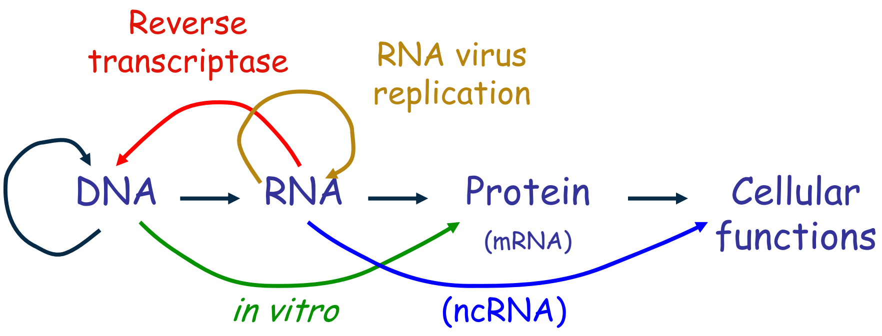
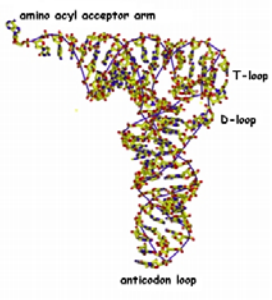
Current Trends:
Non-coding RNAs

“Central Dogma” of molecular biology



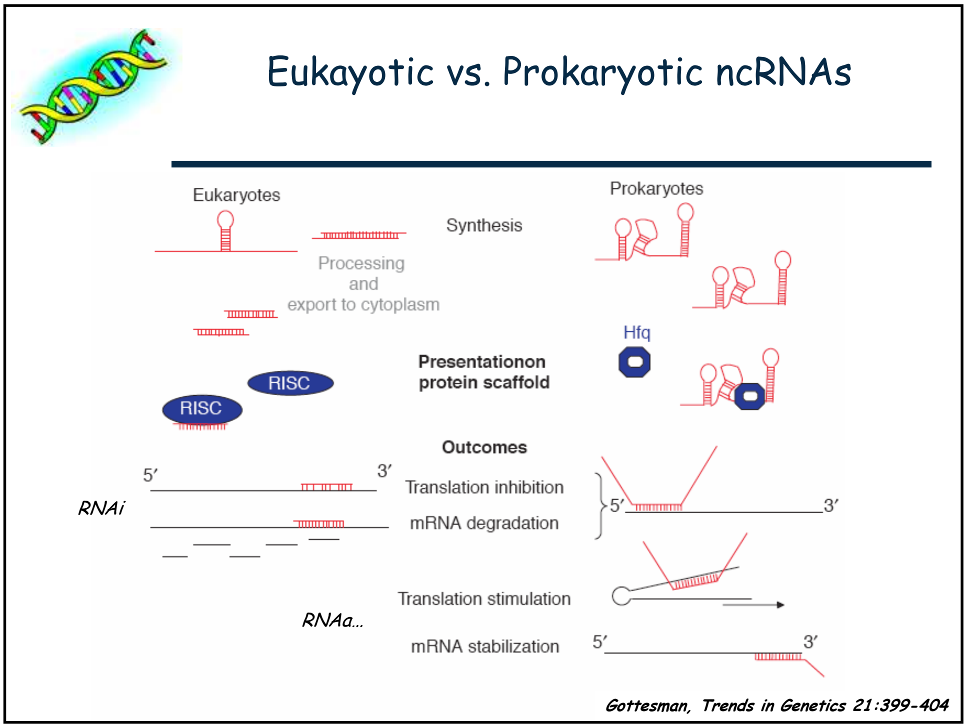
Non-coding RNAs

- Found in prokaryotes (small RNAs) and eukaryotes (non-coding RNAs).
- Well-characterized examples: tRNA, rRNA



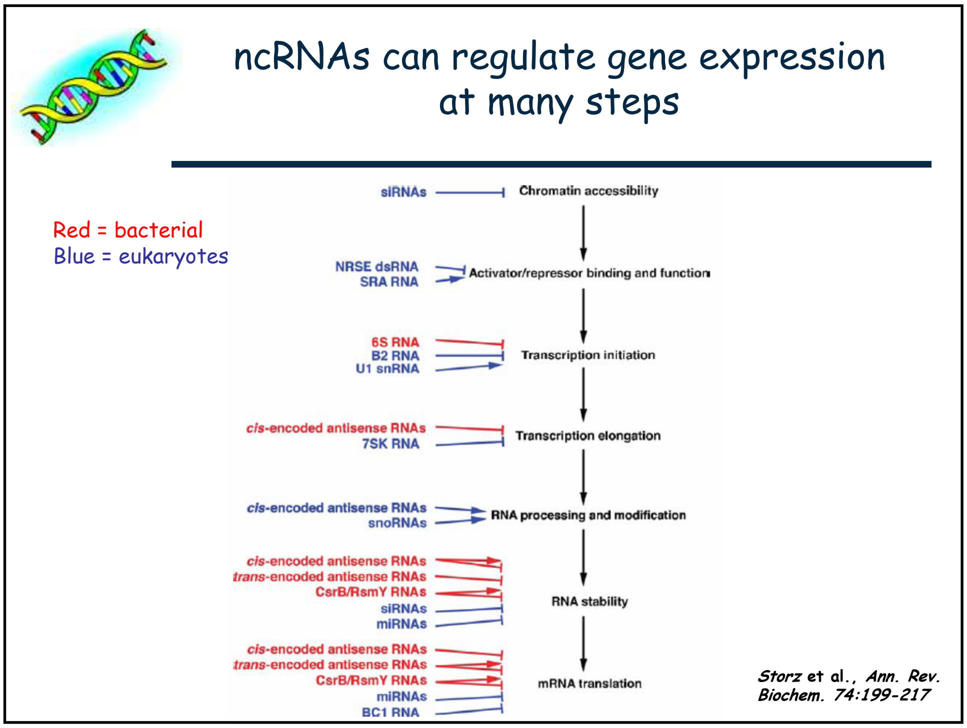
Non-coding RNAs

- Enzymatic activity
  - self-splicing introns
  - peptidyl transfer
  - viral replication
- Regulation of other genes
  - eukaryotes: 21-25 nts; micro RNAs
  - prokaryotes: 50-550 nts; small RNAs



Eukayotic vs. Prokaryotic ncRNAs

Gottesman, Trends in Genetics 21:399-404



ncRNAs can regulate gene expression at many steps

Red = bacterial
Blue = eukaryotes

siRNAs → Chromatin accessibility

NRSE dsRNA
SRA RNA
→ Activator/repressor binding and function

6S RNA
B2 RNA
U1 snRNA
→ Transcription initiation

cis-encoded antisense RNAs
7SK RNA
→ Transcription elongation

cis-encoded antisense RNAs
snoRNAs
→ RNA processing and modification

cis-encoded antisense RNAs
trans-encoded antisense RNAs
CsrB/RsmY RNAs
siRNAs
miRNAs
→ RNA stability

cis-encoded antisense RNAs
trans-encoded antisense RNAs
CsrB/RsmY RNAs
miRNAs
BC1 RNA
→ mRNA translation

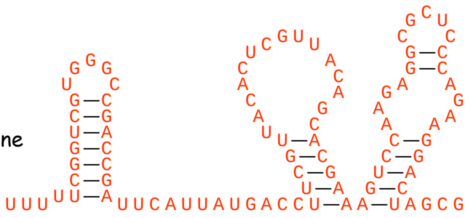
Storz et al., *Ann. Rev. Biochem. 74:199-217*

Targets of RNA Gene Regulation

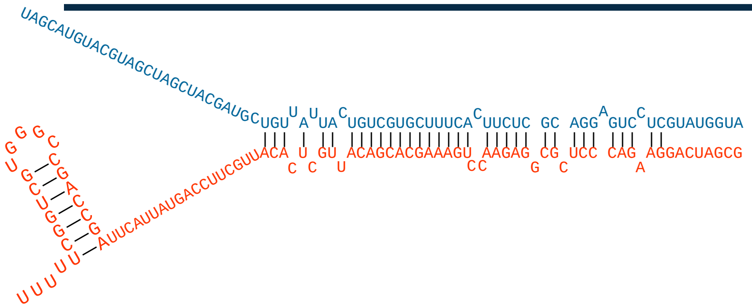
messenger RNA

UAGCAUGUACGUAGCUAGCUACGAUUGUUAUUACUGUCGUGCUUUCACUUCUCGCAGGAGUCCUCGUAUGGUA

RNA gene

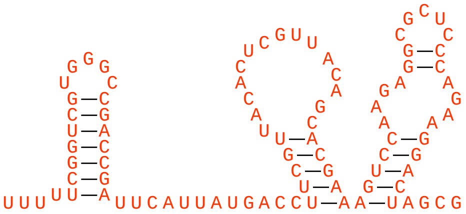


Targets of RNA Gene Regulation



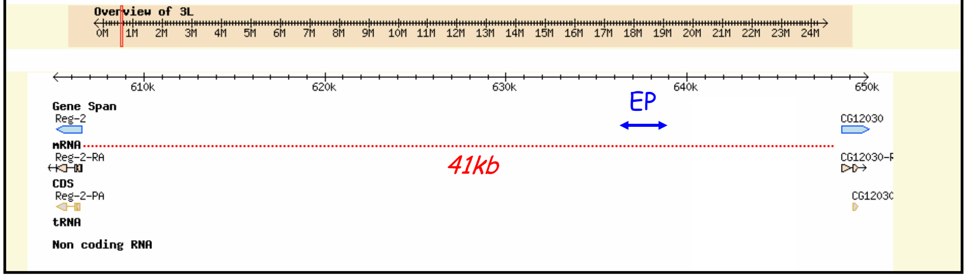
Non-coding RNAs are elusive

- Not annotated in genomes: lack of defined sequence features
- Small, often missed in genetic studies
- Missed in assays for protein function
- None of 70-100 *E. coli* ncRNAs found by mutation

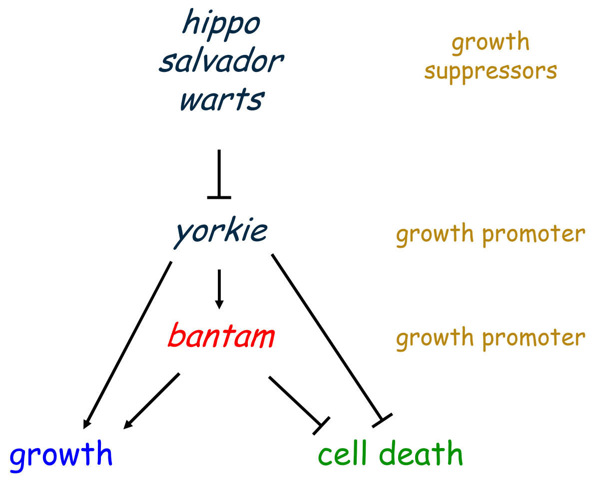


Drosophila bantam gene discovery

- Overexpression of an intergenic region causes cell and tissue overgrowth
- Deletion of intergenic region surrounding EP element results in slow growth



bantam encodes a miRNA that regulates a conserved growth pathway



hippo
salvador
warts

growth suppressors

yorkie

growth promoter

bantam

growth promoter

growth

cell death

Why study ncRNAs in bacteria?

- We live in a bacterial world
- Bacteria serve as useful model organisms
- Bacteria are diverse
- Understanding bacteria is useful in many important applications

Food Poisoning
Strep Throat
Tuberculosis
Diptheria
Dysentery
The Black Plague
Yaws
Whooping cough
Typhoid Fever
Lyme Disease
Botulism
Pneumonia
Gastroenteritis
Meningitis
Scarlet Fever
Syphilis
Gonorrhea
Dental Cavities
Cholera
Peptic Ulcers
Anthrax
Rheumatic Fever
Rocky Mountain Spotted Fever
Tetanus
Leprosy

Shewanella oneidensis

- Gram-negative γ-proteobacterium
- Found primarily in deep water anaerobic habitats
- Can use a wide variety of compounds as terminal electron acceptors
- Bioremediation potential: reduces soluble chromium and uranium to insoluble forms

---

S. oneidensis genome overview

45.9% G-C content; 85.5% of genome is coding

5131416 bp total: chromosome is 4969803 bp; pMR-1 is 161613 bp

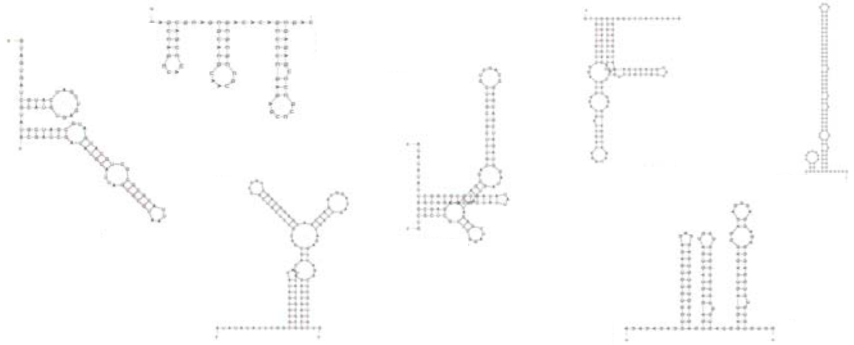
Total genes	**5066**	
tRNA and rRNA genes	**128**	**(2.5%)**
Protein-coding genes	**4938**	**(97.5%)**
Genes assigned function	**2915**	**(59%)**
Conserved hypothetical genes	**864**	**(17.5%)**
Hypothetical genes	**1159**	**(27%)**

Computational ncRNA prediction

- Most ncRNAs are intergenic, function *in trans*
- Bacterial ncRNAs have promoters, terminators
- Genes have distinct nucleotide composition
- Conserved secondary structures (stem-loop)
- Tiling microarray data
- *Data can be integrated into a set of predictions*

ncRNA Genes

```
GTCAGTATAGTCGCATTATAGCCGATCTGAGTCAGTCAGTCGTAGTATCGTAGTCAGTCGTACGTAGTCAGTCGTATCAGTCGAGTCAGTCGA
GCTAGTCGATCGTATCACTATCATCGTACGTAGTGCTAGTCAGTGTCATCGATGCGTACGTAGTCAGTTACGTAGCATCGTACGTAGTCATGC
ATGCTAGCTAGCTAGCTAGCTAGCTACGCGATCGTGCGTATGCGTATATTATATGCGCTAGCAGTCGTAGTACGTAGTACTATGTATGCGTAC
GTGATGCTAGTTGCGTACGATAGCGATACGATCAGTCGTATCGATCGTATGCATCGAGAGTCGTAGTAGCGATTAGCGCTAGTCATTATAGTC
GTACTTAGGTCGCGGCGATTACGGATAGTCTGATCACGACGTATGAGCTGACGCGGCGATCAGGAAGACCCTCGCGGAGAACCTGAAAGCACG
ACATTGCTCACATTGCTTCCAGTATTACTTAGCCAGCCGGGTGCTGGCTTTTTGTACGTACTGAGTCGGCATTATAGCGTATGCATACGGAGT
ACGAGTCGTACGGACAGTCGTAGTCAGTCTGATCAGTCAGTCGTAGTCGTATGCAGTCGACGAGTCGTACGTATGCAGTCGATCGTTAGGACT
CGTAAGTCGTATCATATCGGATTATAGCATGCTAGAGCTAGTCGTATAGTCTACGAGTTATACGTCTAGTGGCTAGTGTACGTCAGTCGTACG
ATGCAGTTAGTAGTCTAGTATTACGATTAGTCGTGATCTGAGTAGTTACGTCGATGGTAGCCATTATACGTACTTAC
```



1) Nucleotide composition

GTCAGTATAGTCGCATTATAGCCGATCTGAGTCAGTCAGTCGTAGTATCGTAGTCAGTCGTACGTAGTCAGTCGTATCAGTCGAGTCAGTCGA
GCTAGTCGATCGTATCACTATCATCGTACGTAGTGCTAGTCAGTGTCATCGATGCGTACGTAGTCAGTTACGTAGCATCGTACGTAGTCATGC
ATGCTAGCTAGCTAGCTAGCTAGCTACGCGATCGTGCGTATGCGTATATTATATGCGCTAGCAGTCGTAGTACGTAGTACTATGTATGCGTAC
GTGATGCTAGTTGCGTACGATAGCGATACGATCAGTCGTATCGATCGTATGCATCGAGAGTCGTAGTAGCGATTAGCGCTAGTCATTATAGTC
GTACTTAGGTCGCGGCGATTACGGATAGTCTGATCACGACGTATGAGCTGACGCGGCGATCAGGAAGACCCTCGCGGAGAACCTGAAAGCACG
ACATTGCTCACATTGCTTCCAGTATTACTTAGCCAGCCGGGTGCTGGCTTTTTGTACGTACTGAGTCGGCATTATAGCGTATGCATACGGAGT
ACGAGTCGTACGGACAGTCGTAGTCAGTCTGATCAGTCAGTCGTAGTCGTATGCAGTCGACGAGTCGTACGTATGCAGTCGATCGTTAGGACT
CGTAAGTCGTATCATATCGGATTATAGCATGCTAGAGCTAGTCGTATAGTCTACGAGTTATACGTCTAGTGGCTAGTGTACGTCAGTCGTACG
ATGCAGTTAGTAGTCTAGTATTACGATTAGTCGTGATCTGAGTAGTTACGTCGATGGTAGCCATTATACGTACTTAC

Frequency of nucleotides

A	0.28
C	0.22
G	0.22
T	0.28

A	0.23
C	0.27
G	0.27
T	0.23

Frequency of dinucleotides

	A	C	G	T
A	0.0562	0.0481	0.0732	0.0464
C	0.0661	0.0720	0.0641	0.0545
G	0.0501	0.0631	0.0844	0.0784
T	0.0382	0.0723	0.0528	0.0801

2) Comparative Genomics: Mutation Patterns

Genome (1)

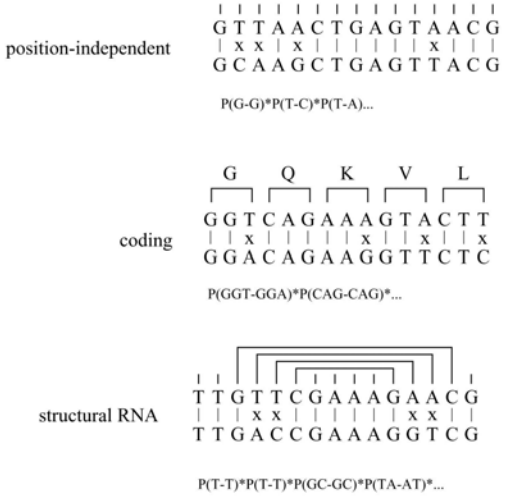
GTCAGTATAGTCGCATTATAGCCGATCTGAGTCAGTCAGTCGTAGTATCGTAGTCAGTCGTACGTAGTCAGTCGTATCAGTCGAGTCAGTCGA
GCTAGTCGATCGTATCACTATCATCGTACGTAGTGCTAGTCAGTGTCATCGATGCGTACGTAGTCAGTTACGTAGCATCGTACGTAGTCATGC
ATGCTAGCTAGCTAGCTAGCTAGCTACGCGATCGTGCGTATGCGTATATTATATGCGCTAGCAGTCGTAGTACGTAGTACTATGTATGCGTAC
GTGATGCTAGTTGCGTACGATAGCGATACGATCAGTCGTATCGATCGTATGCATCGAGAGTCGTAGTAGCGATTAGCGCTAGTCATTATAGTC
GTACTTAGGTCGCGGCGATTACGGATAGTCTGATCACGACGTATGAGCTGACGCG**GCGATCAGGAAGACCCTCGCGGAGAACCTGAAAGCACG**
ACATTGCTCACATTGCTTCCAGTATTACTTAGCCAGCCGGGTGCTGGCTTTTTGTACGTACTGAGTCGGCATTATAGCGTATGCATACGGAGT
ACGAGTCGTACGGACAGTCGTAGTCAGTCTGATCAGTCAGTCGTAGTCGTATGCAGTCGACGAGTCGTACGTATGCAGTCGATCGTTAGGACT
CGTAAGTCGTATCATATCGGATTATAGCATGCTAGAGCTAGTCGTATAGTCTACGAGTTATACGTCTAGTGGCTAGTGTACGTCAGTCGTACG
ATGCAGTTAGTAGTCTAGTATTACGATTAGTCGTGATCTGAGTAGTTACGTCGATGGTAGCCATTATACGTACTTAC

Genome (2)

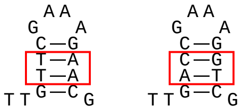
GATATCGGTACGTGTCTAGCATGAGTCTATCTATACTGTCGGCGTATCGTACGTATGCGTAATCGATCAGTGTCGTATCGAGTTACGATGCAT
GAGTCGTACGTATCGTAGCATGCTAGCTACGATGCTAGCATGCTAGCATCGATGCATGCATGCTGACTAGATCGTACGTAGCTACGTAGTCGT
AAGTCGTAGTCGTAGCTAGTTAGCGCGTATAGCGTACGTAGTACGTATCGATGCGTAGTCATTACGACTGATCGTAAGTCGA**GCGATCAGCAA**
GACCCACGAGGAGAACCTGAAGCACGACATTGCTCAATTGCTTCCAGATTACGTAGCCAGGGCCGGGTGCTGGTTTTTCAGTCGTACGTAGCT
AGTAGTCGTACTGAGCAGTCTAGCATCGTAGTCATGATTGCGTACGTATCGATCGAGTCGATGCATGTATATATGCCGCGTACTGACGTACGT
AGTCTAGCTAGTCATGCTATATACGGCGCTAGTCGTAGTACGTCGTAGTCAGTGTCAGTATCGAGTCATGCATGTCGTACGTATGGCATGGCT
AGTCATGGACTAGCTAGTAGCGTACGTAGTCATTATACGTACGTCGTATGATATATTAGCGCCGCGGTGTACTGCGTCGTGTCGTATACTACT
GATCTGATCGTAGTACTGCTACGTAGTCGTAGCAGTCGATCGTATGCATGCGTAGTCGTAGTCTAGCTGATCTACGTAGTCGTAGTATGCGTA
GTCTAGTCTATGCATTATATGCTATAGTCATGCTAGCATACGT

Shewanella amazonensis, Shewanella baltica, Shewanella denitrificans,
Shewanella frigidimarina, Shewanella loihica, Vibrio cholerae,
Yersinia pestis, Photorhabdus luminescens, Photobacterium profundum

2) Mutation Patterns that Conserve RNA Structure

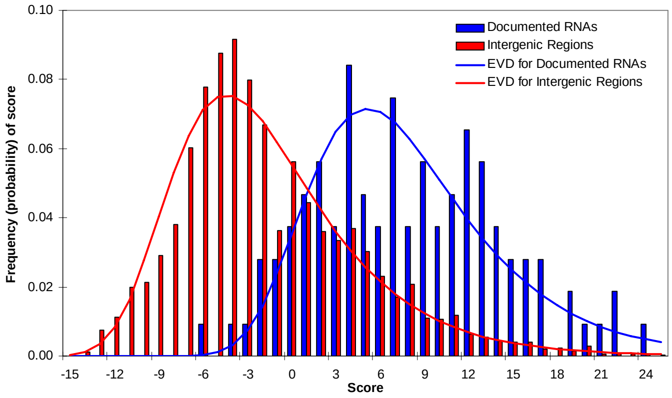


Derive score based on:
\# of compensatory mutations
Length of sequence
Sequence structure

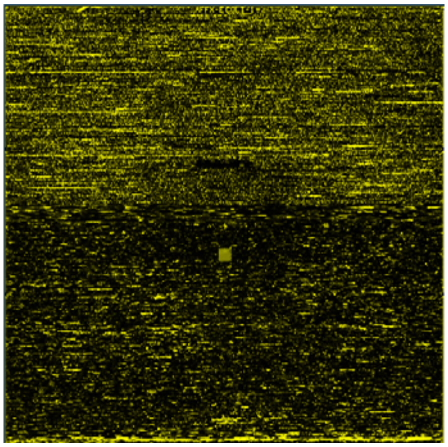


Rivas and Eddy, *BMC Bioinformatics* 2:8

2) Score of Conserved RNA Structure



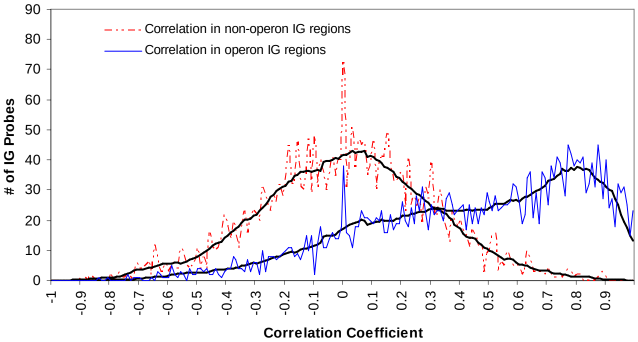
3) DNA Microarray Data

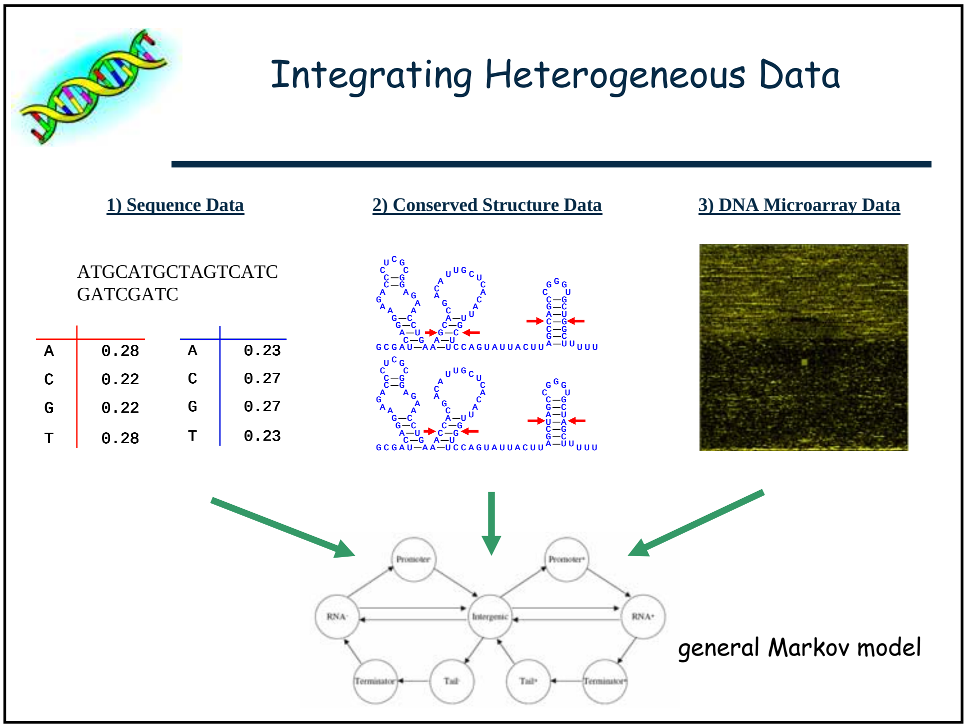


Examine correlation of expression for probes near one another in the genome:

1) intergenic regions likely to produce RNA
2) for those much less likely to produce RNA

3) Correlation of Transcript Expression





Integrating Heterogeneous Data

1) Sequence Data

ATGCATGCTAGTCATC
GATCGATC

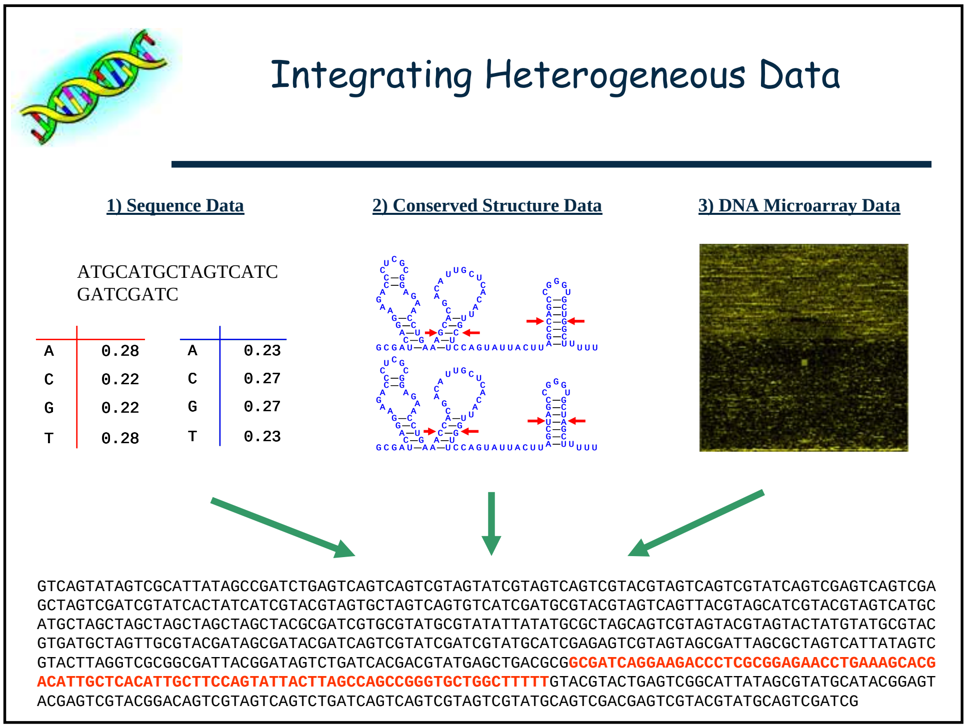
A	0.28
C	0.22
G	0.22
T	0.28

A	0.23
C	0.27
G	0.27
T	0.23

2) Conserved Structure Data

3) DNA Microarray Data

general Markov model



Integrating Heterogeneous Data

1) Sequence Data

ATGCATGCTAGTCATC
GATCGATC

A	0.28
C	0.22
G	0.22
T	0.28

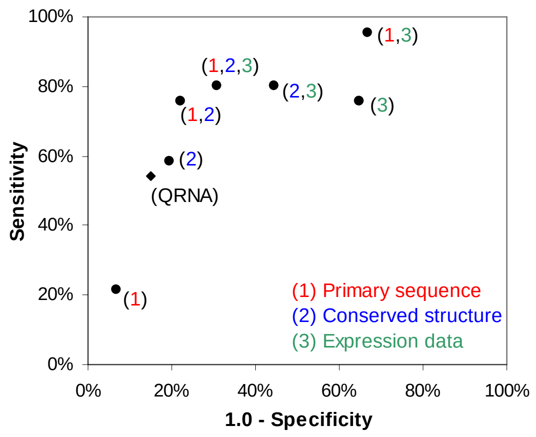
A	0.23
C	0.27
G	0.27
T	0.23

2) Conserved Structure Data

3) DNA Microarray Data

GTCAGTATAGTCGCATTATAGCCGATCTGAGTCAGTCAGTCGTAGTATCGTAGTCAGTCGTACGTAGTCAGTCGTATCAGTCGAGTCAGTCGA
GCTAGTCGATCGTATCACTATCATCGTACGTAGTGCTAGTCAGTGTCATCGATGCGTACGTAGTCAGTTACGTAGCATCGTACGTAGTCATGC
ATGCTAGCTAGCTAGCTAGCTAGCTACGCGATCGTGCGTATGCGTATATTATATGCGCTAGCAGTCGTAGTACGTAGTACTATGTATGCGTAC
GTGATGCTAGTTGCGTACGATAGCGATACGATCAGTCGTATCGATCGTATGCATCGAGAGTCGTAGTAGCGATTAGCGCTAGTCATTATAGTC
GTACTTAGGTCGCGGCGATTACGGATAGTCTGATCACGACGTATGAGCTGACGCG**GCGATCAGGAAGACCCTCGCGGAGAACCTGAAAGCACG**
ACATTGCTCACATTGCTTCCAGTATTACTTAGCCAGCCGGGTGCTGGCTTTTTGTACGTACTGAGTCGGCATTATAGCGTATGCATACGGAGT
ACGAGTCGTACGGACAGTCGTAGTCAGTCTGATCAGTCAGTCGTAGTCGTATGCAGTCGACGAGTCGTACGTATGCAGTCGATCG

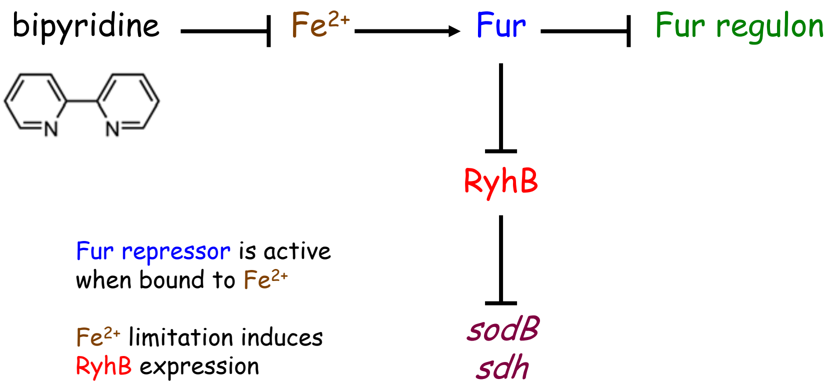
Performance on known *E. coli* ncRNAs



Predicting ncRNAs in *Shewanella*

- Have robust tiling microarray data set: 144 experiments, wide variety of growth conditions
- Generated predictions of ~160 ncRNAs
- Some may be orthologous to characterized ncRNAs
- Some may be novel ncRNAs
- Some may not be ncRNAs at all

Regulation of *RyhB*



bipyridine ⊣ Fe^{2+} → Fur ⊣ Fur regulon

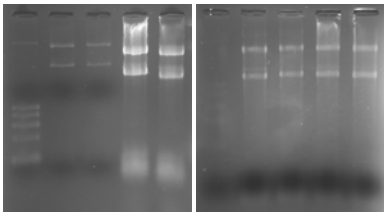
Fur ⊣ RyhB ⊣ *sodB* *sdh*

Fur repressor is active when bound to Fe^{2+}

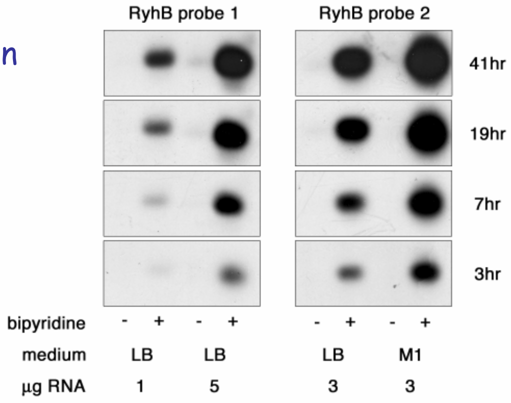
Fe^{2+} limitation induces RyhB expression

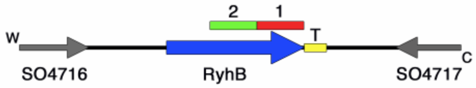
Validating *Shewanella* ncRNA predictions

- Putative *ryhB* northern blotting experiments



Shewanella ryhB ~100 nt





Regulation of *spot42* in *E. coli*

Glucose ⊣ CRP cAMP ⊣ *spot42* ⊣ *galK*

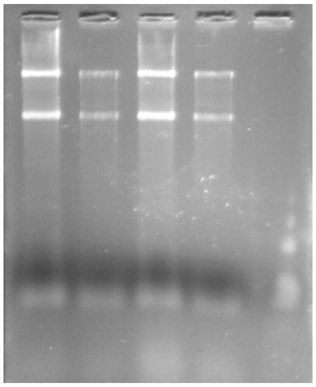
spot42 negatively regulates translation of *galK* but does not affect *galE* translation

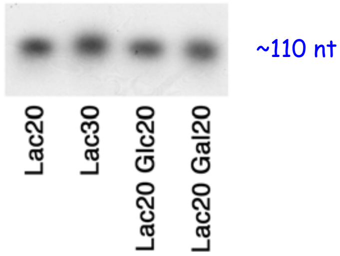
spot42 expression increases the GalE:GalK ratio

Thus, glucose *induces* *spot42* expression

Validating *Shewanella* ncRNA predictions

- Putative *spot42* northern blotting experiments





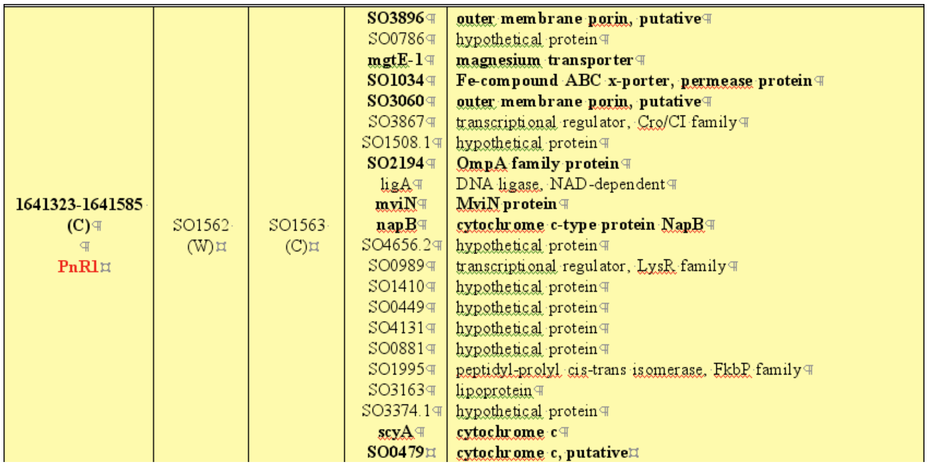
Blot probed with *spot42* probe 1

Predicting ncRNA targets

- Target prediction is an inexact science
- ncRNA sequences not exact matches to targets
- May be multiple targets

- Vaildate targets using northern blots, western blots, exogenous expression

Putative *Shewanella* ncRNA targets



1641323-1641585 (C) **PnR1**	SO1562 (W)	SO1563 (C)	**SO3896**	**outer membrane porin, putative**
			SO0786	hypothetical protein
			mgtE-1	**magnesium transporter**
			SO1034	**Fe-compound ABC x-porter, permease protein**
			SO3060	**outer membrane porin, putative**
			SO3867	transcriptional regulator, Cro/CI family
			SO1508.1	hypothetical protein
			SO2194	**OmpA family protein**
			ligA	DNA ligase, NAD-dependent
			mviN	**MviN protein**
			napB	**cytochrome c-type protein NapB**
			SO4656.2	hypothetical protein
			SO0989	transcriptional regulator, LysR family
			SO1410	hypothetical protein
			SO0449	hypothetical protein
			SO4131	hypothetical protein
			SO0881	hypothetical protein
			SO1995	peptidyl-prolyl cis-trans isomerase, FkbP family
			SO3163	lipoprotein
			SO3374.1	hypothetical protein
			scyA	**cytochrome c**
			SO0479	**cytochrome c, putative**

Questions

- Is the interaction between a ncRNA and its target RNA positive or negative?
- What conditions regulate ncRNA expression?
- What can we learn that will improve ncRNA prediction and understanding of function?