DRAFT June 7, 2011

Evidence-based annotation of proteins and transcripts in the sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough

Morgan N Price*[1] , Adam M Deutschbauer[1] , Jennifer V Kuehl[1] , Haichuan Liu[2] , H. Ewa Witkowska[2] , Adam P Arkin[1]

[1]Physical Biosciences Division, Lawrence Berkeley National Lab, 1 Cyclotron Road Mailstop 977-152, Berkeley California 94720, USA

[2]UCSF Sandler-Moore Mass Spectrometry Core Facility, Department of Obstetric, Gynecology Reproductive Sciences, Univeristy of California San Francisco, 521 Parnassus Ave, Box 0665, San Francisco California 94143, USA

Email: Morgan N Price*- funwithwords26@gmail.com; Adam M Deutschbauer - amdeutschbauer@lbl.gov; Jennifer V Kuehl - jvkuehl@lbl.gov; Haichuan Liu - Haichuan.Liu@ucsf.edu; H Ewa Witkowska - witkowsk@cgl.ucsf.edu; Adam P Arkin - aparkin@lbl.gov;

*Corresponding author

Abstract

Background: *Desulfovibrio vulgaris* Hildenborough is a model sulfate-reducing bacterium. Despite previous proteomics and gene expression studies, little was known about the accuracy of its genome annotation or the structures of its transcripts.

Results: We used high-resolution tiling microarrays and 5' RNA sequencing to identify transcripts. We combined the transcripts with comparative analysis and proteomics data to make 505 revisions to the original annotation of 3,531 proteins: we removed 255 (7.5%) proteins, changed 123 (3.6%) start codons, and added 127 (3.7%) proteins that were missed. We identified the first nucleotide position for 1,124 transcripts. Sequence analysis of these promoters showed that *D. vulgaris* σ^{70} prefers a different -10 box and -35 box than *Escherichia coli* σ^{70} does. For 72 genes, a major transcript began within the upstream protein-coding gene, which confounds measurements of the upstream gene's expression. 549 transcripts ended at intrinsic (rho-independent) terminators, but most other transcripts seemed to have variable ends. We found low-level antisense expression of most genes, and the 5' ends of these transcripts mapped to promoter-like sequences. Because antisense expression was reduced for highly-expressed genes, we suspect that elongation of non-specific antisense transcripts is suppressed by transcription of the sense strand.

Conclusions: Annotating the proteins in this GC-rich genome was challenging. Tiling data had higher coverage than shotgun proteomics and hence led to most of the corrections, but many errors probably remain. Aspects of the *D. vulgaris* transcriptome that did not match the classical operon model include transcripts that begin at the start codon, promoters within protein-coding genes, non-specific 3' termini, and non-specific antisense transcription. Our data are available at http://genomics.lbl.gov/supplemental/DvHtranscripts2011/.

Background

Desulfovibrio vulgaris Hildenborough can obtain energy by reducing sulfate to sulfide while oxidizing organic material such as lactate or pyruvate. Such sulfate-reducing bacteria play a major role in the global sulfur and carbon cycles and are key drivers of biocorrosion [1]. Sulfate-reducing bacteria are also important in the bioremediation of heavy metal ions such as uranyl, chromate, or zinc, which they can reduce to insoluble forms [1–3]. *D. vulgaris* Hildenborough (henceforth DvH) has become a model for studying sulfate-reducing bacteria, as it was the first sulfate-reducing bacterium sequenced [4] and has been the subject of many studies of proteomics, gene expression, and gene regulation (e.g., [5–10]; hundreds of DvH gene expression experiments are available [11]). We are continuing to analyze the response of DvH to environmental stresses as part of ENIGMA – Ecosystems and Networks Integrated with Genes and Molecular Assemblies – which seeks to understand how environmental conditions affect the bioremediation of heavy metals [12].

As DvH is quite distantly related to well-studied bacteria such as *Escherichia coli* or *Bacillus subtilis*, relatively little is known about gene regulation in this organism. Tiling arrays and next-generation sequencing have been used successfully to map transcripts in other prokaryotes [13], so we undertook to characterize the transcripts of DvH. This should reveal how genes are expressed and should help to infer their regulation. We used two genome-wide methods to analyze DvH transcripts: a high-resolution tiling array with 60-nucleotide probes spaced at 2-4 nucleotides on each strand, which shows the extent of transcripts, and "5' RNA-Seq" to identify their precise 5' ends [14, 15].

Preliminary analysis of our transcript data suggested that there were many errors in the genome annotation (the predicted list of proteins encoded by the genome). Although DvH has been the subject of many proteomics studies, we are not aware of any efforts to use proteomics data to correct its genome annotation. Thus, we combined the transcript data with shotgun proteomics data and homology evidence to revise the genome annotation.

Results

To illustrate our approach, Figure 1 shows the tiling data and the 5' RNA-Seq data for a six-kilobase region of the genome, along with revisions to the genome annotation. We will first discuss the reliability of our data and the identification of transcript starts and ends, and then discuss our changes to the genome annotation and some biological implications.

Reliability of tiling data

We obtained tiling data for mRNA from cells grown with lactate as the carbon source and sulfate as the electron acceptor. We used both a defined minimal medium (LS4D) and a rich medium supplemented with yeast extract (LS4). We also hybridized an array to genomic DNA to measure the strength of each probe. We used this genomic control and the nucleotide content of the probes to normalize the tiling data and to estimate the $\log_2$ level of expression at each probe. Log levels for rich and minimal media were quite similar, with a linear correlation of 0.93 across 2.004 million probes.

Probes for the coding regions of genes usually had higher raw intensity than antisense probes for the opposite strand (Figure 2A). To quantify the difference between the two distributions, we used the Kolmogorov-Smirnov D statistic, a non-parametric measure which ranges from 0 if two distributions are identical to 1 if they do not overlap. The D statistic improved upon normalization: for rich media it improved from 0.72 to 0.74. The overlap between the distributions is primarily due to poorly expressed genes rather than noise in individual probe measurements. For example, if we use only the most highly-expressed two-thirds of genes, then D improves to 0.95 for rich media. The poorly-expressed genes can be seen in the left shoulder of the coding distribution (Figure 2B) and the left of Figure 2C. Hundreds of genes have little expression or are expressed primarily on the antisense (non-coding) strand, but genes that are expected to be essential are well expressed. As discussed below, the annotation of many of the poorly-expressed regions as proteins seems questionable.

Figure 2A also shows that most antisense probes were expressed above control probes that did not match the genome sequence. This might reflect non-specific transcription across the genome, as has been reported in *Escherichia coli* [16, 17]. We will discuss non-specific transcription in more detail below. The presence of non-specific transcripts complicates the determination of a region as "expressed" or not. However, if we assume that the entire genome is transcribed at physiologically relevant levels on one strand or the other, then the median across both strands is the boundary between expressed and not. As our tiling data is normalized to a median of zero, we will use zero as the threshold for expressed (similar to [18]).

Transcript ends

If a transcript has a specific end, then the log level should drop sharply. To quantify whether or not there was a sharp drop at a given location, we asked whether the log levels around that point were correlated with a step function, that is, a sequence of high values before the drop followed by a sequence of low values after the drop [19]. We defined a sharp drop as having a "local correlation" of at least 0.8 and at least a two-fold drop in expression level. We identified 771 sharp drops in tiling data from rich media and 483 sharp drops in tiling data from minimal media. For comparison, based on our updated operon predictions (see below), we estimate that DvH has about 1,200 transcript ends.

When we compared these drops to predictions for intrinsic (rho-independent) terminators from TransTermHP [20], we found that the majority of sharp drops were located at intrinsic terminators (61% in rich media and 75% in minimal media). As shown in Figure 2D, the drop tends to be at about −30 relative to the end of the terminator's stem-loop. Because the probes are 60 nucleotides long, this implies that the drop usually occurs around a probe which ends near the termination site. Overall, we confirmed 771 of 2,978 predicted terminators (Additional file 1), but the predicted terminators often overlap. If we combine the overlapping predictions, then we confirmed 549 distinct terminators.

There were just 25 sharp drops that were found in both rich and minimal media but were not predicted by TransTermHP. We examined these manually and removed three questionable ones, leaving 22 unexplained terminators. There was a terminator prediction on the other strand for 13 of these. These included 5 terminators with uncertain strandedness (TransTermHP's "opp_overlap" flag), and most of the other 8 terminators had a plausible T-rich stretch. So, we believe that most of these 13 terminators that were predicted on the "wrong" strand are bidirectional. To understand the termination of the remaining transcripts, which lack sharp drops, we examined a random sample of 10 genes out of 342 that are well expressed in both conditions (median log level of 1 or higher), are expected to be at the end of their operon (based on revised predictions below), and lacked confirmed terminators. For 6 of these 10 genes, transcription downstream of the gene dropped gradually, without any specific end being apparent; the remaining cases included 3 putative intrinsic terminators with weak drops that were below our thresholds and just one unexplained sharp drop. Overall, virtually all of the specific transcript ends are accounted for by intrinsic terminators, but a significant fraction of transcripts have heterogeneous 3' ends.

The other major mechanism for terminating transcription in bacteria involves the rho protein (reviewed by [21]). Although rho is not well understood, it could account for the heterogeneous ends, and it is estimated to account for about 20% of termination in *E. coli* [22]. However, we suspect that rho activity is weaker in DvH than in *E. coli.* First, we observed an operon which contains the antisense strand of an entire protein-coding gene (gidB, DVU1250; see Supplementary Figure 1). Similar cases of operons extending through antisense of an entire gene have been observed in *B. subtilis*, which has weak rho activity, but not, as far as we know, in *E. coli* [23]. Second, a transposon mutagenesis project in *Desulfovibrio alaskensis* G20 (formerly *D. desulfuricans* G20) found several insertions in rho, which suggests that rho is not required for growth in Desulfovibrios (Arkin lab, unpublished data). In contrast, rho is essential in *E. coli* [24, 25]. Finally, large numbers of transcripts with heterogeneous 3' ends have been reported in the archaeon *Halobacterium salinarium* [26] but not, as far as we know, in other bacteria. Thus, we wonder if DvH has another mechanism for non-specific termination. Further study of whether rho can be knocked out in DvH and what effect this has will clarify this.

Inferring transcript starts from 5' RNA-Seq and tiling data

We extracted mRNA from cells grown in minimal LS4D media and used 5' RNA-Seq to map the 5' ends of the RNAs [14, 15]. We constructed and sequenced two libraries; for one library we used an exonuclease treatment which was intended to remove degradation products. However the two libraries gave similar results (the Spearman correlation of their counts was 0.67) and manual examination suggested that the exonuclease treatment had little effect. We analyzed these libraries together and counted the total number of reads that were uniquely mapped to each location in the genome.

As shown in Figure 1, the number of 5' RNA-Seq reads shows steep peaks (note log scale). Sometimes we see a large number of reads at one position, and a much smaller number of reads at locations within 1-2 nucleotides; these could reflect variation in initiation of transcription from the "same" promoter, or they might arise from minor errors in mapping the 5' ends of the transcripts. In any case, each local peak corresponds to a potential transcription start. Some of these peaks may reflect degradation products rather than genuine transcription starts, but peaks in 5' RNA-Seq that correspond to sharp rises in the tiling data should be genuine transcription starts. Just 2.2% of 5' RNA-Seq peaks with 20-500 reads lie within 30 nucleotides of a sharp rise, while 32% of peaks with over 500 reads do. (We defined a sharp rise as having a local correlation of 0.8 or above in tiling data from rich media.) Most of the peaks with many reads but no corresponding rise in the tiling data lie within highly expressed regions and probably reflect degradation products. In some cases, there are multiple 5' RNA-Seq peaks near each other and the tiling data shows a more complicated or gradual rise, which might reflect multiple start sites, but we cannot rule out that these are degradation products.

We combined our 5' RNA-Seq data with the sharp rises in rich media to obtain a preliminary set of 1,618 transcription starts that were likely to be genuine. (For comparison, based on the revised operon predictions below, there should be around 1,900 transcript starts in DvH.) We searched upstream of these starts for promoter motifs. As shown in Figure 3, we were able to reconstruct the motifs for σ^{70}, rpoN (also known as σ^{54}), and fliA (also known as σ^{28}). Furthermore, we found a σ^{70} site at two thirds of these locations, which is the same rate as in a compilation of transcript starts in *E. coli* [27]. Thus the transcript starts that we identified arise primarily from the initiation of transcription and not from RNA degradation. The DvH genome contains one other sigma factor, rpoH, but we were unable to detect this motif, nor did we detect transcription starts at predicted rpoH-dependent promoters [28, 29], so we suspect that rpoH does not have significant activity under our growth conditions.

To predict which of the 5' RNA-Seq peaks correspond to genuine transcription starts, we used a machine learning approach that took into account the number of reads, the correspondence with tiling data, and the presence of a promoter-like sequence. Because we do not have any known promoters to train our predictor with, we used high- and low-confidence subsets of the data, according to the other features, to train a model for each feature. We then combined the models into a naive Bayesian classifier. The classifier selected 1,124 of the 13,822 peaks as high-confidence transcription starts having a log-odds of 4 or higher ($e^4 \approx 55$; Additional file 2). If we randomize the data then the same model predicts just 31 promoters, so we estimate that $31/1{,}124 = 3\%$ of these transcription starts are false positives.

When we compare the locations of these high-confidence transcription starts in the 5' RNA-Seq data and the tiling data, the 5' RNA-Seq peak tends to be at +20 relative to the center of the rise in the tiling data (Figure 2E). The central tendency confirms that most of the transcription starts are genuine. If 60 nucleotides of hybridization were required for strong signal, then we would expect a overlap of +30, so the location at +20 suggests that hybridization of 50 out of 60 nt suffices for signal.

Revising gene models

The tiling data suggested that there were numberous errors in the genome annotation, as 246 putative proteins from the original annotation were expressed on the wrong strand and lacked homology to other proteins (e.g., DVU1640 and DVU1642 in Figure 1). Furthermore, we sometimes found open reading frames with homology support on the expressed strand (e.g., the DUF497 genes in Figure 1). We found other suspicious patterns in the tiling data as well, such as strong terminators within putative genes, genes that were expressed only near their 3' ends, and genes whose transcripts began downstream of their annotated start codons (Figure 4). Together, these results showed that we needed to reconsider the genome annotation.

To complement the transcript data, we used peptide spectra for DvH from shotgun proteomics efforts within the ENIGMA project (Additional files 3 and 4). We reanalyzed this data using the translation of the genome in all six reading frames, without relying on the genome annotation. To maximize coverage, we considered proteins as detected if they had a single high-confidence peptide; to prevent this from leading to errors in the annotation, we visually inspected the spectra if a proposed change was based on a single peptide (Additional file 5). As shown in Figure 2F, most genes from known families are detected in the tiling data, and a majority of genes from known families were also detected by proteomics. The other genes, which are harder to annotate, were much more likely to be detected by tiling than by proteomics, probably because these proteins tend to be less highly expressed and shorter, which reduces the number of peptides that could be detected. To estimate the rate of false positive identification, we used proteins from the original annotation that were unlikely to be genuine. Specifically, we used proteins that were expressed at least two times higher on the antisense than the sense strand and lacked homology support. Of these 106 proteins, just five were detected, each with one peptide. Manual examination of the spectra suggested that these were false positives (H. L. and A. M. Redding-Johanson, personal communication; these are the removed "proteins" that were detected by proteomics in Figure 2F). Thus, the proteomics data confirmed that these "proteins" should be removed, and our automated protein identification had a false positive rate of around 5%.

We also re-examined the annotation of the genome by homology, as additional genomes from the (rather broad) genus of *Desulfovibrio* have been sequenced since the original annotation, and there have also been improvements to the gene family databases. We used two automated gene finders that consider homology information (CRITICA [30] and RAST [31]) as well as several types of BLAST.

While examining the transcript data for suspicious patterns, we used the following rules for how genes should be expressed. First, the coding strand should be expressed more highly than the non-coding strand. Second, the transcript should start at or before the start codon. If moving the start codon would lead to a very short gene, the gene was removed. Third, the transcript should end after the stop codon. Finally, we ignored genes not transcribed on either strand, as these could be expressed under some other condition. Similarly, we cannot rule out that a transcript satisfying these rules would have been seen under some other growth condition, so genes with clear homology or proteomics support were retained despite violating these rules.

Overall, we made 505 corrections to the genome annotation (Additional file 6). We removed 255 putative proteins, including 154 that were expressed primarily on the wrong strand, 44 that were only expressed (in the tiling data) for a small 3'-terminal portion, and 31 with internal terminators. The remaining 26 proteins were removed because we identified strongly overlapping ORFs with homology or proteomics support in another frame. As shown in Figure 5A, most of the removed proteins were relatively short, with a median length of 54 amino acids, but we removed 43 putative proteins of 100 or more amino acids. We added 128 proteins, including 62 that were identified by both CRITICA and RAST. 32 of the new proteins have informative annotations; the remainder are hypothetical proteins. 13 of the new proteins were identified by neither gene calling program or were originally annotated as pseudogenes: 4 of these were were detected by proteomics and their spectra were validated by inspection (H. L. and A. M. Redding-Johanson, personal communication) and the other 9 had strong homology support. Finally, we changed 123 start codons. We moved 35 of them upstream, mostly due to proteomics (for proteins with a single upstream peptide, we inspected the spectra to verify the change). A few start codons were moved well upstream after examining genes with long gaps between the transcript start and the start codon and checking for conservation of the intervening sequence. We moved 88 start codons downstream, usually because the gene's transcript started downstream of the original start codon. As shown in Figure 5B, many of the changes to the start codon were quite large, with a median absolute difference of 37 amino acids. Overall, 80% of proteins in our revised annotation were covered from start codon to stop codon by transcripts in the tiling data, and 54% of proteins were detected in the proteomics data.

We were surprised at extent of these corrections and the lack of agreement between the two automated tools. CRITICA missed 12% of genes in our revised annotation and RAST missed 7% of genes in our revised annotation. Among genes predicted by both tools, the start codon differed 32% of the time. Conversely, 0.9% of CRITICA calls and 5.6% of RAST calls were not included in our revised annotation and are likely to be false positives. As we were rarely able to correct start codons that were too far downstream, we expect that many of the start codons in our revised annotation are still erroneous. An accurate annotation will require proteomics with higher coverage or targeted to N-terminal peptides

[32].

Leaders and untranslated transcribed regions

We identified 5' and 3' untranslated transcribed regions (UTRs) by checking whether the entire region between a transcript's boundary and the nearest gene was expressed. As discussed above, many transcripts show non-specific ends, which makes it problematic to define the 3' UTR, so we only analyzed the 3' UTRs for transcripts with intrinsic terminators. We defined 983 5' UTRs and 494 3' UTRs (Additional file 6).

One surprise was the presence of "leaderless" promoters where the transcript begins at the first nucleotide of the start codon. Leaderless transcripts were first identified in archaea, but they have been identified in various bacteria in genome-wide studies, including in *Geobacter sulfurreducens* PCA, which like DvH is a δ-Proteobacterium [33]. However, given the high rate of error in start codon annotations, we wondered if the leaderless promoters in DvH were genuine. We checked the start codons for candidate leaderless promoters from a preliminary version of our analysis by asking whether homology extended to the very N-terminal end of the annotation. 43 of 49 of our preliminary candidates were confirmed by BLASTp and 21 of these were further confirmed by alignments to known families. The remaining 6 N-terminal regions were not conserved and might be erroneous, but most of the leaderless promoters must be genuine. Our final analysis gave 54 proteins with leaderless promoters out of 954 proteins that are at the beginning of transcripts with clearly defined starts.

As shown in Figure 6A, the median length of the 5' UTR is 55 nucleotides, but some genes have very long 5' UTRs. Two operons that are central to sulfate reduction have particularly long 5' UTRs: dsrABD, which encodes three subunits of the dissimilatory sulfite reductase, has a 5' UTR of 289 nucleotides, and apsB, which encodes a subunit of adenylylsulfate reductase, has a 5' UTR of 208 nucleotides. However, in general, we could not find a clear pattern for which types of genes had long 5' UTRs. Among 5' UTRs of over 100 nucleotides for genes on the main chromosome, about half (106/208) had some conservation in another strain, *D. vulgaris* Miyazaki B, according to a genome alignment [34]. (The Miyazaki B strain is sufficiently diverged from DvH that there should be no neutral conservation of non-functional DNA.) This suggests that many of these 5' UTRs contain functional elements; however we cannot be certain that these function as RNA elements rather than as alternative promoters.

For the 494 genes with a confirmed terminator downstream, the median length of the 3' UTR was 68 nucleotides (Figure 6B). Of 147 3' UTRs of over 100 nt, just 16 contained segments that were conserved in *D. vulgaris* Miyazaki B; thus, we predict that few of these 3' UTRs contain functional sequences.

Finally, we found little evidence of transcribed regions that are not associated with annotated genes. We found just 26 unannotated transcribed regions with high-confidence promoters, and after removing antisense transcripts this dropped to just 4. However, both our experimental protocols and our analysis methods are probably biased against RNAs of under 100 nucleotides, so this does not imply that DvH lacks small RNAs.

Revising operon structures

Before we began this project, we had predicted operons from the distances between genes on the chromosome, how conserved the proximity of the genes was, whether the genes had similar expression patterns across a large collection of microarray experiments, and whether they were likely to have related functions [11, 35]. Here, we use the transcript data to update the operon predictions. We classifed each adjacent pair of genes on the same strand as a simple operon pair, as a complex operon pair with an internal operon or internal attenuator, or a non-operon pair. (Examples of operons with internal attenuators or internal promoters are shown in Supplementary Figure 2.) We began with our original predictions (which predict whether pairs are ever cotranscribed or not) and reclassified pairs with clear signals in our data. Ambiguous cases occurred if there was a weak drop and then a high-confidence transcription start just downstream of the drop – this could be a genuine terminator followed by a promoter, but the tiling data lacks the resolution to distinguish the drop clearly, or it could be noise in the tiling data.

Relative to our original predictions, which had 1,558 operon pairs and 838 non-operon pairs, we reclassified 188 non-operon pairs as simple operons; we reclassified 14 operon pairs as non-operons; we identified 169 complex operon pairs with internal promoters, about half of which were originally classified as operons; and we identified 17 complex operon pairs with internal attenuators, 12 of which were originally classified as operons (Additional file 7). We were surprised at the number of pairs that were reclassified from non-operons to simple operons. These tended to be widely spaced (median separation of 108 nucleotides) and moderately coexpressed (median Pearson correlation of 0.19), which explains why the were classified as non-operon pairs in our original predictions. The wide spacing and the moderate coexpression also suggested that these might contain internal promoters that were missed by our automated analysis. However, only 30 of these 188 pairs had potential internal transcript starts according to our classifier (log odds values of 0 to 4). Manual examination of 10 randomly selected cases found potential internal promoters for just two out of the ten. The weak coexpression could be due to internal promoters that are not active under our growth conditions or due to noise in the expression compendium. Comparison of tiling data from a wide range of growth conditions [26] would be one way to distinguish these alternatives.

As shown in Figure 7, genes that are cotranscribed but also have an intergenic promoter between them show little coexpression. We suspect that this is because we can only identify internal promoters with high confidence if they are stronger than the upstream promoter, so that the upstream gene is transcribed only from the upstream promoter and the downstream gene is transcribed primarily from the intergenic promoter. If there is an internal promoter that is within the upstream gene, however, we see a much stronger coexpression ($P < 10^{-4}$, Wilcoxon rank-sum test). Because the expression data was collected with 1-2 probes per gene and thus lacks spatial resolution, we suspect that this coexpression is an artefact – the probe for the upstream gene hybridizes to the internal transcript, which does not include the upstream gene's start codon and cannot lead to its expression. Thus, the gene expression data for the upstream gene is misleading. Knowledge of transcript structures will allow for better design of gene expression arrays.

Non-specific transcription and termination

As mentioned above, the tiling data suggested that there is weak and potentially non-specific expression of the antisense strand of most genes. To confirm this pattern, we examined the 5' RNA-Seq data. Because we were interested in non-specific effects, we used all local peaks in the data, rather than only the stronger peaks considered while analyzing transcript starts. 1.4% of the mapped reads began within coding regions on the antisense strand of genes in our updated annotation. (For comparison, 33% of the reads corresponded to high-confidence promoters and 15% of reads began within coding regions on the sense strand.) Only 46 of these 3,983 antisense starts were classified as high-confidence starts by our statistical model. To verify that the weak antisense transcript starts are genuine, we asked if they were located at promoter-like signals (similar to [17]). The antisense starts were quite enriched in weak promoter signals; for example, 35% of the antisense locations, but only 12% of randomized locations, had 4 bits or more of similarity to a promoter motif ($P < 10^{-15}$, Fisher exact test). In contrast, starts within coding regions had little enrichment in promoter signals, which suggests that most of them are degradation products.

As shown in Figure 2C, antisense expression tends to be less for genes that are more highly expressed on the sense strand. If we average across the two growth conditions, the correlation is -0.40 ($P < 10^{-15}$; using the rank correlation gave similar results). If we consider only genes that are expected to be essential, then the correlation is -0.60 ($P < 10^{-15}$), which shows that the effect is not due to misannotated genes or to genes that are not expressed at all.

In most prokaryotes, promoter-like sequences within genes are selected against and occur a bit less frequently than expected by chance [36]. We wondered how promoter-like sequences might relate to our evidence for non-specific transcription. To avoid artefacts due to annotation errors or the edges of genes, we considered only longer genes (300 nucleotides or longer) that belong to known families. We found that highly-expressed genes contained fewer internal promoter-like sequences (≥ 4 bits) per kilobase on the sense strand (Spearman $\rho = -0.23$, $P < 10^{-15}$). However, expression level had little effect on the rate of promoter-like sequences on the antisense strand ($\rho = -0.04$, $P = 0.06$). Because the rate of promoter-like sequences on either strand is strongly correlated with GC content ($\rho = 0.80$ for the sense strand and $\rho = 0.67$ for the antisense strand), we also tested the relationship using partial correlations. The effect of expression levels on sense-strand promoter motifs remains after controlling for GC content (partial $\rho = -0.09$, $P < 10^{-5}$).

Consistent with the sequence analysis, the rate of antisense transcript starts (in reads per kilobase) was not significantly reduced for highly expressed genes ($\rho = -0.02$, $P > 0.2$).

We propose that promoter-like sequences on the sense strand are selected against to prevent expression of truncated proteins, while transcription on the antisense strand is suppressed by transcription on the sense strand. Because we see antisense suppression in the tiling data but not in the 5' RNA-Seq data, it appears that elongation, rather than initiation, is suppressed. Although a promoter on one strand can suppress transcription intiation from the opposite strand, this seems to rely on a specific site where the RNA polymerase pauses [37] and would not occur in most situations. We do not know what suppresses the elongation of antisense transcripts for highly expressed genes. One possibility is that elongation of antisense transcripts is suppressed because the RNA polymerase backtracks when it collides with RNA polymerase on the sense strand [38]. Such collisions would occur more frequently for highly-expressed genes, and the RNA polymerase on the sense strand might "win" these collisions because translating ribosomes follow closely behind the RNA polymerase on the sense strand and prevent the RNA polymerase from backtracking [39].

Discussion

Evidence-based annotation of proteins

We identified far more changes to the protein annotation (505) than we had expected from transcriptomics efforts in other bacteria. We have also collected transcript data for *Desulfovibrio alaskensis* G20 and have found a similar number of errors in the original annotation there as well (Arkin lab unpublished data). For comparison, the annotation of *Geobacter sulfurreducens* PCA was updated recently using transcript data and shotgun proteomics, which resulted in only 144 changes [33]. *Desulfovibrio* genomes are GC-rich, which increases the number of spurious long reading frames, and there are relatively few genome sequences for *Desulfovibrios*, which makes comparative gene-finding tools such as CRITICA less effective, but both of these challenges apply to *G. sulfurreducens* as well. Our preliminary analysis suggests that many plausible corrections to the *G. sulfurreducens* annotation remain: we found 39 protein-coding genes in the updated annotation that lack homology support, were not in the proteomics data, and were only expressed on the "wrong" strand. Nine of these "genes" mask unannotated proteins with homology support on the opposite strand. As the tiling and RNA-Seq experiments described in this paper cost less than sequencing a genome did a few years ago, transcriptomics could be used broadly to improve genome annotation.

We were also surprised at the number of changes we made based on the tiling data that, in retrospect, could have been made by homology alone. There were 24 proteins that we removed because they lacked homology support and a conflicting frame had homology support, and there were 10 proteins with homology support that were missed in the original annotation and by both RAST and CRITICA. Neither RAST nor CRITICA uses the full range of approaches to detecting protein homology: RAST relies primarily on pairwise protein comparisons to representatives of known families and CRITICA relies on nucleotide BLAST hits. We found additional proteins by comparing sequences to families with PSI-BLAST [40] or HMMer [41], which can find highly diverged members of known families, and also by comparing to hypothetical proteins that were annotated in other organisms. Faster tools (e.g., HMMer 3 [41] or FastBLAST [42]) should allow for more exhaustive searches and hence more accurate automated annotation.

Our large-scale revisions to the DvH annotation will change the results of many computational analyses and will also affect the design of experiments. For example, changes to the predicted start codons will also affect attempts to clone a gene or tag a protein. We also identified a technical challenge with interpreting gene expression patterns when there are promoters within genes, which will affect the design of gene expression microarrays.

Tiling array design

We used very high-resolution microarrays, with probes every 2-4 nucleotides. We had hoped that such a high density would let us place promoters and terminators very accurately, but that turned out not to work: the resolution of promoters from tiling alone was just 20 nucleotides or so. This is probably because of non-full-length hybridization to 60-mer probes. The high density of the arrays is still beneficial, as nearby probes are a form of replicates.

Still, it would be more cost-effective to use arrays with probes every 6-10 nucleotides and to combine results from more conditions.

Transcript structures

We hope that our revisions to operon structures, along with the 1,124 transcript starts that we identified at nucleotide resolution, will aid the elucidation of gene regulation in DvH. They have already led to revisions in the regulons of the sigma factors rpoN and fliA [43]. As regulatory sequences tend to be near the promoter, promoter locations should aid in the identification of transcription factor binding sites more broadly. In particular, we have used the transcript starts to help us interpret data on where transcription factors bind in the genome (L. Rajeev and A. Mukhopadhyay, unpublished data).

Conclusions

We combined tiling microarrays, 5' RNA-Seq, and proteomics to reannotate the genes and transcripts of *D. vulgaris* Hildenborough. We corrected hundreds of errors in the genome annotation but many more errors probably remain, particularly in the identification of start codons. Although our 5' RNA-Seq protocols returned a mixture of true transcript starts and likely degradation products, we were able to identify 1,124 genuine transcription starts at a false positive rate of a few percent by combining the reads with tiling data and sequence analysis. We found that DvH σ^{70} prefers a different motif than its *E. coli* counterpart and that many transcripts have non-specific 3' ends. Finally, we found non-specific transcription of the antisense strands of protein-coding genes in both the tiling and the 5' RNA-Seq data; elongation of these non-specific antisense transcripts seems to suppressed by transcription of the sense strand. All of our results, including raw data, processed results, modifications to the annotation, and source code, are available from our web site [44]; the data is also available at the Gene Expression Omnibus (GSE29560).

Materials and methods

Strains and growth conditions

Experiments were conducted within a Coy anaerobic chamber with an atmosphere of about 2% H_2 and 5% CO_2, with the remainder being N_2. *Desulfovibrio vulgaris* Hildenborough was inoculated from a 1% glycerol stock and grown in glass bottles with lactate-sulfate media at 30°C. Cells were collected at an optical density of around 0.3. Tiling data was collected from cells grown in two conditions: defined LS4D medium [8] and LS4, which is LS4D supplemented with 0.1% w/v yeast extract. 5' RNA-Seq data was collected from defined LS4D medium.

RNA collection

Bacterial pellets were collected by centrifuging cultures for 10 minutes at 10,000 g at 4°C in RNAse free 50ml polypropylene tubes. Supernatant was immediately poured off and pellets were stored at −80°C. After thawing, RNA was extracted with RNeasy miniprep columns (Qiagen) with the optional on-column DNase treatment. RNA quality was confirmed with an Agilent Bioanalyzer; only samples with an RNA integrity number of around 9 or better were used. Ribosomal RNA was depleted with the MICROBExpress kit (Ambion), which uses magnetic beads coated with oligonucleotides that hybridize to ribosomal RNA. These mRNA-enriched samples were analyzed with tiling arrays or 5' RNA-Seq.

Tiling experiments

First-strand cDNA was synthesized with random hexamer primers using SuperScript indirect cDNA labeling system (Invitrogen); the reaction buffer was supplemented with actinomycin D to inhibit second-strand synthesis [45]. First-strand cDNA was labeled with Alexa 555. About 2 μg of labeled first-strand cDNA was hybridized to the Nimblegen array. For the genomic control, we used DNA from cells in stationary phase to minimize copy number variation across the chromosome. Genomic DNA was extracted using the DNeasy blood tissue kit (Qiagen) and labeled with Nimblegen's comparative genomic hybridization protocol. Briefly, genomic DNA was sonicated to 200-1000 bp and amplified using Klenow fragment and Cy3-labeled random nonamer primers. Nimblegen slides were scanned on an Axon Gene Pix 4200A scanner with 100% gain and analyzed with Nimblescan, with no local alignment and a border value of −1. For rich media, we used the average of log intensities from two independent experiments, while for minimal media and the genomic control we did just one experiment.

To remove probes that might cross-hybridize, we mapped the probes to the genome (NC_002937 and NC_005863) with BLAT [46] and we ignored any probes whose second-best hit matched at 50 or more nucleotides. We computed normalized log levels by using the genomic control and by using each probe's nucleotide content, followed by setting the median value to zero. First, we used a linear regression to model the $\log_2$ intensity as a function of the log intensity in the genomic control and the probe's nucleotide content. To compute this model, we used only probes within the sense strands of genes because of differences in nucleotide composition between coding and non-coding regions and even between the coding and antisense strands of genes. The prediction of this model is the expected bias of the probe, so we subtracted this from the (raw) log intensity. We also removed the data for the 1% of probes with the lowest intensities in the genomic control, as these probes gave poor discrimination between coding and non-coding regions. Finally, we adjusted the normalized values so that their median was 0.

5' RNA-Seq experiments

Given an mRNA-enriched sample, we converted 5'-triphosphate ends to 5'-monophosphate with Tobacco Acid Pyrophophatase, we blocked the 3' ends with sodium periodate, and we added a sequencing adaptor (5'-ACACUCUUUCCCUACACGACGCUCUUCCGAUCU-3') onto the 5' end with Ambion T4 RNA ligase [15]. We used random hexamer primers with a sequencing adaptor on their 5' end (5'-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTNNNNNN-3') to obtain first-strand cDNA. We size-selected products of 150-500 bases from an agarose gel. We PCR amplified the library to enrich for products that contained both adaptors and to complete the 5' adaptor, using primers 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' and 5'-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT-3'. We purified the PCR products and removed unincorporated nucleotides, primers, and adaptor-only products with AMPure XP Beads (Agencourt). We also made a second library in which we used Terminator 5'-Phosphate-Dependent Exonuclease (Epicentre) to try to remove 5'-monophosphate (degraded) transcripts and then converted 5'-triphosphate ends to 5'-monophosphate ends with RNA 5' Polyphosphatase (Epicentre) [14]. Ligation, cDNA and PCR amplification conditions were similar in both libraries. Between each enzymatic reaction of the exonuclease library, RNA was purified using Agencourt RNAClean XP beads. Molecules smaller than 100 nucleotides and unligated adaptors were mostly lost in these clean-up reactions.

For each library, the 32 nucleotides at the 5' end were sequenced with a lane of Solexa by the University of California at Davis sequencing center and the reads were mapped to the genome with Eland. From the first library, 7.5 million reads mapped uniquely to the genome; from the second library, 15.5 million reads mapped uniquely to the genome. (Reads from ribosomal RNAs would not map uniquely, as DvH contains 5-6 nearly identical copies of each ribosomal RNA.) We identified local peaks (within 50 nucleotides) in the combined number of reads in the libraries. Peaks with at least 10 reads in each library were considered as potential transcript starts.

Identifying features in the data

We identified rises or drops in each tiling condition by computing the "local correlation" [19]. We used the data for 50 probes on either side of a potential rise or drop and we asked how similar this pattern was to a step function by measuring the absolute value of the correlation between this subset of the data and a series of -1 values followed by an equal number of 1 values. We measured the local correlation around every probe; to identify the center of the rise or drop, we used the local maximum of the local correlation within 21 probes.

To confirm an intrinsic terminator, we required a sharp drop within 60 nucleotides of the end of the stem loop that had a local correlation of 0.8 or better. We also required a change in the average log level between either side of the drop of at least 1 (i.e., a two-fold change in intensity).

To identify a break in transcription within a potential operon or between a transcript boundary and a gene, we smoothed the normalized log level over five adjacent probes. If the minimum of the smoothed values was below zero, we identified a break in the transcript under that condition. To identify breaks in putative operons, we also required a difference of at least 1 between the expression level of the upstream gene and this minimum.

Transcribed regions were defined by smoothing over 40 adjacent probes (roughly 150 nucleotides) and requiring a smoothed value of above 0.

Promoter sequence analysis

We began with a preliminary set of 1,618 moderate-confidence transcription starts, based on a rise in rich media occuring within 30 nucleotides of a local peak in the 5' RNA-Seq data. We extracted positions -40 to +1 relative to these putative promoters and analyzed only the strand in the orientation of transcription. We used BioProspector [47] to search for a bipartite motif with blocks of width 10 and 8 separated by 10 to 18 nucleotides, and kept the best of 12 runs of its Gibbs sampler. We used MEME [48] to search for ungapped motifs of 30-35 nucleotides under the zero-or-one-occurence-per-site model and found four significant motifs. We used patser [49] to scan the entire genome for hits to any of the four MEME motifs and to correct for the high GC content of the DvH genome. For most analyses we used only hits of 7 bits or above, which across the four motifs gave a hit every 111 nucleotides on each strand of the genome. We associated a motif hit with a 5' RNA-Seq peak if the peak was within one nucleotide of the expected location.

Distinguishing transcription starts from RNA degradation products

We used a semi-supervised machine learning approach to classify local peaks in the 5' RNA-Seq data as transcription starts or other. For each local peak, we computed four features: (1) n_{tot}: the total number of 5' RNA-Seq reads mapped to starting at that location; (2) r_{rich}: whether the 5' RNA-Seq peak was associated with a transcribed region and with a rise in the rich media tiling data with a local correlation of 0.6 or above, and if so, what the local correlation was; (3) r_{min}: the corresponding value for minimal media tiling data, but with a threshold of 0.7 and without consideration of whether it was associated with a transcribed region; and (4) b, the bit score of the best hit to any of the four MEME promoter motifs that occurred within 1 nucleotide of the putative transcription start, if any (weak hits of under 7 bits were ignored). For each feature, we inferred a model (a log-odds score for any given value) by comparing the distribution of the feature for transcription starts that were high- or low-confidence according to the other features. Specifically, we inferred a model for r_{rich} by comparing starts with both a r_{min} and b value to starts with neither. We inferred a model for r_{min} in the analogous way. We inferred a model for b by comparing starts with both r_{rich} and r_{min} values to starts with neither. Finally, we inferred a model for n_{tot} by comparing starts with a total log odds score, from the other features, of above 4, to starts with a log odds score of under -4. ($e^4 \approx 55$ so these are about 55 times more likely to be genuine or false starts.)

Models were inferred using binned subsets of the data, pseudocounts, and smoothing [35]. The R code for this step is available in the MicrobesOnline code base, see BinnedBinaryFit2() and BBFPredict() in util/utils.R [50]. We summed the log odds for each feature to get a final log odds score. This corresponds to assuming that the features are independently distributed amongst the false positives and amongst the true transcription starts, as in a naive Baysian classifier. Values above zero indicate that the transcription start resembles the high confidence transcription starts, and the magnitude of the log odds indicates the level of confidence. We considered starts above a log odds of 4 to be high confidence starts. The distributions of the features for the high-confidence starts and the other starts are shown in Supplementary Figure 3. To estimate the false positive rate, we used a randomized data set: we replaced the locations of all 5' RNA-Seqs peaks with random locations, we recomputed all features, we shuffled the resulting values to eliminate the (biological) agreement between them, and we applied the model.

Shotgun proteomics

Mass spectra were collected for peptides derived from a variety of protein fractions from the ENIGMA project. We also used spectra from previously-published whole-cell proteomics experiments with DvH grown under several stress conditions [9, 10, 51]. All spectra were analyzed against the six-frame translation of the genome. For protein fractions, spectra were analyzed with the Paragon algorithm in ProteinPilot 3.0 [52], and peptides were considered confidently identified at a posterior probability of 0.95, resulting in 22,503 different peptides for 1,866 reading frames. For complete-proteome experiments, reading frames were considered confidently identified if they had a MASCOT score of 32 or greater, resulting in 1,556 reading frames detected. If a change to the annotation derived from a single peptide, then the relevant spectra were checked by hand.

Correcting gene annotations

Candidate proteins to change or remove were selected based on the tiling data, the transcript starts, and the rules discussed above, and were checked manually. Data was viewed in Artemis [53]. Homology evidence for a protein was examined using the domains and homologs tools on the MicrobesOnline web site [11]: these show HMMer 3 [41] or PSI-BLAST [40] hits to known families and FastBLAST [42] hits to other proteins.

Potentially missed proteins were identified by examining proteomics data, by using CRITICA and RAST, by using PSI-BLAST to search for homology to conserved domains [54] in the six-frame translation of the genome, by using blastx [55] to compare the six-frame translation of unannotated transcribed regions to annotated proteins in other organisms, and by checking candidate open reading frames with MicrobesOnline's sequence search and with the PFam web site [56].

After initial changes to the annotation, we examined genes with significant overlaps. In particular, if a gene with homology or experimental support overlapped a gene without support, we removed the unsupported protein. Conversely, a few proteins with odd patterns in the tiling data were retained because they had strong homology or proteomics support (using the proteomics data discussed above and also data from [7]).

We began our analysis with an older RefSeq annotation (from 2007) and all of the changes discussed above are relative to this annotation. As of December 2010, the RefSeq annotation had changed, presumably based on comparative genomic analyses. On the main chromosome, RefSeq had removed 41 of the 255 ORFs that we removed, along with removing one other ORF that is masked by RefSeq's change to a start codon. RefSeq had added 28 of the 127 ORFs that we added, along with 1 more ORF that seems questionable to us. Finally, RefSeq had changed 68 start codons, which includes 26 changes in the same direction as our changes, 1 change in the opposite direction, and 41 changes to genes that we did not modify. Of those 41 changes, 22 agree with both CRITICA and RAST or agree with at least one of the two and have some homology support; we lack confidence in the other 19 but they are consistent with our transcriptome data and could be correct. Overall, there was moderate overlap between our changes and the RefSeq changes, which illustrates the difficulty of annotating proteins in DvH. An update to the DvH annotation that includes the changes from RefSeq that we accept is available from our web site [44].

Revising operon structures

We classified adjacent pairs of genes on the same strand based on whether there was a high-confidence "internal" transcript start (that is, between the upstream gene's start codon and the downstream gene's start codon) and whether there was a break in expression between the genes. Simple operon pairs had neither an internal transcript start nor a 2-fold drop in expression in the intergenic region. Non-operon pairs had a drop by at least two-fold to below a log-level of 0 or had both an internal transcript start and a confirmed terminator. However, we classified pairs as having attenuators if they had a confirmed terminator but a log level of at least 0.25 throughout their intergenic region. If a pair had an internal promoter and no drop then it was classified as a complex operon pair. The various thresholds were validated by manually examining the results.

Statistics

All statistical tests and regressions were conducted in R [57].

Authors' contributions

AMD and MNP designed transcriptomics experiments. JVK conducted transcriptomics experiments. HL and HEW analyzed peptide spectra. MNP analyzed data and wrote the paper. APA and AMD supervised the project.

Acknowledgements

We thank Alyssa M. Redding-Johanson for reanalyzing previously-published proteomics data against the six-frame translation of the genome. We thank John-Marc Chandonia, Evelyn Dora Szakal, and Allen Simon for providing proteomics data and we thank the ENIGMA protein complexes group for providing protein fractions.

We thank Marcin Joachimiak for designing the tiling array. We thank Judy Wall, Terry Hazen, and their research groups for advice on working with DvH. This work conducted by ENIGMA – Ecosystems and Networks Integrated with Genes and Molecular Assemblies – was supported by the Office of Science, Office of Biological and Environmental Research, of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231. The funding agency was not involved in the study.

References

1. Muyzer G, Stams AJM: **The ecology and biotechnology of sulphate-reducing bacteria**. *Nat Rev Micro* 2008, **6**:441–454.

2. Lovley DR, Phillips EJ: **Reduction of Chromate by Desulfovibrio vulgaris and Its c(3) Cytochrome**. *Appl Environ Microbiol.* 1994, **60**:726–8.

3. Wall JD, Krumholz LR: **Uranium reduction**. *Annu Rev Microbiol* 2006, **60**:149–66.

4. Heidelberg JF, Seshadri R, Haveman SA, Hemme CL, Paulsen IT, Kolonay JF, Eisen JA, Ward N, Methe B, Brinkac LM, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Fouts D, Haft DH, Selengut J, Peterson JD, Davidsen TM, Zafar N, Zhou L, Radune D, Dimitrov G, Hance M, Tran K, Khouri H, Gill J, Utterback TR, Feldblyum TV, Wall JD, Voordouw G, Fraser CM: **The genome sequence of the anaerobic, sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough.** *Nat Biotechnol.* 2004, **22**:554–9.

5. Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS: **Reconstruction of regulatory and metabolic pathways in metal-reducing δ-proteobacteria**. *Genome Biol.* 2004, **11**:R90.

6. He Q, Huang KH, He Z, Alm EJ, Fields MW, Hazen TC, Arkin AP, Wall JD, Zhou J: **Energetic consequences of nitrite stress in Desulfovibrio vulgaris Hildenborough, inferred from global transcriptional analysis.** *Appl Environ Microbiol.* 2006, **72**:4370–81.

7. Nie L, Wu G, Zhang W: **Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: A multiple regression to identify sources of variations**. *Biochem Biophys Res Commun.* 2006, **339**:603–610.

8. Mukhopadhyay A, He Z, Alm EJ, Arkin AP, Baidoo EE, Borglin SC, Chen W, Hazen TC, He Q, Holman HY, Huang K, Huang R, Joyner DC, Katz N, Keller M, Oeller P, Redding A, Sun J, Wei JWJ, Yang Z, Yen HC, Zhou J, Keasling JD: **Salt stress in Desulfovibrio vulgaris Hildenborough: an integrated genomics approach.** *J Bacteriol.* 2006, **188**:4068–78.

9. Mukhopadhyay A, Redding AM, Joachimiak MP, Arkin AP, Borglin SE, Dehal PS, Chakraborty R, Geller JT, Hazen TC, He Q, Joyner DC, Martin VJ, Wall JD, Yang ZK, Zhou J, Keasling JD: **Cell-wide responses to low-oxygen exposure in Desulfovibrio vulgaris Hildenborough.** *J Bacteriol.* 2007, **189**:5996–6010.

10. Zhou A, He Z, Redding-Johanson AM, Mukhopadhyay A, Hemme CL, Joachimiak MP, Luo F, Deng Y, Bender KS, He Q, Keasling JD, Stahl DA, Fields MW, Hazen TC, Arkin AP, Wall JD, Zhou J: **Hydrogen peroxide-induced oxidative stress responses in Desulfovibrio vulgaris Hildenborough**. *Environ Microbiol.* 2010, **12**:2645–57.

11. Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin AP: **MicrobesOnline: an integrated portal for comparative and functional genomics.** *Nucleic Acids Res.* 2009, **database issue**.

12. **ENIGMA – Ecosystems and Networks Integrated with Genes and Molecular Assemblies**[http://enigma.lbl.gov/].

13. Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity**. *Nature Rev. Genet.* 2010, **11**:9–16.

14. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BO: **The transcription unit architecture of the Escherichia coli genome**. *Nat Biotchnol.* 2009, **27**:1043–9.

15. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R: **A single-base resolution map of an archaeal transcriptome**. *Genome Res.* 2010, **20**:133–141.

16. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM: **RNA expression analysis using a 30 base pair resolution Escherichia coli genome array**. *Nat. Biotechnol.* 2000, **18**:1262–8.

17. Dornenburg JE, DeVita AM, Palumbo MJ, Wade JT: **Widespread antisense transcription in Escherichia coli**. *mBio* 2011, **1**:e00024–10.

18. Halasz G, van Batenburg MF, Perusse J, Hua S, Lu X, White KP, Bussemaker HJ: **Detecting transcriptionally active regions using genomic tiling arrays**. *Genome Biol.* 2006, **7**:R59.

19. Güell M, van Noort V, Yus E, Chen W, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin A, Bork P, Serrano L: **Transcriptome Complexity in a Genome-Reduced Bacterium**. *Science* 2009, **326**:1268–71.

20. Kingsford C, Ayanbule K, Salzberg S: **Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake.** *Genome Biol.* 2007, **8**:R22.

21. Ciampi MS: **Rho-dependent terminators and transcription termination**. *Microbiology* 2006, **152**:2515–28.

22. Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R: **Rho directs widespread termination of intragenic and stable RNA transcription**. *Proc Natl Acad Sci U S A* 2009, **106**:15406–11.

23. de Hoon MJ, Makita Y, Nakai K, Miyano S: **Prediction of Transcriptional Terminators in Bacillus subtilis and Related Species.** *PLoS Comput. Biol.* 2005, **1**:e25.

24. Bubunenko M, Baker T, Court DL: **Essentiality of Ribosomal and Transcription Antitermination Proteins Analyzed by Systematic Gene Replacement in Escherichia coli**. *J Bacteriol.* 2007, **189**:2844–53.

25. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, Furubayashi A, Kinjyo S, Dose H, Hasegawa M, Datsenko KA, Nakayashiki T, Tomita M, Wanner BL, Mori H: **Update on the Keio collection of Escherichia coli single-gene deletion mutants.** *Mol Syst Biol.* 2009, **5**:335.

26. Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo F, Pratap A, Deutsch EW, Peterson A, Martin D, Baliga NS: **Prevalence of transcription promoters within archaeal operons and coding sequences**. *Mol Syst Biol* 2009, **5**:285.

27. Hershberg R, Bejerano G, Santos-Zavaleta A, Margalit H: **PromEC: An updated database of Escherichia coli mRNA promoters with experimentally identified transcriptional start sites**. *Nucleic Acids Res.* 2001, **29**:277–00.

28. Chhabra SR, He Q, Huang KH, Gaucher SP, Alm EJ, He Z, Hadi MZ, Hazen TC, Wall JD, Zhou J, Arkin AP, Singh AK: **Global analysis of heat shock response in Desulfovibrio vulgaris Hildenborough**. *J Bacteriol.* 2006, **188**:1817–28.

29. Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I, Rodionov DA: **RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes**. *Nucleic Acids Res.* 2010, **Database issue**:D111–8.

30. Badger JH, Olsen GJ: **CRITICA: coding region identification tool invoking comparative analysis.** *Mol Biol Evol.* 1999, **16**:512–24.

31. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O: **The RAST Server: Rapid Annotations using Subsystems Technology**. *BMC Genomics* 2008, **9**:75.

32. Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, Klein C, Martens L, Staes A, Timmerman E, Damme JV, Siedler F, Pfeiffer F, Vandekerckhove J, Oesterhelt D: **Large-Scale Identification of N-Terminal Peptides in the Halophilic Archaea Halobacterium salinarum and Natronomonas pharaonis**. *J. Proteome Res.* 2007, **6**:2195–2204.

33. Qiu Y, Cho BK, Park YS, Lovley D, Palsson BO, Zengler K: **Structural and operational complexity of the Geobacter sulfurreducens genome.** *Genome Res.* 2010, **20**:1304–11.

34. Darling AC, Mau B, Blatter FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements**. *Genome Research* 2004, **14**:1394–1403.

35. Price MN, Huang KH, Alm EJ, Arkin AP: **A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes**. *Nucleic Acids Res.* 2005, **33**:880–92.

36. Hahn MW, Stajich JE, Wray GA: **The Effects of Selection Against Spurious Transcription Factor Binding Sites**. *Molecular Biology and Evolution* 2003, **20**(6):901–906.

37. Palmer AC, Ahlgren-Berg A, Egan JB, Dodd IB, Shearwin KE: **Potent Transcriptional Interference by Pausing of RNA Polymerases over a Downstream Promoter**. *Molecular cell* 2009, **34**:545–555.

38. Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH: **Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy**. *Nucleic Acids Res.* 2006, **34**:5416–24.

39. Proshkin S, Rahmouni AR, Mironov A, Nudler E: **Cooperation Between Translating Ribosomes and RNA Polymerase in Transcription Elongation**. *Science* 2010, **328**(5977):504–508.

40. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al.: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res.* 2001, **29**:2994–3005.

41. **HMMer 3**[http://hmmer.janelia.org/].

42. Price MN, Dehal PS, Arkin AP: **FastBLAST: Homology Relationships for Millions of Proteins**. *PLoS ONE* 2008, **3**:e3589.

43. **RegPrecise**[http://regprecise.lbl.gov/RegPrecise/].

44. **DvH transcriptome data**[http://genomics.lbl.gov/supplemental/DvHtranscripts2011/].

45. Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM: **Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D**. *Nucleic Acids Res.* 2007, :e128.

46. Kent WJ: **BLAT–the BLAST-like alignment tool**. *Genome Res.* 2002, **12**:656–64.

47. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput.* 2001, :127–38.

48. Bailey TL, Elkan C: **Unsupervised Learning of Multiple Motifs in Biopolymers using Expectation Maximization.** *Machine Learning* 1995, **21**:51–80.

49. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**:563–77.

50. **MicrobesOnline Source Code**[http://www.microbesonline.org/programmers.html#source].

51. Redding AM, Mukhopadhyay A, Joyner DC, Hazen TC, Keasling JD: **Study of nitrate stress in Desulfovibrio vulgaris Hildenborough using iTRAQ proteomics**. *Brief Funct Genomic Proteomic.* 2006, **5**:133–43.

52. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA: **The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra.** *Mol Cell Proteomics* 2007, **6**:1638–1655.

53. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**:944–5.

54. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, et al.: **CDD: a curated Entrez database of conserved domain alignments.** *Nucleic Acids Res.* 2003, **31**:383–7.

55. Gish W, States D: **Identification of protein coding regions by database similarity search.** *Nature Genet.* 1993, **3**:266–72.

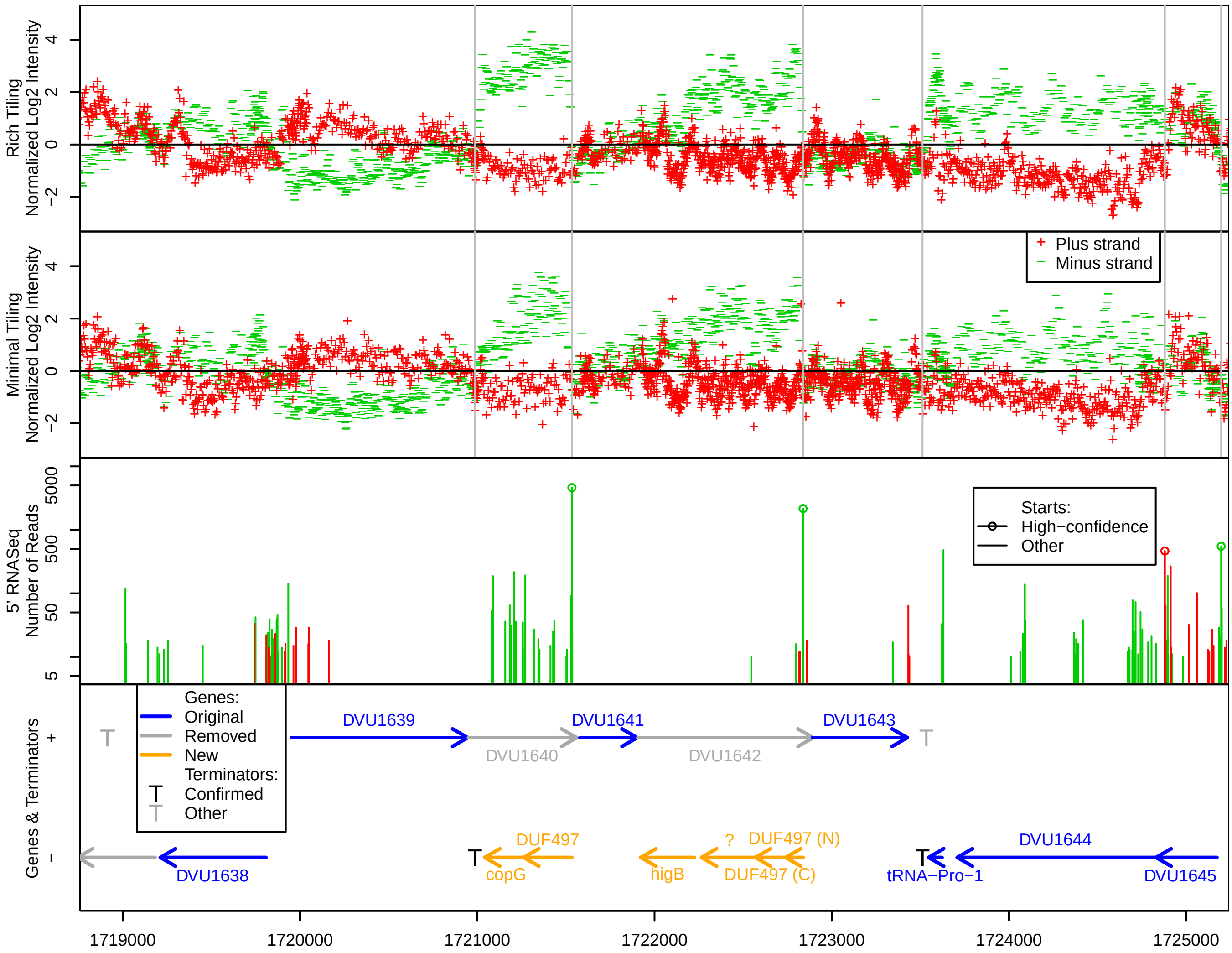
56. **PFam**[http://pfam.janelia.org/].

57. **R**[http://r-project.org/].

58. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: A sequence logo generator**. *Genome Research* 2004, **14**:1188–1190.

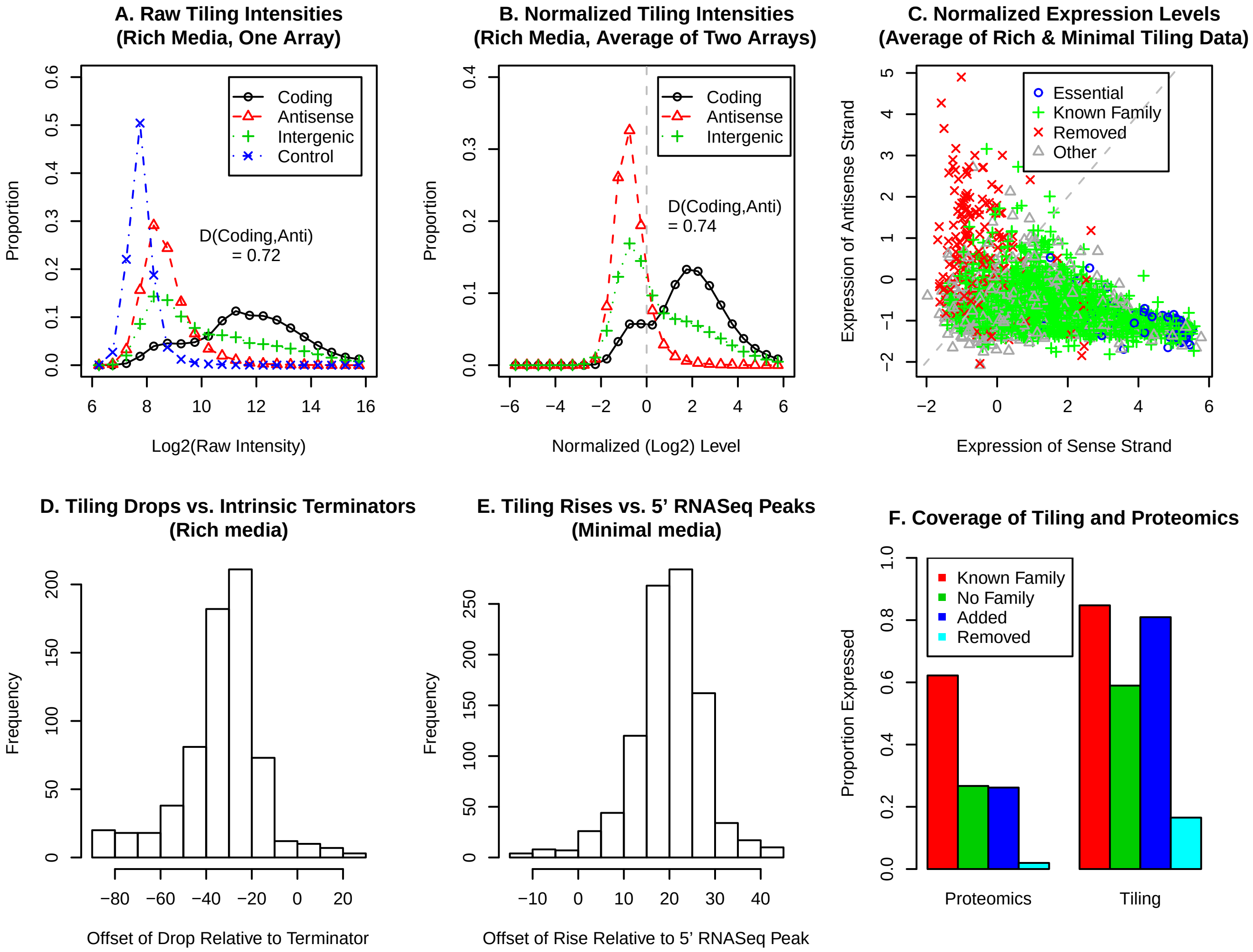
Figures

Figure 1 - Data for a region of the genome.



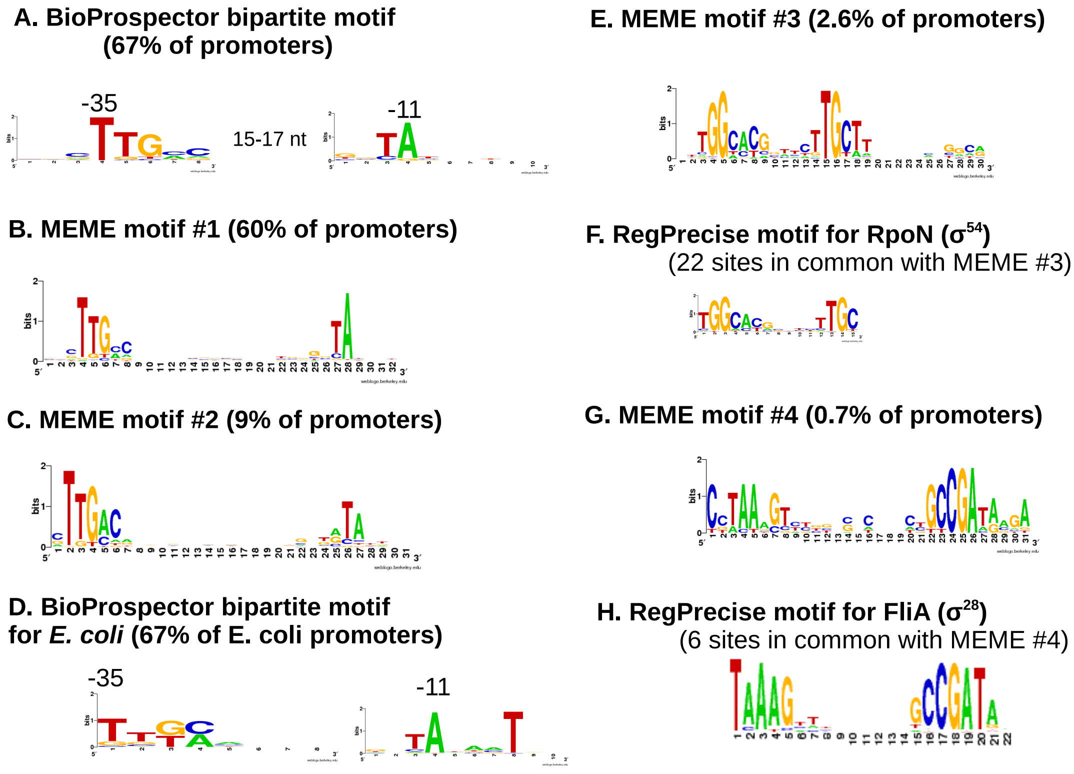
We show the tiling and 5' RNA-Seq data for 1719 to 1725 kilobases on the main chromosome, along with gene annotations, transcript starts, and terminators. The top two panels show normalized $\log_2$-levels from tiling data, with each probe plotted at its center. The genome-wide median value of 0 is shown as a horizontal black line, and vertical grey lines highlight the locations of key features from other panels, namely high-confidence starts and confirmed terminators. The third panel shows the number of reads starting at each location across two 5' RNA-Seq libraries from minimal media; note the log y axis. The bottom panel shows annotated genes (arrows) and predicted intrinsic terminators [20]. For newly annotated genes we show which gene family they belong too, if any (DUF is short for domain of unknown function). Two of the newly annotated DUF497 genes have leaderless promoters; the start of the transcripts for DVU1638 and for DVU1639 is ambiguous; DVU1645's transcript starts 24 nucleotides upstream of its start codon; and there is an antisense transcript for DVU1645 (an arsR-like regulator). The tiling data confirms the terminators for DUF497-copG and for tRNA-Pro-1.

Figure 2 - Quality and coverage of data.



(A) The distribution of raw $\log_2$ intensities, as a function of probe type, for a single array hybridized to cDNA from rich LS4 media. Probes were classified as coding, antisense, or intergenic using the original genome annotation; control probes have random sequences that do not match the DvH genome but have about the same GC content (63%). (B) The distribution of normalized $\log_2$ intensity for rich media (average of two replicates). The median value for the probes (excluding the random control probes) is zero and is shown with a vertical line. (C) The median normalized expression level for the sense and antisense strands of each protein-coding gene from the original annotation. The dashed line shows $x = y$. (D) The distribution of offsets between drops in the tiling data and the end of the intrinsic terminator's stem-loop. (E) The distribution of offsets between rises in the tiling data and peaks in 5' RNA-Seq. (F) The proportion of different types of protein-coding genes that were detected by tiling or by shotgun proteomics. Genes were considered detected by tiling if their smoothed intensity was above 0 throughout. Genes with a single high-confidence peptide were considered detected by proteomics.

Figure 3 - Promoter motifs.



A. BioProspector bipartite motif (67% of promoters)

B. MEME motif #1 (60% of promoters)

C. MEME motif #2 (9% of promoters)

D. BioProspector bipartite motif for *E. coli* (67% of E. coli promoters)

E. MEME motif #3 (2.6% of promoters)

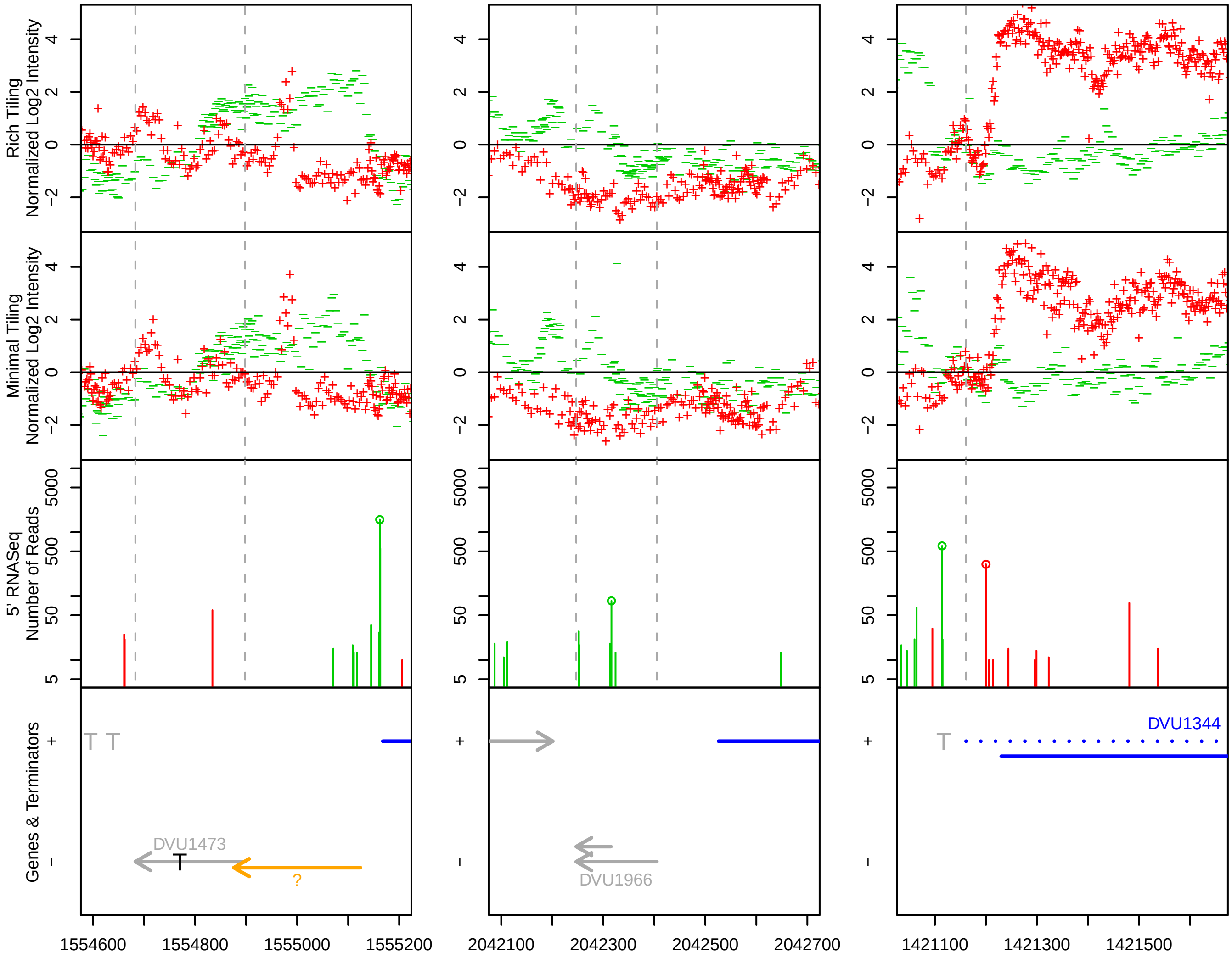
F. RegPrecise motif for RpoN (σ^{54}) (22 sites in common with MEME #3)

G. MEME motif #4 (0.7% of promoters)

H. RegPrecise motif for FliA (σ^{28}) (6 sites in common with MEME #4)

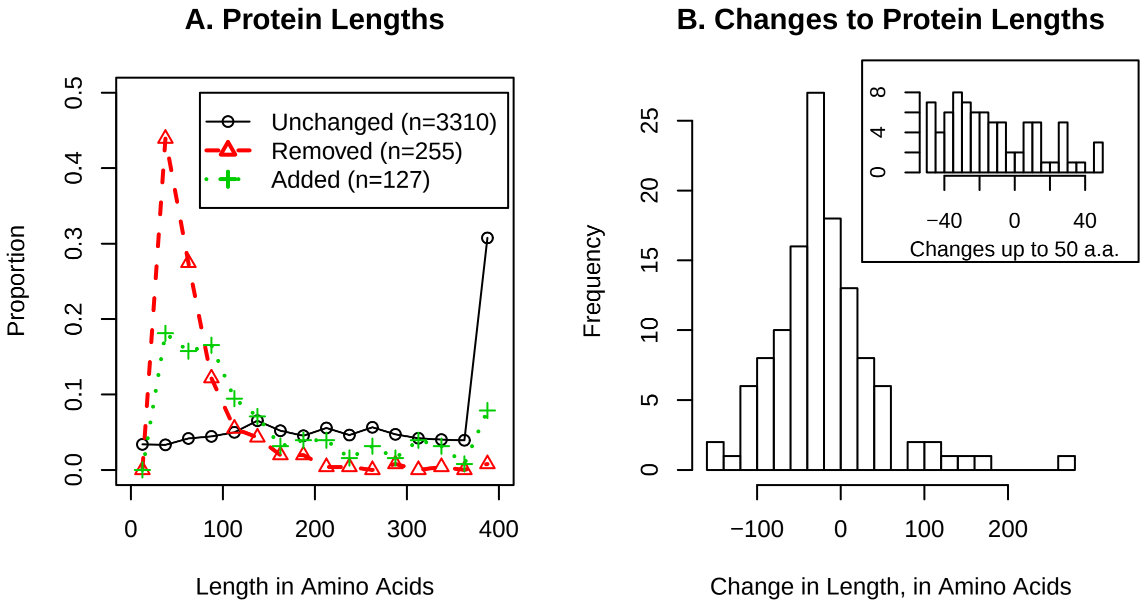
We show motifs from analyzing the -40 to +1 regions of 1,618 moderate-confidence DvH transcription starts using (A) BioProspector [47] and (B,C,E,G) MEME [48]. For comparison, we also show a motif (D) from analyzing 370 known promoters in *E. coli* K12 [27] with BioProspector and motifs from RegPrecise [29] (F,H) for alternate *Desulfovibrio* sigma factors that were inferred by comparative genomics. Each motif is shown as a sequence logo: at each position, the height of a nucleotide is proportional to its information content in bits [58].

Figure 4 - Examples of modified protein annotations.



We show the data and modifications to the annotation for three regions of the genome. Vertical lines show the extents of the original gene annotations; the other plotting symbols are as in Figure 1. The left panel shows that DVU1473 contains a terminator, while an ORF in another reading frame is expressed start-to-stop. This ORF does not belong to a known family but is homologous to other proteins, so it replaced DVU1473 in our annotation. The middle panel shows that only the C-terminal part of DVU1966 is transcribed; the upstream-most start codon that is consistent with the data is shown but would reduce the ORF to just 22 amino acids, so we removed it from our annotation. The right panel shows that DVU1344, as originally annotated (in dotted line), begins upstream of its promoter; we selected a new start codon downstream of the promoter.

Figure 5 - Lengths of proteins.



(A) The distribution of lengths of unchanged, removed proteins, and added proteins. Values above 400 are show in the rightmost bin. (B) The distribution of changes in length for the 123 proteins whose start codons were modified.

Figure 6 - Lengths of 5' and 3' untranslated regions.

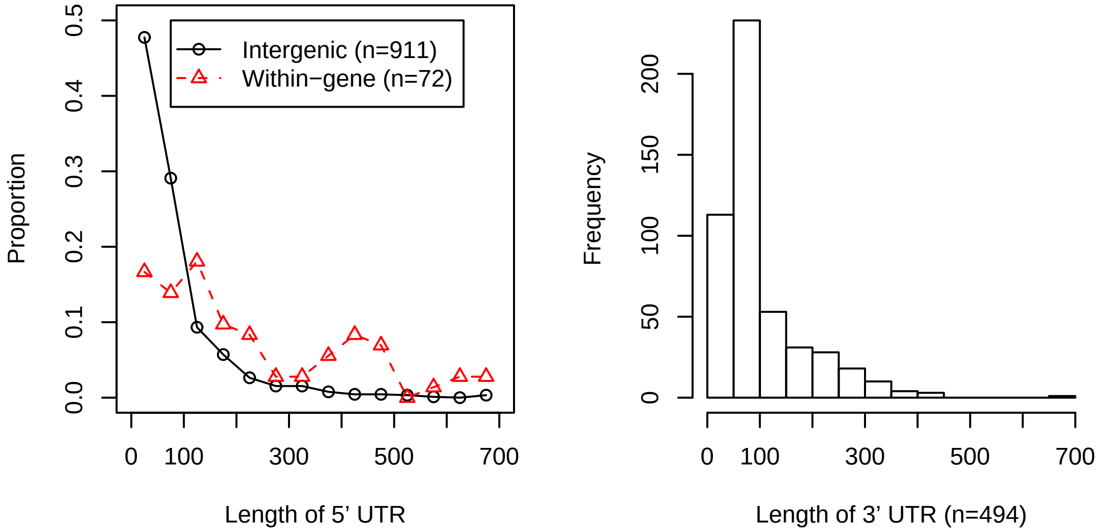
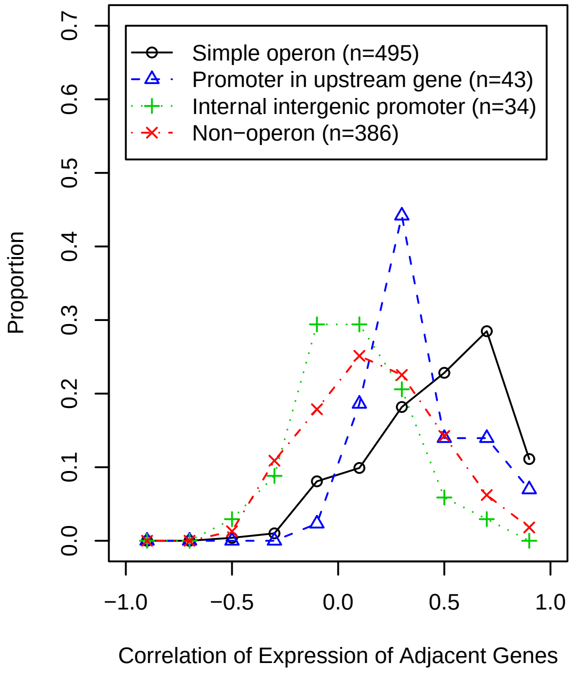
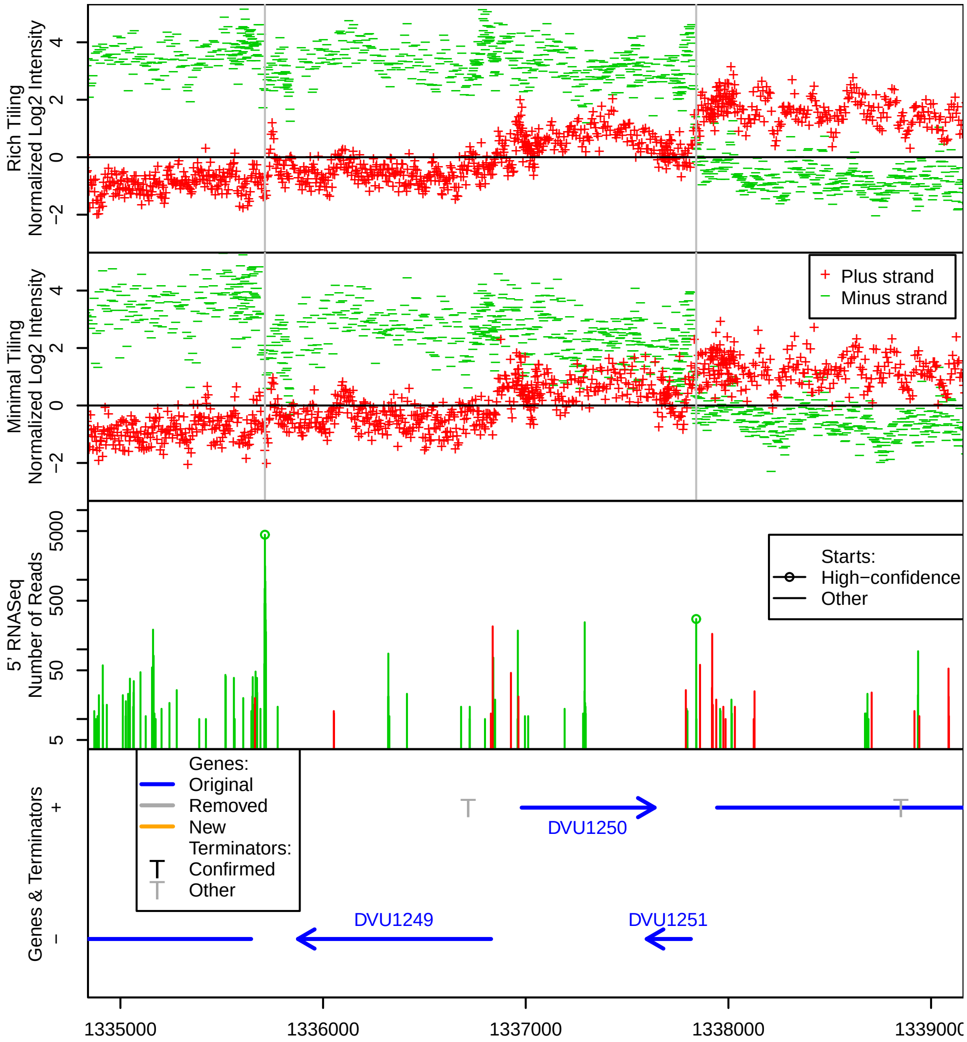


Figure 7 - Promoters within genes can confound gene expression measurements.



We show the distribution of co-expression for different types of pairs of adjacent genes on the same strand. The correlation of expression log-ratios was computed from 785 comparisons in MicrobesOnline [11].

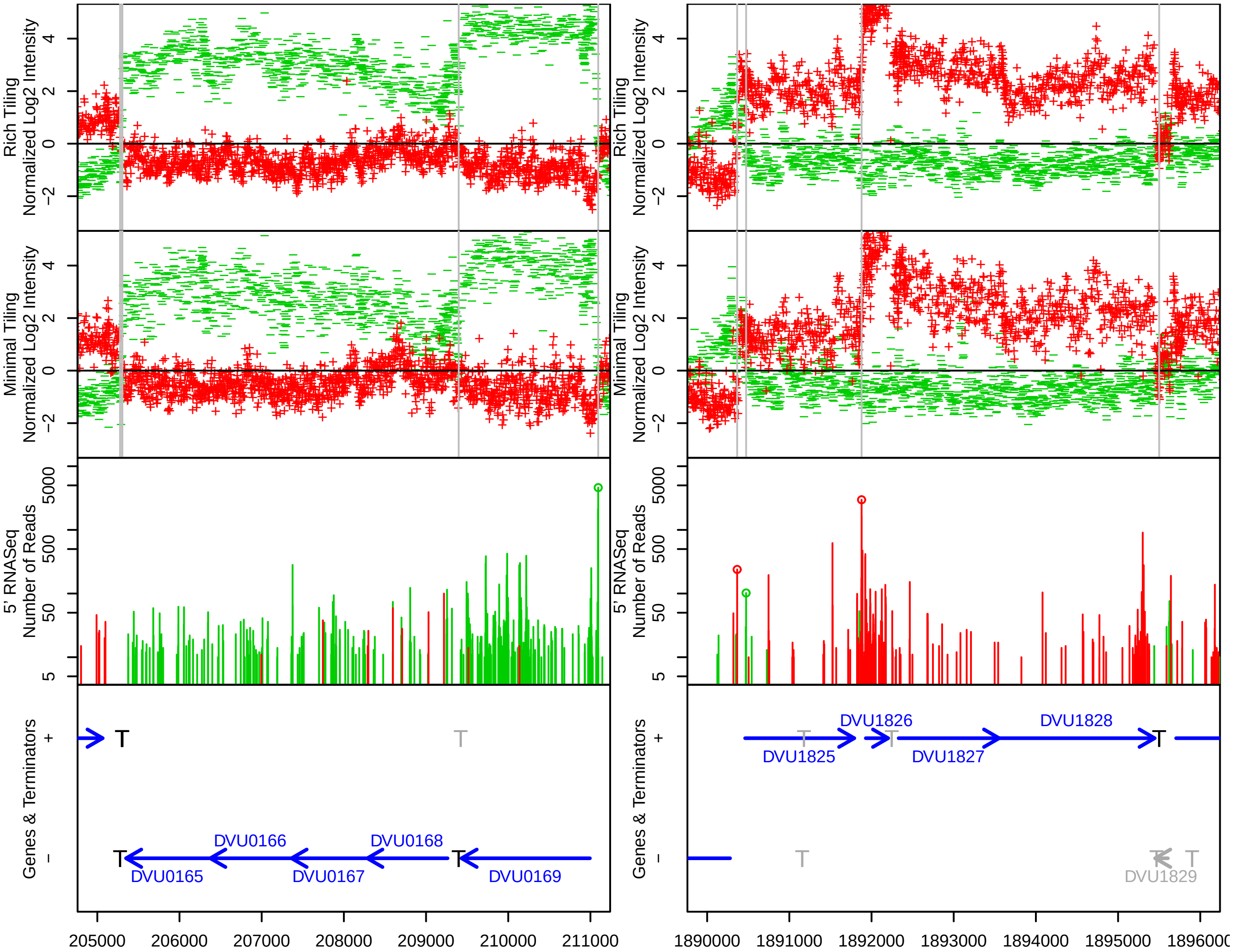
Supplementary Figure 1 - The DVU1251-DVU1249 operon containing the antisense strand of DVU1250 (gidB).



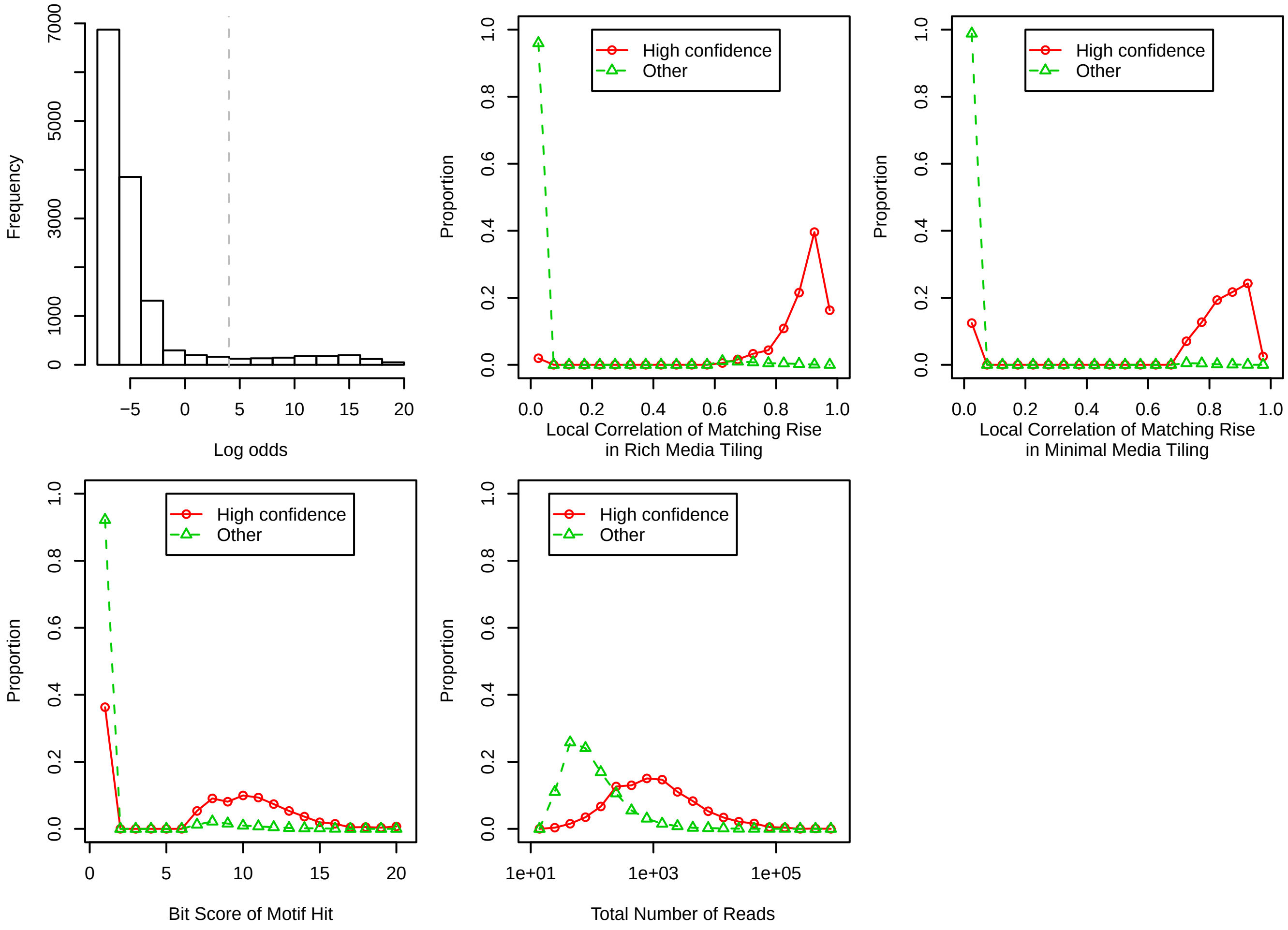
We are fairly confident that all three proteins are genuine: DVU1251 and DVU1249 were detected by proteomics, and DVU1250 belongs to a well-known family (COG0357; PF02527) and shows moderate expression as RNA.

Supplementary Figure 2 - Examples of complex operons

Plotting symbols are as in Figure 1; the vertical lines highlight predicted transcript starts and ends. On the left, we show a predicted attenuator between DVU0169 and DVU0168. On the right, we show an internal promoter just upstream of DVU1826. There may also be an attenuator downstream of DVU1826 but this was not selected by our automated approach.


Rich Tiling
Normalized Log2 Intensity
Minimal Tiling
Normalized Log2 Intensity
5' RNASeq
Number of Reads
Genes & Terminators
−2
0
2
4
5
50
500
5000
+
−
T
DVU0165
DVU0166
DVU0167
DVU0168
DVU0169
DVU1825
DVU1826
DVU1827
DVU1828
DVU1829
205000
206000
207000
208000
209000
210000
211000
1890000
1891000
1892000
1893000
1894000
1895000
1896000

Supplementary Figure 3 - Classifying local peaks in 5' RNA-Seq data as promoters or not.



We show the distribution of log odds for all 13,822 potential transcription starts, and we compare distribution of each feature for the 1,124 starts that have log odds above four and the 12,698 starts that have log odds below zero. If a potential start does not correspond to a rise in the tiling data or does not match any of the promoter motifs then the corresponding value is shown as a zero.

Additional Files

Additional file 1 – Intrinsic terminators

terminators.tab (tab-delimited file)

- scaffoldId, start, stop, strand – the terminators predicted by TransTermHP.
- stop – the end of the stem-loop, a few nucleotides upstream of the expected termination location.
- scaffoldId – 1944 for the main chromosome; 1945 for the megaplasmid.
- confirm – TRUE is this terminator was confirmed by our tiling data.

Additional file 2 – Classification of transcript starts

transcript_starts.tab (tab-delimited file)

- scaffoldId, start, and strand – the potential promoter location.
- nf, nppp, and ntot – the number of reads at that location in the 1st library, the 2nd (exonuclease) library, and the total.
- midrich – the middle of the corresponding rise in the tiling data from rich media (if any).
- atrich – the rise, corrected by 15 nt for the typical offset between the rise and the transcript start (if any).
- rich – the local correlation of the rise (if any).
- midmin, atmin, min – similarly for minimal media.
- startm – the start of the corresponding hit to a promoter motif (if any)
- motif – which motif (1 and 2 for sigma 70 with different spacings; 3 for rpoN; 4 for fliA).
- bits – the bit score of the motif hit.
- The various logodds* values gives the log odds values for each individual feature.
- lo – the total log odds; lo $\geq$ 4 means high confidence.

Additional file 3 – Peptides detected in ENIGMA experiments with fractionation or pull-downs

peptides.tab (tab-delimited file)

- peptide – the sequence of the peptide detected.
- nFractions – the number of different fractions or experiments that this peptide was detected in.

Additional file 4 – Peptides detected in ENIGMA complete-proteome experiments

peptides2.tab (tab-delimited file)

- peptide – the sequence of the peptide detected.
- nFractions – the number of different fractions or experiments that this peptide was detected in.

Additional file 5 – Peptide spectra from protein fractions that were examined by hand

MS_Spectra_Details.xls (Excel format)

Additional file 6 – Gene annotations and revisions and lengths of the 5' and 3' UTRs

genes.tab (tab-delimited file)

- sysName – also known as locus tag; identifiers beginning with DVU or DVUA are from the original annotation.
- locusId – the MicrobesOnline or VIMSS id (if from the original annotation).
- scaffoldId, strand, start, stop – the location of the gene.
- type – type=1 means protein-coding gene; type=7 means pseudogene derived from a protein-coding gene; types 9 and 10 are CRISPR repeats and spacers; other types are various kinds of non-coding RNAs.

- desc – description of the gene.
- removed – if present, the reason the gene was removed.
- changed – if present, the reason the gene was changed.
- start.orig – the original start.
- start.critica – the gene start from CRITICA, if this frame was selected by CRITICA.
- start.rast – the gene start from RAST, if this frame was selected by RAST.
- critId and rastId – ids for the CRITICA and RAST predictions; these are used in some of the Artemis feature files.
- UTR5 – length of the 5' UTR for this gene, if there is a confident transcript start upstream; 0 indicates a leaderless transcript.
- UTR3 – length of the 3' UTR for this gene, if there is a confirmed terminator downstream.

Additional file 7 – Operon predictions

operons.tab (tab-delimited file)

- scaffoldId, strand – which scaffold and strand the potential operon pair is on.
- upg and dng – the locusIds for the upstream and downstream genes in the pair.
- start.up, stop.up, start.dn, stop.dn – start and stop for the upstream and downstream genes.
- min5 and rich5 – smoothed minimum expression between the genes from a tiling array (or missing if it cannot be computed because there is little space between the genes).
- ttConfirm – non-zero if there is a confirmed terminator between the genes.
- rich.c.up – median expression of the coding strand of the upstream gene in rich media.
- rich.n.up – median expression of the non-coding (antisense) strand of the upstream gene in rich media.
- min.*.up – similarly for minimal media.
- *.dn – similarly for the downstream gene.
- start and lo – the location and log odds score of the most confident promoter between the two genes (if any).
- code – classification of the operon pair.
- ExprSim – Pearson correlation of gene expression patterns of the two genes.