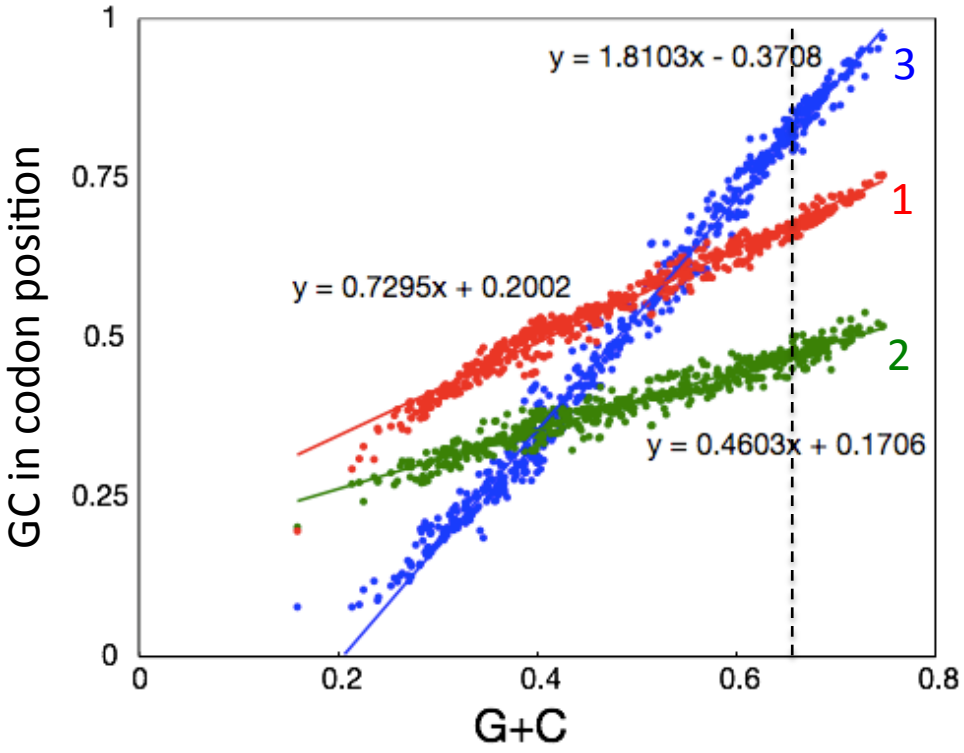
The coding potential of *Pseudomonas aeruginosa*: Comparative genomics and ribosome profiling

Luciano Brocchieri

Department of Molecular Genetics & Microbiology and Genetics Institute
University of Florida, Gainesville, FL 32610

Gene finding and GC content

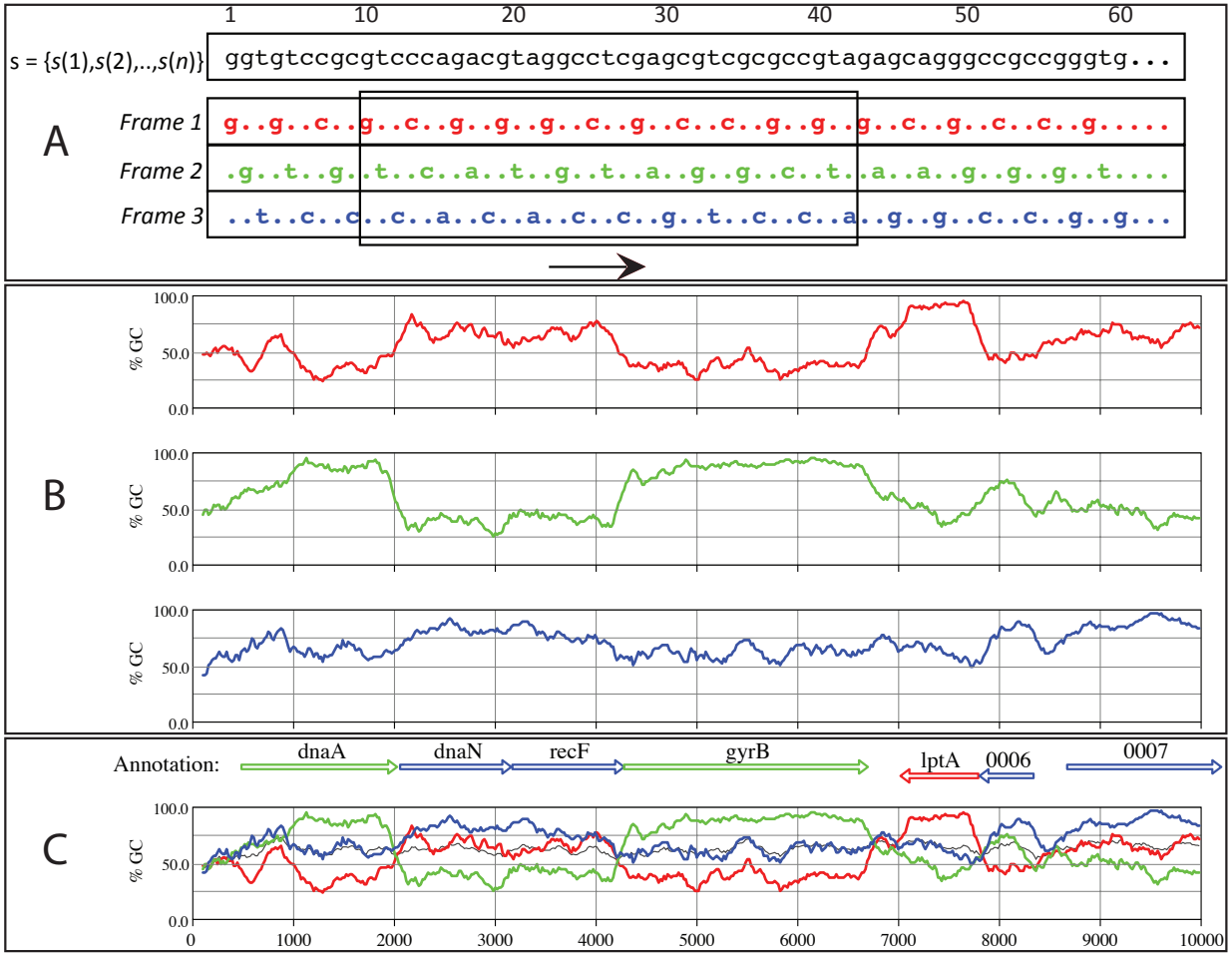


Pseudomonas aeruginosa:

67.7% GC

Coding regions are characterized by typical 3-base periodicity in GC content, depending on the overall GC content of the sequence.
(Bibb et al. 1984; Borodovsky and McIninch 1993; Besemer et al. 2001)

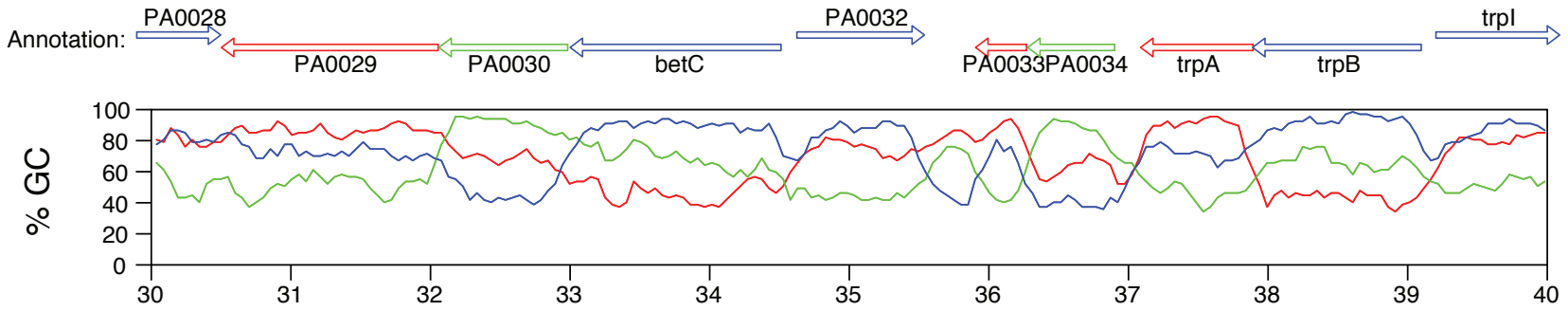
Frame analysis (Bibb et al 1984)



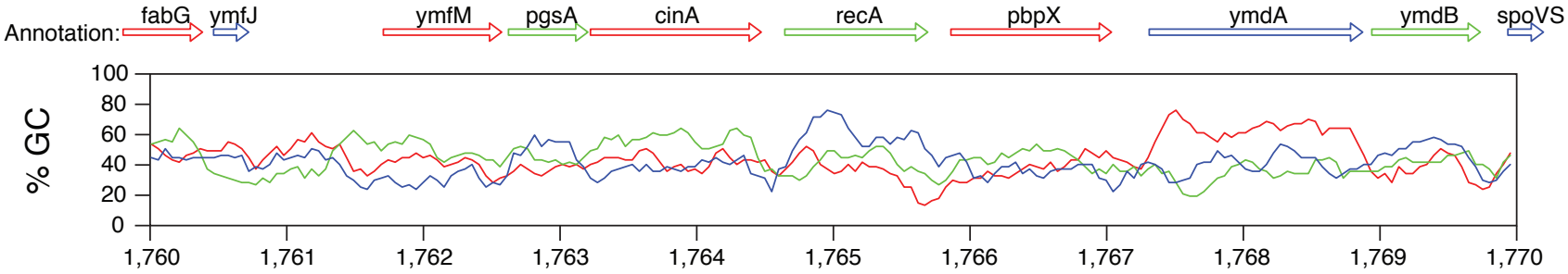
GC content is measured every third nucleotide in three phases. Compositional contrasts among S-profiles of GC content indicate presence and frame of coding regions.

Frame analysis and sequence GC content

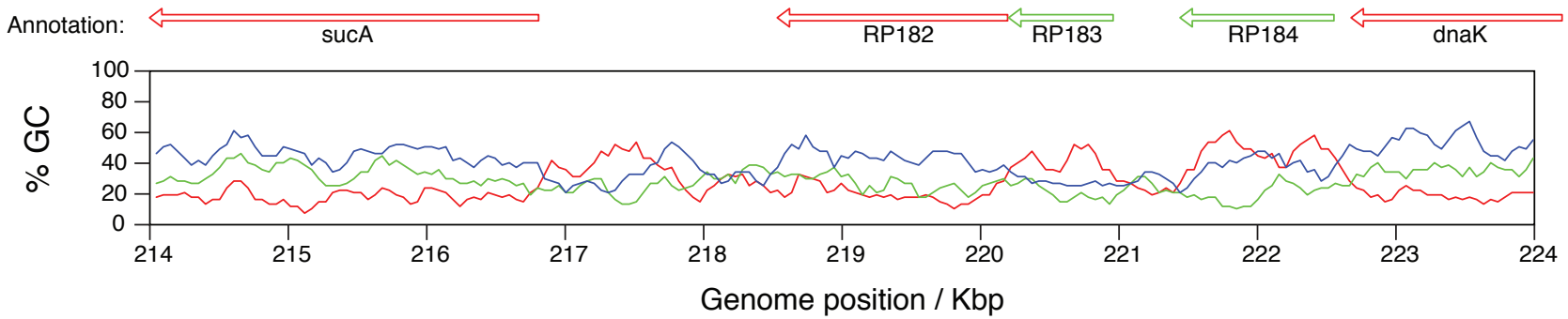
NC_002516 *Pseudomonas aeruginosa* PAO1 (%GC = 68.2)



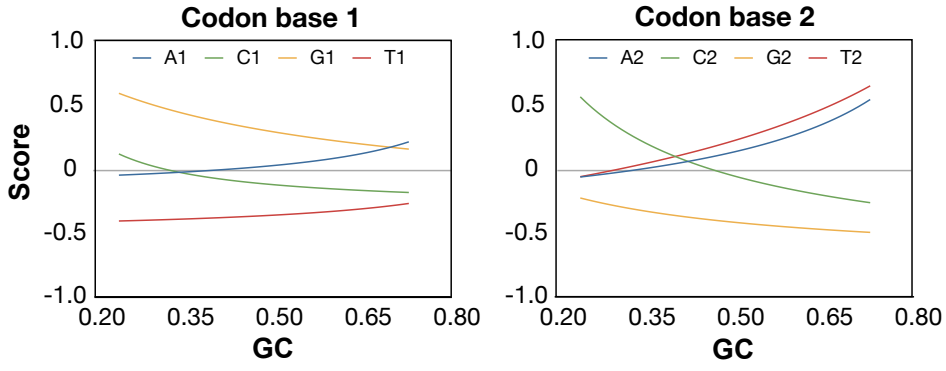
NC_000964 *Bacillus subtilis subtilis* 168 (%GC = 43.1)

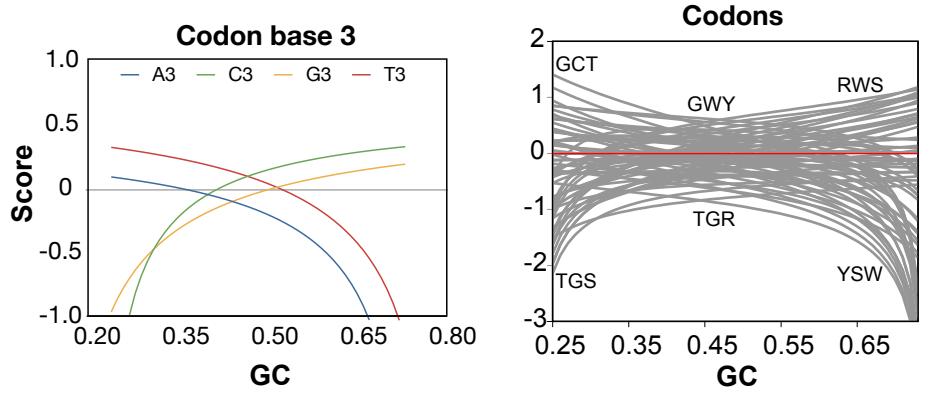


NC_000963 *Rickettsia prowazekii* Madrid E (%GC = 32.2)



Relative representation of nucleotides at each codon position can be used as scores for nucleotide usage

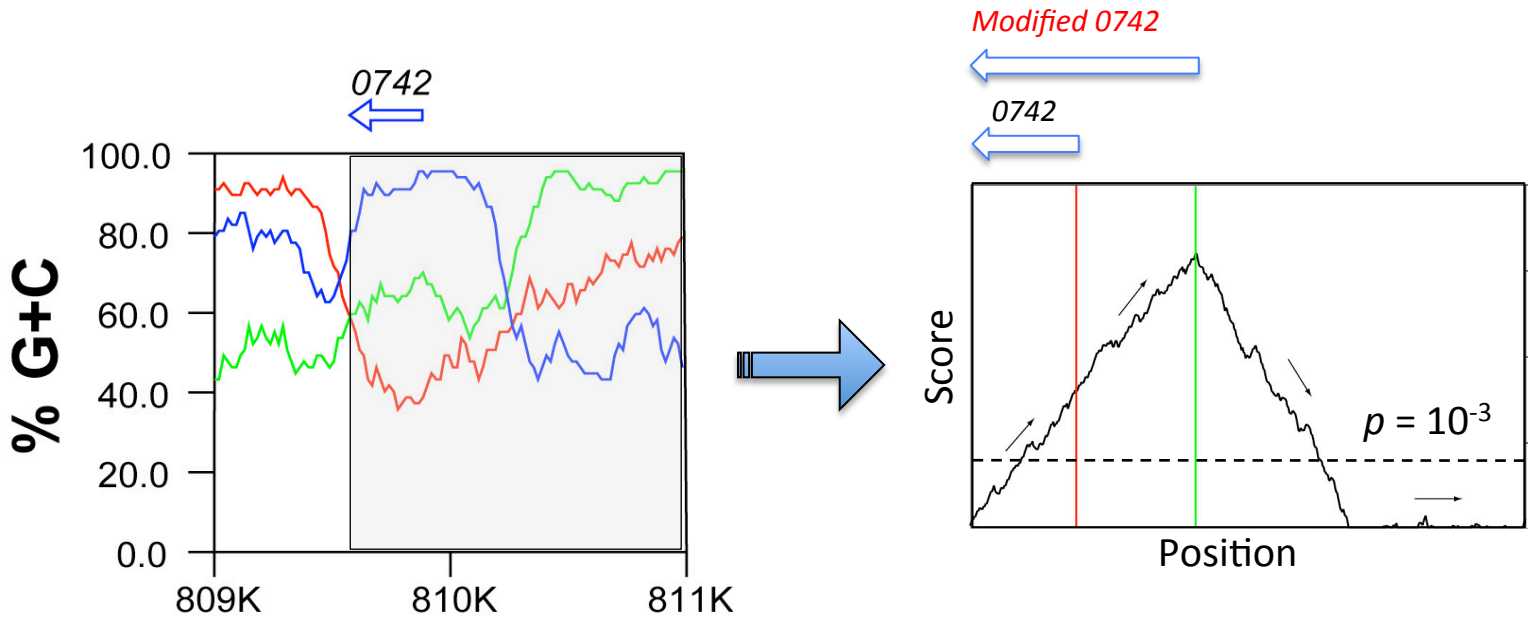




$$\text{Score}(XYZ) = \ln \frac{p_{X_1} p_{Y_2} p_{Z_3} \left(1 - p_{stop}^{(0)}\right)}{p_X p_Y p_Z \left(1 - p_{stop}^{(1)}\right)}$$

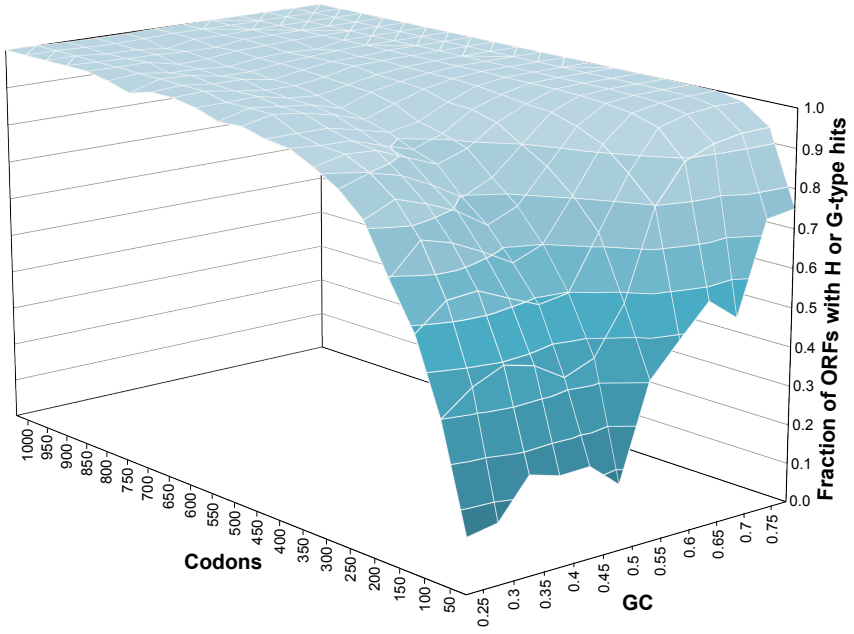
Scores that depend on global nucleotide composition are calculated for each trinucleotide

Cumulative scores and H-type hits

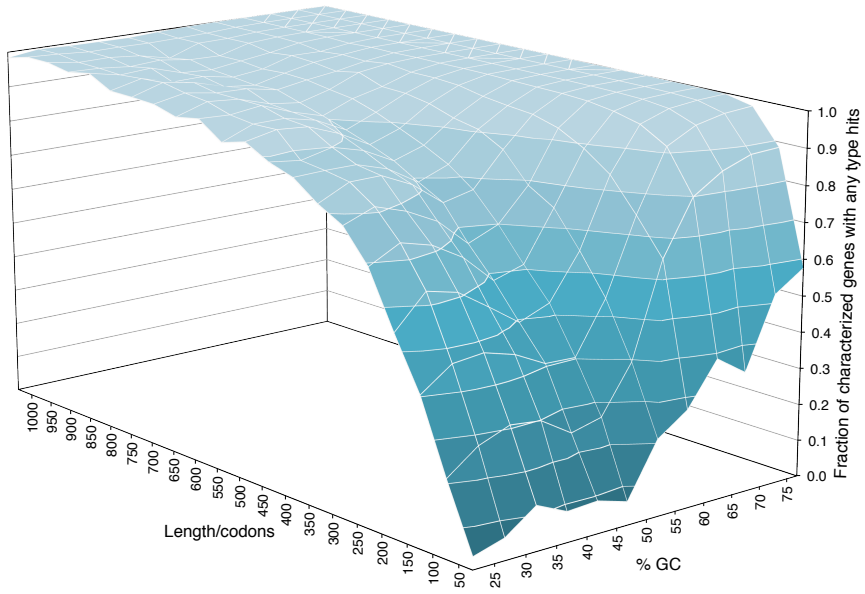


Cumulative scores can be used to translate qualitative visual information on compositional contrasts (left) into statistically characterized (Karlin and Altschul 1990, Karlin 1994) measures of coding potential for precisely defined sequence segments (right).

Significant contrasts in characterized genes



$p \leq 0.01$



$p \leq 0.001$

Power of the tests is high for all long sequences and decreases more with shorter sequences of low or intermediate GC content

ORF structure and local composition: The reading-frame prediction-unit

Reading frame

The sequence is read frame-by-frame independently considering sequence segments included between consecutive stop codons (i.e., analyzed segments do not contain stop codons).

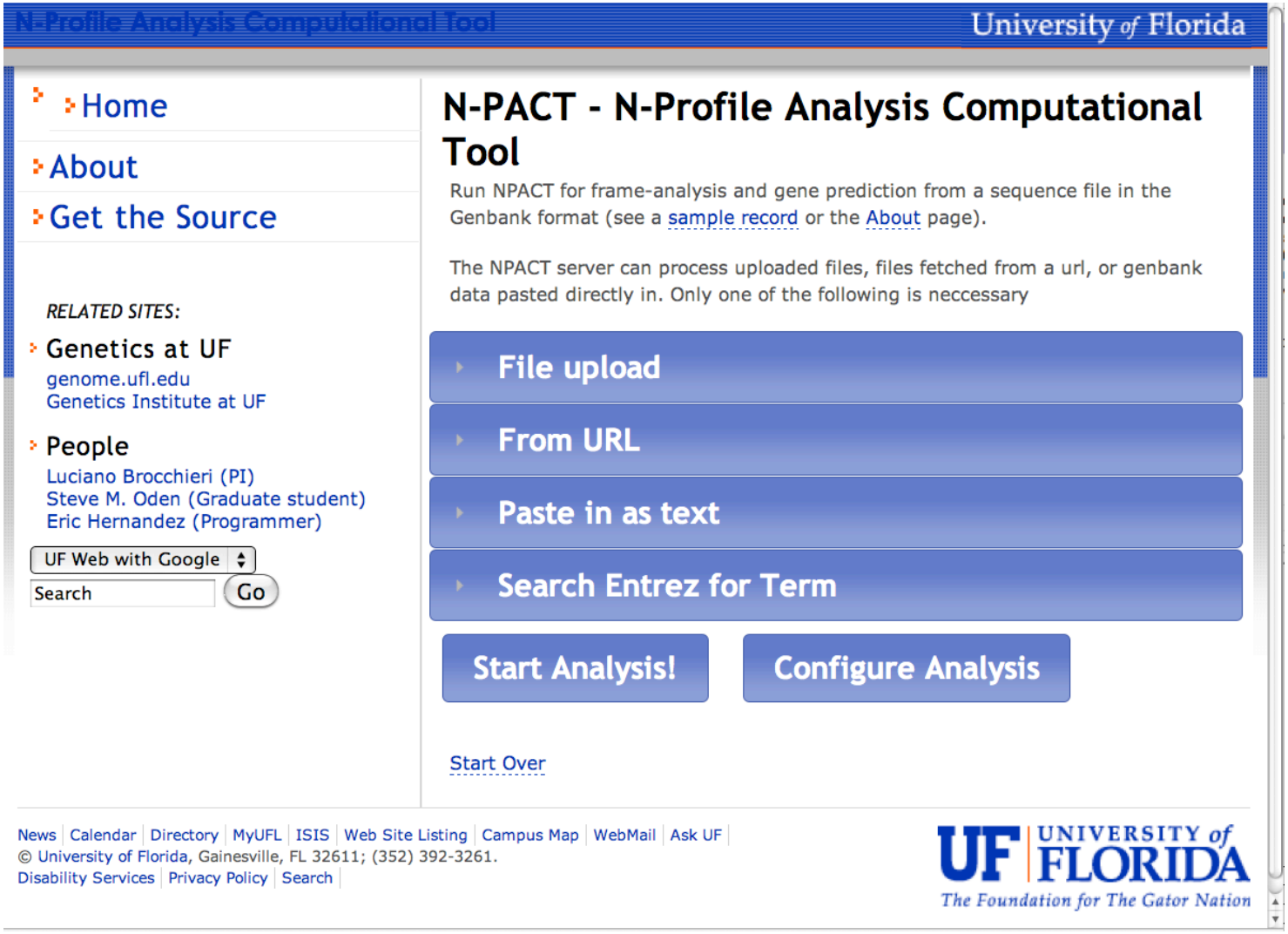
Local composition

The local composition of the each sequence segments is independently calculated from each selected reading frame and segment-specific scores are built based on its composition.

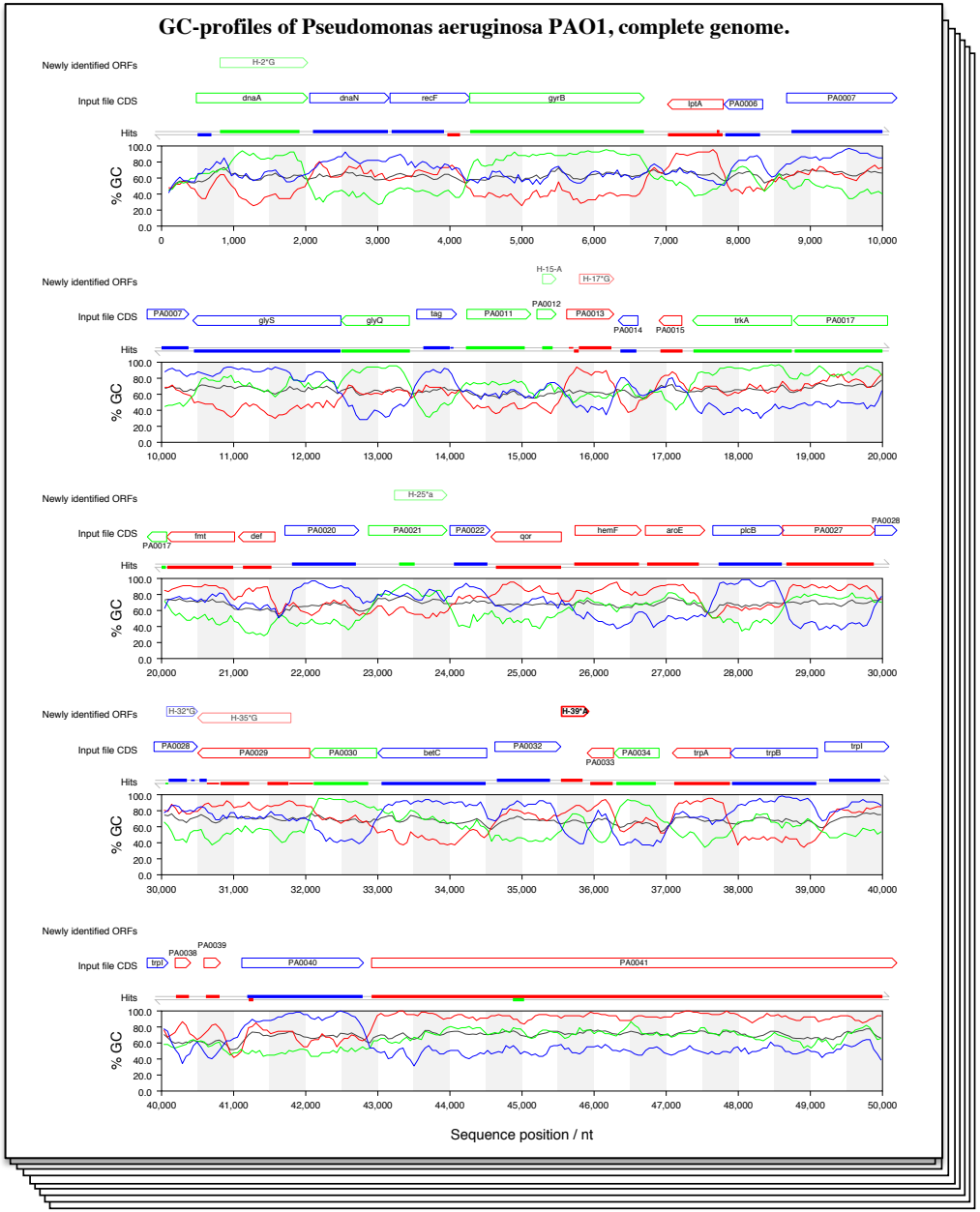
ORF structure

Only hits included within ORFs are considered (i.e., a potential start codon must be identified).

NPACT home web page (http://genome.ufl.edu/npact/)



NPACT graphical output



Newly-identified ORFs in *P. aeruginosa* strains

Strain	ID	Genome lengh	GC%	Annotated CDS	Newly-identified ORFs
PAO1	NC_002516	6264404	66.56	5572	179
UCBPP-PA14	NC_008463	6537648	66.29	5892	173
PA7	NC_009656	6588339	66.45	6286	189
LESB58	NC_011770	6601757	66.3	5925	258
M18	NC_017548	6327754	66.5	5684	161
NCGM2_S1	NC_017549	6764661	66.14	6268	250
DK2	NC_018080	6402658	66.27	5883	157
B136_33	NC_020912	6421010	66.42	5828	160
RP73	NC_021577	6342034	66.46	5762	187

From 157 to 258 ORFs with significant ($p < 0.001$) compositional periodicity and not corresponding to annotated genes are identified in different strains of *P. aeruginosa*.

Conservation of newly-identified ORFs between *Pseudomonas aeruginosa* strains

Strain	Cat	B136_33	DK2	LESB58	M18	NCGM2_S1	PA7	PAO1	RP73	PA14	Cons	Non Cons.
B136_33	ORF		66	47	49	78	22	48	59	66	130	30
	CDS	27	43	47	44	45	46	44	43	48		
DK2	ORF	66		60	63	57	28	61	69	55	135	22
	CDS	48	30	54	51	49	56	52	49	59		
LESB58	ORF	47	60		66	65	22	108	62	59	212	46
	CDS	128	130	51	131	121	130	79	127	120		
M18	ORF	49	63	66		56	27	62	58	62	134	27
	CDS	53	49	56	26	48	49	44	53	54		
NCGM2_S1	ORF	78	57	65	56		30	58	55	66	175	75
	CDS	79	72	57	64	31	79	51	71	65		
PA7	ORF	22	28	22	27	30		27	28	32	98	91
	CDS	47	49	39	45	45	34	45	51	49		
PAO1	ORF	48	61	108	62	58	27		58	68	187	11
	CDS	120	121	70	115	107	113	32	125	104		
RP73	ORF	59	69	62	58	55	28	58		53	171	17
	CDS	91	98	96	99	93	93	101	49	104		
PA14	ORF	66	55	59	40	66	32	68	53		146	27
	CDS	66	64	44	55	54	60	38	65	32		

Homologs for the majority of newly identified ORF are identified among annotated genes or newly-identified ORFs of other *P. aeruginosa* strains.

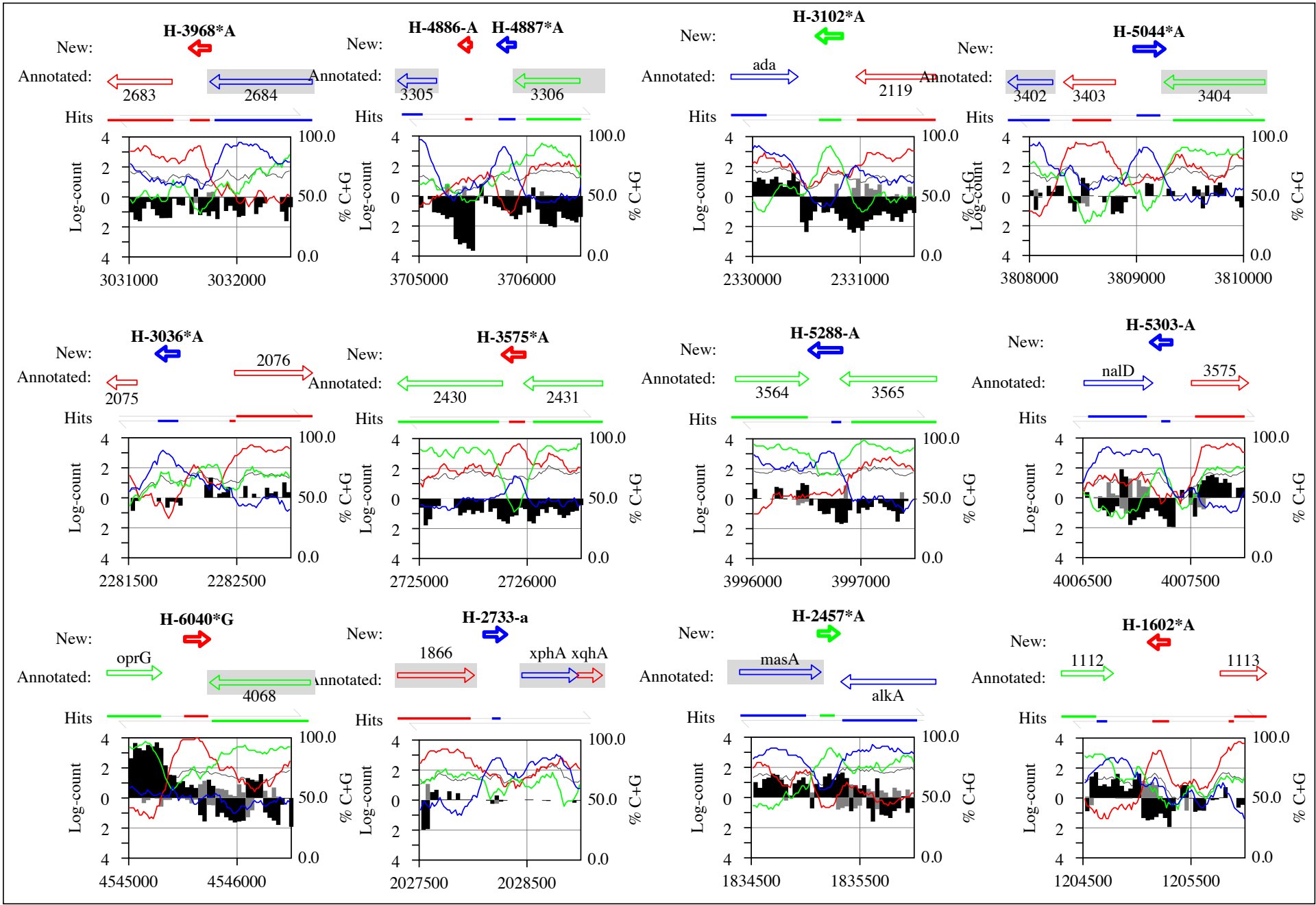
Long-range onservation of ORFs newly-identified in *Pseudomonas aeruginosa* strains

Name	Phyla	Genera	Species	Strains	Tot_con	Not_con	Tot	Tot Strain
PAO1	59	26	21	13	119	60	179	109
UCBPP-PA14	59	10	17	11	97	76	173	89
PA7	67	11	15	17	110	79	189	95
LESB58	128	19	25	10	182	76	258	174
M18	41	17	23	18	99	62	161	88
NCGM2_S1	105	15	27	12	159	91	250	146
DK2	42	16	18	11	87	70	157	75
B136_33	34	11	13	13	71	89	160	62
RP73	84	17	11	7	119	68	187	110
Total	619	142	170	112	1043	671	1714	948

Most newly identified ORF conserved among *P. aeruginosa* strains are conserved over different phyla

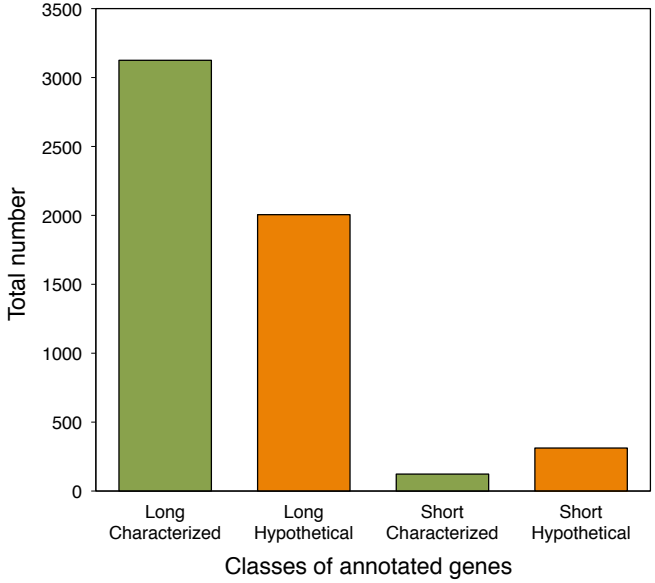
Verification of computational gene prediction by transcriptome analysis in *Pseudomonas aeruginosa* PAO1: RNA-seq and Ribosome Footprinting

The *P. aeruginosa* transcriptome: RNA-seq

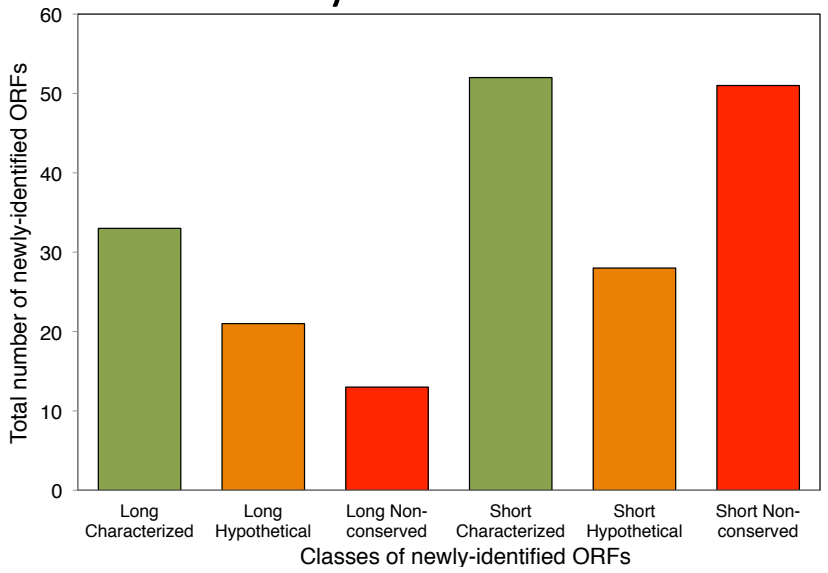


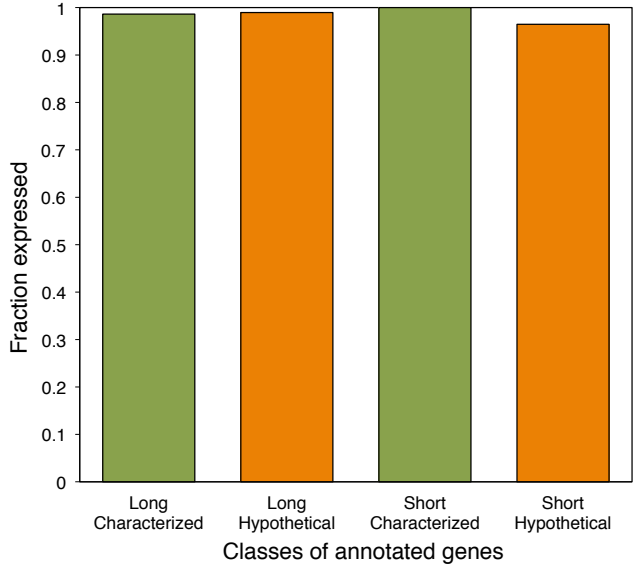
Expression of predicted genes by length and conservation classes

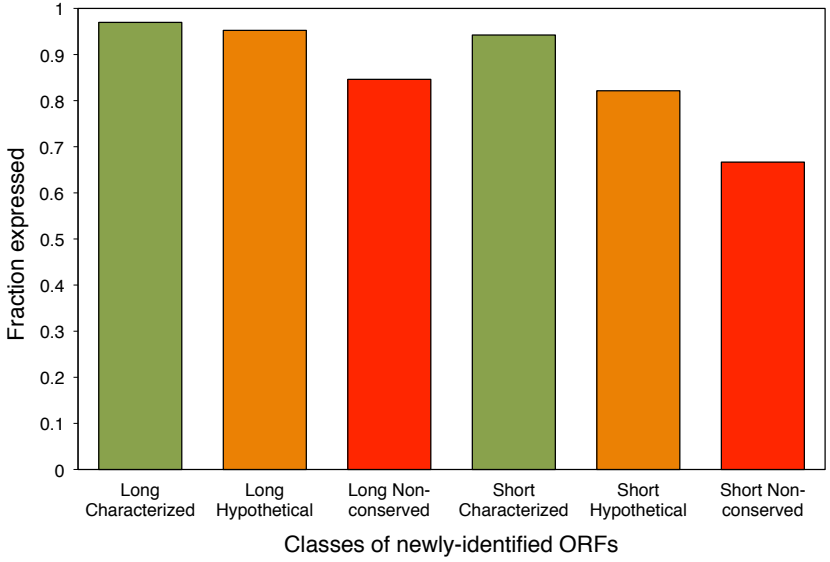
Published annotation



Newly identified ORFs



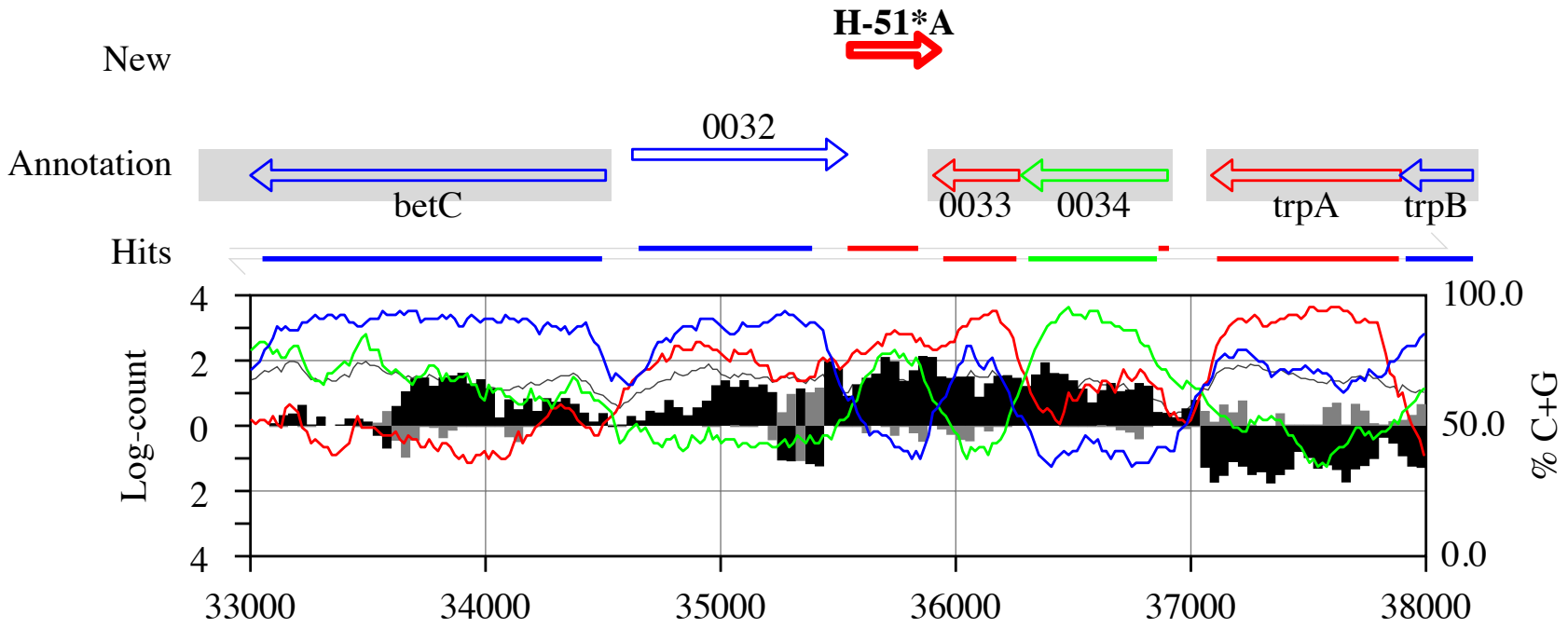




ORFs with RNA-seq reads

What do we learn about gene predictions from transcription in bacteria?

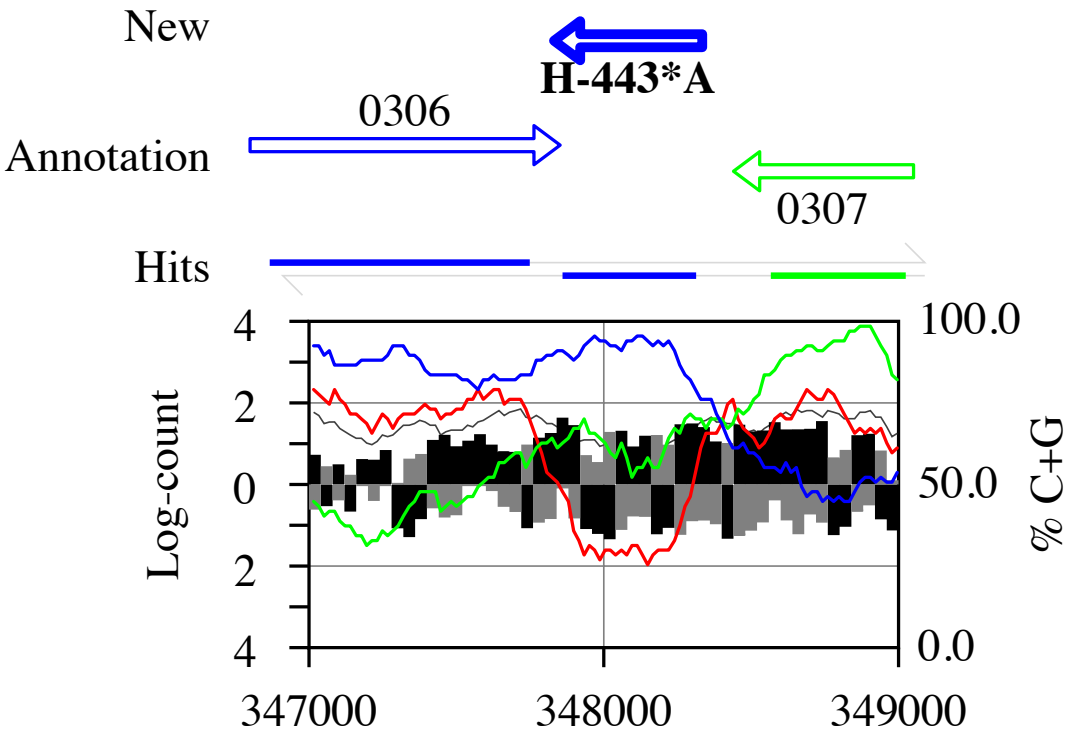
Unexpected patterns



Contradictory patterns of expression of well defined protein coding genes

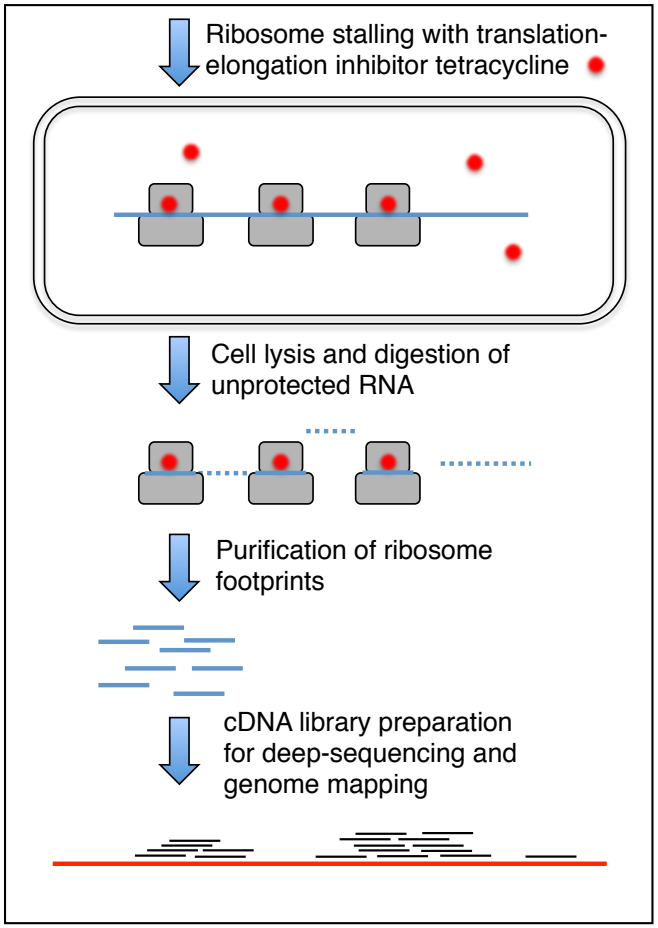
What do we learn about gene predictions from transcription in bacteria?

The problem of antisense transcription



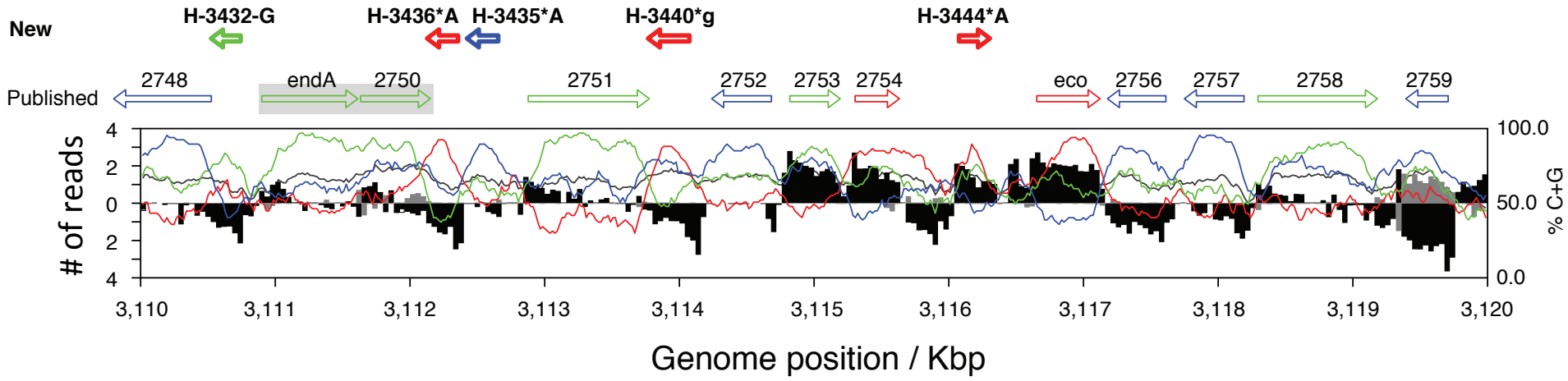
In the case of prediction of H-443*A , sequence features appear to be more convincing than RNA expression data

Ribosome footprinting
(Ingolia et al, *Science* 2009)



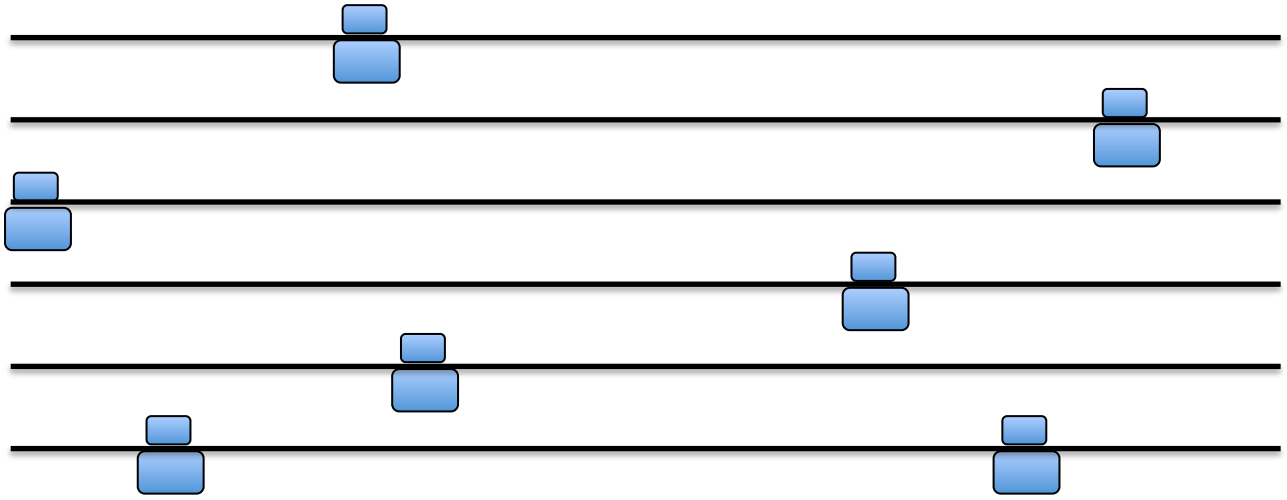
Schematic representation of the ribosome footprinting procedure applied to *P. aeruginosa*

Ribosome footprint coverage in *P. aeruginosa*



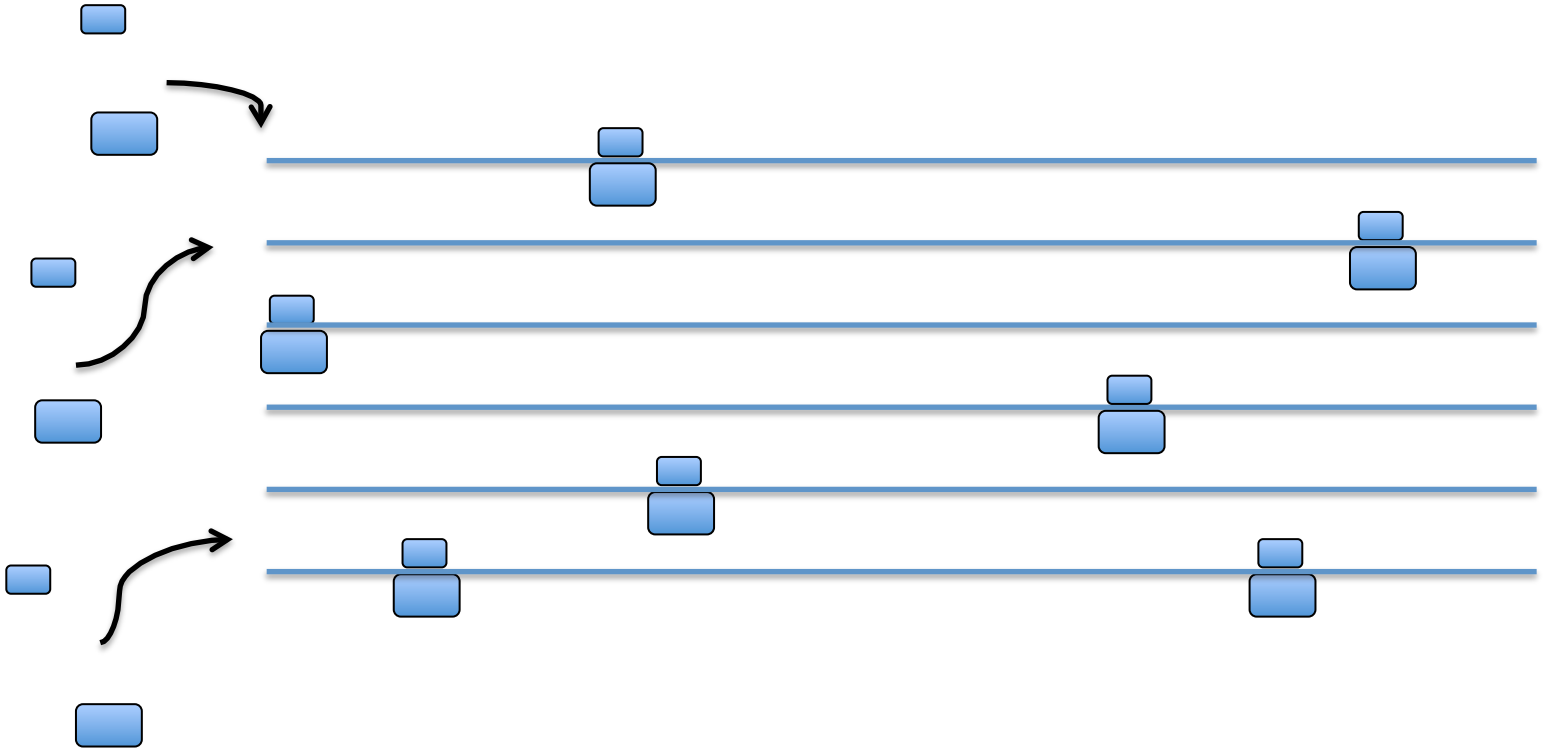
Example of ribosome footprint coverage in *P. aeruginosa* PAO1 showing relation with S-profiles, annotated genes and newly identified ORFs.

Ribosome footprints of initiation sites



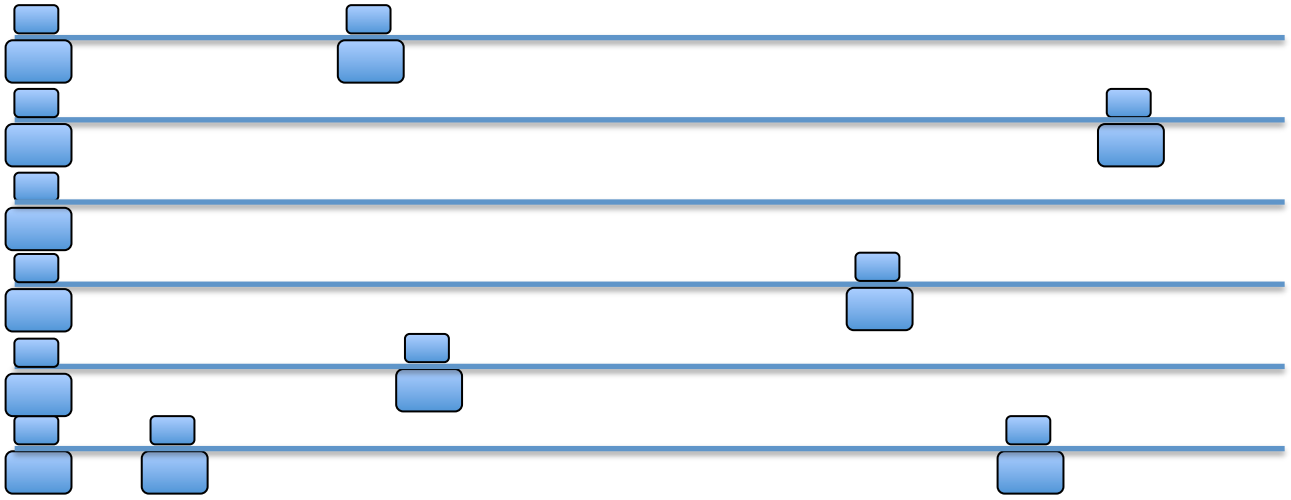
The antibiotic tetracycline inhibits translation-elongation stalling actively-translating ribosomes

Ribosome footprints of initiation sites



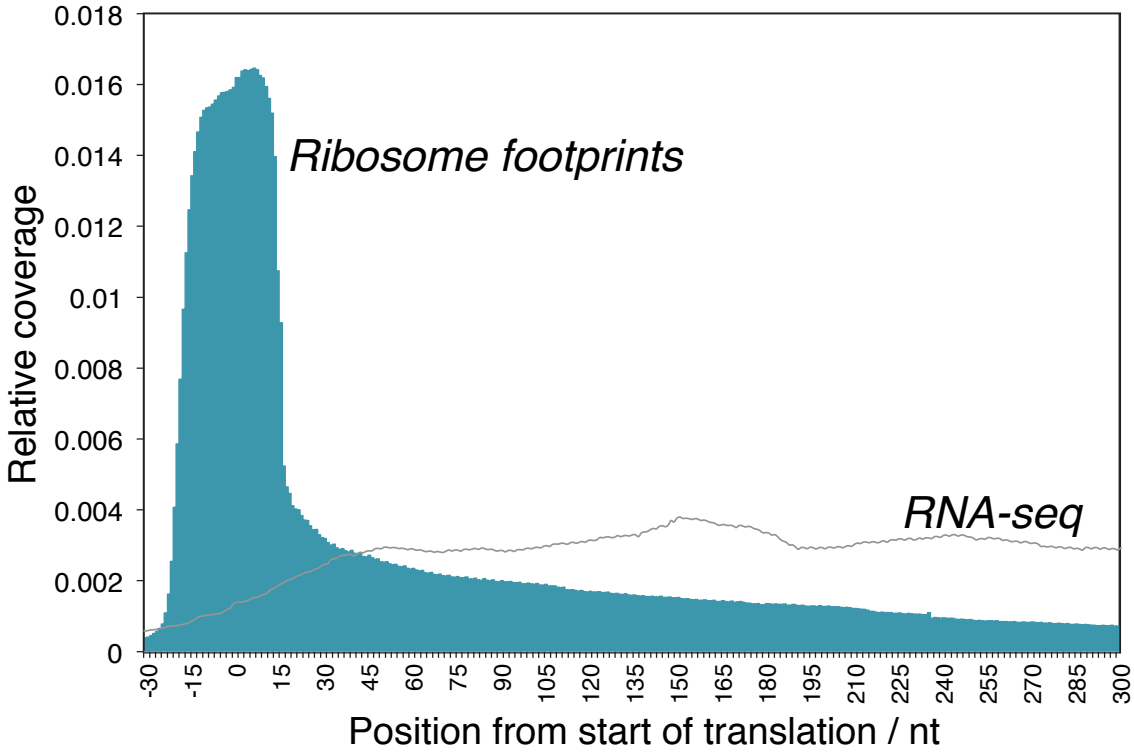
However, tetracycline does not prevent more ribosomes to be recruited at the initiation site.

Ribosome footprints of initiation sites



The accumulation of ribosomes will result in increased numbers of profile-reads corresponding to the initiation site.

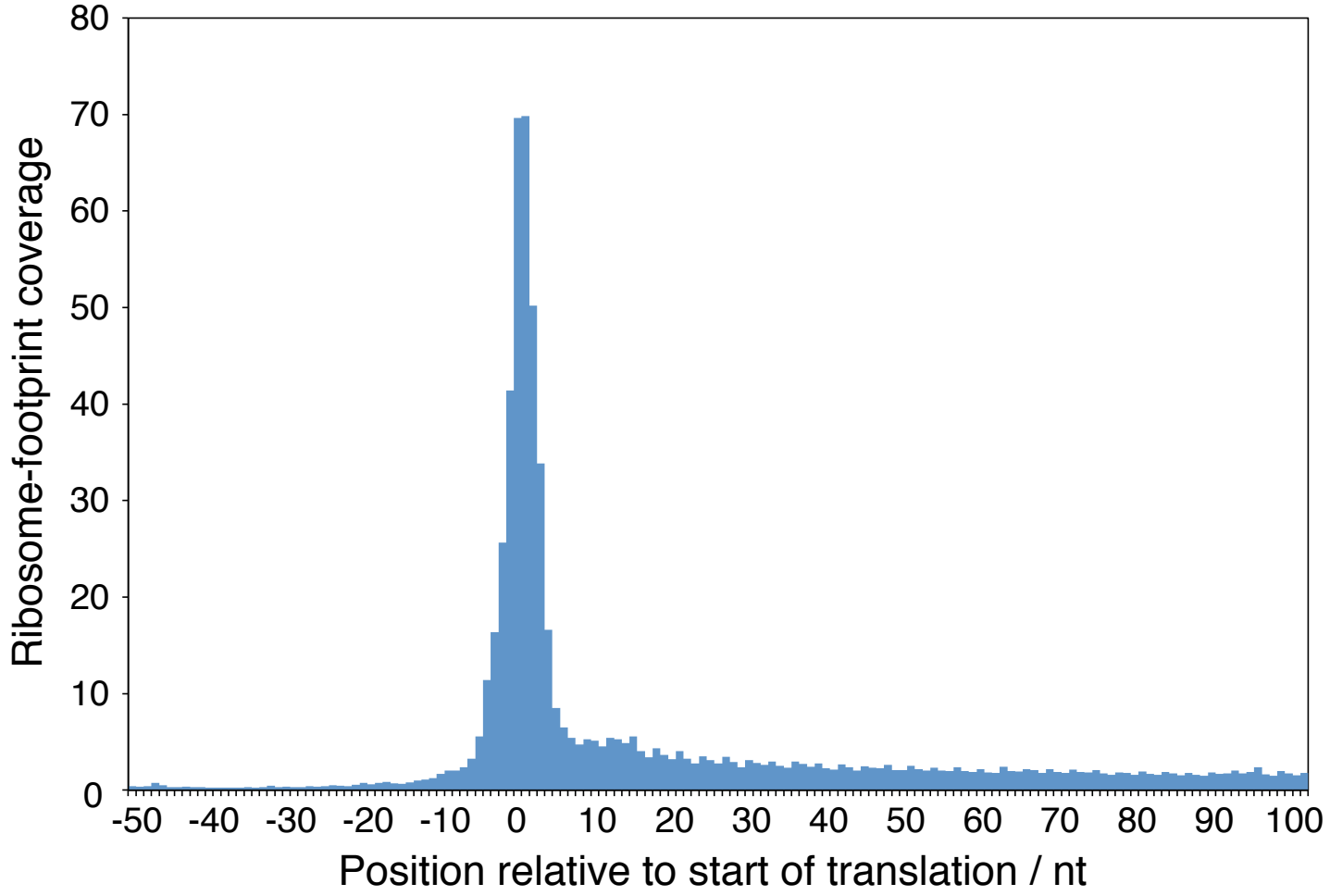
Ribosome footprint coverage by codon position



Metagene analysis of ribosome-footprint coverage

Coverage is averaged over all genes, relative to the start of translation

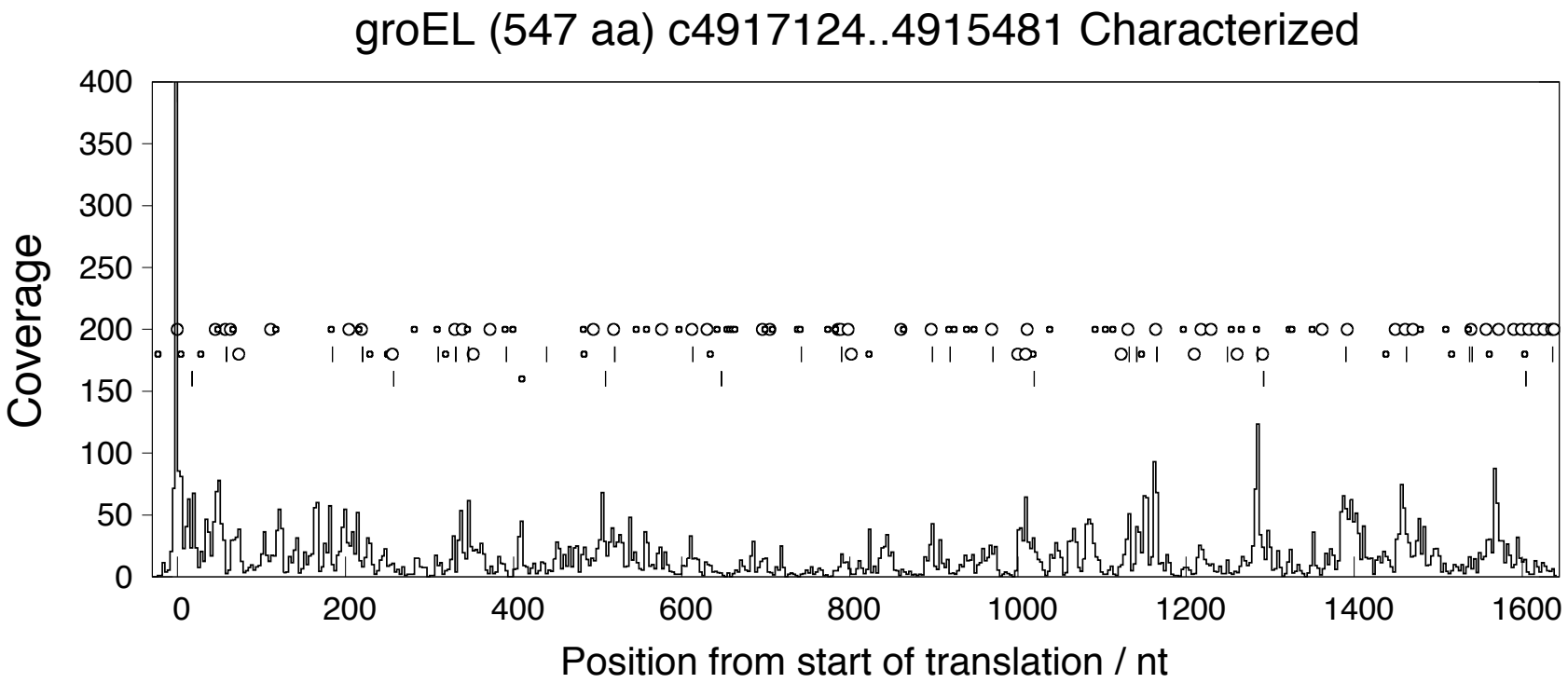
Ribosome footprint coverage by codon position: center of reads



Metagene analysis of coverage by read center + 2 nt

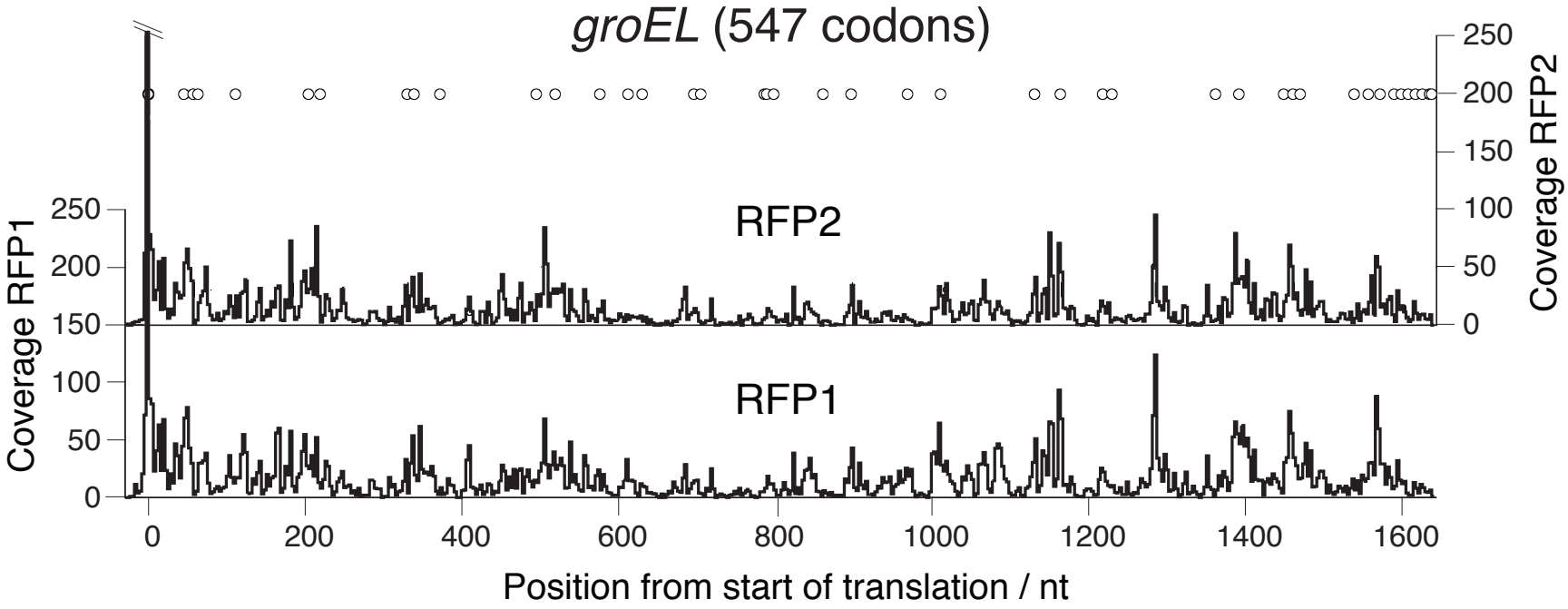
Coverage is averaged over all genes, relative to the start of translation

Translational evidence by ribosome footprinting in *P. aeruginosa*



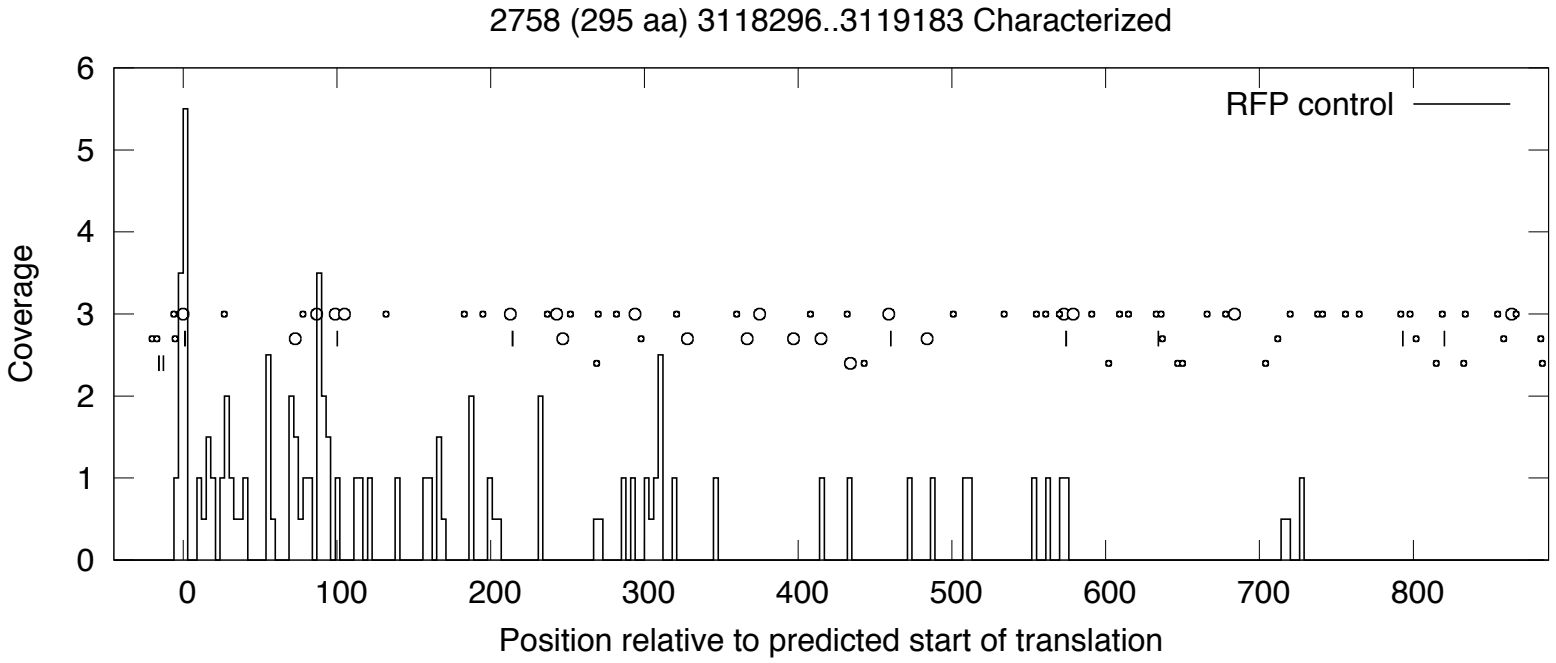
Ribosome-footprint read-count patterns identify mRNA translation, translation initiation site, and translation pausing.

Translational evidence by ribosome footprinting in *P. aeruginosa*



Similar coverage patterns are observed in different biological replicates

Scoring RFP expression



“Strength” of evidence decreases for poorly translated mRNA.

Scoring RFP expression

$$\text{Expression Index} = C_0 \ln \frac{C_0}{C_1}$$

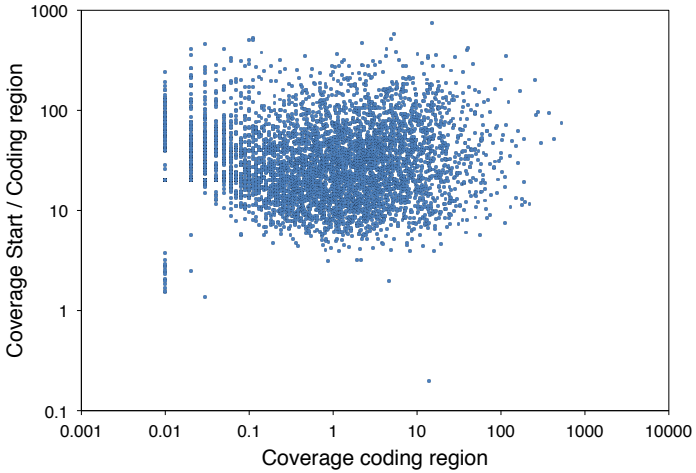
C_0 : Count of RFP reads in codon positions [-2,+3] / 5;

C_1 : Count of RFP reads in codon positions [+9, len/2] / (len/2 -9);

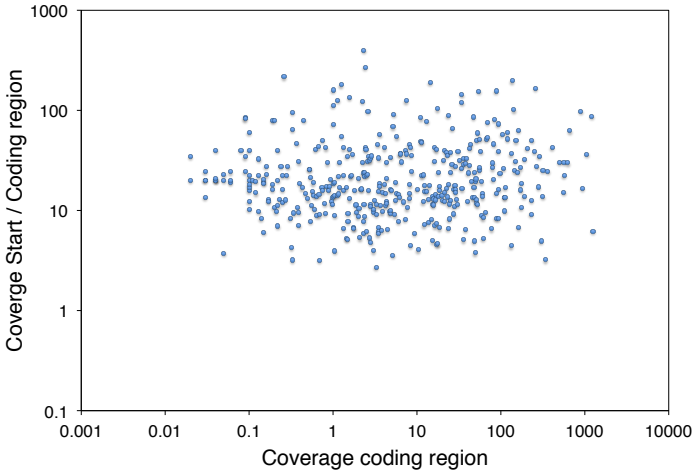
"Strength" of the evidence of expression is measured by an "Expression Index".

Expression of published and newly-predicted genes in *Pseudomonas aeruginosa*

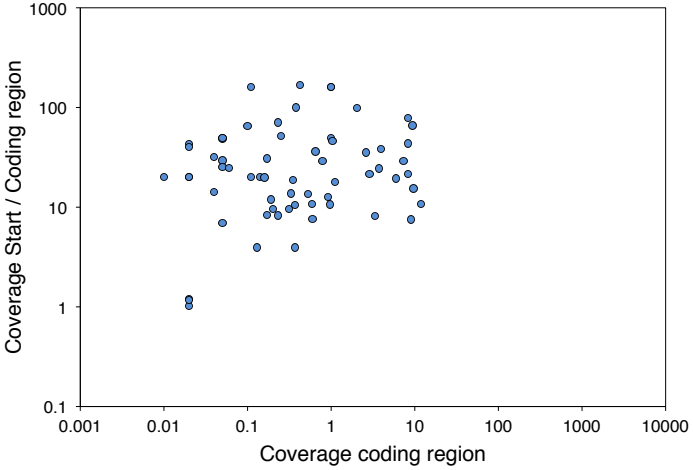
Long published genes



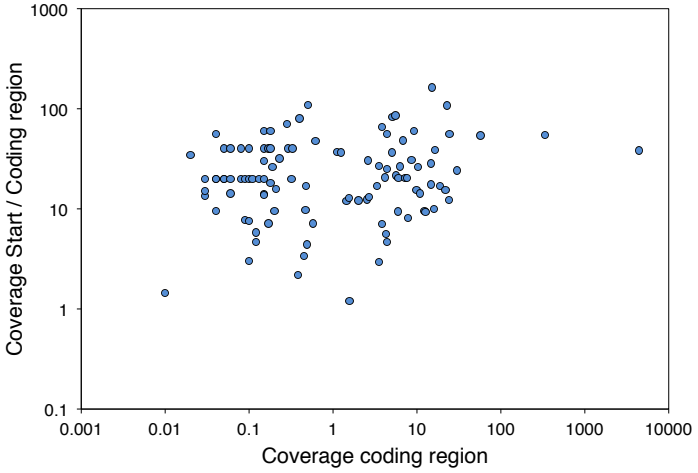
Short published genes



Long newly identified ORFs



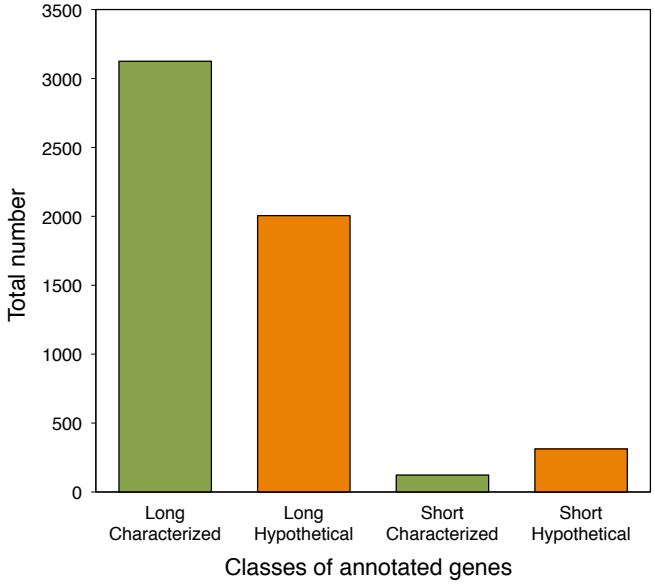
Short newly identified ORFs



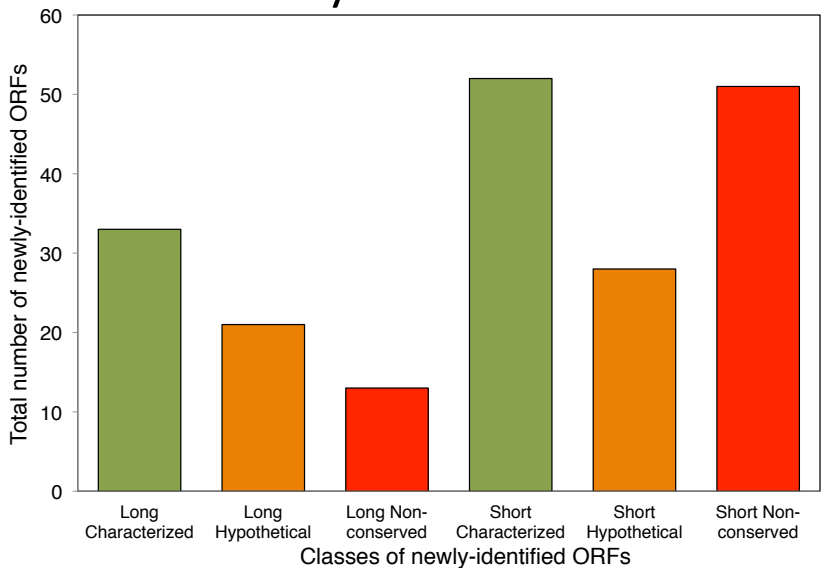
* Coverage normalized by the number of positions

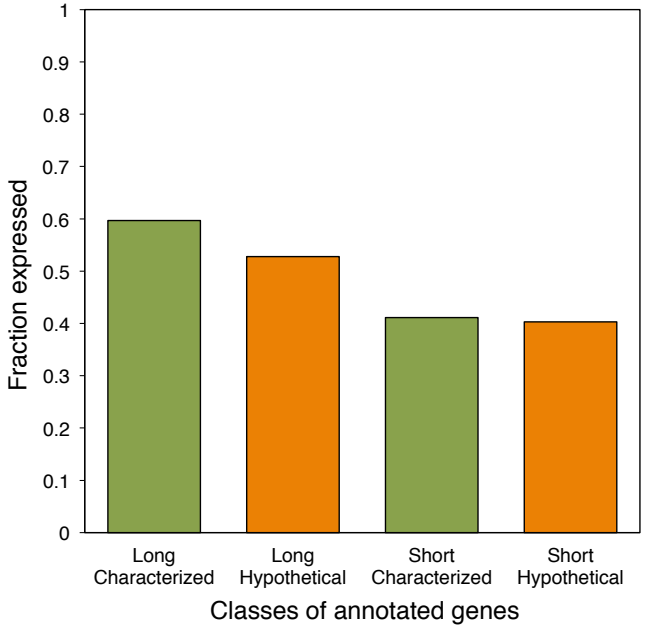
Expression of predicted genes by length and conservation classes

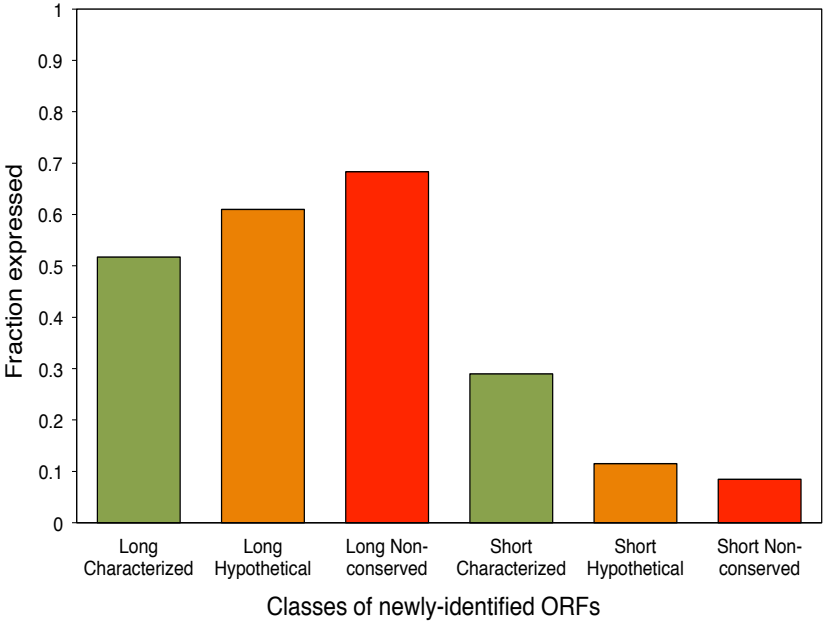
Published annotation



Newly identified ORFs

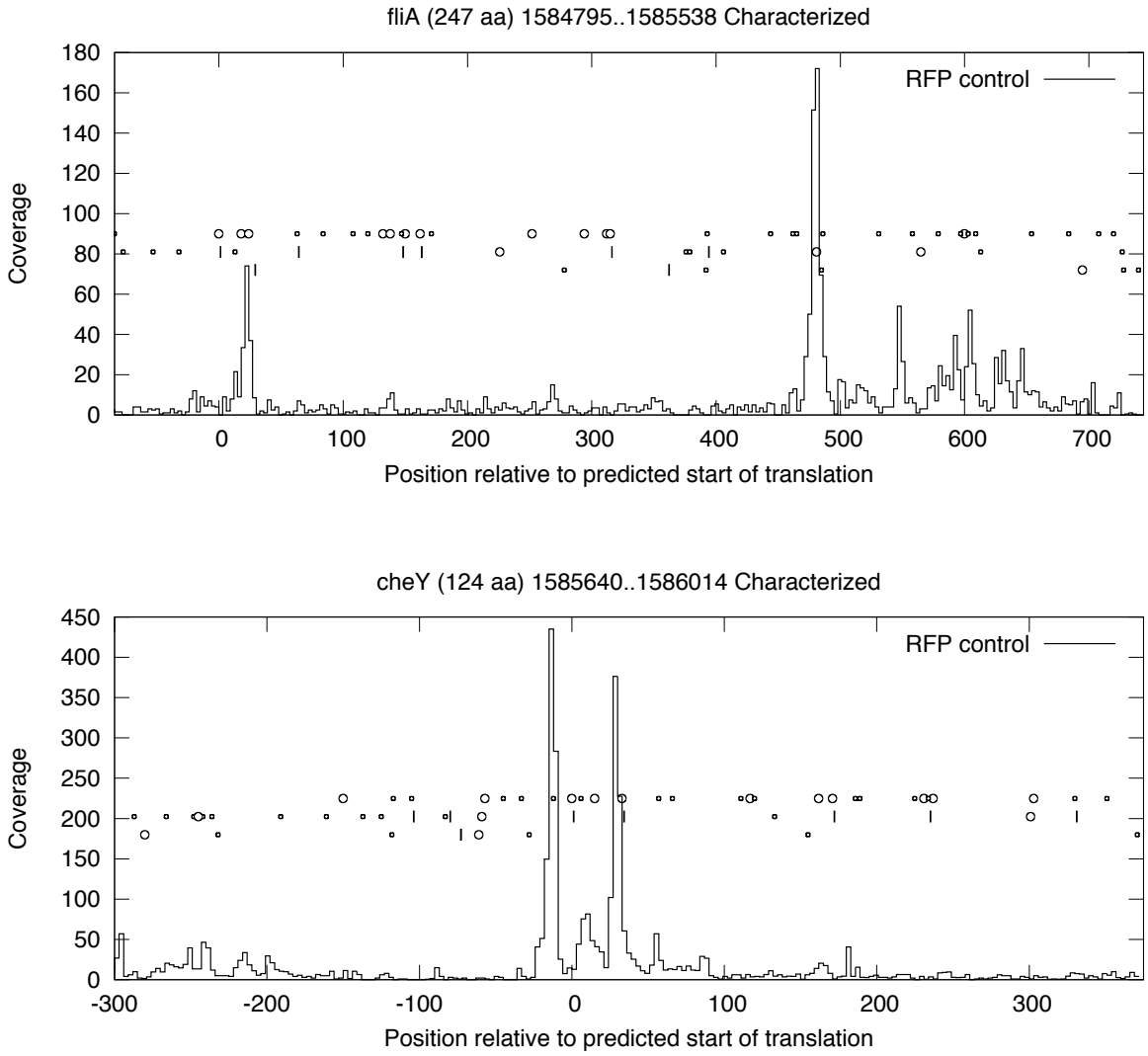






ORFs with Expression Index ≥ 12.0

Identification of translation-initiation sites by ribosome-footprinting



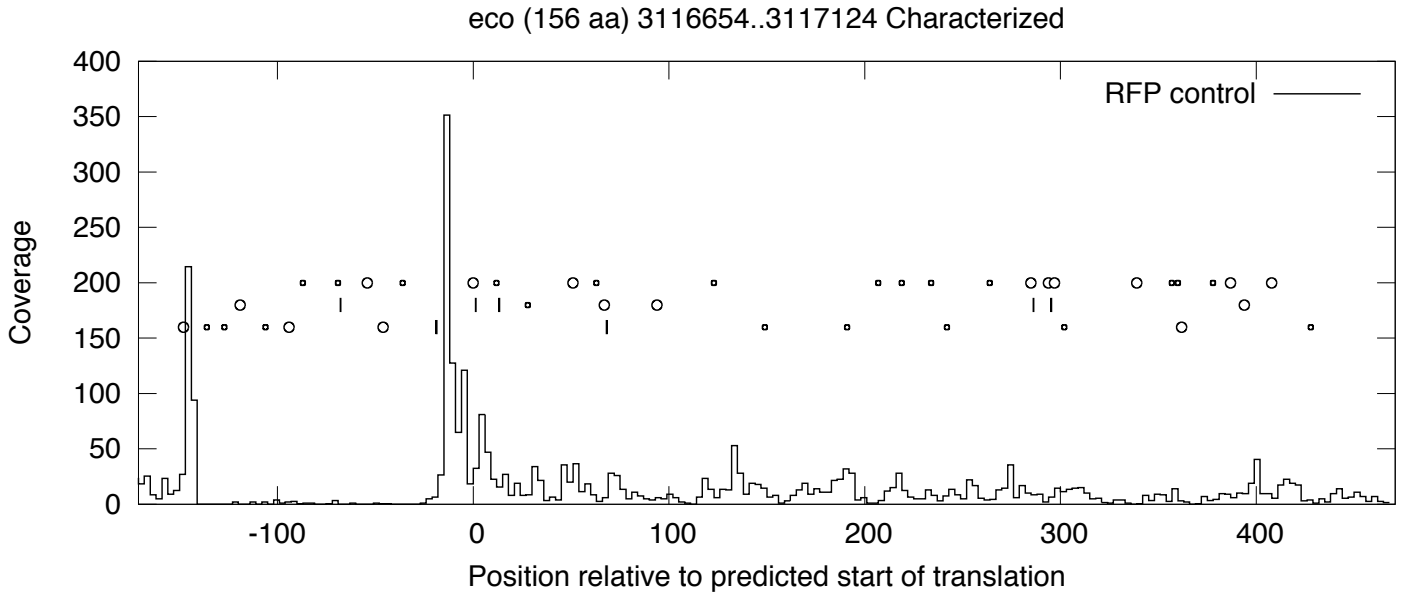
FliA, sigma factor of RNA polymerase for flagellin gene transcription.
CheY is involved in transmission of sensory signal to the flagellar motor.

Start of translation identification by RFP read accumulation

	Annotated	Newly identified
Same start	0.850	0.778
Different start	0.150	0.222

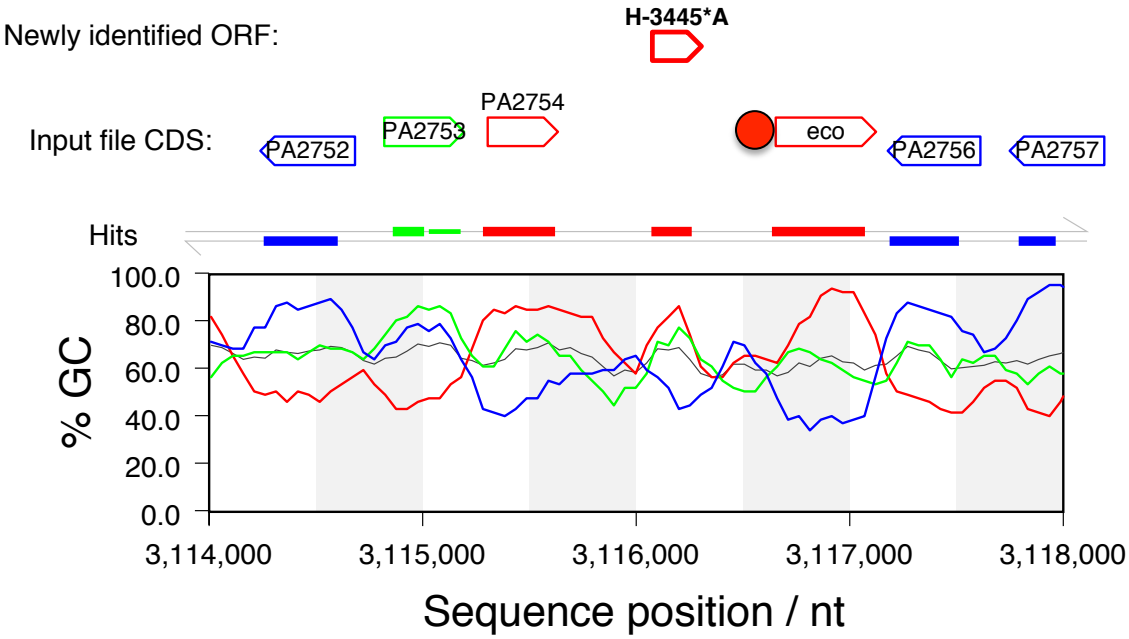
Ribosome footprints confirm the predicted start of translation for 85% of the genes, and of 78% of the newly-identified ORFS, among those with evidence of translation.

Identification of new genes by ribosome-footprint evidence



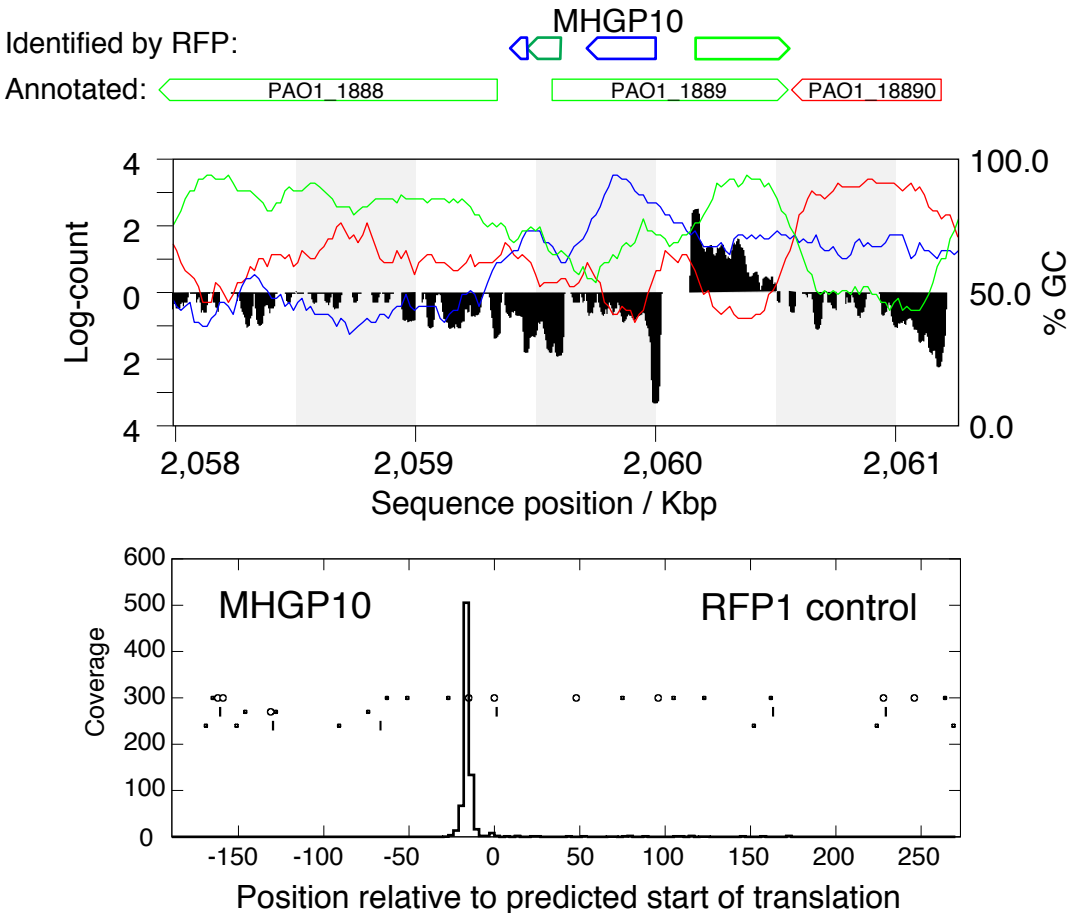
A new gene is found to be expressed 5’ of the gene *eco* for Ecotin, a protease inhibitor localized to the periplasmic space.

Identification of new genes by ribosome-footprint evidence



The newly-identified translated ORF (red circle) corresponds to a region of weak compositional 3-base periodicity.

Ribosome footprint coverage in *P. aeruginosa*



Examples of RFP-based gene discovery in *P. aeruginosa* PAO1 showing relation with S-profiles and annotated genes.

Thanks to

Lab members

- Steve Oden – Postdoctoral associate. Development of gene finding methods and software, gene content analysis in human and prokaryotes.
- Eric Hernandez – Programmer.
- Dr. Anna Picca– Postdoctoral associate. RNA-seq and ribosome profiling
- Dr. Ying Zhang – Postdoctoral associate. RNA-seq

Collaborators

- Dr. Shouguang Jin (Molecular Genetics and Microbiology). *P. aeruginosa*
- Dr. Silvia Tornaletti (Medicine). Transcription and RNA.

Sequencing facility and support

- Dr. Rolf Renne and MGM.
- Dr. Jianhong Hu (Research Scientist)

Funding

- NIH R01 GM087485-01A2
- MGM, Genetics Institute, College of Medicine.