Analysis of the ARS-seq data

Processing the reads

The ARS-seq experiment yielded 5,201,753 read pairs referred to as S1 and S2 below, each 76 bases long. 4 bases were trimmed at the 5' end of every S1 read to remove adjacent vector sequence. BT (Bowtie version 0.12.7) was applied to the processed read pairs with the additional parameters "-S -5 3 -3 0 --best -y --solexa1.3-quals -m 1 -X 3000" aligning 3,607,211 of the read pairs to the *S. cerevisiae* genome. All references to the *S. cerevisiae* genome are to the October 2003 version of the S288C strain genome (Saccer1).

The read pairs that were not initially aligned include :

- 231,198 read pairs that aligned to the 2-micron plasmid (these were excluded from further analysis).
- 411,461 read pairs (or 8% of all pairs) mapped to two different parts of the genome and hence most likely came from plasmids with chimeric inserts. These alignments were also excluded from downstream analysis.
- Alignments where the insert is shorter than the read length, in which case some vector sequence is added to the 3' end of the read. A related problem is that BT does not report alignments where one mate's alignment is entirely contained within the other's. To recover such lost alignments we took all the unaligned read pairs after the initial BT run and trimmed a further 24 bases from the 3' end of the S1 reads and trimmed 28 bases from the 3' end of the S2 reads. This procedure generated an additional 159,128 alignments which were added to the 3,607,211 initial alignments for a total of 3,766,339 alignments.

The 3,766,339 aligned read pairs were mapped to 6,054 unique contiguous genomic fragments. The highest read pair count per fragment was 227,992 and the lowest was 1 (1,496 read pairs were uniquely mapped to a genomic fragment yielding a read count of 1).

Some of the genomic fragments did not start and end as expected with sites of one of the four 4-cutter restriction enzymes that were used to digest the insert DNA (likely due to DNA end degradation prior to ligation): "GATC", "AGCT", "GGCC", "GTAC". Therefore, we used a python

script to find the minimal number of indels required so that the inferred genomic fragment would start and end with sites of one of our four 4-cutters.

- Limited to 7 deleted bases and to 2 inserted bases at each end as well as to a total of no more than 8 indels at both ends, the script was able to uniquely assign 3,766,339 alignments (over 99.5%) to one of our 4-cutters: AGCT 27.4%, GTAC 30.36%, GGCC 17.4%, GTAC 24.4%.
- Looking at how the indels are distributed across the 4-cutters we note that less than 1% of the fragments attributed to the sticky ends 4-cutter GATC had any such indels at their ends whereas over 54% of the fragments attributed to the blunt ends 4-cutter GTAC were subjected to an indel event at their ends (29.5% and 45% for the other two blunt 4-cutters). We conclude that the indels were most likely generated during the plasmid construction stage rather than during sequencing. Moreover this finding is consistent with our observation that GATC libraries gave much lower cloning efficiencies than the rates observed with the 3 blunt cutters.

Defining the ARS-seq segments

We next grouped together all alignments that are mapped to the same genomic fragment flanked by the sites of one of our 4-cutters. This grouping left us with 926 genomic fragments that start and end with a unique 4-cutter site. For each such genomic fragment we keep track of the number of read pairs that mapped to it as well as the number of orientations, or whether the insert was found in a single or both orientations relative to the vector (which is also indicative of the number of different plasmids that contained this genomic fragment).

- 206 of those 926 fragments had a single read pair mapped to them and were removed from further analysis.
- All the remaining 720 fragments passed a quality filtering step demanding that the average read quality score at both ends of the insert (across all positions of all reads that mapped to that fragment) is greater than 10.
- Most of these 720 fragments overlap another fragment as they correspond to different 4-cutter digestion patterns of the same ARS. We wrote a Python script to assemble these genomic fragments into 366 disjoint components or contigs of overlapping fragments keeping track of the total number of fragments that define each contig as well as the accumulated read depth and insert orientation counts. The table below gives the distribution of the number of fragments per ARS-seq contig:

# of fragments per contig	# of ARS-seq contigs
1	139
2	121
3	88
4	15
5	3

Defining the Inferred Functional Core

The contigs represent the union of the genomic fragments at each locus; however, we are also interested in the minimal intersection region of the fragments at each locus which should yield better resolution around the minimal functional element. In practice, due to false positives we devised a more sophisticated method to define what we refer to as the inferred functional core.

- We wrote a Python script that implemented a dynamic programming approach to find the minimal number of fragments that needs to be excluded from each contig in order for the intersection of the remaining fragments to be at least 50 bases long. When excluding a fragment, its weight is set to its orientation count, so a fragment with an orientation count of 2 counts twice as much as one with an orientation count of 1.
- In 364 of the 366 contigs the intersection spanned at least 50 bases hence the script defined these contigs' inferred cores as their intersections (no fragments were excluded). The inferred cores of the remaining 2 contigs were defined while excluding a total of 3 fragments: the ARS-seq contig spanning Chr 11, 151996 - 154496 and the contig spanning Chr 13, 633628 - 634930 are both made of 5 fragments whose corresponding intersections are empty.

The ARS-seq cores we defined are significantly shorter than the contigs from which they were inferred: the median contig length is 1002bp whereas the median length of the inferred core is 387 bp. For comparison, the median length of the 720 genomic fragments is 702 bp. To assess whether this reduction in length is associated with some loss of function we checked how it affects our retrieval of OriDB ARSs. Using the entire contig we find that 312 of the 366 ARS-seq contigs contain the top scoring match of a non-dubious OriDB ARS (ndoARS) to either the core 17 bp ACS

or the extended 33 bp ACS[1]. Reassuringly, the same 312 top ACS matches are also contained in the inferred cores of these contigs. Therefore in the rest of the analysis we restrict our attention to the inferred cores of the ARS-seq contigs.

- Note: generally, each ARS-seq contig overlaps at most one confirmed OriDB ARS, however 3 contigs overlap more that one:
  - Ch 3, 13802-15620 includes the top ACS matches of *ARS302*, *ARS303* & *ARS320*
  - Ch 10, 374144-376006 overlaps *ARS1012* (includes ACS) & *ARS1013* (misses ACS), with the inferred core overlapping only *ARS1012*
  - Ch 16, 562996-565162 overlaps *ARS1622* (includes ACS) & *ARS1622.5* (includes ACS) with the inferred core overlapping only *ARS1622*

Precision analysis

To systematically assess the functionality of each one of the 366 ARS-seq contigs we relied on 3 sources of information: OriDB, the score of the best ACS match and manual biological verification. We already mentioned the 312 ARS-seq inferred cores which share a top ACS match with ndoARSs. As 50 of those top matches were in "likely" rather than "confirmed" ARS categories we selected 10 of the 50 inferred cores for verification (a subsequence of the ARS-seq inferred core was used for verification in 3 of those 10 cases, the rationale for this is explained below). All 10 were found to be functional albeit one was designated weakly functional based on visibly slower yeast colony growth on selective media. The remaining 262 are confirmed OriDB ARSs (coARSs) of which 22 were selected for control verification and all were found to be functional albeit one weakly and one very weakly functional (in 9 of the 22 verification we again used a subsequence of the inferred core).

Another two ARS-seq inferred cores contain ACS matches but not top ones in one likely (Ch. 7, 15327-19624) and one confirmed (*ARS229*, Ch. 2, 801930-802617) OriDB ARSs. In the latter case the corresponding ARS-seq core was verified to be positive and in the likely ARS case a fragment which heavily overlaps the inferred core and contains the same ACS match was also verified to be functional (see Supplementary Table 3 for details). We therefore presumed that all

---

[1] Both the 17bp and the 33bp PWM were learned by GIMSAN applied to the non-redundant (non-repetitive) set of 334 confirmed OriDB ARSs as of April 2011. We then used SADMAMA to scan the set of all OriDB ARSs for matches to those two PWMs.

314 of the 366 ARS-seq inferred cores that contain an ACS match (core 17bp or extended 33bp) that is fully contained in a ndoARS are functional.

The 52 of the 366 ARS-seq contigs that do not include a ndoARS ACS match can be further divided into two sets: those with a good ACS match (score >= 8, Supplementary Table 4) and those without a good ACS match. The first set contains 5 ARS-seq inferred cores all of which were verified as functional (in two of the cases a smaller fragment was verified positive). The second set contains one ARS-seq inferred core with a score of 6.86 which we could not clone and therefore left its presumed ARS function as unknown. Of the remaining 46 ARS-seq contigs we verified that 36 are non-functional (highest ACS score is 6.1, most below 5). The remaining 10 were therefore presumed non-functional as well (highest ACS score 5.98, most below 5).

To summarize, we posit that 319 of the 366 ARS-seq inferred cores are indeed functional (39 of these were verified as such) and 46-47 or about 12.5% are non-functional (of which 36 were verified as such).

Improving the precision

We explored two approaches for reducing the false-positive (FP) rate among the ARS-seq contigs. The first involves removing contigs that do not have ample support. For example, if we arbitrarily remove all ARS-seq contigs with an accumulated orientation count of 1 and read count less than 100 (that is, contigs defined by a single orientation insert which is supported by no more than 100 read pairs) as well as all contigs with read count less than or equal to 2 then we remove 28 presumed non-functional ARS-seq contigs as well as 30 which are presumed functional so the new FP rate is 18-19 out of 308 or 6%.

The second approach involves enlisting the ACS to tease out the FP contigs. We did not assume prior knowledge of the ACS but instead ran the motif finder GIMSAN on the set of 366 ARS-seq inferred cores in ZOOPS (zero or one occurrence per sequence) mode. Visually inspecting the results of a few selected widths we find that width 33 is a good predictor and discard all the 71 ARS-seq inferred cores that were determined by GIMSAN to have 0 occurrences of the ACS. These include *all* 46 inferred cores that are presumed to be non-functional as well as 25 that are presumed functional. Therefore this approach leaves us with 0-1 presumed non-functional cores out of 295 ARS-seq inferred cores.

Coverage analysis

We turned to OriDB to estimate how exhaustive is the coverage of our ARS-seq segments. Specifically, we looked at the 351 confirmed ARSs in OriDB (coARSs) as of January 2012. We found 82 coARSs that do not have an overlap with any ARS-seq fragment (including fragments with a read count of 1) representing a substantial initial false-negative (FN) rate of 23.4%. However, when we individually verified 67 of these 82 missed coARSs we found that 26 of them were non-functional in our tests. In addition, of the 41 that we did verify as functional, 20 were only very weakly so and another 12 were weakly functional. These results should be compared with the results mentioned above of our verification of coARSs with an ACS that is included in an ARS-seq inferred core: all 23 were found to be functional with only one very weakly and one weakly so (Fisher Exact Test p-value of 0.0002 if we keep the barely functional separate and of 5.8e-8 if we lump the barely together with the weakly functional).

Further evidence that the ARSs that are missed by the ARS-seq procedure are somewhat weaker is found by comparing the strength of the set of 262 ACS matches that lie in an ARS-seq core and that coincide with a top ACS score of coARSs with the set of 86 top coARSs ACS that are not covered by an inferred core of an ARS-seq (note that one ARS-seq core covers 3 coARSs and another coARS has no ACS at all). A Mann-Whitney test comparing these two sets of ACS scores rejects the null with a 2-sided p-value of 8.6e-14. Thus, it seems that in general the ARSs that are missed by the ARS-seq procedure are weaker but in any case the revised FN rate is between 41-56 / 325 ~ 12.5%-17%

Double inserts

As noted above 411,461 read pairs (8%) were mapped to two different parts of the genome indicating at least that many cases of a double insert occurred. In and of themselves these are not alarming as they are filtered out by BT but they do show that this construct is present in a non-negligible quantity. A more troubling hypothetical construct is a non-contiguous double insert made of two copies of the plasmid vector interleaved with two inserts. Such a construct could generate “silent double inserts” or ones that we cannot detect. This situation can be particularly problematic if only one of the inserts is actually functional as in this case the other would be a FP. There are two

verified FP ARS-seq contigs that we suspect could have arisen this way. Both of these are very short contigs and stand out due to their large read count:

- Ch. 11, 456966-457051 has 10,086 read pairs and a single orientation alignment.
- Ch. 14, 686525-686595 has 31,833 read pairs and a single orientation alignment.

Analysis of the miniARS-seq data

Processing the reads

Four separate miniARS-seq experiments were performed:

- miniARSseq1 was the initial experiment.
- miniARSseq2: plasmid pools resulting from miniARSseq1 were re-screened in yeast, isolated and sequenced.
- miniARSseq3 was a biological replicate of miniARSseq2.
- miniARSseq4 was a technical replicate of miniARSseq3 (same plasmid pools sequenced separately).
- miniARSseq1 differed from the 3 subsequent runs in both the read lengths and the sequencing primers. Similarly to the analogous primer for the ARS-seq primer, the S1 primer of miniARSseq1 started the sequencing reaction with GATC and therefore 4 bases were trimmed from the 5' end of all S1 reads of miniARSseq1.

Bowtie (BT) was used to align the read pairs of each of the 4 experiments using the parameters "-S -5 3 -3 0 --best -y -m 1 -X 500" ("-X 1000" was used for the three later sequencing runs) generating the number of alignments specified in the third row of the below table.

As the inserts here are generally shorter than in the ARS-seq experiment, the reads often read through the insert into the vector. Therefore, after the initial BT run of the unprocessed read pairs we applied two rounds of progressively trimming more bases at the 3' end of both reads for each read pair that was not previously aligned.

A feature unique to the miniARS-seq data is that some of the fragments contain a piece of the ARS-seq vector which results from the ARS amplification step prior to shearing and ligating the ARS sub-fragments. If the fragment is oriented such that the vector piece is at the 5' end of the read then none of the previous three alignment steps would be able to align such a fragment. We therefore used a Python script that scanned the hitherto unaligned reads for the longest prefix matching the suffix of one of the three ARS-seq vectors, more precisely, one of the three 5' flanks or one of the three 3' flanks that are adjacent to the vector-insert boundary. We then removed any

such common pieces that were longer than 5bp provided the remaining fragment was at least 25 bases long. We then applied BT to these 5' ARS-seq-vector trimmed read pairs obtaining over 5.5% of new alignments in the original run and over 10% in the three subsequent sequencing runs.

Finally, we took the 5' ARS-seq-vector trimmed reads that failed to align and trimmed their 3' ends as well as those of their mate reads so as to obtain a mate pair of 45 bp long reads. We then applied BT to the resulting set of read pairs (last row of table below).

	miniARSseq1	miniARSseq2	miniARSseq3	miniARSseq4
S1/S2 read length	76	101/89	101/89	101/89
# of read pairs	24,815,006	8,994,784	10,203,755	9,521,202
# of aligned raw reads	8,825,920	3,629,471	3,759,133	3,695,659
initial 3' ends trim: S1/S2	145,787 22/26	206,101 36/24	304,946 36/24	191,736 36/24
deeper 3' ends trim: S1/S2	N/A	79,247 56/44	100,658 56/44	57,155 56/44
5' ARS-seq vector trim	504,809	397,186	389,545	433,279
5' vector trim + 3' end trim	4,568	2,634	2,307	2,147
total # of aligned reads	9,481,084	4,314,639	4,556,589	4,379,976

Defining the miniARS segments

The aggregated outcome of our aforementioned iterative alignment procedure is the alignment of 22,732,288 of the 53,534,747 read pairs into 21,592 distinct, though often overlapping, genomic fragments. For each such distinct miniARS genomic fragment we keep a record of the number of read pairs mapped to it as well as the total insert count: the sum, over the 4 sequencing runs, of the number of distinct inserts that were made of this fragment (1 or 2

orientations and possibly another distinct insert coming from the addition of a piece of the ARS-seq vector).

The maximal read count per genomic fragment was 971,476, while 7,214, or about a third of the fragments, had a read count of 1. The latter fragments whose total read count was less than 0.05% of the aligned read pairs were removed from further analysis. The remaining 14,378 fragments were further subjected to a quality control test, requiring that the maximum quality score of the S1 reads as well as of the S2 reads is greater than 15. This filtering left us with 13,991 genomic fragments with a median read count of 17 (mean 1,624) and a median length of 148 bps.

Removing chimeric inserts

Inspecting the data, we found a fairly high number of suspected FPs: 661 of the 13,991 mini genomic fragments have no ACS match at all. Moreover, 131 of these 661 fragments have a read count of over 1000 and they show up in multiple sequencing runs. On closer inspection of some of these suspected FPs we found that many share an end with the parent ARS-seq insert and moreover they have the same orientation as that of the ARS-seq insert. This structure suggested that these FPs arise from silent double insert constructs where ARS-seq vector sequences adjacent to the insert boundary are carried over into the miniARS plasmid. Since the vector sequences adjacent to the insert serve as templates for Illumina sequencing primers, during the miniARS sequencing reaction this fragment of the ARS-seq vector masks any additional insert that might be located just 5' to it. If this additional insert is in fact the functional one whereas the piece that we see is not functional we will have a FP due to this silent double insert. There are several pieces of evidence supporting this conjecture:

- In 2 out of the 4 sequencing runs that we checked for those (miniARSseq2 and miniARSseq3) we found 22-24% of all read pairs are made of double inserts. This estimate was established by running BT against the entire DNA content of *S. cerevisiae* (including the mitochondria and the 2-micron plasmid) and looking for alignments where the 2 reads were inconsistently oriented or the genomic distance between the aligned reads was larger than 1000 bps. This estimate is roughly half the rate of properly aligned read pairs (see table above) suggesting the double insert phenomenon is prevalent in the miniARS-seq experiments (compared with 8% of observed double inserts in the ARS-seq experiment). We suspect that the reason behind these high rates of double inserts is the increased efficiency at which ligase joins very short DNA fragments and incomplete de-phosphorylation of insert fragments.

- The observed double insert rate among those read pairs that were subjected to the 5' ARS-seq-vector trim was extremely high for some of the vectors: 71-73% for GATC, 47-48.5% for TCG, and 23.6-28% for TGG (miniARSseq1 and miniARSseq2). We suspect that these even higher rates of double inserts are observed since conditioning on the mini fragment aligning with an end of the arsseq insert it is less likely to be functional on its own and requires another, functional, insert to be present.
- We examined the cases where any of the 13,991 miniARS genomic fragments start or end at the same position as an ARS-seq insert with a single orientation. Specifically we looked at the ratio of the number of miniARS inserts that conserve the unique ARS-seq orientation to the number that reverse the orientation. At 465 to 146 (3.2 to 1) this ratio is consistent with the conjecture that many of those miniARS fragments are in fact part of a silent double insert as they inherit the ARS-seq insert orientation together with the appropriate inherited piece of the ARS-seq vector. As we increase the distance d between the start/end points of the ARS-seq and the mini inserts the ratio drops rapidly: for d=1 it is 1.5, d=2: 1.3, d=3: 1.1.

Taking into account the ratios above, we decided to filter out 1,653 mini genomic fragments (11.8%) that fit the description of a suspected silent double insert: shares the orientation of an ARS-seq insert with which it also shares an end or a start up to a slack of 1bp (corresponding to d <= 1 above). As further indication that many of the fragments removed by this criterion are indeed FPs we note that:

- 281 of the 1,653 (17%) removed fragments do not have an ACS. A fraction that is significantly higher than the 661 of the previous list of 13,991 mini fragments that do not have an ACS (4.7%).
- 85 of the 131 (65%) ACS-less genomic inserts that have a read count of over 1,000 are removed as part of this group of suspected double inserts.

The miniARS-seq contigs and their inferred cores

After the previous filtering step we were left with 12,338 miniARS genomic fragments were assembled into 181 unique contigs. We first defined the inferred core of each miniARS-seq contig in the same way we defined the inferred core of the ARS-seq contigs: looking for the minimal (weighted) number of genomic fragments that need to be thrown out so that the intersection of the remaining fragments will contain at least 50 bp. Here the multiplicity, or weight, of each genomic fragment was taken as the total insert count defined above.

Focusing on the ACS we noticed that a few inferred cores miss the ACS match by up to 3 bp and in another case the inferred core completely missed the obvious ACS match. With an average of less than 2 ARS-seq fragments per contig it seems that the combinatorial approach to deduce the inferred core is well suited for the less complex ARS-seq data. However, the miniARS-seq data presents us with an average of over 68 fragments per contig so we implemented an alternative, statistical approach for deriving its cores as explained below.

The left end (start) of the inferred core is set to the 0.05 quantile of all left ends of the genomic fragments that constitute the considered contig where each fragment multiplicity is taken according to its total insert count. Similarly, the right end (end) of the inferred core is set to the 0.95 quantile of all right ends of the genomic fragments that constitute the contig (again, each counted according to its total insert count). If the right end point is at least 50 bp to the right of the left end point the core boundaries are set here. Otherwise, we again use a dynamic programming approach to find the minimum total count of *additional* genomic fragments that need to be thrown out so that the inferred core would be at least 50 bp long.

Judging by the location of the ACS relative to the inferred core we concluded that the latter statistical approach performed better than the combinatorial approach in determining the inferred core of the mini contigs.

Once again we find a substantial difference between the median lengths of the contigs and their inferred cores: 230 bp vs. 92 bp. We reuse the OriDB argument to show that the reduction of length is apparently not detrimental to ARS function: using the entire miniARS contig we find that 169 of the 181 mini contigs contain the top scoring match of a ndoARS (to either the core 17 bp ACS or the extended 33 bp ACS) and, importantly, the same holds if the contigs are replaced with their inferred cores.

Finally, we note that during the process of defining the ARS-seq cores we discarded genomic fragments in only 2 of the 366 contigs; in defining the miniARS cores we ended up discarding fragments in 93 of the 181 contigs. This last number does not include all the fragments that were thrown out as suspected double inserts and indicates the higher level of noise in the miniARS-seq data compared to the ARS-seq data. No doubt part of this higher level of noise can be attributed to the much larger number of fragments generated by the miniARS-seq experiments.

Precision of the miniARS-seq contigs

To determine the functionality status of each of the inferred miniARS cores we again integrate OriDB data with ACS scores and manual biological verification. We have already mentioned the 169 inferred cores that contain a top ACS match of a ndoARS (18 likely, and 151 confirmed OriDB ARSs). We verified that 12 of these 169 cores (overlapping 3 likely and 9 confirmed OriDB ARSs) are functional albeit we allowed 5 of those 12 cores to extend 3 extra bps on either side of the core. In 2 more cases (overlapping coARS) we verified as functional a fragment that overlaps the critical ACS but in one case was larger than the core and in another overlapped the core. Thus, to various degree of accuracy we verified that 14 of these 169 cores are functional and we therefore presumed that all 169 are.

An additional 2 miniARS cores contain a non-top ACS match of a ndoARS (1 likely and 1 confirmed). Both were verified to be functional albeit in one case we used the mini contig rather than its core. Another 2 miniARS cores with no overlap to a ndoARS were verified as functional new telomeric ARSs. This bring the total number of presumed functional mini cores to 173 of the 181 cores. The remaining 8 miniARS cores are presumed non-functional. Five were verified to be non-functional: in 3 of those we verified the core itself, and in two of those the entire contig was verified as non-functional. Each of the additional 3 mini cores that are presumed non-functional have a read count of 2, two do not have an ACS and the third has a very low score match. Therefore the FP rate of our mini segments is estimated at 8 of 181 or 3.9%. Of note is that one of the verified non-functional miniARS cores is inherited from its ARS-seq ancestor which is non-functional as well.

We can significantly increase the precision of our screen if we remove all miniARS contigs with a read count of less than 10: this will remove 5 of the 8 presumed non-functional mini segments (and no functional one) yielding a FP rate of 3 / 176 or 1.7%. In principle we can also follow the idea of de novo ACS discovery to remove suspected FPs as we did in our analysis of the ARS-seq data. A couple of alternative approaches for reducing the FP rate are suggested below.

ARS-seq segments and their inferred miniARSs

Correlations between ARS-seq and miniARS-seq experiments

All but 3 of the 181 miniARS contigs are fully contained (up to a 3bp slack) by an ARS-seq contig. Each of the remaining 3 miniARS contigs heavily overlaps, but also extends significantly beyond, a corresponding ARS-seq contig. None of these extensions seem to be functionally important (the ACS is already included in the corresponding ARS-seq contig).

There are 5 ARS-seq contigs, all of which are presumed functional, that generate 2 mini contigs each. In all 5 cases one of the mini cores is presumed functional (3 were verified as such) while the other mini core is presumed non-functional (3 were verified as such). In 4 of these 5 pairs the presumed functional mini core has much larger read count than the presumed non-functional core but in the fifth case the non-functional core has the larger read count (and both cores are verified). Hence if we adopt the strategy of keeping only the popular core from each pair we would be rid of 4 of the 8 non-functional miniARS cores at the cost of erroneously discarding one functional miniARS core.

Consistently, there are 176 ARS-seq contigs that overlap a miniARS contig and all but one of these ARS-seq contigs are presumed functional, suggesting that if we restrict attention to the ARS-seq contigs that yield miniARSs the *ARS-seq* FP rate will be as low as 0.6% (1/176). However, that would be at the substantial cost of throwing out the other 144 of the 319 (54.9%) presumed functional ARS-seq contigs that are not covered by any miniARS fragment.

There are 12 miniARS cores that are not fully contained in the corresponding ARS-seq cores. Of these, 6 are among the 8 miniARSs which are presumed non-functional. This indicates that if we throw out all miniARS segments whose inferred core is inconsistent with that of the corresponding ARS-seq we would reduce the miniARS FP rate to 2 / 169 or 1.2% at the cost of removing 6 presumed functional miniARS cores.

As expected, ARS-seq contigs with a higher read count are more likely to generate a miniARS fragment: the median read count per ARS-seq contig with a mini overlap is 8,513 whereas the median read count for an ARS-seq contig without a mini overlap is only 214 (Supplementary Figure 2). Moreover, when an overlap exists, there is a substantial positive correlation between (log of) the read counts in the two related contigs: 0.44.

A less expected observation is that ARS-seq contigs that overlap confirmed OriDB ARSs are more likely to generate miniARS fragments than ARS-seq contigs that overlap likely OriDB ARSs. Recall that there are 262 ARS-seq contigs that essentially include a top ACS match of a *confirmed* OriDB ARS. Of these, 151 or 57.6% generate a mini contig with the same property. However, only 18 of the 55 (32.7%) ARS-seq contigs whose core essentially includes a top ACS match of a *likely* OriDB ARS generate a miniARS contig with the same property. The difference in these ratios is statistically significant (Fisher Exact Test p-value 0.001). A possible explanation is that currently designated likely ARSs are generally weaker than currently designated confirmed ARSs. Consistent with this explanation a Mann-Whitney 2-sided test comparing the read count of the 262 coARS-overlapping ARS-seq contigs with that of the 55 loARS-overlapping contigs gives a p-value of 0.01 (with the confirmed-overlapping ARS-seq contigs having the higher count).

Related observations are that there is a small but non-negligible positive correlation between the top ACS score in an ARS-seq contig and its read count (0.19) and an even higher correlation between the top ACS score in a miniARS contig and its read count (0.30).

Finally, there is a statistically significant difference between the top ACS scores of ARS-seq contigs that are covered by miniARS contigs and those that are not (2-sided Mann-Whitney p-value of 1.289e-05).

All these observations can be roughly summarized by: stronger ACS is positively correlated with higher ARS-seq read count and the latter is positively correlated with higher miniARS read count. In addition (current) likely OriDB ARSs are generally weaker than confirmed OriDB ARSs.

The miniARS-seq segments are refining the ARS-seq segments

With a median core length of 92 bp the miniARS screen generally offers a significant refinement of the ARS-seq core whose median length is 387bp. Indeed, 169 of the 181 mini cores are completely contained in the corresponding ARS-seq core (and, as mentioned above, 6 of the remaining 12 mini cores are presumed non-functional). While the full details are available in Supplementary Table 4, of particular interest are short miniARS cores and especially ones that start or end very close to the end of the contained ACS.

Specifically, we looked at miniARS cores that contain, up to 2 bp slack, the top ACS match (17 bp core or 33 bp extended) of a likely or confirmed OriDB ARS and start or end no more than 5bp away from one end of the ACS. Interestingly, when the top ACS match is on the forward strand (T-rich) there are 35 miniARS cores that start at most 5 bps to the left of the ACS but there are only 2 miniARS cores that end at most 5 bps to the right of the ACS (2-sided binomial test p-value is 1.0e-08). Consistently, when the ACS match is on the reverse strand (A-rich) there are 36 miniARS core that end at most 5 bp away from the right of the ACS but only 6 mini cores that start at most 5 bp away from the left of the ACS (2-sided binomial test p-value is 2.8e-06). Taken together there is strong evidence that an ARS is less likely to be functional if trimmed very close to 3' end of the T-rich ACS than if trimmed very close to the 5' end of the T-rich ACS.

It is however important to note that we tested five of the mini cores that are trimmed very close to the 3' end of the T-rich ACS and all five turned out functional albeit one weakly and one barely so. Similarly, one of the ARS-seq cores (Chr. 5, 145676 - 145839) starts only 6 bp 5' to the reverse complement (RC) end of the top (extended) ACS match but was verified to be functional, albeit barely.

The extended ACS can form a viable ARS on its own

The previous observations led us to examine whether the 33 bp extended ACS which spans the 17 bp core ACS and the B1 element is enough to initiate replication of our plasmid. Somewhat surprisingly the answer was positive in 3 of the 4 cases that we tested. Note that Marahrens and Stillman (Science, 1992) showed that in the case of *ARS1* (*ARS416*) the extended ACS does not initiate replication of the plasmid they used.

All four inserts that we tested were derived from the aforementioned set of miniARS cores that include a top ndoARS ACS match which starts or ends very close to an end of the core. Three of the cores were selected as they were found to be functional even though the core ends very close to the 3' end of the T-rich ACS and one of those three cores (Chr. 9 in the table below) even included only the core 17 bp ACS ending just after its end. The fourth selected core (Chr. 4 in the table below) was in the more popular set of cores with an end close to the 5' end of the T-rich ACS and has a high scoring ACS.

The three mini cores that included the extended ACS were further trimmed so that they include just the 33 bp extended ACS. The mini core that included the 17bp ACS and not the 33 bp was trimmed from the left to start with the ACS and extended to the right to include the 33bp extended ACS. Details are available in the table below.

Chr	ACS start-end	ACS orientation	ACS score	ACS function	mini core	miniARS function
4	1447471-1447503	f	13.2	weak	1447462-1447555	ND
6	68824-68856	f/r	10.9/ 10.6	weak	68759-68856*	strong
9	214732-214764	f	9.6	no	214623-214748*	barely
10	612741-612773	r	12.5	strong	612742-612793*	strong

* The verified mini cores included an additional 3 bp on each side: Chr. 6 68756-68859, Chr. 9 214620-214751, and Chr. 10, 612739-612796

Analysis of the mutARS-seq data

Processing the reads

Our reference sequence is the 100bp of ARS1, or ARS416 ch. 4, 462510-462609. We used a beta version of Bowtie2 (BT2, 2.0.0-beta5) which, unlike the original Bowtie, can handle alignments with indels. We used the global alignment mode (--end-to-end) of BT2 in combination with extending the 100bp reference sequence by adding about 35bp from its flanking vector on each side. More precisely, the BT2 reference database was made of the insert plus flanking pieces of the vector in both possible orientations. The addition of the vector fragments was done to improve alignment of reads that include deletions (recall that the read length is 101bp and the WT insert is 100bp long). We also adjusted BT2's scoring scheme, in particular, significantly lowering the relative cost of extending a gap in the reads (but not in the reference). The full set of parameters we used is: " --no-discordant --no-mixed --dovetail --dpad 50 --gbar 1 --rdg 44,1 --ma 16 --mp 40,16 --np 8 --rfg 60,24 --score-min L,0,-4.8 --end-to-end -X 1000". We wrote a Python script to process the BT2 SAM output files, concentrating on the information in the CIGAR string, so as to find the copy number of, or the number of read pairs aligned to, each of the following variants:

- The WT sequence
- All sequences with a single substitution (and no indels on either one of the paired reads)
- All sequences with *exactly* 2 substitutions (and no indels on either of the paired reads)
- All sequences with a single deletion of 1bp (and no consistent substitutions)
- All sequences with a single deletion of 2bp (and no consistent substitutions)

Both reads in the pair have to agree on the position of the substitution and they should also agree on the substituted base or, if they do not agree, then the accepted substitution should have a quality score >= 30 while the rejected base should have a quality score < 15. If only one of the reads has a substitution at a position then that substitution is ignored whereas if both reads have a substitution at the position but none of the conditions above is met then the read pair is rejected as having an inconsistent substitution.

Similarly, a reference base needs to be deleted in both reads in order to be counted as a deletion whereas if it is deleted in only one read of a pair then the pair is thrown out as having

inconsistent deletions. Note that identifying consistent deletions between the mate pairs demands some care. The problem is best understood by an example: consider the 2 possible deletions of length 3 in the reference sequence TXYT. Either one of these deletions will result in a single T so we cannot tell whether the deletion started with TXY or with XYT. The version of BT2 we used often determines the location of the gap inconsistently between the pair of reads so by simply parsing the CIGAR string it looks as if the deletion is inconsistent while in fact it most likely is not. Our script specifically looks for these potential inconsistencies and resolves them on the side of agreement.

Note that we did not analyze variants that include insertions as the number of reads mapped to those variants was deemed too small.

	plasmid libraries	sample 1 t=0	sample 2 t=20h	resample 1 t=0	resample 2 t=12h
# of mate pairs	2,844,996	264,859	1,083,027	6,930,570	3,706,055
# of aligned mate pairs	2,676,160	237,379	1,040,502	6,745,420	3,612,710
% of aligned pairs	94.07%	89.62%	96.07%	97.33%	97.48%
WT count	444,631	44,624	216,768	1,207,507	685,589
single sub count (0 indels)	763,329	74,475	343,331	1,972,634	1,078,934
2 subs count (0 indels)	656,190	61,670	268,014	1,633,918	861,536
1bp del count (0 sub)	23,345	2,155	11,040	61,303	36,905
single 2bp del count (0 sub)	3,139	269	1,128	8,039	4,099

Analysis of single substitution mutant variants

Estimating the relative growth rate of *any* variant

We make the simplifying assumption of ignoring the distinction between the plasmid population and the yeast population. We therefore assume there is a growth rate associated with each plasmid that is conferred by the exact composition of its insert as well as by the environmental conditions. To estimate the growth rate of a specific variant between two measurement points we look at the ratio of the variant's read counts at those two points. To normalize this ratio we divide it by the ratio of the analogous counts for the WT insert and finally we take its (base 2) logarithm. Thus, a variant that grows faster than the WT has a positive relative growth whereas one that grows slower has a negative relative growth.

Analysis of single substitution variants on plates

We began with analyzing the relative growth rates of the 300 (100 x 3) single substitution mutants between the first measurement taken as the plasmid libraries were scraped from the *E. coli* plates and purified in pools ("plasmid libraries" in table above) to the second measurement taken as the transformed yeast colonies were scraped off their plates ("sample 1" and "resample 1" in the table above). This captures the relative growth rate of each variant as it grows to a yeast colony on the plate.

As can be seen from the top chart in Supplementary Figure 3 there is not a lot that can be learned from these measurements. With very few exceptions all variants that deviate from the WT are penalized with a reduced growth rate. Even more troubling is the fact that this reduced growth rate is not generally more pronounced in variants that include mutations in the known functional elements such as the ACS. We suspected that the physical constraints of growth on a plate is constraining the yeast transformants' growth rate and therefore took the scraped yeast colonies and inoculated them for competitive growth in liquid cultures.

Analysis of single substitution variants in competitive liquid growth

We have two independent sets of measurements that allow us to estimate the relative growth rate in liquid. One between times t=0 and t=20h (sample 1 and sample 2 in table above) and another, using a biological replicate experiment between times t=0 and t=12h (resample 1 and resample 2 in table above). We find that these relative growth rates estimated from competition in liquid are much more informative than the ones estimated from the plates.

First, the two sets of measurements are generally in good agreement with one another. This can be seen in the apparent linear relation between the pairs of growth rate estimates for all single substitution mutants (Supplementary Figure 4, correlation coefficient is 0.92), as well as by visually comparing the 2 bottom positional growth rate graphs in Supplementary Figure 3. At the same time the correlation with the growth rates measured from the plates is a rather poor 0.07 (12h resample estimate compared with the plates estimate).

Second, the positional growth rate graphs are quite effective in delineating the functional elements within the ARS: note the negative growth rates associated with most of the mutations that coincide with the core 17bp ACS (positions 5-21), the B1 element (positions 35-37) and the B2 element (positions 67-77). Interestingly, the only substitutions in the core ACS that give any substantial positive relative growth rates are in positions 6 and 20 where the WT base does not coincide with the consensus base indicated by the PWM. Moreover, all the mutations in those positions that confer positive growth rates are indeed more favored by the ACS PWM than the corresponding WT bases. Similarly, the positive growth rates conferred by the mutations in positions 69 and 72 of the B2 element turn it into a string that perfectly matches the 11bp (RC) ACS consensus.

Third, we note that the linear regression of the growth rates estimated from the 20hr experiment on the growth rates estimated from the 12h experiment is given by: 1.76*x - 0.0012. Note that the constant term is almost vanishing as expected and moreover the linear coefficient of 1.76 is in good agreement with the temporal ratio of 18/10 = 1.8 which factors in the fact that it takes the yeast about 2 hours to transition into the exponential growth phase.

To combine the two sets of measurements into a single graph we computed yet a third ratio: the sum of the relative growth rates of the 20h experiment over the corresponding sum of the 12h experiment. The result of 1.78 equals to the linear coefficient of a regression where we constrain the constant term to be 0. We then multiplied the 12h experiment measurements by 1.78 and took a simple average of each single substitution variant's pairs of measurements (Figure 2B).

Analysis of variants with exactly 2 substitutions

Our biggest difficulty in studying the growth rates of the 2-substitution variants is the great increase in the number of possible variants. There are only 300 single substitution variants,

ensuring a deep average coverage of about 250 read pairs per variant even in the sample with the lowest read count (sample 1). However, there are 403,650 double substitution variants yielding an average coverage of about 0.15 read pairs per variant from that same sample. Therefore we decided here to use only the 12h experiment samples as they had considerably more reads.

We first examined whether growth rates are additive, that is, whether the relative growth rate of a 2-substitution variant is given by the sum of the relative growth rates of each of the two corresponding single substitution variants. Supplementary Figure 5 compares the measured relative growth rate with the additive model for each of 8,017 double-substitution variants for which the read count at t=0 was at least 20 and the read count at t=12h was at least 10 (in addition to a plasmid library read count of at least 5). The positive correlation is a non-negligible 0.57 yet it is clear that the additive model is not completely satisfactory. This is further evidenced when we consider the same correlation separately within the set of 1,026 double substitution variants whose both corresponding single substitution variants have positive relative growth rate and within the set of 3,239 double substitution variants whose single substitution variants both have negative growth rate. In both cases the correlation decreases significantly (to 0.39 and 0.31 respectively) and consistently with increased noise in corresponding scatter plots (Supplementary Figures 5B-C).

Regardless of whether or not the additive model offers a reasonable fit, we note that the linear regression of the measured double substitution growth rate on the 1,026 sums of pairs of positive single substitution growth rates indicate the former growth rate is typically sub-additive. Indeed, in 548 of these 1,026 cases the sum is larger than the double substitution growth rate (a modest yet significant 2-sided binomial test p-value of 0.03). This is much more pronounced in the 3,239 cases where both individual growth rates are negative: in 2,226 of those the absolute value of the sum is larger (p-value < 2.2e-16 which is the machine precision) indicating again a sub-additive effect.

Looking specifically at a few double-substitution variants we find that

- The variants combining 2 positive rate substitutions from the three positions 20 (core ACS), 69 and 72 (B2) yield an averaged super-additive growth: for each such pair of positions, the average measured 2-substitution rate is larger than the average sum of the individual rates.
- Unfortunately, the other positive substitution at the core ACS (position 6) yields no variants that pair with the above three positions.
- Position 67 in the B2 element has a very slight positive substitution which exhibits a sub-additive effect when paired with position 6 but super-additive when paired with positions 20 and 72 (no information regarding position 69).

- We looked at all variants with a positive relative growth rate and with both substitutions falling in the extended ACS (positions 5-37). Excluding the obvious pairs that include positions 6 or 20 the highest growth rate was found for the 'AA' substitution at positions 36 and 37 (the end of the B1 element). Interestingly, this is consistent with the fact that the motif finder GIMSAN applied to the set of 334 (repetitively filtered) confirmed OriDB ARSs finds these two positions dependent. Specifically, GIMSAN finds an over abundance of 'TT' (ARS1 WT) and 'AA' (the observed positive growth double substitution) but a lack of the mixed 'AT' and 'TA'.

Analysis of single base deletion variants

When compared with the single substitution variants there is significantly less data available to analyze the single base deletion variants. Still at an average of, depending on the sample, 22-369 read pairs per deleted position (see Table above) there is enough data to draw some meaningful conclusions. Note that because of the gap position identifiability problem referred to earlier, in practice there are even less than 100 single deletion variants. In essence, when a single base deletion occurs in a monomeric repetition we attribute that gap to the rightmost position of the repeat hence we do not have any single base deletion variants with a deletion inside a monomeric repetition.

We find that the growth rates learned from the same variants grown in liquid are in good agreement between the same 12h and 20h experiments and are informative (Supplementary Figure 6). Note that positions for which no bar is visible include, in addition to positions within a monomeric repeat, positions for which the read count was deemed too low (10 for plasmid libraries, 20 for resample 1, 10 for resample 2, 10 for sample 1 and 20 for sample 2). Concentrating on Supplementary Figure 6A for which more data is available we would like to point out:

- The positive relative growth rate associated with the deletion at position 6. This is not surprising given this deletion will slide the much more favorable 'A', currently at position 5, into position 6.
- The abundance of negative rate deletions and the lack of any positive rate deletion in the rest of the core ACS positions: 7-21.
- The same applies to positions 67-77 associated with the B2 element.
- The positive deletions observed between positions 23-35 seem troubling at first as they apparently change the rigid distance between the core ACS and the B1 element or worse, alter the B1 element (position 35 is the start of the B1). However, the "extra" T at position 38 means that the B1 element will continue to consist of a TTT in any one of those single base deletions hence these deletions are neutral as far as the B1 element is concerned.

- Note that there are more substantially positive deletions extending all the way to position 52 which is close to the start of the B2 element. Taken together these positive deletions strongly suggest that the WT distance between the extended ACS and the B2 element is not optimal and shortening it would generate a more efficient ARS.

We offer an alternative visual presentation of the deletion data in Supplementary Figure 6C. The (x, y)-entry in that heat map is the $\log_2$ of the WT normalized ratio between the read pair count in resample 2 and the respective count in resample 1 of all deletions that occur between positions x and y. Red indicates positive entries and green indicates negative ones. The core ACS and the B2 elements are visible quite clearly as are the aforementioned suggestive positive deletions between them.

Relative growth rates of single 2-base deletion variants (in liquid)

Supplementary Figure 7 gives the relative growth rate of each variant with a single 2-base deletion that starts at the specified position. Note that the graph is rather sparse, partly due to the gap location identifiability problem mentioned earlier but mostly due to the low number of single 2-base deletion reads available (again we impose the minimum read count of 10 for resample 2 and 20 for resample 1).

The sparseness of the graph makes it difficult to come up with any sweeping claims however we would like to point out:

- The core ACS again rejects deletions and, as expected, the growth rates of a 2-base deletion are generally more negative than the rates of 1-base deletions.
- The positive single base deletions between the core ACS and the B1 element have been replaced by negative 2-base deletions. This is consistent with the fact that unlike a single bp deletion a two bp deletion would alter the B1 element.
- There is a positive 2bp deletion starting at position 40 and an even more pronounced one starting at position 47 indicating that the WT distance between the extended ACS and the B2 element could be reduced by up to 2bp without any detrimental effects or indeed possibly with increased replication efficiency.
- The B2 element is again resistant to deletions.
- The highly negative 2bp deletion starting at position 50 is suggestive that the B2 element might extend to that region. Interestingly, if we think of the B2 as a reversed core ACS then positions 48-50 would correspond to its B1 element.