at which base 1 of contig 5 was labeled as the 5’ clone end.
High Quality Discrepancies

In contigs 5 and 6, there were several discrepancies between reads and the consensus sequence with a phred score greater than forty. These high quality discrepancies, as shown in Figures 6 and 7, were then analyzed individually. The majority of these discrepancies were caused by the phred miscalls at the start of the read. In these cases, I either changed the base to low quality or changed the pad to match the consensus base if there were sufficient reads with high quality to support this change. There were a few high quality discrepancies that indicated possible single nucleotide polymorphisms (SNPs). These cases had high quality data that corroborated the discrepancy with the consensus sequence. These instances were labeled with a comment tag stating that the base discrepancy could be a SNP. The one high quality discrepancy that suggested a growth difference turned out to be a case of Consed
incorrectly overlapping a group of reads onto the wrong part of the assembly. A table of all the high quality discrepancies and how each was resolved is provided below (Table 1).

<table>
<thead>
<tr>
<th>Contig</th>
<th>Position</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4238</td>
<td>Changed to low quality on read</td>
</tr>
<tr>
<td>5</td>
<td>14497</td>
<td>Changed to x b/c part of vector</td>
</tr>
<tr>
<td>5</td>
<td>19895</td>
<td>Base change from * -&gt; t</td>
</tr>
<tr>
<td>6</td>
<td>982</td>
<td>Tagged as possible SNP</td>
</tr>
<tr>
<td>6</td>
<td>1097</td>
<td>Changed to x b/c part of vector</td>
</tr>
<tr>
<td>6</td>
<td>1271</td>
<td>Tagged as possible SNP</td>
</tr>
<tr>
<td>6</td>
<td>1958</td>
<td>Told Phrap not to overlap at this position, pulled out three reads from main contig</td>
</tr>
<tr>
<td>6</td>
<td>7382</td>
<td>Changed to low quality on read</td>
</tr>
<tr>
<td>6</td>
<td>18121</td>
<td>Changed to low quality on read</td>
</tr>
</tbody>
</table>

Table 1: High quality discrepancies with their position and resolution.

The discrepancy at position 4238 is characteristic of high quality discrepancies located at the beginning of reads. At this position, a “*” was found on read 22470011D17.g1(D17.g1) when the consensus sequence indicates an “A” at this position. When evaluating the area surrounding the discrepancy, the peaks were neither equally spaced nor distinct (Figure 8). The other high quality reads
show distinct peaks that agree with the consensus and had equal spacing, indicative of a good quality read. Read 22470011D17.g1 at this position had not started producing good quality data, as shown by the areas of low quality surrounding this discrepancy. I chose to change the base at this position to low quality, but this base could have also been changed to a low quality “a” as well.

![Graph showing high quality discrepancy at position 4238 of the consensus](image)

Figure 8: The high quality discrepancy at position 4238 of the consensus. Read 22470011D17.g1 has low quality around the discrepancy because of its location at the start of the read.

The discrepancy at position 1271 on contig 6 is an example of the other common type of high quality discrepancy that occurred during finishing. Read 35232711PO3.b1(P03.b1) has an extra A at this position compared to the other reads. When evaluating the traces, the quality of the data for P03.b1 and other high quality reads, there was even peak spacing and distinct peaks with little background noise. Because of the high quality of the data at this position, this high quality discrepancy was marked with a comment tag labeling the discrepancy as a possible SNP or growth difference. Both possibilities would seem to be equally likely. Further reads are needed to determine which case is present.
Figure 9: Possible SNP or growth difference on read 35232711PO3.b1 at consensus position 1287.

**Low Quality Areas and Closing the Gap**

When trying to resolve the low quality areas in main contigs, I found that the low quality areas showed a pattern of being located near the gap between contigs 5 and 6. For Contig5, the low quality area consisted of the last 69 base pairs at the 3’ end (Figure 10a). For contig 6, the areas of low quality spanned the first 800 base pairs at the 5’ end(Figure 10b). The clustering of the low quality areas around the main contig gap allowed me to try to resolve the low quality areas in my project while creating reads to resolve the gap between contig 5 and 6.

Figure 10a: Areas of low quality on contig5.
Figure 10b: Areas of low quality on contig 6.

When designing primers to resolve the gap between contig 5 and 7 (previously contig 6), I followed a number of guidelines. These included limiting the primers to 17-28 bases in length, a GC content of 50-60%, a Tm of 55C-80C, no more than 2 or 3 C’s or G’s at the 3’ end, at least one C or G at the 3’ end, and annealing site of the primer to be as close to the gap as possible. For the first round, I decided to be conservative and start the reactions further back than suggested by Autofinish so that the reaction did not start in a region of low quality (Table 2, Figure 11, and Figure 12).

<table>
<thead>
<tr>
<th>Contig</th>
<th>Start-End</th>
<th>Direction</th>
<th>Tm</th>
<th>Oligo</th>
<th>Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>923-903</td>
<td>3’-&gt;5’</td>
<td>59C</td>
<td>cgccgtgtcaagggcttacctc</td>
<td>dGTP, BigDye, 4:1</td>
</tr>
<tr>
<td>5</td>
<td>20520-20520</td>
<td>5’-&gt;3’</td>
<td>58C</td>
<td>cagggtctttgaccaagt</td>
<td>dGTP, BigDye, 4:1</td>
</tr>
</tbody>
</table>

Table 2: Sequencing Primers used for first round of sequencing.

Figure 11: Autofinish suggested primers to resolve the low quality areas.
Figure 12: Primer placement for the 1st round oligos (red) and the autofinish suggested oligos (blue).

This first round of sequencing proved to be ineffective at resolving the low quality area or the gap between the main contigs. The reactions seemed to have started correctly and then began to fail when the polymerase reached the areas of low quality. This problem occurred in the past, as shown by the reads from the past two years. A second round of sequencing was done with primers located directly next to the low quality areas (Simon Hsu) and about 300-500 bases away from the low quality areas (Table 3 and Figure 13). These reactions also proved unsuccessful. Reads from the trace archive were then evaluated in an attempt to gain some added data in the areas of low quality. These reads also proved unable to resolve the low quality areas or the gap.

<table>
<thead>
<tr>
<th>Contig</th>
<th>Start-End</th>
<th>Direction</th>
<th>Tm</th>
<th>Oligo</th>
<th>Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1317-1336</td>
<td>3’-5’</td>
<td>56C</td>
<td>aagctctgcttaactgca</td>
<td>dGTP, BigDye, 4:1</td>
</tr>
<tr>
<td>5</td>
<td>20397-20415</td>
<td>5’-&gt;3’</td>
<td>55C</td>
<td>gagcttggaggcttaga</td>
<td>dGTP, BigDye, 4:1</td>
</tr>
</tbody>
</table>

Table 3: Primers created 300-500 bp away from low quality area for second round of reactions.

Figure 13: Placement of primers for second round of sequencing reactions. Primers closest to the gap were ordered by Simon Hsu.
In an attempt to resolve the size of the low quality area and gap, a combination of PCR and sequencing was used. A PCR reaction was performed to try to produce DNA fragments containing the low quality and gap region for sequencing. The PCR reactions for my project used the two pairs of primers that were used in the first and second rounds of sequencing. The products are shown in lanes 1 and 2 (Figure 14). The reads that correlate to the primers began at the correct positions on the contigs and did provide high quality data leading up to the gap. The two primers on contig 7 began to sequence around position 1300, and provided high quality data until base 800. The two primers on contig 5 primed produced sequencing reads starting positions 20,420 and 20,530 and provided high quality data until about position 20,800. This indicated that the primers did initially anneal at the correct locations and the PCR did initially amplify out the correct products. The main PCR products of about 1300 and 1400 bases (lanes 1 and 2 respectively from Figure 14) provided an estimate of the main contig gap size to about 700 to 800 bases.

It was believed that there was enough correct PCR product compared to the other incorrect products to provide a strong enough sequencing signal for high quality data. Unfortunately, the Sanger Sequencing reads from the PCR failed once the polymerase reached the areas of low quality. I was therefore unable to resolve the main contig gap with any of the reads.
Figure 14: Gel electrophoresis run of the PCR products that were sequenced.

**Restriction Digests**

Despite our inability to resolve the main contig gap, the restriction digests for the assembly supported the arrangement of the major contigs. The restriction digests using EcoRV, EcoRI, SacI, and HindIII provided a size range of the gap between contigs 18 and 7. By examining the discrepancies between the real and *In Silico* digest fragments, I was able to range the gap to be from 41bp to 991 bp in size.
Figure 17: a) EcoRI digest with scaffold arrangement of 18-7. The “g” indicates a gap-spanning fragment.

b) Text output of EcoRI restriction digest.

The restriction digest fragment sizes of both the real and In Silico EcoRI digests (Figure 17a and 17b) support the arrangement of contig 18 proceeding contig 7. The discrepancies between the fragment sizes are within the error range of gel band sizes. The only inconsistencies present involve fragments below 1kb in size. The difference between the real and In Silico fragments that span contig 18 and 7 indicates a minimum gap size of 122 bases. I then estimated the low quality areas around the gap to be 869 bases in size and subtracted it from the In Silico fragment size to estimate the maximum size gap. The EcoRI digest indicates a maximum size of 991 bases for the gap between contigs 18 and 7.

The EcoRV digest also supports the current assembly arrangement. The real digestion fragments correspond to the In Silico digest fragments. The two bands below 1kb are vector fragments and the fragment that spans the gap between contigs 18 and 7. However, there is no real digest fragment that
corresponds to the gap *In Silico* digest fragment because it is smaller than 1kb in size. Thus, I was unable to estimate the gap size from this digestion since there could be a real digest fragment for the gap that was not reported.

Figure 19: a) *HindIII* digest with scaffold arrangement of 18-7. The purple band indicates that two fragments are located at the same position. The “g” indicates a gap-spanning fragment. b) Text output of *HindIII* restriction digest.

The *HindIII* digest has one discrepancy between the real and *In Silico* fragments that include the gap between contigs 18 and 7. The real *HindIII* digest fragment is 109 bases larger than the *In Silico* *HindIII* fragment. This variance indicates that the minimum gap size is 109 bases and the maximum gap size is 978 bases, assuming the low quality area is about 869 bases in size. The discrepancy can be resolved once better sequencing data is obtained to resolve the low quality areas affected by the hairpin structure. The other discrepant bands are below 1kb and do not have corresponding real fragments because of the small size.
The *SacI* digest has no discrepancies and supports the current arrangement of the contigs. There is one band below 1kb, but this band was recognized as part of the vector. The real digest fragment is 41 bases larger than the *In Silico* fragment, indicating a minimum gap size of 41 bases. Using the 869 base size estimation of the low quality areas of contig 18 and 7, the maximum gap size from this digest is 910 bases.

**BLAST**

With the assembly as contiguous as possible, I then ran a BLAST search to check for any contamination or lateral transfer of DNA into the project. For contig 18, no matches were found, suggesting that the contig was free from contamination or any other lateral transfer of DNA. When looking at contig 7, 131 hits were found. Most of these clustered around the 10,000 to 13,500 bp region of the contig, but these matches had a maximum identity (similarity in sequence) of 80% and below. These hits were spread across a variety of bacteria including *Streptomyces griseus* and *Mycobacterium tuberculosis*. However, the digest suggests that these hits were not contamination from the sequencing reactions. The lack of discrepant fragments between the real and *In Silico* digests supports a lack of contamination.
Figure 21: Blast search results for contig 7.

**Discussion**

The failure of the normal sequencing reactions and the PCR/sequencing reactions led me to believe that there is a secondary hairpin (stem-loop) structure (Figure 15) preventing the polymerase in the sequencing reaction from correctly functioning. When performing a Search for String on areas of high quality nearby the low quality areas near the main contig gap, there were complementary matches on the other main contig. These complementary matches indicated the presence of a secondary hairpin structure. The hairpin causes an abrupt stop in sequencing that could lead to the polymerase disassociating with the original DNA strand and re-associating on a different DNA strand. The polymerase could then begin replication again, leading to sequencing of unwanted products.

Figure 15: Stem-loop structure that is believed to be preventing sequencing of the low quality regions and the gap. Contig 18 was previously contig 5.

A potential way to eliminate the problem posed by the hairpin secondary structure, in the future, will be to use nanopore sequencing (Figure 16) to avoid DNA transcription. This technology unravels the DNA into single strands using a helicase and passes one of the strands into a specialized protein with a small hole (nanopore). The helicase might be able to remove the hairpin secondary structure before the DNA is inserted into the nanopore, resolving the issue of sequencing within the low quality and gap region. The flow of the ions from the DNA strand creates a current, which is altered by each base blocking the flow of ions to a certain degree. These changes in current are then analyzed to produce sequencing data without the use of DNA Polymerase.
Future Directions

The unresolved areas in this project are the low quality region around the gap, the gap itself, three paired reads too far apart from each other, and one unsupported single chemistry region towards the 3’ end of contig 7. Considering that there are low quality data on the 3’ end of contig 18 and the 5’ end of contig 7, the three mispaired reads are only about 30-40 bases further apart than they should be and will resolve themselves once the low quality data around the gap is elucidated. The one unsupported single chemistry region should be within the overlap between projects. This area should be within the area that is overlapped by the next project, which should increase the confidence of our consensus in this area.
Figure 22: The final fosmid assembly with crossmatch.

Acknowledgements

I would like to thank Professor Elgin, Professor Schaffer, Wilson Leung, Lee Trani, Jenn Hodges, Harry Senaldi, and all the people at the Genome Institute for their help in this project.
References


