Electronic Supplementary Material (ESI) for Molecular BioSystems

Supplementary Materials to Zarei et al. "Gene silencing and large-scale domain structure of the E. coli genome"

January 12, 2013

Description of Supplementary Figures

Mock-IP control

We analyzed the large-scale organization of H-NS binding regions data from Kahramanoglou *et al*[1] as a function of growth phase. These binding regions were obtained by comparing the number of reads mapped to each region, normalized by the total number of reads obtained for that sample, with the corresponding value from the Mock-IP experiment using a binomial test[1]. The Mock-IP was available only for the mid-exponential phase sample, where the gene dosage effect is highest. Since in stationary phase the gene dosage effect is small or absent, the applied filter might create a bias at large scales towards the stationary phase dataset. In order to control for this, we compared the results from binding regions found before Mock-IP control and after the control. Supplementary Figure 1 shows the sliding window sums of the length and number of H-NS binding regions along the genome for binding regions obtained before and after Mock-IP control in the early-exponential (EE), mid-exponential (ME), transition to stationary (TS) and stationary (S) phases of growth (window size=500 Kb). The total length is roughly unaffected by the Mock-IP control, except for minor changes around Ori. The change around Ori is larger for the number of binding regions, due to the presence of a large number of short spurious regions in the exponential and early exponential phase data. However, the results in a wide region around Ter are robust to the Mock-IP filter for all data sets, for both the number and the total length of bound regions, indicating that the observed increase of total length of polymerized H-NS is not a spurious result of the Mock-IP filter.

Growth-phase dependent H-NS binding regions

Supplementary Figure 2 shows the result of the linear aggregation analysis for genes located in the H-NS binding regions from early exponential to stationary phase. One can see that the genes located in the H-NS binding regions in early exponential phase are clustered close to the macrodomain boundaries and that this cluster pattern is preserved as the cells go from early exponential to stationary phase. In stationary phase however, new clusters appear inside the Ter macrodomain.

H-NS binding regions

We performed the clustering analysis for the list of genes associated to regions of H-NS binding obtained by Oshima *et al*[2] using a high-density oligonucleotide chip (ChIP-chip analysis). The experiment was performed with the W3110 strain of E. coli K12 grown in LB medium. The H-NS genome-wide binding was assessed on exponentially growing cells. The W3110 strain is very similar to the MG1655 strain and in their data analysis the genome coordinates of MG1655 were used. The analysis shows similar clusters to the results presented in the main text (Supplementary Figure 3).

Horizontally transferred genes found using phylogenetic information are not clustered along the genome.

The most accurate methods of HGT identification rely on phylogenetic tree information because it allows one to estimate the relative date of the transfers. The study of Lercher and Pal[3] defined a list of horizontally transferred genes according to the number of branches that separate *E. coli* K-12 from the node of the tree where the transfer is detected. We divided the list of transferred genes into two groups, “recent” and “old” transfers. In order to produce a sufficiently large data set, we defined horizontally transferred genes with an age less than 5 branches as the recent transfers. The results of this analysis show that recently transferred genes are clustered along the genome, while the whole set of horizontally transferred genes is distributed uniformly along the genome, as shown by Supplementary Figure 4.

Genes overlapping with heEPODs are clustered along the genome.

Supplementary Figure 5 shows that the genes overlapping with heEPODs are also clustered at different observation scales along the E. coli genome. The clusters of genes overlapping with heEPODs include ribosomal and flagella genes, which

are highly transcribed, as shown by Supplementary Table 6. Comparison with Figure 4 in the main text suggests that the clusters of highly transcribed EPODs are located at the larger distances from macrodomain boundaries with respect to the tsEPODs clusters.

Most of the clusters found have a lower mean expression level than the genomic average.

Supplementary Figure 6 shows the ratio of the average expression level for each cluster and the average expression level of all genes along the genome. Microarray data sets are extracted from the ASAP database[4] at https://asap.ahabs.wisc.edu. The data set used for this analysis has the transcript copy number of 4220 genes in wild-type E. coli. The strain MG1655 was cultured in MOPS minimal with glucose at 37 degrees to log phase (OD600=0.2). Most of the clusters are expressed at a significantly lower level than the average of E. coli transcriptome. There are some genes that are highly expressed and affect the average expression level for the clusters. For example, ompt and nmpC are outer membrane related genes which are highly expressed in the data set that we used. On the other hand, nmpC is reported to be silent in E. coli K-12[5]. rpmB and rpmG are ribosomal proteins, which are highly transcribed. rpmB and uspA are associated to adaptation to stress.

Description of Supplementary Tables

Coordinates of the macrodomains and chromosomal sectors

Supplementary Tables 1 and 2 show the coordinates of the macrdomains and chromosomal sectors used in this study.

Summary of the clustering for H-NS binding regions and horizontally transferred genes

Supplementary Tables 3,4,5 summarize the clustering results for different data sets.

heEPODs clusters are enriched by flagella and ribosomal genes.

Supplementary Table 6 shows the results of a hypergeometric test for the enrichment of MultiFun functional categories within the lists of genes located in heEP-

ODs clusters. The results indicate that, as expected, these clusters are enriched by highly expressed genes, such as flagella, ribosomal proteins, rRNA and tRNA.

Intersection between data sets

Supplementary Tables 8 and 9 show the intersection between different gene lists used in this analysis. The result of a hypergeometric test (assessing the statistical significance of the intersection) can be found in Supplementary Table 8. The overlap of the lists is large, but the lists do not coincide, making the coincidence of the clusters nontrivial.

Summary of the clustering for pseudo-genes

Supplementary Table 7 summarizes the clustering results for pseudo-genes.

Some of the long non-lethal deletions and prophages are close to the macrodomain boundaries.

Supplementary Table 10 represents the correlation between the position of clusters found in our analysis and the position of long non-lethal deletions or prophages. Some of the clusters correlate with the prophages and long non-lethal deletions.

Most of the clusters, that are reported in Table 1, are enriched by Prophage related functions.

Supplementary Table 11 shows the results of systematic hypergeometric testing for enrichment of MultiFun functional categories within the lists of functional genes located in each cluster that is reported in Table 1. Considering clusters close to or far from the macrodomain boundaries separately, we did not see particular differences between the two groups. Not surprisingly, most of the clusters are enriched by prophage-related functions. The clusters with coordinates 1-17-1.25 Mb and 3.62-3.68 Mb show enrichment in the term membrane. The clusters within coordinates 3.6-4.3 Mb show enrichment in the terms membrane, surface antigens, lipopolysaccharide synthesis, carbon compounds, and anaerobic respiration.

References

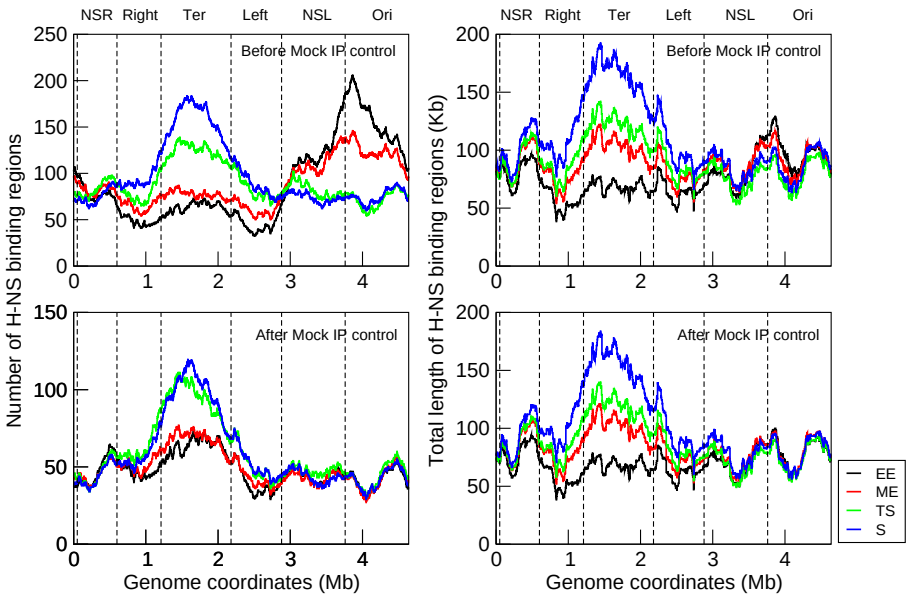
[1] C. Kahramanoglou, A. S. N. Seshasayee, A. I. Prieto, D. Ibberson, S. Schmidt, J. Zimmermann, V. Benes, G. M. Fraser and N. M. Luscombe, *Nucleic Acids Research*, 2011, **39**, 2073–2091.

[2] T. Oshima, S. Ishikawa, K. Kurokawa, H. Aiba and N. Ogasawara, *DNA research an international journal for rapid publication of reports on genes and genomes*, 2006, **13**, 141–153.

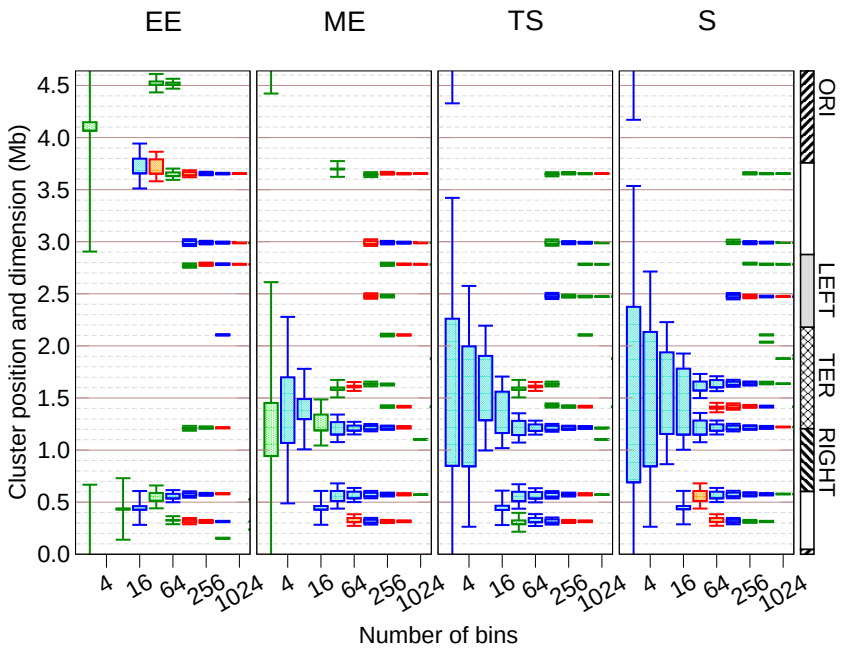
[3] M. J. Lercher and C. Pál, *Molecular biology and evolution*, 2008, **25**, 559–567.

[4] T. E. Allen, M. J. Herrgard, M. Liu, Y. Qiu, J. D. Glasner, F. R. Blattner and B. . Palsson, *J Bacteriol*, 2003, **185**, 6392–6392.

[5] M. S. Hindahl, G. W. Crockford and R. E. Hancock, *J Bacteriol*, 1984, **159**, 1053–1055.



Supplementary Figure 1. The observed growth of H-NS binding regions in the Ter region in stationary growth phase is unaffected by the Mock-IP control. Total length (right panels) and number of H-NS binding regions (left panels) along the genome, before and after Mock-IP control in early-exponential (EE), mid-exponential (ME), transition to stationary (TS) and stationary (S) phases of growth. Because of the gene dosage effect in the exponential phase, many spurious binding regions around Ori appear, while the Ter region is weakly affected. Additionally, the data for the stationary and transition to stationary phase are weakly affected by the Mock-IP control along the whole genome. This indicates that the observed average growth of H-NS binding regions around Ter is not a consequence of the Mock-IP control.



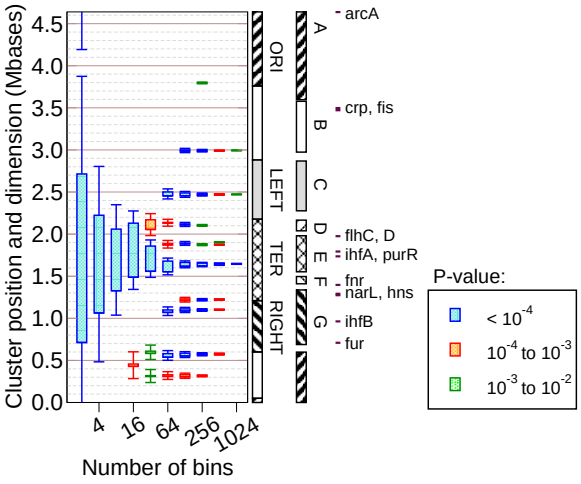
Supplementary Figure 2. The number of clusters for genes located in the binding regions increase while cells progress from early-exponential to stationary growth phase. Cluster diagrams of genes located in the H-NS binding regions in the early-exponential (EE), mid-exponential (ME), transition to stationary (TS) and stationary (S) phases of growth. The clusters close to the boundaries of the Ter macrodomain are preserved during different growth phases, while new clusters emerge as the cells progress from early-exponential to the stationary phase.

Supplementary Table 1. Macrodomain coordinates.

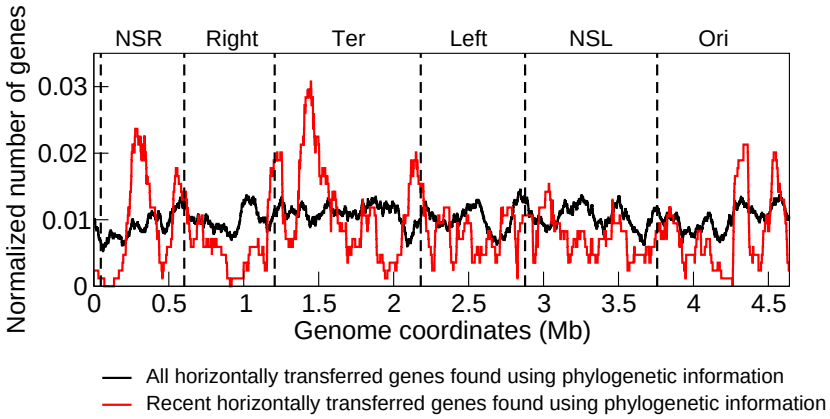
Ori	NSR	Right	Ter	Left	NSL
3.76-0.05	0.05-0.60	0.60-1.21	1.21-2.18	2.18-2.88	2.88-3.76

Supplementary Table 2. Chromosomal sectors.

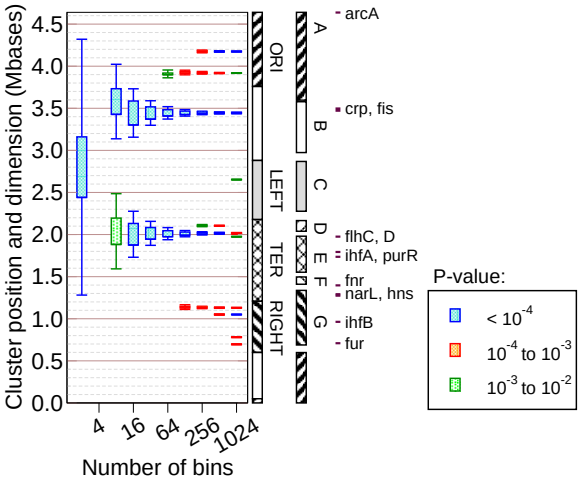
A	B	C	D	E	F	G
3.59-0.59	2.97-3.57	2.27-2.86	2.03-2.16	1.54-1.97	1.40-1.49	0.68-1.33



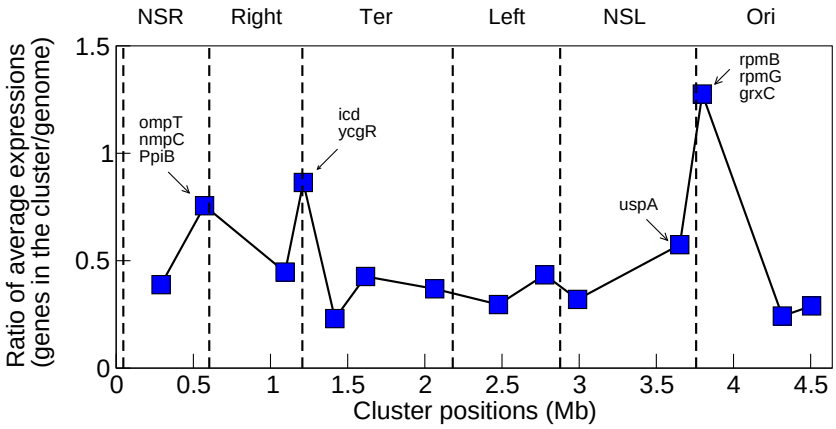
Supplementary Figure 3. The H-NS binding regions found by Oshima *et al.*[2] are clustered along the genome. Diagram of the statistically significant H-NS bound genes clusters. The plot shows that there are significant clusters close to the boundaries of macrodomains or Mathelier sectors. The position of the clusters correlate with the clusters shown in Figure 2 of the main text.



Supplementary Figure 4. Horizontally transferred genes are distributed uniformly along the genome. Comparison of the sliding-window histograms (window size 100 Kb) of the whole list of horizontally transferred genes found using phylogenetic information and of the transferred genes that occur before divergence between E. coli and Salmonella[3]. The figure shows that only recently transferred genes are clustered along the genome.



Supplementary Figure 5. Genes overlapping with heEPODs are clustered along the genome. Diagram of the statistically significant clusters. The plot shows that there are significant clusters of genes overlapping with heEPODs along the genome. Three of these clusters include flagella and ribosomal genes.



Supplementary Figure 6. Most of the clusters found have a lower average expression level compared to the average for the E. coli transcriptome. Ratio of the average expression level for each cluster found analysing different datasets and the average expression level of all genes along the genome. The estimated transcript copy number of genes in wild-type E. coli were obtained from the ASAP data base[4], and refers to MG1655 cultured in MOPS minimal medium with glucose at 37 degrees to log phase (OD600=0.2). The genes in the clusters that are more highly expressed are labeled in the figure. Most of the other clusters have a significantly lower mean expression level compared to the genomic average.

Supplementary Table 3. H-NS Binding regions in the early exponential growth phase.

Number of clusters (bin-size=128)	6
Total length of binding regions	648991
Total length of binding regions in the clusters	343221
Total number of genes in the H-NS binding regions	441
Number of H-NS binding genes in the clusters	112
Number of annotated genes in the clusters	374

Supplementary Table 4. H-NS Binding regions in the stationary growth phase.

Number of clusters (bin-size=128)	7
Total length of binding regions	944735
Total length of binding regions in the clusters	438379
Total number of genes in H-NS binding regions	748
Number of H-NS binding genes in the clusters	217
Number of annotated genes in the clusters	510

Supplementary Table 5. Horizontally transferred genes found using nucleotide composition.

Number of clusters (bin-size=128)	11
Total number of HGT	350
Number of HGT in the clusters	201
Number of annotated genes in the clusters	773

Supplementary Table 6. Enrichment analysis for functional categories of genes located in the heEPODs clusters. List of significant MultiFunfunctional categories for genes found within the heEPODs clusters. The MultiFun data annotate 3382 genes out of 4667, and out of 406 genes located in the heEPODs clusters, 326 are annotated. The first column shows the MultiFun class number. The second column shows the number of genes in the MultiFun class (M). The third/fourth column show the number of genes/operons of MultiFun class found in the heEPODs clusters (K_1/K_2). The fifth column shows P-values obtained from a test with parameters K_1, M, 3382, 757. The last column shows the terms related to the MultiFun class. We reported the results with $K_2 \geq 3$ and P-values< 0.01.

Class	M	K_1	K_2	P-value	Related term
1.3.7	155	4	3	0.000368	Anaerobic respiration
1.6.3.1	14	10	4	0.000001	O antigen
1.6.12	38	31	9	0.000001	Flagella
1.7.10	14	8	4	0.000011	Sugar nucleotide biosynthesis
2.2.5	91	18	11	0.001319	tRNA
2.2.6	26	11	4	0.000010	rRNA, Stable RNA
2.3.2	101	47	11	0.000001	Translation
2.3.8	57	33	6	0.000001	Ribosomal proteins
3.1.3.1	10	4	4	0.009770	Translation attenuation and efficiency
4.9.B	63	15	5	0.000483	Putative uncharacterized transport protein
4.S.12	12	6	3	0.000391	amino acid
4.S.160	32	15	4	0.000001	protein
5.3	59	33	11	0.000001	Motility
6.1	851	67	44	0.006859	Membrane
6.3	67	14	8	0.002512	Lipopolysaccharide
6.4	44	35	10	0.000001	Flagellum
6.6	95	46	12	0.000001	Ribosome
7.1	989	144	73	0.000001	Products location is cytoplasm
7.3	560	38	25	0.002379	Products location is inner membrane
10	43	17	5	0.000001	cryptic genes

Supplementary Table 7. Pseudo-genes.

Number of clusters (bin-size=128)	5
Total number of pseudo-genes	212
Number of pseudo-genes in the clusters	68
Number of annotated genes in the clusters	358

Supplementary Table 8. Intersection between different datasets. Result of a hypergeometric test assessing the statistical significance of the intersection between datasets. The P-value represents the probability of obtaining an intersection of the given size selecting two random gene lists (of the same length of the lists in consideration) from the total number of genes in the genome.

	HGT	H-NS binding	tsEPODs	Pseudo-genes
HGT	0	5.48797e-98	3.41112e-62	1.87785e-06
H-NS binding	5.48797e-98	0	1.76438e-189	7.49369e-18
tsEPODs	3.41112e-62	1.76438e-189	0	1.44551e-09
Pseudo-genes	1.87785e-06	1.44551e-09	7.49369e-18	0

Supplementary Table 9. Intersection between different datasets. Number of common genes between two different datasets, divided by the size of the smallest one. The datasets have many genes in common but they are not identical. Here, we considered genes located in H-NS binding regions in early exponential phase.

	HGT	H-NS binding	tsEPODs	Pseudo-genes
HGT	350/350	175/350	107/241	35/212
H-NS binding	175/350	441/441	203/241	62/212
tsEPODs	107/241	203/241	241/241	33/212
Pseudo-genes	35/212	62/212	33/212	212/212

Supplementary Table 10. Some of the significant clusters found in this study overlap with the long non-lethal deletions and with the gene from temperate prophages. Comparison of the position of clusters found analysing different datasets with large non-lethal deletions and prophages. The tick marks represent overlap between a cluster and a large deletion or a prophage. The position of the clusters correlate with the position of some long non-lethal deletions and prophages. The clusters that are close to macrodomain boundaries (distance ≤ 0.15 Mb) are colored in red.

Clusters coordinates (Mb)	Long deletions	Prophages
0.233-0.348	✓	✓
0.537-0.609	✓	✓
1.068-1.121	✓	-
1.173-1.252	-	✓
1.387-1.448	✓	✓
1.553-1.679	✓	✓
2.041-2.087	✓	✓
2.444-2.510	-	✓
2.747-2.807	✓	✓
2.956-3.024	-	-
3.619-3.685	✓	-
3.765-3.831	✓	-
4.284-4.349	✓	-
4.471-4.541	✓	✓

Supplementary Table 11. Enrichment analysis for functional annotations of genes located in each cluster that is reported in Table 1. Summary of the significant MultiFun gene classes for each cluster found. The first column shows the coordinates of the clusters. The second column shows the numbers of functional genes located in each cluster. The third column shows the MultiFun class number and related term. Here, we considered the results with P-values< 0.01. We removed the results when the number of operons of MultiFun class found in each cluster is less than 3.

Clusters coordinates (Mb)	Functional gene count	Functional classes and terms
0.233-0.348	77	7.1 (Products location is cytoplasm), 7.3 (Products location is inner membrane), 8.1 (Prophage related functions), 8.3 (Transposon related)
0.537-0.609	61	7.1 (Products location is cytoplasm), 7.4 (Products location is outer membrane), 8.1 (Prophage related functions)
1.068-1.121	43	8.1 (Phage related functions)
1.173-1.252	63	6.1 (Membrane), 7.1 (Products location is cytoplasm), 8.1 (Prophage related functions),
1.387-1.448	52	8.1 (Prophage related functions)
1.553-1.679	100	5.5.4 (PH response), 7.1 (Products location is cytoplasm), 8.1 (Prophage related functions)
2.041-2.087	36	2.2.5 (tRNA), 8.1 (Prophage related functions)
2.444-2.510	55	7.1 (Products location is cytoplasm), 8.1 (Prophage related functions)
2.747-2.807	47	8.1 (Prophage related functions)
2.956-3.024	36	1.7.1 (Unassigned reversible reactions)
3.619-3.685	35	3.1.2.2 (Activator), 6.1 (Membrane)
3.765-3.831	53	1.6.3.2 (Lipopolysaccharide), 2.1.4 (DNA repair), 6.3 (Surface antigens), 7.3 (Products location is inner membrane)
4.284-4.349	46	1.1.1 (Carbon compounds), 1.3.7 (Anaerobic respiration)
4.471-4.541	53	2.1.3 (DNA recombination), 8.1 (Prophage related functions), 8.3 (Transposon related)