Supplemental Information for "Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi"

Jade Vinson, et al.

1 Nucleotide sequence data for *Ciona savignyi*:

1.1 WGS data

Table 1: The WGS libraries used to sequence the Ciona savignyi genome. All libraries were constructed using DNA from the same individual.

		Reads (millions)			Bases	
Vector	Insert size	Attempted	Passing	Assembled	Total	Q>20
Plasmid	4.7 kb	4.39	3.72	3.04	2.32 Gb	1.96 Gb
Fosmid	40 kb	0.35	0.19	0.14	0.10 Gb	0.09 Gb

Total	--	4.74	3.91	3.18	2.42 Gb	2.05 Gb

The trace files for these reads have been deposited in the NCBI Trace Archive and are also available for download at http://www.broad.mit.edu/annotation/ciona/.

1.2 WGS assembly

The haplotype assemblies and the reference assembly are available at Genbank under the accession number AACT01000000, and also at http://www.broad.mit.edu/annotation/ciona/.

The two versions differ slightly, in that scaffolds comprising a single contig were excluded from the Genbank submission to comply with Genbank formatting policy.

1.3 Finished sequence from the WGS-sequenced individual

Table 2: Finished and nearly finished sequence from the same Ciona savignyi individual.

Locus	Accession	Phase	Note
Locus 1	AC129897	3	figure 1a, figure 2
	AC131245	3	figure 2
Locus 2	AC129899	3	
	AC129904	2	
Locus 3	AC129900	3	
	AC131244	3	
Locus 4	AC129901	2	
	AC129903	2	
Locus 5	AC130812	3	
	AC126540	3	
Locus 6	AC129896	3	
	AC126602	2	
Locus 7	AC129902	3	
	AC130813	2	

These Fosmid clones were selected manually in pairs from a preliminary diploid assembly. This preliminary assembly was created using completely different methods, which are not reported here in detail.

1.4 Finished sequence from another individual

Seven BACs were selected at random from an existing library that was constructed using sperm from 25 Ciona savignyi individuals from the San Francisco Bay (see

http://bacpac.chori.org/ciona301.htm). We generated finished sequence for these seven selected BACs. Their accession numbers are: AC092520, AC092560, AC092561, AC102146, AC117993, AC117995, and AC153355. The clone shown in Figure 1, Panel B is AC153355.

1.5 EST sequences:

We sequenced 84544 *Ciona savignyi* ESTs from several specimens collected from the Sea of Japan. The accession numbers at DDBJ are BW509979-BW594280.

Note that the ESTs were collected from specimens of a different geographic source (Japan) than that of the specimens from which the WGS libraries were collected (California).

2 Detailed assembly statistics

2.1 Read usage statistics

- 73% of passing, non-excluded reads are present in the haplotype assemblies.
- 81% of passing, non-excluded reads are present in the haplotype assemblies.
- Of the reads in the haplotype assemblies, we estimate that 23% are in the unpaired scaffolds, 40% are in the reference assembly, and 37% are in the mirror image of the reference assembly
- Thus, 62% of passing non-excluded reads are represented in either the reference assembly or its mirror, and 32% are in the reference assembly.

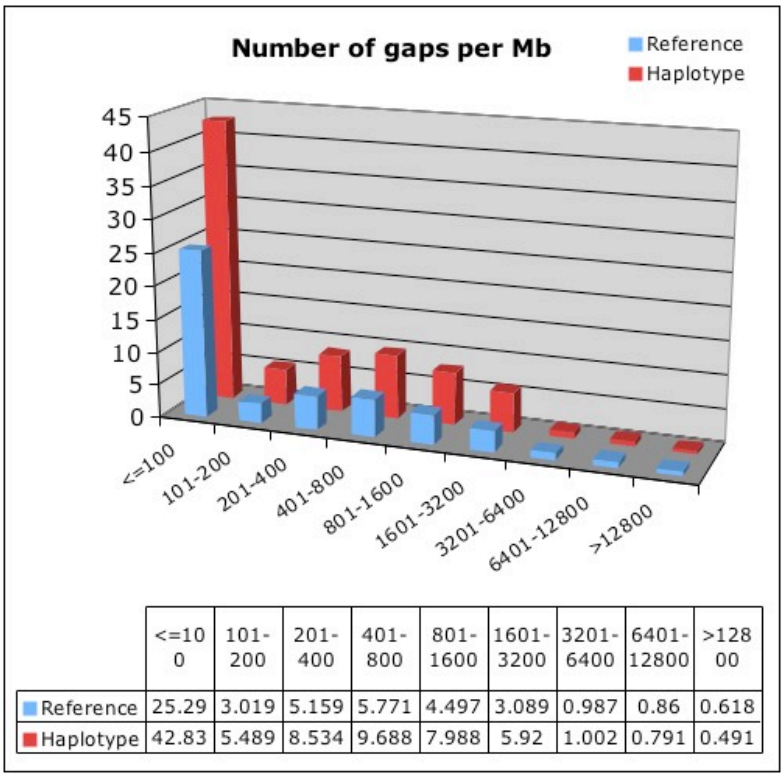
2.2 Proportion of nucleotides vs. gaps in scaffolds

The N50 scafffold sizes listed in Figure 4 include only base-pairs in contigs not the gaps (the unpadded lengths). Of the padded lengths, the proportions corresponding to pads, for the various sets of scaffolds, are:

- 7.0% for the default assembly
- 5.6% gaps for haploid assembly
- 4.3% for reference assembly
- 6.2% for unpaired scaffolds
- 6.6% for mirror image of reference assembly

2.3 Gap size chart

For the reference assembly and the haplotype assembly, we calculated the frequency of gaps, per megabase, and then plotted a histogram of the gap sizes. See below.



	<=100	101-200	201-400	401-800	801-1600	1601-3200	3201-6400	6401-12800	>12800
Reference	25.29	3.019	5.159	5.771	4.497	3.089	0.987	0.86	0.618
Haplotype	42.83	5.489	8.534	9.688	7.988	5.92	1.002	0.791	0.491

Note that all gaps of length <=100bp have been rounded-up to be exactly 100bp in the released assemblies.

2.4 Haplotype segment statistics:

We define a haplotype segment in the reference assembly as a contiguous segment nucleotide sequence in the reference assembly that originates from the same contig in the haplotype assemblies. Thus:

- Each contig in the reference assembly is either a single haplotype block or a concatenation of several haplotypes segments.
- The total length of the haplotype segmentss is equal to the total length of contigs in the reference assembly.

The length statistics of the haplotype segments are:

- Total length is 157151546
- Number of haplotype segments: 13490
- The N50 length is 21421
- The median length is 6751

3 Detailed analysis of EST alignments

3.1 Overview

We 84544 sequenced *Ciona savignyi* ESTs. We filtered these ESTs to exclude a total of 10821 ESTs for one or more of the following reasons: EST is mitochondrial (10246), EST is a duplicate sequence of a clone already sequenced (242), or the EST is short (365). The number of non-exlucluded ESTs, 73723, becomes the denominator for the following analyses.

We aligned each of the 73723 non-excluded ESTs to each of the following sets of sequence, using BLAT with default parameters:

1) the reference assembly
2) the unpaired scaffolds
3) the haplotype assemblies
4) the WGS reads not assembled in the haplotype assemblies
5) the default Arachne assembly.

For each EST we recorded the number of times it aligns to each set over 80% of the length of the EST, whether zero times, once, twice, or more than twice. Thus, we obtain a joint table with 3^5 = 243 entries whose sum is 73723.

The marginal values of this table (the exact values corresponding to Figure 4 in the main text) are:

	Reference assembly	*Unpaired scaffolds*	*Haplotype assemblies*	*Unassembled reads*	*Default Arachne*
0	9027 12.24%	63601 86.27%	3740 5.07%	68185 92.49%	5915 8.02%
1	59844 81.17%	6840 9.28%	9675 13.12%	2547 3.45%	14366 19.49%
2	2238 3.04%	1472 2.00%	53248 72.23%	869 1.18%	43252 58.67%
>2	2614 3.55%	1810 2.46%	7060 9.58%	2122 2.88%	10190 13.82%

Nearly all of the ESTs aligning to the reference assembly also align to the haplotype assemblies, to which 95% of ESTs align. For completeness, the haplotype assemblies are also available for download alongside the reference assembly. See Genbank accession number AACT01000000 or the Broad website http://www.broad.mit.edu/annotation/ciona/. In addition, the unpaired scaffolds are available for download at the Broad website.

Of the 3740 of ESTs not aligning to the haplotype assembly, most (3413) align to none of the categories listed above (including the unassembled reads). These nonaligning ESTs are due to either aligner inefficiencies, sequence not present in the sequenced individual due to strain differences, or contamination within the EST libraries.

There are 9027 ESTs (12.2%) that did not align to the reference assembly, those of greatest concern are the 5375 (7.3%) that do align to the haplotype assemblies.

Of the 5375 ESTs (7.3%) aligning to the haplotype assemblies but not to the reference assembly, 3116 (4.2%) are in the unpaired scaffolds. We used very conservative criteria used for pairing haploid scaffolds with their partner of the opposite haplotype (described in Assembly Methods), leading to a high proportion of unpaired scaffolds for which we could not unambiguously determine the haplotype partner. These conservative criteria use both local similarity and long-range synteny to minimize the possibility of artificially duplicating loci (presenting both haplotype A and haplotype B in the reference assembly) or incorrectly associating two paralogs (only presenting one of the two paralogs in the reference assembly). More aggressive criteria would have had the opposite trade-offs – few unpaired scaffolds, but a high rate of artificial duplication or artificial collapse of loci.

Of the 5375 ESTs aligning to the haplotype assemblies but not to the reference assembly or unpaired scaffolds, there are 1514 ESTs (2.1%) which align to the mirror-image of the reference assembly, but not to the unpaired scaffolds. Because these ESTs align to fewer raw WGS reads than other ESTs (see below), many of them may represent sequence that is specific to only one of the two sequenced haplotypes. (The mirror image of the reference assembly is not discussed in our manuscript, but is also available for download at http://www.broad.mit.edu/annotation/ciona/.) [Note that if we only ask how many ESTs align to the mirror image of the reference assembly but not to the reference assembly, there are 2178 such ESTs (3.0%).]

Of the 5375 ESTs aligning to the haplotype assemblies but not to the reference assembly, there are 745 ESTs (1.0%) that align to neither the unpaired scaffolds nor the mirror image of the reference assembly. These may represent regions fragmented by the choice of a single haplotype path through each diploid scaffold to present as the reference assembly. These loci represent prime targets for future algorithmic improvement – until now the ESTs have not been used to guide the assembly, only to test the completeness and the unique representation of loci in the assembly.

3.2 Analysis of ESTs that align multiply, or only to the non-chosen haplotype

We seek to determine whether ESTs aligning twice to the reference assembly represent actual duplicated loci in the genome, or loci duplicated artificially due to the assembly process (e.g. both alleles are present in the haplotype assemblies). Our analysis indicates that these are real, because they align to the raw WGS reads about twice as often as the ESTs which align exactly once to the reference assembly.

We similarly analyzed the ESTs which align to the mirror image of the reference assembly but not to the reference assembly itself. We find that these have about 75% as many alignments to WGS reads as is typical, suggesting that some of these represent sequence that is specific to one of the two haplotypes, and some represent assembly artifacts.

Note that approximately one third of ESTs have no alignment directly to WGS reads over 80% of their length, probably because the ESTs represent multiple exons whose total transcript length (including introns) exceeds the size of typical WGS reads. As shown in the Figure below, the conclusions of the analyses are roughly the same whether or not such ESTs are included. We exclude as outliers those ESTs aligning more than 50 times to the WGS reads, because some aligned thousands of times and skewed the mean.

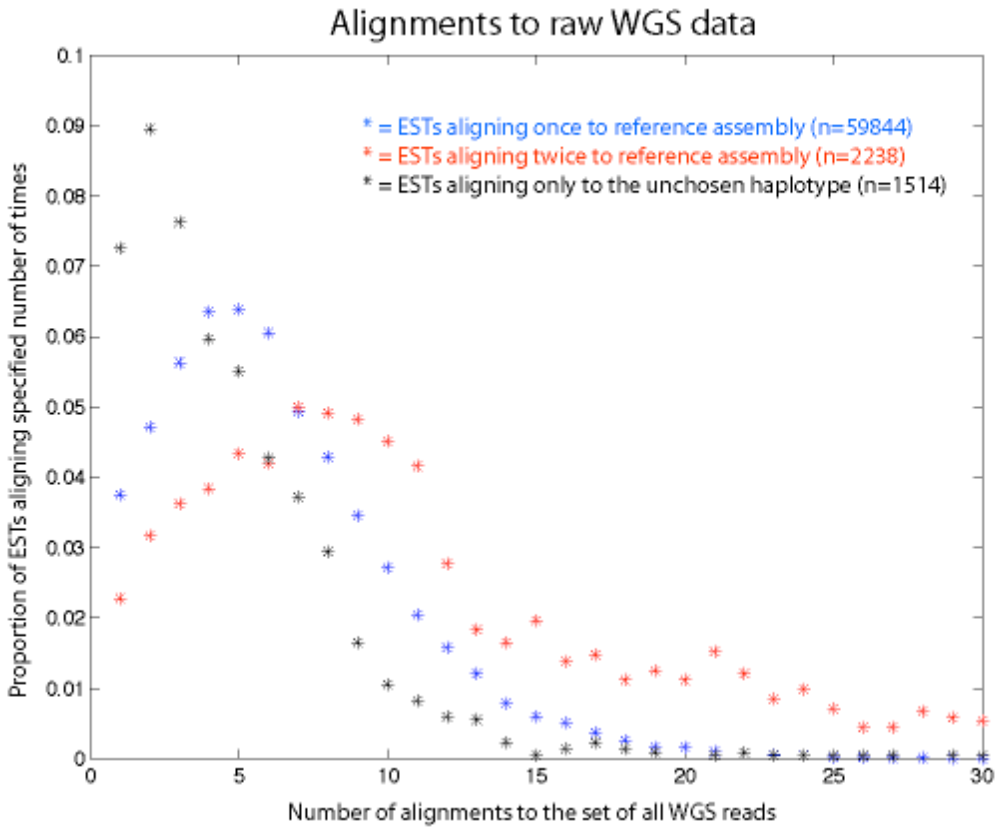


Figure above: Histogram of the number of alignments to WGS reads for ESTs that align exactly twice to the reference assembly (red), ESTs that align exactly once to the reference assembly (blue), and ESTs that align to the mirror image of the reference assembly, but not the reference assembly itself. We exclude as outliers ESTs aligning more than 50 times (some align thousands of times and skew the mean). Including the zero point, the three means are: 2.89 (black) 3.64 (blue) 8.34 (red), a ratio of 0.79 - 1.00 – 2.29. Excluding the zero point, the three means are 4.97 (black) 6.44 (blue) 11.64 (red), a ratio of 0.77 – 1.00 – 1.81. These results are consistent with most of the ESTs aligning twice to the reference assembly representing actually duplicated loci, and nearly half of the ESTs aligning only to the mirror image representing loci that are present in only one of the two sequenced haplotypes.

4 Validation using finished sequence

Table 3: Finished and nearly finished sequence from the same Ciona savignyi individual.

Locus	Accession	Phase	Length	Aligned
Locus 1	AC129897	3	38707	32823
	AC131245	3	39400	38063
Locus 2	AC129899	3	38548	38572
	AC129904	2	32803	29481
Locus 3	AC129900	3	38504	38504
	AC131244	3	43713	42113
Locus 4	AC129901	2	38703	35035
	AC129903	2	40689	38372
Locus 5	AC130812	3	40437	39985
	AC126540	3	39109	38850
Locus 6	AC129896	3	39512	39521
	AC126602	2	35567	34819
Locus 7	AC129902	3	38410	38155
	AC130813	2	38600	38264
TOTAL:			**542702**	**522557**

These Fosmid clones were selected manually in pairs from a preliminary diploid assembly. This preliminary assembly was created using completely different methods, which are not reported here.

5 Further analyses of repetitive sequence

In addition to the method presented of measuring repetitive sequence content that was described in the main text (method 1), we present below a second independent analysis based on over-occurring 48mers (method 2) that confirms the original result that the unpaired scaffolds are highly enriched for repetitive sequence.

For this, we listed all 48mers which occur more than 1000 times in either the Ciona savignyi WGS trimmed reads or their reverse complement. Since the coverage is 12.7x, this means that the 48mer is present about 40x more often than is expected at random.

We used this set of overoccurring 48-mers, which were based solely on the WGS reads, to mask the assemblies. A base is masked if it is part of one of these 40x overoccurring 48mer, thus masking only the most highly repetitive sequence. We applied this test to the

reference assembly, the unpaired scaffolds, and the unassembled reads. Their repetitive content by this measure is:
Reference assembly: 2.2%
Unpaired scaffolds: 17.9%
Unplaced reads (not placed in haplotype assemblies): 62.2%

Thus, by this measure, the repeat content of the unpaired scaffolds is eight times greater than the repeat content of the reference assembly.

6 Rate of heterozygous insertions and deletions:

To calculate the rate of insertions and deletions, we counted the number of gaps within the set of BlastZ alignments scoring >100,000 between finished clones of opposite haplotype.

There were 1689 distinct gaps interspersed within 158k aligning bases, or 1 every 94 bp, which was rounded up to 1 every 100bp because the sample size is too low to have two significant figures.

7 Estimate of genome size

We estimate that the euchromatic genome size of *Ciona savignyi* is about 190 Mb, with two methods yielding similar estimates. First, assuming from the EST alignments that the reference assembly represents 88% of the euchromatic sequence, so the overall size can be calculated as ~186 Mb =164 Mb/(0.88). Second, the largest haploid scaffolds all have a depth of coverage tightly clustered around 8480 assembled sequencing reads per Mb. Since the total number of assembled reads is 3.21 million, this leads to a size estimate of 189Mb per haploid copy. Since there were a total of 3.91 million passing reads, of which 82% were assembled in the haplotype assemblies, the total genome size (for which there is no independent estimate) could be as large as 230 Mb (= 3910000*1000000/8480), much larger than the euchromatic portion.

8 Confirming large-scale haplotype differences using PCR

8.1 Overview:

See text of paper for context.

We listed all examples of a collinear block of alignments ending at least 40kb before the end of either of the haploid scaffolds it relates and, based on dot-plot comparisons of the haploid scaffolds at these loci, manually selected ~20 critical junctions – points at which a collinear block terminates far from the end of either of the haploid scaffolds it relates – that manifest varied examples of large-scale differences, such as inversions over 100kb.

At this stage, a critical junction is comprised of a position and orientation on two haploid scaffolds so that they agree to the left and disagree to the right.

We then list all Fosmid clones spanning the junction in either haploid scaffold (Fosmids were the 40kb WGS library), based on the placement of end-reads in the haploid scaffolds. We manually select two 40kb Fosmid clones from each haploid scaffold so that in each scaffold, both Fosmids extend >10kb in either direction from the critical junctions (and note an exception if in one direction only one Fosmid clones from a haplotype extends >10kb), and so that the end reads of the two Fosmids from a haplotype are not coincident (about 13% of the Fosmid clones in our WGS library are exact duplicates, perhaps double picked from the selection plate; this way we don't test the same Fosmid clone twice). In addition, 30% of the glycerol stocks for our Fosmid library had been misplaced, and we had to select Fosmid clones from the remaining 70%. 14 critical junctions are acceptable by these criteria and we proceeded with their analysis.

If the region of agreement, we selected two assays that should amplify in all four clones; in the region of disagreement we selected two assays specific to haplotype A and two assays specific to B. Because of the high heterozygosity, both haplotypes must be considered in the design of the primers for the shared assays; see Subsection 8.2.

The experimental design and ideal results are described in Subsection 8.3.

We made predictions for a pilot experiment of 2 critical junctions, tested, analyzed data refined algorithm for picking primers and added sequencing to the experimental design, designed tests and made predictions for 12 more junctions. See attached data file.

We then tested the 12 junctions of the main experiment, for a total of 14 critical junctions See Subsection 8.4 for photographs.

We interpreted the PCR product lengths blindly (without consulting the predictions). See attached data file. We then recorded all deviations from the predicted results (Subsection 8.6). Most of these deviations were easily attributable to lab error (for example, a Fosmid clone completely missing, in that none of its products were present; we retested a few cases to confirm this suspicion), and such cases are recorded in Figure 5 as a lack of data.

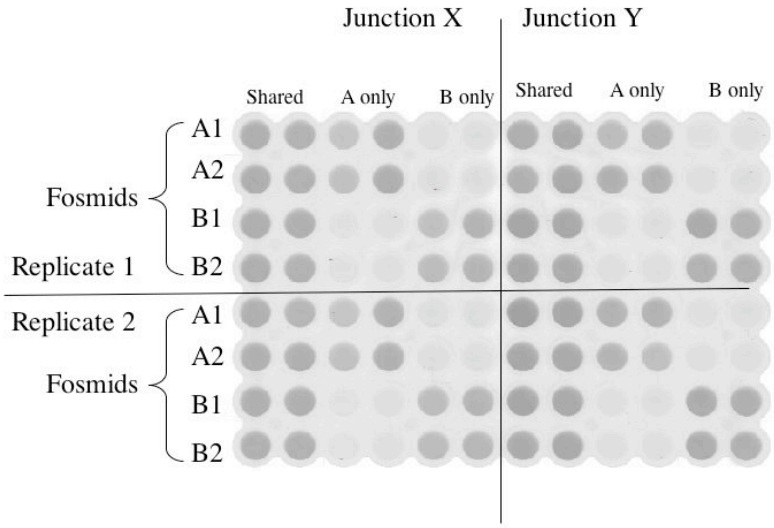
8.2 Method for picking primers

We wrote a script which picks PCR primers, based on the predicted sequence of the Fosmid clones, at random based these criteria:

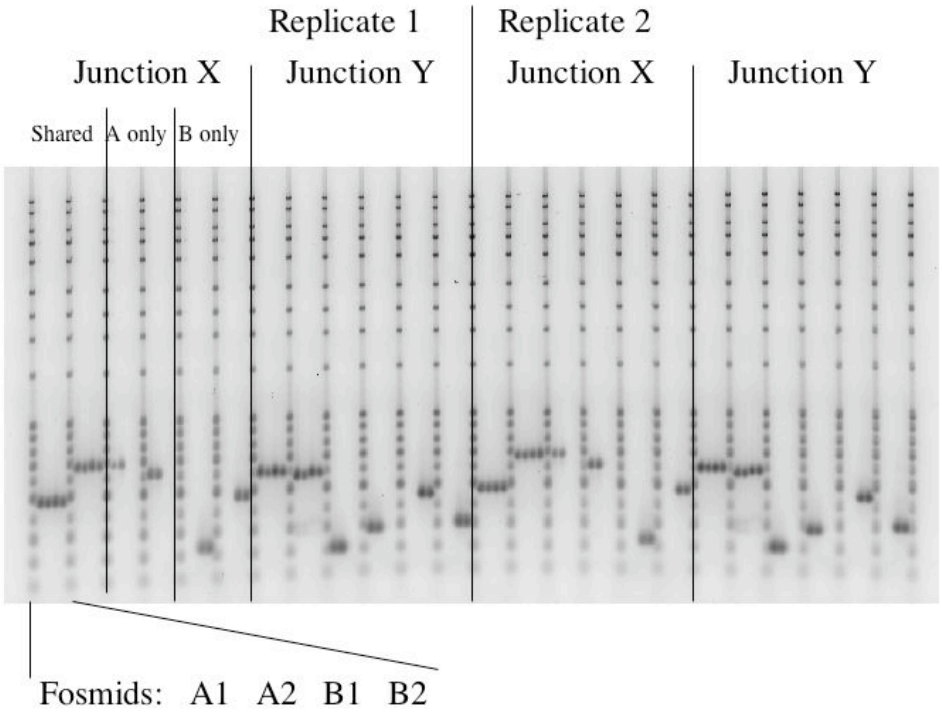
1) Ideal size for the PCR product (the distance between the outer ends of the primers) is 200-1000 bp. (Sometimes this condition was relaxed if we couldn't find primers)
2) Exactly 22 bp, of which 10/22 are a G or C.
3) The 3' base, and two of the last four, should be a G or C.
4) No four in a row of the same nucleotide.
5) For assays that should amplify in all four clones, the primer sequences must be identical in the two haploid scaffolds. Note that, at a 4.6% heterozygosity rate, this is a very restrictive condition.
6) No perfect 8-mer matches of the 3' end to anywhere else in the predicted sequence of the clone. (Sometimes relaxed to 9 or 10 if needed to find primers).
7) The 3 bases at the 3' end do not reverse-complement anything in the primer.
8) We pick primers which do not have gap-free imperfect matches to other locations in the predicted clone sequence; by counting the maximum number of matches to 22bp intervals in the predicted clone sequence or its reverse complement, and seeking to pick the primer to minimize this. In most cases, we got down to 16/22 or lower.
9) We do not pick primers within 1kb of either the critical junction or the outermost edge of the fosmid end reads in the haploid assembly, since, if we had done so, the resulting assay would have been uninformative.

8.3 Experimental design and ideal results:

The primers and fosmids for two junctions and two replicates are arranged on a 96 well plate (or a quadrant of a 384-well plate) as follows, and an example of ideal results (hand-picked from the actual data) are also indicated:

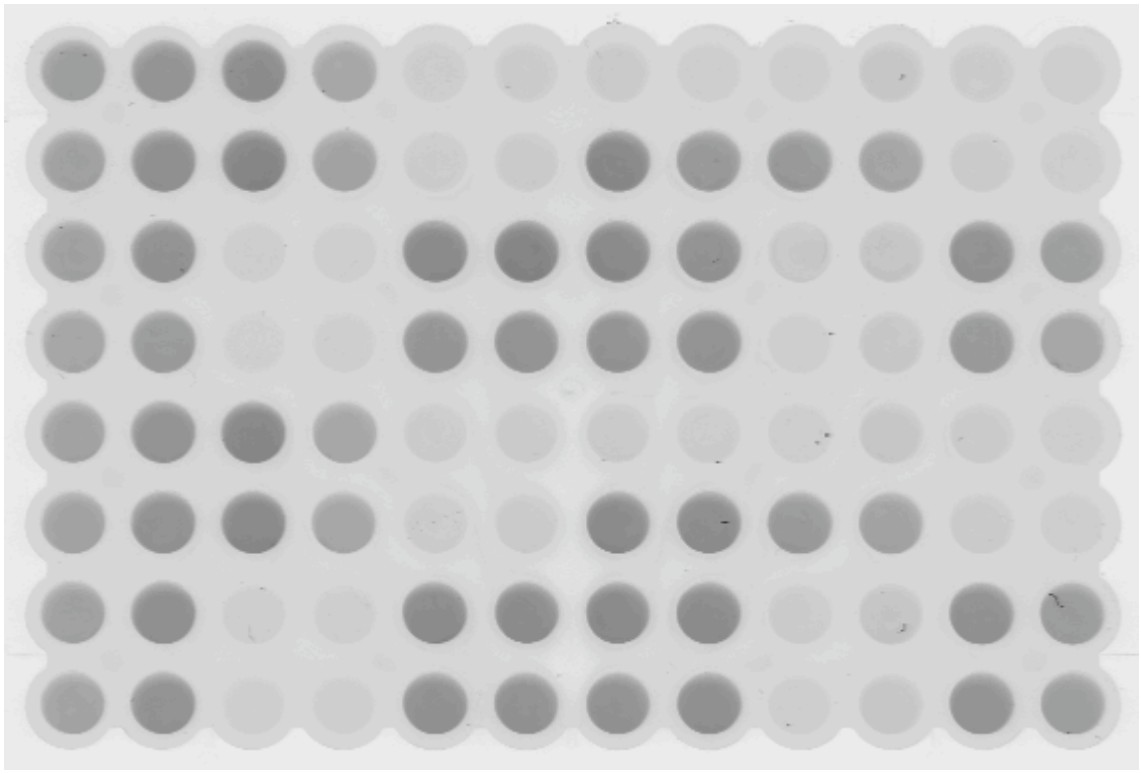


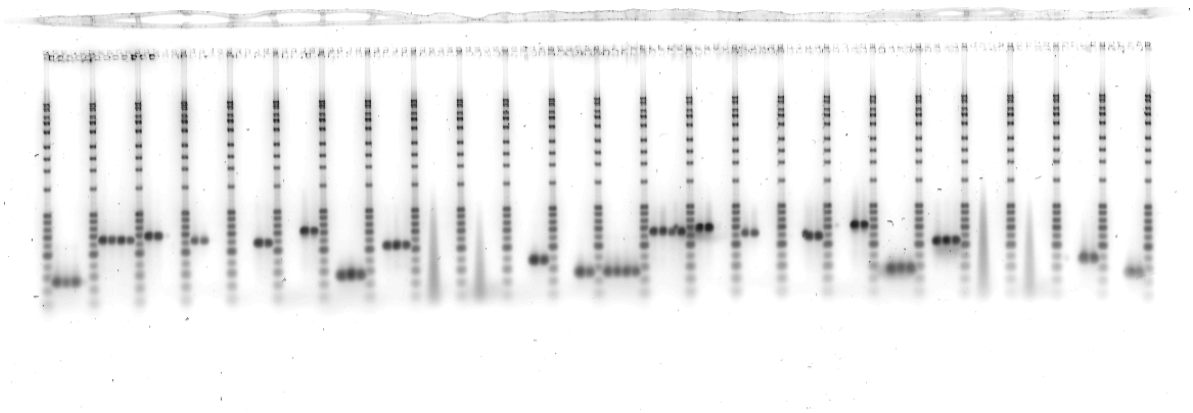
For each of the reactions in which we predict a product, we also predict a length. We check the length by running out the products of all reactions on a gel, in the following order:



8.4 Detailed raw results (pictures), junction by junction

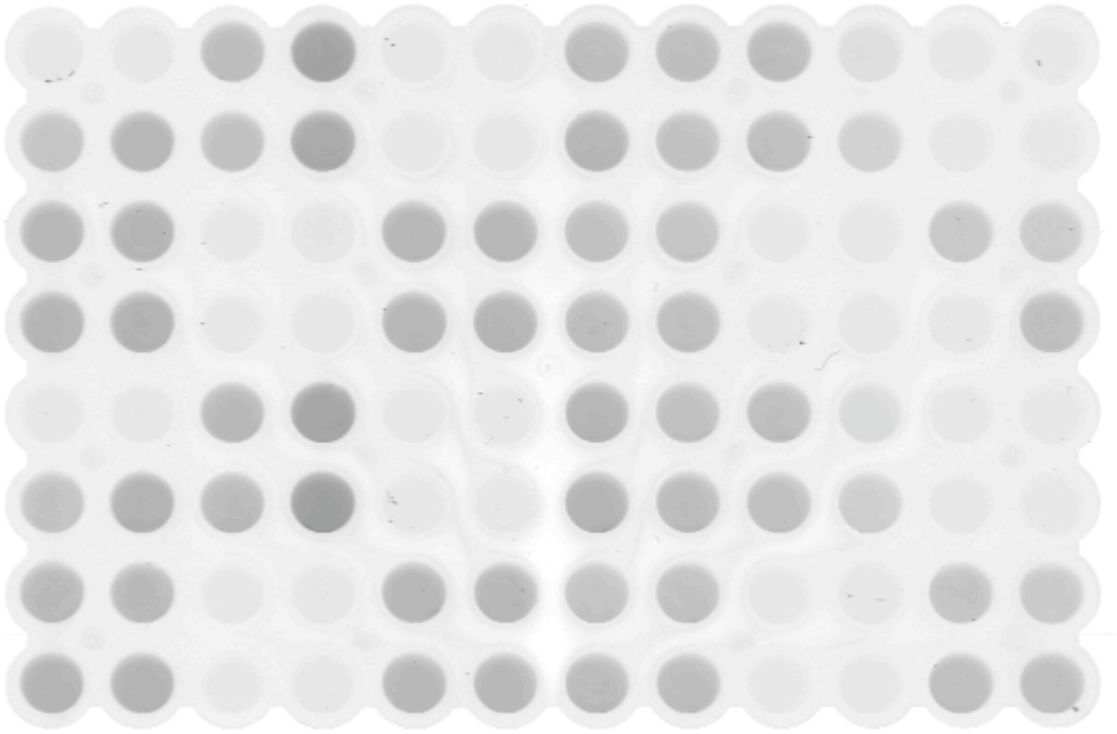
8.4.1 Junctions Pilot1 and Pilot2:



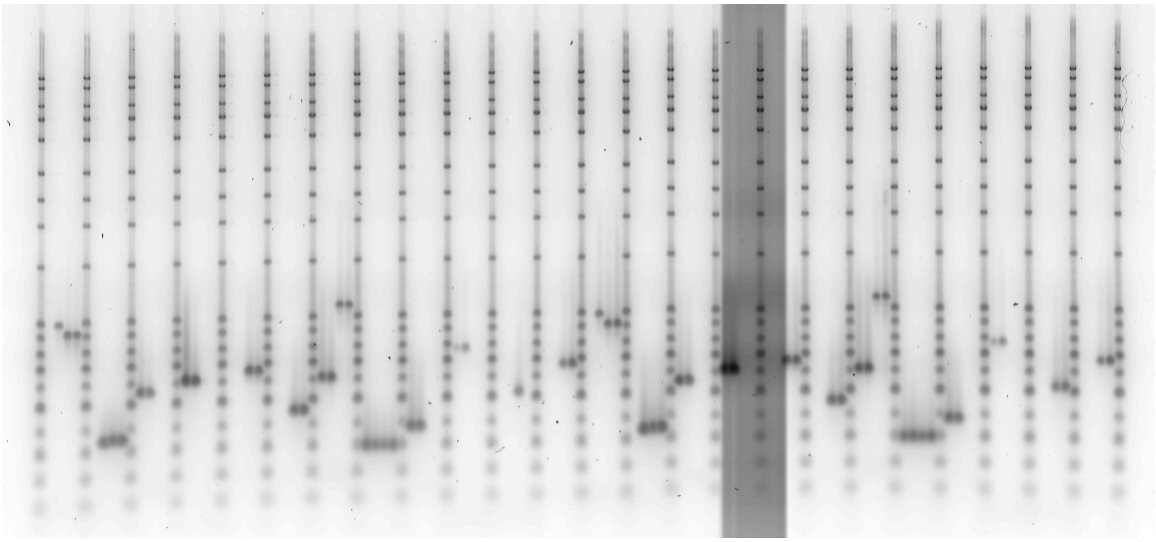


8.4.2 Junctions 1 and 2:

Ciona PCR pico Quad A1 of 1st 384-well plate 9-2-04

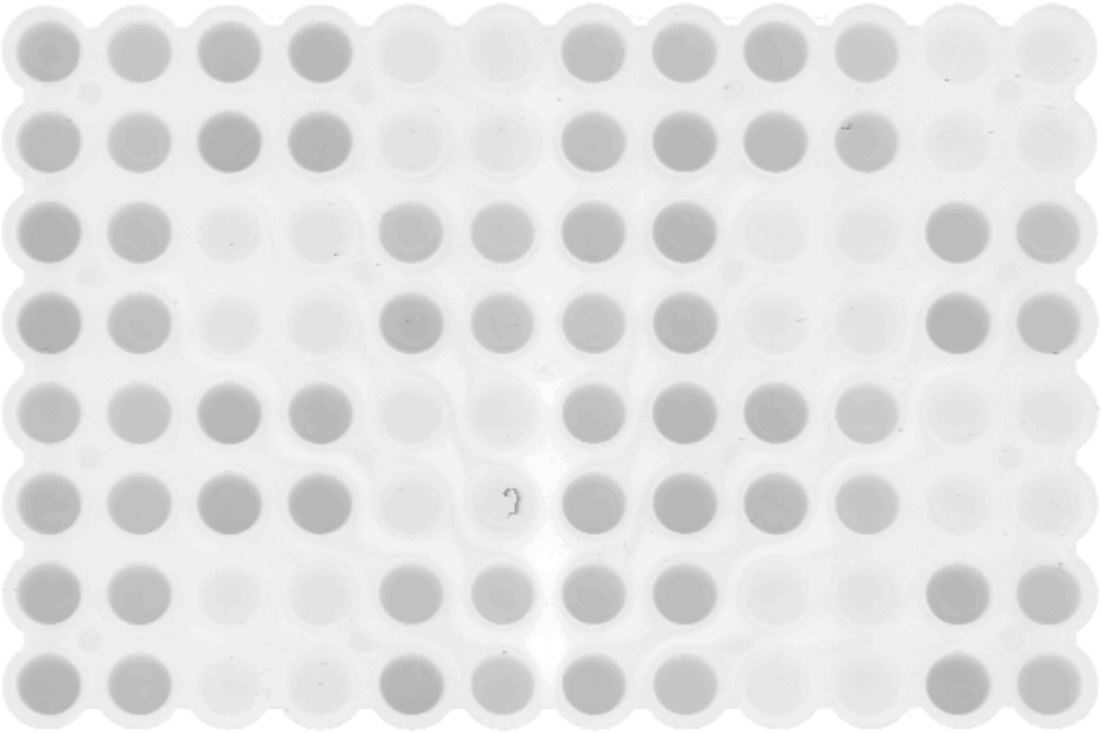


Ciona PCR gel image Quad A1 of 1st 384-well plate 9-2-04

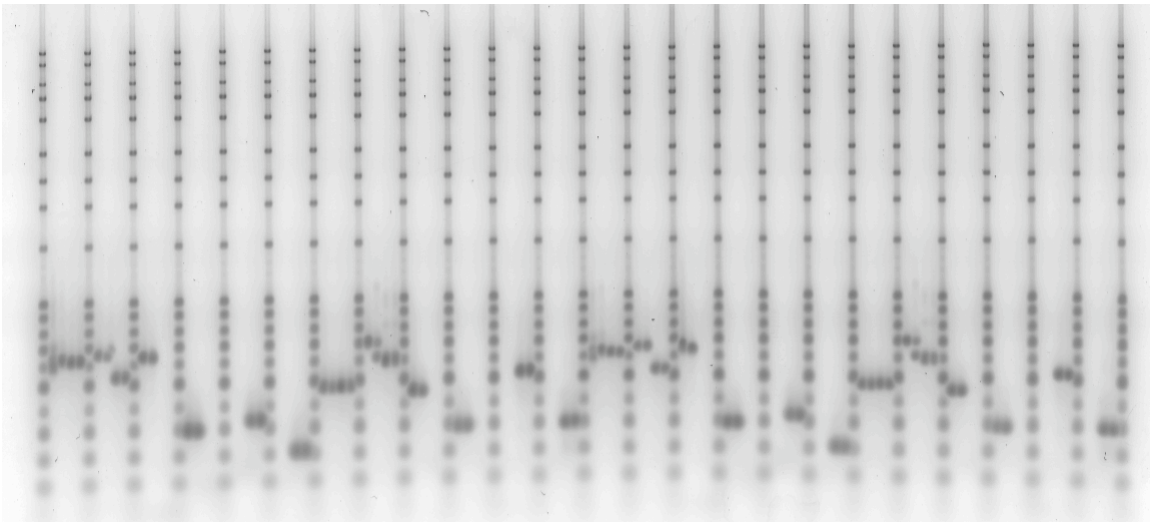


8.4.3 Junctions 3 and 4

Ciona PCR pico Quad A2 of 1st 384-well plate 9-2-04

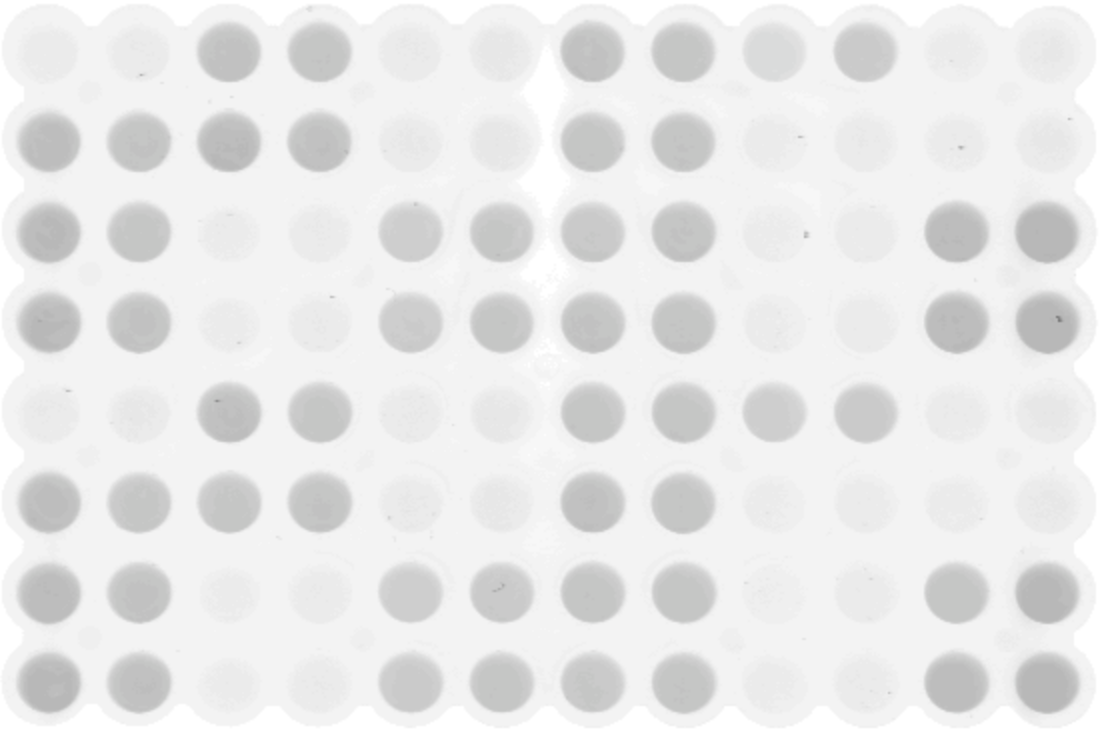


Ciona PCR gel image Quad A2 of 1st 384-well plate 9-2-04

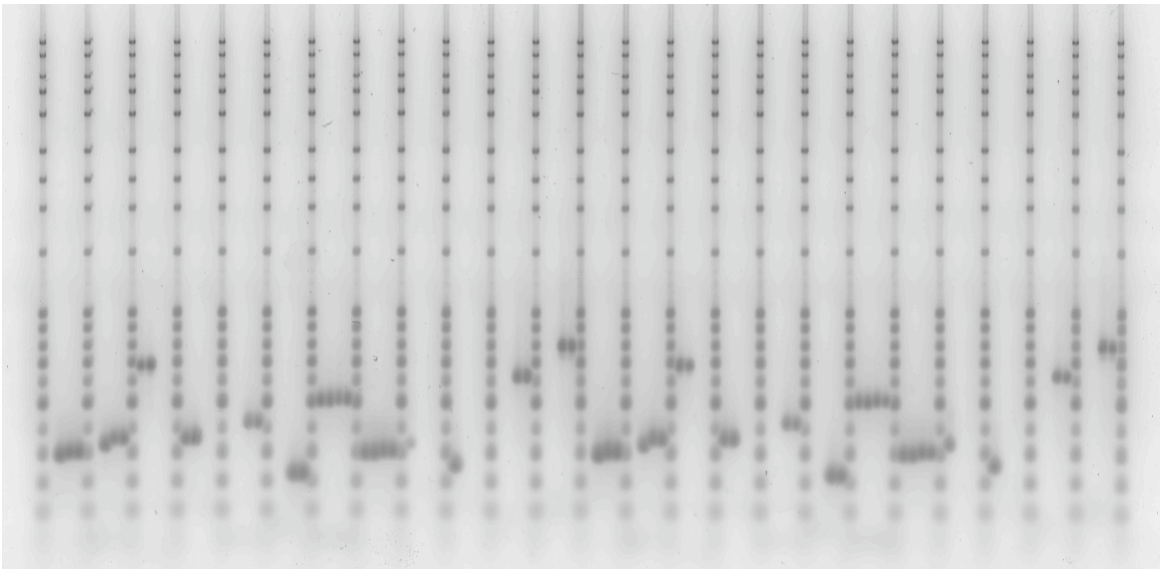


8.4.4 Junctions 5 and 6

Ciona PCR pico Quad B1 of 1st 384-well plate 9-2-04

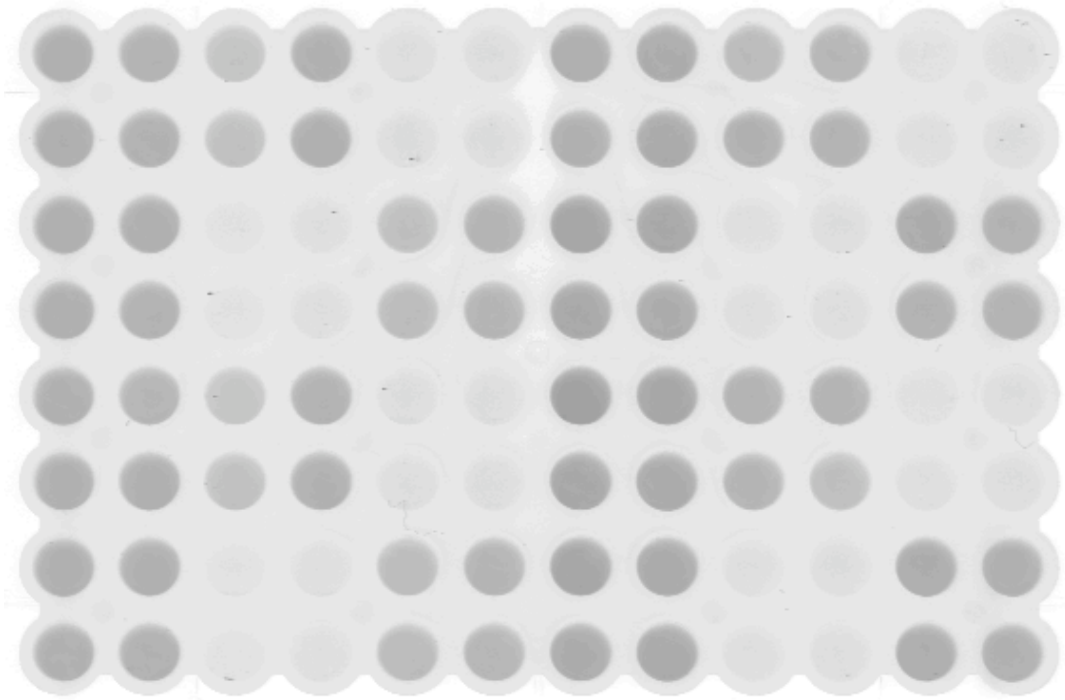


Ciona PCR gel image Quad B1 of 1st 384-well plate 9-2-04

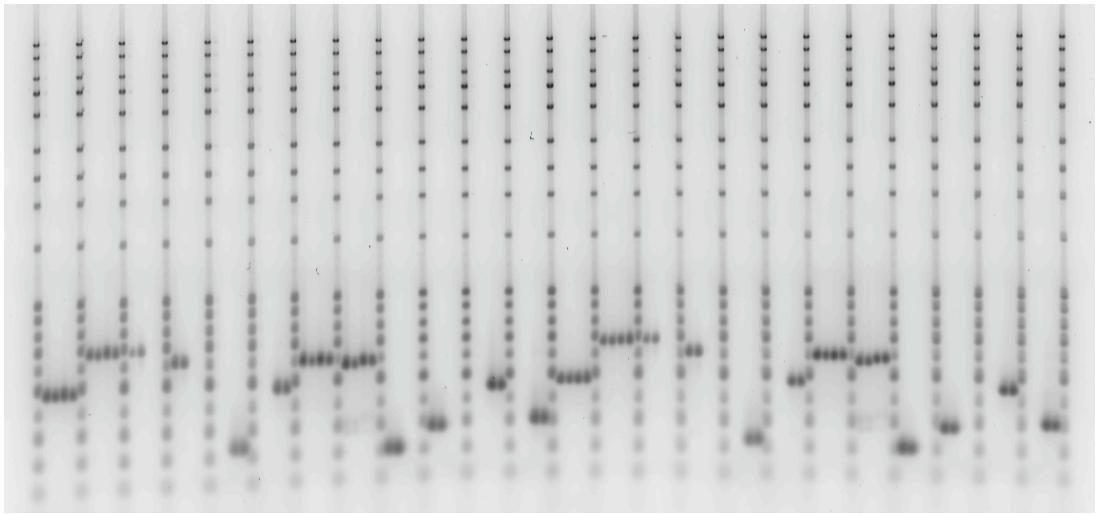


8.4.5 Junctions 7 and 8

Ciona PCR pico Quad B2 of 1st 384-well plate 9-2-04

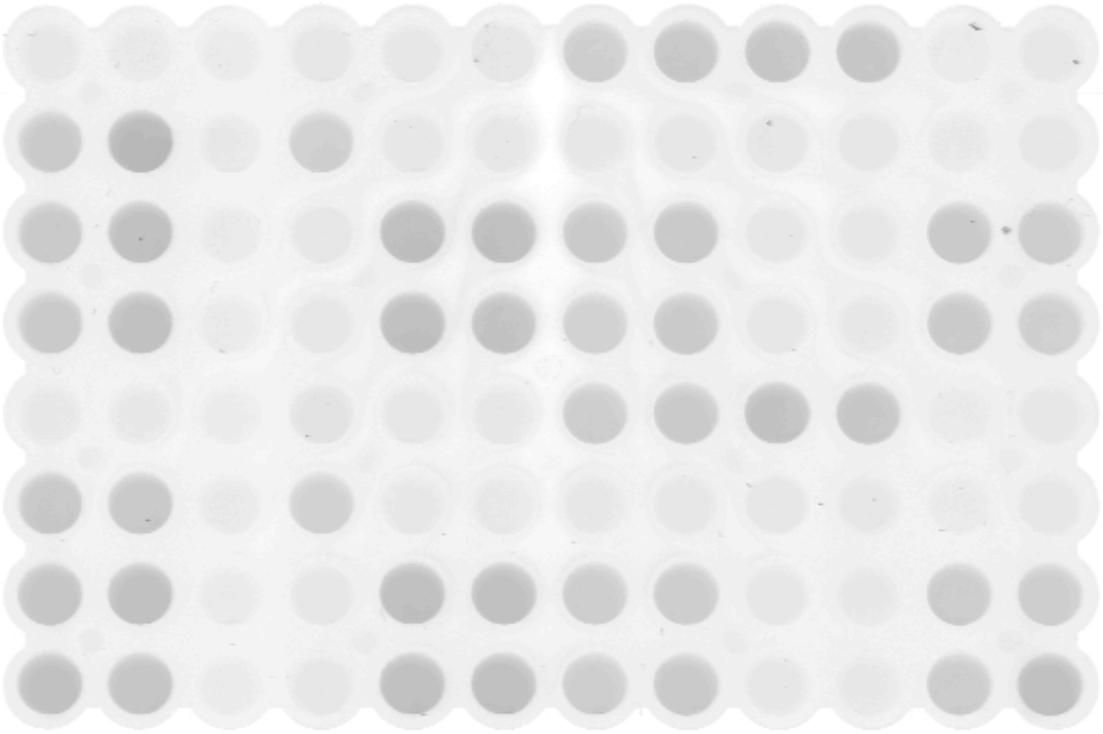


Ciona PCR gel image Quad B2 of 1st 384-well plate 9-2-04

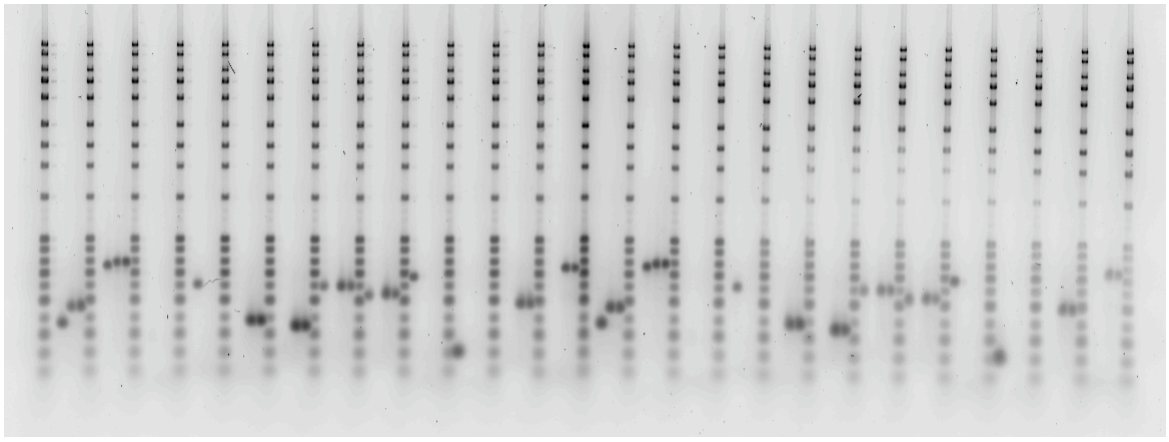


8.4.6 Junctions 9 and 10

Pico image of 2nd 384-well plate Quad A1 for Ciona PCR 9-3-04

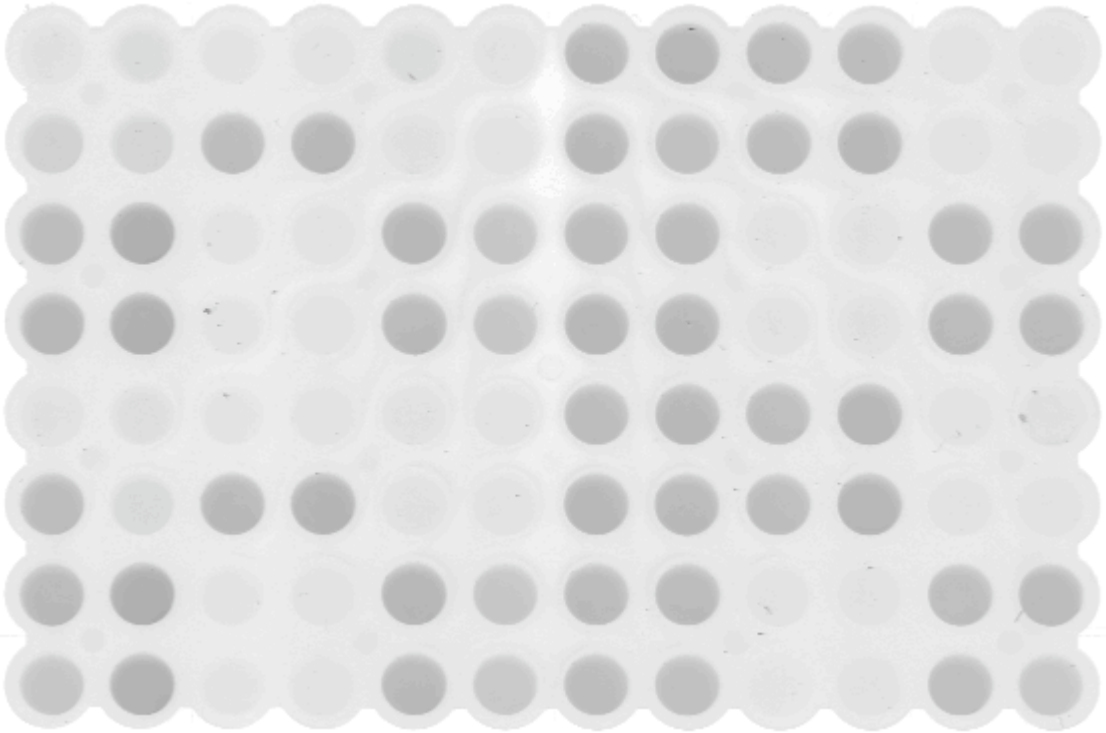


Ciona PCR 384 #2 Quad A1 gel image 9-3-04

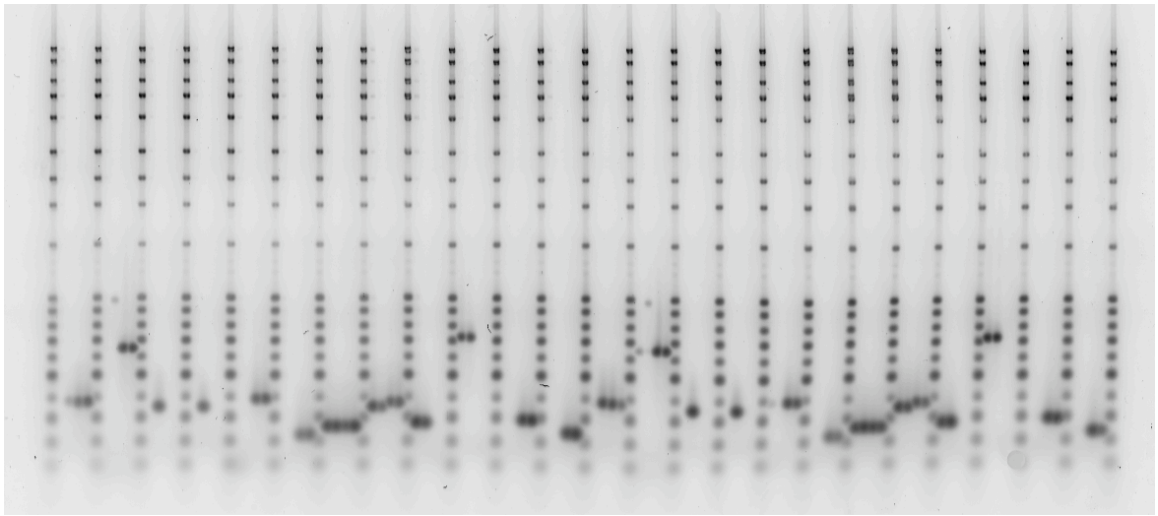


8.4.7 Junctions 11 and 12

Pico image of 2nd 384-well plate Quad A2 for Ciona PCR 9-3-04



Ciona PCR 384 #2 Quad A2 gel image 9-3-04



8.5 Sequencing and alignment of PCR products

We also sequenced all of the PCR products, and aligned these sequences to the two haploid assemblies. The sequence should align best to the haploid scaffold corresponding to the Fosmid clone from which the PCR product was amplified. This tests whether the two haploid scaffolds might be crossing over, getting the haplotypes mixed up, in between the Fosmid end reads. This test is most informative for the shared primers in the case that the PCR product sizes are nearly equal. Sometimes for the shared primer, even the sequences of the PCR products are identical, and the sequence alignment test is uninformative (e.g. the PCR product aligns equally well to the predicted sequenced of both haplotypes). Such uninformative cases are recorded in Figure 5 as no data.

Sequences of PCR products were aligned to the haploid scaffolds using Blastn using default parameters; we observed whether the top-scoring alignment was from the correct haplotype. There are 16 products giving a sequence, but they come in 8 pairs of identical sequences, so that's 8 tests not 16 which we indicate in Figure 5.

We ignored sequence data for PCR products with no visible band, because the sequencing machines are extremely sensitive and will produce sequence data corresponding to trace contaminants.

We sequenced the PCR products for 10 of the 14 critical junctions: this had not been part of the experimental design for the 2 junctions from the pilot experiments; and the sequences from 2 of the 12 junctions of the main experiment were lost due to lab error.

8.6 Differences between predictions and observations:

Unless noted below, everything was perfect and the observation exactly matched the prediction in at least one of the two replicates.

The two discrepant PCR products (these comprise the four black squares in Figure 5):

1) The "double band": Location: junction 4, shared primer 2, fosmid A2. Prediction: 701bp. Gel: double band at approx 710bp and 610bp in both replicates, where 610 is the same length as the B haplotype product. Sequence: aligns well to draft fosmid A2 but not A1.
2) The "smeared band": Location: junction 3, shared primer 1, fosmid A1. Prediction: 618bp. Gel: band smeared over 150bp, center at 610bp, same as other fosmids or maybe a little shorter. Sequence: aligns well to fosmid A1 but not A2.

These discrepant products are not lab errors; they are present in both replicates and have been confirmed by retesting. The two discrepant situations are very similar, and occur at two different critical junctions at opposite ends of a 280kb inversion. My guess (the only explanation of the observations below that I can think of) is that in both cases PCR is

amplifying at multiple locations on the Fosmid due to short duplications within the Fosmid clone, and that this happens twice independently.

The two cases were very similar. In both cases, there is a shared primer pair which amplifies in one of the A-haplotype Fosmids to produce something other than a sharp band of the predicted length (in one case fuzzy 150bp band centered at the correct length; in the other a double band with one of the correct length). The other A-Fosmids and the two B-Fosmids produced products of the expected length using this primer pair. Surprisingly, the PCR products produced sequence, and this sequence aligned well to the abnormal A-Fosmid but to neither the normal A-Fosmid nor the two B-Fosmids.

These segmental duplications are visible in dotplots comparing the two haploid scaffolds. It appears from the dotplots that the abnormal A-Fosmid extends into the second (distracting) copy of the segmental duplication whereas the normal A-Fosmid does not. Given that there are segmental duplications at either end of this inversion (and they are related), there is no way to be confident of its correct assembly, even with the PCR results. It is also, biologically, a very likely spot for an inversion.

Deviations of PCR product sizes from prediction, resolved through retesting:

1) Fosmid A1 is missing from Pilot2
    a. This fosmid was regrown and tested against the six primer pairs. It behaved just as did the A2 fosmid: it amplified specifically in the two shared primer pairs, non-specifically for the A primer pairs, and no product for the B primer pairs.
2) J2-PA2-FA1: observed and predicted length are very different.
    a. Predicted length is identical to predicted length of previous interaction. I checked, it was a typo, and correcting it resolved the discrepancy.
3) J7-PA2-FA1: I predicted 546 and observed 610.
    a. I checked the sequence between the primers. There was a contig gap, and contig gaps are only estimates to begin with.
    b. The sequence from the PCR reaction aligned well.
4) The "faint band": Location: junction 11, shared primer 2, fosmid A1.
    a. Prediction: 985bp. Gel: entire fosmid was missing for all primers, so nothing expected. However, in one replicate, we observe a faint band of length 640, the predicted and observed length for the B haplotype. Sequence: not available. Guess: contamination.
    b. Retested and observed no product.

Deviations of PCR product sizes from prediction, but seems pretty clear what happened and not retested

1) Pilot2 A1 and A2 primers amplify nonspecifically in the fosmids FA1 and FA2.
    a. These primers were picked in a microsattelite. In retrospect, that is clear from the sequence.

b. Modified primer picking algorithms to avoid this in main expt.

2) FA1 seems to be completely missing from J9.
3) FA1 seems to be completely missing from J11.
    a. Observation: all Fosmids that have been completely missing have been from position FA1.

Minor exceptions of PCR product sequences not aligning best to the correct haploid scaffold using Blastn at default settings:

1) Uninformative alignments of sequences of PCR products:
    a. That is, the blastn scores of the alignments to the draft A and B sequences are identical.
    b. Junction 2, shared primer 2.
    c. Junction 8, shared primer 1.

1) At Junction 2, shared primer 1, the sequences for the fosmid B1 haplotype had a higher score aligning to the draft A1 and A2 fosmids than draft B1 and B2 fosmids. However, a closer analysis at the sequence level found 9 heterozygous positions at which the sequence of the PCR product agrees with the draft B but not draft A haplotypes. Further, I did a multiple alignment of the predicted and two observed sequences, and saw near perfect agreement along the entire length excepting the very endpoints. The sequence for the fosmid FB2 had very close scores, but with B scoring higher than A as predicted.
2) Junction 4, shared primer 2, fosmid B2: the sequence does not align well to any of the four draft fosmids at this junction. The sequence is of low quality. This is true of both replicates.

9 List of supplemental data files

Filename	**Description**
pcr_primers_predictions.xls	Spreadsheet of primers and clones, and predictions
pcr_observed_lengths.xls	Observations of PCR products, lengths, and sequence alignment (observations were blind and then copied over to a spreadsheet with the predictions).