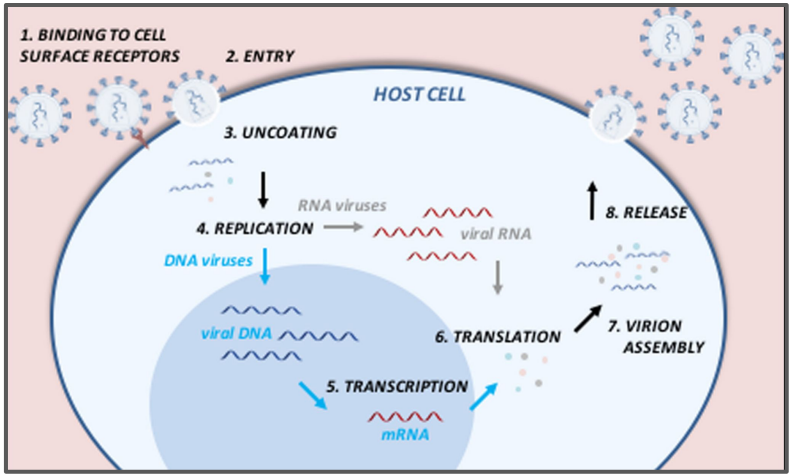
Analysis and evolution of immunogenic dinucleotide motifs in SARS-CoV-2

Andrea Di Gioacchino

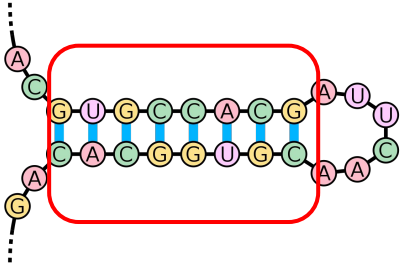
Work in collaboration with: P. Sulc, A. Komarova, B. Greenbaum, R. Monasson and S. Cocco

How to recognize viral mRNA

Pathogen-associated molecular patterns (CpG motifs, long double stranded regions) are used by the innate immune system to detect viral RNA.



From Ruggiero, Richter, *Nucleic Acid Res.* 2018.



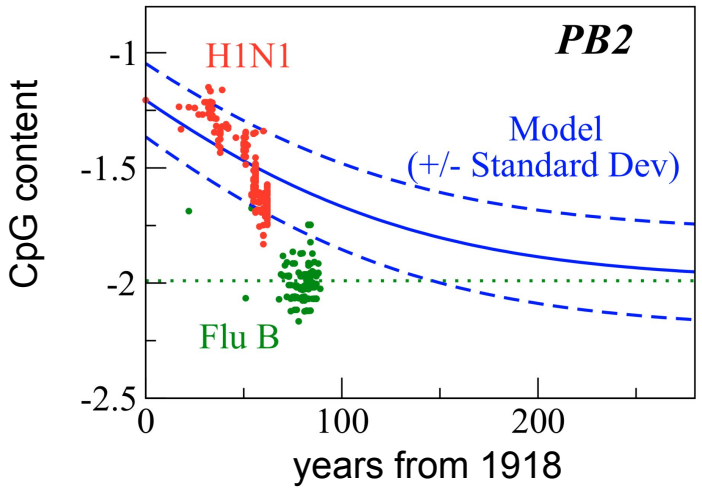
...ACT**CG**AAC...

Sauter, Kirchhoff, *Plos Biol*. 2021.

Known histories of viral mimicry

The H1N1 (PB2 segment) CpG content evolution:

...ACT**CG**AAC...



From Greenbaum, Cocco, Levine, Monasson, *PNAS* 2014.

Time to become “self” is ~100-300 years (for H1N1).

CpG motifs in SARS-CoV-2

We investigated the genomic content of pathogen-associated dinucleotide motifs possibly involved in the innate immune system response.

Why this can be useful:

1 - Possibility to build better antiviral therapies.

2 - Optimize vaccine strains.

3 - Understand virus evolution (in contact with the human host).

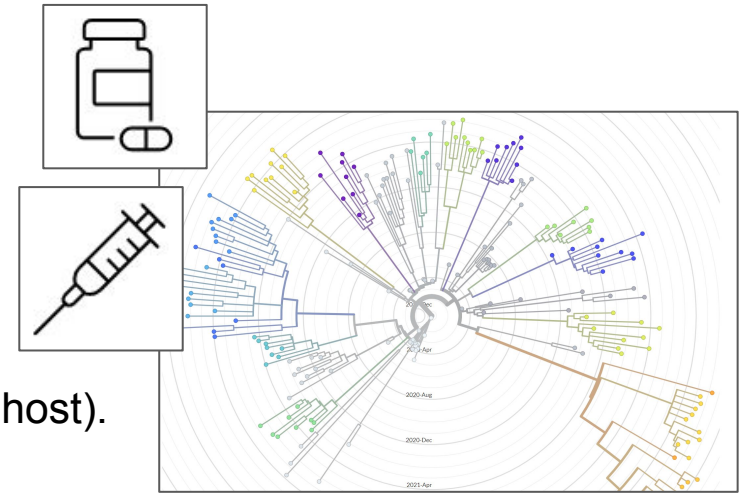
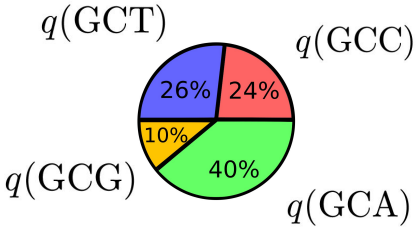


Image from nextrain.org

ADG, P Šulc, A Komarova, B Greenbaum, R Monasson, S Cocco, *Mol. Biol. Evol.* 2021.

How to estimate CpG content?

CpG force [Greenbaum, Cocco, Levine, Monasson, *PNAS* 2014]:



Alanine codon bias

$$p(\boldsymbol{s}) = \frac{1}{Z}\left(\prod_{i=1}^{L/3} q(c_i)\right) e^{f N_{\mathrm{CG}}(\boldsymbol{s})}$$

Additional pressure

Null model / Prior information

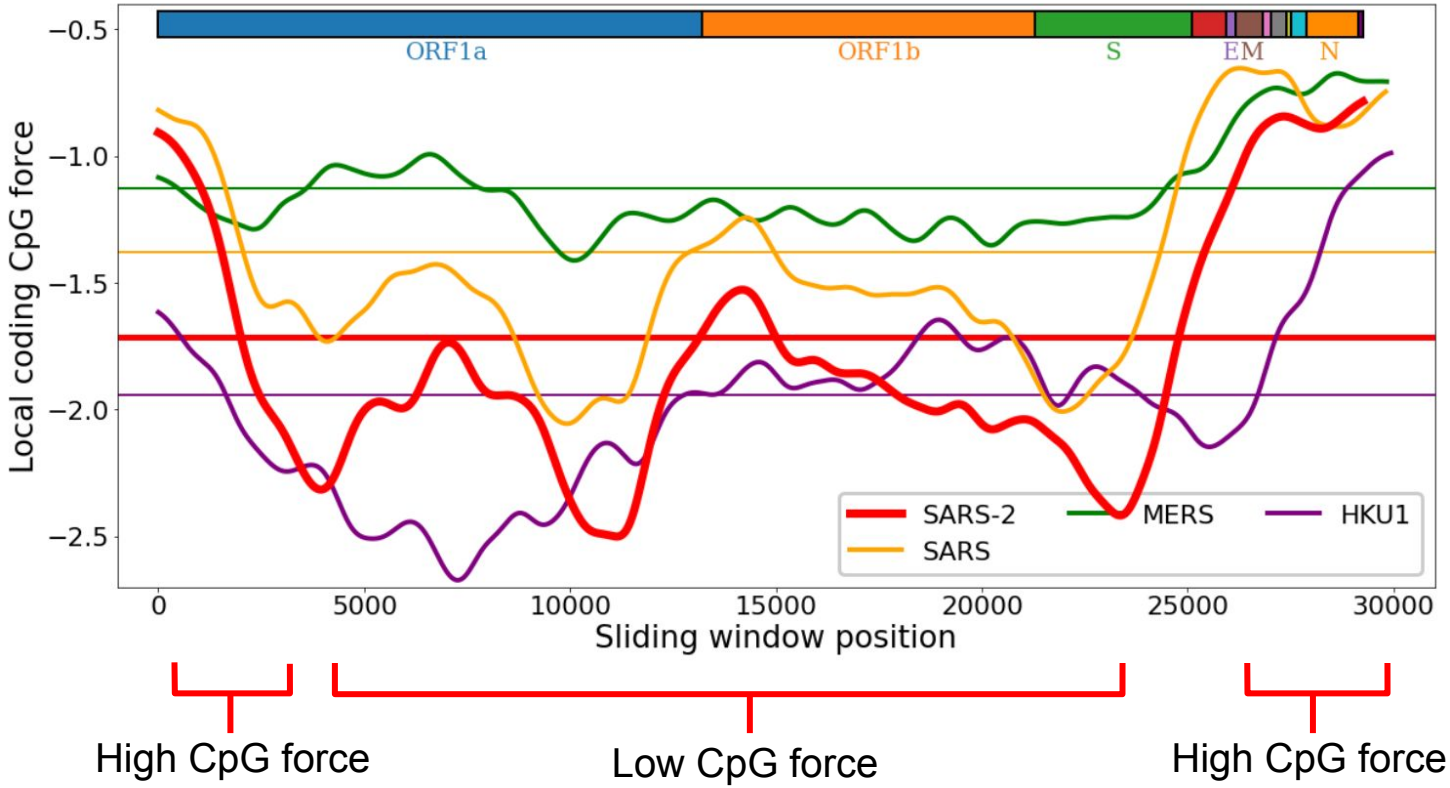
$$Z = \sum_{\boldsymbol{s}\in\mathcal{S}} \left(\prod_{i=1}^{L/3} q(c_i)\right) e^{f N_{\mathrm{CG}}(\boldsymbol{s})}$$

- Formalism borrowed from statistical physics.
- The partition function Z is computed through the transfer matrix method.

$f > 0 \rightarrow$ more CG than in the null model

$f < 0 \rightarrow$ less CG than in the null model

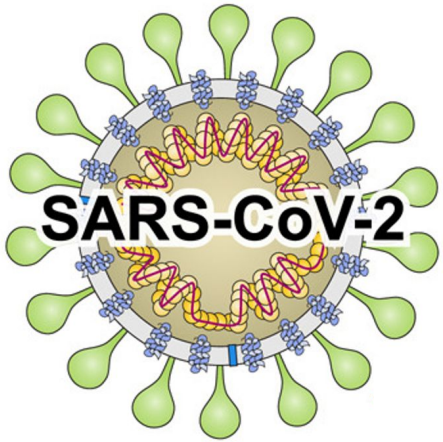
CpG motifs in SARS-CoV-2 genome: local view

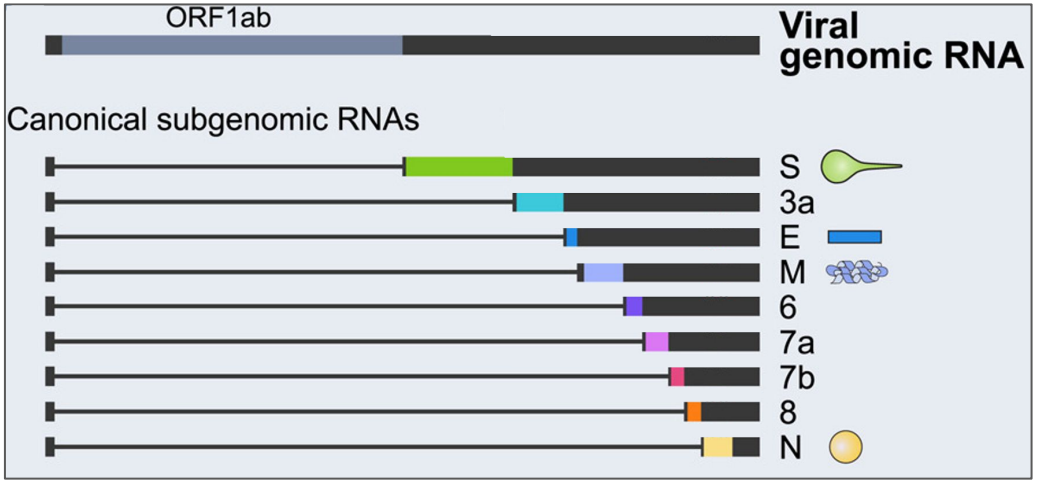


ADG, P Šulc, A Komarova, B Greenbaum, R Monasson, S Cocco, *Mol. Biol. Evol.* 2021.

The viral genome is not uniformly expressed

Coronaviruses use a mechanism of discontinuous transcription to express subgenomic RNAs.

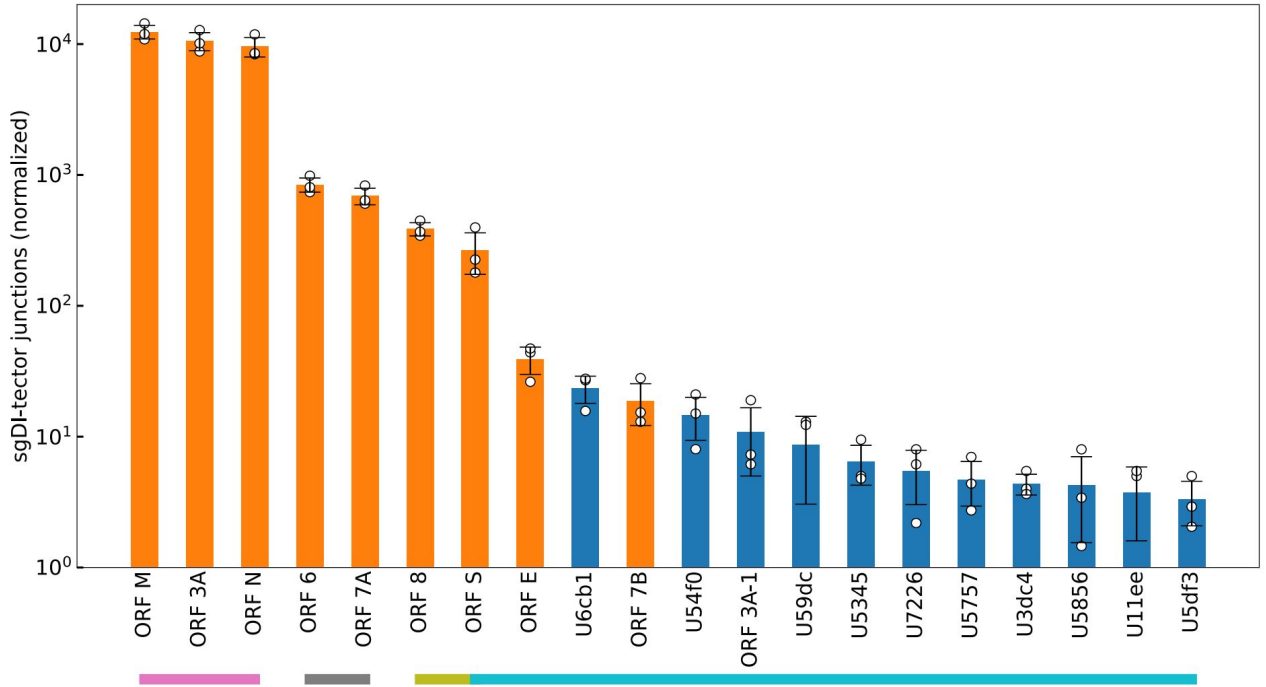




[image adapted from Kim *et al*., "*The architecture of SARS-CoV-2 transcriptome*", Cell 2020]

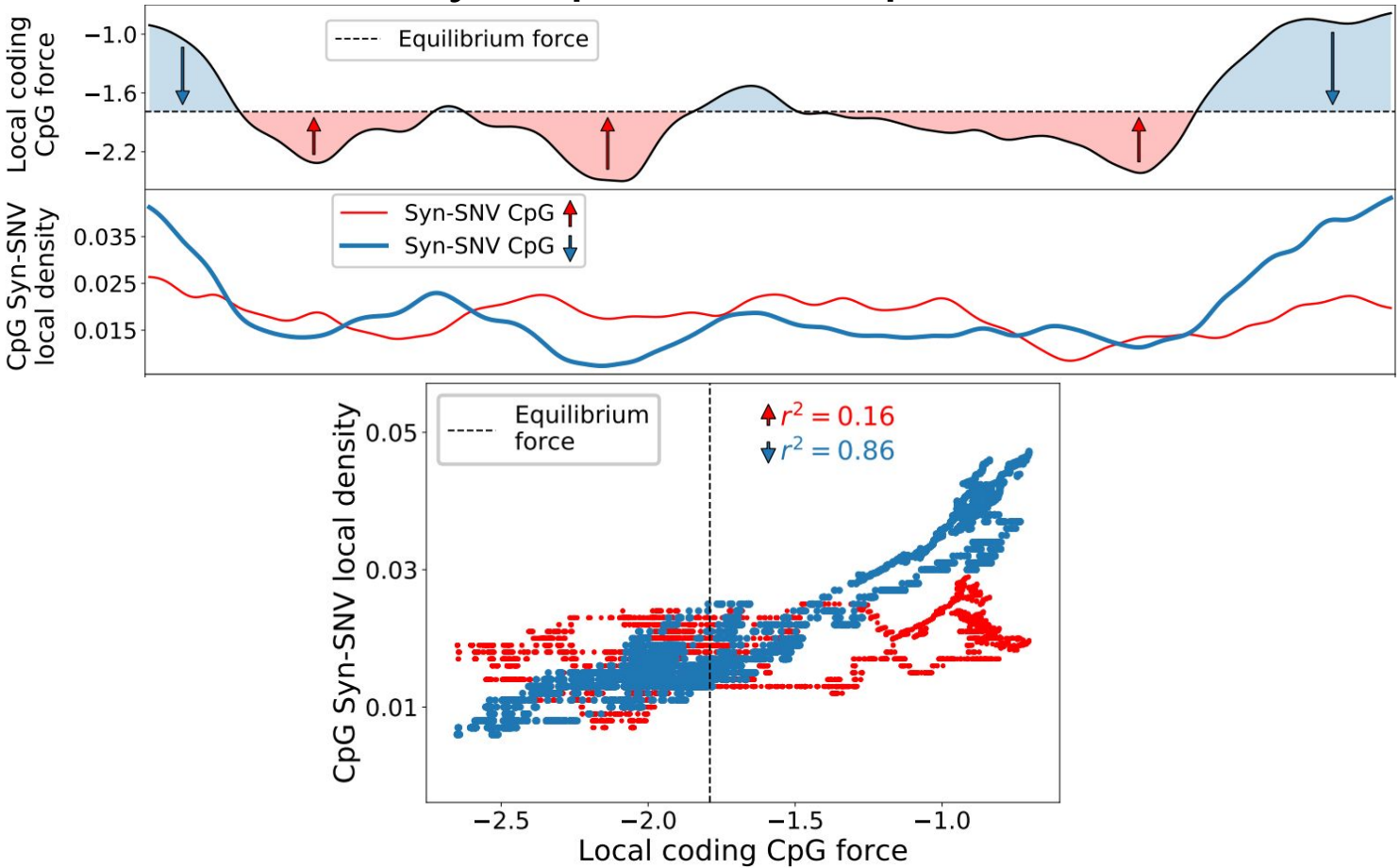
High-CpG regions are among the most expressed

We developed sgDI-tector to discover and quantify subgenomic RNAs from raw short-read data.



ADG, R Legendre , Y Rahous, V Najburg, P Charnau, B Greenbaum, F Tangy, S Van der Werf, S Cocco, *RNA* 2022.

Back to viral mimicry: equilibrium CpG force



ADG, P Šulc, A Komarova, B Greenbaum, R Monasson, S Cocco, *Mol. Biol. Evol.* 2021.

Introducing the synonymous mutation score

- Evolve towards “equilibrium” CpG content (from HKU1, OC43, NL63, 229E);
- Follow the virus codon usage bias.
- Favour transitions (A<->G, C<->T) over transversions (others).

$$\mathrm{SMS}(c \to c') = (f_{\mathrm{eq}} - f)\Delta N_{\mathrm{CG}} + \log\left(\frac{q(c')}{q(c)}\right) + (n_{\mathrm{tn}} - n_{\mathrm{tv}})\log 2$$

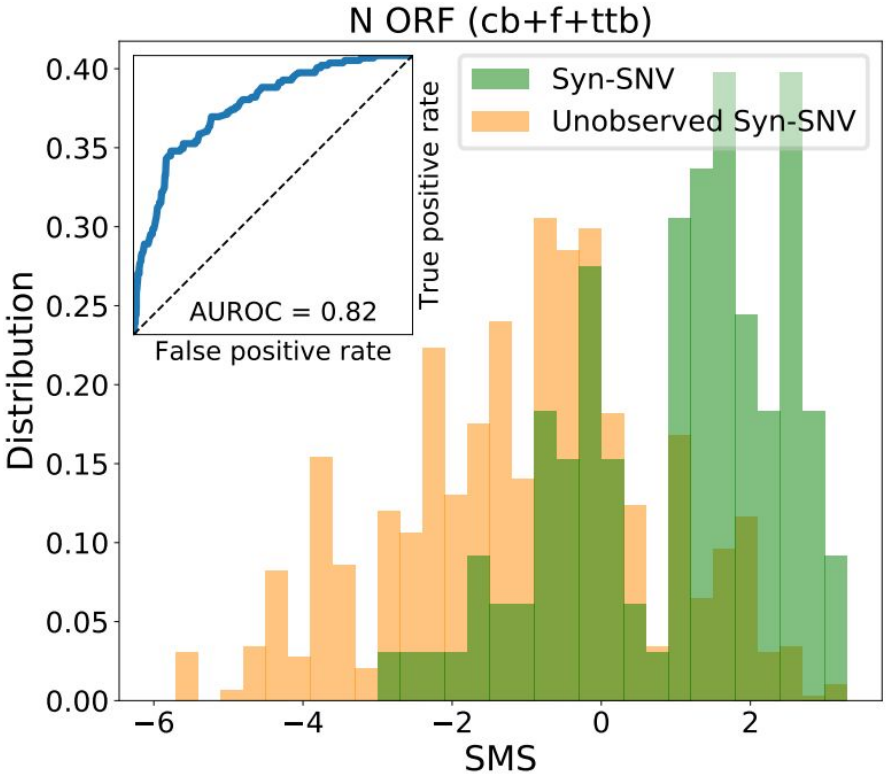
CpG drive

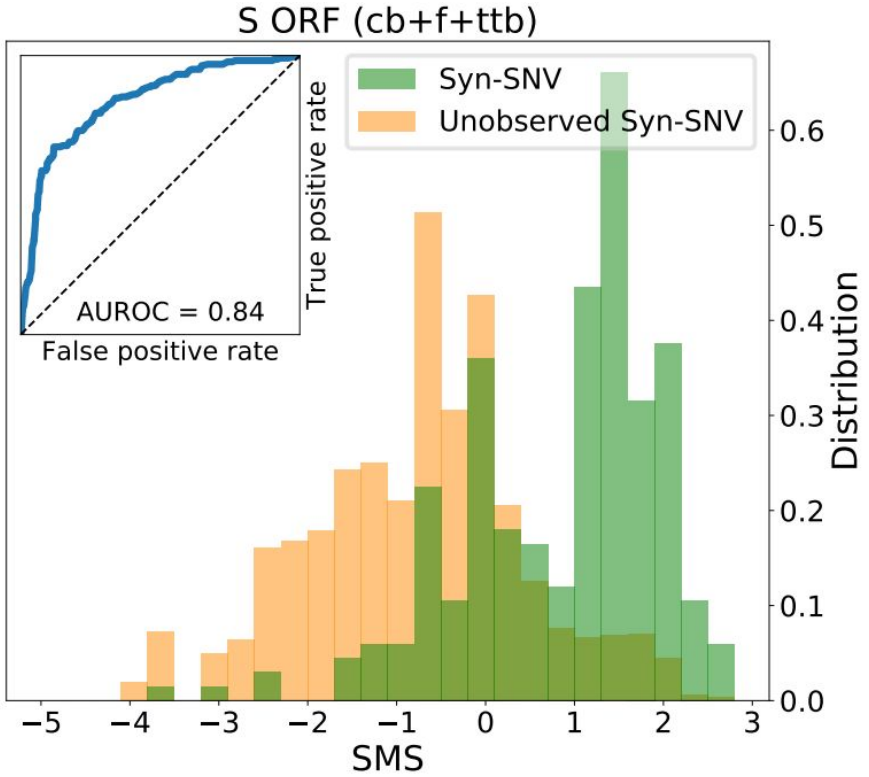
Virus codon bias

Transitions vs transversions bias

ADG, P Šulc, A Komarova, B Greenbaum, R Monasson, S Cocco, *Mol. Biol. Evol.* 2021.

“Predicting” observed synonymous mutations





ADG, P Šulc, A Komarova, B Greenbaum, R Monasson, S Cocco, *Mol. Biol. Evol.* 2021.

Conclusions

SARS-CoV-2 has low CpG force, with large variance (compatible with a mosaic genome [Temman *et al., Nature* 2022]):

1. The overall genome (and in particular the S protein) has low CpG content.
2. The N protein coding region has large CpG force.
3. Regions of SARS-CoV-2 genome are differentially expressed, N being the most common.

Our model provides some predictive power for synonymous mutations:

1. CpG number is expected to decrease in the N ORF, and to increase in the S ORF (over long timescales).
2. Mutations observed in the first year of the pandemics have large SMSs.

Thank you for your attention!

Backup

What about more recent sequences?

The Slowing Rate of CpG Depletion in SARS-CoV-2 Genomes Is Consistent with Adaptations to the Human Host

Akhil Kumar, Nishank Goyal, Nandhini Saranathan, Sonam Dhamija, Saurabh Saraswat, Manoj B Menon, Perumal Vivekanandan ✉ Author Notes

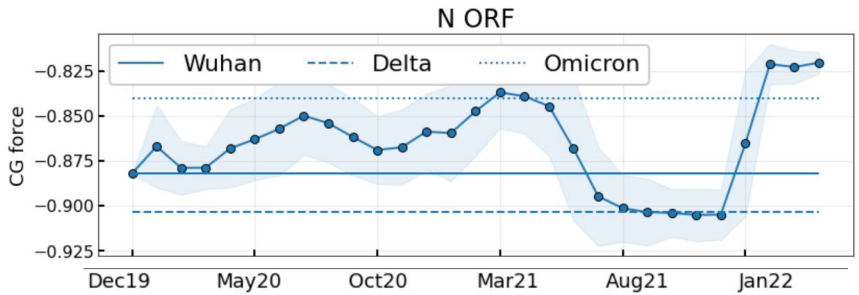
Molecular Biology and Evolution, Volume 39, Issue 3, March 2022, msac029, https://doi.org/10.1093/molbev/msac029
Published: 03 February 2022

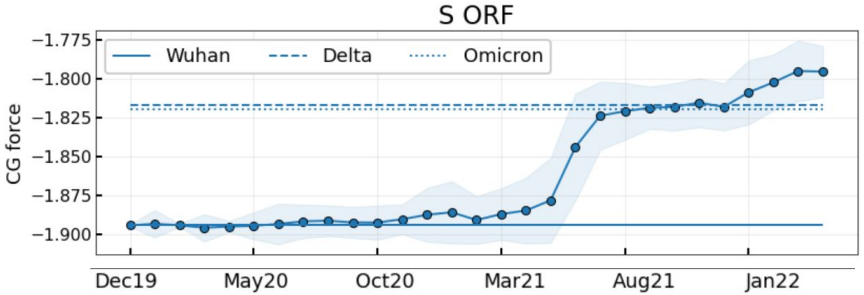
Article | Open Access | Published: 14 February 2022

The low abundance of CpG in the SARS-CoV-2 genome is not an evolutionarily signature of ZAP

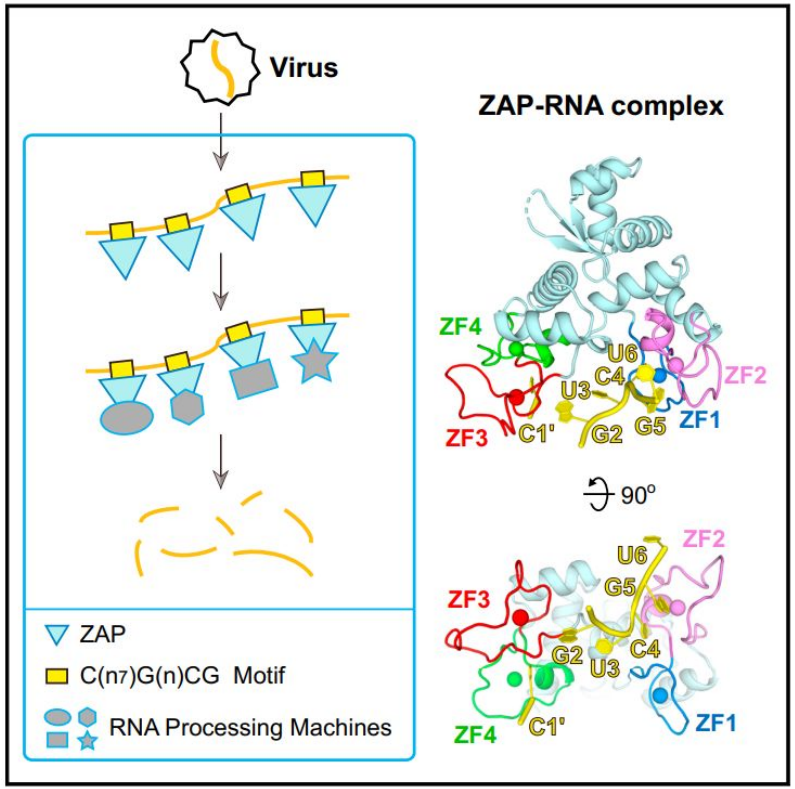
Ali Afrasiabi, Hamid Alinejad-Rokny, Azad Khosh, Mostafa Rahnama, Nigel Lovell, Zhenming Xu & Diako Ebrahimi ✉

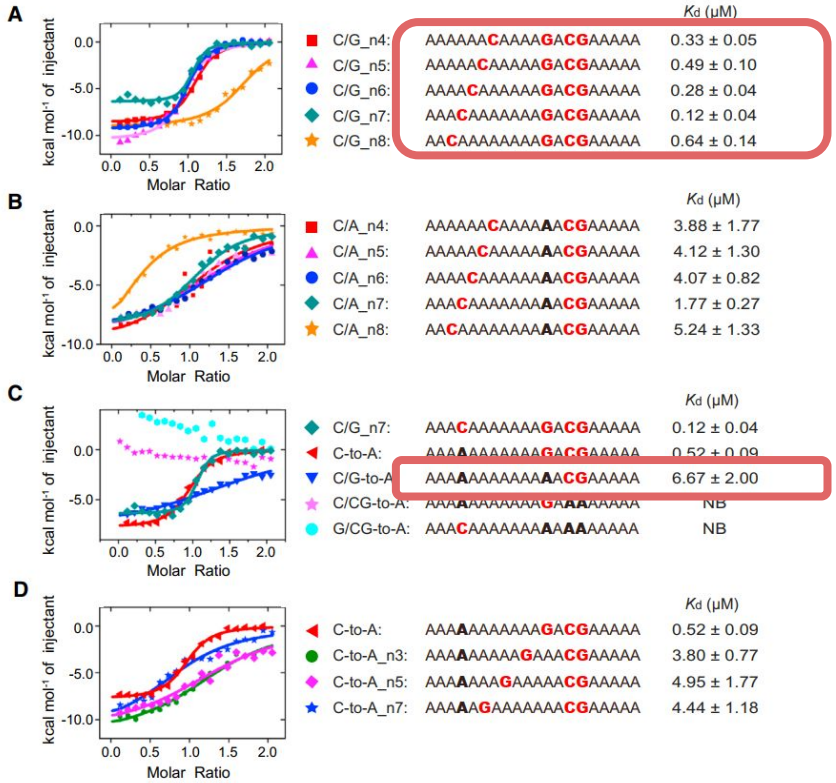
Scientific Reports **12**, Article number: 2420 (2022) | Cite this article
748 Accesses | **1** Altmetric | Metrics





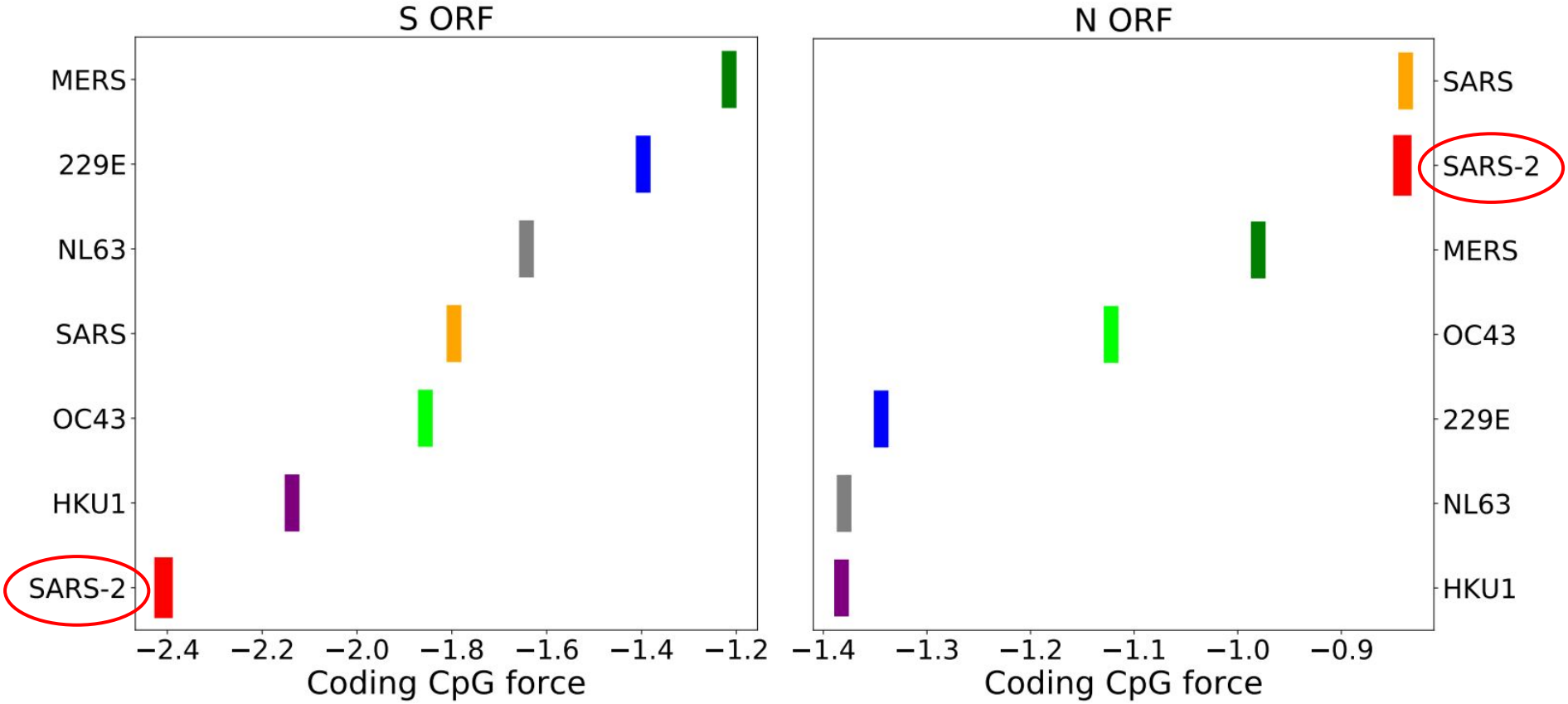
The role of Zinc-finger Antiviral Protein (ZAP)





Luo, Wang, Gao, Zhu, Liu, Gao, Gao, Cell Reports 2019.

CpG forces on structural proteins: comparison



The importance of CpG motifs - lesson from HIV

Synonymous variants of HIV with increased number of CpG motifs cannot replicate efficiently when the Zinc-finger Antiviral Protein (ZAP) is present.

Published: 27 September 2017

CG dinucleotide suppression enables antiviral defence targeting non-self RNA

Matthew A. Takata, Daniel Gonçalves-Carneiro, Trinity M. Zang, Steven J. Soll, Ashley York, Daniel Blanco-Melo & Paul D. Bieniasz

Nature **550**, 124–127 (2017) | Cite this article

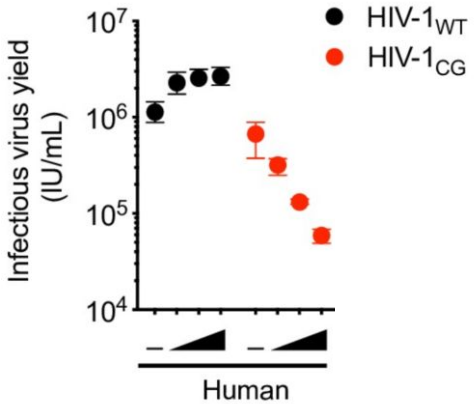


Image from Goncalves-Carneiro, Takata, Ong, Shilton, Bieniasz, *Plos Path*. 2021.