Finishing *Drosophila grimshawi* Fosmid Clone DGA26H21

Andrew Stein
March 23, 2009
Abstract

In this paper, I discuss my work on finishing *Drosophila grimshawi* fosmid clone DGA26H21. My initial consensus sequence was split over four contigs, and the major problems included three gaps and many low quality regions. Over the course of three rounds of sequencing reactions, I was able to significantly improve the consensus sequence of this fosmid. I closed two of the three gaps, leaving my final project with two contigs and only one gap. Furthermore, although I was unable to close one of the gaps, I extended over 500 base pairs of sequence into the gap. This will make it much easier to close this gap in future work. Additionally, the only low quality regions present within my contigs are within the gap region. In short, the consensus sequence of fosmid clone DGA26H21 has been significantly improved, but it still contains one gap.

Introduction

In Biology 4342, the goal is to finish and annotate the fourth (dot) chromosome of many species of *Drosophila*. The dot chromosome is present in most species of *Drosophila*, and it contains a very high density of repetitious elements, has no meiotic recombination, and exhibits other characteristics of heterochromatin. Despite these observations, the dot chromosome still contains about 80 genes that are clustered in a small region, and these genes appear to be conserved in many species of the *Drosophila* genus. Although the same genes are expressed across these varying species of *Drosophila*, the chromatin structure of these organisms can be quite different. Studying the dot chromosome could be very informative in terms of the evolution of chromatin structure within the dot chromosome. This year, Biology 4342 students are focusing their work on finishing the dot chromosome of *Drosophila grimshawi*. Each student hopes to bring up the consensus sequence of their fosmid to a Phred score of 25 for double stranded regions and 30 for single strand regions. My specific portion of this project was to finish fosmid clone DGA26H21. In this paper, I highlight the steps I took to finish my fosmid.

Project Work Flow

*Initial Assembly Assessment*

To begin work on finishing fosmid DGA26H21, I analyzed the initial assembly view as shown below in figure 1. There were a total of four contigs and three gaps within my fosmid. Also, discrepant forward-reverse pairs were present between contigs 2 and 4 since they were located too far apart, and crossmatch indicated repetitive sequences shared between the two contigs. Consistent forward-reverse pairs were present between contigs 5 and 3 as well as between contigs 2 and 4.
Initial Work

Based upon my initial assembly view, I attempted a force join between contigs 4 and 2 as my first line of attack. I ran a search for string using twenty unique nucleotides from the end of contig 2, and they matched up perfectly with the beginning of a region from contig 4. After aligning the sequences (figure 2), it was evident that a force join was reasonable. However, as figure 2 shows, there were some discrepant bases present in the alignment.

To ensure that the force join was reasonable in light of the apparent discrepant bases, I looked at the trace windows for both of the contigs (figure 4). I noticed that within these regions contig 2 had poor sequence reads. Therefore, I felt comfortable creating the join between contigs 2 and 4. These contigs were now represented as contig 6 in my assembly view.
Next, I defined the beginning and end of my fosmid clone to make sure my contigs were ordered in the correct manner. Using the reads containing “.dga”, I was able to find vector sequences present at the start of contig 6 and the end of contig 3. Therefore, I tagged the end and beginning of the vector sequences and reordered my contigs as shown below in figure 5.

![Figure 5: Assembly view after force join and determination of fosmid clone ends](image)

**Analysis of Remaining Gaps**

After closing the first gap, I wanted to determine the gap sizes of the two remaining gaps. In order to analyze this, I looked at the real vs. *in silico* digests of my fosmid.

![Figures 6 and 7: Digest of my fosmid with HindIII along with text output information](image)

The *HindIII* and *EcoRV* digests proved to be the most informative and showed similar information. As figures 6 and 7 display (arrows and box), there was a gap of around 800 base pairs between contigs 5 and 6. Also, the gap between contigs 5 and 3 was around 100 base pairs. This information indicated that I needed to design oligos to extend my consensus sequence into these gaps. Although it appears that the real digest has an unmatched band at 19249 base pairs, I deduced that this was a miscalled band since the actual gel digest (figure 8) showed a very thick band here that could have been miscalled as two separate bands.
Comparison of Round One Oligos and Autofinish

It was apparent from my initial work that I needed new sequencing reactions to close the gaps and resolve low quality regions (table 1). Fortunately, most of the low quality regions were near the gaps, so the oligos I designed to close the gaps also covered these low quality areas. However, there was one section of contig 6 where I needed to design specific oligos to resolve low quality. I ordered all oligos using 4:1, dGTP, and Big Dye sequencing chemistries in order to increase the read depth across the gap regions.

After choosing my oligos, I compared my selections to those chosen by Autofinish (table 2). The first, second, and fourth primers called by Autofinish were similar to oligos that I called to span the two remaining gaps. Moreover, the fifth and sixth oligos selected by Autofinish would help close the gap between contigs 5 and 6 as well as resolve the low quality region present within contig 6. I called three oligos for these purposes. Lastly, the third primer that Autofinish
called would resolve single strand chemistry in contig 5. However, this single strand portion has a Phred score over 30, so there was no reason to call new oligos in this region.

**Incorporation of Round One Oligos**

Initially, I incorporated all of the new reads from my first round of oligos using the “Add New Reads” feature of Consed. Oligo DGA26H21.4 successfully sequenced into the gap between contigs 5 and 3. Although this read was of high quality, it was discrepant with all the other reads in this region. Upon further investigation, it was apparent that the other reads were all low quality, so I changed the consensus sequence to match the sequence from this new read (figures 9 and 10). A similar situation arose with the reads from oligo DGA26H21.1. However, as in the other situation, the earlier reads were of low quality, so I changed the consensus sequence to match the high quality reads from my new oligo.

Overall, the reads that were incorporated from the first round of oligos extended the consensus sequence into the gap regions. Therefore, I decided to rerun phredPhrap to see if any of my gaps would be closed. I was able to close the gap between contigs 3 and 5. These two contigs were now represented as contig 3 in my new assembly view (figure 11).
Although phredPhrap closed the gap between contigs 5 and 6, it also created a new gap between contigs 1 and 2. Fortunately, this was the same gap that I had closed previously (between contigs 2 and 4 in initial work). Therefore, I performed a search for string in the same manner as before and based upon the alignment of the sequences I was able to close this gap again (figure 12).

![Figure 12: Assembly view after force join](image)

I ran crossmatch on the new assembly to see if there were any repetitious regions present between contigs 3 and 4 that would allow me to perform a force join. Crossmatch did not return any information along these lines. I also attempted to find similar nucleotides between the two contigs by using “search for string.” My searches came up dry, and there was no way for me to join these contigs with my present reads. Thus, my main concern with this assembly was the gap present between contigs 4 and 3, so I checked the digests to see if I could learn anything more about the gap size. The *HindIII* digest provided interesting information (figure 13).

![Figure 13: HindIII restriction digest text output information](image)

From the text output data, it appeared that my contigs actually contained too much sequence information. The *in silico* band of 9164 base pairs should actually be at 8730 base pairs. This indicates that my contigs have around 400 nucleotides of extra information. Therefore, this made me confident that I had the necessary sequence information to create a force join between these two contigs; however, the reads were of too low quality for a force join to be feasible.

**Round Two Oligos**

After reviewing the digest data, I realized I could not improve my consensus sequence any further. Therefore, I proceeded to generate my second round of oligos. I searched contigs 3 and 4 for low quality regions to re-sequence. Also, I designed new primers to close the gap between contigs 3 and 4. Due to the fact that my first round of oligos had extended the consensus sequence of my contigs into the gap, I was hopeful that these new reads would allow me to close
the gap. The oligos presented in table 3 were ordered using all three chemistries to increase read depth across the gap and low quality regions.

<table>
<thead>
<tr>
<th>Sequence (5’-3’)</th>
<th>Name</th>
<th>Orientation</th>
<th>Purpose</th>
<th>Position</th>
<th>Contig</th>
</tr>
</thead>
<tbody>
<tr>
<td>tgggccccacctga</td>
<td>DGA26H21.9</td>
<td>--&gt;</td>
<td>Close gap contigs 4/3</td>
<td>14134</td>
<td>4</td>
</tr>
<tr>
<td>ccggagcccgccctag</td>
<td>DGA26H21.10</td>
<td>--&gt;</td>
<td>Close gap contigs 4/3</td>
<td>14292</td>
<td>4</td>
</tr>
<tr>
<td>cctctctctccctgcaata</td>
<td>DGA26H21.11</td>
<td>&lt;--&gt;</td>
<td>Low quality</td>
<td>14161</td>
<td>4</td>
</tr>
<tr>
<td>tcaaaaaattccccgagagata</td>
<td>DGA26H21.13</td>
<td>--&gt;</td>
<td>Low quality/single strand chem</td>
<td>12919</td>
<td>4</td>
</tr>
<tr>
<td>gcttgccgctctctcttt</td>
<td>DGA26H21.12</td>
<td>&lt;--&gt;</td>
<td>Close gap contigs 4/3</td>
<td>659</td>
<td>3</td>
</tr>
<tr>
<td>atttttcgccgagctt</td>
<td>DGA26H21.14</td>
<td>&lt;--&gt;</td>
<td>Low quality</td>
<td>1070</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3: Round two oligos ordered

When I received the reads from the second round of oligos, I incorporated them by using “Add New Reads.” Out of the 18 reads, only six were actually incorporated into my contigs. However, none of these reads proved to be very informative. Most of them were of very low quality and did not add to the consensus sequence nor read depth of my contigs (figure 14).

![Figure 14: Low quality traces from round two oligos (reads from oligo DGA26H21.9)](image)

Overall, the second round of oligos was unsuccessful and did not help me resolve any low quality regions nor make progress on closing the gap between contigs 4 and 3. I began to think that there must be some sort of secondary structure or other difficult region to sequence within this gap that caused all of my primers to fail. It is also possible that my primers were not specific enough and annealed at many locations, leading to poor sequence reads.

**Round Three Oligos**

My main goal with the third round of oligos was to design primers that would allow me to close the gap between contigs 4 and 3 (table 4). I made the primers as specific as possible by performing a search for string on the eight base pairs at the 3’ end of each oligo to ensure that they did not match elsewhere. I designed two new sequencing primers to close the gap along with PCR primers that would also be used as a sequence template for the gap. Furthermore, I designed oligos to resolve a single subclone region in contig 3 and a polymorphism in contig 3 at base pair 18545 (figure 15).
I incorporated the reads from my third round of oligos by using the “Add New Reads” function. First, I noticed that the single subclone region that was in contig 3 had been eliminated by my new reads. Additionally, the read depth was increased at the site of the putative polymorphism. Based on all of the traces, it appeared that this polymorphism was actually due to an *E. coli* cloning problem since only one of the traces was different from all of the others (tagged this base with a comment).

Unfortunately, the primers called to close the gaps were unable to accomplish their ultimate goal. I tried joining contigs 4 and 3 both by using the search for string function with unique nucleotides from each contig and rerunning phredPhrap; however, neither of these methods proved to be successful. Therefore, I decided to analyze the traces at the end of contig 4 (figure 16) and beginning of contig 3 (figure 17) to see if any overlap was present. It is evident from these traces that there is a large repeat of AGGAGG. This repeat makes it difficult for me to sequence across my gap and join these contigs. The beginning of the contig 3 sequence contains this repeated region, but it is flanked by unique sequences that inhibit the force join (figure 17). Furthermore, the sequences from contig 4 extend well into this repeat, but there are not any unique sequences present at the end of contig 4 that are needed to create the force join.
Although I failed to close the gap, I was able to extend the consensus sequence from contig 4 into my gap (about 500 base pairs). I accomplished this by manually changing the consensus sequence to match that of oligo DGA26H21.18 (PCR oligo) since this is a high quality read whereas all the others in the same region are low quality (figure 18).

![Figure 18: Traces showing high quality read from oligo DGA26H21.18 and other low quality reads](image)

**Additional Low and High Quality Discrepancies**

To improve the consensus sequence of my fosmid, I decided to examine any high quality discrepancies present with contigs 4 and 3. From contig 4, there are high quality discrepancies from base pairs 2368-2402. However, all of these high quality discrepancies are from read 02366640J04.b1. This read contains double sequence information, and therefore it is unreliable.
I decided to remove this read and place it into its own contig. Furthermore, Consed showed that base pairs 2031-2108 contained unaligned high quality sequences. In reality, there is a trace present within this region that is of low quality. I placed a tag on this region to note that Consed had improperly noted the presence of unaligned high quality reads. Finally, there are two additional high quality discrepancies present at base pairs 3483 and 6882 in contig 4. Both of these discrepancies are a result of a miscalled base at the start of the read (figure 19), so I made comments to note this. In contig 3, there is a high quality discrepancy at base pair 8191 similar to those just described in contig 4, so it was resolved in the same manner.

Finally, in both contigs 3 and 4, I noticed a few low quality bases embedded within high quality regions. In order to resolve these areas, I checked the traces, and the traces made me confident that the bases were actually high quality (figure 20). Therefore, I changed the consensus sequence at these locations to match the high quality bases. This arose in contig 4 at base pair positions 11633, 13595-13599, and 14099. In contig 3 this was evident at base pairs 726-727.

**Restriction Digest Analysis**

In order to see if my final assembly was reasonable, I decided to view the restriction digests (figures 21 and 22). The *HindIII* and *EcoRV* digests match up very well aside from a discrepant
band present in the middle of *HindIII* and at the top of *EcoRV*. Importantly, the text output data from both digests show that this discrepant band includes the gap region between contigs 4 and 3. The other discrepant band present in the *EcoRV* digest around 4000 base pairs does not concern me because the text output data shows that these bands are only off by about 70 base pairs. Also, all of the matching bands have consistent small shift calibration errors present. Considering the fact that the rest of the digest aside from the gap matches up perfectly (when taking into account these small shift calibrations), I am confident that my sequence is indeed correct. However, in order for the digests to match up perfectly overall, the gap between contigs 4 and 3 needs to be closed.

![Figures 21 and 22: *EcoRV* (left) and *HindIII* (right) digests](image)

**Final Comments**

In the end, I was able to make significant progress on finishing fosmid clone DGA26H21. My final assembly consists of two contigs with one gap remaining (figure 23). Although I was unable to close this gap, my sequence reads extended over 5000 base pairs into the gap. Most of the additional high quality sequence information that has been extended into the gap is from the end of contig 4. It was much more difficult to sequence into the gap from contig 3.

![Figure 23: Final assembly view of fosmid clone DGA26H21](image)

In the final fosmid, there are low quality regions at the end of contig 4 and the beginning of contig 3. Both of these areas are presumably part of the gap region. Single subclone regions and
single strand regions of low quality are also present within these areas. Additionally, there is a region in contig 4 that only contains sequence reads from my PCR reaction (tagged this region with a comment). Furthermore, the consensus does not contain any X’s or N’s; however, both contigs contain a mononucleotide run of A’s. These regions are not problematic since the traces below (figures 24 and 25) clearly show the presence of a properly called mononucleotide run.

![Figures 24 and 25: Mononucleotide run present in contigs 3 and 4](image)

All high quality discrepancies and putative polymorphisms have been appropriately tagged. Lastly, a BLAST analysis was run, and it showed that neither contig 4 nor contig 3 contain any bacterial contamination (from *E. coli* DNA for example). Overall, the fosmid has been thoroughly scanned and tagged in order to provide detailed information about its present status.

**Future Work**

Further work on this fosmid is needed to close the gap between contigs 4 and 3 and bring the entire consensus sequence up to high quality. Based upon the fact that all of the low quality sequence information comes from the oligos I called to close this gap, it could be beneficial to survey these reads and remove them if the traces show very poor read quality. Additionally, more reactions need to be called off the end of contig 4 since this is the area where I was able to successfully sequence into the gap. Either new oligos could be called off of the extended regions of contig 4, or oligos could be called off of the PCR fragment since it covers the gap. Overall, I am certain that I have extended far into the gap, and the addition of new reads from the end of contig 4 will most likely close the remaining gap and fix the low quality regions within fosmid clone DGA26H21.