Fatih Ozsolak **03-05-04**

Project: XAAA106

After the assembly of the first set of reads that I generated and the reads that were generated by the GSC, I check the assembly picture of my fosmid (Figure 1). There were two contigs, contig 7c (14176bp) and contig 8 (26385bp), and a ~700bp gap between them (determined later from the restriction digest analysis). I checked the ends of contig 7c and 8 to verify the cloning sites. Both contigs were ending with GATC sequence, indicating that these two contigs were spanning the whole insert in the fosmid.

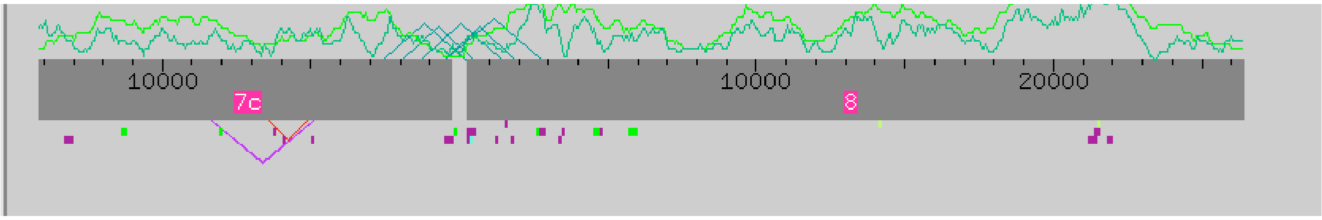


Figure 1: Assembly view of the initial assembly. There were two contigs, contig 7c and contig 8, as well as a ~700bp gap in between.

I observed the following major problems with my project:

1) There was a ~700bp spanned gap between contigs 7c and 8.
2) There was a low-quality region on contig 8 from 3380 to 3690 bp. There were 8 reads covering this region, however none of them provided sufficient amounts of high-quality data to deduce the consensus sequence.
3) There was another low-quality region on contig 8 from 23300 to 23700 bp. There were only 3 reads covering this region, but none of them provided data with sufficient quality.
4) The last low-quality region was at the beginning of contig 7c from 1 to 70 bp.

I also observed the following minor problems:

5) The region between 6957 and 6979 on contig 7c was covered by only one chemistry. However, although this region seemed to be consistent of multiple 'GA' repeats and it was possible that errors could have been present, there were three reads covering this region with very high quality sequence (quality values above 70). Therefore there was no need to re-sequence this region.
6) There were also such regions on contig 8 covered by only one sequencing chemistry (420-866, 1375-1474, 5037-5120, 6999-7338, 12406-12476, 17535-17704, and 25353-25474). However, all of these regions were covered by at least 2 high-quality reads. Therefore, there was no need to suspect the accuracy of data in these regions.
7) There was a 5bp region (12406-12410) on contig 8 covered by a single read. However, the region was not in a repetitive region and qualities of all 5 bases were 68. Therefore, there was no need to re-sequence this region.

I called the following reads on Table 1 to close the gap as well as to resolve the three low quality regions identified:

Reaction Name	Oligo Used	Purpose of Reaction	Quality of Reads
uua73b11_1.b1	XAAA106.Oligo1 gggatatcgatatcgatagatttat	To resolve the low quality region on contig 8 (3380-3690)	Low
uua72b06_2.b1	XAAA106.Oligo2 aatctgaagcaaagatatattcgt	To resolve the low quality region on contig 8 (3380-3690)	High
uua73h04_3.g1	XAAA106.Oligo3 ttccacacatgcccac	To resolve the low quality region on contig 8 (23300-23700)	High
uua73b08_4.g1	XAAA106.Oligo4 gttaaatataatgagcagttgacg	To resolve the low quality region on contig 8 (23300-23700)	Low
uua73h10_5.g1	XAAA106.Oligo5 cctcttattccaacaccaaat	To close the gap	Low
uua72a06_6.g1	XAAA106.Oligo6 tccagccgggacctc	To close the gap	Low

Table 1: The reads called during the first round of finishing

The problems identified by autofinish were the same issues that I identified as major problems above. To close the gap, autofinish picked two different oligos and two different subclones. XAAA106.2 oligo and corresponding subclone picked by autofinish was the same oligo and subclone that I picked (Oligo6 and uua72a06.b1). The other oligo that autofinish picked (XAAA106.1) was not the same as the primer that I picked (Oligo5). I wanted my oligos to be closer to the gap region, therefore I had to pick oligo 5 in a rather low-quality and repetitious region. Before ordering oligo5, I looked at the available sequence information from uua72a06 to minimize the possibility that oligo5 might have had more than one binding site.

XAAA106.4 oligo and uua72b06 subclone that autofinish picked to resolve the low quality region on contig 8 (3380-3690) was the same oligo (oligo2) and subclone that I picked for this low-quality region.

XAAA106.3 oligo that autofinish picked to resolve the low-quality region on contig 8 (23300-23700) was not the same oligo that I picked (Oligo 4). XAAA106.3 was about 600bp away from the low-quality region while Oligo 4 that I designed was 150bp away.

Only two of the six reactions ordered yielded high-quality sequence (uua72b06_2.b1 and uua73h04_3.g1) and were included in the assembly. Information obtained with uua72b06_2.b1 reaction was sufficient to resolve the low quality region on contig 8 (3380-3690), but it was still covered with only one read. However, the other low-quality region on contig 8 as well as the gap remained unresolved. To solve these two problems, the reactions on Table-2 were called in round2 of finishing.

Reaction Name	Oligo	Purpose of Reaction	Quality of reads
uua73h04_4.g1	XAAA106.Oligo4	To resolve the low quality region on contig 8 (23300-23700)	High
uua73h04_g4.g1	XAAA106.Oligo4	To resolve the low quality region on contig 8 (23300-23700)	Low
uua73h04.g2		To resolve the low quality region on contig 8 (23300-23700)	High
uua73b08_g3.g	XAAA106.Oligo3	To resolve the low quality region on contig 8 (23300-23700)	Low
uua73b08_3.g1	XAAA106.Oligo3	To resolve the low quality region on contig 8 (23300-23700)	Low
uua73b08.g2		To resolve the low quality region on contig 8 (23300-23700)	High
uua72d03_t7.b1	XAAA106.Oligo7 aatttgcaatagttggaatgac	To resolve the low quality region on contig 8 (3380-3690)	High
uua72d03.b2		To resolve the low quality region on contig 8 (3380-3690)	High
uua72b02_t8.g1	XAAA106.Oligo8 cttccaacataaaatgcatgta	To close the gap	High
uua72b02_g8.g1	XAAA106.Oligo8	To close the gap	Low
uua72b02_t5.g1	XAAA106.Oligo5	To close the gap	Low
uua72b02.g2		To close the gap	High
uua72a06_t6.g1	XAAA106.Oligo6	To close the gap	High
uua72a06.g2		To close the gap	High
uub25d01_t6.g1	XAAA106.Oligo6	To close the gap	High
uub25d01.g2		To close the gap	High

Table 2: The reads called during the second round of finishing

11 of the 16 reads ordered yielded high-quality data and were included in the assembly. Although the low quality region on contig 8 (23300-23700) was now resolved, gap still remained unresolved. However, uua72a06_t6.g1 and uub25d01_t6.g1 reads

yielded low quality data for the 500bp region of the gap on contig 8 indicating that there might be a lot of GC repeats in the gap region. This might have been the reason why so many reads failed to close the gap. Additional reads were ordered mostly by using the dGTP-sequencing chemistry, since this chemistry tends to perform better in compressed regions containing repeats (Table 3).

Read Name	**Oligo**	**Purpose of Reaction**	**Quality of Reads**
uub25d01_g6.g1	XAAA106.Oligo6	To close the gap	High
uua72a06_g6.g1	XAAA106.Oligo6	To close the gap	High
uua72a01_g6.b1	XAAA106.Oligo6	To close the gap	Low
uua72a01_t6.b1	XAAA106.Oligo6	To close the gap	Low
uua73h10_g5.g1	XAAA106.Oligo5	To close the gap	High
uua72b02_g5.g1	XAAA106.Oligo5	To close the gap	High
uub25d01_g5.g1	XAAA106.Oligo5	To close the gap	High
uua72b02_g8.g1	XAAA106.Oligo8	To close the gap	High
uub25d01_g8.g1	XAAA106.Oligo8	To close the gap	High
uua73h10_g8.g1	XAAA106.Oligo8	To close the gap	High
uua73h10_t8.g1	XAAA106.Oligo8	To close the gap	High

Table 3: The reads called during the third round of finishing

Luckily, 9 of the 11 reads called during the third round of finishing worked quite well and provided sufficient data to close the gap. In the final consensus sequence, the gap was present between 13890 and 14570, a 680bp region as expected, and was rather rich in G and C's. The final size of the contig was 40734bp determined from 454 reads (Figure 2).

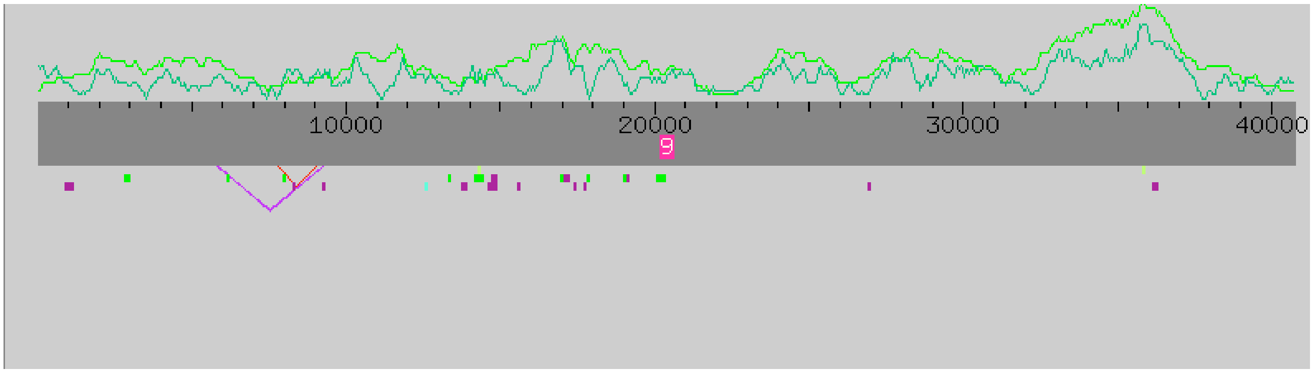


Figure 2: Assembly view of the final assembly. There is one contig spanning the entire fosmid insert.

One interesting problem was the existence of two subclones, uub25g02 and uua72h01, assembled inconsistently, since their forward and reverse pairs were pointed towards the same direction (indicated in figure 1 and 2 with purple and red lines). There were no high-quality discrepancies, indicating that the problem was not simply due to a sequence alignment mistake. I reassembled the contig by using the mini-assembly feature and changing the minmatch value to 30, minscore value to 60 and forcelevel to 15. The new assembly was exactly the same as the old assembly. After looking at the trace files, I understood that these two subclones were essentially the same subclone, that is, p0 vectors were containing the same inserted fragment. The reason for the inconsistency is probably the chimerical nature of the insert. In addition, there were many other subclones covering the region with proper forward and reverse reaction orientation. Therefore, this problem was considered to be not important (Figure 3).

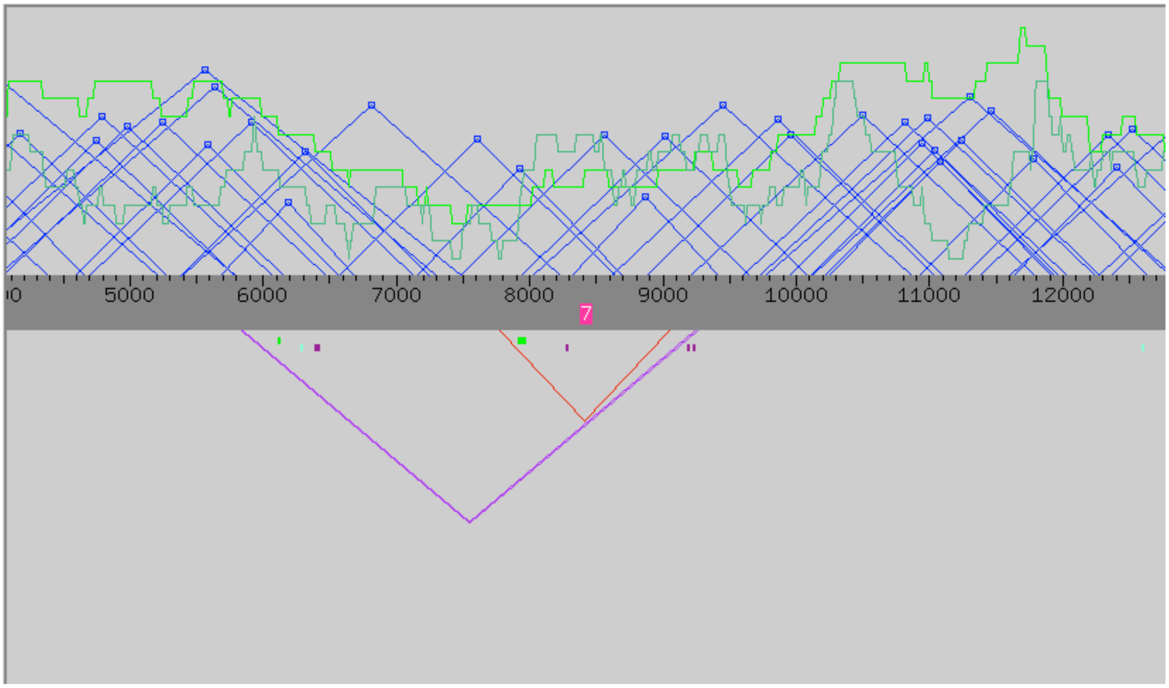
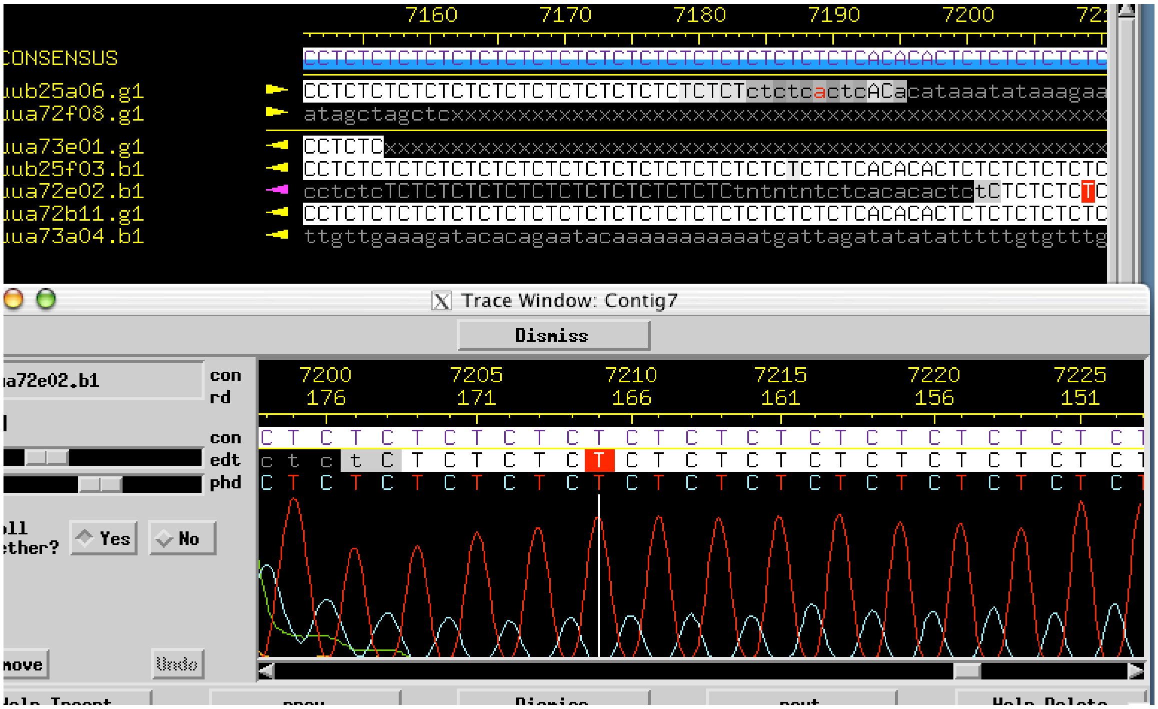


Figure 3: The region containing two inconsistent reads as well as the consistent subclones.



```
                         7160      7170      7180      7190      7200      72
CONSENSUS                CCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCACACACTCTCTCTCTCTC
uub25a06.g1              CCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTctctcactcACacataaatataaagaa
uua72f08.g1              atagctagctcxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
uua73e01.g1              CCTCTCxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
uub25f03.b1              CCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCACACACTCTCTCTCTCTC
uua72e02.b1              cctctcTCTCTCTCTCTCTCTCTCTCTCTCTCTCtntntntctcacacactctCTCTCTCTC
uua72b11.g1              CCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCACACACTCTCTCTCTCTC
uua73a04.b1              ttgttgaaagatacacagaatacaaaaaaaaaaatgattagatatatattttttgtgtttg
```

```
con  C T C T C T C T C T C T C T C T C T C T C T C T C T C T C
edt  c t c t C T C T C T C T C T C T C T C T C T C T C T C T C
phd  C T C T C T C T C T C T C T C T C T C T C T C T C T C T C
```

Figure 4: 46 ‘CT’ tandem repeats observed (7130-7238)

There were also a couple repeat regions. The first was the 46 CT repeats observed between 7238 and 7130 (Figure 4). However, as seen from the trace file, the data obtained is at very high quality (>58), therefore the sequence obtained is likely to be accurate. The other repeat pattern was 11 GATC simple repeats from 25645 to 25699. There were also multiple C repeats on contig 6, but since contig 6 was not a part of the final contig, it was not a major problem.

Findid analysis indicated the presence of Tn10 transposoan on contig 1. However, contig 1 was not a part of the final assembly. Contig 1 contains only 1 read, which indicates that it might be simply due to bacterial contamination.

There were no problems in the EcoRV digest map. In the HindIII digest map, there were four fragments (3000, 1200, 1300 and 800 bp), which were identified as doublets in the real digest, but identified as singlets in the in silico digest. This is not a major problem, since band signals detected when the gels are run are not as accurate as one would hope.

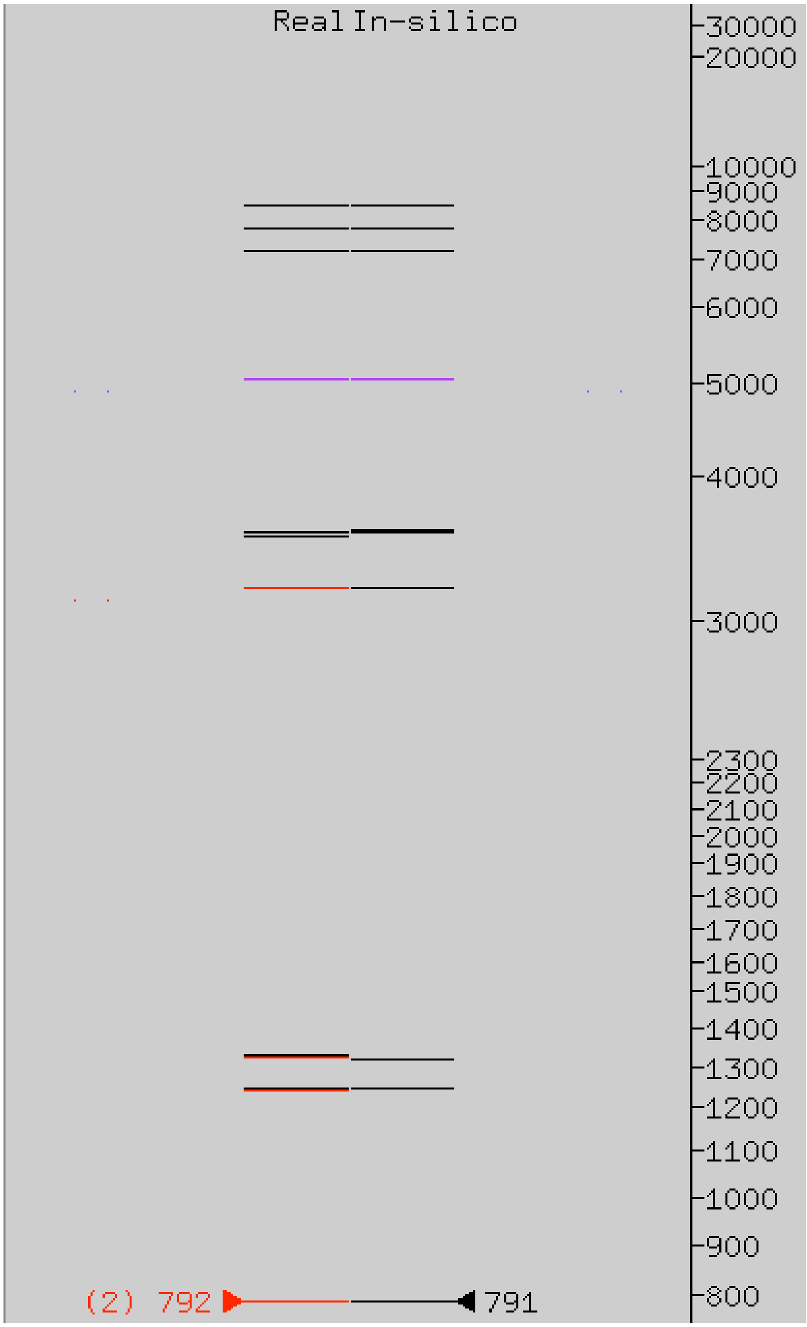


Figure 5: HindIII digest map of XAAA106 fosmid.

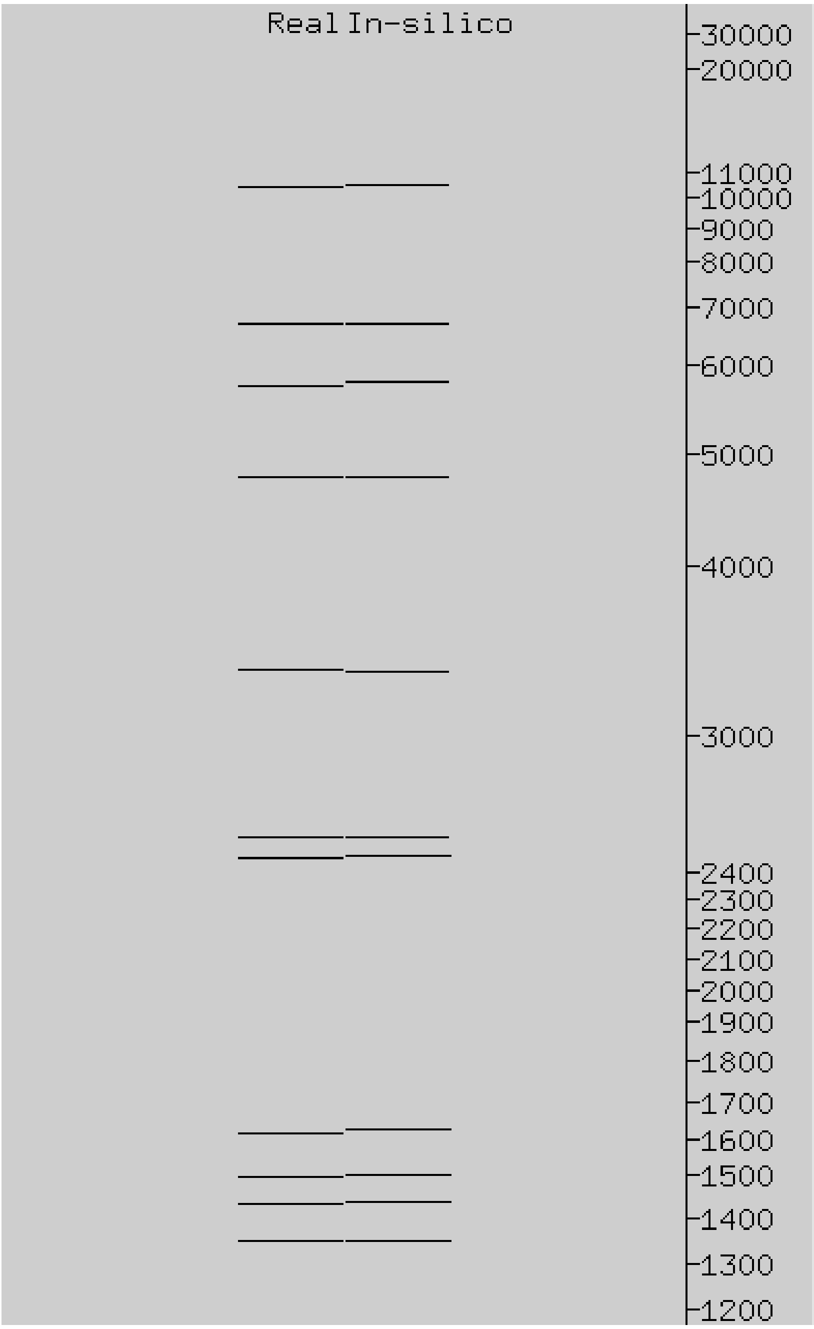


Figure 6: EcoRV digest map of XAAA106 fosmid.