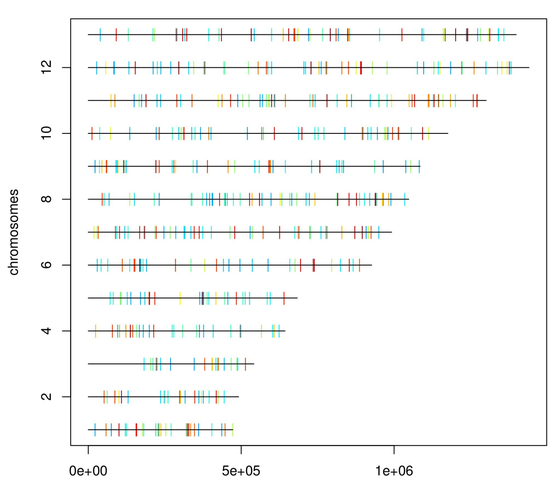
# Genome periodicity analysis using Fourier transforms in *E. coli* and yeasts

Raphaël Champeimont, Alessandra Carbone
raphael.champeimont@upmc.fr, Alessandra.Carbone@lip6.fr
Laboratoire de Génomique des Microorganismes, CNRS - Université Pierre et Marie Curie

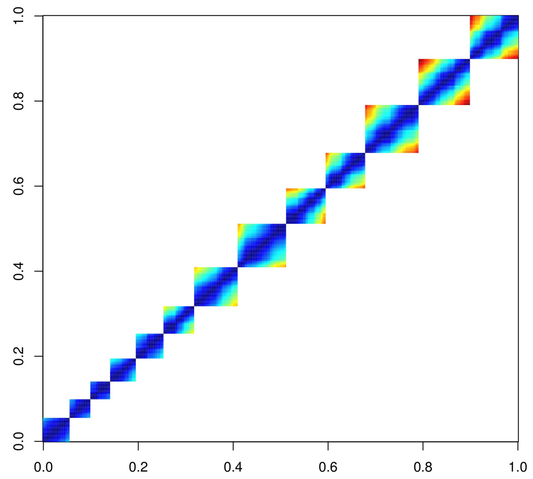
## General methodology

### Studied positions on genome



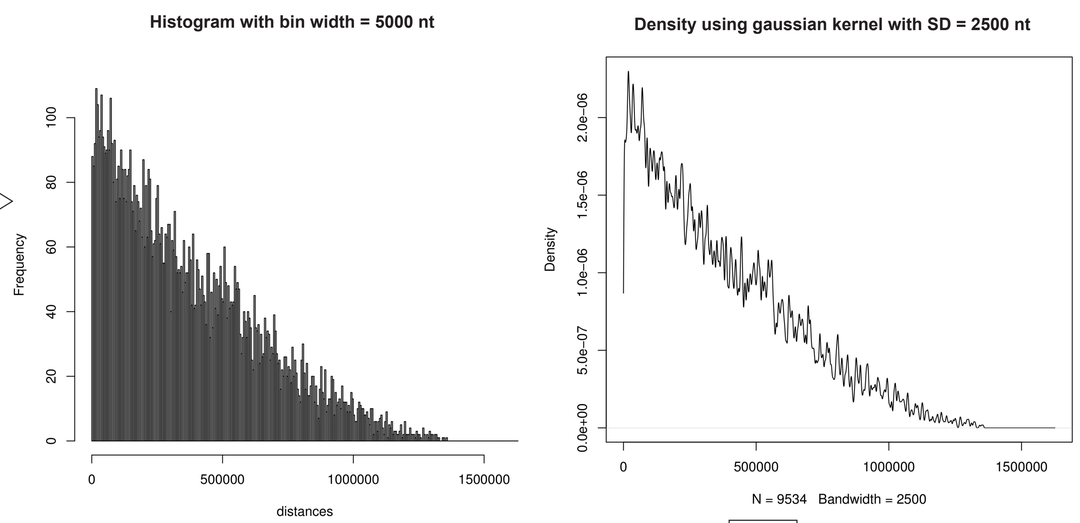

Our input data is a set of positions on the genome.
In the case of yeast, they are on several chromosomes.

### Distances between these positions




We compute the distances between every pair of genes in the same chromosome. In circular genomes we take the smallest distance of the two strands.

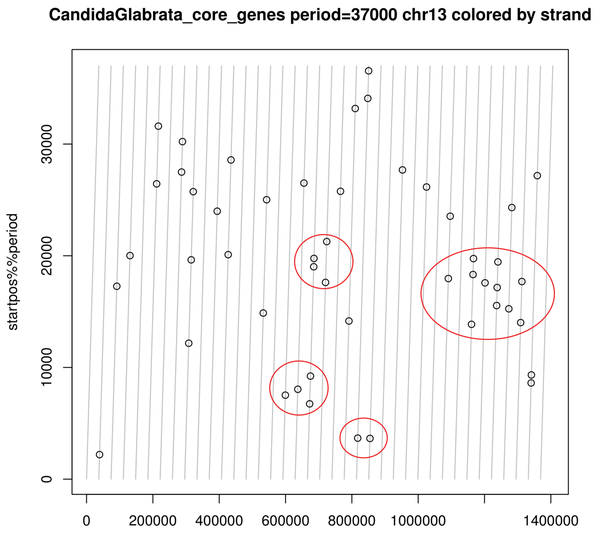
### Estimate distribution (with an histogram or a kernel density estimator)




We compute an histogram with a bin width of 5000 nucleotides, or use a kernel density estimator based on a gaussian kernel with a standard deviation of 2500 nucleotides. With this second method, the height of the peaks in the spectrum is more robust.

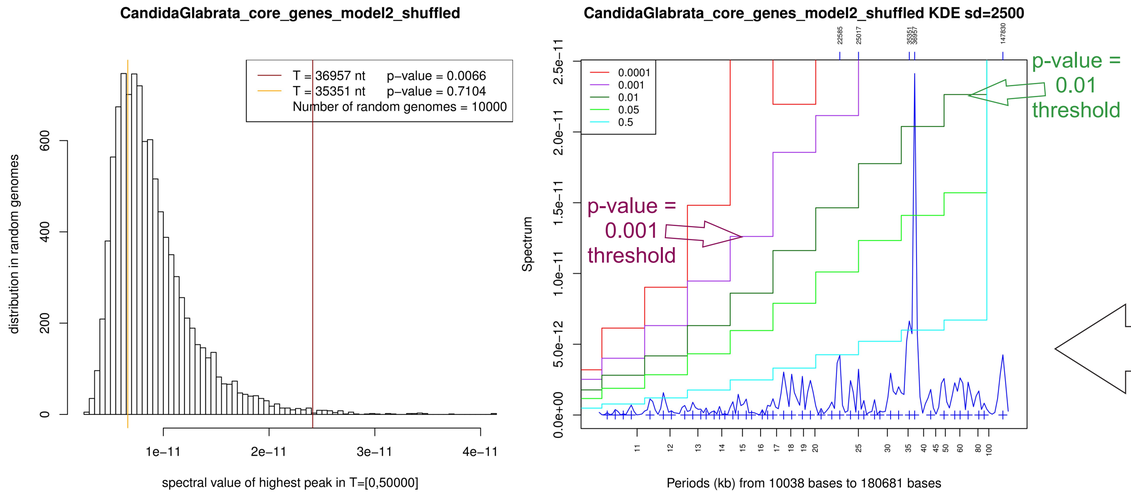
### Fast Fourier Transform

### Project positions modulo the period



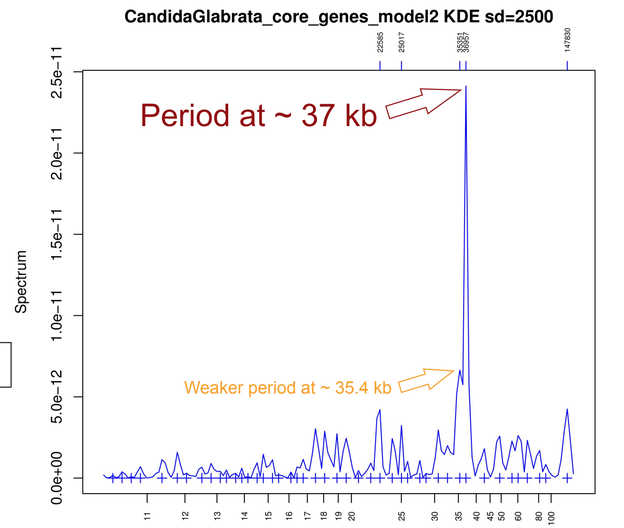

We plot our studied positions: The X axis is the position on the genome, while the Y axis is the position modulo 37 000 nucleotides (the period we found). Projected positions tend to cluster, ie. groups of consecutive genes are separated by distances multiples of 37 000 nt (examples highlighted in red).

### Assess significance using random models




We generate a **random model of the genome**. We try both a uniform model (positions are put randomly on the genome with a uniform distribution) and a "shuffled" model (we shuffle the genes and intergenic regions, so that the gene length and intergenic region length distribution are preserved). For 10000 randomly generated data sets, we run the same procedure as for the real data, and compare the intensity of the peaks to those of the real data. On the left, we show the distribution of the height of the highest peak in the [0,$10^5$ nt] period interval. On the right, for several period intervals (the width of the steps), we show different quantiles (0.0001 red, 0.001 purple, 0.01 dark green, 0.05 light green, median cyan). As can be seen, peaks in random data are higher for high periods, so we need to take that in account when assessing significance of our period.

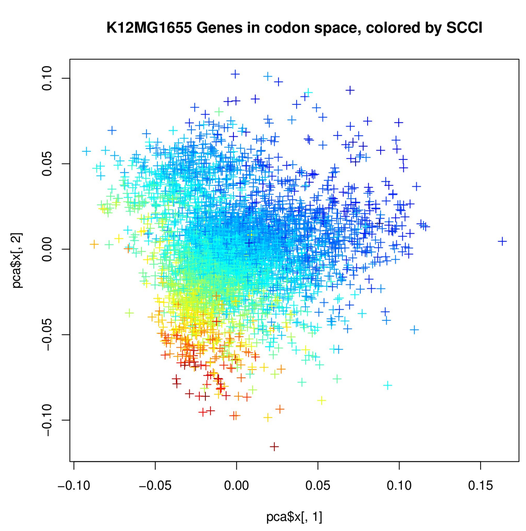
### Fast Fourier Transform




Here we show the periodogram, ie. for each period we show its weight in the input signal.

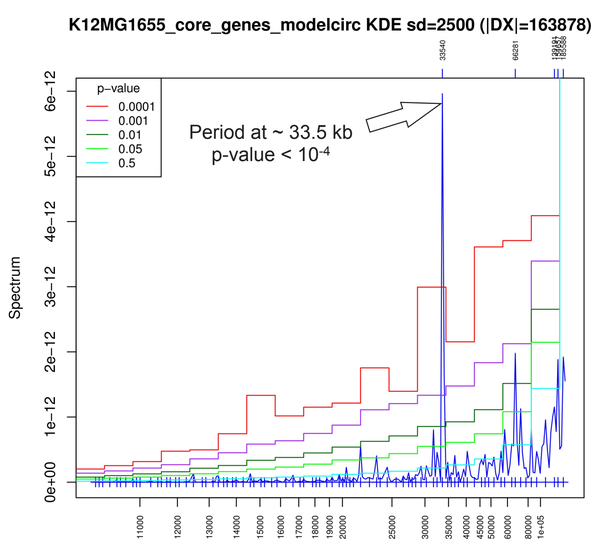
## Application: codon biased genes in *E. coli*

### Codon Adaptation Index of *E. coli* genes



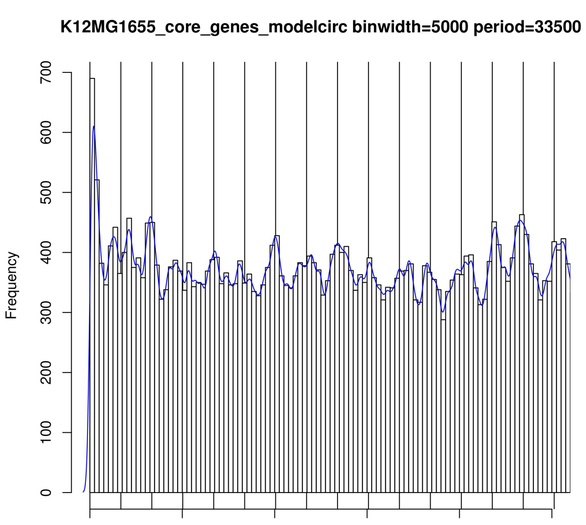

We compute the Self-Consistent Codon Index (a measure of the Codon Adaptation Index that does not need a reference set [1]) and consider the set of genes with the highest score (> mean + SD).

### Periodicity analysis with Fourier Transform



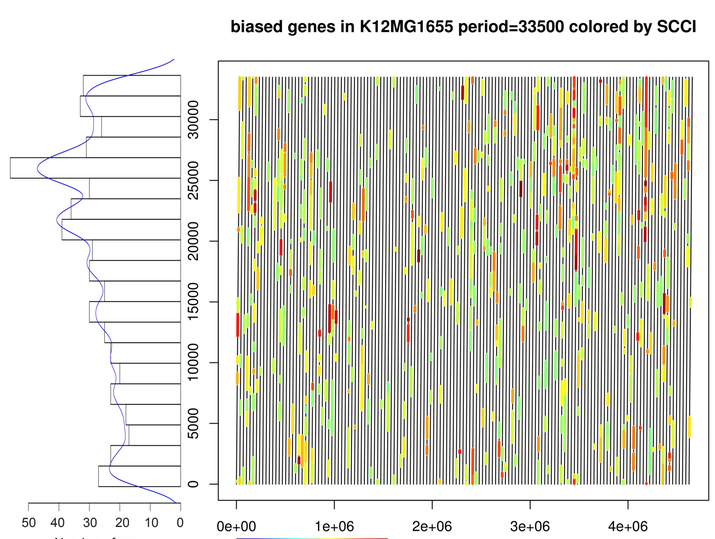

After applying our method, we get this periodogram. We clearly see a peak around 33.5 kb.

### Codon biased genes distance distribution



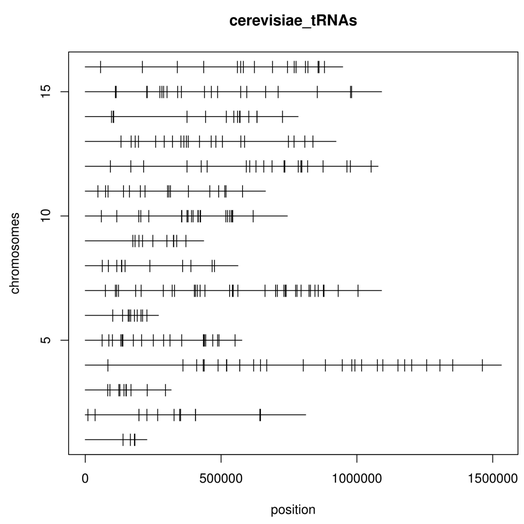

Now we have found this period, we can go back at the distance distribution and use an appropriate bin width for the histogram to see our period. Here we have added vertical lines at distances multiples of 33.5 kb, so we can see distances tend to occur more at these values.

### Projected positions of codon biased genes

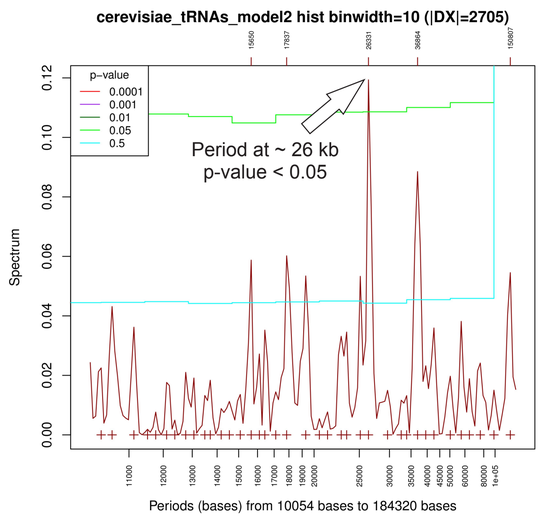



Here we project the codon biased genes positions modulo the period we found. The X axis is the absolute position while the Y axis is the position modulo 33.5 kb. Genes are colored differently depending on how much they are biased. On the left an histogram shows the distribution of these projected positions.

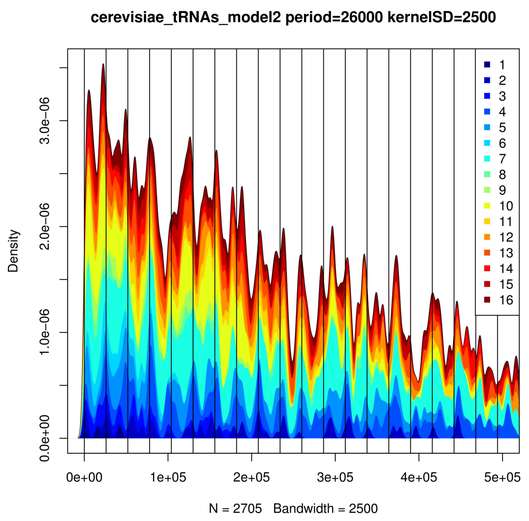
## Application: tRNAs in *S. cerevisiae*




tRNAs positions in *S. cerevisiae*, on the 16 chromosomes




Periodogram computed by Fast Fourier Transform, with p-value thresholds in green and cyan.




Distribution of distances between tRNAs, colored by contribution from each chromosome.

## References

[1] Codon adaptation index as a measure of dominating codon bias. Carbone A, Zinovyev A, Képès F. *Bioinformatics*. 2003

[2] Chromosomal periodicity and positional networks of genes in Escherichia coli. Mathelier A, Carbone A. *Mol Syst Biol.* 2010 May 11

[3] Chromosomal periodicity of evolutionarily conserved gene pairs. Wright MA, Kharchenko P, Church GM, Segrè D. *Proc Natl Acad Sci USA*. 2007 Jun 19;104(25):10559-64. Epub 2007 Jun 11.