Computational identification of noncoding RNAs in *E. coli* by comparative genomics
Elena Rivas, Robert J. Klein, Thomas A. Jones and Sean R. Eddy

Some genes produce noncoding transcripts that function directly as structural, regulatory, or even catalytic RNAs [1, 2]. Unlike protein-coding genes, which can be detected as open reading frames with distinctive statistical biases, noncoding RNA (ncRNA) gene sequences have no obvious inherent statistical biases [3]. Thus, genome sequence analyses reveal novel protein-coding genes, but any novel ncRNA genes remain invisible. Here, we describe a computational comparative genomic screen for ncRNA genes. The key idea is to distinguish conserved RNA secondary structures from a background of other conserved sequences using probabilistic models of expected mutational patterns in pairwise sequence alignments. We report the first whole-genome screen for ncRNA genes done with this method, in which we applied it to the “intergenic” spacers of *Escherichia coli* using comparative sequence data from four related bacteria. Starting from >23,000 conserved interspecies pairwise alignments, the screen predicted 275 candidate structural RNA loci. A sample of 49 candidate loci was assayed experimentally. At least 11 loci expressed small, apparently noncoding RNA transcripts of unknown function. Our computational approach may be used to discover structural ncRNA genes in any genome for which appropriate comparative genome sequence data are available.

The key idea of our approach for distinguishing a small number of structural RNAs from a large background of other conserved sequences is to exploit a distinctive pattern of mutation, as opposed to simple conservation, as sketched in Figure 1. A conserved structural RNA tends to show a pattern of compensatory mutations consistent with some base-paired secondary structure. A conserved coding region tends to show a pattern of synonymous codon substitutions [9, 10]. Other types of conserved regions may be approximated by a “null hypothesis” that mutations occur position independently, with no special pattern. Thus, we should be able to use position-specific mutational models to separate conserved regions into three types: probable structural RNAs, probable coding regions, and probable “other” sequences.

Badger and Olsen [9] described a two-model protein gene-finding approach that distinguishes coding alignments from conserved noncoding alignments. We extended this idea by including a third model of RNA structure evolution, by allowing the input alignments to be gapped, and by using full Bayesian probabilistic models. Our three probabilistic models (RNA, COD, and IND) are stochastic “pair grammars”: a pair stochastic context-free grammar (SCFG) as the RNA model, and pair hidden Markov models for both COD and IND [11]. The secondary structure model in the pair-SCFG is a full probabilistic analog of the Zuker MFOLD algorithm [12], including terms for base stacking, loop lengths, etc. that have been trained on a database of tRNA and rRNA secondary structures [3]. The evolutionary distance component of all three models is derived directly (for COD) or indirectly (for RNA and IND) from a single choice of amino acid substitution matrix (BLOSUM62) [13, 14]. It is essential that all three models are at the same evolutionary distance. Otherwise, models might distinguish alignments solely based on their level of conservation rather than on the pattern of mutation.
We started with all annotated intergenic sequences with a length of ≥50 nt in E. coli according to the University of Wisconsin’s annotation of 115 ncRNA genes and 4290 coding ORFs [21]. This gave 2367 intergenic sequences, totaling about 500 kb. The average sequence length was 211 nt; the longest was 1729 nt. Four known ncRNA genes (csrB, oxyS, micF, and rprA) [2, 22] were unannotated and were left in the dataset as positive controls.

Each “intergenic” region was used as a query to search the four comparative genome databases using BLASTN [23, 24] (version 2.0MP-WashU/12 Feb 01, using default parameters and scoring matrix). All alignments with an E value of <0.01, a length of ≥50 nt, and an overall identity of ≥65% were collected, giving a database of 23,674 pairwise BLASTN alignments (12,037 from S. typhi, 5,239 or a structural RNA (bottom). Marks above each alignment indicate how it is scored to calculate each model’s likelihood: one position at a time for IND, one codon at a time for COD (integrated over all six possible frames), and as a combination of base-paired positions and single positions for RNA (integrated over all possible secondary structures).

Given an input pairwise sequence alignment (for instance, from a BLASTN comparison of two related genomes), we score the alignment with each of the three models. The scoring algorithms are all dynamic programming algorithms; the rate-limiting one is the O(L^3) memory pair SCFG alignment algorithm used to score the RNA model, which must be done as an SCFG Inside algorithm [11], summing over all possible RNA secondary structures. The scores are log likelihoods that are then used to calculate a final log odds score for the RNA model compared to the other two models. Non-RNA alignments get negative scores; increasingly positive scores indicate increasingly strong comparative evidence that the alignment contains a conserved structural RNA.

A satisfactory mathematical description of the approach is beyond the scope of this paper and will be published elsewhere (E.R. and S.R.E., submitted). We have implemented the complete approach in a program, QRNA [15]. The goal of this brief communication is to describe the first whole-genome screen we have done with QRNA.

We used QRNA to screen the complete genome of Escherichia coli K12 MG1655 (version M52) [16]. We chose E. coli because there is a complete genome sequence, it is a simple model genetic system in which we could readily test our predictions experimentally, and there is extensive comparative sequence coverage from at least 15 related species [17]. We chose four related enterobacterial genomes for comparison: Salmonella typhi [18], S. paratyphi A [19], S. enteriditis [20], and Klebsiella pneumoniae [19].

The complete screen took about 20 hr on a single SGI Origin200 R10K processor. A variety of tests of QRNA’s performance suggest that, under the conditions above, it has a sensitivity of about 80% on known structural RNAs (E.R. and S.R.E., submitted). As a test of specificity, we shuffled each of the 23,674 input alignments by aligned columns (preserving % identity, while scrambling the sequences and any position-specific mutational pattern in the alignment) and applied the same procedure, which produced 73 false-positive “RNA” loci with scores over 5 bits. Therefore, about 85% of our 556 candidate loci should be “true positives”, in the sense that they are not just the result of expected statistical noise.

Because QRNA screens for conserved RNA secondary structure, we expected it to detect various nongenic sequences with conserved RNA structure, including rho-independent terminators, rRNA spacers, transcriptional
Expression is suggestive but not entirely sufficient to define a candidate as a new ncRNA gene. For example, the cis-acting transcriptional attenuator of the his operon [31] is detected by QRNA, is flanked by a strong consens-promoter and an obvious rho-independent terminator, and has no significant coding potential (the hisL leader peptide is only 22 aa long); Northern analysis detects the his leader RNA as a distinct 170-nt transcript (data not shown). Candidates t44, tpk1, and tpk2 are directly upstream of coding genes in the same orientation and may be attenuators. We also cannot exclude the possibility that expressed candidates are protein-coding genes with small ORFs. For example, candidate k4 is classified as RNA in a *Klebsiella* alignment but is classified as coding Total *E. coli* RNA was isolated from log-phase cultures growing in rich (LB) medium at 37°C, run on a 6% polyacrylamide gel, electroblotted, and probed with a 5'-32P-labeled oligonucleotide specific to each strand of the predicted locus. (a,b) A typical positive result: tpe7 hybridizes to a 65-nt RNA transcript only with a 5'-strand probe (a), but not with a 3'-strand probe (b). (c,d,e) Confirmed transcripts for candidates tpk2, tke1, and tp2. (f) A typical result for a candidate (tpe1, downstream of *rpsA*) that most likely is part of an mRNA.

 attenuators in ribosomal protein and amino acid biosynthetic operons [25], other cis-regulatory RNA structures [26], and even certain repetitive elements [27, 28]. We removed 281 loci that plausibly fell into one of these nongenic classes, leaving a total of 275 candidate loci (see Table S1 in the Supplementary material available with this article online for a list). It must also be noted that not all ncRNA genes conserve an intramolecular secondary structure; for example, QRNA does not detect C/D box small nucleolar RNAs [29] in yeast or *Pyrococcus* alignments.

We expected these 275 loci to be a mix of ncRNA genes, cis-regulatory RNA structures, and false positives. A criterion that tends to distinguish an ncRNA gene from either cis-regulatory structures or false-positive signals is the expression of a distinct transcript independent of adjacent coding genes. We assayed 49 candidate loci for expression by Northern blotting (see Table S2 in the Supplementary material for a list). For 11 loci, we observed discrete small RNA transcripts <400 nt (Figure 2). Six others showed larger products that were interpreted as coding mRNA transcripts. Two showed multiple discrete bands that we could not interpret. The remaining 30 loci were not expressed under these growth conditions. We cannot interpret a negative result because several known ncRNAs are expressed only under specific conditions (for example, OxyS RNA is expressed under oxidative stress conditions but not in normal lab growth [30]). The 11 loci that express small RNAs are listed in Table 1.

While this paper was in preparation, two groups reported exciting results of different screens for small ncRNAs in the intergenic regions of the *E. coli* genome sequence [32, 33]. Wassarman et al. used sequence conservation coupled with microarray expression analysis and found 17 new ncRNAs [32]. Argaman et al. used sequence conservation coupled with promoter and rho-independent terminator prediction and found 14 new ncRNAs [33]. Both groups report extensive experimental characterization of the new loci. The overlap of these experimentally confirmed ncRNA genes with our results gives us additional confidence in QRNA’s sensitivity. Of the 14 RNAs reported by Argaman et al., 10 are in our list of 275 candidate loci; of the 4 that we do not detect (sraD, sraH, sral, sral), 3 have scores only slightly below our 5-bit cutoff, and only sral was completely missed. Of the 17 RNAs reported by Wassarman et al., 14 were detected by QRNA; of the 3 that we missed (ryeA, ryhA, and ryjA), 2 were just below our cutoff, and 1 (ryeA) was detected in the initial list of 556 QRNA candidates, but we mistakenly discarded it, thinking it was just a rho-independent terminator. On the other hand, only 4 out of 11 of our confirmed candidates were detected and confirmed by one of the other screens (Table 1), which suggests that QRNA’s sensitivity is higher than either the Argaman et al. or the Wassarman
et al. screens and that neither of these screens saturated the *E. coli* genome for novel ncRNAs.

These data, though experimentally preliminary, nonetheless validate QRNA as a powerful and general means for identifying candidate structural ncRNA loci. Because we use no organism-specific information (such as promoter or terminator consensus sequences), QRNA will be applicable in any organism for which appropriate comparative genomic data are available. We have already anecdotaly observed that signal/noise is sufficient to screen the human genome using low-pass mouse shotgun sequence coverage; for example, a QRNA screen of the 196-kb draft sequence of a human BAC (GenBank accession AL357874) spanning the cartilage hair hypoplasia (CHH) locus [34], using unassembled 1.7X mouse shotgun coverage (Mouse Sequencing Consortium, unpublished), predicts two ncRNA loci (data not shown), one of which corresponds to the 265-nt RNase MRP ncRNA locus that has recently been implicated as the gene responsible for CHH [34].

Table 1

Candidate loci expressing small RNAs.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Strand</th>
<th>Transcript size, nt</th>
<th>QRNA score, bits</th>
<th>Candidate start size, nt</th>
<th>Candidate Adjacent ORFs</th>
<th>Note</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>tpke11</td>
<td>+</td>
<td>370</td>
<td>22.3</td>
<td>14,080</td>
<td>88</td>
<td>dnak-&gt; dnakJ-&gt;</td>
<td>Overlaps ORF(s)</td>
</tr>
<tr>
<td>tp2</td>
<td>-</td>
<td>60; 120</td>
<td>25.0</td>
<td>122,857</td>
<td>160</td>
<td>pdhr-&gt; aceE-&gt;</td>
<td></td>
</tr>
<tr>
<td>t44</td>
<td>+</td>
<td>135</td>
<td>15.3</td>
<td>189,628</td>
<td>251</td>
<td>map-&gt; rpsB-&gt;</td>
<td>Attenuator?</td>
</tr>
<tr>
<td>tpe7</td>
<td>-</td>
<td>65</td>
<td>21.3</td>
<td>1,762,685</td>
<td>190</td>
<td>suLA-&gt; ydhL-&gt;</td>
<td>rydB</td>
</tr>
<tr>
<td>tpk70</td>
<td>-</td>
<td>40</td>
<td>12.3</td>
<td>2,494,649</td>
<td>435</td>
<td>lpxP-&gt; ydpZ-&gt;</td>
<td></td>
</tr>
<tr>
<td>tp1</td>
<td>+</td>
<td>300</td>
<td>29.1</td>
<td>2,651,802</td>
<td>373</td>
<td>sseA-&gt; sseB-&gt;</td>
<td>PAIR3 [35]</td>
</tr>
<tr>
<td>tke1</td>
<td>+</td>
<td>150; 180</td>
<td>19.5</td>
<td>2,689,183</td>
<td>212</td>
<td>yhiK-&gt; purL-&gt;</td>
<td></td>
</tr>
<tr>
<td>tp8</td>
<td>-</td>
<td>110; 140</td>
<td>18.3</td>
<td>3,192,705</td>
<td>254</td>
<td>ygiK-&gt; rfaC-&gt;</td>
<td>QUAD1d [35]</td>
</tr>
<tr>
<td>tpk1</td>
<td>+</td>
<td>120; 180</td>
<td>46.8</td>
<td>3,235,948</td>
<td>269</td>
<td>ygiR-&gt; ygdT-&gt;</td>
<td>Attenuator?</td>
</tr>
<tr>
<td>k4</td>
<td>-</td>
<td>200</td>
<td>19.3</td>
<td>3,436,082</td>
<td>197</td>
<td>mecL-&gt; zntR&lt;</td>
<td>sraF</td>
</tr>
<tr>
<td>tpk2</td>
<td>+</td>
<td>250</td>
<td>26.1</td>
<td>4,048,659</td>
<td>268</td>
<td>yihA&lt; yihb&lt;</td>
<td>Attenuator?</td>
</tr>
</tbody>
</table>

Candidate names are arbitrary codes, not final gene names. Candidate positions are for the computational prediction, not the observed transcript, and are relative to the M52 version of the genome. Transcriptional direction of adjacent ORFs in indicated by > or <. The “gene” column indicates gene names assigned to candidates that were also experimentally confirmed by either Argaman et al. [33] or Wassarman et al. [32].

References

18. Sanger Centre Pathogen Sequencing Group. Genome sequence