Prediction of highly expressed genes in microbes based on chromatin accessibility

BMC Molecular Biology 2007, 8:11
Willenbrock & Ussery

Overview
by Age Tats

Journal Club Feb 26th, 2007

Background

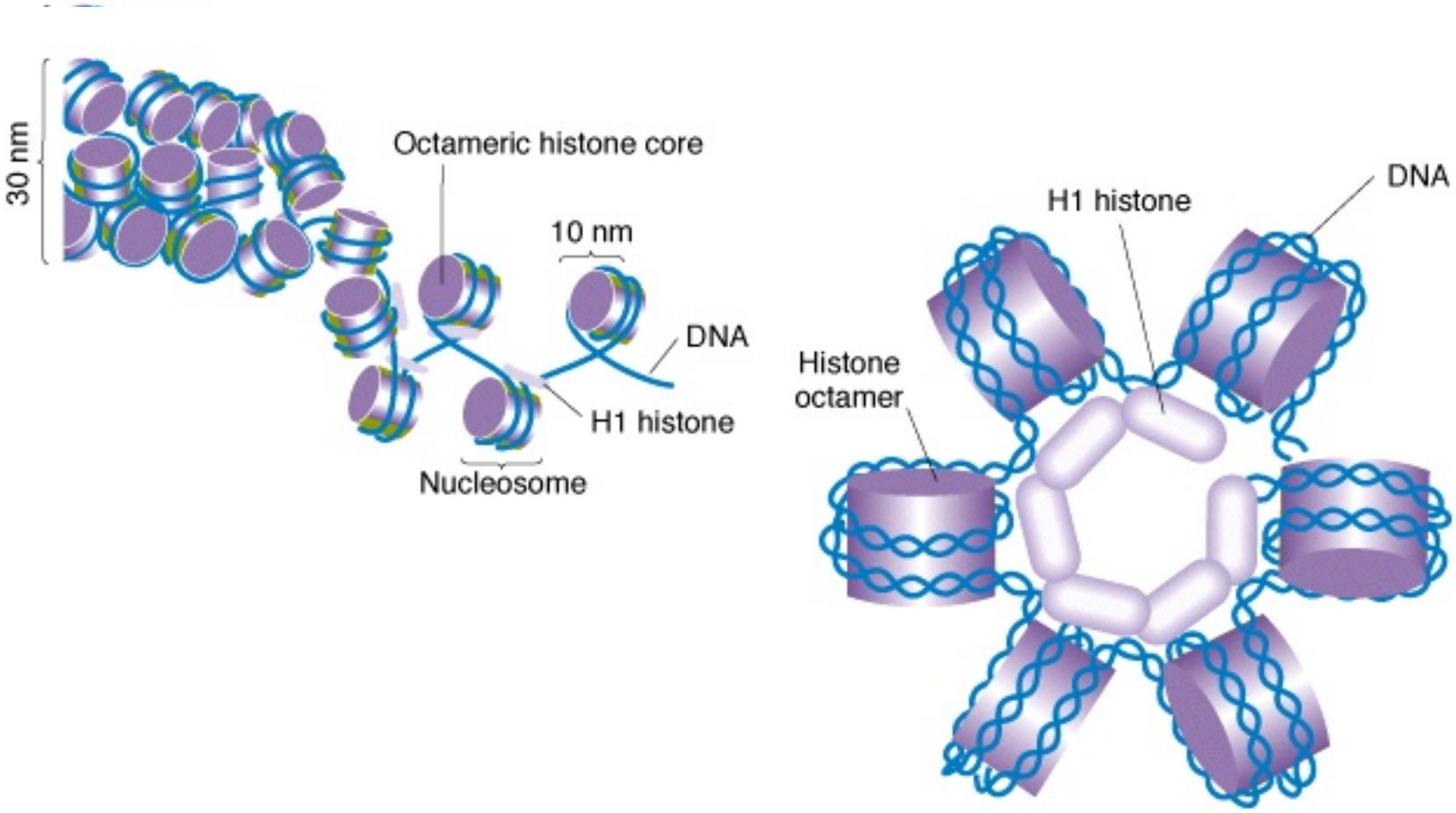
- Gene expression is dependent on chromatin structure in eukaryotes.

- The '**position preference**'(**PP**) measure is a DNA structural measure, which reflects the preference of a given di-/trinucleotide for being found in a region where the DNA minor groove faces either towards or away from the nucleosome histone core.

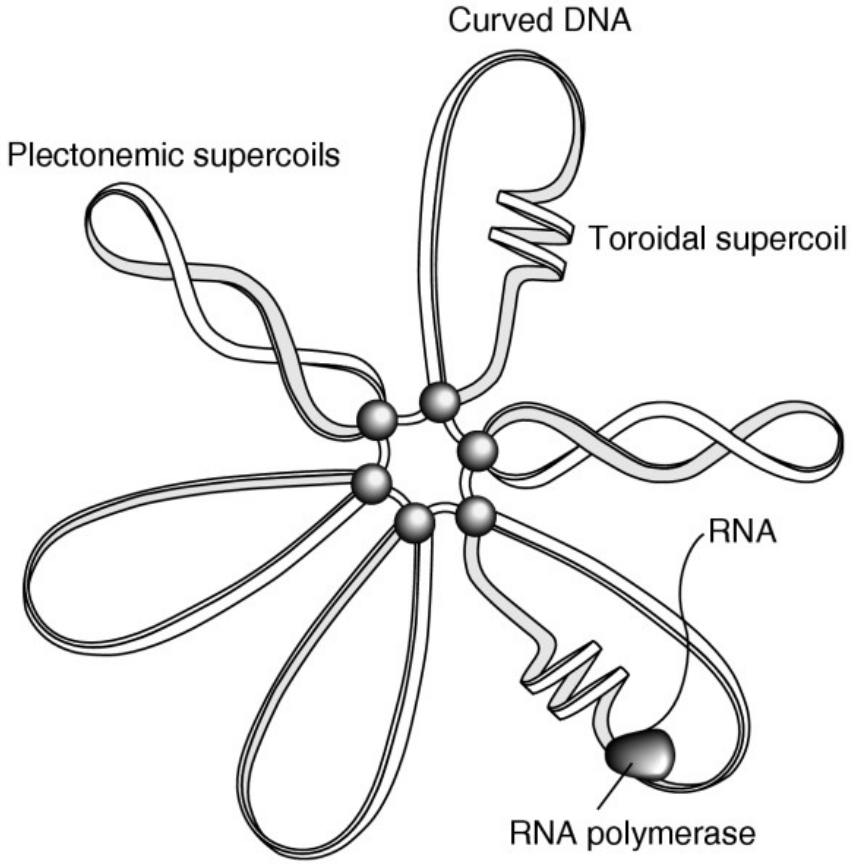
- High absolute PP reflects a high preference for nucleosomes, while low absolute PP reflects di-/trinucleotides which tend to exclude nucleosomes.

- The PP measure also describes a more general structural property of DNA – that is, how easily can it be wrapped around chromatin proteins.

Nucleosome structure in eukaryotes



Prokaryotes do not have nucleosomes but they also have chromatin, and the DNA is compacted to similar levels (i.e., more than 1000x) in both prokaryotes and eukaryotes.



An illustration of DNA supercoiling domains in the *E. coli* chromosome.

Willenbrock & Ussery 2004

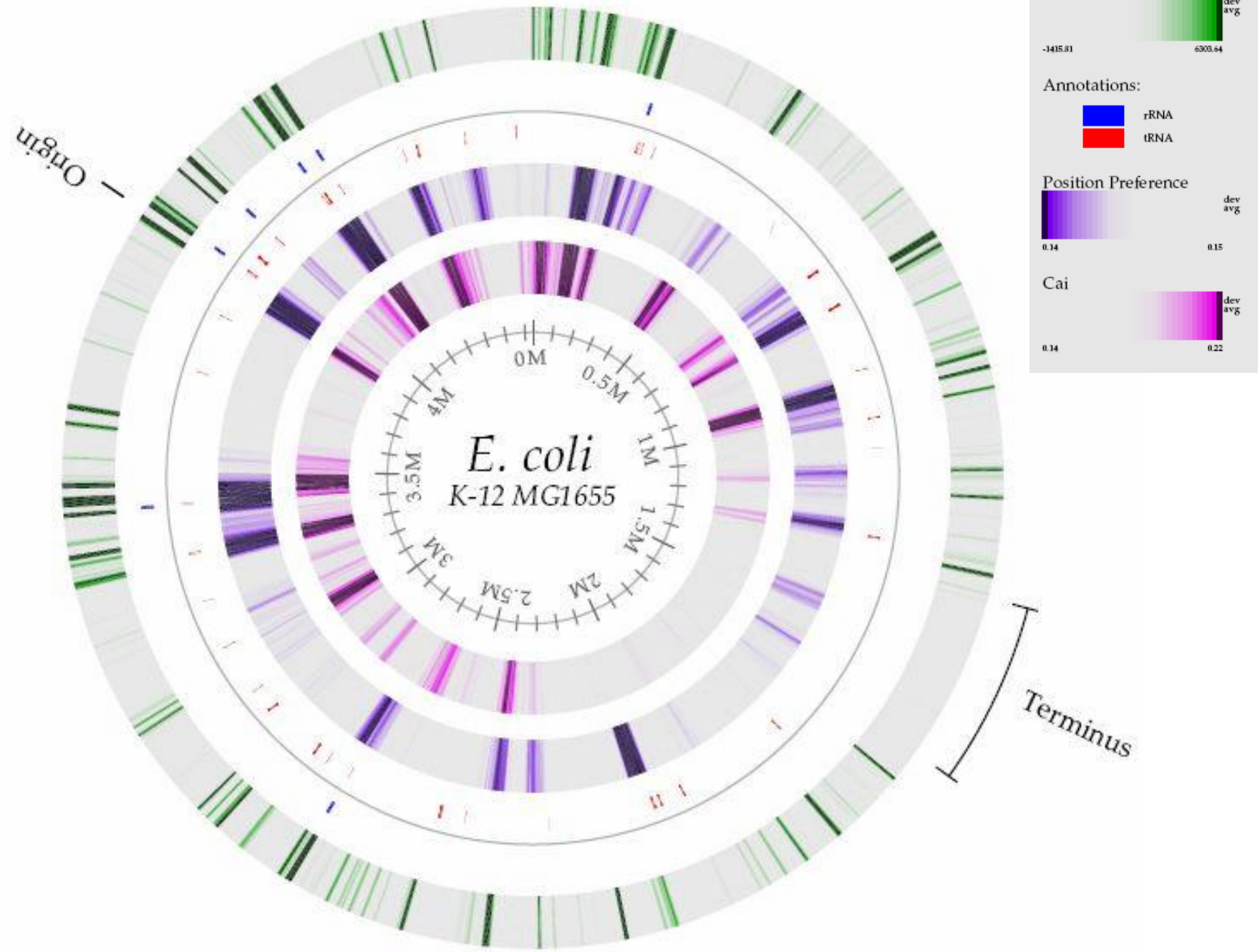
Productive binding of DNase I requires DNA to be bent.

DNase I interacts with a 6 bp contact surface of the minor groove and bends the DNA molecul away from the enzyme, towards the major groove. Therefore, base sequences that are flexible or inherently bent towards the major groove should be more accessible to DNase I cleavage.

Table I. DNA bending and/or bendability parameters as revealed by DNase I: parameters of trinucleotide steps[a]

Trinucleotide step	No. of occurrences in dataset	DNase I-derived trinucleotide parameter (ln *p*)
AAT/ATT	89	−0.280
AAA/TTT	278	−0.274
CCA/TGG	45	−0.246
AAC/GTT	81	−0.205[b]
ACT/AGT	77	−0.183[b]
CCG/CGG	73	−0.136
ATC/GAT	112	−0.110
AAG/CTT	110	−0.081
CGC/GCG	84	−0.077
AGG/CCT	101	−0.057
GAA/TTC	117	−0.037
ACG/CGT	84	−0.033
ACC/GGT	87	−0.032
GAC/GTC	81	−0.013
CCC/GGG	141	−0.012
ACA/TGT	52	−0.006[b]
CGA/TCG	84	−0.003
GGA/TCC	71	0.013
CAA/TTG	74	0.015[b]
AGC/GCT	35	0.017
GTA/TAC	83	0.025
AGA/TCT	127	0.027
CTC/GAG	102	0.031[b]
CAC/GTG	55	0.040
TAA/TTA	99	0.068[b]
GCA/TGC	34	0.076
CTA/TAG	64	0.090
GCC/GGC	57	0.107
ATG/CAT	71	0.134[b]
CAG/CTG	61	0.175
ATA/TAT	80	0.182
TCA/TGA	127	0.194

Brukner *et al* 1995


Expression
dev avg
-1415.81
6303.64
Annotations:
rRNA
tRNA
Position Preference
dev avg
0.14
0.15
Cai
dev avg
0.14
0.22
Origin
Terminus
0M
0.5M
1M
1.5M
2M
2.5M
3M
3.5M
4M
E. coli
K-12 MG1655

- Codon adaptation index is highly correlated with the expression level in fast growing bacteria, but
  - it cannot consider tRNAs, ribosomal RNAs and other non-coding RNAs;
  - it is less effective predictor in slow growing organisms;
  - it requires the identification of a representative subset of highly expressed genes in an organism (problematic for more distant microbes).
- Alternative - using the PP for the prediction of highly expressed genes in microbial genomes?

- The average position preference for ribosomal protein encoding genes is lower than for other protein encoding genes (Wilcoxon P-value 4e-11).

- rRNAs, tRNAs, and miscellaneous RNAs have significantly lower position preference values than translated genes (P-value = 6e-34).

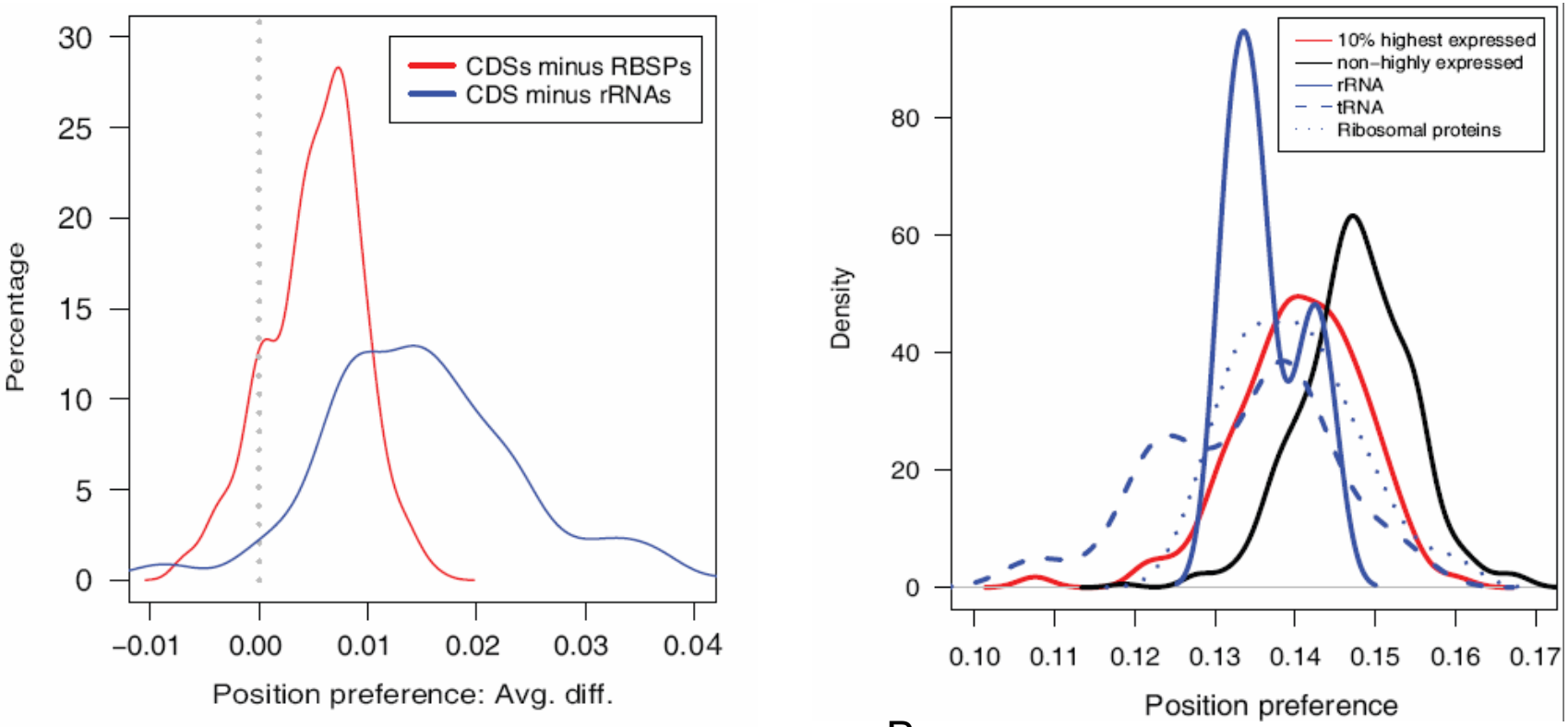


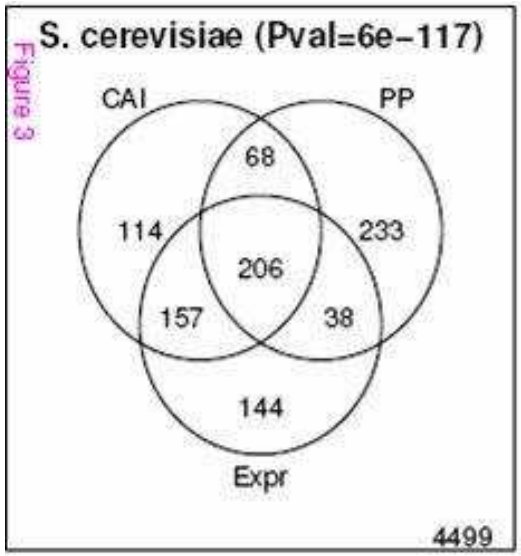
Figure 2. Gene density plots. (A) Density plot of position preference differences for 328 microbial genomes. Differences between mean position preference of translated coding sequences (CDSs) and ribosomal proteins (red) or between mean position preference of CDSs and ribosomal RNA (rRNA) (blue). Most microbial genomes CDSs have a higher mean position preference values than ribosomal proteins and rRNA (mean above 0). (B) Position preference densities for the 10% most highly expressed genes, non-highly expressed genes, rRNAs, tRNAs and ribosomal proteins in *E. coli*.

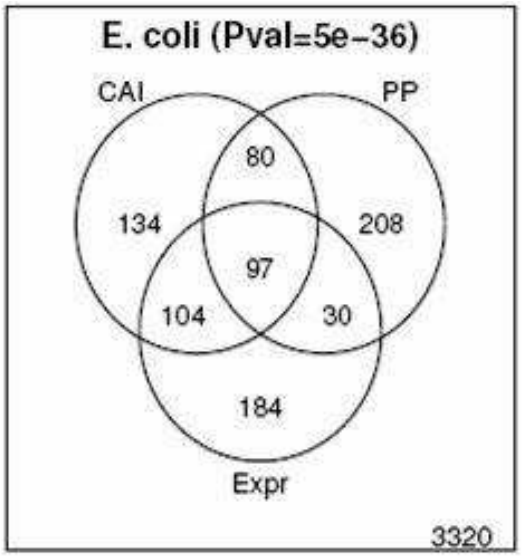
- There was no significant correlations between CAI triplet weights and PP scores.

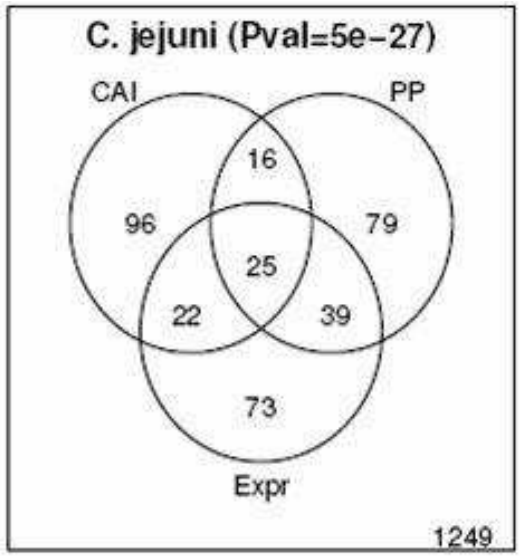
- The correlation between CAI weights and PP triplet values did not increase for fast replicating bacteria, indicating that PP may be a useful supplement for predicting highly expressed non-translated genes even in slow-growing microbes.

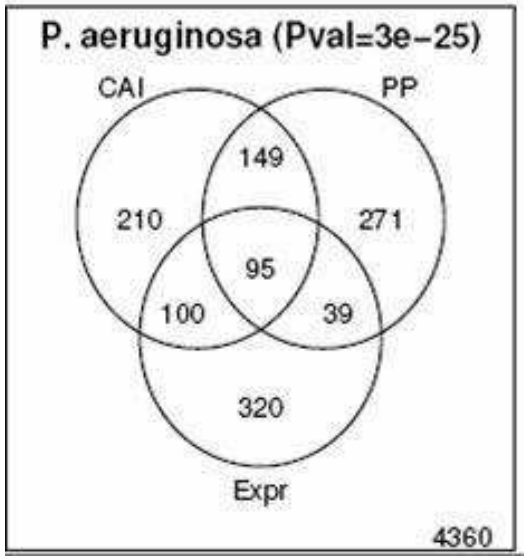
- The PP measure could be useful for identifying rRNA, tRNA and other non-coding RNA genes in pre-annotated DNA sequences, because those genes tend to have lower PP than the genomic average.

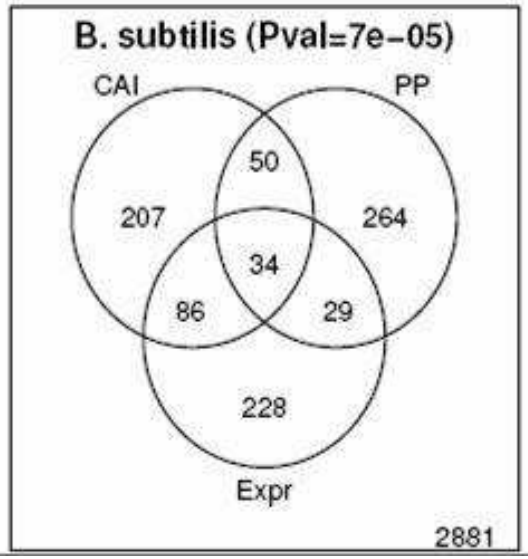
Figure 3


S. cerevisiae (Pval=6e−117)
CAI
PP
68
114
233
206
157
38
144
Expr
4499


E. coli (Pval=5e−36)
CAI
PP
80
134
208
97
104
30
184
Expr
3320


C. jejuni (Pval=5e−27)
CAI
PP
16
96
79
25
22
39
73
Expr
1249


P. aeruginosa (Pval=3e−25)
CAI
PP
149
210
271
95
100
39
320
Expr
4360


B. subtilis (Pval=7e−05)
CAI
PP
50
207
264
34
86
29
228
Expr
2881

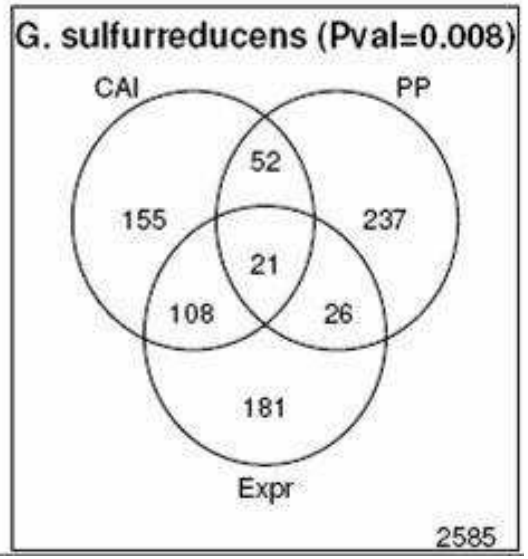
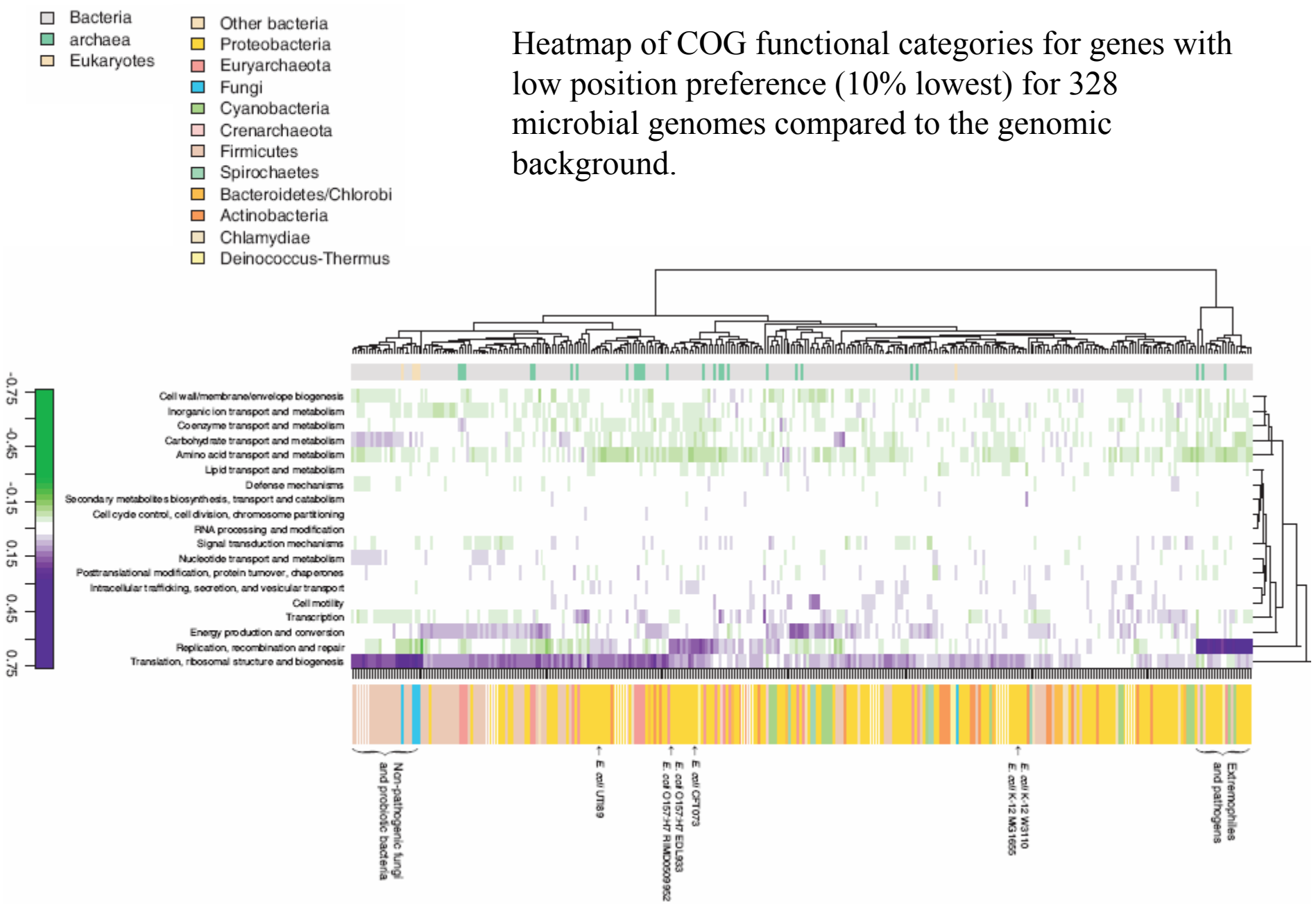

G. sulfurreducens (Pval=0.008)
CAI
PP
52
155
237
21
108
26
181
Expr
2585

Table 1. Predicted highly expressed non-translated *E. coli* genes by the position preference measure.

Gene	*bnumber*	*Type*	*Gene expression rank*	*PP*
asnT	**b1977**	**tRNA**	**122**	**0.1076**
asnW	**b1984**	**tRNA**	**121**	**0.1076**
asnU	**b1986**	**tRNA**	**101**	**0.1076**
asnV	**b1989**	**tRNA**	**127**	**0.1076**
thrV	b3273	tRNA	898	0.1154
valW	b1666	tRNA	1544	0.1172
valT	**b0744**	**tRNA**	**114**	**0.1212**
valZ	**b0746**	**tRNA**	**130**	**0.1212**
valX	**b2402**	**tRNA**	**124**	**0.1212**
valY	**b2403**	**tRNA**	**109**	**0.1212**
aspU	**b0206**	**tRNA**	**219**	**0.1215**
aspV	**b0216**	**tRNA**	**233**	**0.1215**
aspT	**b3760**	**tRNA**	**150**	**0.1215**
serU	b1975	tRNA	1420	0.1221
leuU	b3174	tRNA	3814	0.1229
valV	b1665	tRNA	2022	0.1233
thrU	**b3976**	**tRNA**	**66**	**0.1235**
ileU	**b3277**	**tRNA**	**274**	**0.1239**
ileT	**b3852**	**tRNA**	**236**	**0.1239**
argQ	b2691	tRNA	455	0.1241
argZ	b2692	tRNA	765	0.1241
argV	**b2694**	**tRNA**	**401**	**0.1241**
ffs	b0455	misc_RNA	686	0.1255
trpT	b3761	tRNA	1483	0.1276
pheV	b2967	tRNA	3047	0.1285
pheU	b4134	tRNA	1995	0.1285
leuT	**b3798**	**tRNA**	**171**	**0.1287**
leuQ	b4370	tRNA	502	0.1287
leuP	**b4369**	**tRNA**	**404**	**0.1294**
selC	b3658	tRNA	617	0.1295
thrT	**b3979**	**tRNA**	**143**	**0.1300**
serW	**b0883**	**tRNA**	**196**	**0.1310**
rrsH	**b0201**	**Ribosomal**	**15**	**0.1318**
rrsA	**b3851**	**Ribosomal**	**1**	**0.1318**
tyrU	**b3977**	**tRNA**	**172**	**0.1318**
rrsG	**b2591**	**Ribosomal**	**6**	**0.1319**
rrsC	**b3756**	**Ribosomal**	**2**	**0.1320**
rnpB	**b3123**	**misc_RNA**	**73**	**0.1332**
leuZ	b1909	tRNA	920	0.1338
rrfF	**b3272**	**Ribosomal**	**14**	**0.1338**
rrlH	**b0204**	**Ribosomal**	**4**	**0.1348**
metU	**b0666**	**tRNA**	**309**	**0.1351**
metT	**b0673**	**tRNA**	**229**	**0.1351**
rrlA	**b3854**	**Ribosomal**	**12**	**0.1351**
rrlC	**b3758**	**Ribosomal**	**18**	**0.1352**
rrlE	**b4009**	**Ribosomal**	**5**	**0.1352**
rrlG	**b2589**	**Ribosomal**	**3**	**0.1354**
rrlD	**b3275**	**Ribosomal**	**30**	**0.1355**
lysT	**b0743**	**tRNA**	**123**	**0.1357**
lysW	**b0745**	**tRNA**	**142**	**0.1357**
lysY	**b0747**	**tRNA**	**94**	**0.1357**
lysZ	**b0748**	**tRNA**	**131**	**0.1357**
lysQ	**b0749**	**tRNA**	**113**	**0.1357**
dicF	b1574	RNA; Cell	3004	0.1361
proK	b3545	tRNA	3076	0.1367



Heatmap of COG functional categories for genes with low position preference (10% lowest) for 328 microbial genomes compared to the genomic background.

Conclusions

- Absolute gene expression levels are highly correlated with low PP in multiple microbial genomes.
- PP may be exploited for predicting the expression of non-translated genes and highly expressed genes in slow growing microbes.
- Genes often encoded by DNA with low position preference values were mostly involved in ‘translation, ribosomal structure and biogenesis’, ‘energy production and conversion’, and transcription.
- For pathogens and microbes living in extreme environments, the predominant functional category was ‘replication, recombination and repair’.

References

- Willenbrock & Ussery (2007) Prediction of highly expressed genes in microbes based on chromatin accessibility. *BMC Mol Biol* 8:11

- Brukner et al (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J*, 14:1812-1818

- Willenbrock & Ussery (2004) Chromatin architecture and gene expression in *Escherichia coli. Gen.Biol.* 5:252

- Postow et al (2004) Topological domain structure of the *Escherichia coli* chromosome. *Genes and Dev*. 18: 1766-1779