Karen Zhou
Biology 434W
March 6, 2013

Finishing Fosmid 8283H01

Abstract

In this project, I aimed to finish the fosmid clone 8283H01 from the third chromosome of *Drosophila ananassae*. The initial assembly was problematic, as there were gaps, multiple high quality discrepancies, and mis-assemblies. By ordering new reads and making joins where possible, these issues were resolved and a final assembly with no gaps was achieved. These initial issues were found to have been caused by reads that likely did not belong in the project, but were mistakenly incorporated due to having a similar sequence with repetitive regions.

Introduction

In our study of chromosome structure, we have found that DNA is generally found in one of two forms – heterochromatic and euchromatic. Heterochromatic DNA is tightly packaged and not available to transcription machinery, and is commonly found in centromeric regions. Euchromatic DNA is not silenced and is therefore able to be transcribed. The fourth chromosome, also known as the dot chromosome, of *Drosophila ananassae* is unique in that it possesses both heterochromatic and euchromatic properties. Therefore, the dot chromosome of *Drosophila annanassae* is of research interest in order to better understand the role of DNA packaging in gene transcription and expression. In this study, we finished both clones from the dot chromosome and a presumed control, the third chromosome.

Initial Assembly

The initial state of my project can be seen in Figure 1 below. There were five major contigs, the longest of which were contigs 4 and 5. Contig 1 contained a single read. The red lines below the contig bars represent inconsistent forward/reverse pairs, as seen between contigs 4 and 5, contigs 2 and 4, and contigs 3 and 4. These suggested that there is a mis-assembly in the project. There was also a gap between contigs 4 and 5, with consistent mate pairs spanning the gap.

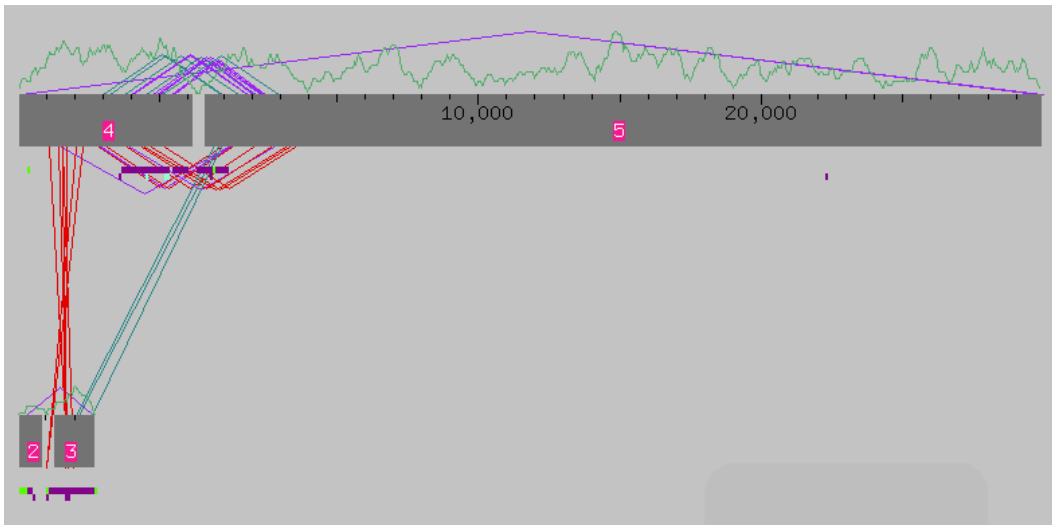

Figure 1. Initial Assembly View of the project.

I first searched for the clone ends. The left end was found on the left end of contig 4, and the right end was found on the right end of contig 5. I changed any vector sequences to x's, and added clone end tags to both clone ends. Cross-match was run on the project. The assembly view with cross-match is shown in Figure 2.

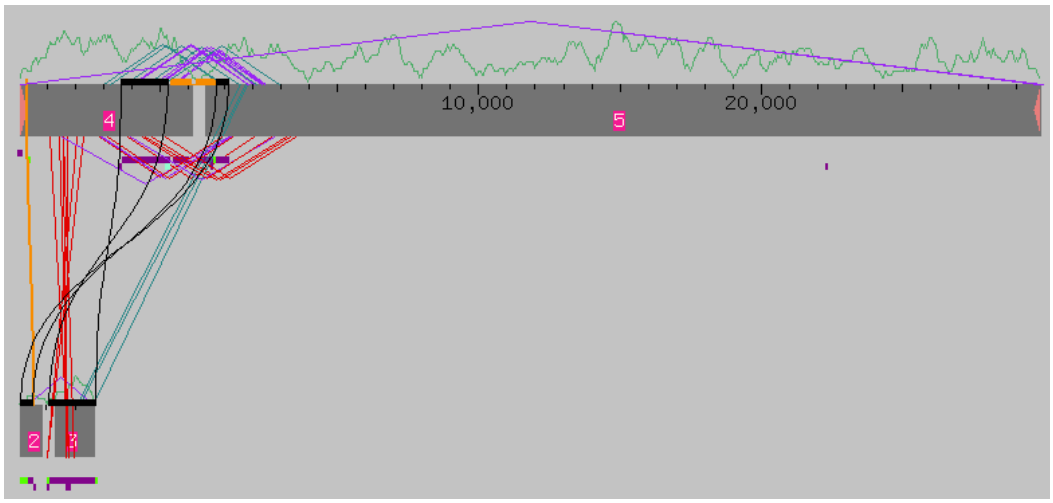

Figure 2. The results of calling cross-match on the assembly.

As we can see from the figure, there are inverted regions that matched contig 3 to contig 4. There is also an inverted region that matched contig 2 to a region near the beginning of contig 5. There is a match that connects the end of contig 4 to the beginning of contig 5.

High Quality Discrepancies

Next, I navigated to the high quality discrepancies. These are shown in the figure below.

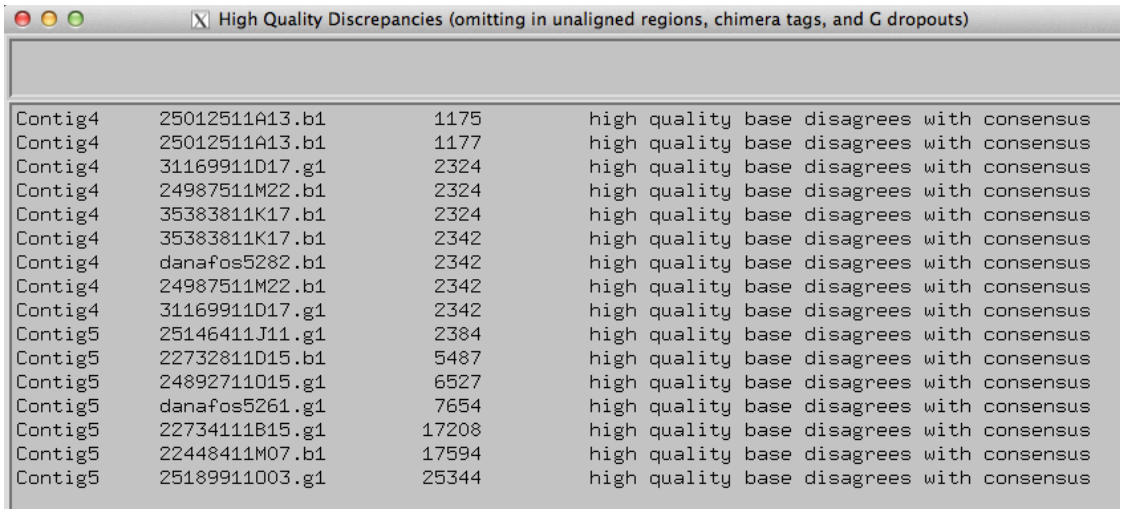

Figure 3. High quality discrepancies in the main assembly.

The high quality discrepancies found in contig 4 at bases 1175 and 1177 were within a vector sequence, GAATTC. These bases were at the beginning of the same read. These bases, and those to the left of these positions, were marked as vector sequence in order to resolve these high quality discrepancies.

Next, I examined the high quality discrepancy found in multiple reads at base 2324 in contig 4. About half of the reads covering this region had a T at this position, whereas the other half had a C. Examination of the traces of the reads showed that the quality of the reads at this position was good, so the discrepancy is not due to contamination or poor quality. This implies that the high quality discrepancies at this position may be due to a polymorphism, due to the 1:1 ratio of the different bases. Some reads may not belong in the project and may have been errantly associated with this project. This position was marked to tell phrap to not overlap discrepant reads.

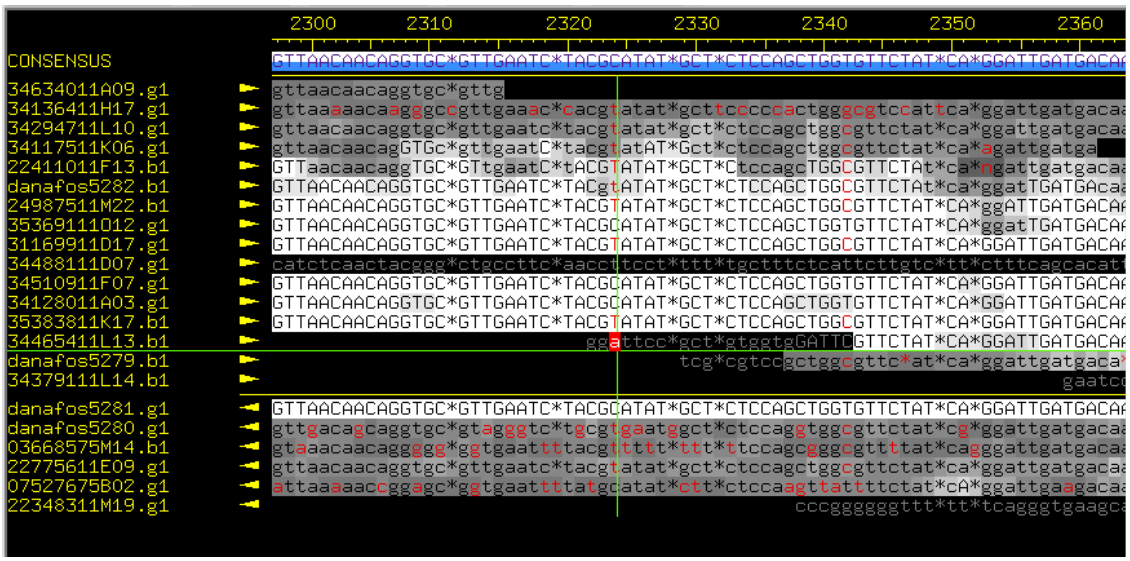

Figure 4. Aligned Reads at position 2324 of contig 4, where multiple high quality discrepancies were found.

The multiple high quality discrepancies at position 2342 in contig 4 were examined next. Much like at position 2324, approximately half of the reads had base T, whereas the other half had base C. When the traces of the reads were examined, the quality was good. This position was also marked to tell phrap to not overlap discrepant reads. Position 2342 was within the same region as 2324 that was tagged as being repetitive. The reads that were discrepant from the consensus at base 2324 were also discrepant at position 2342. I hypothesized that these reads may not belong in my project, due to the multiple high quality discrepancies, and given the presence of a repetitive region in these reads. It was also possible that these reads were due to polymorphisms, since approximately half of the reads in this region showed one base, while the other half showed another base. However, there were multiple inconsistent mate pairs in these regions that were anchored in repetitive regions, so I believed that it was more likely that reads were errantly pulled into my project.

In contig 5 at position 2384, there was only one read discrepant from the consensus. The trace showed that the bases called were evenly spaced and the peaks were smooth. To resolve this discrepancy, I removed the read from the contig and placed into its own contig. Since there was only one discrepant read at this position, the discrepancy may be due to a growth difference.

At position 5487 of contig 5, the discrepant read had a quality of 40 at this postion. This was lower than the quality of the base in other reads, which had phred scores of about 68. When I examined the trace, I found that the spacing was irregular at this position. Furthermore, the base called was not a single peak. I deduced that one A had been called

instead of the two A's that should have been called. I changed the discrepant base in this read to an A in order to resolve the high quality discrepancy.

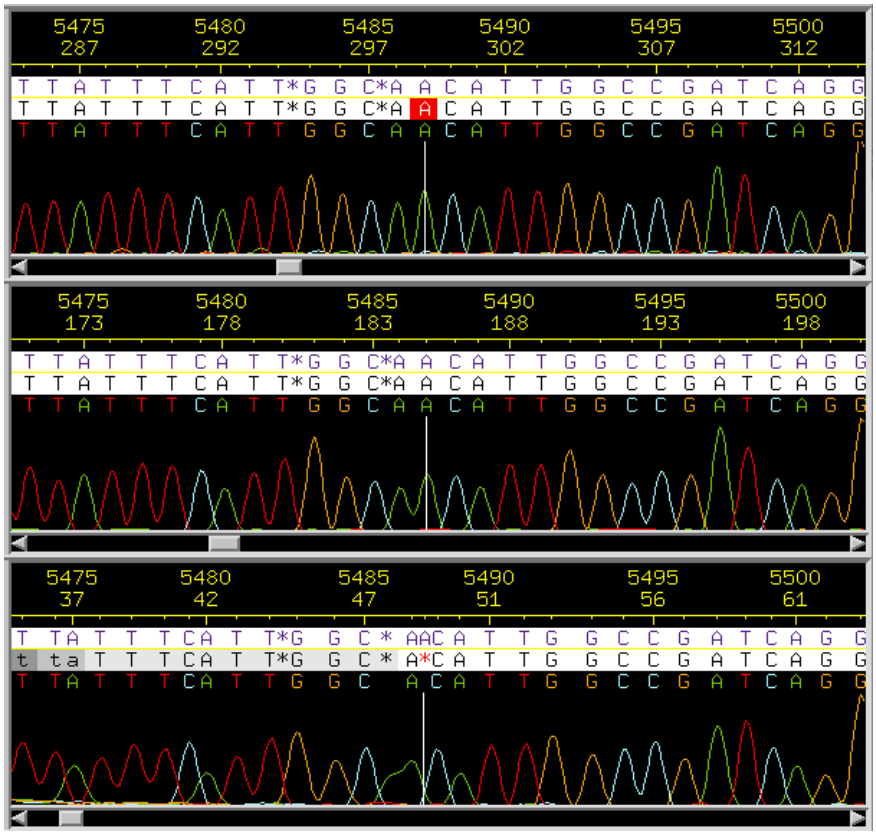

Figure 5. Comparison of the traces of reads at position 5487, with the discrepant read on the bottom.

The rest of the discrepancies are detailed in the table below, as well as how they were resolved.

Contig	Position	Problem	Resolution
5	6527	Vector sequence at end of read	Bases on discrepant read changed to vector sequence (x)
5	17594	Vector sequence at end of read	Bases on discrepant read changed to vector sequence (x)
5	7654	Number of bases miscalled	Base changed manually on read
5	17208	Number of bases miscalled	Base changed manually on read
4	25344	Number of bases miscalled	Base changed manually on read

Table 1. High quality discrepancies in the assembly.

After I finished examining and resolving the high quality discrepancies in my project, I ran Miniassembly on the contigs that contained the discrepancies, contigs 4 and 5. This created new contigs. The reads that had been discrepant at positions 2324 and 2342 in contig 4 were placed into their own contig, contig 8. The remaining reads from contig 4 became contig 9.

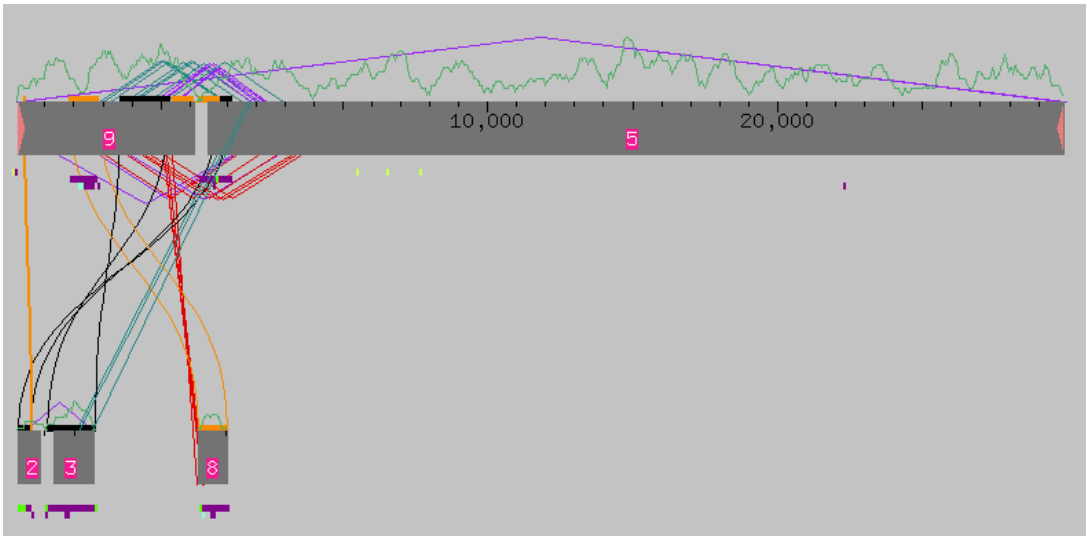

Figure 6. Assembly View after addressing high quality discrepancies.

I then reoriented the scaffold containing contigs 2 and 3.

Examining the Digests

In order to begin resolving the gap between contigs 9 and 5, I viewed the digests of my main scaffold, contigs 9 and 5. The four digests chosen were *Eco*RV, *Eco*RI, *Sac*I, and *Hind*III. When these digests were viewed, the *in-silico* digest matched up fairly well to the real data. However, in both *Eco*RV and *Hind*III there was one major difference between the *in-silico* digest and the real digest. As seen in Figure 7, the *in-silico* digest showed a band that was approximately 900 – 1000 bases longer than the real digest. This band corresponded to the segment of the scaffold where contig 9 and contig 5 were joined.

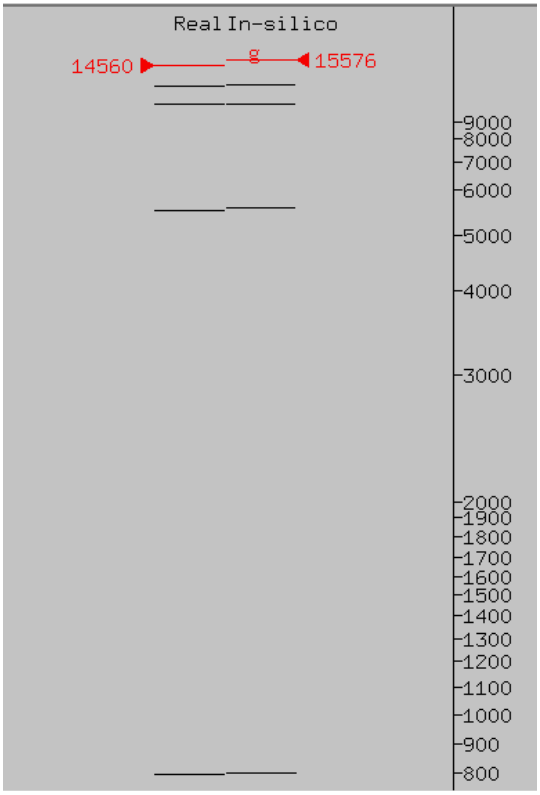

Figure 7. The *Eco*RV digest of the scaffold of contigs 9 and 5.

Looking at the results of the digests, I concluded that my assembly had 1 kb too much data. The repeat matching the end of contig 9 to the beginning of contig 5 was about 700 bases long. I hypothesized that joining the two contigs might be a solution to addressing the issue of the 1 kb extra data found by examining the digest. In addition, the gap in the assembly may be resolved.

Inconsistent Mate Pairs and Resolving the Gap

After looking at the digests, I began examining the inconsistent mate pairs found throughout the project. I removed the reads with the inconsistent mate pairs, and ran Miniassembly on those reads. Some of these mate pairs were placed in the same contig, but others were placed into their own contig. The results of cross-match can be seen in Figure 8 below. There were many matches between contigs. I examined these sequence matches to determine whether a join could be made between them.

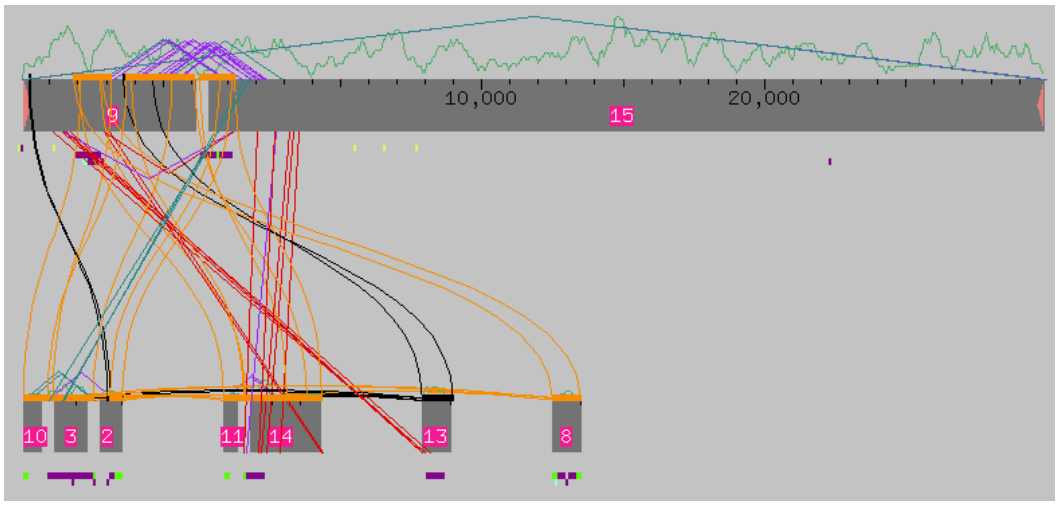

Figure 8. Assembly View after removing inconsistent mate pairs from main scaffold.

After examining sequence matches, I decided to join contigs whose alignments showed no high quality discrepancies. In this manner, I joined contig 9 to 10, contig 8 to 11, contig 2 to 15, and contig 3 to 13.

Finally, the two major contigs of the project's main scaffold, contigs 9 and 15, were joined. To verify that the join was good, I again examined the digests. The digest was better than it had been previously, now with a difference in 120 bases, or approximately 0.8%, between the *in-silico* band to the real band at the region where the join had been made. My main scaffold now consisted of one contig.

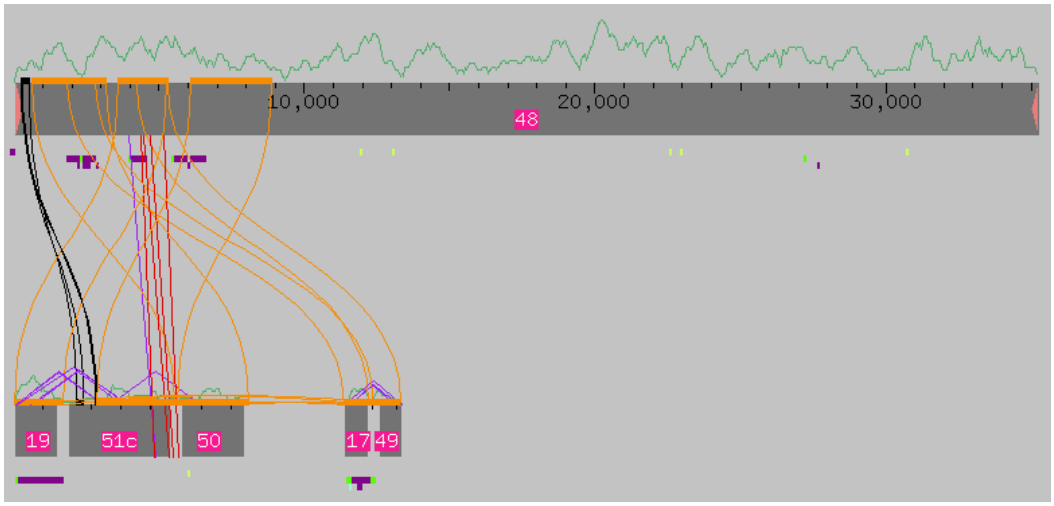

Figure 9. Assembly View after the join was made to resolve the gap.

Ordering Reactions – First Reactions

I decided to order reactions for additional reads during the class's second round of ordering.

Contig	Start / End	Direction	Oligo	Chemistry	Reason
Contig 48	9035 9058	=>	taacagagcattaatttc tgatagga	BigDye, 4:1, dGTP	Low coverage
Contig 48	9694 9718	<=	gaagactagtctcaac caaatcata	BigDye, 4:1, dGTP	Low coverage
Contig 48	3492 3509	=>	tttccgggcaatagtag g	BigDye, 4:1, dGTP	Confirm consensus
Contig 48	5356 5378	<=	attgcttgaagttattga gaatg	BigDye, 4:1, dGTP	Confirm consensus
Contig 48	2464 2482	<=	cgtttgccaggcatata at	BigDye, 4:1, dGTP	One high quality read
Contig 48	2209 2226	=>	tgtttcggctcatgtacg	BigDye, 4:1, dGTP	One high quality read

Table 1. The first reads ordered.

There were three regions in which I desired more data. In the 9 kb region of the main contig, there were few reads covering the region. In the 2 kb – 3 kb region, although there were many reads, only one was high quality. In order to confirm the consensus sequence, I decided to order more sequencing reactions from this region.

The region around 3.5 kb to 4.5 kb on the main contig was matched to another contig not in my main scaffold. Due to the steps I had taken when addressing the high quality discrepancies at the start of the project, the other contig consisted of reads that were removed from the main scaffold due to multiple high quality discrepancies. My hypothesis was that these discrepant reads did not belong in my project and were mistakenly added to the project. I decided to order reads over this region to confirm the consensus sequence.

Autofinish

Autofinish suggested reactions to help close the gap between contigs 19 and 51, based on the forward/reverse pairs linking the two. In addition, these two contigs have regions covered by only one read at their ends. Autofinish suggested reactions for these low coverage regions, as well. Autofinish did not suggest any other reactions on other contigs.

```
(contig) (left) (right) (type),(strand),(first base of read),(exp id),(template)
Contig19      -777     170 catttggtgtagtctctgatga,55,8283H01.10,<-,170,4,34117511K06 (fwd),5,25012511A13 (rev),6,24794611E01 (rev)
Contig19       940    1887 tggctctcttacgcttactc,55,8283H01.11,->,940,7,34117711A07 (rev),8,24864611M04 (fwd),9,35233111N18 (fwd)
Contig19      1611    2558 ttgtcaaattcggagtaactgt,56,8283H01.9,->,1611,1,34117711A07 (rev),2,24864611M04 (fwd),3,35233111N18 (fwd)
Contig51      -788     159 caaatggagcttcgtgatga,56,8283H01.13,<-,159,13,22241311J06 (fwd),14,25011411J20 (fwd)
Contig51      -149     798 ttaccttgggcgaga,57,8283H01.18,<-,798,26,22241311J06 (fwd),27,25011411J20 (fwd)
Contig51       840    1787 cagtgaaaatacgcatcaacaca,56,8283H01.16,<-,1787,21,22241311J06 (fwd),22,25011411J20 (fwd)
Contig51      2202    3149 cgatcaatcagttatatggca,56,8283H01.15,<-,3149,18,35233111N18 (fwd),19,24864611M04 (fwd),20,34117711A07 (rev)
Contig51      3107    4054 ctttctgttgcaggcagta,55,8283H01.14,->,3107,15,34302011M04 (fwd),16,34297211H21 (fwd),17,34117711A07 (fwd)
Contig51      3592    4539 tttgttttggctttgaacc,56,8283H01.17,->,3592,23,34302011M04 (fwd),24,34297211H21 (fwd),25,34117711A07 (fwd)
Contig51      3926    4873 tgcgcgtttaaatgtttc,56,8283H01.12,->,3926,10,34302011M04 (fwd),11,34297211H21 (fwd),12,34117711A07 (fwd)
```

Figure 10. The primers suggested by Autofinish.

However, as I believe that these two contigs do not belong in my project to due to multiple high quality discrepancies in their alignment with the main contig, I decided not to order reactions on contigs 19 and 51.

Incorporating New Data

From the round of reactions that I had ordered, only one read was unable to add successfully to existing contigs. The read that could not be added was low quality.

Next, I navigated high quality discrepancies.

```
Contig48    selgin13XBAD-8283H01_t7.b1    2324         high quality base disagrees with consensus
Contig48    selgin13XBAD-8283H01_t7.b1    2342         high quality base disagrees with consensus
Contig48    22346711C16.b1          5987         high quality base disagrees with consensus
Contig48    25182311C16.b1          5987         high quality base disagrees with consensus
Contig48    31170511L07.g1          5987         high quality base disagrees with consensus
Contig48    22445511E04.g1          5987         high quality base disagrees with consensus
Contig50    selgin13XBAD-8283H01_7.b1    1750         high quality base disagrees with consensus
Contig50    selgin13XBAD-8283H01_t8.b1    1768         high quality base disagrees with consensus
```

Figure 11. High quality discrepancies after adding new data

There are high quality discrepancies on multiple reads, found on the new reads added. Examination of these regions indicated that they are genuine high quality discrepancies, and are not due to base mis-calls.

The read selgin13XBAD-8283H01_t7.b1 was discrepant at both base 2324 and 2342 on contig 48, the main contig. An alignment of contigs 48 and 17 reveals that contig 17 has the same discrepancies as the new read. This suggests that earlier, I had removed the wrong reads from the main contig. As such, the reads in contig 17 belong in the main contig, and the reads with different bases at these positions do not belong in the main contig. There is also the possible that all of these reads should be included in the assembly, but with the discrepant bases marked as possible polymorphisms.

To test my hypothesis, I then joined these two contigs. Joining the contigs introduced many inconsistent forward/reverse mate pairs. I then removed the reads that were previously in the main contig. Miniassembly grouped these reads into a single contig. After the reads were removed, there were no more inconsistent mate pairs. This confirms my hypothesis that discrepant reads at this region were due to reads errantly pulled into

the project, rather than because of polymorphisms. I also joined any other contigs where there were no high quality discrepancies in their alignments.

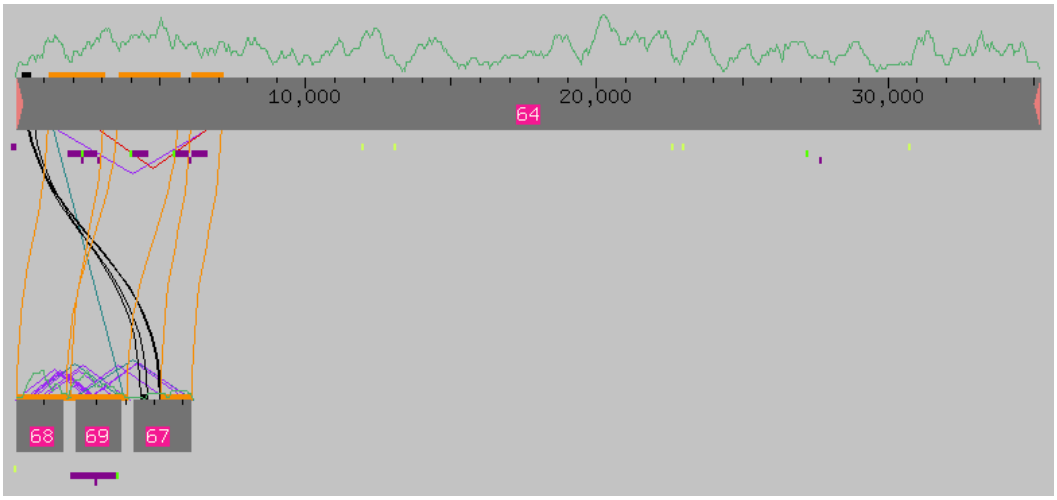

Figure 12. Assembly View after incorporation of new reads and joining contigs.

Conducting a BLAST search

To determine whether my project was contaminated by bacterial DNA or viral DNA, I ran a BLAST search on my main contig, contig 64. I exported the consensus sequence and entered this into the BLAST site. I used blastn, set the expectation threshold to 1 e-10, and unchecked the "Low complexity regions".

BLAST returned five hits. Four of the matches were to "Thioalkalivibrio nitratireducens DSM 14787, complete genome," a bacteria. These hits mapped to the same 88 bp region in my assembly. If my project were contaminated by this bacteria, this DNA contamination should manifest as extra data in my project. An *in-silico* band should be of this length greater than the real band. However, the digests of my main contig did not show bands that were 80 bp more than the real data provided. Therefore, it's unlikely that this bacteria contaminated my project.

The last hit returned by BLAST was to "Francisella philomiragia subsp. philomiragia ATCC 25017 chromosome, complete genome." This match was about 84 bases long. Again, the digest of my main contig did not show any bands that were this many bases longer than the real data at this region.

The hits returned matched my assembly to the bacterial genomes in low complexity regions, GC rich regions. One of the hits can be seen in the figure below. From the examination of the digest data and the sequences found by BLAST, I believed that there was no contamination in my clone.

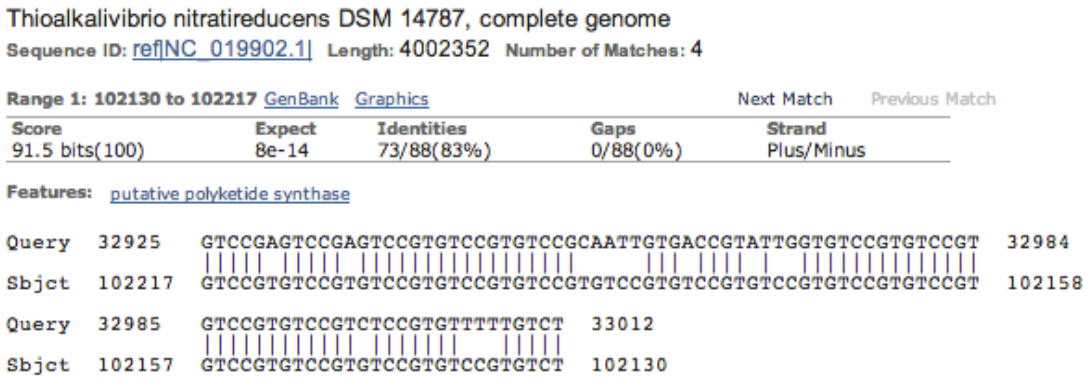

Figure 13. One of the hits returned by BLAST.

Ordering Reactions – Second Reactions

When ordering more reactions, I decided to call for more reads around the beginning of my fosmid, since there were fewer reads covering the beginning of the clone.

Contig	Start / End	Direction	Oligo	Chemistry	Reason
Contig 72	146 165	<=	ggtgcttatgatggatg gat	BigDye, 4:1, dGTP	Low coverage
Contig 72	321 341	<=	caccaaagttatgcca aagat	BigDye, 4:1, dGTP	Low coverage

Table 1. The second group of reactions called.

Incorporating New Data – Second Reactions

All of the reads that I had ordered were added to the main contig.

Searching for x's within consensus

I found that the first 55 bases of my sequence were marked x's in the consensus, which denotes a vector sequence. To resolve this, I tore the contig downstream at approximately position 4 kb. The new left contig was then run through Miniassembly. This placed all vector sequence before the 0 position, and no more x's were in the consensus sequence. This contig was then rejoined to the right contig.

Final Assembly

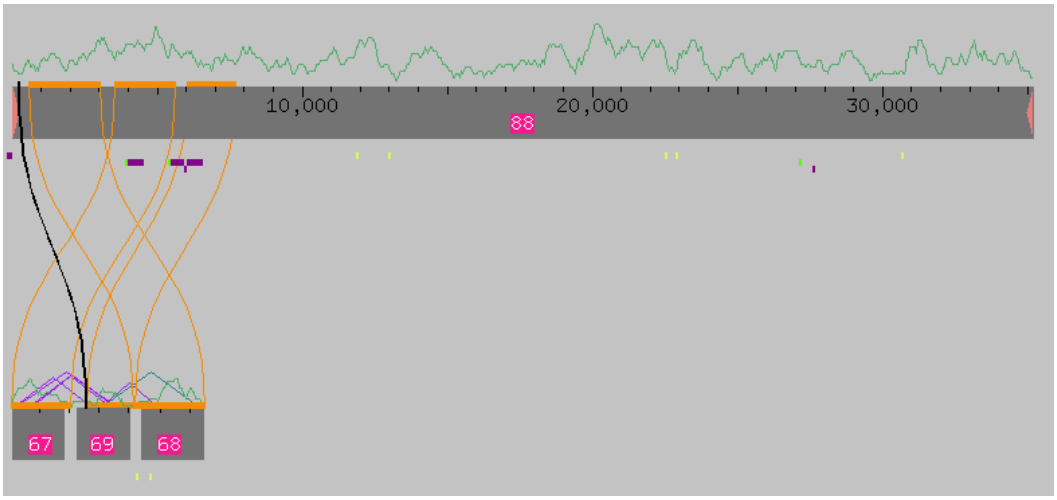

Figure 14. The final assembly for clone 8283H01.

The final assembly for this project consists of one contig, contig 88. As you can see from the figure above, there are three contigs over 2 kb that are not in the final assembly. These three contigs are connected by forward/reverse pairs, and all match to regions on the main contig. However, I believe that these contigs do not belong to the project. When their alignments with the main contig are viewed, there are multiple high quality discrepancies. Furthermore, when trying to join these contigs with the main contig, inconsistent forward/reverse pairs are introduced. Thus, I do not believe that contigs 67, 68 and 69 belong to the project. The main assembly of contig 88 represents the final assembly for this project.

There is a potential single nucleotide polymorphism at base 5932; this was tagged as such.

Finishing Checklist

There were no mononucleotide runs, and no x's nor n's in the consensus sequence. One possible polymorphism was tagged. The three contigs over 2 kb that were not in the assembly were tagged as well. A BLAST search confirmed that there was no contamination.

There were no regions of low consensus quality. There were no single subclone regions in the main assembly. There were four single strand regions, all under 300 bp long, in the main assembly. These regions had multiple high quality reads. The only high quality discrepancies were of the potential SNP at base 5932. There were no mononucleotide runs of 15 bp or more in the assembly.

Restriction Digests

In all four of the restriction digests, the *in-silico* digest matched well with the real digest.

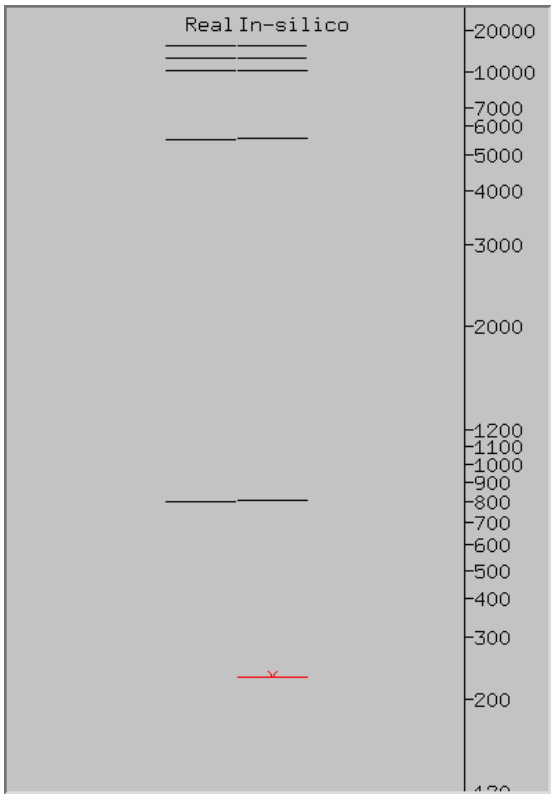

Figure 15. Digest with *Eco*RV.

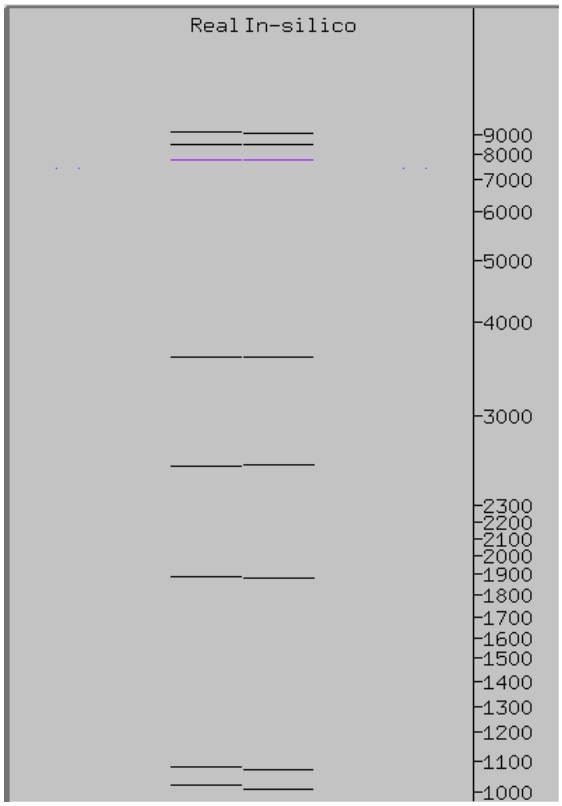

Figure 16. Digest with *Eco*RI.

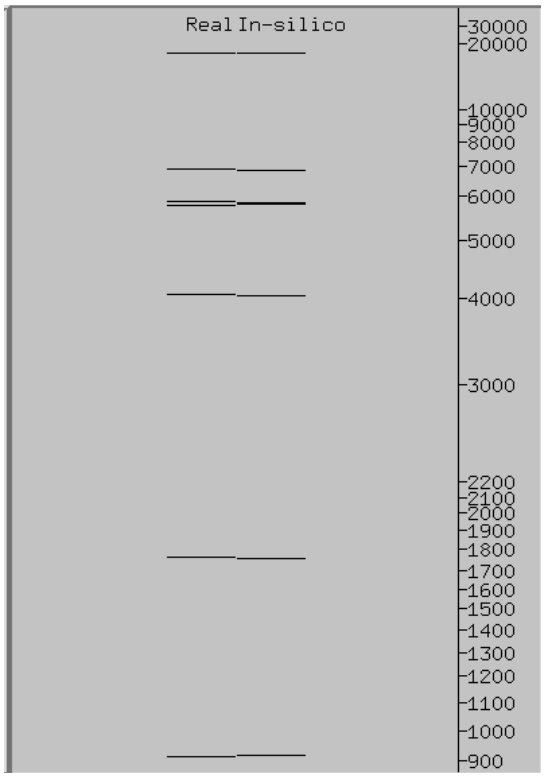

Figure 17. Digest with *Sac*I.

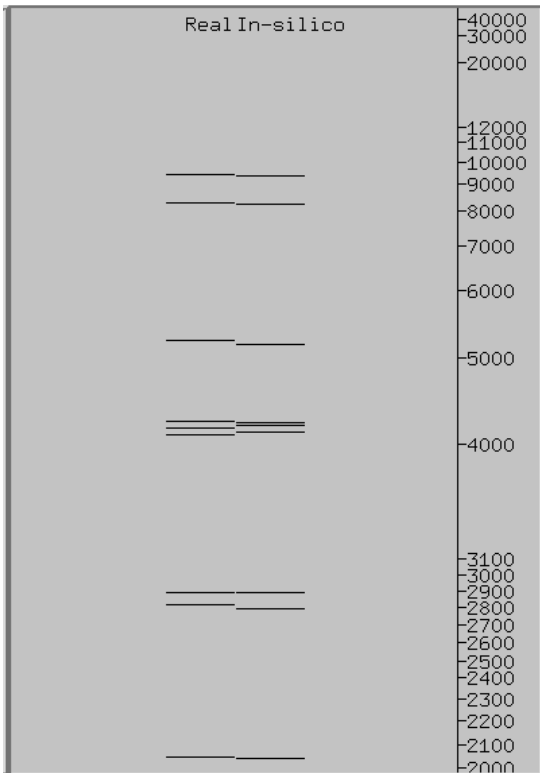

Figure 18. Digest with *Hind*III.

In the digest with *Eco*RI, the *in-silico* digest has one more 135 bp fragment not seen in the real digest. The digest with *Sac*I shows that the *in-silico* digest has one more 154 bp fragment not seen in the real digest. In *Hind*III digest, the *in-silico* digest has one extra 324 bp fragment not seen in the real digest.

Small bands of these sizes might not have been detected by the real digest.

The four in-silico digests were checked for their total size. The sum of all band sizes from each of the four in-silico digests was added; the sum was 43310 bp for each.

Conclusion

My clone fosmid, 8283H01, initially had various issues in its assembly, such as a gap, inconsistent mate pairs, and multiple high quality discrepancies. By separating discrepant reads from the main assembly and by ordering new reads, I was able to finish the fosmid. Three contigs that did not belong in the final assembly may have been mistakenly matched to the project due to a repetitive region. It would be fruitful to determine the type of repeat found in this project, and where else in the genome this repetitive region is also found.

Acknowledgements

Thank you to Professor Elgin, Dr. Shaffer, Dr. Mardis, Wilson Leung, Lee Trani, Jennifer Hodges, and Harry Senaldi for all of their guidance.