Finishing *Drosophila virilis* Fosmid Clone 4N16

David Desruisseau
April 12, 2006

Bio 434W Genomics

Finishing *Drosophila virilis* Fosmid Clone 4N16

David Desruisseau

Abstract

The overarching goal of Bio 4342/W research over the past several years has been to understand a euchromatic region of the *Drosophila virilis* genome well enough to be able to distinguish this domain at the DNA level from the heterochromatic counterparts in its genetic relatives, such as *Drosophila melanogaster*, the common fruit fly. The class will utilize the well-established genome of *D. melanogaster* as a model organism possessing a heterochromatic fourth chromosome for comparison. Current class research focuses on completing a reliable genome sequence for the equivalent "dot" chromosome in *D. virilis* in hopes of discerning sequence domains or gene characteristics that explain this inter-species difference in chromosomal organization. In this report, I describe sequence finishing for fosmid 4N16.

Finishing Workflow

To start off

After running Phred/Phrap to call bases on the raw sequence data and create an initial assembly for my fosmid, I was presented with a Consed assembly consisting of five major contigs with a significant number of inconsistent forward/reverse pairs, a significant low coverage region and extensive misassembly. Running Crossmatch identifies direct sequence repeats, indicated by orange lines, or tandem sequence repeats, indicated by black lines (none visible here), according to a desired percent similarity. Viewing the Crossmatch results with a sequence similarity of 90% revealed very high sequence repetition throughout my entire fosmid.

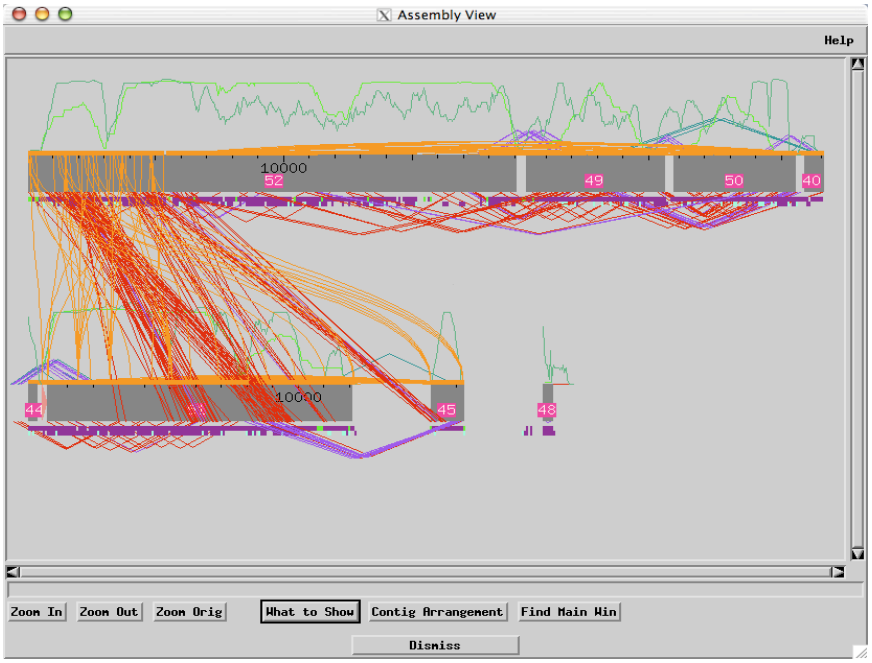

Figure 1.1
Initial assembly

As a first effort to quickly improve assembly quality, multiple high quality discrepancies were tagged with the command *Tell Phred/Phrap not to overlap reads at this location* in order to increase the stringency of assembly joins. After rerunning Phred/Phrap with these parameters set, Assembly View displayed slightly altered contigs with somewhat fewer inconsistent forward/reverse pairs. In order to improve the readability of Assembly View, the *Reorient Contigs* command was used to ensure consistent sequence directionality between all contigs. Finally, the *exclude contig if depth of coverage greater than this* parameter was increased to 80. This tells Consed to display any highly misassembled contigs with unusually high read depths, revealing a greater amount of the available sequence. The resulting assembly is shown in Figure 1.1.

Locating and establishing ends

The next task was to locate and establish two different types of ends: those of the individual reads and those of the entire fosmid. Looking through the aligned reads, my goal was to identify vector sequence that had been represented as clone sequence and eliminate it wherever possible. This process involved the recognition of vector end sequences on individual reads. The ends were as follows: GAATTCGTC—insert, GAATTCGTT—insert, insert—GACGAATTC, and insert—AACGAATTC. After locating regions where vector sequence remained in a read, the regions were excluded using the command *Change to x's to left/right*. This tells Consed to ignore this sequence in order to better calculate consensus sequence for that particular area and therefore construct a better whole assembly.

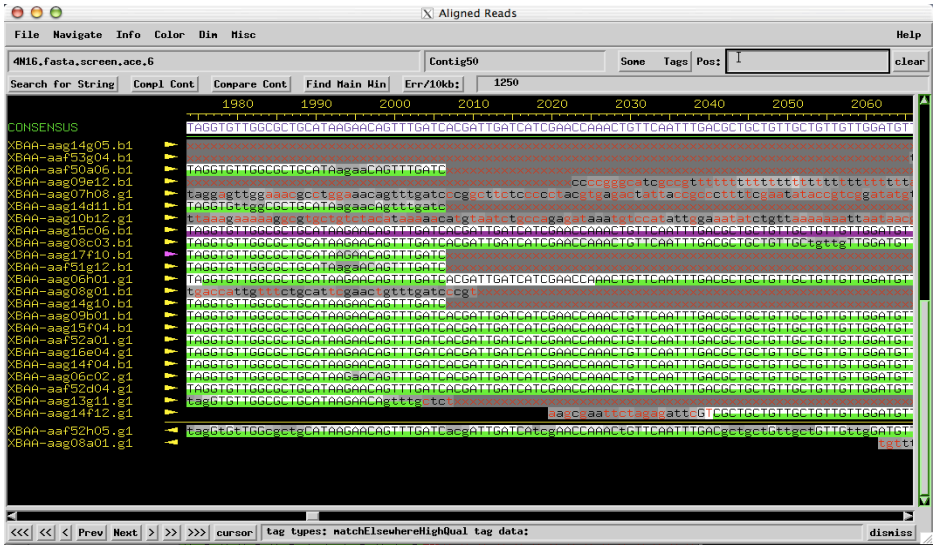

Figure 2.1
Candidate end

Whole clone ends were located in a similar fashion and were identified with a GATC end sequence and then a string of x's in either direction, representing a fosmid end. High quality flanking vector sequence (from pfos1, the vector used here) was almost always correctly identified by Consed, making the process of finding clone ends relatively straightforward. However, if vector sequence was of too low quality for Consed to properly define as a clone end (see Figure 2.1), the trace window was used to easily

compare the general trace patterns for the confirmed and unconfirmed vector sequence. In this way, having to resolve the exact sequence for the unconfirmed trace would be unnecessary. Simply comparing the general trace patterns would give sufficient visual proof that the read in question did indeed contain pfos1 vector sequence and that the region could be safely ignored (Figure 2.2).

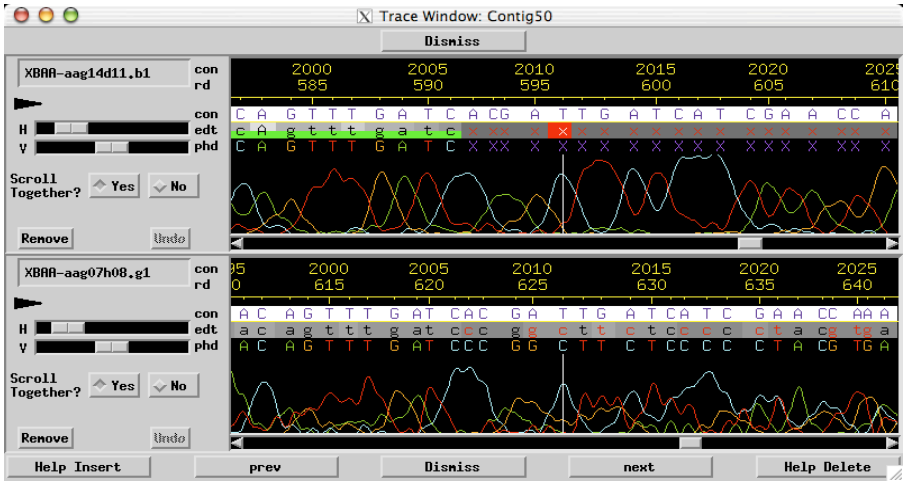

Figure 2.2
Dye trace

Note the high similarity between peak distribution and general trace shape.

Assessing inconsistent forward/reverse pairs

After locating my fosmid ends, the next problem to be addressed was to assess inconsistent forward/reverse pairs. As is apparent in the original Assembly View (Figure 1.1), my fosmid was initially highly misassembled, with a number of mismatches falling almost entirely within highly repetitious regions. Knowing this, as well as understanding the time constraints on this project, a decision was made to access a version of the fosmid that had been partially finished by a WU GSC finisher. The assembly pieces borrowed from this alternate version would serve as a scaffold for my assembly.

The scaffold data from this alternate assembly was prepared as a .phd file, added to the phd_dir of my project folder, and Phred/Phrap was rerun. The assembly data would serve to augment mine, and would provide a relatively trustworthy template to which Phred/Phrap could align my sequence. Incorporating this data improved my Assembly View significantly and produced a large contig of approximately 37 kilobases in length (Figure 3.1). The remaining four significant smaller contigs also exhibited fewer inconsistent forward/reverse pairs with relation to the main contig.

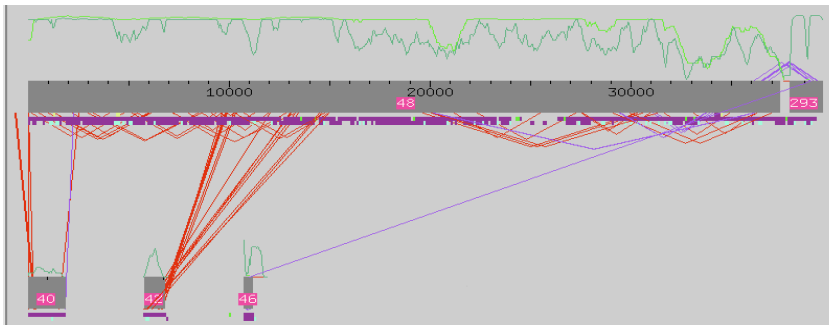

Figure 3.1
Scaffold incorporated

Note about Findid

The GSC ran my assembly against their in-house database to sequence from 14 different organisms, including bacterial, human, yeast, and maize DNA. This scan confirmed that my main contig contained no findid-labeled contamination.

Calling reads

In order to close the gap at the right end of my fosmid, I needed to call reads to obtain additional sequence data for that region. Ordering oligos that would effectively span this region required careful selection of unique sequences that would anneal to only one complementary region during the PCR reaction. This was somewhat difficult with my clone, since it contains so much repetitious sequence, but I was able to define four unique oligos through trial-and-error using the *Search for String*. Autofinish was not used, since the level of misassembly was determined too high for the program to be helpful.

Oligo	Sequence	Template	Reaction Outcome
1	TGCTGTTTGTGCGTTA	aag12e08	Added
1	TGCTGTTTGTGCGTTA	aag13c05	Added
2	GCATAAACAGCATAACTCC	aag16c11	Not added
2	GCATAAACAGCATAACTCC	aag16f12	Added
3	TCATGTGTATTCTTTGCACT	aaf53c05	Not added
3	TCATGTGTATTCTTTGCACT	aag11b06	Not added
4	TCAAAACATTTGTTTTATAGG	aaf52d12	Not added
4	TCAAAACATTTGTTTTATAGG	aag12e08	Not added

Figure 4.1
Reads and outcomes (BigDye Chemistry)

Unfortunately, reaction success was not universal, as five of the eight reactions failed and were not entered by Consed when reaction data was incorporated into the assembly. These results indicate that oligos 3 and 4 did not perform well in hybridization, hence their completely ineffective results. Oligo 2 was slightly more successful, resulting in one read of good data, while oligo 1 clearly worked best here, since Consed added both of its returned reads. These poor hybridization outcomes could have been the result of poor primer-template annealing, which could have been addressed with more time and trial-and-error. However, the reads that were successful provided good quality data to help augment the low-coverage areas near the right end of the clone. The successful reads and the positions where they were added into the primary contigs are indicated in the following table (Figure 4.2).

Read Added	Contig	Start Position	End Position
XBAA-aag12e08_1.b1	Contig 294	32,570	34,022
XBAA-aag13c05_1.b1	Contig 295	37,531	38,949
XBAA-aag16f12_2.b1	Contig 296	36,160	37,510

Figure 4.2
Reads added to assembly

Using Printrepeats output to perform the final contig join

Ghostview presents a graphical representation of the Printrepeats file output and displays tandem and inverted repeats all in a single view. For my clone, the Printrepeats minimum repeat threshold length was specified at 51 bps. This image provides a meaningful visual representation of the overall repetitiveness of the entire fosmid, including relative levels of tandem versus inverted repeats (Figure 5.1). Within Ghostview, inverted repeats are visualized as arches that peak above others; six of this repeat type are visible in the above figure. The remainder of the arches shows tandem repeats. Overall, the information provided in the Ghostview representation is very similar to Consed Crossmatch, and this picture served mostly as an additional visual representation of the repeat distribution along my fosmid. Accompanying this visual representation, Printrepeats provides a text output of the assembly, which provides specific paired sequence match positions if needed.

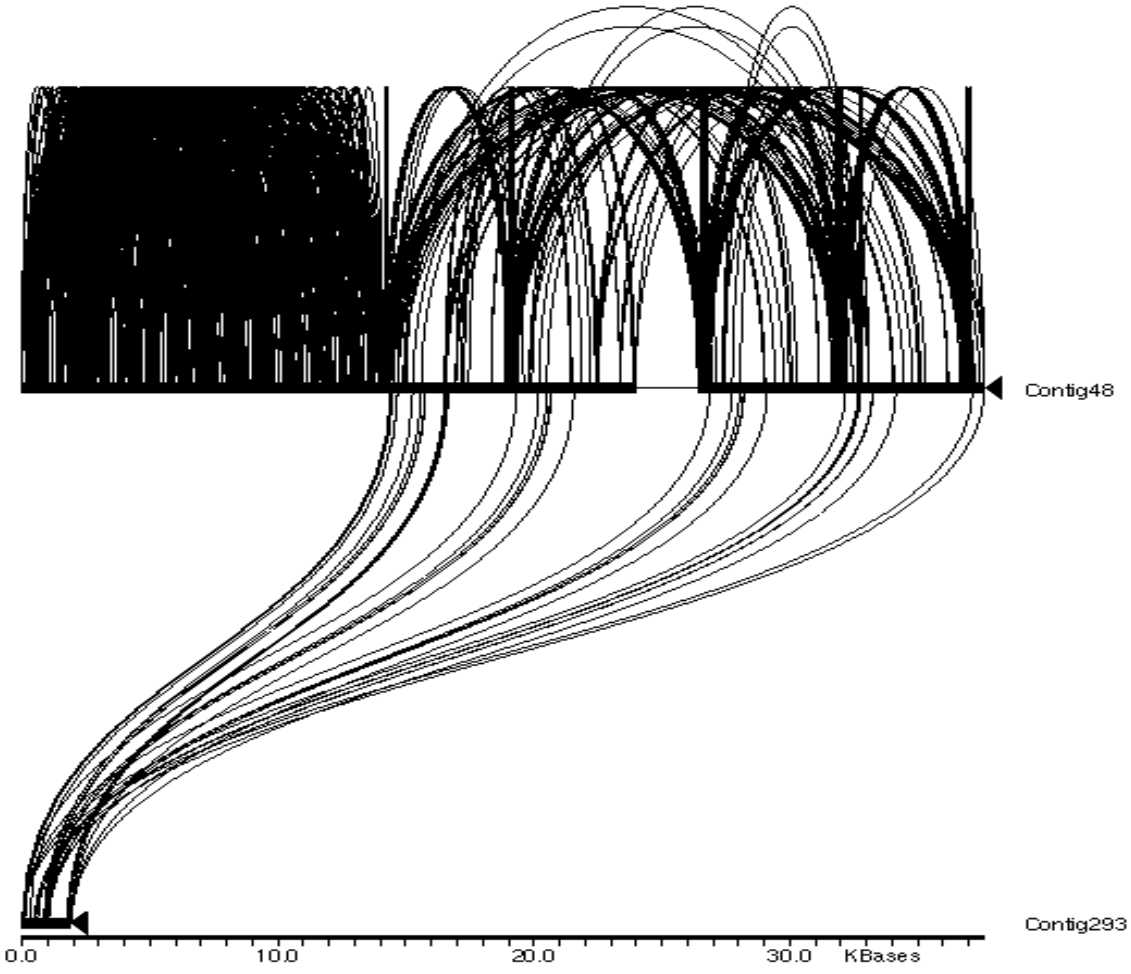

Figure 5.1
Ghostview of Printrepeats

Restriction digests and making the penultimate join

In order to join contigs 48 and 293 (as they appear in Figure 5.1), I used the Consed-based restriction digest analysis tool to view the digests for various enzymes. Some of the enzymes returned very poor results and would not give me any additional information as to where a join should be made. However, *Hind*III and *Eco*RI returned good results for my fosmid and allowed me to effectively compare real fragment and *in silico* fragment lengths. The digest results are provided in Figures 6.1 and 6.2. Of note is the slight discrepancy between the real and *in silico* fragment lengths for the *Hind*III digest. In figure 6.2, there is purple line visible under the lane marked 'Real' which indicates the presence of two digest fragments at that position. However, this is often simply a symptom of incorrect gel band-intensity reading by the computer that processes the original gel captures.

The band intensity on the original gel images varies not only with fragment number but also fragment size. Ideally, the computer normalizes for the relative size differences between these fragments as well as for differences in gel band intensity in order to report correct fragment number. However, the somewhat qualitative nature of intensity classification sometimes introduces mistakes and will cause Consed to report the incorrect fragment band multiple for a given position. Analysis of the original gel image confirms this (Figure 6.3).

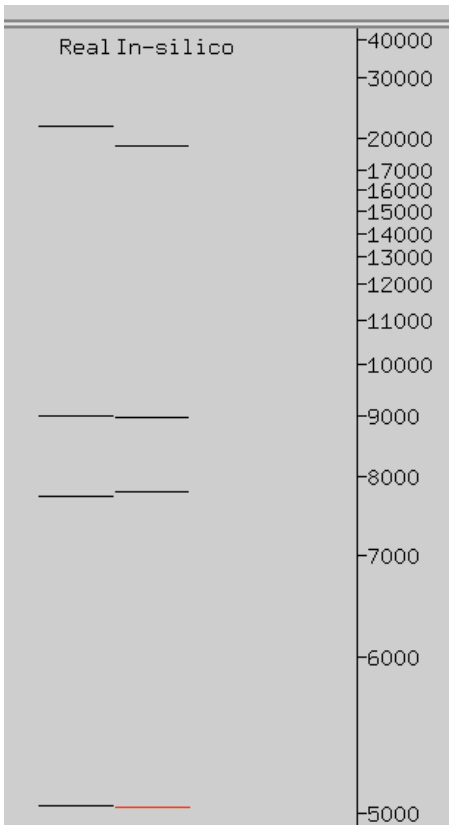

Figure 6.1
*Eco*RI digest

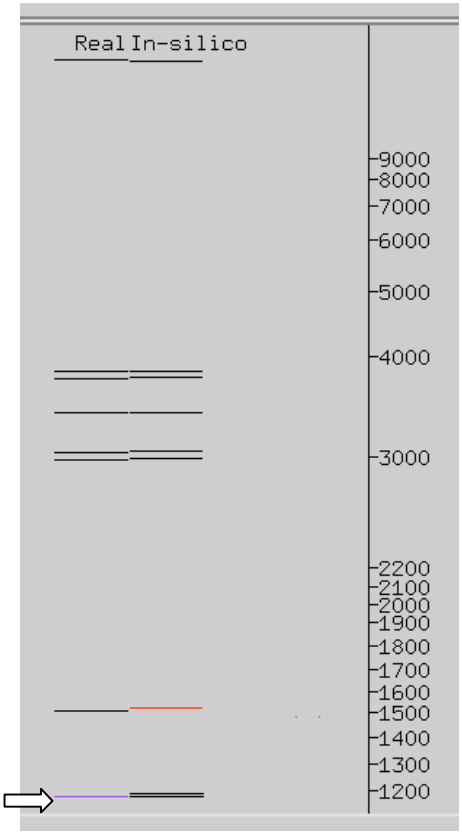

Figure 6.2
*Hind*III digest

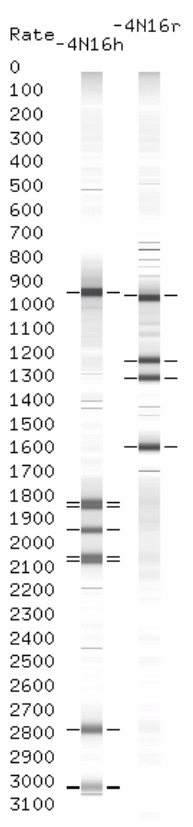

Figure 6.3
Gel image

Finally, based on knowledge of restriction digest cut sequence, the digest fragment sizes were compared to existing contigs, indicating a join was needed between contigs 48 and 293 (Figure 5.1). After determining the approximate region of the join based on these results, *Compare Contigs* was used to confirm the exact pairing and then a join was made, leaving only two remaining contigs (Figure 6.4).

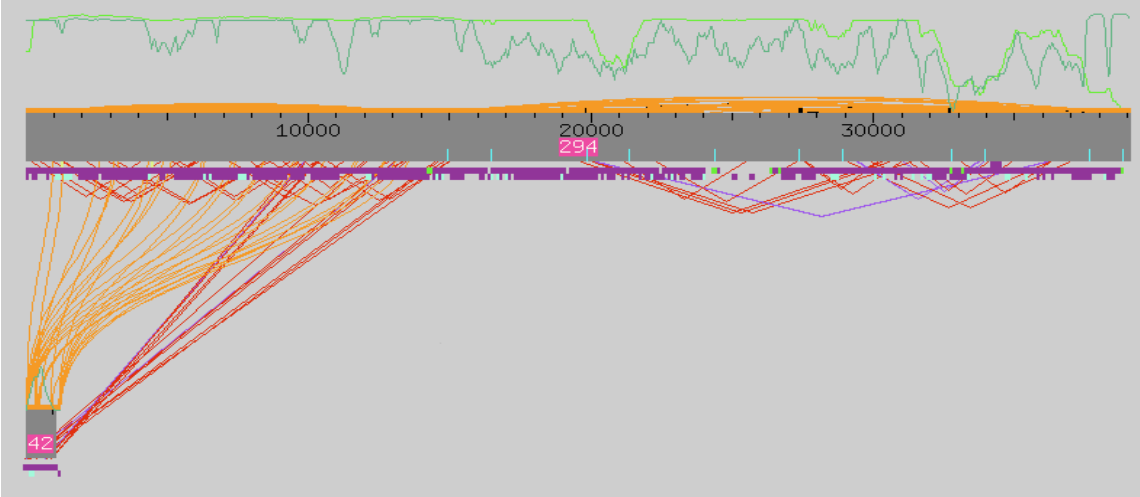

Figure 6.4
Resulting assembly following restriction digest-aided join

Joining the final contigs

Only one more contig join was needed before the fosmid assembly in Figure 6.1 would become a single contig. To accomplish this, Crossmatch was used to determine that the assembly location on the major contig would likely fall somewhere in the range of 10 – 13 kilobases. Using this information, as well as the *Search for String* function returned a likely sequence match at approximately 11 kilobases to Contig 42 in Figure 6.1. *Compare Contigs* confirmed the alignment and the two contigs were joined, providing a nearly finished single-contig assembly of the expected total length of nearly 40 kilobases (Figure 7.1).

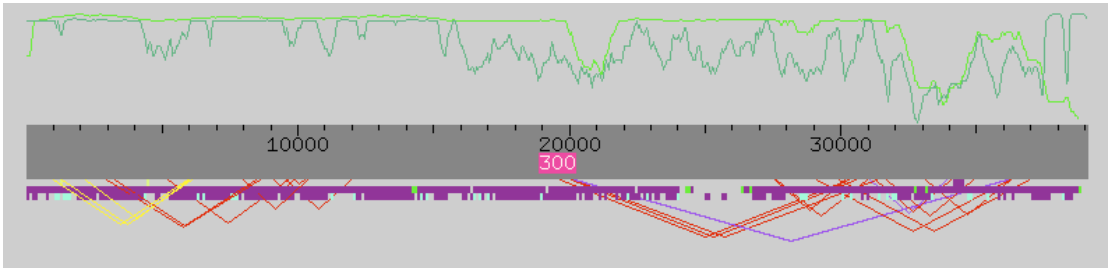

Figure 7.1
Single-contig assembly

Tying up loose ends

After the successful joining of all contigs into one unbroken fosmid assembly, the most significant remaining problems were the inconsistent forward/reverse pairs. These inconsistencies were not especially difficult to resolve and involved tearing out the read of one end of the pair and using the average pair separation provided by Consed to estimate the correct placement of the read. In most cases, there is no logic as to which end to pull out, so the process proceeded mostly by trial and error. However, in cases in which tearing one end would require moving it somewhere off the end of the contig, the choice was clear: tear out the other end. *Search for String* indicated the presence of a match within the estimated region and *Compare Contigs* provided confirmation of this before the reads were joined. Repeating this process quickly resolved these inconsistencies in Assembly View (Figure 8.1), providing a nearly complete fosmid (contrasting sharply with the very messy initial assembly in Figure 1.1).

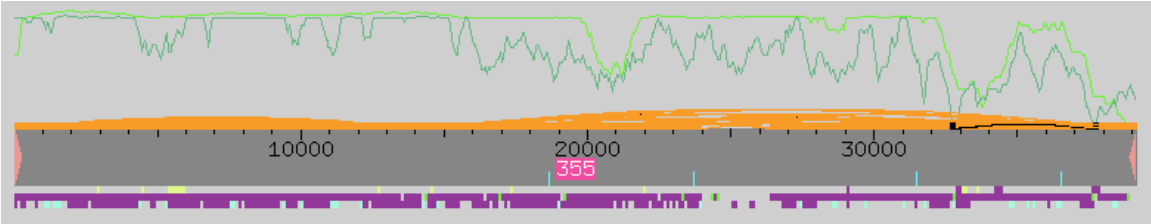

Figure 8.1
Final single-contig assembly

Checklist

As a final check of the completeness of the finished fosmid, a quick search was done to confirm that there were no existing mononucleotide runs (of over 15 bases of a single nucleic acid) or X's or N's. *Search for String* confirmed that these were not present. Also, there were no regions with low consensus quality (a Phred score of under 25), no high quality discrepancies, no single-strand regions, and no single subclone regions. These checks confirmed my fosmid was ready for submission to the GSC.

Final thoughts

Although my fosmid was successfully assembled into a single contig free of inconsistent forward/reverse pairs, there are still some small remaining problems. Due to the extremely high level of repetitive sequence that made this fosmid so challenging to assemble in the first place, there are still several reads that do not properly match in that they contain two different versions of a tandem repeat. These highly similar regions are differentiated by a single high quality discrepancy and are not dissimilar enough to be tagged as mismatched forward/reverse pairs. The most likely explanation for this is the presence of chimeric DNA. Accurate library formation was likely hampered due to the

extremely repetitious nature of my fosmid. This repetitiousness can cause several otherwise unrelated sequences to become joined after sonification, resulting in fragments that contain one end of DNA from one area and DNA from another area at the other end. This can be detected in Consed by examining reads and seeing that two regions of the read match to different areas of the assembly. Resolving this issue is beyond the scope of my work, so it will be up to the GSC to decide whether they will attempt to fix this.

Special thanks

Special thanks to Laura Courtney who guided me through much of my finishing work. Without her expertise, I would not have achieved the same success in assembling my fosmid.