

Chromosomal periodicity and positional networks of genes in *Escherichia coli*

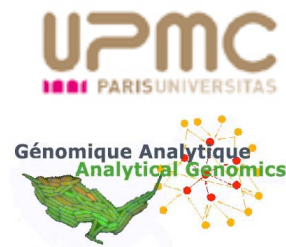
A.Mathelier, A.Carbone

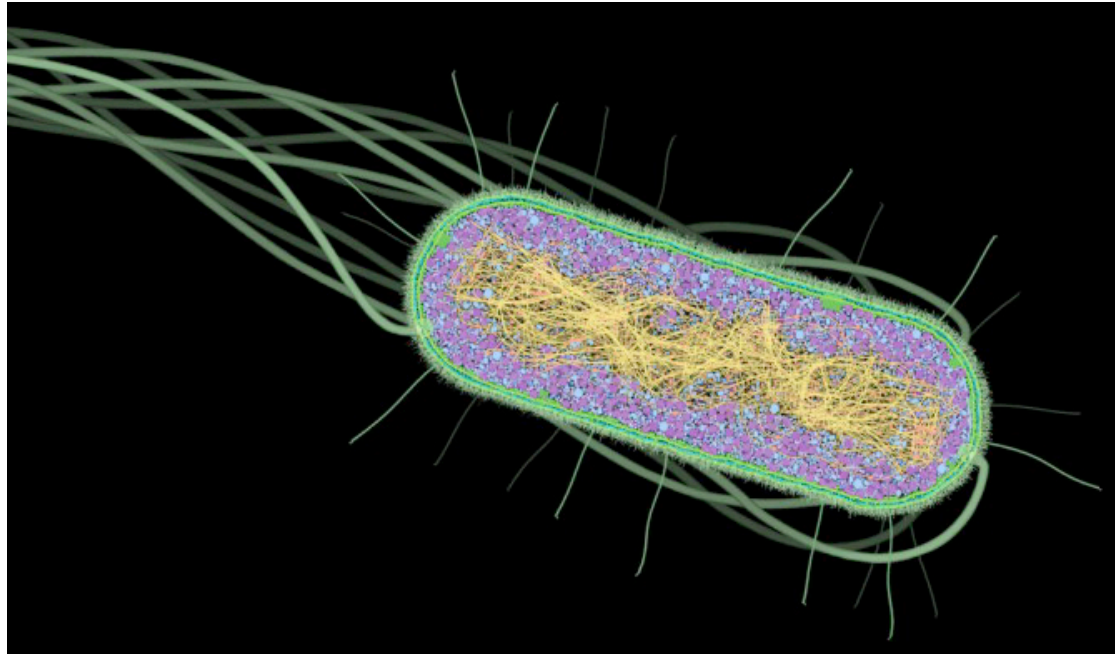
Molecular Systems Biology, 6:366, 2010

Alessandra Carbone

Département d'Informatique, Université Pierre et Marie Curie
Génomique Analytique, FRE 3214 CNRS-UPMC

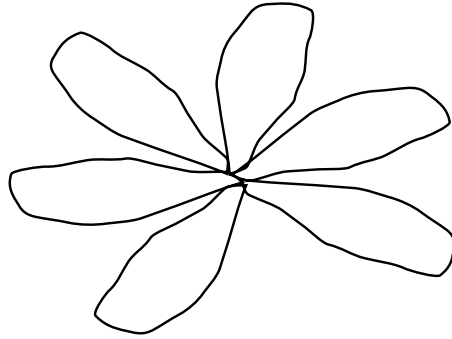
TAGp2010 – October 20, 2010



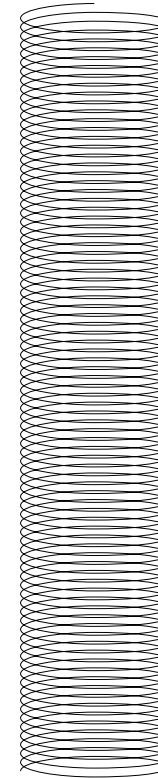
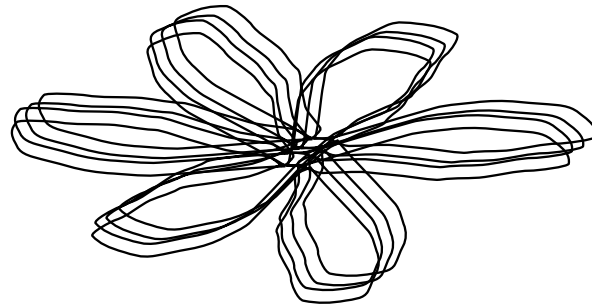


How can we accommodate a chromosome into a *E.coli* container with 0.5-0.75 μM diameter?

Spatial organization



Rosettes

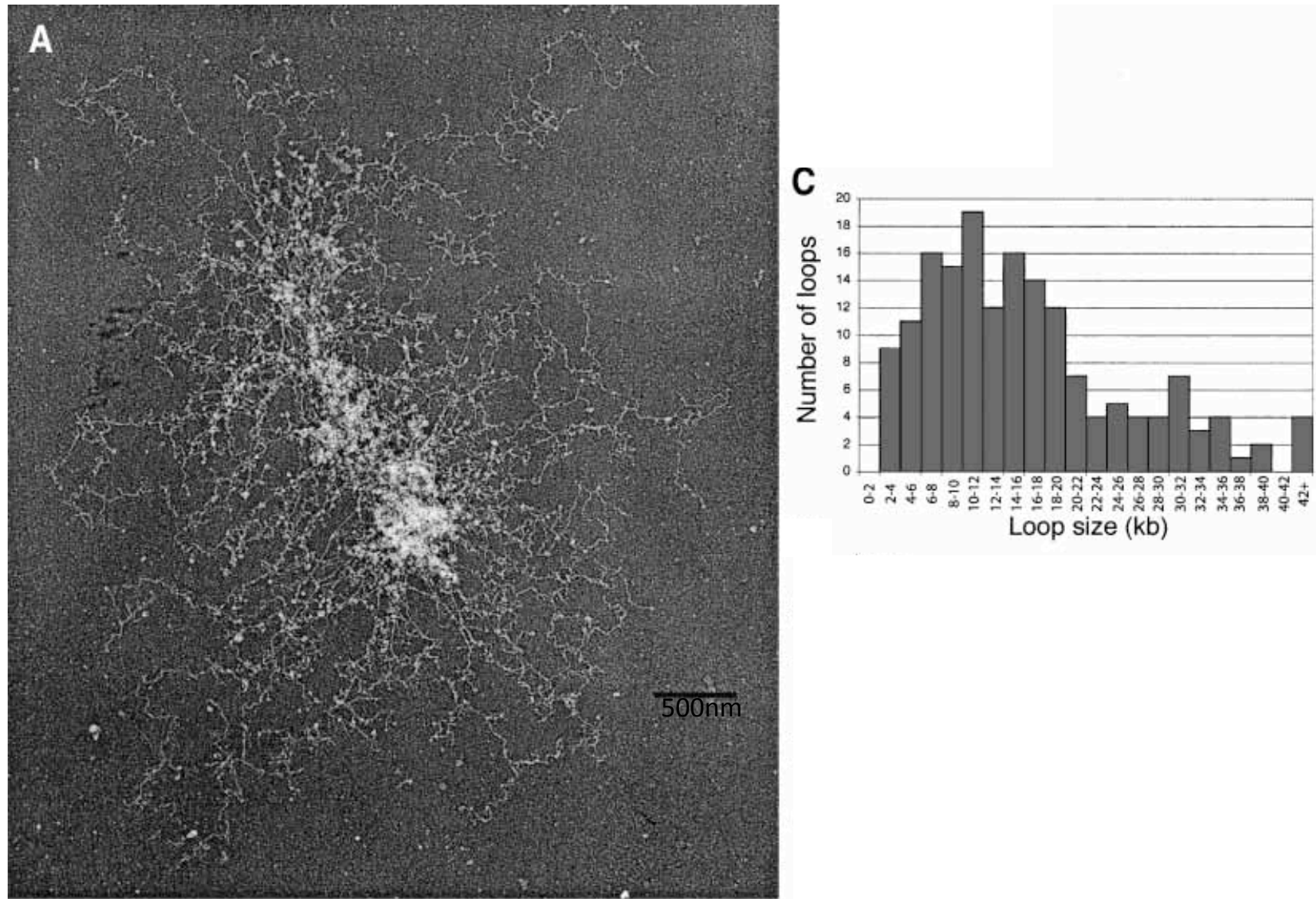


Solenoid

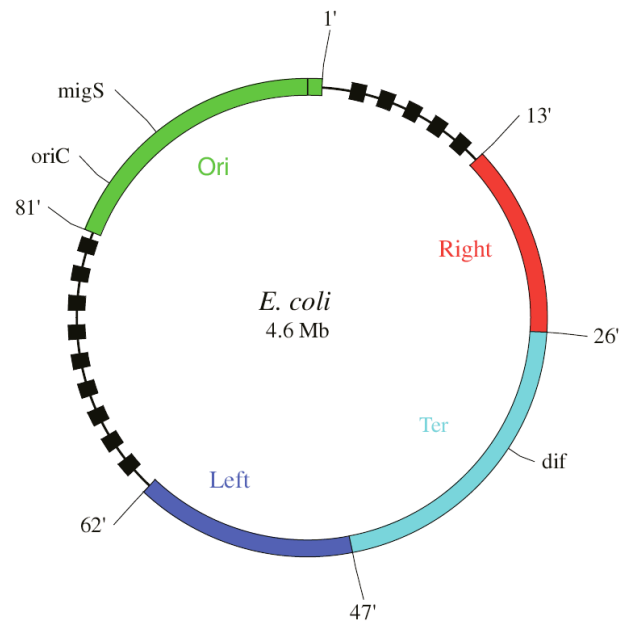
Intuition

Localization of certain groups of genes along structured chromosomes creates spatial chromosomal subregions in which genes can be accessed by limited diffusion of RNA polymerase or RNA polymerase fixed in factories.

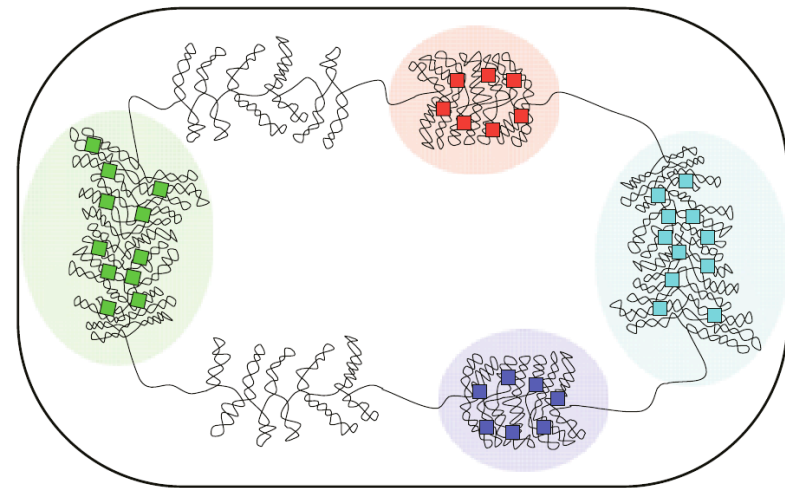
Experimental evidence for *E.coli*



L.Postow, C.D.Hardy, J.Arsuaga and N.R.Cozzarelli (2004) *Genes and Development*, 18:1766-1779.



M.Valens, S.Penaud, M.Rossignol, F.Cornet, F.Boccard
EMBO J, 23(21):4330-4341, 2004.



F.Boccard, E.Esnault, M.Valens
Molecular Microbiology, 57(1):9-16, 2005.

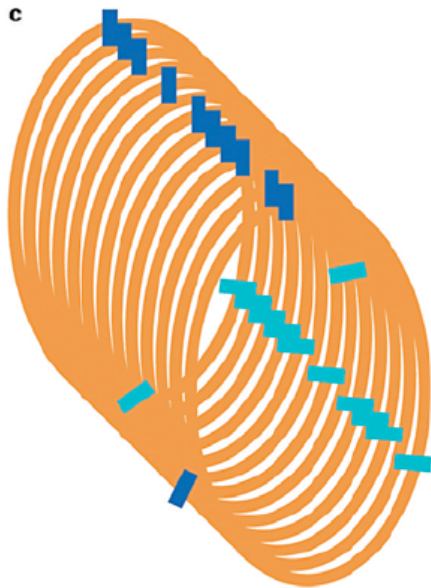
Others

C.D.Hardy and N.R.Cozzarelli (2005) *Molecular Microbiology*, 57(6):1636-1652.

N.Blot, R.Mavathur, M.Geertz, A.Travers and G.Muskhelishvili (2006) *EMBO Rep.*, 7(7): 710-715.

Genomic evidence for *E.coli*

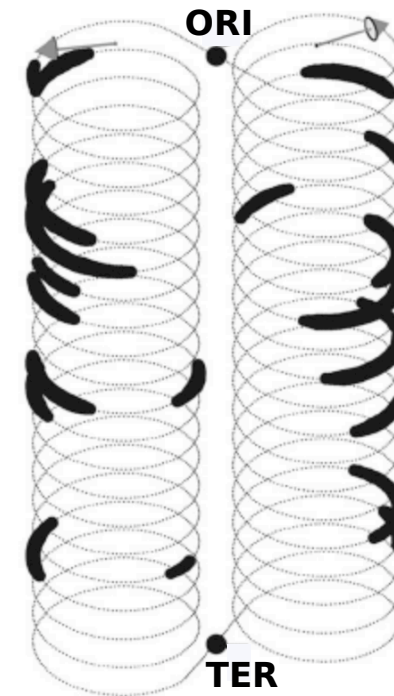
Targets of transcription factors



F.Képès (2004) *Journal of Molecular Biology*, 340:957-964.

90kb

Evolutionary conserved gene pairs



M.A.Wright, P.Kharchenko, G.M.Church and D.Segré (2007) *PNAS*, 104:10559-10564.

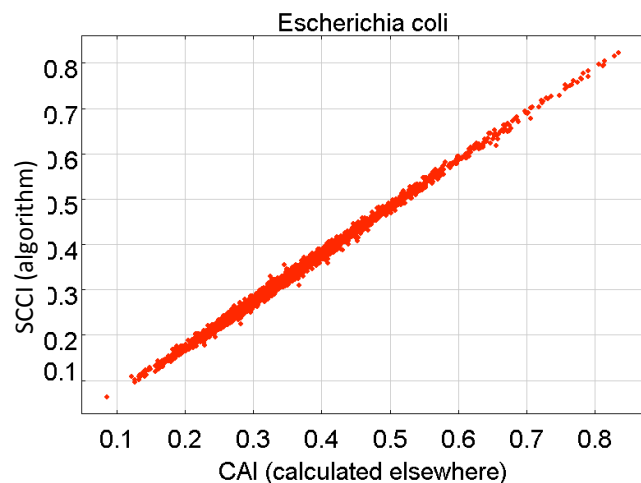
117kb

Search for periodicity from core genes

Core genes

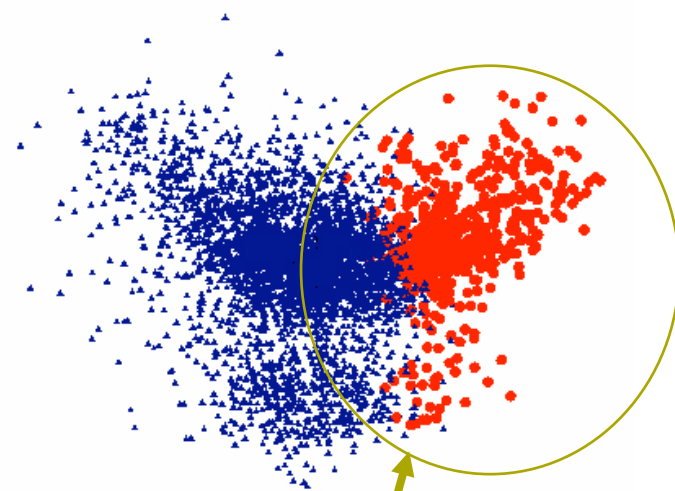
Core genes are selected to have high Self-Consistent Codon Index (SCCI) values. SCCI is highly correlated to Codon Adaptation Index (CAI) in *E.coli*.

The most biased set of genes found in *E.coli*

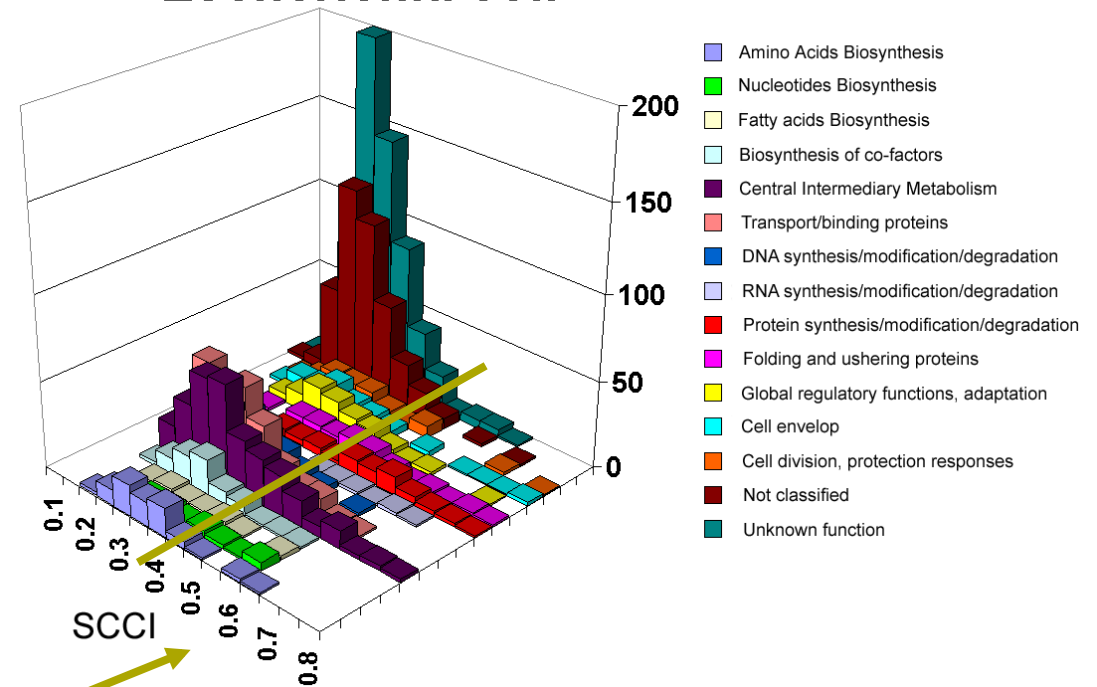


Gene	Annotation
tufA	protein chain elongation factor EF-Tu
tufB	protein chain elongation factor EF-Tu
tsf	protein chain elongation factor EF-Ts
fusA	GTP-binding protein chain elongation factor EF-G
mopA	chaperonin GroEL
dnaK	heat shock protein DnaK
cspA	cold shock protein 7.4
tig	trigger factor
ompA	outer membrane protein
ompX	outer membrane protein
ompC	outer membrane protein
lpp	murein lipoprotein
pal	peptidoglycan-associated lipoprotein
yaiU	putative flagellin structural protein
yfiD	putative formate acetyltransferase
eno	diadenosine tetraphosphatase
tpiA	triosephosphate isomerase
pgk	phosphoglycerate kinase
gapA	glyceraldehyde-3-phosphate dehydrogenase A
fba	fructose-bisphosphate aldolase class II
pykF	pyruvate kinase I
pflB	formate acetyltransferase 1
ahpC	alkyl hydroperoxide reductase C22 subunit
sodA	superoxide dismutase SodA
tktA	transketolase 1/2 isozyme
rpoC	RNA polymerase beta prime subunit
rpsI	30S ribosomal subunit protein S9
rpsA	30S ribosomal subunit protein S1
rpsB	30S ribosomal subunit protein S2
rpsC	30S ribosomal subunit protein S3
rpsU	30S ribosomal subunit protein S21
rplA	50S ribosomal subunit protein L1
rplY	50S ribosomal subunit protein L25
rplI	50S ribosomal subunit protein L9
rplL	50S ribosomal subunit protein L7/L12
rplC	50S ribosomal subunit protein L3
rpmE	50S ribosomal subunit protein L31
rplB	50S ribosomal subunit protein L2
rplK	50S ribosomal subunit protein L11
rpmI	50S ribosomal subunit protein A
rpmA	50S ribosomal subunit protein L27
rplD	50S ribosomal subunit protein L4, regulates expression of S10 operon

Escherichia coli



Core genes



We look at the tail

$$SCCI(g) > \mu + \sigma = 0.42$$

- Amino Acids Biosynthesis
- Nucleotides Biosynthesis
- Fatty acids Biosynthesis
- Biosynthesis of co-factors
- Central Intermediary Metabolism
- Transport/binding proteins
- DNA synthesis/modification/degradation
- RNA synthesis/modification/degradation
- Protein synthesis/modification/degradation
- Folding and ushering proteins
- Global regulatory functions, adaptation
- Cell envelop
- Cell division, protection responses
- Not classified
- Unknown function

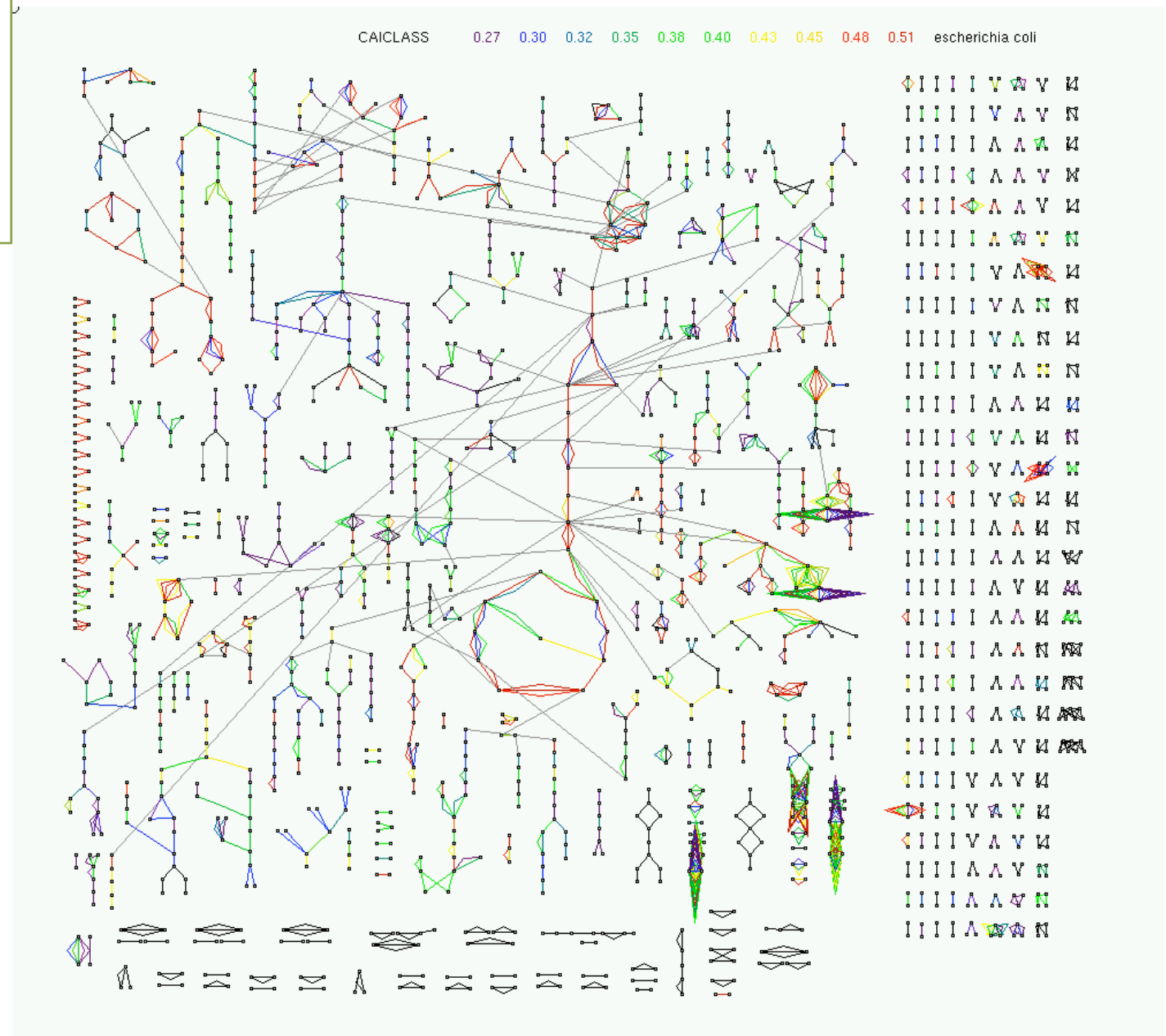
$$SCCI(g) > \mu + \sigma$$

- Genes with uncharacterised function
- Genes dependent on specific environmental conditions
- Stress response genes
- Highly expressed genes (belonging to most species)
- Non-orthologous genes

Core genes

SCCI can be used to detect essential metabolic networks

E.coli



A. Carbone, R. Madden, *Journal of Molecular Evolution*, 2005.

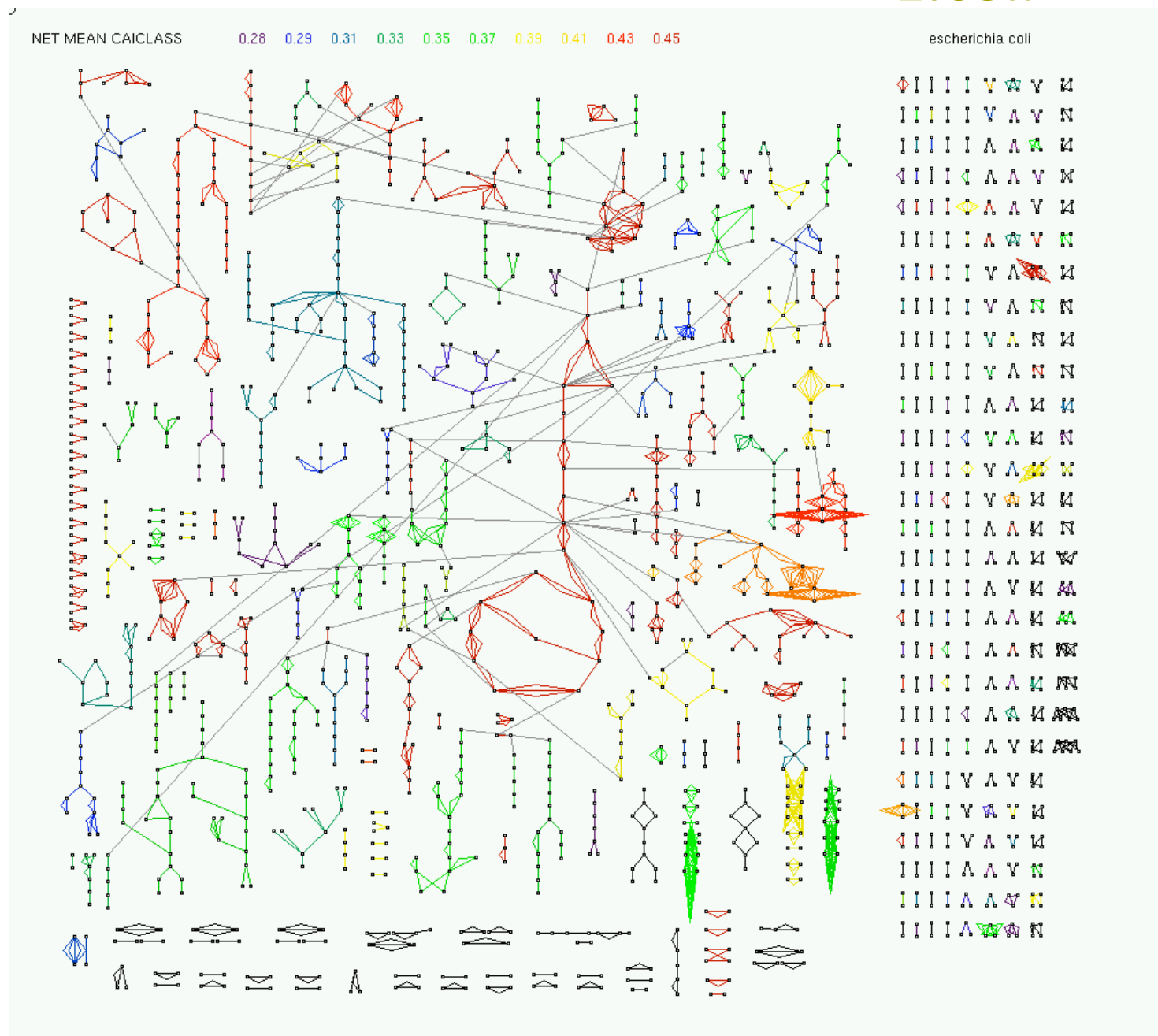
EcoCyc network, P.Karp *et al.*

Pathway Index

$$PI(P) = \text{mean } SCCI(g)_{g \in P}$$

Relative Pathway Index

$$RPI(P) = (PI(P) - \mu_M) / \sigma_M$$



NET MEAN CAICLASS

0.28 0.29 0.31 0.33 0.35 0.37 0.39 0.41 0.43 0.45

escherichia coli

Histidine+purine+
pyrimidine biosynthesis

Non-oxidative branch
of the pentose
phosphate pathway

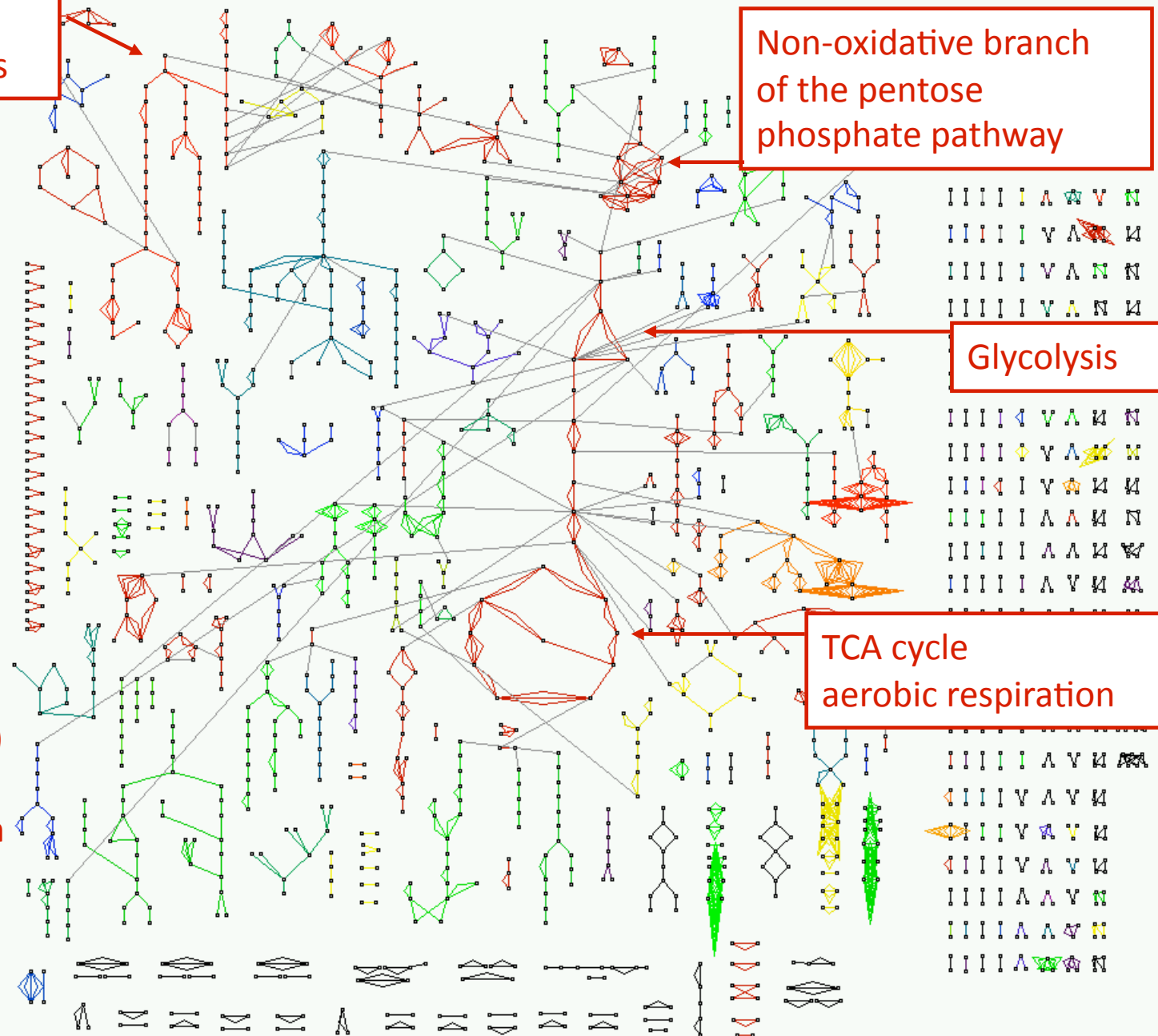
Glycolysis

TCA cycle
aerobic respiration

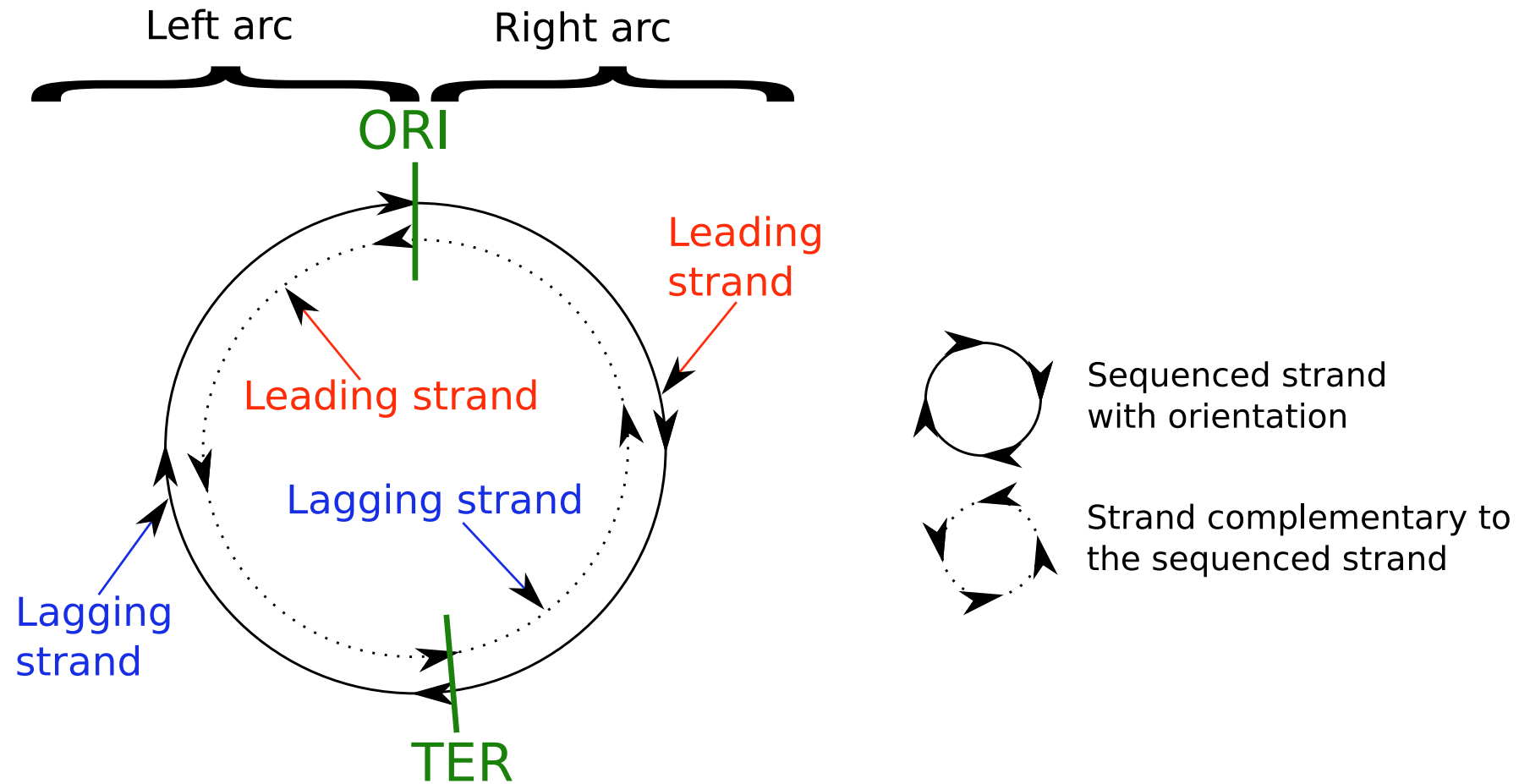
...and also :

L-serine degradation
(Pizer&Potochny 1964)

**Ammonia assimilation
Pathway**
(Reitzer 1986,
Helling 1994)



Circular chromosome

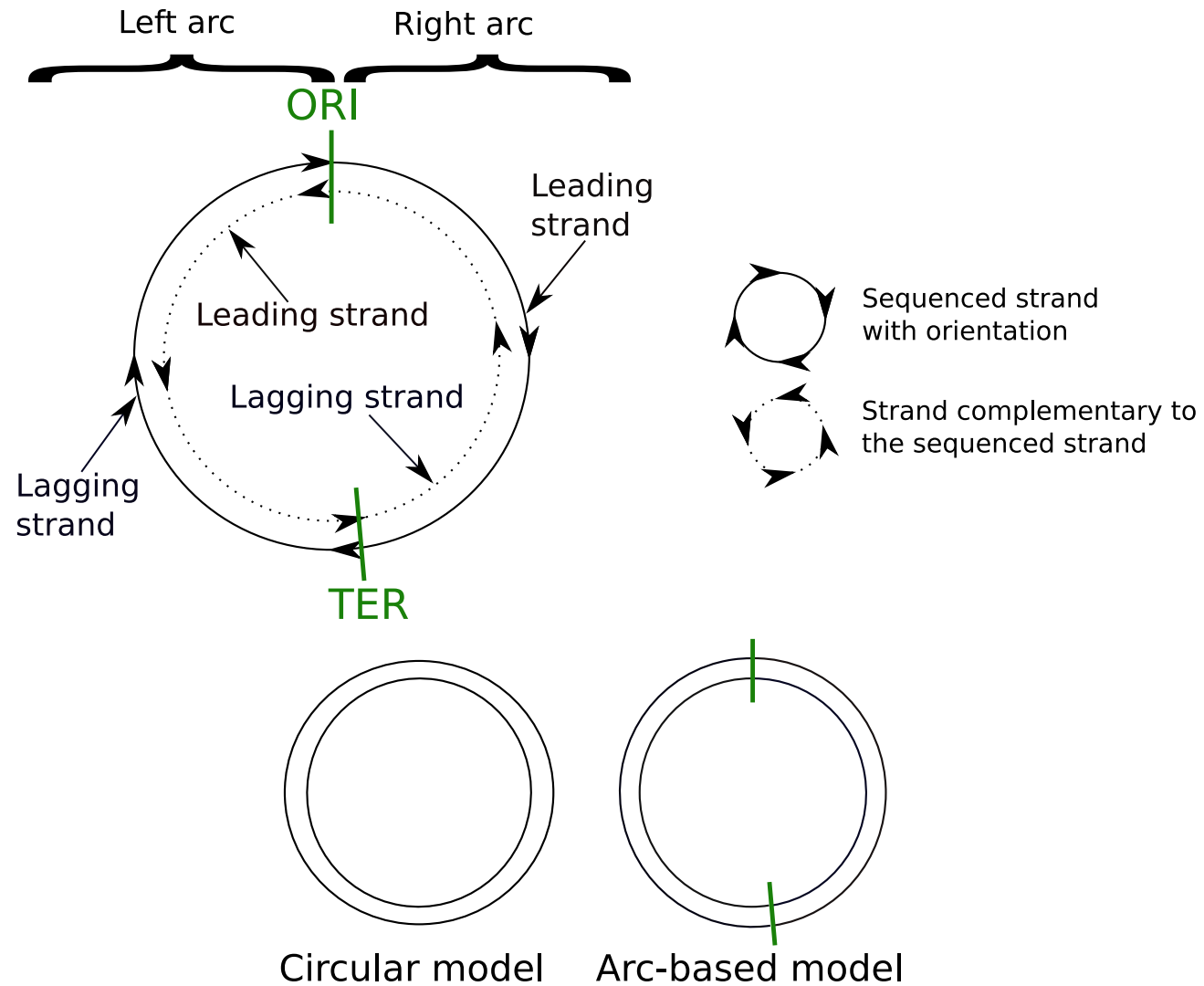


Chromosomal localization of core genes

	Lagging	Leading	Total
Number of genes	1931	2364	4295
Number of core genes	223	340	563
Top 100 core genes	24	76	100
Core genes near to ORI	122	207	329
Core genes near to TER	101	133	234

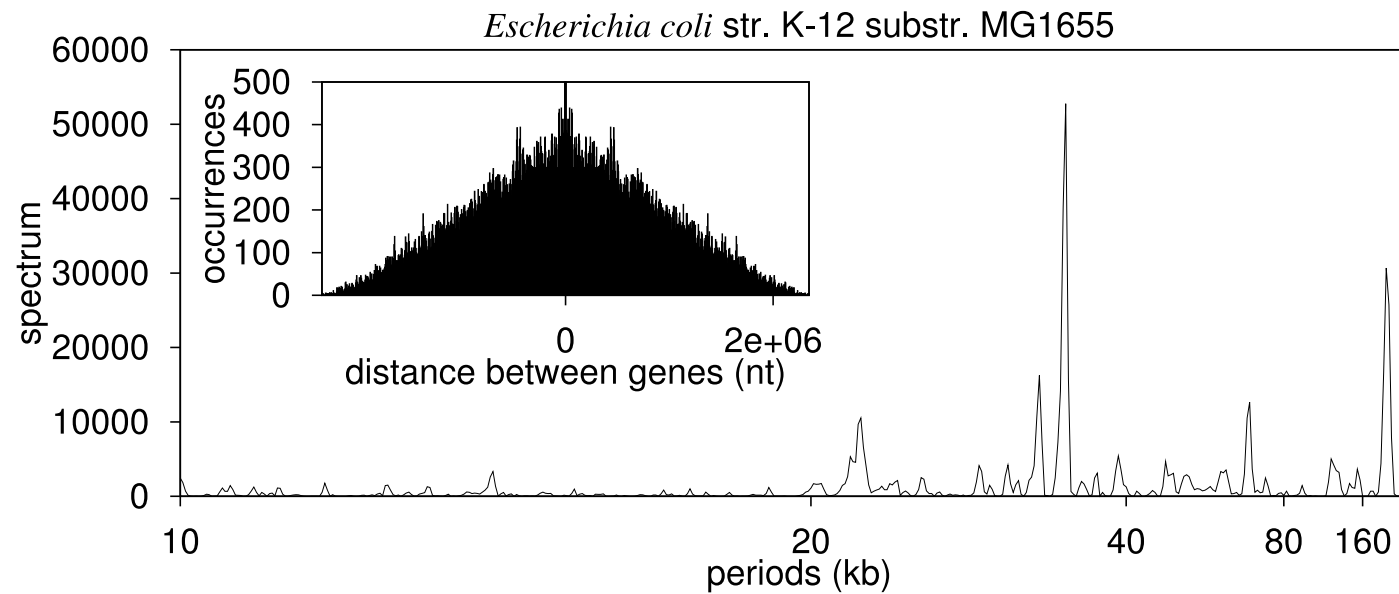
A bias appears on the Leading strand around the Origin of replication (ORI)

Models



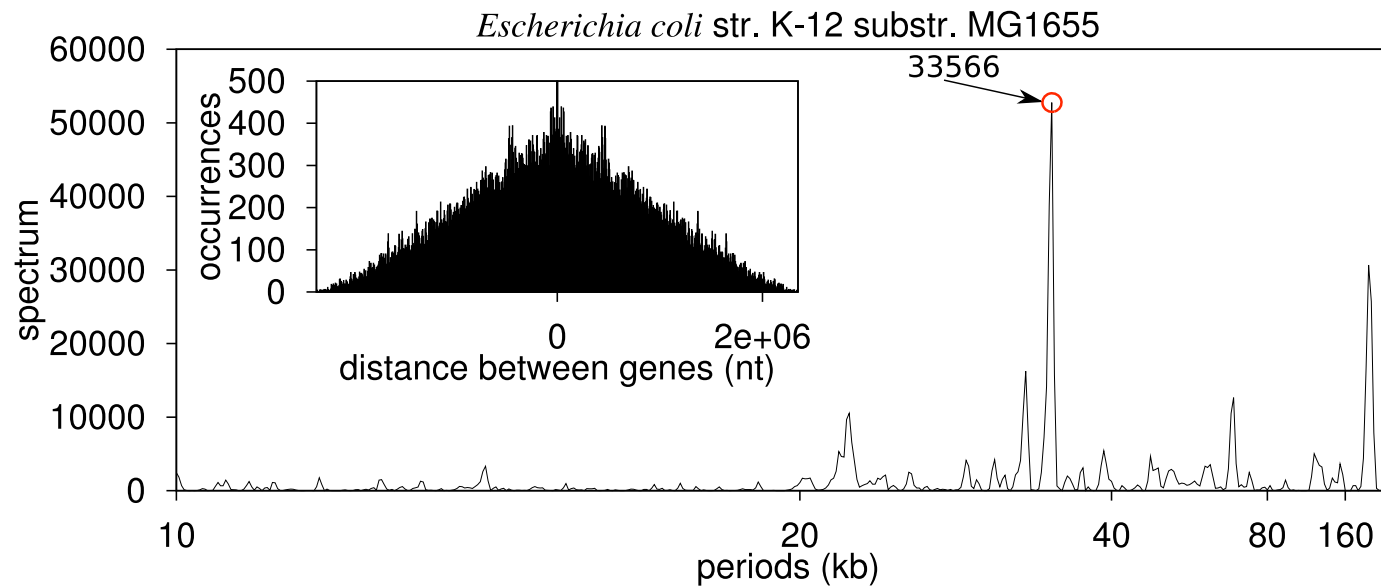
Fourier Analysis: histogram and periodogram

On Arc-based model :



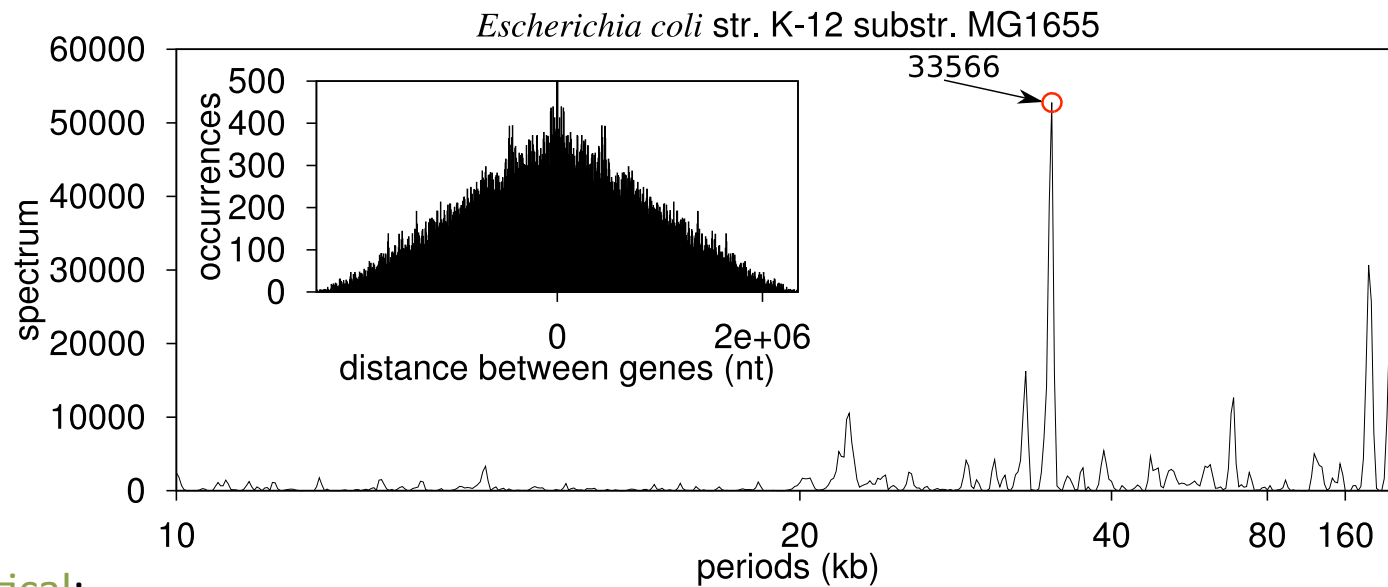
Fourier Analysis: histogram and periodogram

On Arc-based model :

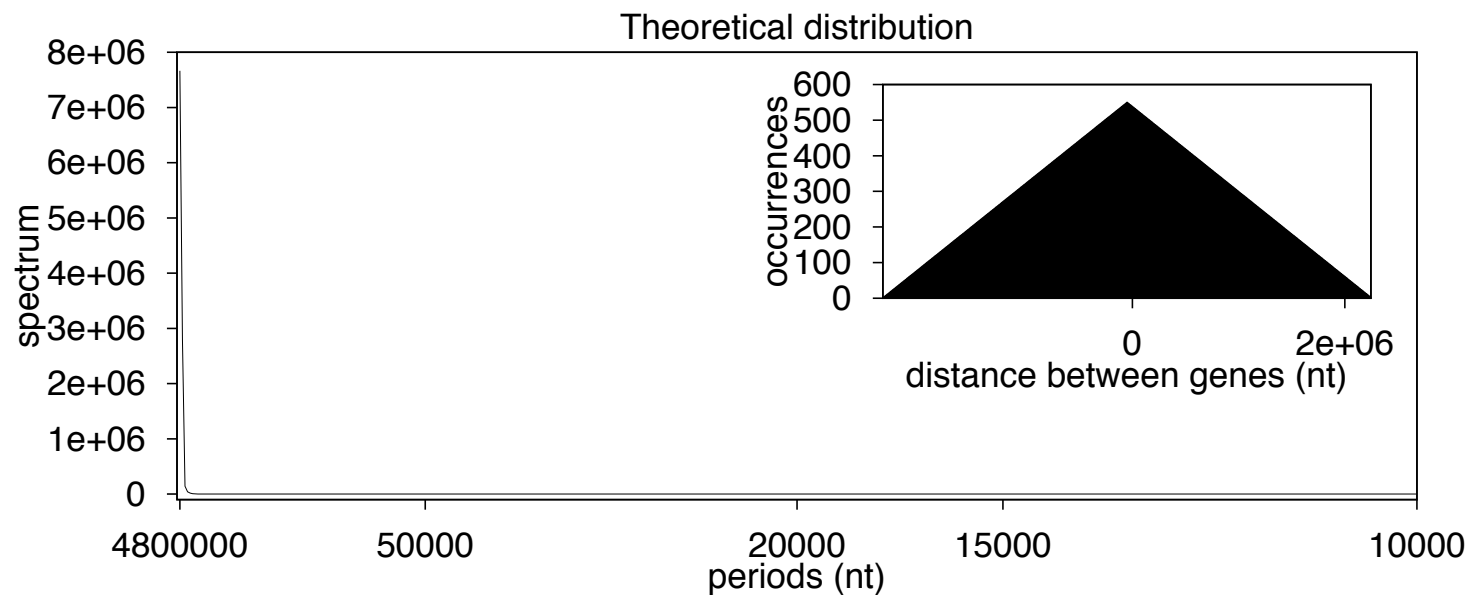


Fourier Analysis: histogram and periodogram

On Arc-based model :

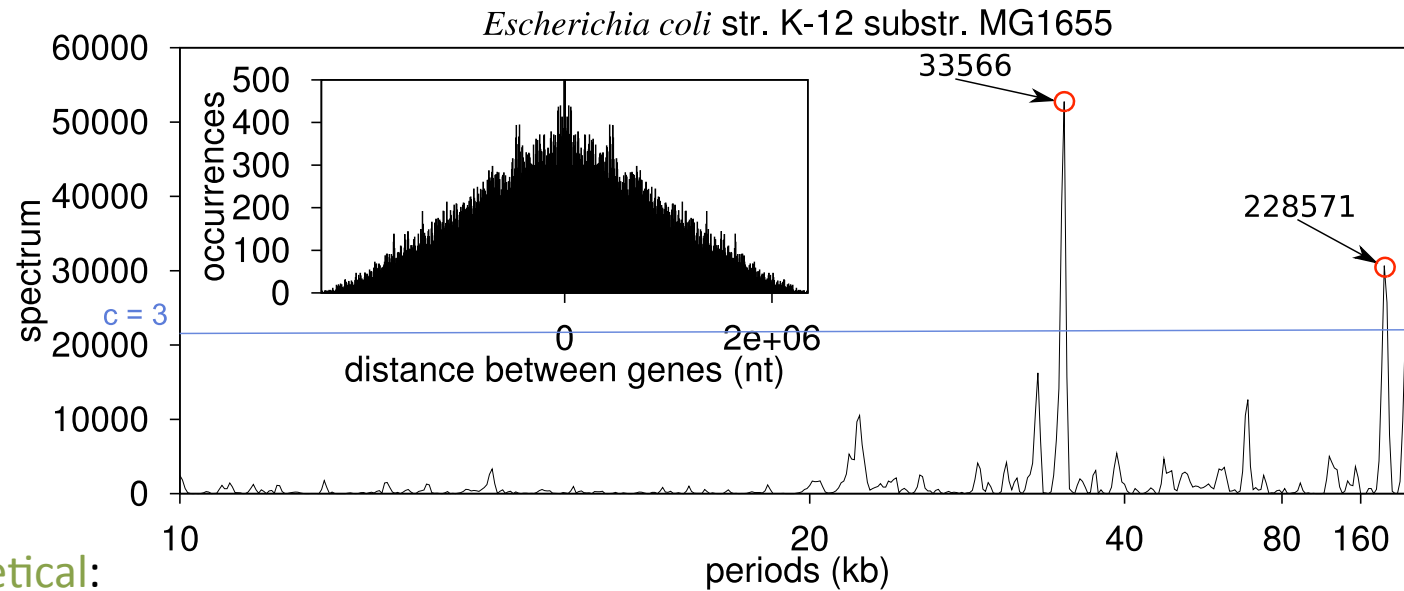


Theoretical:

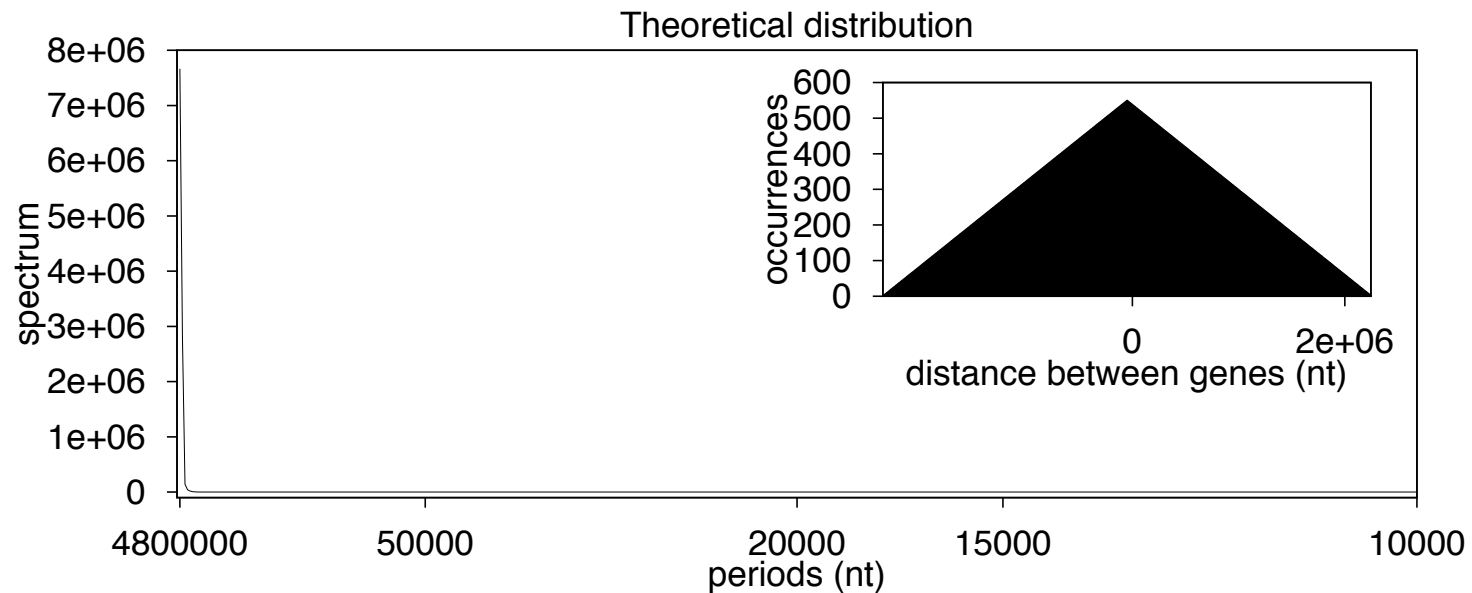


Fourier Analysis: histogram and periodogram

On Arc-based model :



Theoretical:



Periods

Core genes tend to be separated by a periodic distance of **33kb**

Pour $c=3$

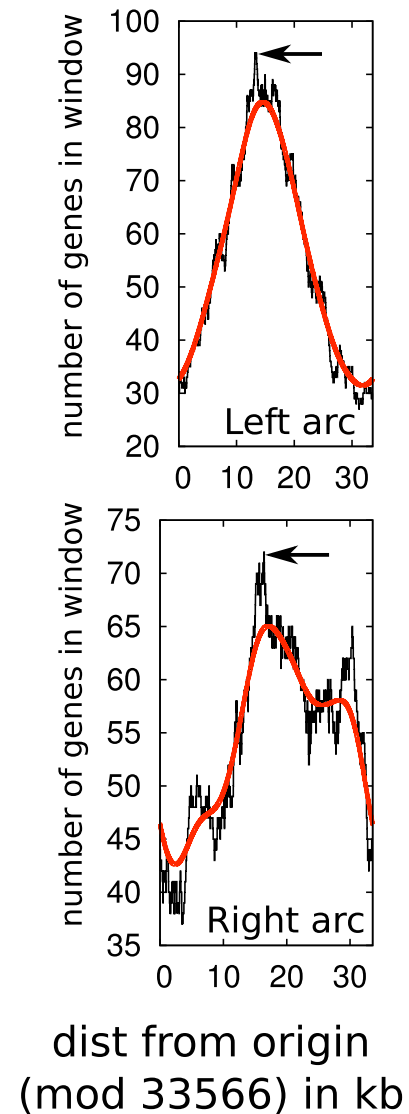
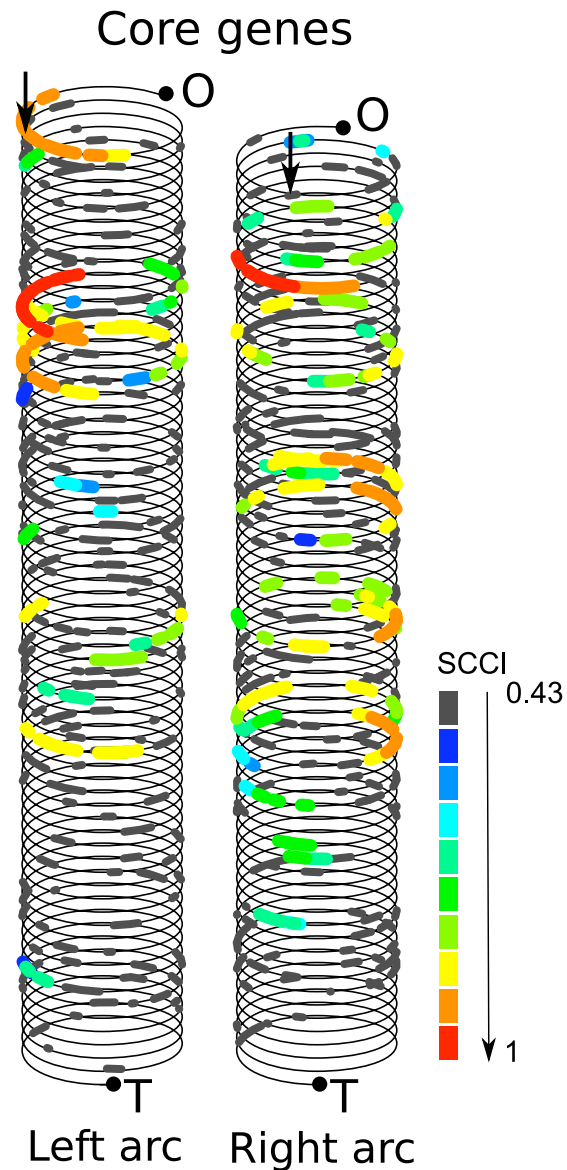
Circular model

33566 ($P < 10^{-4}$) 228571

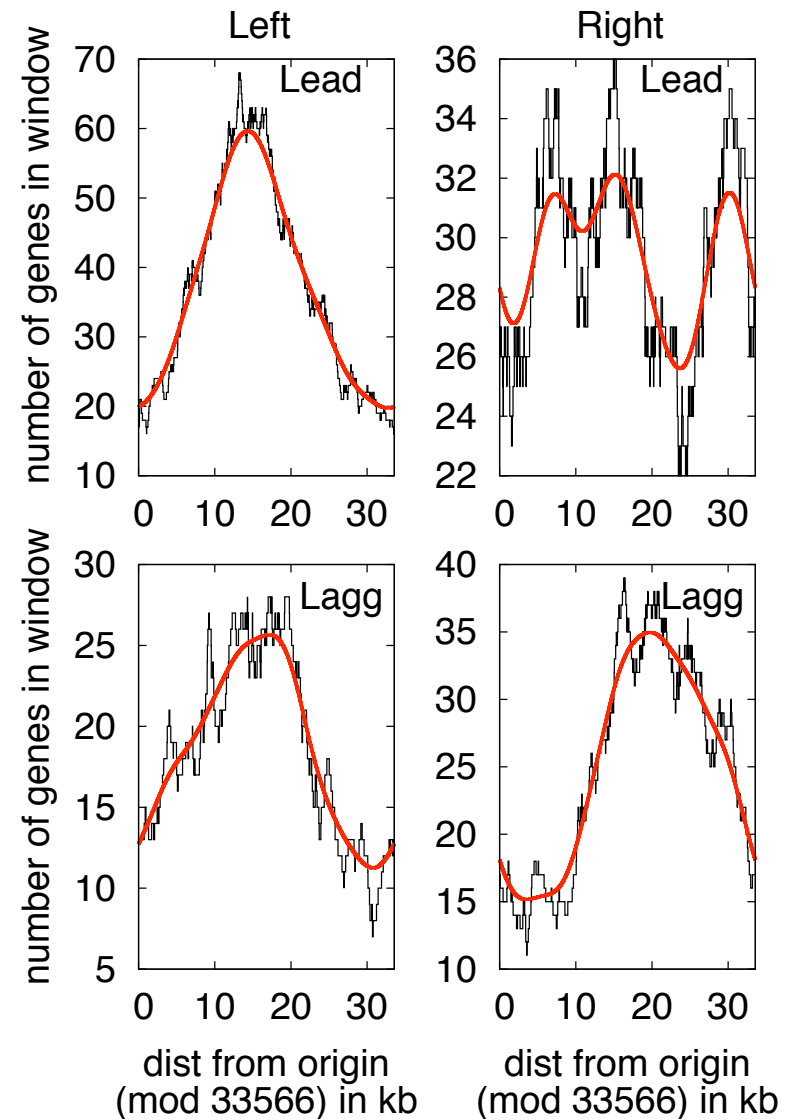
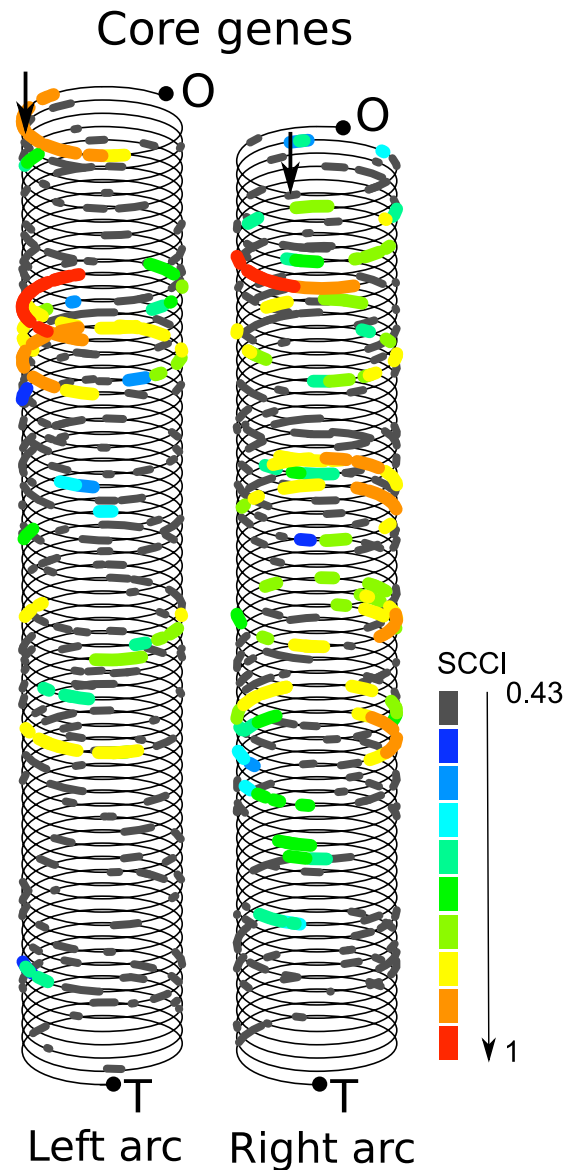
Arc-based model

33566 ($P < 10^{-4}$) 228571

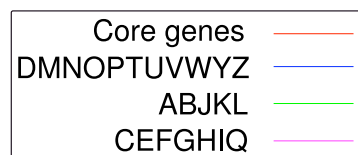
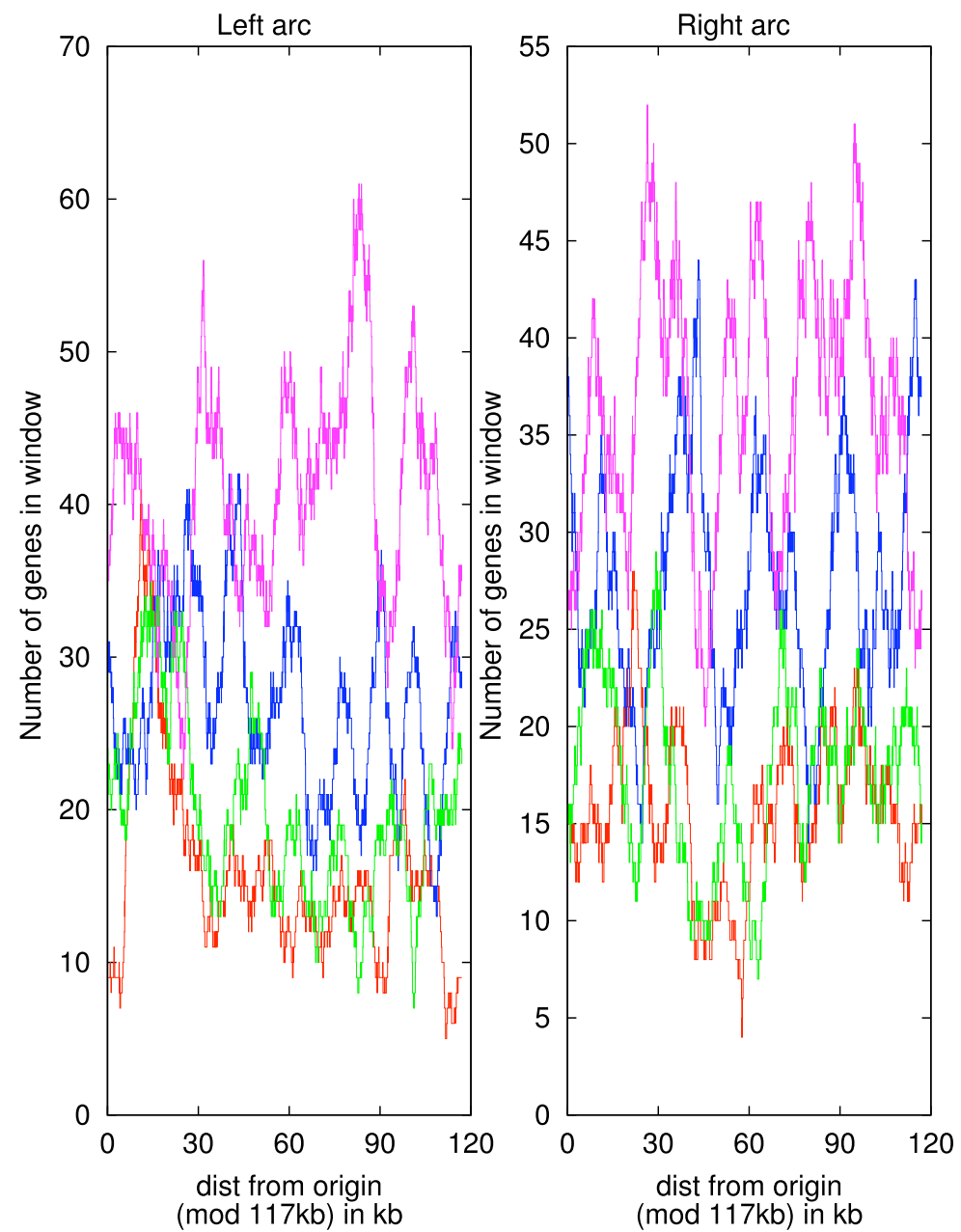
Core genes along the chromosomal arcs



Core genes on leading and lagging strands & LGT



(French, 1992) and (Touchon et al. 2009)



Search for periodicity from Clusters of Orthologous Groups

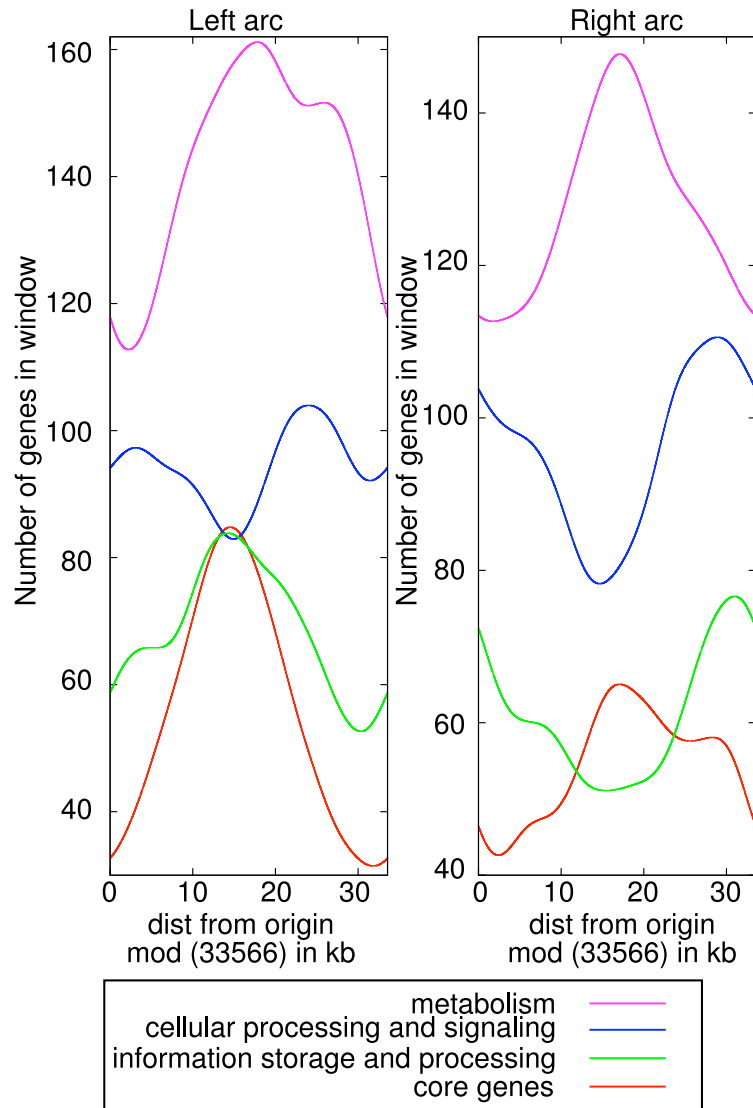
Tatusov RL et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*

COGs functional classes of genes

There are three main COGs classes:

- **metabolism**
energy production and conversion, carbohydrate/amino-acid/nucleotide transport and metabolism, coenzyme/lipid metabolism
- **information storage and processing**
translation, transcription, DNA replication, recombination, repair
- **cellular processing and signaling**
cell division and chromosome partitioning, posttranslational modification, cell envelop biogenesis, cell mobility and secretion

Functional organization of genes along the chromosome



Two positional networks

left arc:

metabolism

information storage and processing

right arc:

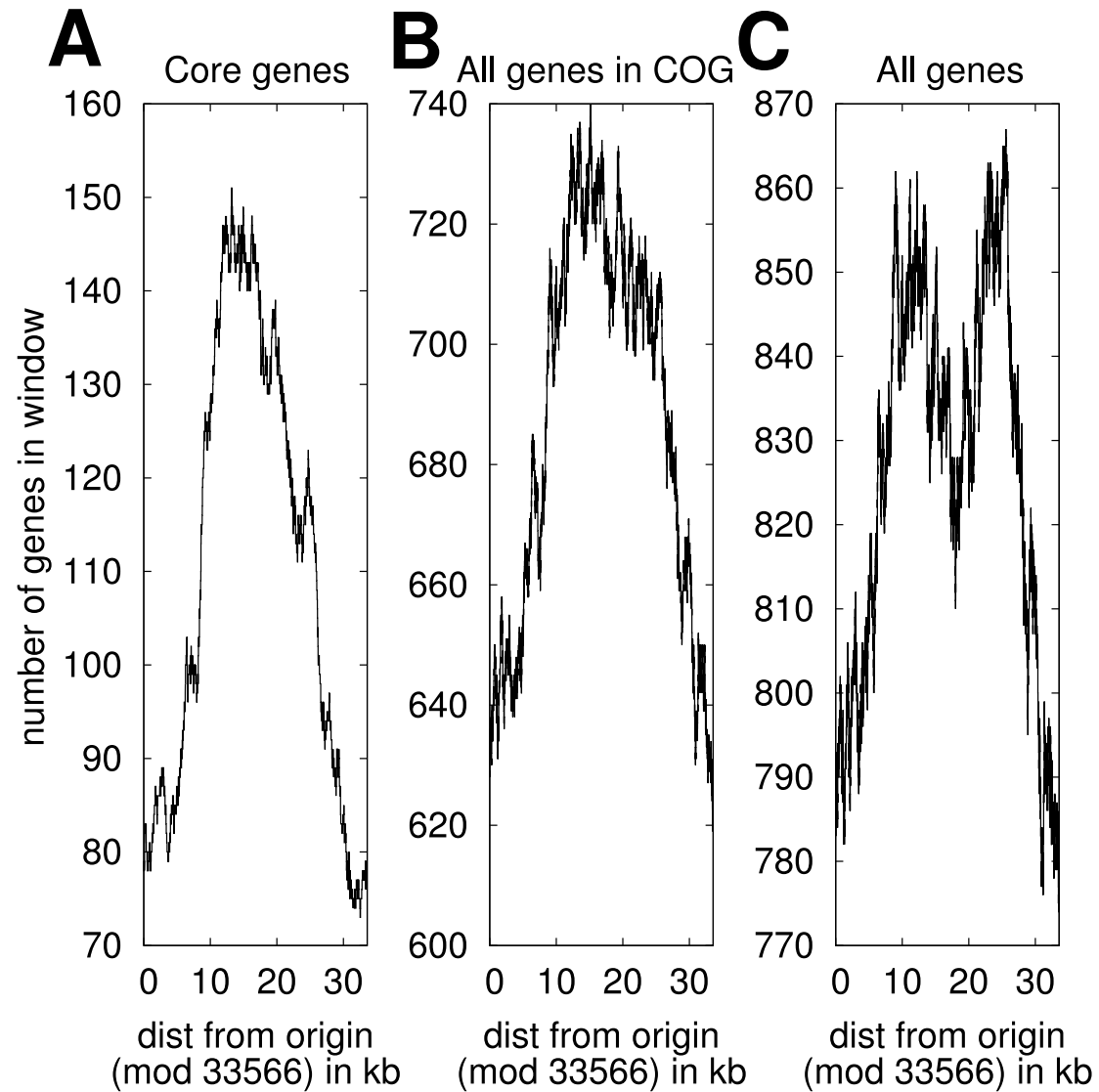
cellular processing and signaling

Information storage and processing

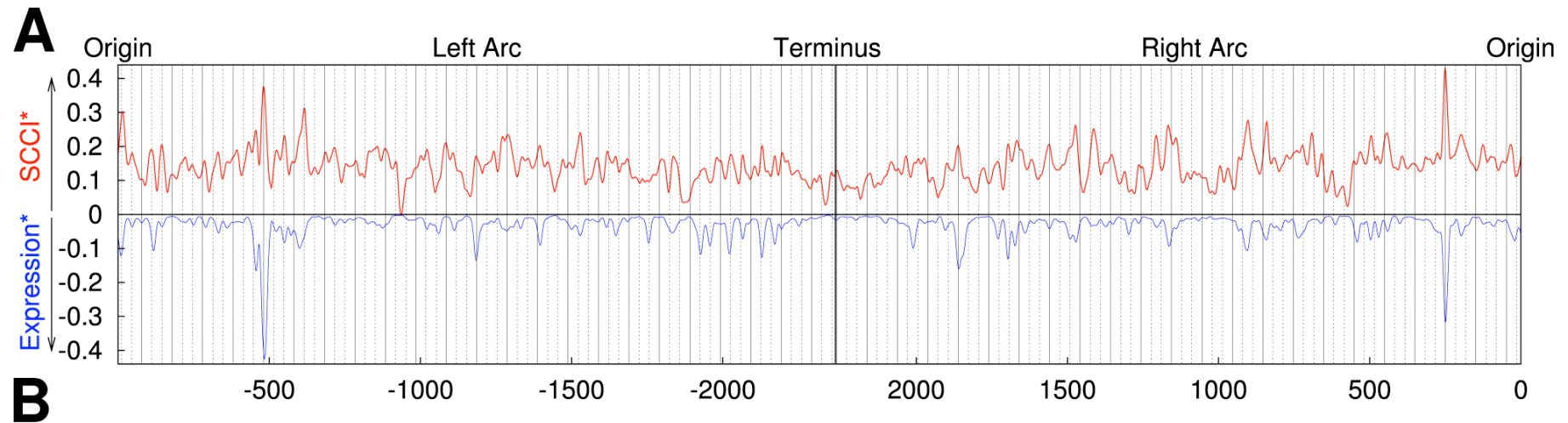
metabolism: first peak at 33kb ($P=0.0041$)

Other classes: 33kb but not as first.

Functional organization of **all** genes

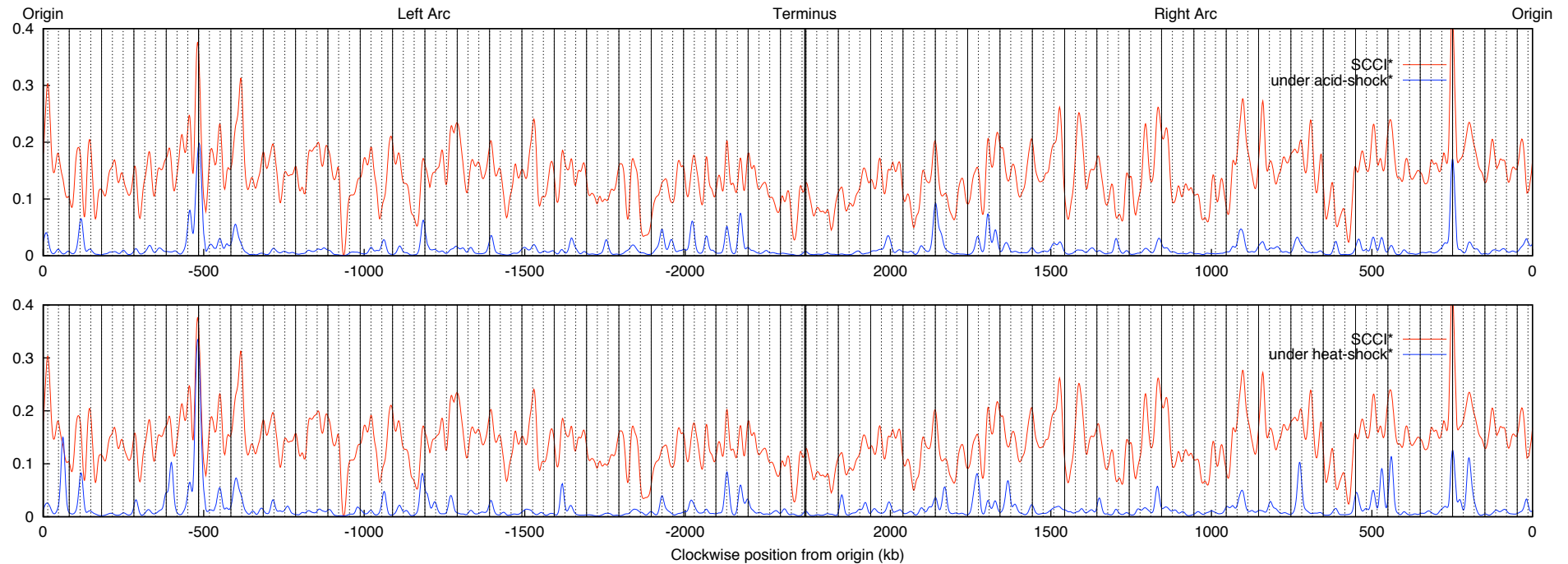


SCCI vs transcription data



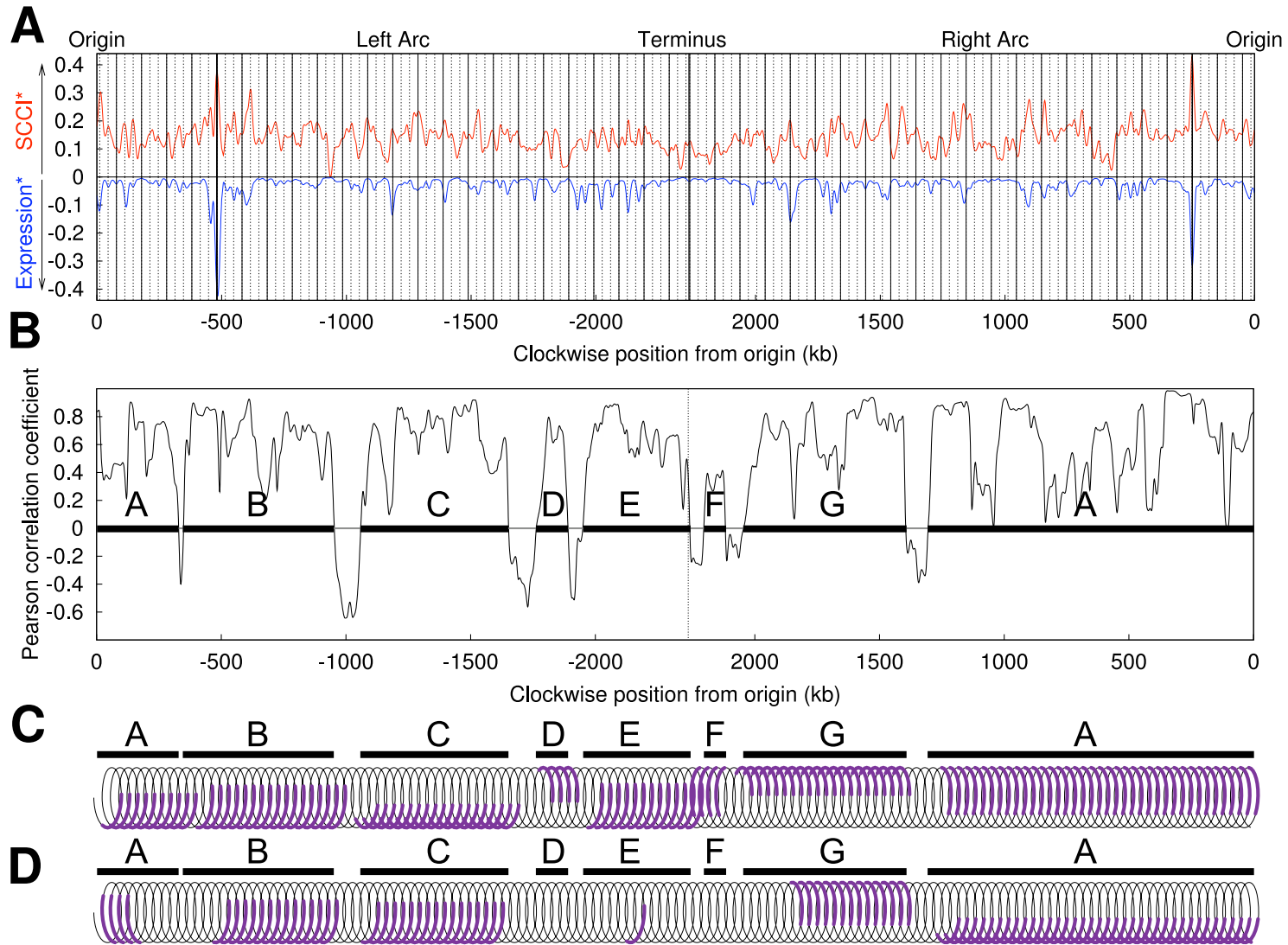
Correlation between SCCI and log-phase transcript data
(global Pearson correlation $R=0.59$, and $P < 2.2e^{-16}$).

SCCI vs acid-shock - SCCI vs heat-shock

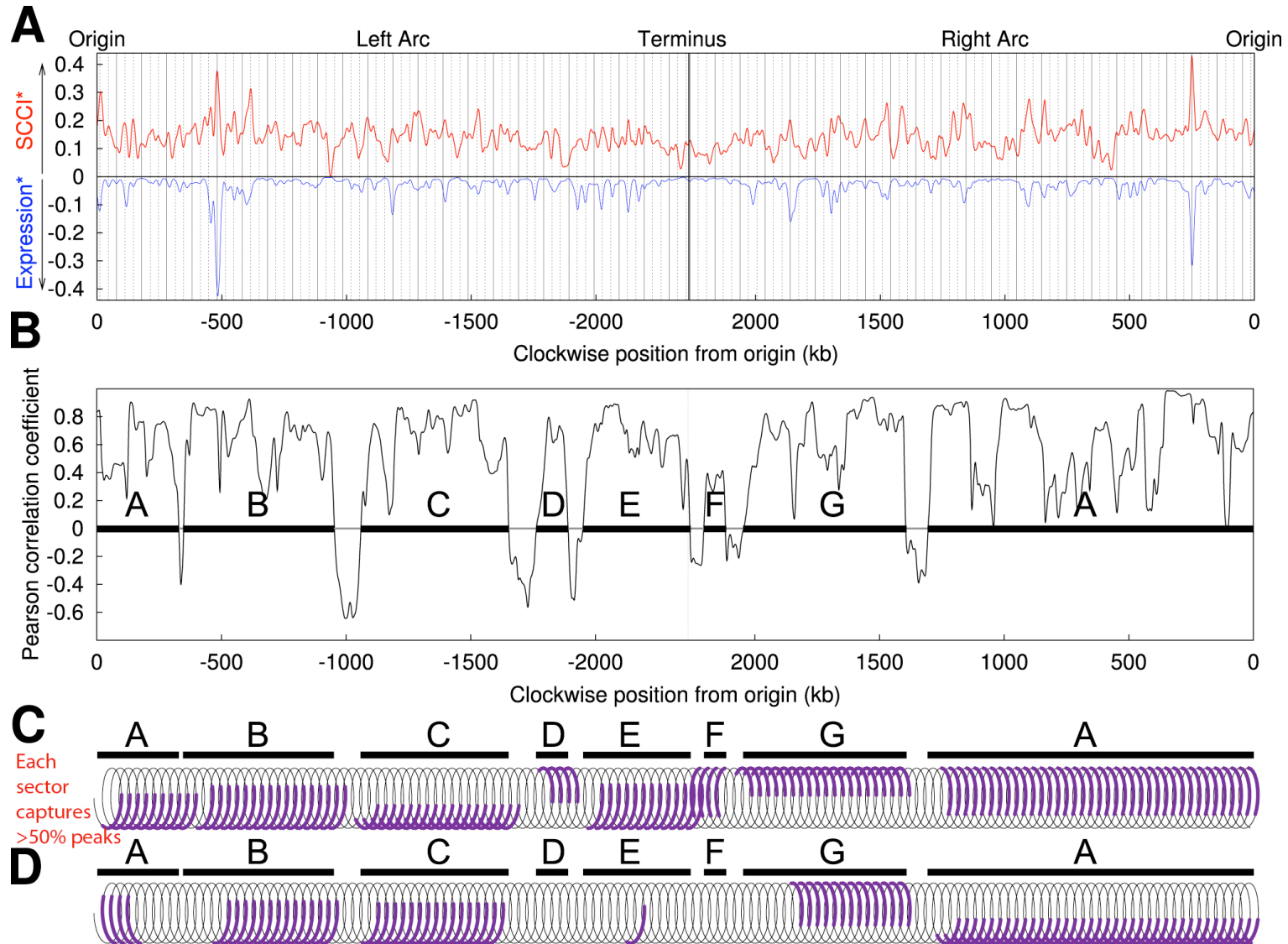


Global Pearson correlation between
SCCI and acid-shock transcript data : $R=0.54$
SCCI and heat-shock transcript data: $R=0.47$

Local Pearson correlation coefficient



Local Pearson correlation coefficient



Position of SCCI* peaks, sectors and strips

	A	B	C	D	E	F	G	A
Sector size (kb)	330	605	593	127	429	87	650	1304
#Peaks in strip	8	12	9	3	8	3	9	18
#Total peaks	9	19	17	3	16	3	18	42

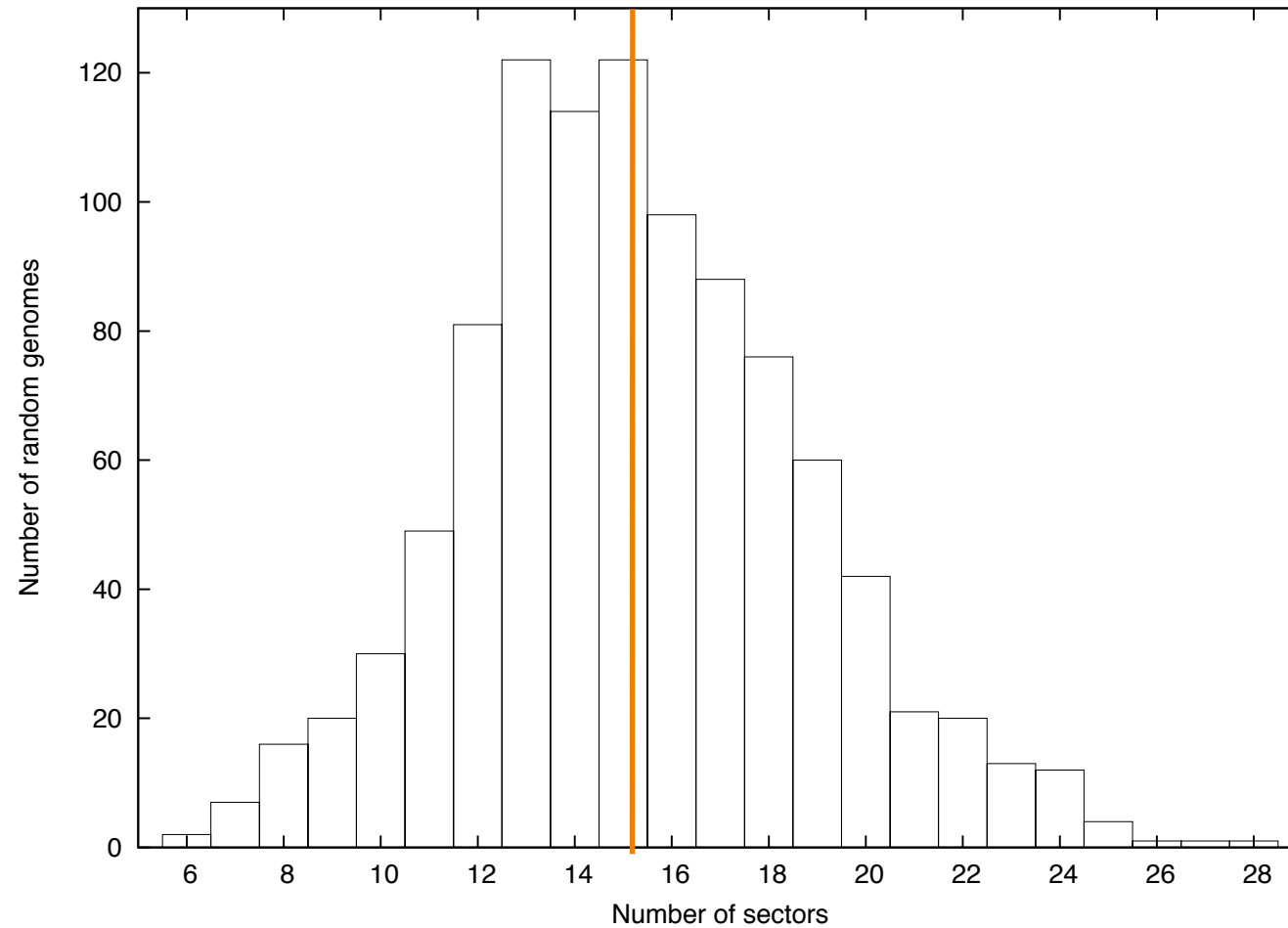
For all sectors, at least 50% of the peaks are covered by a sector

Position of $\text{SCCI}^* > \mu + \sigma$ peaks, sectors and strips

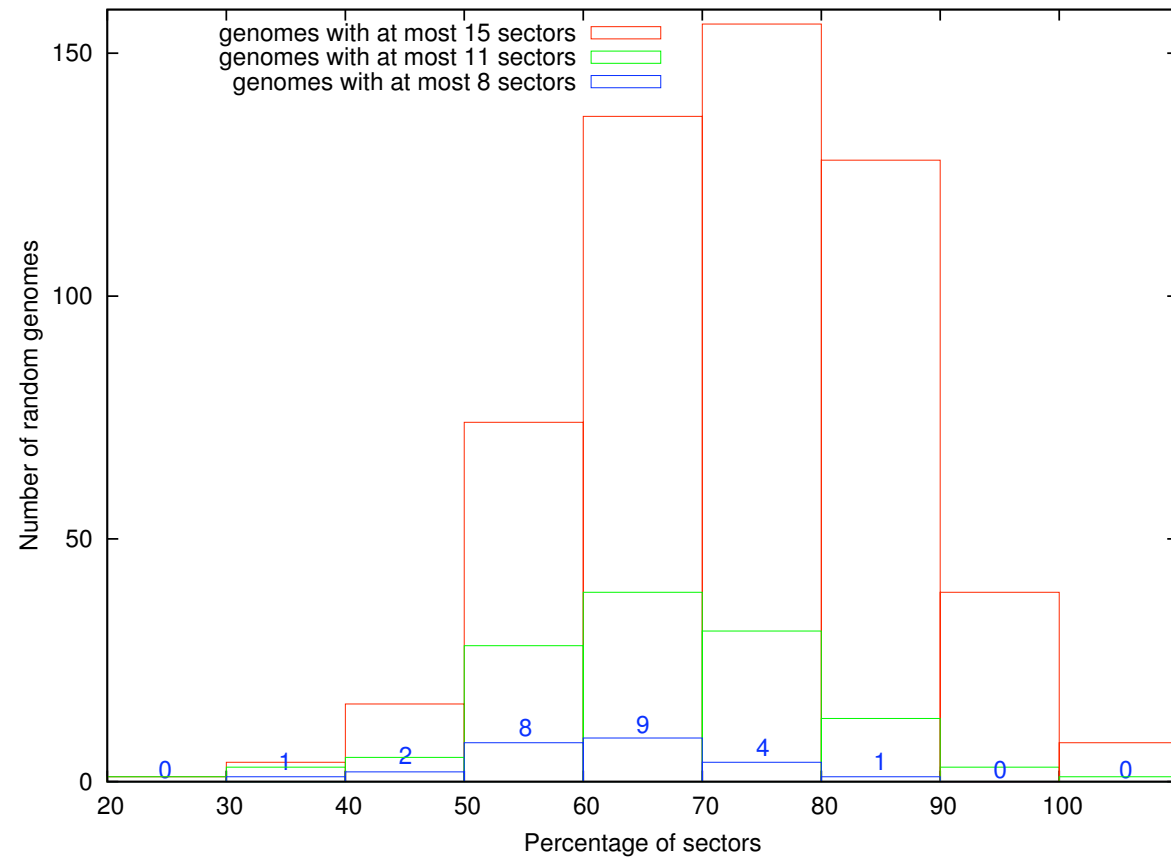
	A	B	C	D	E	F	G	A
Sector size (kb)	330	605	593	127	429	87	650	1304
#Peaks in strip	0	7	4	0	1	0	4	9
#Total peaks	2	8	5	0	1	0	6	13

For all sectors, at least 60% of the peaks are covered by a sector

Sector analysis of 1000 randomly generated genomes

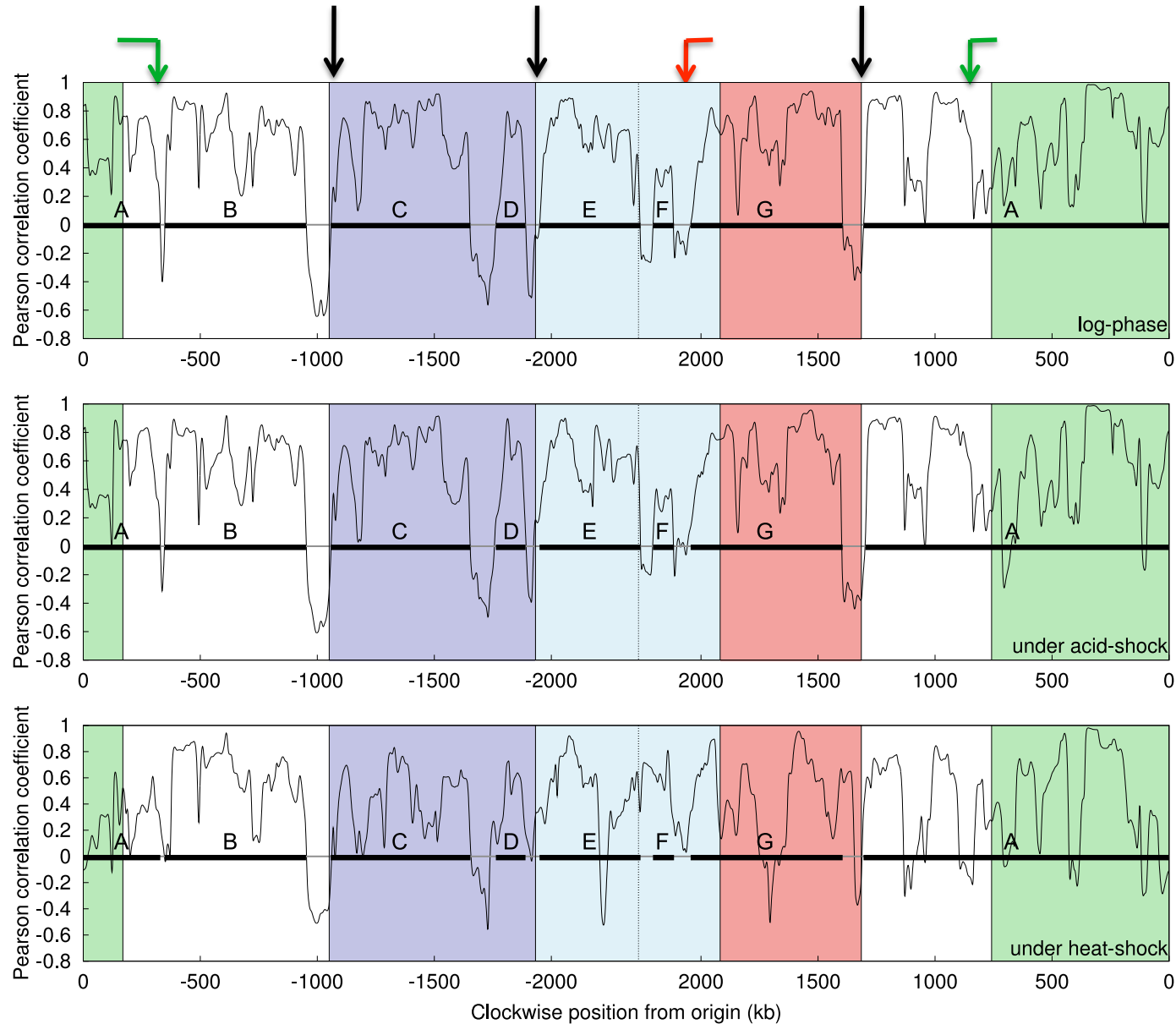


7 sectors for *E.coli*



No random genome with at most 8 sectors displays all sectors where strips cover at least 50% of peaks

Sectors and Boccard's domains

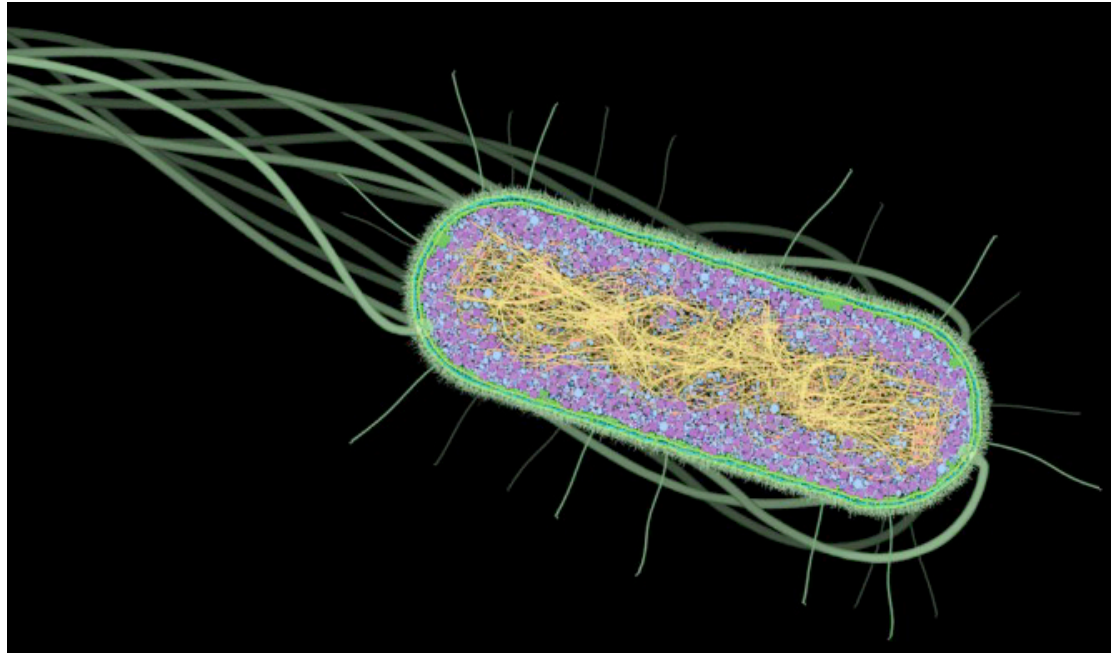


Bounds of Boccard's domains and sectors bounds.

Conclusions

- *E. coli* periodic gene distribution is identified for a periodic interval of 33 kb.
- Two positional networks of genes are discovered by studying gene periodic distribution: one is driven by metabolic genes and the other by genes involved in cellular processing and signaling.
- A functional core of *E.coli* genes drives gene periodic distribution.
- A few chromosomal regions that preserve gene transcription profiles across environmental changes are identified.
- This single genome analysis approach can be taken as a footprint for a large-scale bacterial and archaeal periodic distribution analysis.

How can we accomodate a chromosome into a *E.coli* container with 0.5-0.75 μM diameter?



Consider the largest point-to-point distance of the solenoid with 33kb turn, and assume 50 nm persistence length of DNA. Such solenoid perfectly accommodates in a *E. coli* container with 0.5-0.75 μM diameter.

Acknowledgments



Anthony Mathelier
PhD in september 2010

F.Képès
Institute of Synthetic and Systems Biology
Génopole, EVRY

A.Mathelier, A.Carbone (2010) Chromosomal periodicity and positional networks of genes in *Escherichia coli*, *Molecular Systems Biology*, 6:366.

Algorithm and microbial SSCI codon space :

- **F.Képès**, CNRS & génopole Evry
- **A.Zinovyev**, IHÉS & Institut Curie (Paris)

A. Carbone, A. Zinovyev, F. Képès, Codon adaptation index as a measure of dominating codon bias, *Bioinformatics*, **19**, 2005–2015, 2003.

A. Carbone, F. Képès, A. Zinovyev , Codon Bias Signatures, Organization of Microorganisms in Codon Space, and Lifestyle, *Molecular Biology and Evolution*, **22**, 547–561, 2004.

Metabolic networks comparison :

- **D.Madden**, IHÉS & IGI (USA)

A. Carbone, R. Madden, Insights on the Evolution of Metabolic Networks of Unicellular Translationally Biased Organisms from Transcriptomic Data and Sequence Analysis, *Journal of Molecular Evolution*, **59**, 1–25, 2005.

Minimal gene sets :

A.Carbone, Computational prediction of genomic functional cores specific to different microbes, *Journal of Molecular Evolution*, **63**, 733-746, 2006.