Addressing the Yan report

In September 2020, the above claimed to **be** scientific evidence for SARS-CoV-2 being an engineered bioweapon 😨 Zenodo granting it a Digital Object Identifier (DOI) made the report appear credible, despite a lack of peer review.

The Johns Hopkins University made up for that lack the same month, explaining why the report was unconvincing. But conspiratorial audiences 🤨🤨 value neither authority, nor being pointed to tonnes of reading material (can't blame them for the latter 😄).

They do value critical thinking, which only requires that **the core claim** be verified. Should it prove false, everything else can be dropped 🤷‍♀️ What was the report's postulate, then, that sufficed for its authors to be able to seek asylum in 🇺🇸?

Restriction enzymes around the spike's receptor binding motif

Mikolaj Raszek, PhD, was kind enough to elucidate, in *SARS-CoV-2 coronavirus origins alternative theories – do they hold up against science?*, the core claim of the Yan report.

Two **restriction enzymes** (sequences bacteria use to slash viruses to bits, repurposed by humans to glue parts of different genomes together): EcoRI and BstEII. According to Yan et al, the sequence between them allowed to target mammals larger than 🦇🦇.

Download the earliest known SARS-CoV-2 genome ⬇️ (1 of 2)

Yan et al's image caption cites the isolate **Wuhan-Hu-1** (isolate: a population of organisms having little genetic mixing with other organisms of the same species).

> **Figure 5. Two restriction sites are present at either end of the RBM of SARS-CoV-2, providing convenience for replacing the RBM within the spike gene.** *A. Nucleotide sequence of the RBM of SARS-CoV-2 (Wuhan-Hu-1).*

Viewing the isolate at *NCBI Virus*, the absolutely earliest accession (unique sequence identifier) is MN908947.1, collected in Dec 2019 ➡️ submitted 2020-01-05 ➡️ released 2020-01-12.

That's 2 months until the World Health Organization would declare COVID-19 a pandemic ☣️ (2020-03-11).

Download the earliest known SARS-CoV-2 genome ⬇️ (2 of 2)

In the accession page, switching to the FASTA format (a text format often used for storing reference genomes) allows us to download the troublemaker's genome:
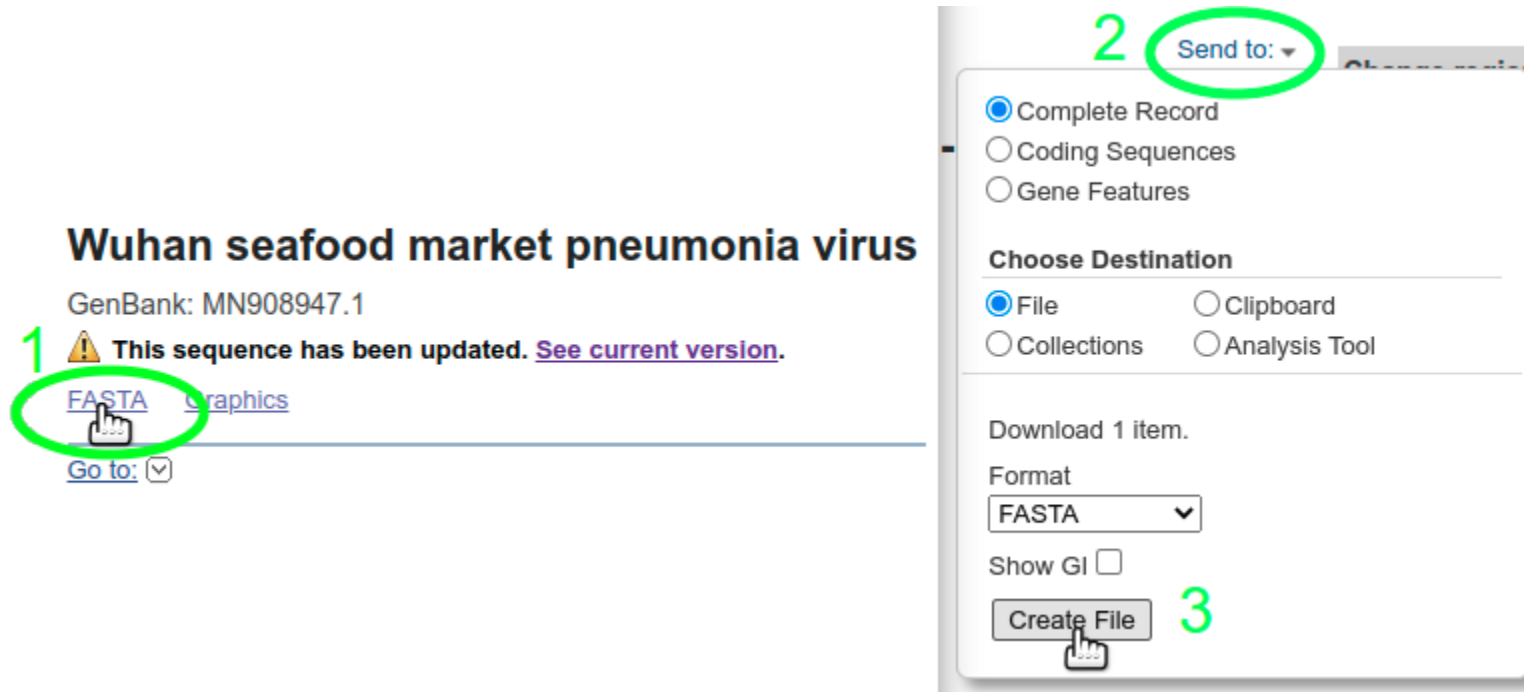
~30k bases (a base is one of `A, C, G, T`) long? What a tiny genome. A human one is 3.1 billion bases, with a single cell taking up between 3.3 GB (reference genome, a measurement standard) and 70 GB (non-reference genome) of your hard drive 🙈

Are EcoRI and BstEII *actually there?* 🔍👀

- **Note:** Sequences identical to those listed **needn't** necessarily come from restriction enzymes - but let's simplify and humour that notion 👌👩

You can open the downloaded SARS-CoV-2 genome in a text editor 📋, and search (`Ctrl+f` / `Cmd+f`) for the occurrences of the EcoRI sequence **GAATTC** yourself. If you fancy a dopamine rush, **stop reading and go ahead now** 😁

The **N** (= whichever base) in BstEII's **GGTNACC** is a tad more problematic, though. If you can locate *regular expression mode* (look for a button marked `.*`) 🤞, this hurdle can be cleared by inputting **GGT[ACGT]ACC**.

# Plotting EcoRI & BstEII sequence matches in the spike gene
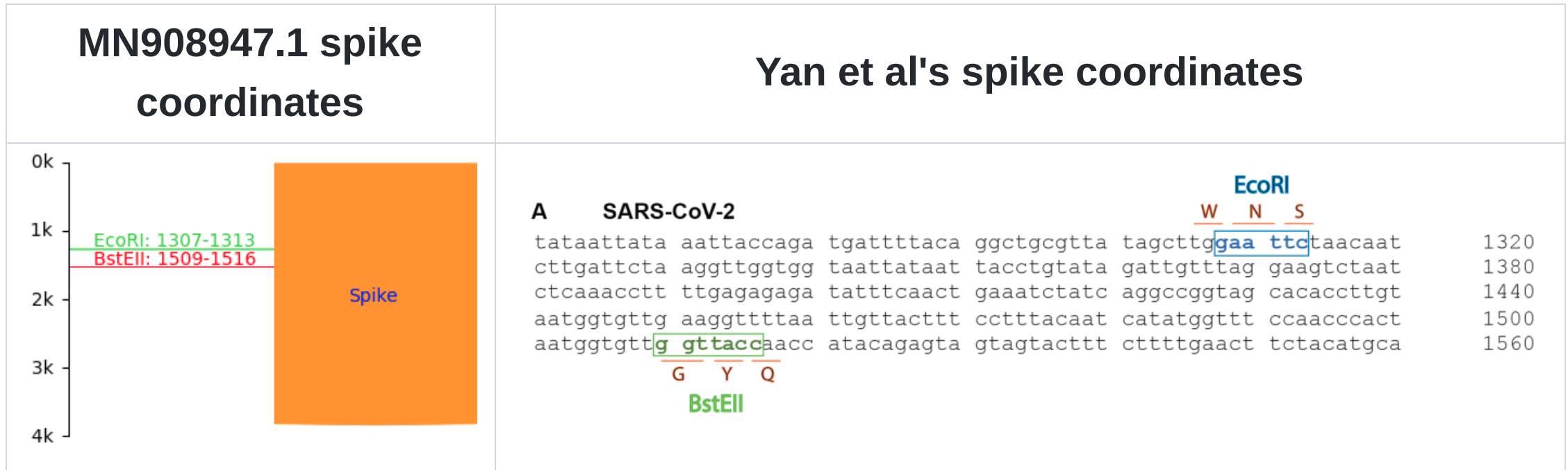
The accession MN908947.1 spike gene **does** contain sequence occurences with 100% identity to EcoRI & BstEII, and that's at the **exact** coordinates specified by Yan et al 🎯 So far so good - let's look at the rest of the genome 🔍

Plotting EcoRI & BstEII matches across the whole genome

But looking at **all the genes** (instead of just the spike), one seems to find more 'genetic modifications' than Yan et al bargained for 🤔

There's even an EcoRI match in the 3' untranslated region (nothing there ever becomes live proteins, hence there's no point in engineering the region).

Sequences identical to EcoRI or BstEII in the SARS-CoV-2 genome (accession MN908947.1)
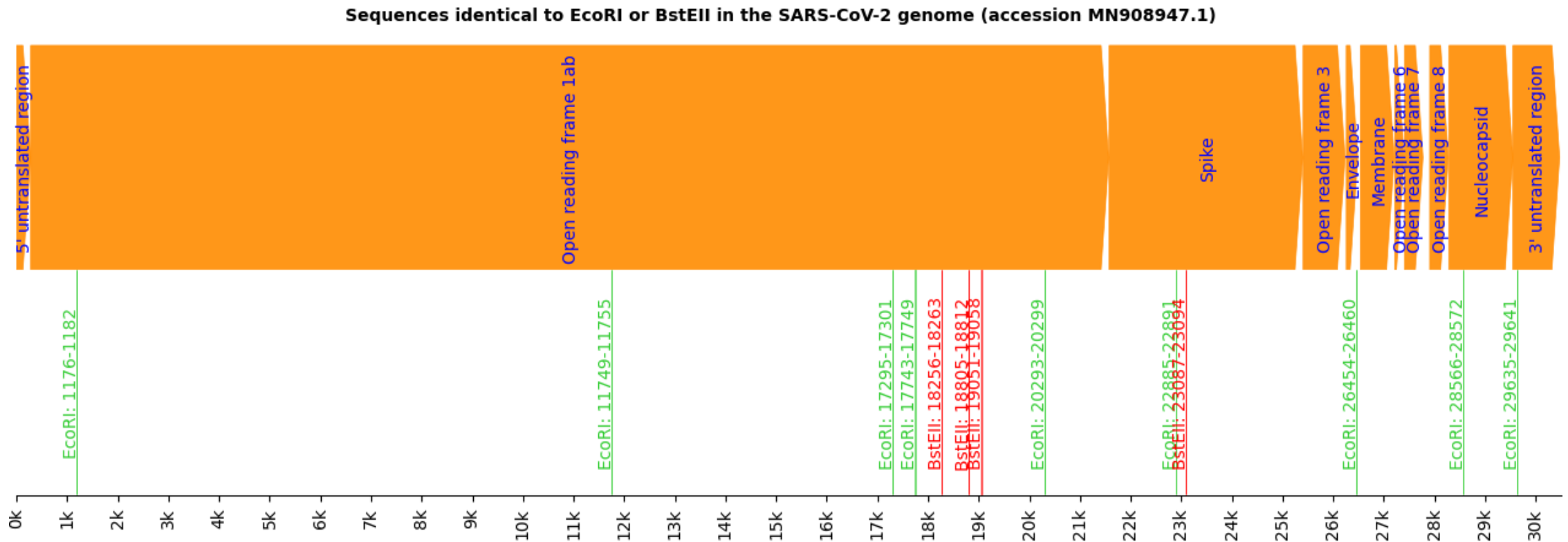
A restriction enzyme cornucopia? 🦄 Let's find out 🤷‍♀️

Bioinformatics Algorithms: An Active Learning Approach gives the formula (search for `approximation`) for approximating the likelihood that a **k-mer** (word of size k) occurs in a text *by random chance alone* 🎲

The **lower** ⬇️ that likelihood, the **more probable** ⬆️ any bioengineering 🧬✂️🧬 Customarily, values with `< 5%` chance of being randomly generated, are worthy of investigation.

Click here for the Python version of the approximation formula 🐍. Its code's been tested, so should be reliable. Let's take it for a spin 🧶🐱

⚔️ Theory vs practice: probabilities along the full genome 🧬

1️⃣ A nice property of our approximation formula: if we seek the probability of **just a single occurrence**, any returned number `> 1.0` is the **expected occurrence count**.

2️⃣ BstEII's middle character (GGT**N**ACC) can be anything, so BstEII is considered to have length 6 (the same length as EcoRI), instead of 7.

Restriction enzyme	Expected occurrences	Actual occurrences
EcoRI (**GAATTC**)	7.44	9 (*...are Yan et al onto something?*)
BstEII (**GGT_ACC**)	7.44	4 (*...no they aren't*)

No conclusive evidence either way yet 🤷‍♀️ Let's concentrate on the spike 👀

Occurrence probabilities within the spike gene 📌

[The accession page](#) informs us that the range of the `"S"` gene is `21579..25400`, which makes for a length of `3821`. Plugging this text length into our formula 🔌, we get:

```
In [3]: ProbabilityOfKmerOccurringNTimesInText(alphabet_size=4)(
   ...:       text_length=3821, kmer_length=6, kmer_occurrence_count=1
   ...: )
Out[3]: 0.931640625
```

There's a 93% probability of at least one sequence of length 6 (doesn't matter if it's EcoRI or BstEII) occurring, in a coronavirus spike gene of that length, just by random chance alone. How about the **joint probability of both of them occurring at once**? 👇 👀

Conclusion 🔓 🔑

Since BstEII and EcoRI are considered the same length (after disregarding BstEII's arbitrary middle character, they're each 6 bases long), the joint probability of them occurring together in the spike is approximately `93% * 93%`:

```
In [4]: 0.931640625 * 0.931640625
Out[4]: 0.8679542541503906
```

➡️ at least **86% of all coronaviruses** are going to have - in their spike protein gene - an EcoRI sequence occurring together with a BstEII sequence. Without the need for **any** genetic engineering 🤷‍♀️

Putting it differently: if SARS-CoV-2 was bioengineered ✂️🧬 the way Yan et al suggested, then 17 in 20 coronaviruses occurring in nature **also were**. Why go through the trouble of bioengineering, when nature has already done the work ❓