The primary transcriptome of *H. pylori*, a major human pathogen

Steve Hoffmann
steve@bioinf.uni-leipzig.de

February 20, 2010

The primary transcriptome of *H. pylori*, a major pain in the ass

Steve Hoffmann
steve@bioinf.uni-leipzig.de

February 20, 2010

Mugshot

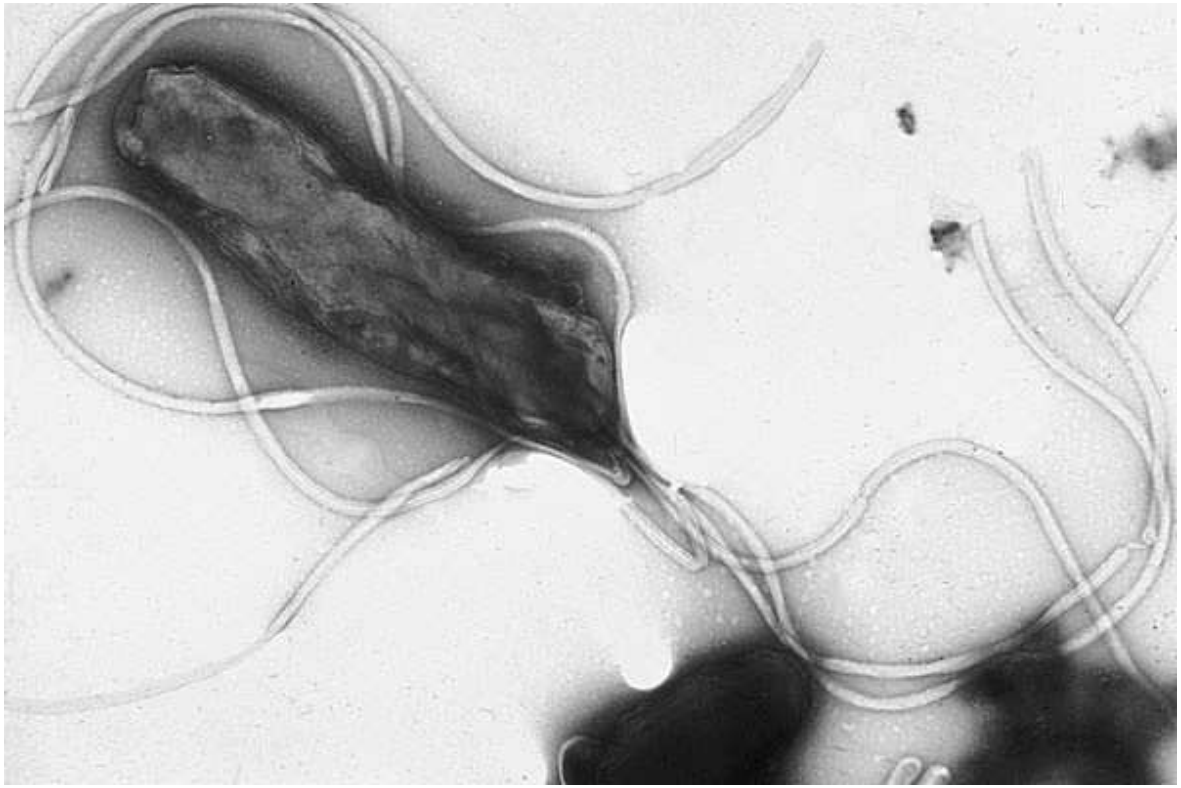


Figure: A real bad hat: *H. pylori* lives at pH 1 and he likes it!

H. pylori during the attack

Figure: A real bad hat: *H. pylori* tries hard to infect!

Wait a second! Did you say pH 1?

(Loading Chimera.mov)

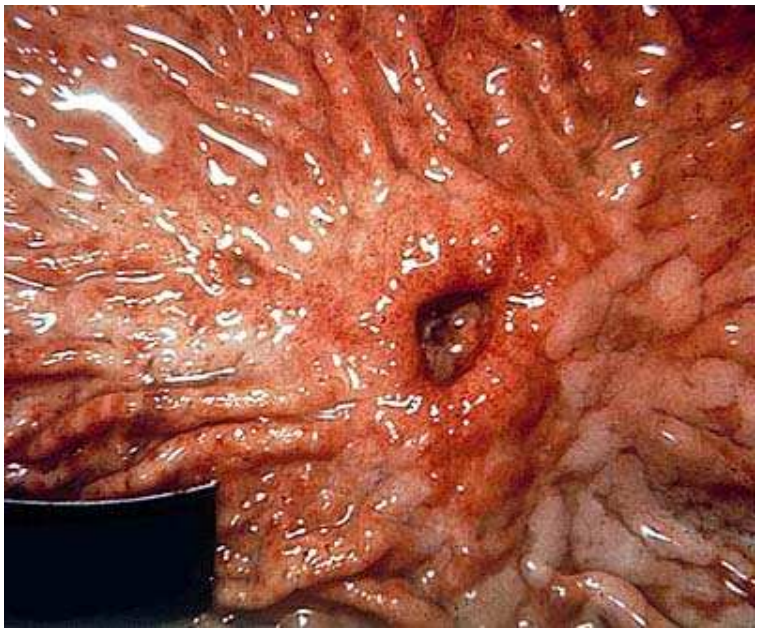
Chemistry

$$(NH_2)_2CO + H_2O \xrightarrow{Urease} CO_2 + NH_3$$

H. pylori: a ubiquitous agent

- seroprevalence up to 55.5%
- seroprevalence correlated to socioeconomic status
- major reason for gastritis (type b)
- major reason for the peptic ulcer
- correlated to the development of gastric cancer

Peptic ulcer



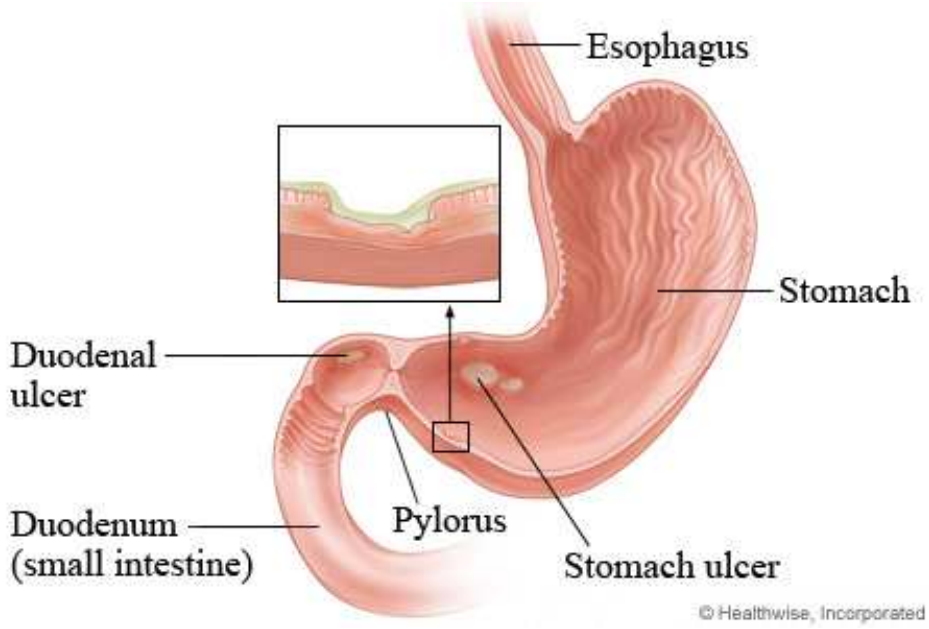


Figure: Lifetime-chance to develop a peptic ulcer: 10%!

As walls crumble ...

The gastric mucosa produces H^+ and maintains the acidic environment of the stomach lumen. Furthermore, it produces pepsin. It really needs to protect itself:

- gastric mucus layer of glyoprotein (a gel 5-200μm)
- slows down hydrogen diffusion
- binds luminal pepsin
- active Cl^-/HCO_3^- transport (bicarbonate!)

As walls crumble (2)

- cag pathogenicity island
- its cagA gene
- and vacA

help H.pylori to hook up gastric cells. cagA and vacA are involved in proliferation, cell vacuolation and cell death.

If the barrier is broken ... acid and pepsin and the host immune system will do the rest.

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

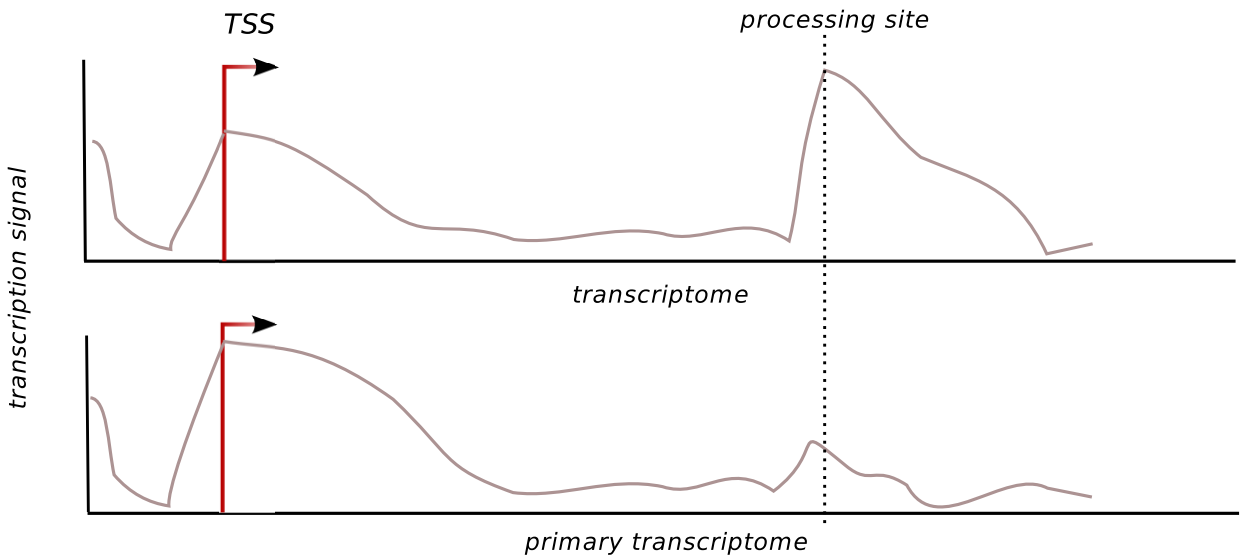
$>$ 600 genes are hypothetical !

HP26695 and its sports genome

genome size	1.6×10^6
GC content	38.87%
tRNA	36
rRNA	7
coding genes	1630
- hypothetical	470
- conserved hypothetical	185
- coding nucleotides	90.81%

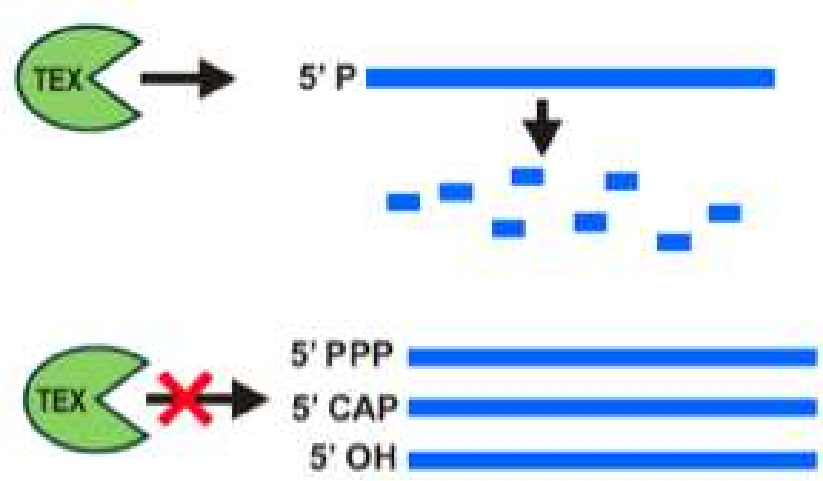
$>$ 600 genes are hypothetical !

A primary transcriptome?



- normal transcriptome contains processing products
- primary transcriptome to allow
- $\longrightarrow$ identification of transcription start sites (TSS)
- $\longrightarrow$ identification of processing sites
- $\longrightarrow$ transcriptional organization

How to get the primary transcriptome?



5-monophosphate dependent terminator exonuclease (TEX) specifically degrades RNAs with 5-monophosphates, while primary transcripts are not affected.

dRNAseq in *H. pylori*

Molecular biology

1. cell cultures for several environmental cond. (AS, ML, INF)
2. generation of TEX(-) and TEX(+) libraries
3. high-throughput sequencing (454 & Illumina)

to finally be able to

1. transcription start site
2. processing sites
3. small RNAs
4. check current gene annotation

Some (easy) bioinformatic questions

... arise in the mapping (short RNAs; mapping with poly-A tails), the normalization(!), the annotation of TSS (dRNAseq) and in the comparison with annotation (classification of TSS).

dRNAseq: when to assume a TSS?

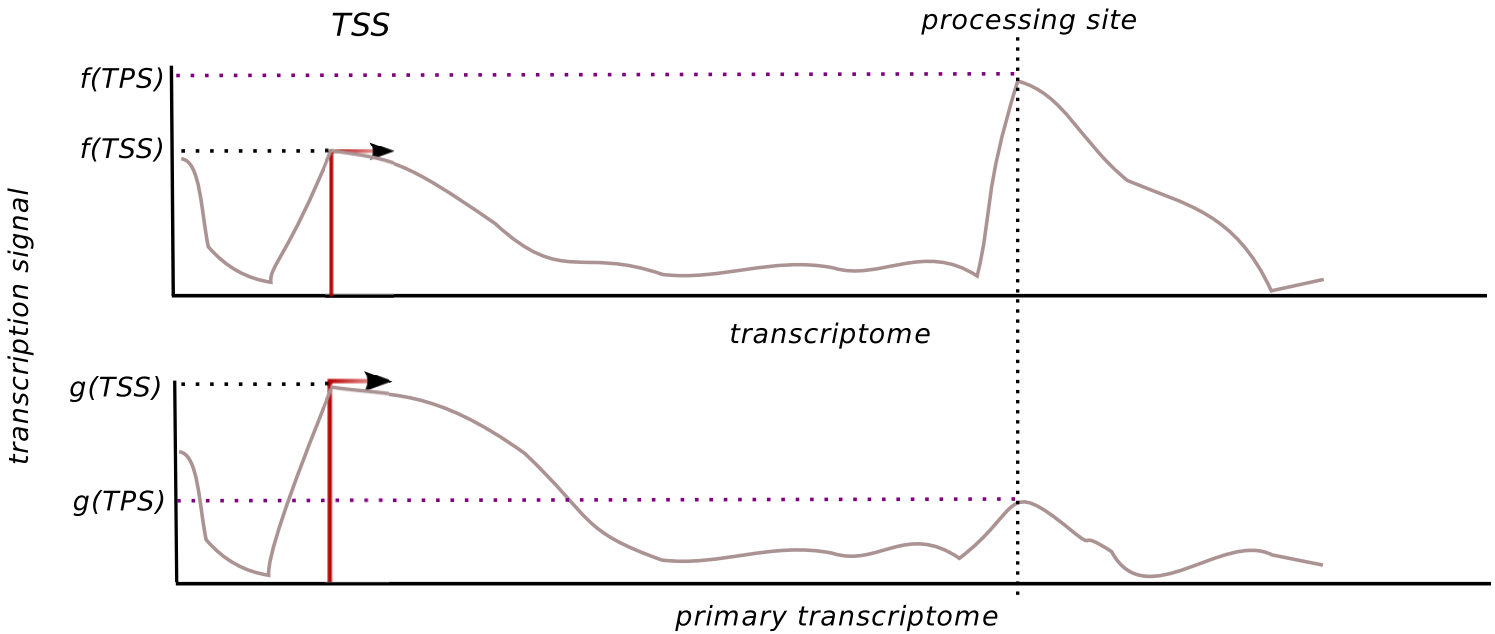


Figure: Threshold business: $g(TSS) > a \times f(TSS)$?

dRNAseq: when to assume a TSS?

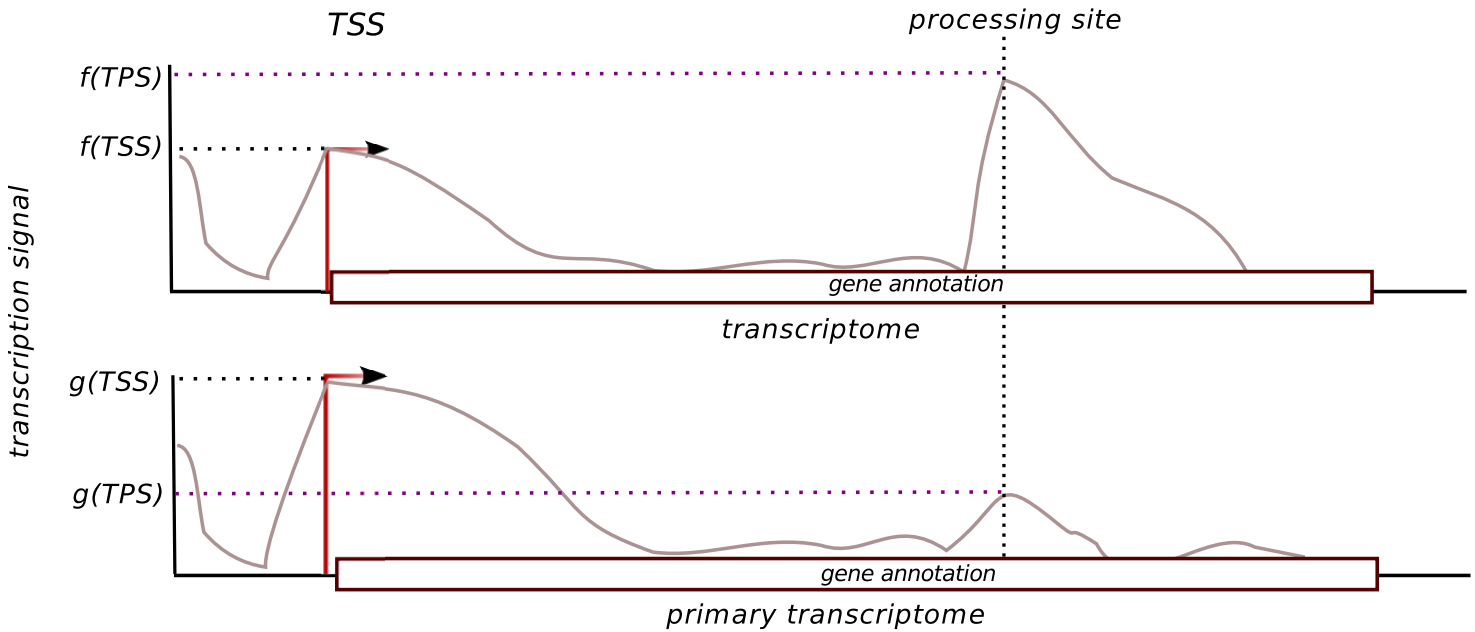


Figure: Threshold business: $g(TSS) > a \times f(TSS)$?

Overlap w/ annotation

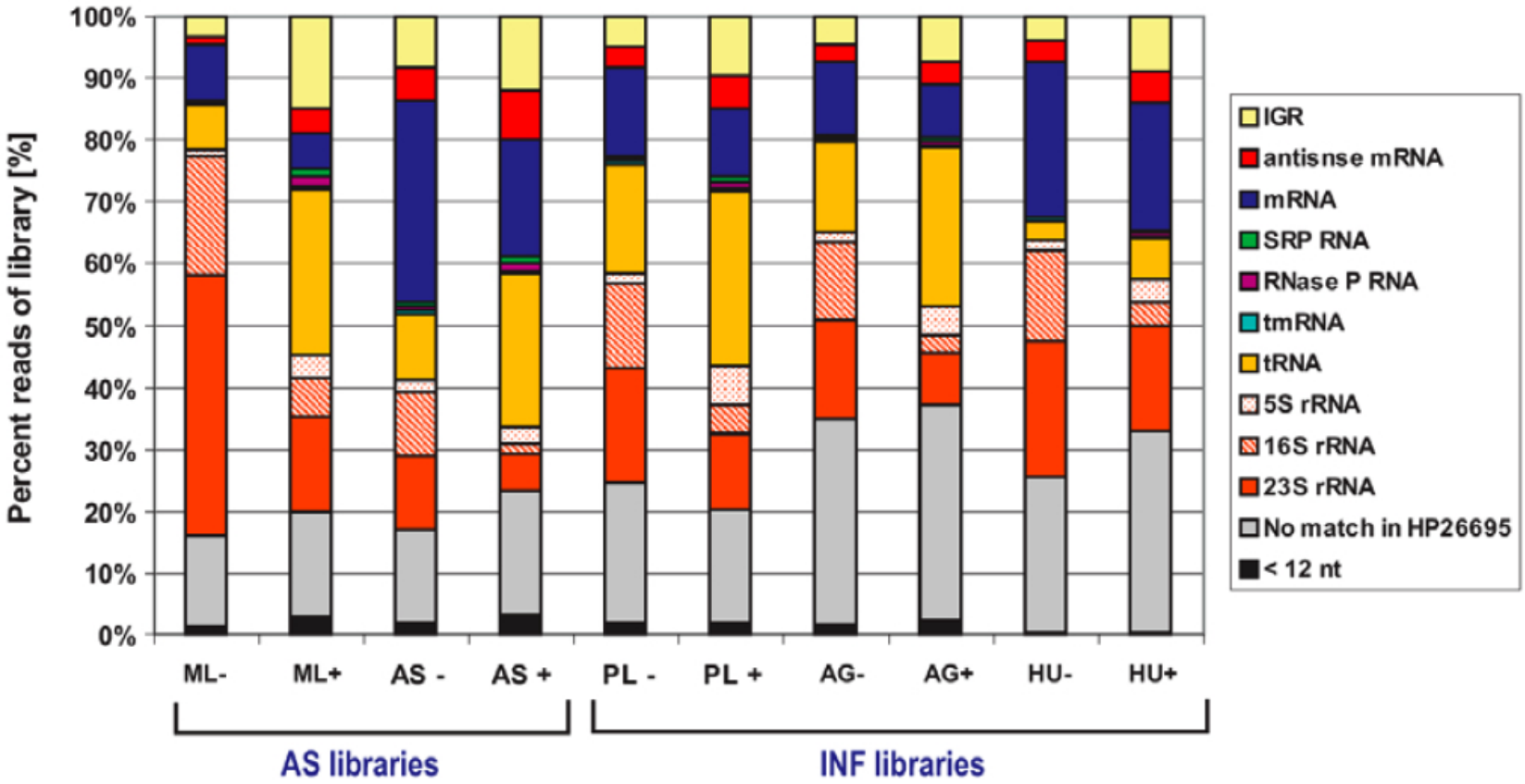
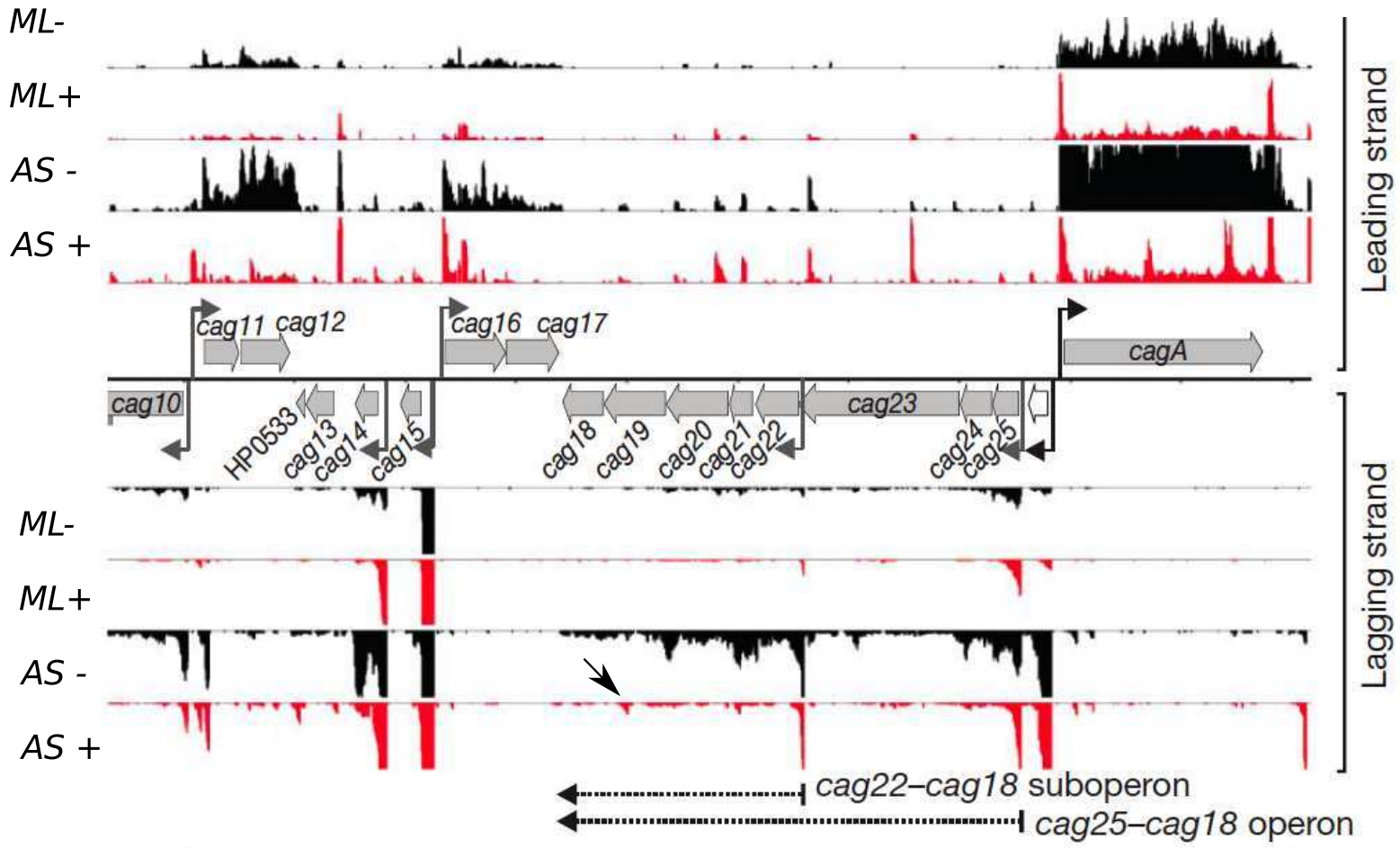


Figure: Enrichment reveals a fairly large fraction of antisense and intergenic transcription.

Annotation of TSS for known "genes"



Classes of transcription start sites

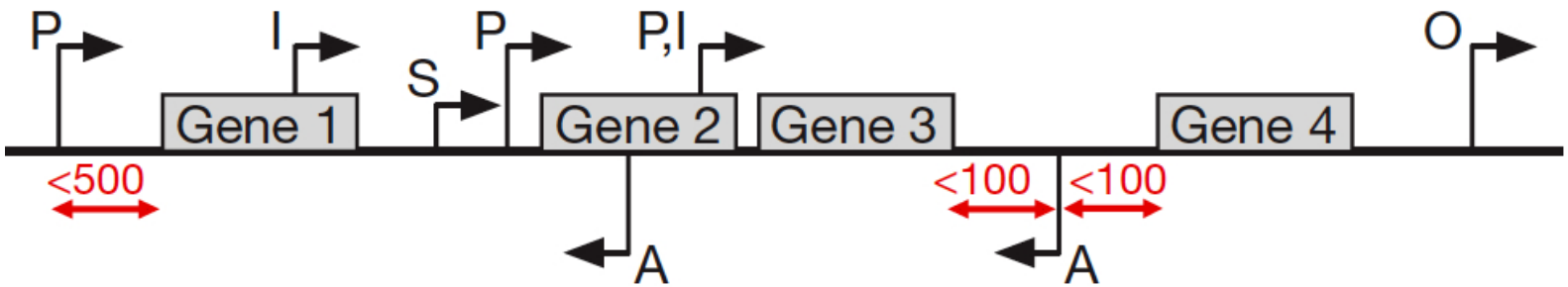
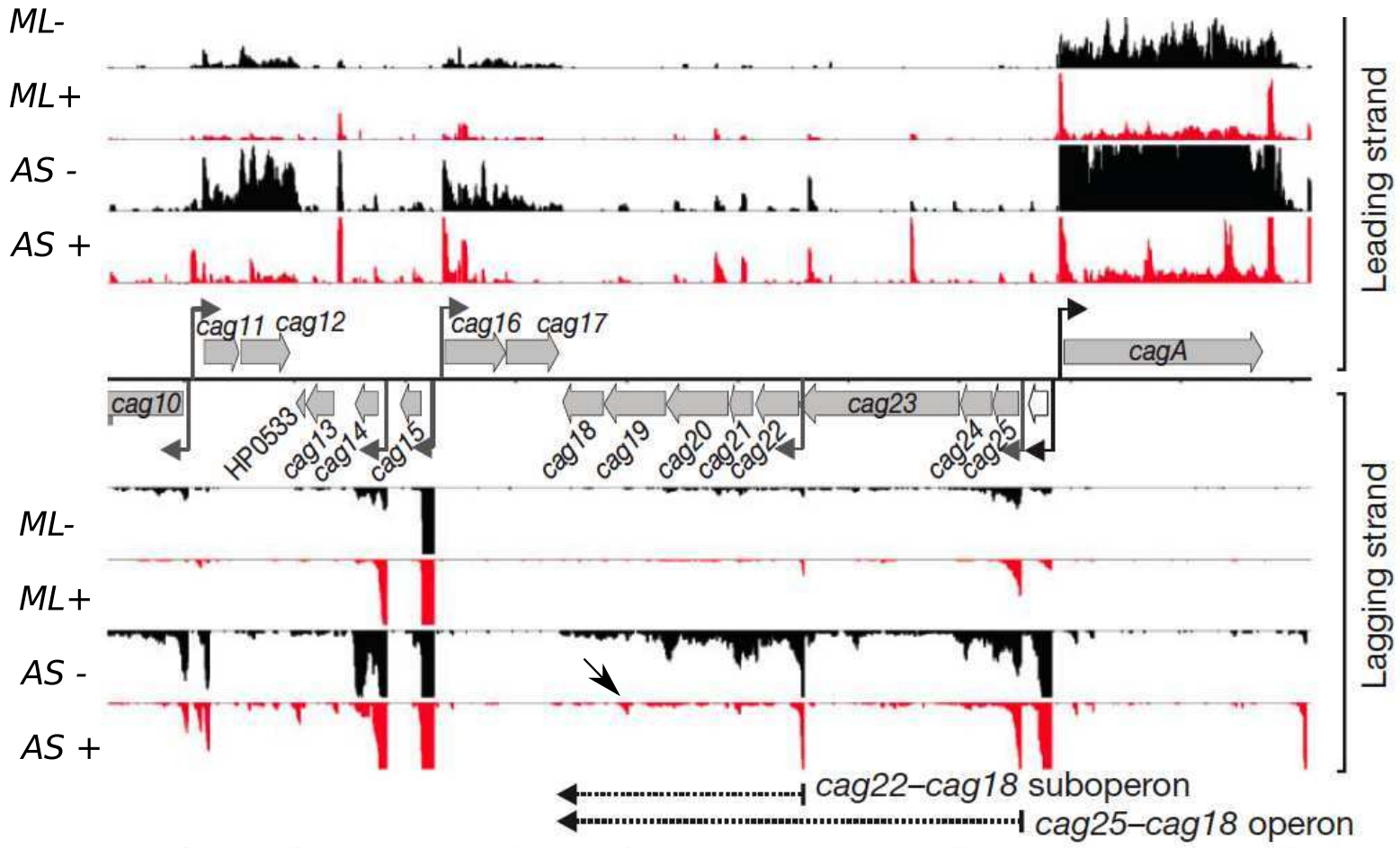


Figure: We classify primary (P), secondary (S), internal (I) and antisense (A) start sites. O denotes an orphan TSS

Annotation of TSS for known "genes"



Annotation of TSS (verification)

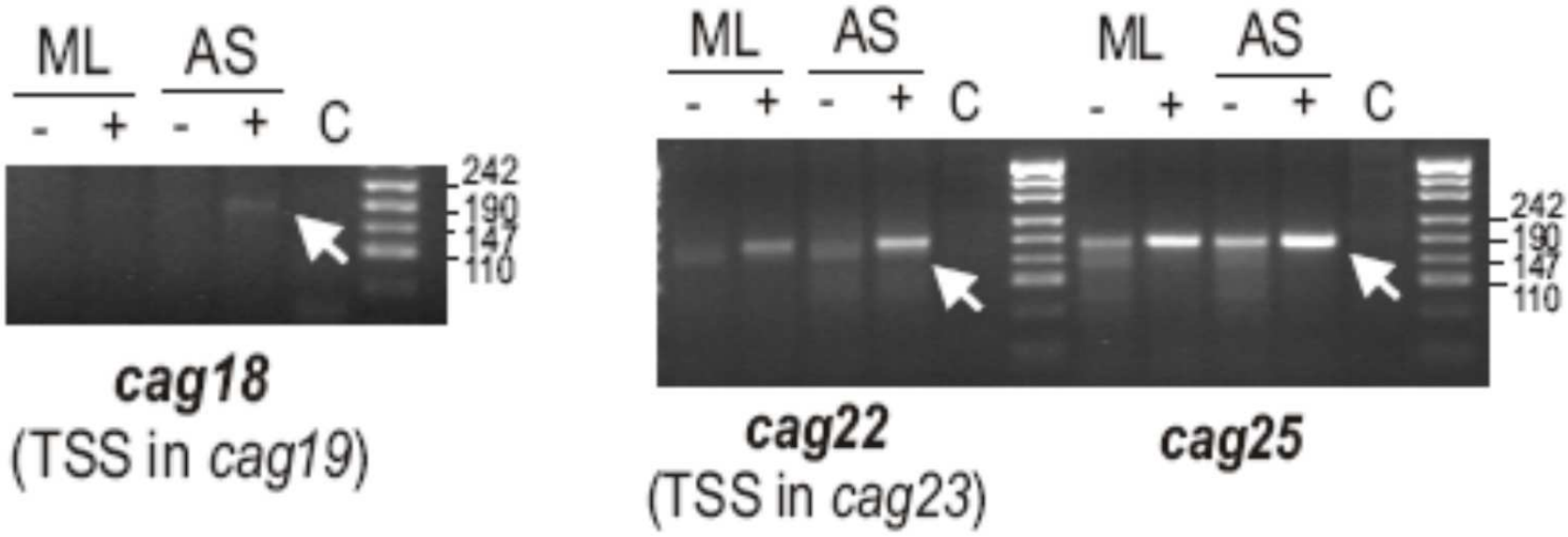


Figure: Confirmation by 5' RACE on RNA from AS and ML libraries.

Annoation of TSS (verification)

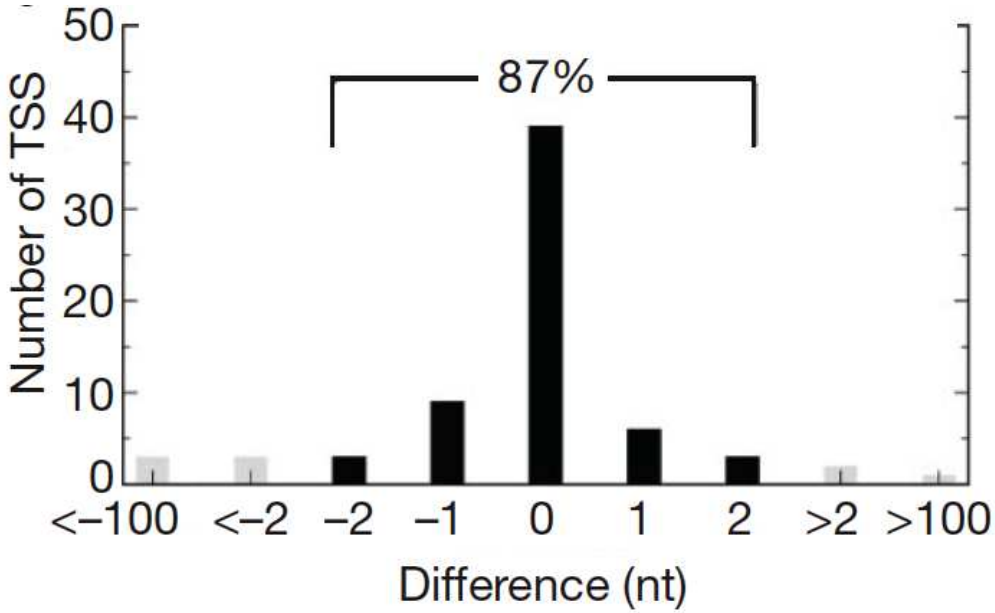


Figure: dRNAseq is accurate.

Riboswitches

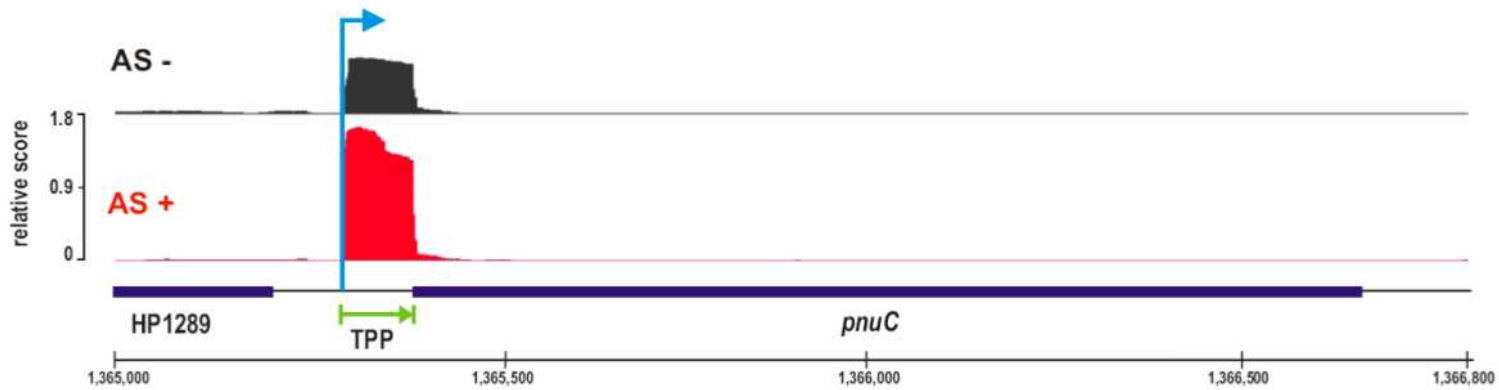


Figure: A predicted riboswitch was confirmed in the dRNAseq approach.

Riboswitches (verification)

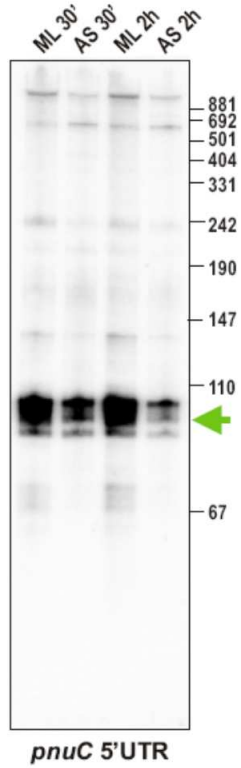


Figure: Confirmation in northern blot.

Small RNAs: 6S RNA and its regulator

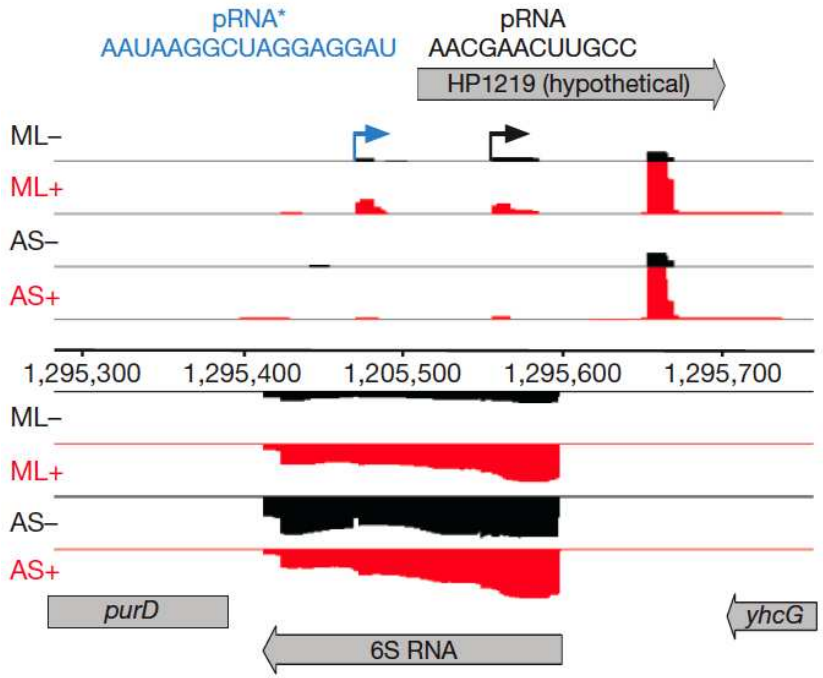


Figure: 6S RNA, interacting with the RNA polymerase holoenzyme, was successfully detected using the dRNAseq approach. Note the pRNAs.

Small RNAs: 6S RNA and its regulator

6S RNA associates with the active site of RNAP and serves as a template for the synthesis of short RNA products (pRNAs) in vitro and in cells. pRNA synthesis destabilizes the complex between RNAP and the 6S RNApRNA hybrid.[1]

[1] Kugel et al., Nature, 2007

Other regulatory sRNAs

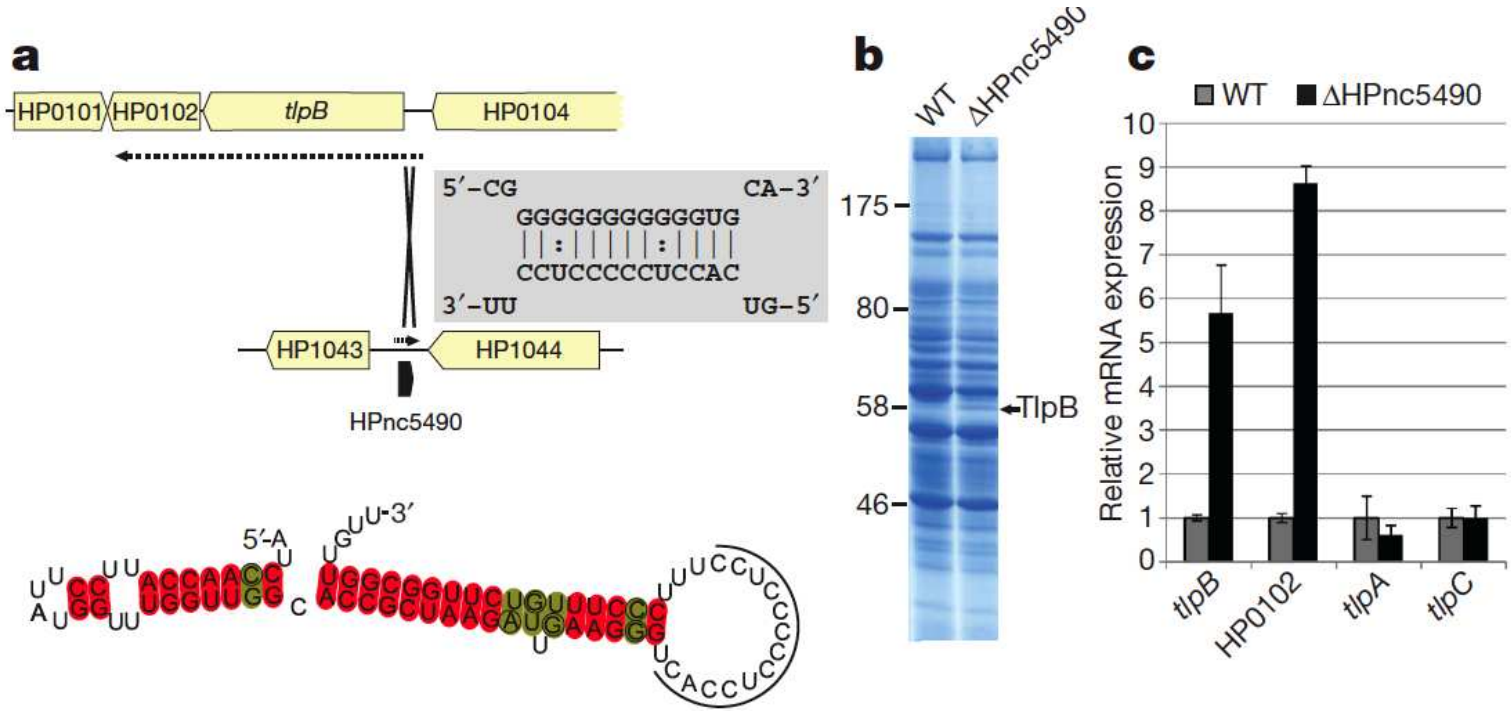


Figure: Trans-encoded regulatory RNAs: tlpB chemotaxis receptor is regulated by a small RNA at a different location.

Reannotation

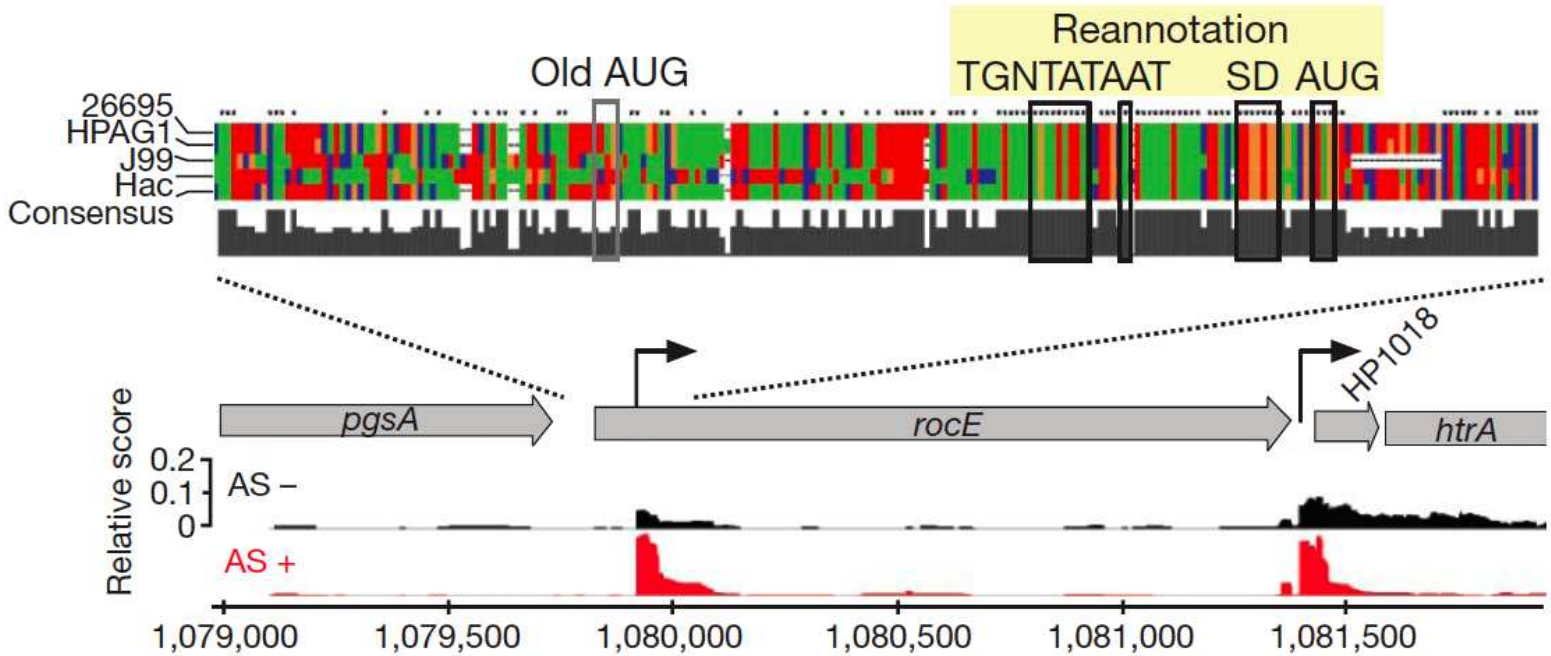


Figure: dRNAseq helps to annotate and correct annotation.

Summary

- annotation of 1907 TSS accross the genome
- massive antisense transcription: 27% aTSS
- identification of ≈ 60 sRNAs (northern blot conf'd)
- regulatory sRNAs in trans and cis (e.g. fucT)
- identification of 6S RNA with pRNAs
- 2.2 % of all mRNAs have a 5'UTR < 10 nt
- 26 coding transcripts (dnaA, recR and hemH) init'd at AUG
- ORF corrections for 19 genes proposed

Herzlicher Dank geht an ...

doi:10.1038/nature08756 nature

ARTICLES

The primary transcriptome of the major human pathogen *Helicobacter pylori*

Cynthia M. Sharma[1], Steve Hoffmann[2], Fabien Darfeuille[3,4], Jérémy Reignier[3,4], Sven Findeiß[2], Alexandra Sittka[1], Sandrine Chabas[3,4], Kristin Reiche[5], Jörg Hackermüller[5], Richard Reinhardt[6], Peter F. Stadler[2,5,7,8,9] & Jörg Vogel[1,10]

Genome sequencing of *Helicobacter pylori* has revealed the potential proteins and genetic diversity of this prevalent human pathogen, yet little is known about its transcriptional organization and noncoding RNA output. Massively parallel cDNA sequencing (RNA-seq) has been revolutionizing global transcriptomic analysis. Here, using a novel differential approach (dRNA-seq) selective for the 5′ end of primary transcripts, we present a genome-wide map of *H. pylori* transcriptional start sites and operons. We discovered hundreds of transcriptional start sites within operons, and opposite to annotated genes, indicating that complexity of gene expression from the small *H. pylori* genome is increased by uncoupling of polycistrons and by genome-wide antisense transcription. We also discovered an unexpected number of ~60 small RNAs including the ε-subdivision counterpart of the regulatory 6S RNA and associated RNA products, and potential regulators of *cis*- and *trans*-encoded target messenger RNAs. Our approach establishes a paradigm for mapping and annotating the primary transcriptomes of many living species.

Figure: ...