Hotspots of silent mutation in BsaI/BsmBI sites in the SARS-CoV-2 genome

Guy Gadboit

Contents

List of Figures

List of Tables

Abstract

A well-known preprint argued (among other things) that there was an anomalously high concentration of silent mutations in BsaI/BsmBI recognition sites in the SARS-CoV-2 genome when compared to RaTG13 and BANAL-20-52, and estimated p-values for the significance of both comparisons. Although the calculated significances (of $p=9\times10^{-8}$ and p=0.004) are overestimates, corrected and controlled calculations still yield some significance ($p=7.66\times10^{-5}$ and p=0.035). It remains unclear whether these observations are better explained by a natural or lab origin hypothesis.

1 Introduction

The lab engineering hypothesis here is that researchers were manipulating the locations of BsaI/BsmBI recognition sites in the SARS-CoV-2 genome in order to make it easier to reconstruct it with reverse genetics. The virus they were manipulating them in would have been SARS-CoV-2 itself, only differing in the handful of locations where they were manipulating these sites. If we had the sequence of that virus it would be obvious whether the sites had been manipulated. But the closest thing we have are SARS-CoV-2's natural close relatives, of which RaTG13 and BANAL-20-52 are the closest.

The approach taken in the preprint then is to see if we can separate the natural mutations that exist between SARS-CoV-2 and those close relatives from any additional artificial ones that may have been introduced. That is why the calculations always involve two viruses: SARS-CoV-2 and either RaTG13 or BANAL-20-52. I have also added a third for comparison, a "Chimeric Ancestor". This is not a real virus but a reconstruction of all the various pieces of other viruses where they are most closely related following the analysis in Temmam et al. 2022).

When trying to understand whether anything we discover in the SARS-CoV-2 genome tells us anything about the origin what we are really interested in is a "Bayes Factor"— what is the ratio of the probabilities of the observations under the natural and lab origin hypotheses?

If we assume that under natural evolution silent mutations happen at random (in some sense) we can use Fisher's Exact Test to estimate the probability that some observed classification of them would arise by chance. If we find what looks like a lot of silent mutations in the BsaI/BsmBI recognition sites, we can ask, how likely is that to have arisen naturally, under the assumption that their positions should be random? If we find this was unlikely we may have discovered something, but not necessarily evidence of lab engineering. For that we *also* need to show that the observations are more probable under a lab hypothesis. There may be other natural hypotheses that explain them better. We may also discover that our assumptions about randomness weren't quite right.

In short there are two different questions here: did we find anything unusual or suprising at all? And, if we did, does it point to lab engineering?

2 Reproducing the calculation

In the text of the preprint, they write:

> *There are 12 silent mutations found in 9 distinct BsaI/BsmBI sites between RaTG13 and SARS-CoV-2, and 882 silent mutations outside of BsaI/BsmBI sites. There are 5 silent mutations found in 7 distinct BsaI/BsmBI sites between BANAL52 and SARS-CoV-2, and 753 silent mutations outside BsaI/BsmBI sites.*

And the calculation in the code is presented as follows:

```
### Fisher Test of Silent Mutations within BsaI/BsmBI sites
## number of nt's within BB sites
BB_sites_RaTG13 <- Pos[Virus %in% c('SARS2','RaTG13'),length(unique(position))]*6
BB_Smuts_RaTG13 <- M[site %in% REsites & Virus=='RaTG13',.N]
SMuts_RaTG13 <- M[Virus=="RaTG13",sum(silent)]-BB_Smuts_RaTG13
RaTG13_genome_nonBB <- nchar(RaTG13)-BB_sites_RaTG13

A_RaTG13 <- matrix(c(BB_Smuts_RaTG13,BB_sites_RaTG13-BB_Smuts_RaTG13,
SMuts_RaTG13,RaTG13_genome_nonBB-SMuts_RaTG13),nrow=2,byrow = T)
fisher.test(A_RaTG13)
# Fisher's Exact Test for Count Data
#
# data:  A_RaTG13
# p-value = 9.087e-08
# alternative hypothesis: true odds ratio is not equal to 1
# 95 percent confidence interval:
# 4.24042 17.25056
# sample estimates:
# odds ratio
# 8.877937
```

From this information we have three out of the four terms we need to reconstruct the contingency table: (12/42) / (882/d).

We assume that by `nchar(RaTG13)` they mean the total number of nucleotides in RaTG13 (MN996532.2), which is 29855. `BB_Smuts_RaTG13` is the total number of silent muts in sites, which is 12, and `BB_sites_RaTG13` is the total number of nucleotides in sites, which is 54. This makes `RaTG13_genome_nonBB` equal to 29855-54, and the final value in the contingency table 29855-54-882=28919.

So our complete contigency table should be (12/42) / (882/28919). This gives me an OR of 9.368 and a p of 5.21×10^{-8}, which does not match their numbers of 8.877937 and p=9×10^{-8}. I suspect the reason for this may be different estimates of the length. Perhaps they're using the length in the alignment rather than the actual genome length?

But I don't think these details matter. In my own calculations, I also get slightly different numbers for the total number of silent mutations (possibly because I am using a different alignment tool) and for the total number of nucleotides as I am only counting those that are in coding sequences.

This gives me ORs and ps for RaTG13 and BANAL-20-52 of:

```
BANAL-20-52: OR=5.085137 p=0.004515 (5/37) / (773/29088)
RaTG13: OR=8.579448 p=1.296e-07 (12/42) / (962/28887)
```

3 The calculated p-values fail basic sanity tests

If we compute some "p-values" using a formula, the question arises: how can we check the result? One way is to try the same calculation on some "controls": datasets on which we would not expect to find anything unusual. If we do find something unusual on those datasets, either we have just made an error in how we are counting things, or perhaps we have found something real and unusual, but perhaps that isn't better explained by our hypothesis (in this case that SARS-CoV-2 shows traces of engineering).

A BsaI recognition site is just the 6 nucleotide pattern GGTCTC or its reverse complement GAGACC (and BsmBI is CGTCTC/-GAGACG).

The two basic controls we will be using here are: calculating odds ratios (ORs) for BsaI/BsmBI sites in *other* pairs of SARS-CoV-2 related viruses, that are not under suspicion of having a lab origin; and calculating odds ratios for mutations in other 6nt patterns (selected at random) between SARS-CoV-2 and RaTG13 (or BANAL-20-52 etc.).

The control patterns we will use are generated by picking two 6nt sequences at random, and their reverse complements. Then we will count silent mutations etc. in the sites defined by those patterns.

The preprint claims a p-value of 9×10^{-8} for the comparison with RaTG13. This is much smaller than the reciprocal of the total number of possible 6 nucleotide sequences (4096) and of pairs of 6 nucleotide sequences (1.68×10^{7}).

So if that p-value were correct, it should be very hard to find *another* pair of 6nt patterns which score a higher odds ratio (using the calculation in the preprint) than the actual BsaI/BsmBI patterns.

But it turns out we can find several, just by trying the calculation on a few thousand random pairs and their reverse complements.

```
$ go run -- *.go -algo original -montecarlo -which-mc=3 -doubles=false
```

Produces the following pairs of 6nt patterns and their odds ratios, which are all higher than the value of 8.579448 for BsaI/BsmBI in RaTG13:

```
CGGCCC/TGCGCC 9.097702
CGCCGG/CGGACT 14.879115
GGGGCA/CGAGTC 9.927909
GGCCGG/CCGAGA 9.927909
CGAGTC/GGCGCA 9.927909
AGGGCG/CCGACC 9.927909
0.001600 are greater than reference OR of 8.579448
```

In other words, the patterns above (which were just selected at random) are all apparently more unusual and anomalous than the BsaI/BsmBI recognition sites, using the same calculation that was supposed to pick those out as having been manipulated.

This gives us a rough estimate that the true significance is somewhere closer to p=0.0016, which is much less significant than $p=1.296\times10^{-7}$. Therefore something must be wrong with the original calculation.

Our other kind of control is to repeat the calculation, using the BsaI/BsmBI recognition sites, in other pairs of genomes, which are not under suspicion of having been manipulated.

```
$ go run -- *.go -algo original -testall -fasta ../fasta/SARS2-relatives.fasta
```

This reveals several high odds ratios for other pairs of viruses and apparently low p-values. Interestingly RaTG13 appears near the top several times (the reasons for this will be clear later). But there are other examples too.

```
RaTG13,BANAL-20-103: 8.058429 p=8.956683643280792e-08
RaTG13,BANAL-20-236: 7.593421 p=3.7636490908469625e-06
BANAL-20-236,RaTG13: 7.585572 p=3.796766324134739e-06
RaTG13,RpYN06: 7.533070 p=9.468263454817954e-10
BANAL-20-236,RaTG13: 7.317192 p=2.1811718530008234e-07
BANAL-20-103,RaTG13: 7.315253 p=6.562885160341414e-07
RaTG13,BANAL-20-236: 7.282421 p=2.2971100021053767e-07
YN2021,PrC31: 7.070457 p=5.197625152971807e-06
PrC31,YN2021: 7.040989 p=5.384177287044883e-06
Rp22DB159,Rp22DB167: 6.895097 p=0.00042395898530785684
```

4 The distribution of ORs under the original calculation is not centered on 1

When we calculate an odds ratio of, say, 8, we are saying that what we observe appears to be 8 times less likely than what we expect under some null hypothesis. This requires that our calculation should give us an odds ratio of about 1.0 on data on which we think the null hypothesis— which in this case is that the mutations in the BsaI/BsmBI sites arose naturally— is the right one.

It's therefore instructive to look at the distribution of odds ratios for our two kinds of control. Let's look at BsaI/BsmBI sites in other (non-engineered) viruses first:

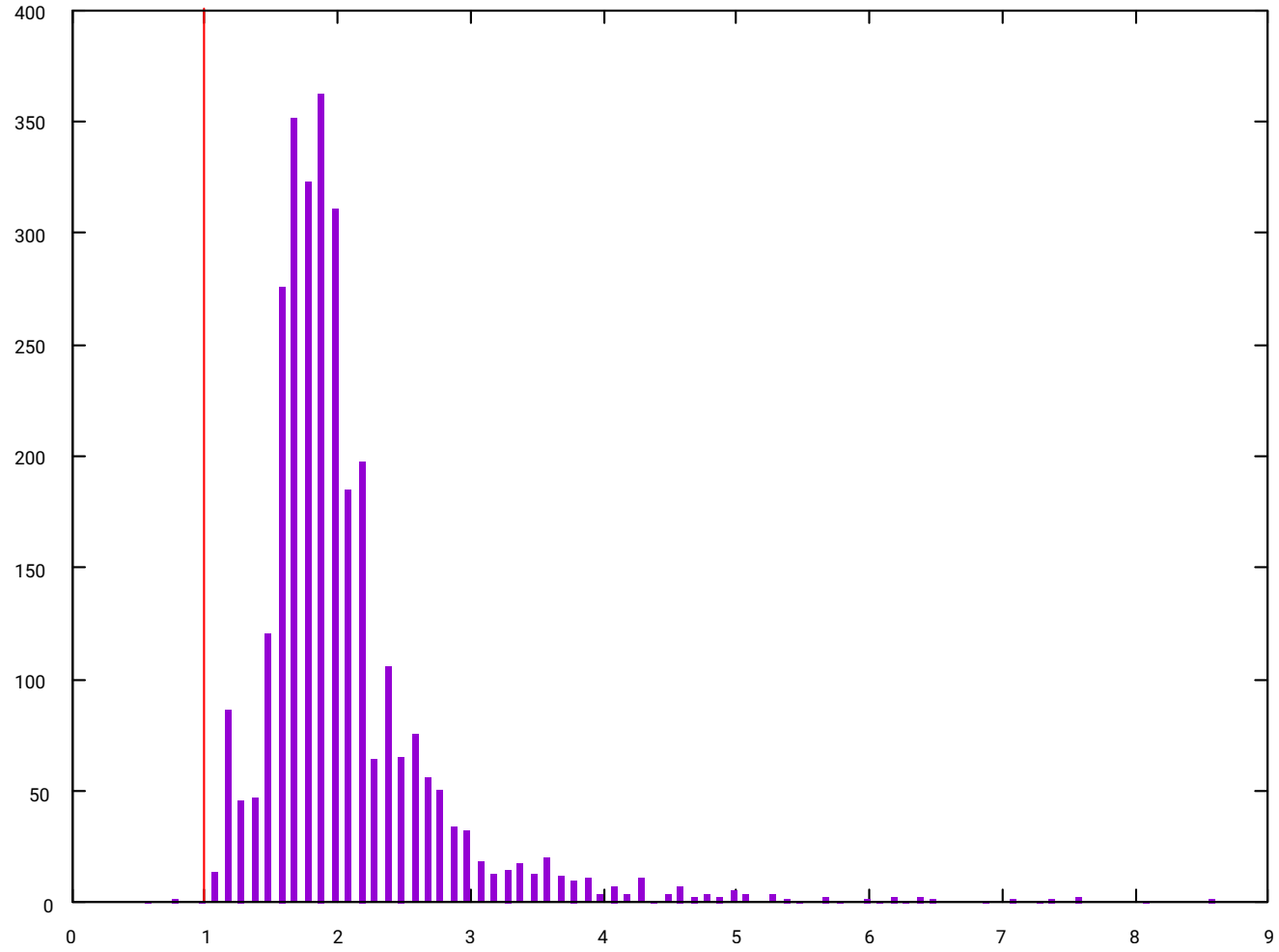

Figure 1: ORs (original calculation) for all combinations of pairs of SARS2r-CoV genomes

This was computed with:

```
$ go run -- *.go -algo original -doubles=false -testall -fasta ../fasta/SARS2- ↩
    relatives.fasta
$ gnuplot --persist distribution.gpi
```

And now for our other control, different random 6nt patterns, but still comparing SARS-CoV-2 and RaTG13. As we can see, the distribution here is not centered on 1 either:

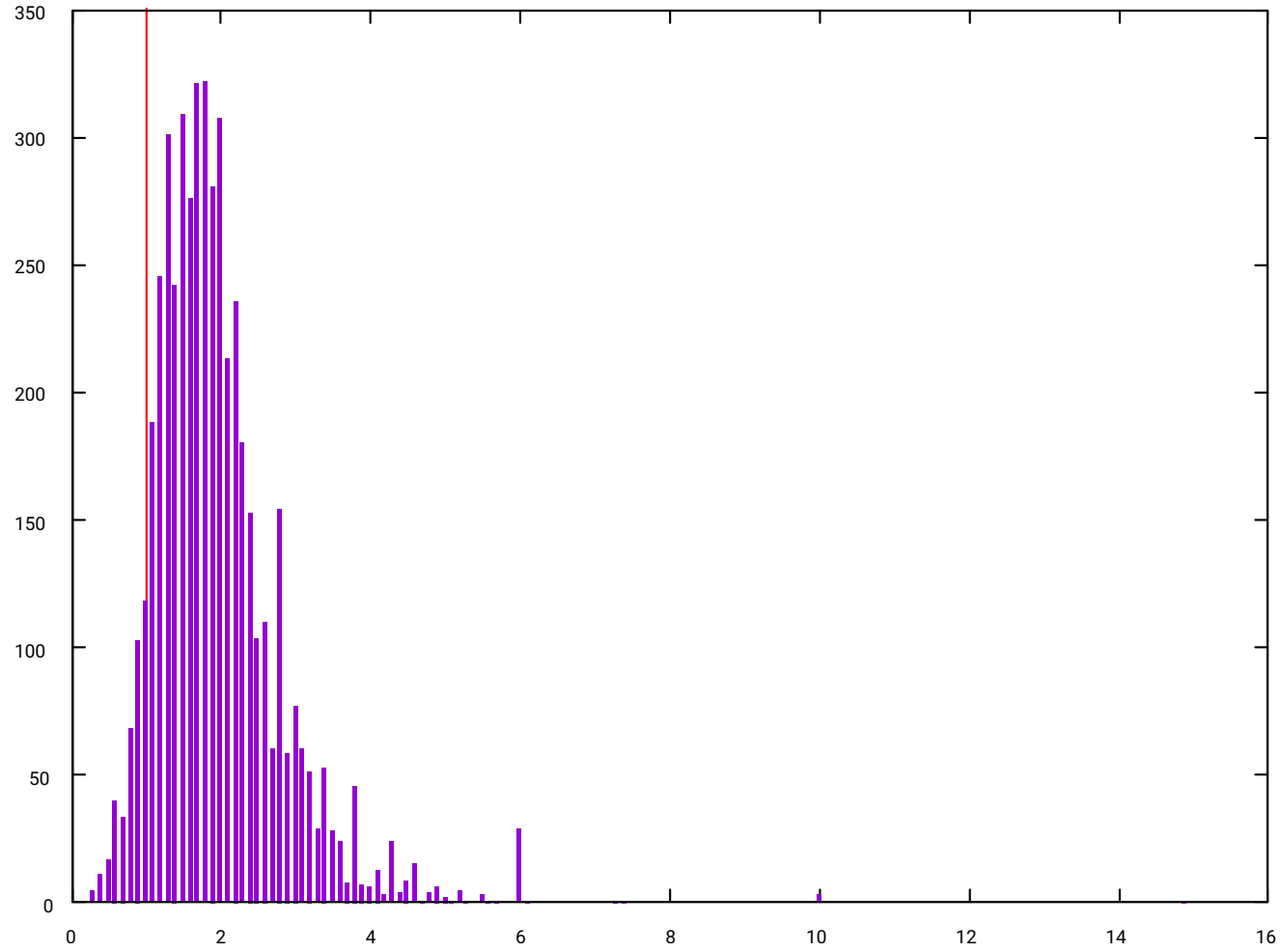

Figure 2: ORs (original calculation) for 5000 pairs of random 6nt patterns in SARS-CoV-2 vs RaTG13

This was computed with:

```
$ go run -- *.go -algo original -doubles=false -montecarlo -which-mc=3
$ gnuplot --persist distribution.gpi
```

In other words, using the original calculation, other random pairs of viruses and other random sites all show about twice the odds we are expecting. Obviously this doesn't imply either that all those viruses have been manipulated (and not just in the BsaI/BsmBI sites but everywhere!) but that the calculation is wrong. We will see how in the next section.

5 The way mutations are counted in sites that appear in both genomes inflates the OR

In the preprint, they define "inside" a site as anywhere a site appears *either* in SARS-CoV-2 *or* in RaTG13. But what about the places where a site is inside *both* of them? Let's look at the numbers. There are 5 sites in SARS-CoV-2 (at 2193, 9751, 17329, 17972 and 24102) and 6 sites in RaTG13 (at 10444, 11648, 17329, 17972, 22922, 24509). Note that two of those sites are in *both*, viz the ones at 17329 and 17972.

In this table the "Alignment" entries show the nucleotide sequences in SARS-CoV-2 and RaTG13, a '*' wherever they are the same, and the amino acid coded for at that location ('R'=Arginine, 'D'=Aspartic Acid etc.).

Position in alignment	Enzyme	Genome	Alignment	Number of silent mutations
2193	BsmBI	SARS-CoV-2	AGAGACGGT AGAGATGGT ***** *** RRRDDDGGG RRRDDDGGG	1
9751	BsmBI	SARS-CoV-2	AAGAGACGT AAAAGGCGT ** ** *** KKKRRRRRR KKKRRRRRR	2
10444	BsaI	RaTG-13	ATGAGGCCC ATGAGACCT ***** ** MMMRRRPPP MMMRRRPPP	1
11648	BsaI	RaTG-13	GGCCTC GGTCTC ** *** GGGLLL GGGLLL	1
17329	BsmBI	Both	GAGACG GAGACG ****** EEETTT EEETTT	0
17972	BsaI	Both	AGAGACCTT AGAGACCTT ********* RRRDDDLLL RRRDDDLLL	0
22922	BsmBI	RaTG-13	AGATTG CGTCTC * * RRRLLL	4
24102	BsaI	SARS-CoV-2	AGAGACCTC AGGGATCTT ** ** ** RRRDDDLLL RRRDDDLLL	1
24509	BsmBI	RaTG-13	AGATTG CGTCTC * * RRRLLL RRRLLL	2

Table 1: BsaI/BsmBI sites compared in SARS-CoV-2 and RaTG13

The odds they are calculating in the preprint are based on the number of nucleotides in these regions that are silently mutated, over the number that aren't silently mutated. So the first question is, how many silent mutations are there in those regions?

Those first 5 sites contain, respectively, 1, 2, 0, 0 and 2 mutations. Those add up to 5. 5 sites of 6nts each is 30nts, so we have 5 silent muts, 30 total, and so 25 locations that aren't mutated (30-5). So the odds would be 5/25 if we were only looking at sites in SARS-CoV-2. Now if we looked instead at the 6 places in RaTG13 where sites occur, and how many mutations are inside those sites, we find, respectively, 1, 1, 0, 0, 4, 1. Those sum to 7. So the odds here are 7/29. Clearly there are 12 silent mutations when we are interested in sites in *either* genome so the numerator of our odds will be 12. But what about the denominator?

There are 9 sites altogether (5 in SARS-CoV-2 plus 6 in RaTG13 minus the two that are in both). On that basis the total number of nucleotides is $9\times6=54$, and so our odds would be 12/42. This is the way they calculate it in the preprint:

```
BB_sites_RaTG13 <- Pos[Virus %in% c('SARS2','RaTG13'),length(unique(position))]*6
```

This inflates the odds ratio because it is only counting the number of nucleotides that aren't silently mutated in those two regions where both genomes share a site once, not once *for each genome*.

If we look at the odds of silent mutations in places where the sites appear in each genome individually, but without trying to combine them, we find 5/25 and 7/29 for where sites appear in SARS-CoV-2 or RaTG13 respectively. We shouldn't expect those odds to magically increase when we look at both together, just because we are looking at both together. This is why combining those odds should give us (5+7)/(25+29)=12/54 (not 12/42).

In other words, if a site appears in both genomes, it needs to contribute to the denominator *twice*. There is a more discussion of this subtlety in this X thread.

6 Another look at the distributions of ORs with correct counting of double matches

So how do things look when we use the correct denominators, taking into account sites that appear in both genomes?

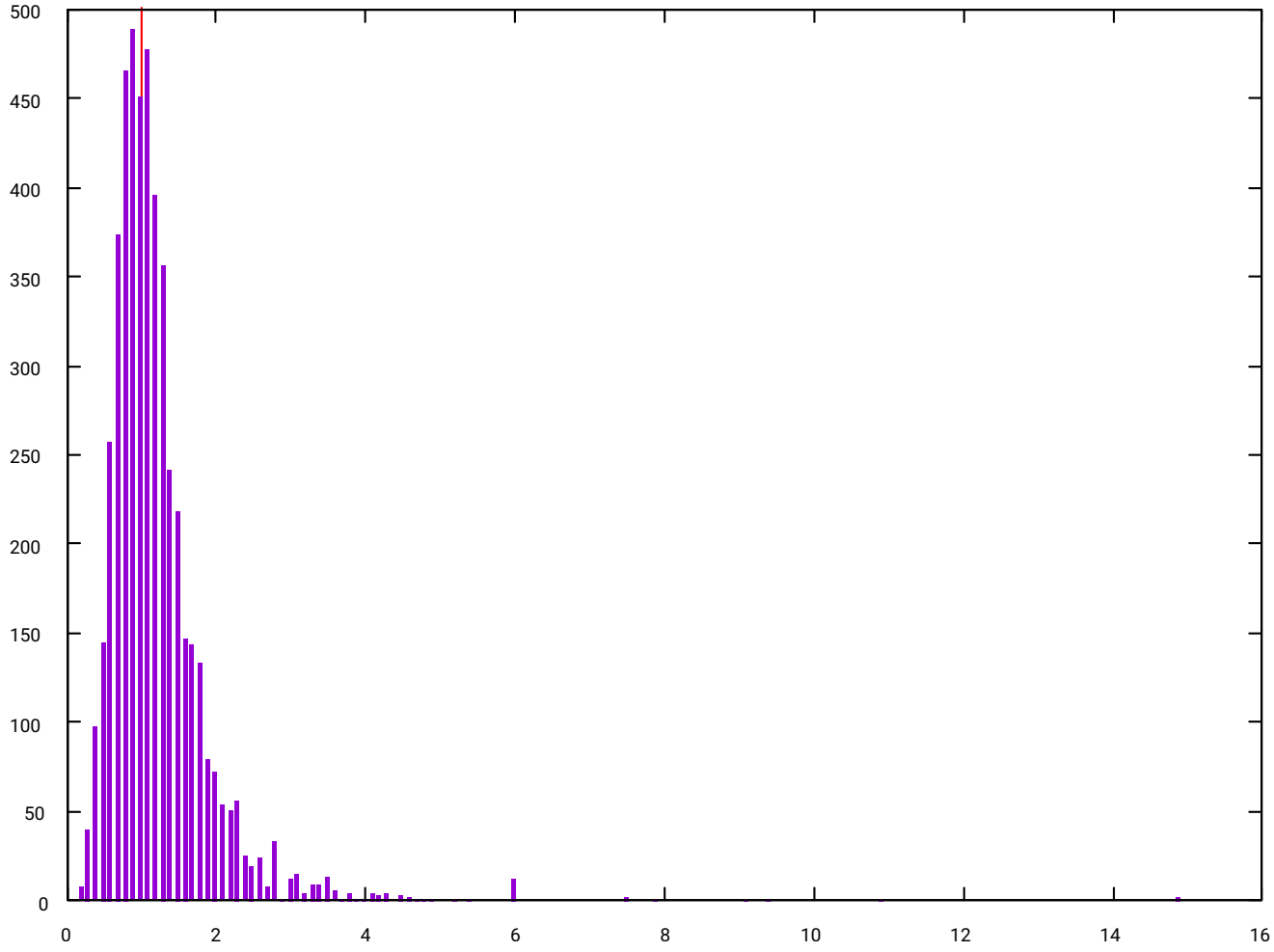

Figure 3: ORs (corrected double matches) for 5000 pairs of random 6nt patterns in SARS-CoV-2 vs RaTG13

This was computed with:

```
$ go run -- *.go -algo original -montecarlo -which-mc=3
$ gnuplot --persist distribution.gpi
```

As we can see (compare with Figure 2) the distribution of ORs for other 6nt sites is now correctly centered on 1.

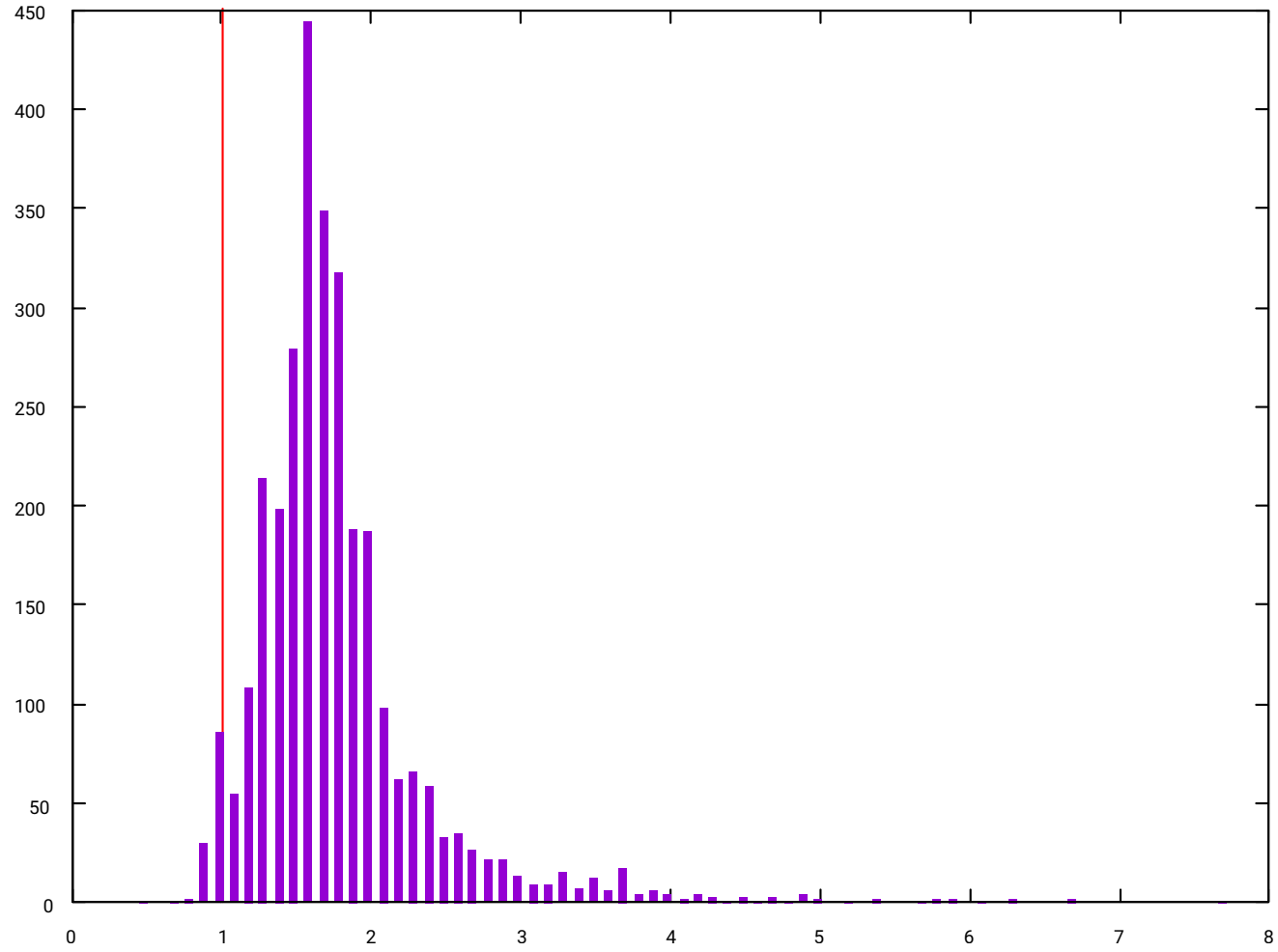

Figure 4: ORs (corrected double matches) for all combinations of pairs of SARS2r-CoV genomes

This was computed with:

```
$ go run -- *.go -algo original -testall -fasta ../fasta/SARS2- ↩
    relatives.fasta
$ gnuplot --persist distribution.gpi
```

But, interestingly, the distribution of the ORs when looking at mutations in BsaI/BsmBI sites in other pairs of viruses is still centered some way above 1.0.

We now have a calculation we can basically trust— because it does yield an OR of about 1.0 on one of our controls (we don't expect patterns other than BsaI/BsmBI to have been manipulated in SARS-CoV-2).

But on the other control (we don't expect BsaI/BsmBI sites to have been manipulated in other viruses that aren't SARS-CoV-2) it still seems to be "reading too high".

What this seems to be telling us is that there is indeed something "special" about BsaI/BsmBI sites, but which we also see in other viruses that are under no suspicion of having been engineered. So whatever that special thing it is, it is not attributable to engineering. But if we can understand what it is attributable to, maybe we can do a corrected calculation to see if there is any remaining significance for the engineering hypothesis.

After counting the sites that appear in both viruses correctly, we end up with the following ORs and ps:

Comparison	OR	p-value
SARS-CoV-2 vs RaTG13	6.672904	1.294×10^{-6}
SARS-CoV-2 vs BANAL-20-52	3.839797	0.013

7 What's special about BsaI/BsmBI?

If you look at Table 1 above, one particular site stands out. This is the site at 22922. In SARS-CoV-2 we have AGATTG, and in RaTG13, CGTCTC (which is the BsmBI site). Remarkably, in this section of 6 nucleotides, there are *4* silent mutations.

That one site contributes a great deal to the high odds ratio (and low p-value) for the comparison. If we change that last site in SARS-CoV-2 to CGTCTT (which differs from RaTG13 silently, as before, but this time with only 1 mutation rather than 4) and repeat the calculation, we get an OR of 4.74 and p=0.00027 (compared with 6.67 and 1.294×10^{-6})

What best explains this unusual site, with 4 silent mutations out of 6? On the face of it, it's unlikely to be engineering. We would expect an engineer to make the minimum number of changes possible to add or remove a BsaI/BsmBI site (so as to minimize the risk of disrupting secondary RNA structure etc.). Why would she remove a site with 4 silent mutations rather than just one?

What about a natural explanation? The BsmBI site happens to be codon-aligned and codes for the amino acid sequence Arginine-Leucine at this point in RaTG13. That particular combination happens to have the largest number of possible synonyms of *any* pair of amino acids in biology.

There are 36 synonyms for Arginine-Leucine, many of which have 3 or 4 mutations from CGTCTC:

Nucleotides	Mutations from CGTCTC
CGTCTC	0
AGACTC	2
CGGCTC	1
CGACTC	1
CGCCTC	1
AGGCTC	2
AGACTG	3
CGGCTG	2
CGACTG	2
CGCCTG	2
AGGCTG	3
CGTCTG	1
AGATTA	4
CGGTTA	3
CGATTA	3
CGCTTA	3
AGGTTA	4
CGTTTA	2
AGATTG	4
CGGTTG	3
CGATTG	3
CGCTTG	3
AGGTTG	4
CGTTTG	2
AGACTA	3
CGGCTA	2
CGACTA	2
CGCCTA	2
AGGCTA	3
CGTCTA	1
AGACTT	3
CGGCTT	2
CGACTT	2

Nucleotides	Mutations from CGTCTC
CGCCTT	2
AGGCTT	3
CGTCTT	1

Under an assumption of strong purifying selection, it is perhaps not too unexpected to find 4 silent mutations in 6 nucleotides in places where 4 silent mutations in 6 nucleotides are *possible*.

In other words, the specialness of BsmBI in this test might be arising not from *where* it is, but from *what* it is. It so happens that in one direction (and when in frame) the BsmBI recognition sequences codes for a protein sequence that has a high number of synonyms with lots of mutations. This might be why we find an unusually high number of silent mutations in BsmBI sites (including in other pairs of viruses, not suspected of being engineered— see Figure 4).

There is another way we can test whether what matters about BsaI/BsmBI sites is their content rather than their location. We can repeat the calculation we used for Figure 4 but before doing each comparison, redistribute the silent mutations.

The way this works is as follows: Before doing the calculation of the odds ratio between two genomes, we first count the number of silent mutations that exist between them. Then we set the second genome equal to the first. Then we find all the positions where a silent mutation is possible. We pick positions from that set at random, and choose, also at random, the actual nucleotide to use, provided that it gives us a silent mutation, and apply that to the second genome. Then we do the odds ratio calculation as before.

If, after randomizing the positions of the mutations, we still find that BsaI/BsmBI sites score generally higher than 1 on our test, it confirms the hypothesis that this is because more mutations are possible in those sites, because of the content they happen to have. It can't be anything to do with position, because we randomized the positions.

```
$ go run -- *.go -algo original -testall -fasta ../fasta/SARS2-relatives.fasta - ↩
    redistrib
```

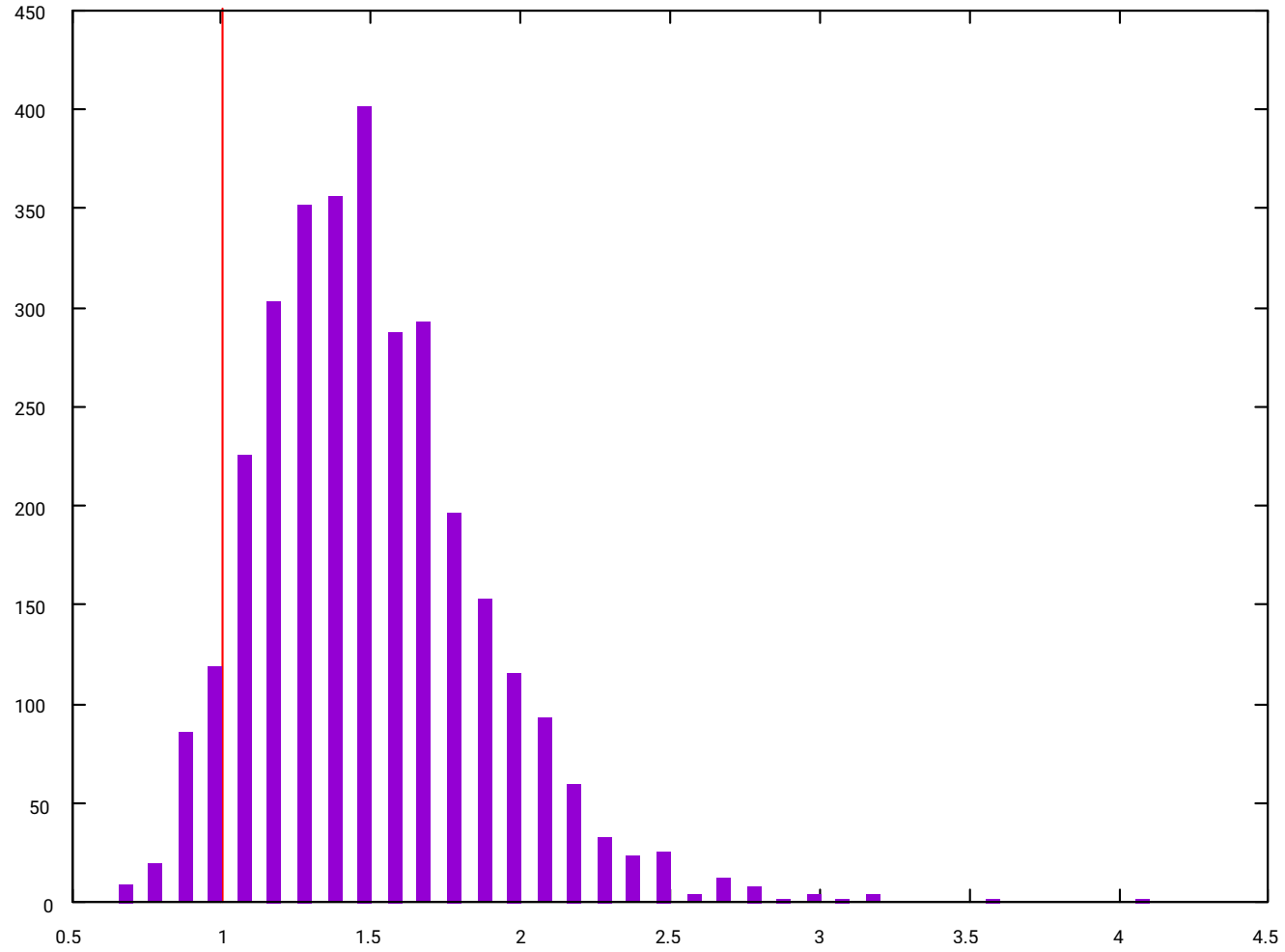

Figure 5: ORs (corrected double matches) for all combinations of pairs of SARS2r-CoV genomes, silent mutations redistributed randomly

The result is a distribution still centered above 1. In other words: you tend to find more silent mutations in BsaI/BsmBI sites even when their positions are randomized, because (at least partly) more silent mutations are possible in those sites.

8 An improved calculation shows that there is a real anomaly

The original calculation in the preprint was of the odds ratio of silently mutated locations, and locations that weren't silently mutated, inside and outside the BsaI/BsmBI sites. There is an implicit assumption there that we would expect those ratios to be about the same. But, as we have seen, that assumption turns out to be false.

Can we do better? It's worth reminding ourselves what Fisher's Exact Test actually exactly tests. If I put 1000 balls in a hat, of which 990 are red, and 10 are blue, and you pick out 5 balls at random, what is the probability that 3 or more of the 5 will be blue? That probability is the p-value that Fisher's Exact Test gives us (there are slightly different calculations for exactly 3, 3 or more, or 3 or fewer).

So what we really want to do here is put *all possible* silent mutations into a "hat", with the ones that are inside BsaI/BsmBI sites "coloured" red. Then we can use the test to tell us the probability that, under the assumption that nature just picks possible mutations at random, we would find the number of silent mutations in the sites that we do find. If that number is still low then we can wonder about alternative hypotheses to explain it (including engineering).

In reality, we know that nature does not exactly pick mutations at random. Some nucleotides are more common than others in particular genomes. Sometimes particular codons are preferred over others. Some regions are subject to stronger purifying selection than others. Nevertheless, keeping these caveats in mind, we can try a modified test and see how it performs on our controls. This will at least give us some kind of estimate of whether there is really anything unusual about the locations of the sites in the SARS-CoV-2 genome.

To recap briefly: the odds ratio calculated in the original preprint was (number of silent mutations in sites / other locations in sites) / (number of silent mutations outside sites / other locations outside sites). This takes no account of how many silent mutations are *possible* in each location, and resulted in inflated values, particularly for one of the BsmBI sites which happens to code for a sequence that has a large number of possible synonyms, as well as defining a site. This coincidence has nothing to do with any lab engineering hypothesis.

Our new calculation will take the form: (number of silent mutations in sites / number of silent mutations outside sites) / (number of possible silent mutations in sites / number of possible silent mutations outside sites).

How do we decide how many silent mutations are possible at a particular location? There are various ways you could do this. You could say a silent mutation is possible if you can change just that nucleotide on its own to one of the other four without changing the protein. You could call that "one possible mutation", or you could count each alternative nucleotide that doesn't change the protein there as a separate possible mutation, and add them all up

These are reasonable models for random point mutations over a relatively short period of time. But they don't capture the situation where multiple adjacent nucleotides can change together (or one after the other) and give you the same protein, which perhaps models better a situation in which you have a lot of mutation time but strong purifying selection. In particular they won't be able to "find" the situation we have with that BsmBI site at 22922, where there are 4 mutations out of 6 and no change to the protein.

A more "aggressive" approach, and the calculation we are using here, is to count all the possible values that a nucleotide could have, while preserving the protein, assuming that surrounding nucleotides change as necessary in order to allow that. Each of those alternative values is a "ball" in our "hat" for Fisher's Exact Test, and the ones that happen to be inside BsaI/BsmBI sites are "coloured red". Now that we know how many balls of each colour are in the hat, we can compare with the actual numbers of silent mutations inside and outside sites, between any two genomes, and calculate an odds ratio and p-value. If a nucleotide can have the same value as part of more than one synonymous codon, we count all of them. For example the third nucleotide in a Leucine codon can be 'A' either because the codon is "TTA" or because it's "CTA". We therefore count a mutation to A in this position as *two* possible mutations, because what we are trying to capture here is some approximation to the "number of ways" you can end up with an A there.

When applied to our first control— random pairs of 6nt patterns and their reverse complements— we see a distribution that looks a bit skewed, but is still centered on 1:

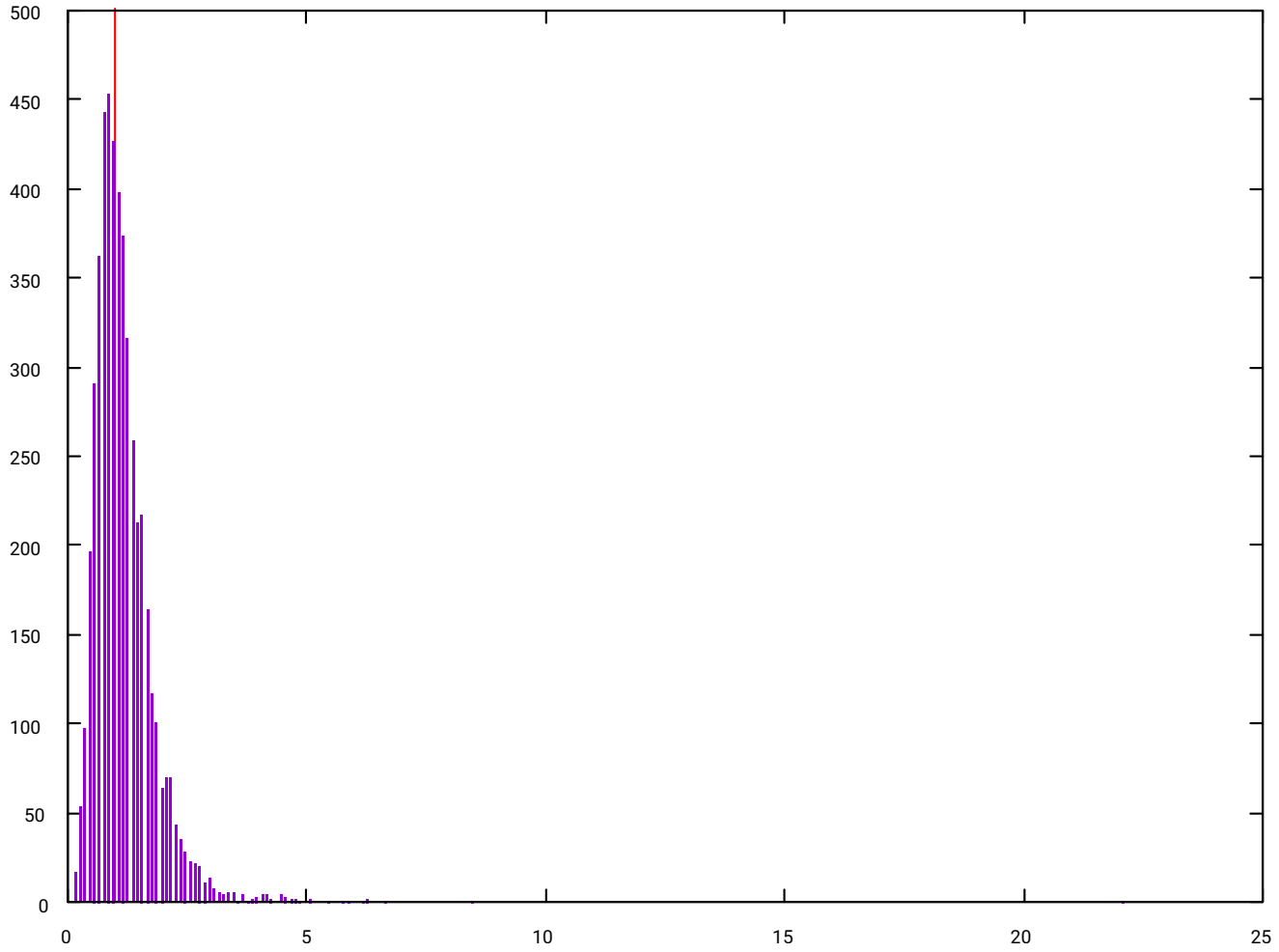

Figure 6: ORs (improved calculation) for 5000 pairs of random 6nt patterns in SARS-CoV-2 vs RaTG13

This was created with:

```
$ go run -- *.go -montecarlo -which-mc=3
$ gnuplot --persist distribution.gpi
```

The first control was always centered on 1. But this time, when applied to our other control— all the pairs of SARS2-related viruses in a large set— this new calculation also gives us a distribution centered on 1. Notably, the comparisons between SARS-CoV-2 and BANAL-20-52, RaTG13, or the "Chimeric Ancestor" are *still* outliers.

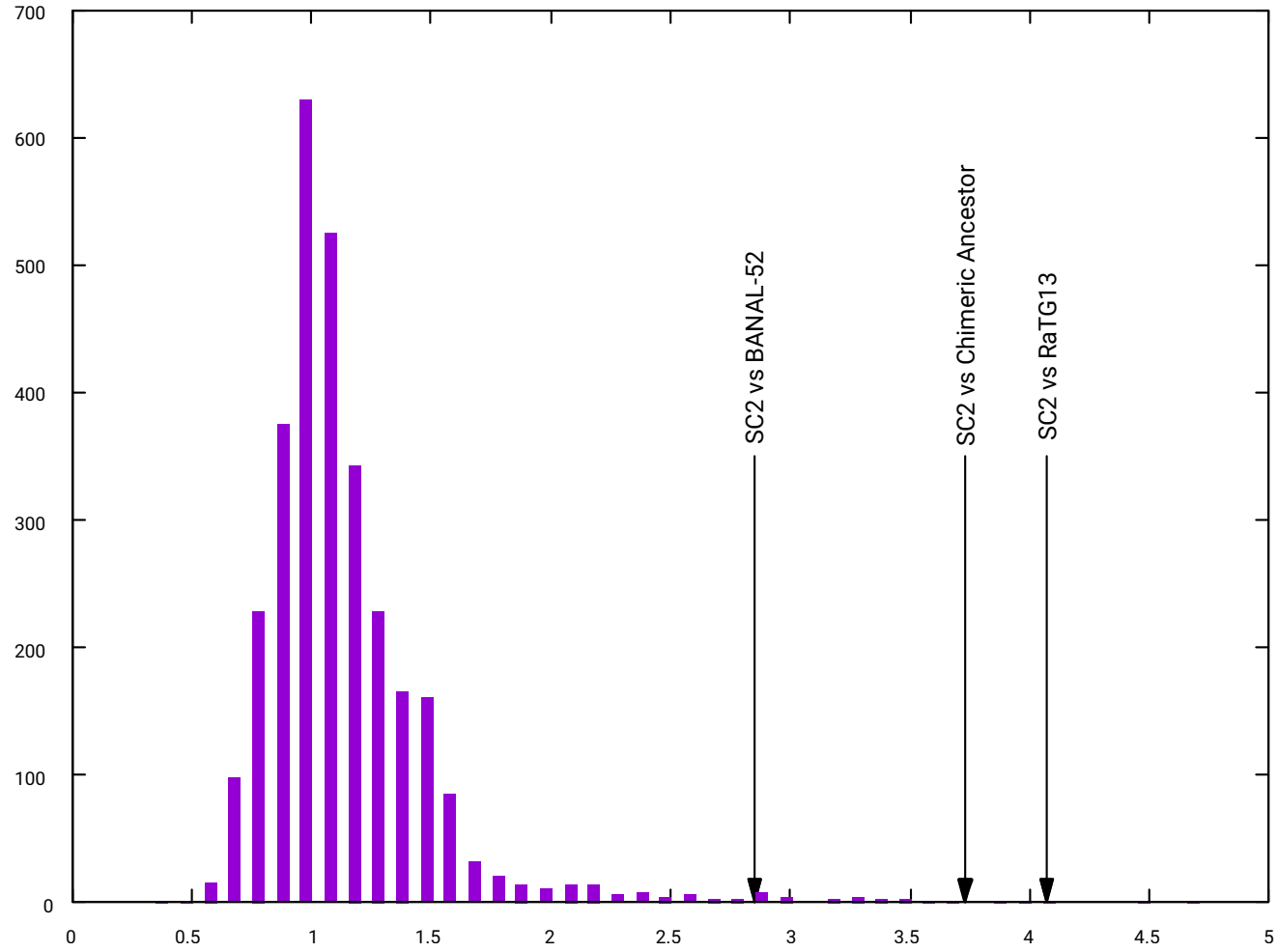

Figure 7: ORs (improved calculation) for all combinations of pairs of SARS2r-CoV genomes

This was created with:

```
$ go run -- *.go -testall -fasta ../fasta/SARS2-relatives.fasta
$ gnuplot --persist distribution.gpi
```

And here are the ORs and p-values for comparisons between SARS-CoV-2 and closely related viruses, using Fisher's Exact Test on the contingency matrix constructed with possible mutations as described above:

Comparison	**OR**	**p-value**
SARS-CoV-2 vs RaTG13	4.066	7.66×10^{-5}
SARS-CoV-2 vs BANAL-20-52	2.850	0.0349
SARS-CoV-2 vs Chimeric Ancestor	3.730	0.00055

Note that these p-values should only be considered approximations. Fisher's Exact Test is assuming a hypergeometric distribution (which is what you will get with actual coloured balls in hats). The actual distribution will be something different, because of, among other things, the caveats mentioned above. Nevertheless we can see from Figure 7 that there is no doubt that the number of silent mutations between SARS-CoV-2 and its closest known relatives in BsaI/BsmBI sites *is* anomalous. That is just a graph of actual values computed on real viruses.

It doesn't automatically follow that what we observe here is more probable under an engineering hypothesis. But it certainly invites suspicion given that:

- The DEFUSE draft proposal mentioned 6 segments using BsmBI (but it doesn't mention BsaI).
- SARS-CoV-2 does divide into 6 segments, none of which are too long, if you split it on BsmBI and BsaI sites. This type of pattern would be expected to occur naturally about 0.6%[1] of the time.
- As demonstrated above, there is some anomaly around the silent mutations in BsaI/BsmBI sites in SARS-CoV-2 compared to its closest known relatives, when compared to controls.

However, other features of the engineering hypothesis still require explanation. If you were using BsaI/BsmBI, why would you leave the sites in the final assembled genome anyway? Why use BsaI as well as BsmBI?

Furthermore, as we shall see in the next section, a different feature of the BsaI/BsmBI sites in SARS-CoV-2 and its relatives seems more consistent with a natural origin, and also raises questions about whether this anomalous odds ratio has anything to do with engineering.

9 What do BsaI/BsmBI sites in related viruses but not in SARS-CoV-2 tell us?

This table shows each location where one or more of the related viruses in the Temmam et al. 2022 paper have a BsaI/BsmBI site but SARS-CoV-2 does not, but where the protein is unchanged.

The list of viruses, and the basic idea here, comes from this project, which I recommend investigating. But we are specifically concentrating on sites *not* in SARS-CoV-2 that were apparently *removed* and on *how many ways there are* of doing that.

The second column shows what SARS-CoV-2 does have in each location instead of the BsaI/BsmBI site.

The third column shows all the different nucleotide values of each site that would preserve the protein, followed by how often each occurs in the set of related viruses. **Bold type** indicates alleles that actually are BsaI/BsmBI sites.

Looking at the first row as an example, one of the viruses in the set (it's RpYN06 in this case) has the BsaI site GGTCTC at position 4665 in the alignment. In this location, SARS-CoV-2 has GATCTC, and so do 10 other viruses in the set. As we can see in the third column, other silent alternatives of nucleotides are only rarely seen.

The striking thing about this data is how in nearly all the places where SARS-CoV-2 has a site removed, the *way* it's removed seems to copy the exact nucleotides that are present in a significant majority of relatives.

Under an engineering hypothesis, our engineer working in the lab would need to move some sites around to achieve the desired restriction map. This would involve adding sites in some locations, and removing them in others.

If she were adding a site, she would probably look for a place close enough to where she wanted that could be achieved with the minimum number of mutations. There are lots of places you can put sites with just one silent mutation, and the exact spacing for making a reverse genetics system is not that important. This implies that it would be very hard to see whether this had been done. The only trace would be a handful of single silent mutations (and these are dotted around the place anyway).

But if she were *removing* a site, there are many more choices. I would expect her just to put an arbitrary single nucleotide change into the site, probably in the third position of a codon, to eliminate it without changing the protein. So she would have to be "lucky" to end up with an allele that happened to be shared by a significant number of natural relatives.

This table gives us an idea of how lucky. In the first case, there are four ways she could have removed the site. She picked the one that happened to be shared by the majority of relatives-- about a 1/4 chance. It's about 1/4 for all the rows in the table (except the location at 10444).

One question is how many sites did she have to remove? If it was four, then there is a probability of only 0.004 that she would get this "lucky". But if it was only 1, we have an insignificant result of of 0.25.

Looking at the table one location stands out— the one at 10444. Here the site is removed in SARS-CoV-2 with an allele only shared with one other relative (RpYN06). So maybe only that one was engineered? The site has been removed with a single nucleotide change, which is the sort of economical way we might expect an engineer to do it. Maybe it's just luck that it matches

[1] Based on my own simulation, the code for which is in the `resite_sim` directory.

Position	Allele in SARS-CoV-2	Shared with	Alternatives and their frequencies
4665	GATCTC	10	GATCCC (2), **GGTCTC (1),** GATCAC (1), GATCGC (0)
4772	GAAACC	14	GAAACG (0), **GAGACC (0),** GAAACA (0), GAAACT (0)
6403	AGTCTC	10	AGTTTC (0), AGTGTC (0), AGTATC (0), TGTCTC (1), **CGTCTC (2), GGTCTC (0)**
7699	AAGACC	11	AAGGCC (0), **GAGACC (4),** ACGACC (0)
10444	GAGGCC	1	GCGGCC (0), **GAGACC (14),**
11064	GGTCTT	14	GGTCCT (0), GGTCAT (0), GGTCGT (0), **GGTCTC (1)**
11648	GGCCTC	6	GGCCTT (0), GGCCTG (0), GGACTC (0), **GGTCTC (7),** GGGCTC (0), GGCCTA (1)
12423	GAGATG	12	**GAGACG (1),** GGGATG (2)
15644	GAGATG	13	**GAGACG (1),** GGGATG (1)
15752	CATCTC	10	CATCCC (0), CCTCTC (0), **CGTCTC (2),** CTTCTC (0), CATCAC (2), CATCGC (0)
15844	GAGACT	14	**GAGACG (0), GAGACC (1),** GAGACA (0), GAAACT (0)
16299	TAGACC	11	TAGGCC (0), CAGACC (2), **GAGACC (1),** AAGACC (1), TCGACC (0)
16663	GAGACA	10	**GAGACG (0), GAGACC (3),** GAAACA (2), GAGACT (0)
18467	GAGATC	9	**GAGACC (3),** GTGATC (1), GCGATC (0), GGGATC (0)
20967	TGTCTC	10	TGTTTC (3), TGTGTC (0), TGTATC (0), **CGTCTC (2)**
24509	CGTCTT	7	CGCCTT (2), CGGCTT (0), CGACTT (0), CGTCTG (0), **CGTCTC (2),** CGTCTA (3)

Table 2: BsaI/BsmBI sites removed in SARS-CoV-2 compared to close relatives

RpYN06? This is possible, but note that in the Temmam et al. paper this whole region of the SARS-CoV-2 genome was found to be closest to RpYN06 out of all the relatives examined (suggesting it may have gained the allele at 10444 from a recent recombination event).

It is quite possible that in order to achieve the desired restriction map you would only need to remove one site. But we can't have our cake and eat it. If you only removed one site, with a single silent mutation, that would be too small a change to make much difference to the anomalous odds ratio reported above. We can test this easily. If I remove that site and repeat the "improved" odds ratio calculation I get:

Comparison	**OR**	**p-value**
SARS-CoV-2 vs RaTG13	3.463	0.00055
SARS-CoV-2 vs BANAL-20-52	2.067	0.1344
SARS-CoV-2 vs Chimeric Ancestor	3.830	0.00024

The ORs have all reduced, but there are still significant anomalies compared to RaTG13 and the Chimeric Ancestor.

The fact that each added site can also be found in a relative is not very significant, because when you add a site you have less choice about how to add it. You would expect to find any arbitary single nucleotide variation somewhere else in a collection of related viruses of about 90-95% sequence similarity. Only removed sites have the potential to tell us something. But they all look "natural". Even if we suppose, for the sake of argument, that the site at 10444 was engineered (and it's a coincidence that it's shared by RpYN06), one site is not enough to explain the anomalous odds ratio.

10 How easy is it to detect simulated engineered genomes?

To summarize the problem so far: we have an anomalous statistic concerning mutations in BsaI/BsmBI sites in SARS-CoV-2 compared to its close relatives. The mutations in those sites score significantly higher on our calculation than in other pairs of related viruses. But when we look at the sites themselves it doesn't look like any of them have been removed compared to related viruses (except perhaps one coincidence).

The question we have is: is the anomalous statistic caused by engineering, or some unknown natural factor? We already found one coincidental natural factor (BsmBI sites naturally have a lot of synonyms) that affected the original calculation. There may be others. Nature is full of surprises.

The fact that the anomalous statistic apparently has high significance might actually argue *against* engineering for the following reason: would we *expect* engineering to result in something detectable with such high significance? If I did want to tweak a genome to give me 6 segments shorter than 8000 nucleotides between BsaI/BsmBI sites I would only need to introduce a small number of silent mutations, sprinkled among the noise of the 700 or so others that exist between the genomes we are comparing anyway. Would those few extra ones show up on our test? And with how strong a signal? An analogy is: suppose I use a sensitive microphone to try to detect a water leak underground. But instead of finding the small hissing noise I'm looking for, it finds a really loud one. This would make me wonder if the signal was caused by something else like a car driving past.

We can approach this question with a simulation. On the computer we can make thousands of simulated genomes with the right numbers of "natural" silent mutations between them, selected from a representative nucleotide distribution. Then in about half of these simulated genomes we introduce a few additional silent mutations to add and remove BsaI/BsmBI sites. We compute the odds ratio as discussed (using our final "improved" calculation— that centres on 1 in both controls) and then we can see how much those odds ratios differ between the "tampered" and "untampered" genomes on average.

Simulations like this are only an approximation. But they are based on similar assumptions to those behind the idea that the odds ratio we are calculating could detect tampering with BsaI/BsmBI sites. We are asking: does the test work, even under its own assumptions?

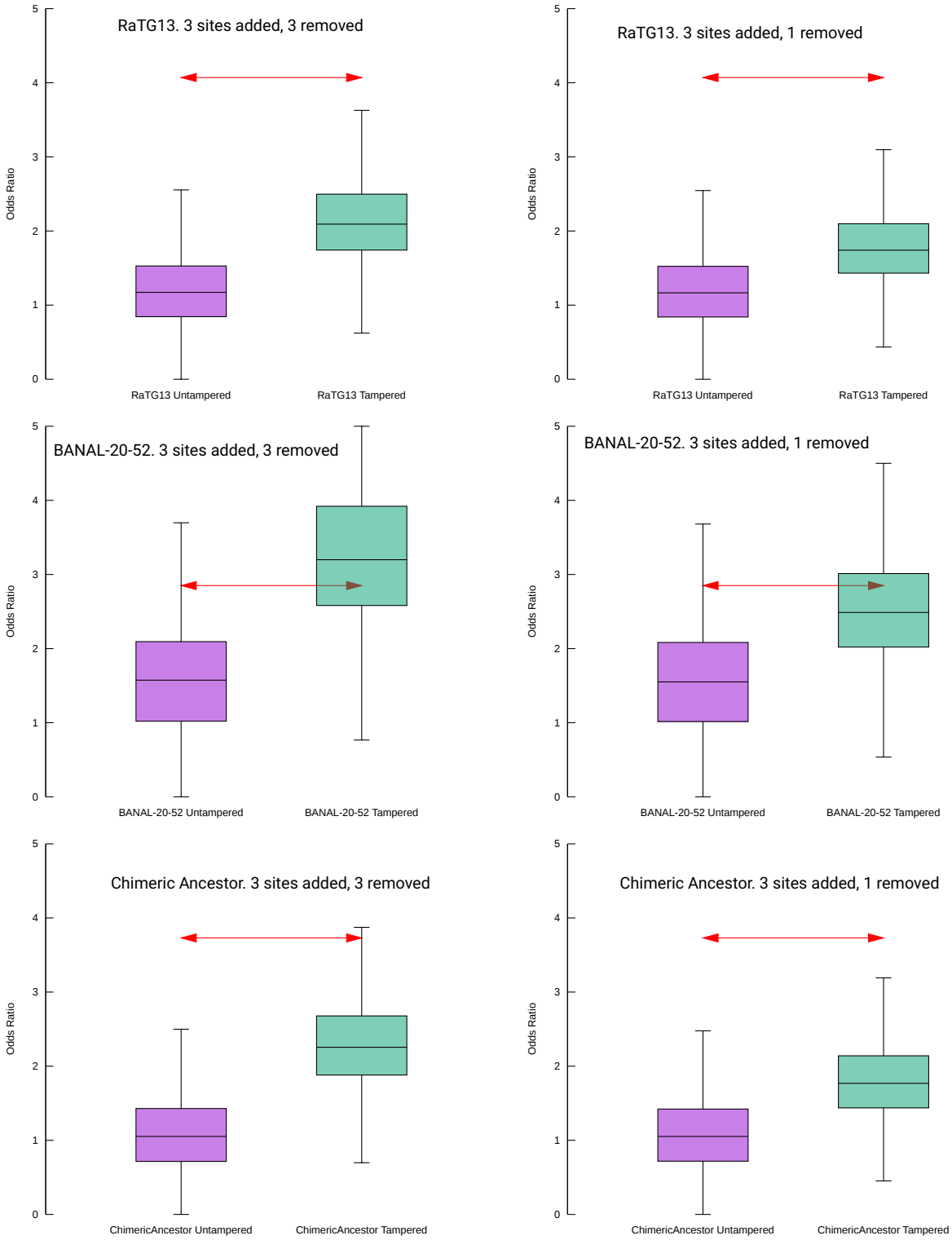

Figure 8: Simulation boxplots

These are boxplots showing the distribution of ORs computed for mutants derived from each of the specified genomes. The pink boxes show the ORs for mutants that had not had extra silent mutations added in BsaI/BsmBI sites, and the green ones those that

had. The red arrows show the actual values of the OR for SARS-CoV-2 compared with each virus.

The boxes are drawn around the regions between the first and third quartiles, and the whiskers extend as far as 1.5 times the interquartile range in each direction. There is good diagram of how this works here.

As we can see, even with only one site removed, the test is able to distinguish engineered genomes from those that are not engineered, but only with modest accuracy, especially in the case where only one site was removed. In the right-hand plots the interquartile ranges overlap.

This seems to fit with the intuition that a handful of extra silent mutations should not be possible to detect with such high confidences as p~=10^{-5}. In the comparisons with RaTG13 (the top row of boxplots) the red line does seem to be "too high". Its value is outside the whiskers for both boxes: it would be an outlier *whether SARS-CoV-2 was engineered or not*.

This implies that our test *is* picking up something else. Another factor that might contribute to different mutation rates inside BsaI/BsmBI sites could be codon usage bias. In other work I have found some evidence that sequential double mutations are more common in close relatives of SARS-CoV-2 than you would expect if they just occurred individually at random. We don't know what might be causing this, but we can't just assume that when something is anomalous that it's caused by engineering.

A caveat to the above is how these mutants were simulated. The simulator works by picking a nucleotide at random from the distribution it finds in the starting genome. It then tries to put that nucleotide in at a random position to replace what was there. But if that would change the protein it gives up and tries again. It keeps doing this until it's achieved the same number of individual silent mutations that actually exist between SARS-CoV-2 or RaTG13 (or whichever virus we're running the test on). The model here is that mutations are close to uniformly distributed in the genome. But, as we found when looking at the odds ratio calculation, a better model seems to take into account that in some places more silent mutations are *possible* than in others.

That is harder to simulate directly, because many of those possible silent mutations require nearby changes to keep them silent. But given that we expect there to be more silent mutations in the sites *anyway* (just because of what they happen to code for) in reality, we might expect that adding extra mutations in the sites would affect the odds ratio less. In other words a more realistic simulation might bring the red arrow down a bit in the RaTG13 simulation. In short: a more realistic simulation might mean that tampering is harder to see, and thus at the same time solve the problem that our test seems to be "reading too high".

When we look at the boxplots for BANAL-20-52 however, the actual OR hits squarely in the green box and nicely outside the pink one. That looks like a good positive result validating both the method and the result. It seems to indicate that this test works to detect tampering in the BsaI/BsmBI sites, and that it has detected it, but only with modest confidence (eyeballing the chart maybe it's somewhere around p=0.05%, which seems consistent with our estimate from Fisher's Exact Test which was p=0.035).

11 Discussion

The main points are:

- After correcting errors, and adjusting the calculation, we have a test that gives us an OR of about 1 on controls, but significantly outlying values on SARS-CoV-2 (See Figure 7). Therefore we can say that there are indeed hotspots of silent mutation in BsaI/BsmBI sites in SARS-CoV-2.
- However extremely low p-values that seem too good to be true probably are: it just isn't that easy to spot such a small number of extra silent mutations (see Figure 8).
- Is the evidence for engineering from this odds ratio calculation enough to balance out the evidence against it in Table 2?

12 Conclusion

The conclusion is left as an exercise for the reader.