Comparing SARS-CoV-2 variants between the Spanish capital and the Balearic Islands

Alexandra Palacios

10 May, 2021

Background and Overview

Here, I analyzed SARS-CoV-2 variants in regions within the different Spanish Islands and compared them with variants found in the Spanish capital, Madrid.

This is a report on SARS-CoV-2, including some variant analysis (Koyama *et al.*, 2020).

Methods

Data Collection

I selected and downloaded a total of 150 SARS-CoV-2 samples from the PRJEB43166 SRA Bioproject located in the NCBI SRA SARS-CoV-2 Bioproject list. This bioproject collected SARS-CoV-2 variant data from individuals of different ages, sex, and geographic region within Spain. 100 of the samples I selected came from localities within two of the Balearic Islands in Spain (Ibiza and Mallorca) and were collected by Servicio de Microbiología, Hospital Universitario Son Espases and SeqCOVID-Spain consortium. The other 50 samples came from the Spanish capital, Madrid, and were collected by Hospital General Universitario Gregorio Marañón and SeqCOVID-Spain consortium.

Variant Analysis

Using a bash pipeline created by (Koyama *et al.*, 2020) and modified by Naupaka Zimmerman, I downloaded all the raw Ilumina fastq data selected from the SRA Bioproject, checked the data quality, trimmed unwanted sequence data, indexed and mapped sequences against the SARS-CoV-2 reference genome, sorted and processed reads, and configured reads to be processed in R as vcf files. The SARS-CoV-2 reference genome came from the NCBI This entire pipeline was driven by a Makefile.

Next, I loaded in, tidied, and stacked all of the vcf files in R, loaded in a gff file with genome annotations

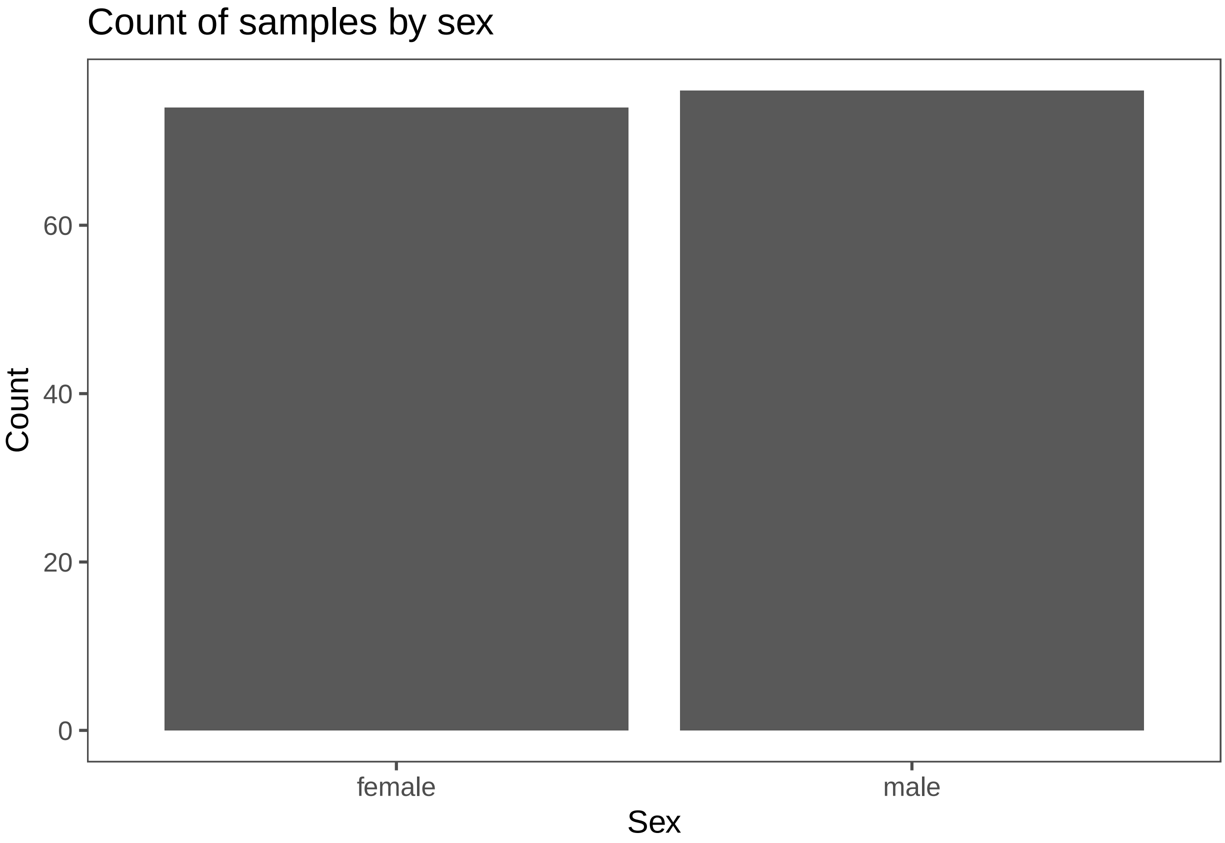
See the set of tutorials on the vcfR package website.

Results

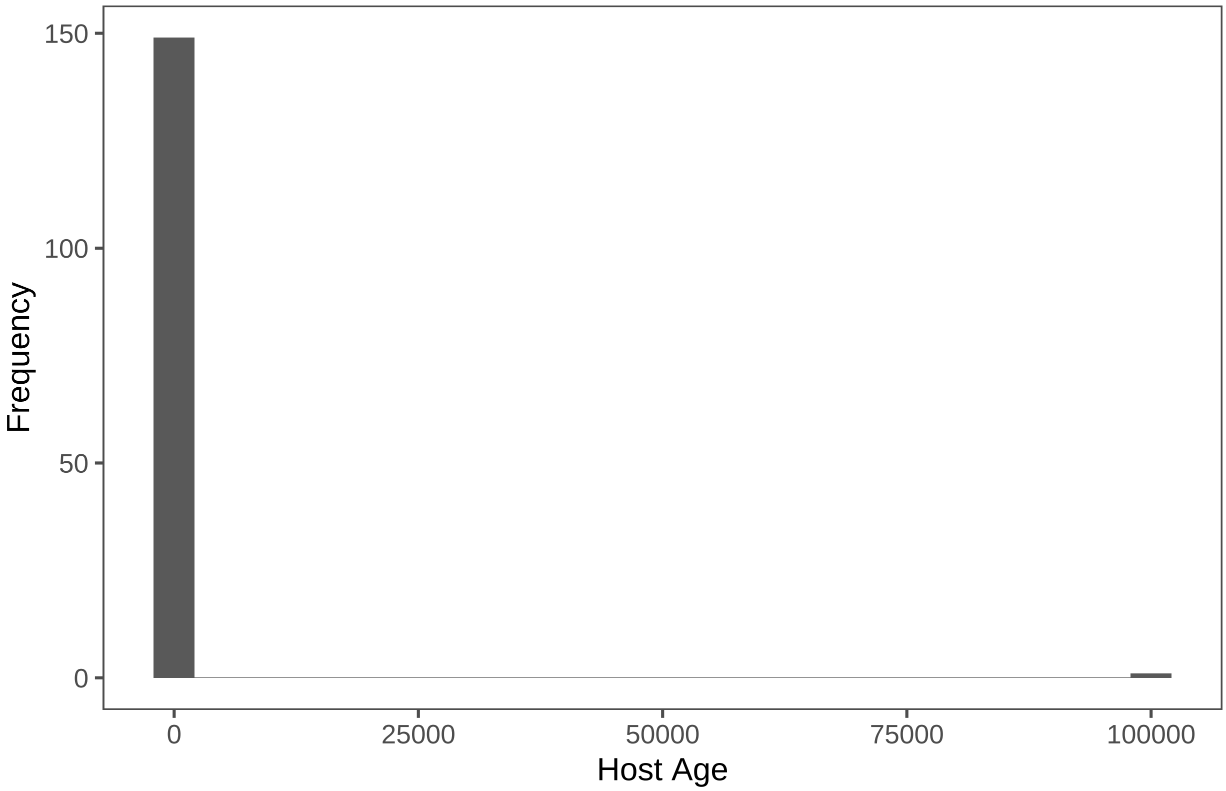
Data from the

Discussion

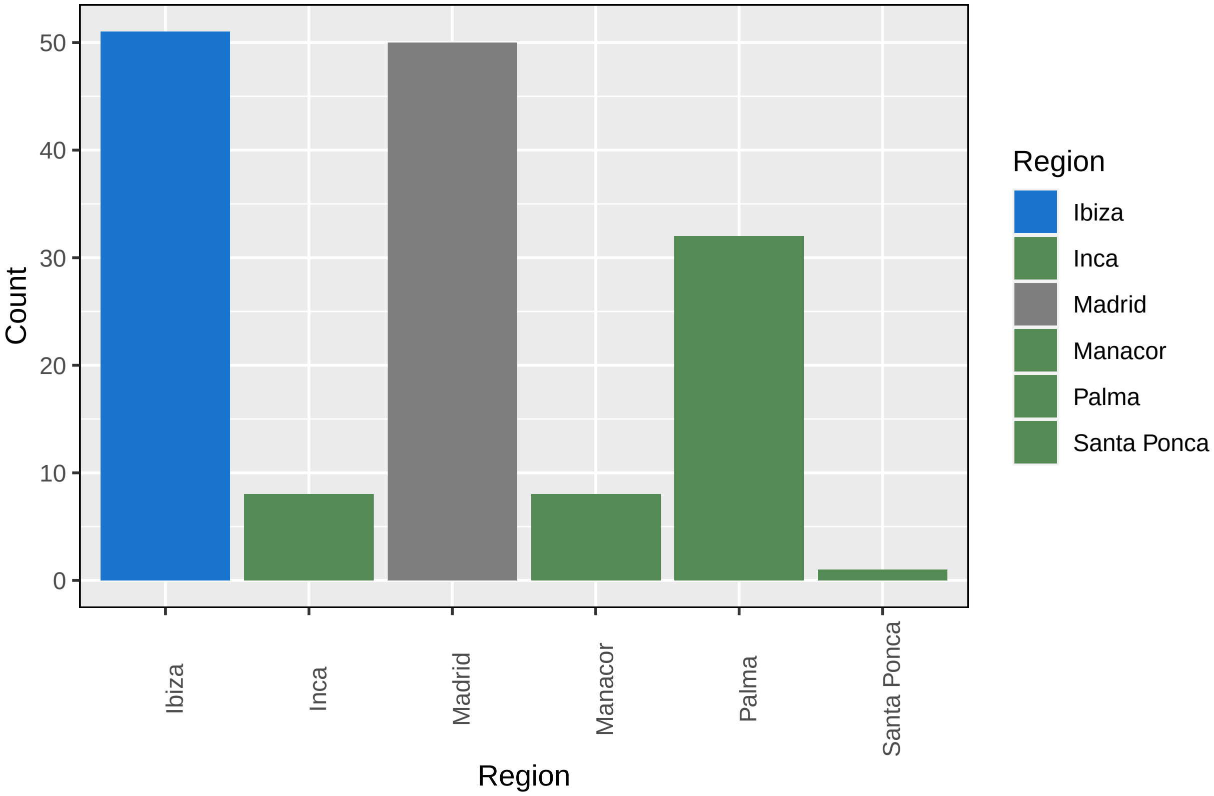
Figures



Histogram of samples by host age



Count of samples by geographic region



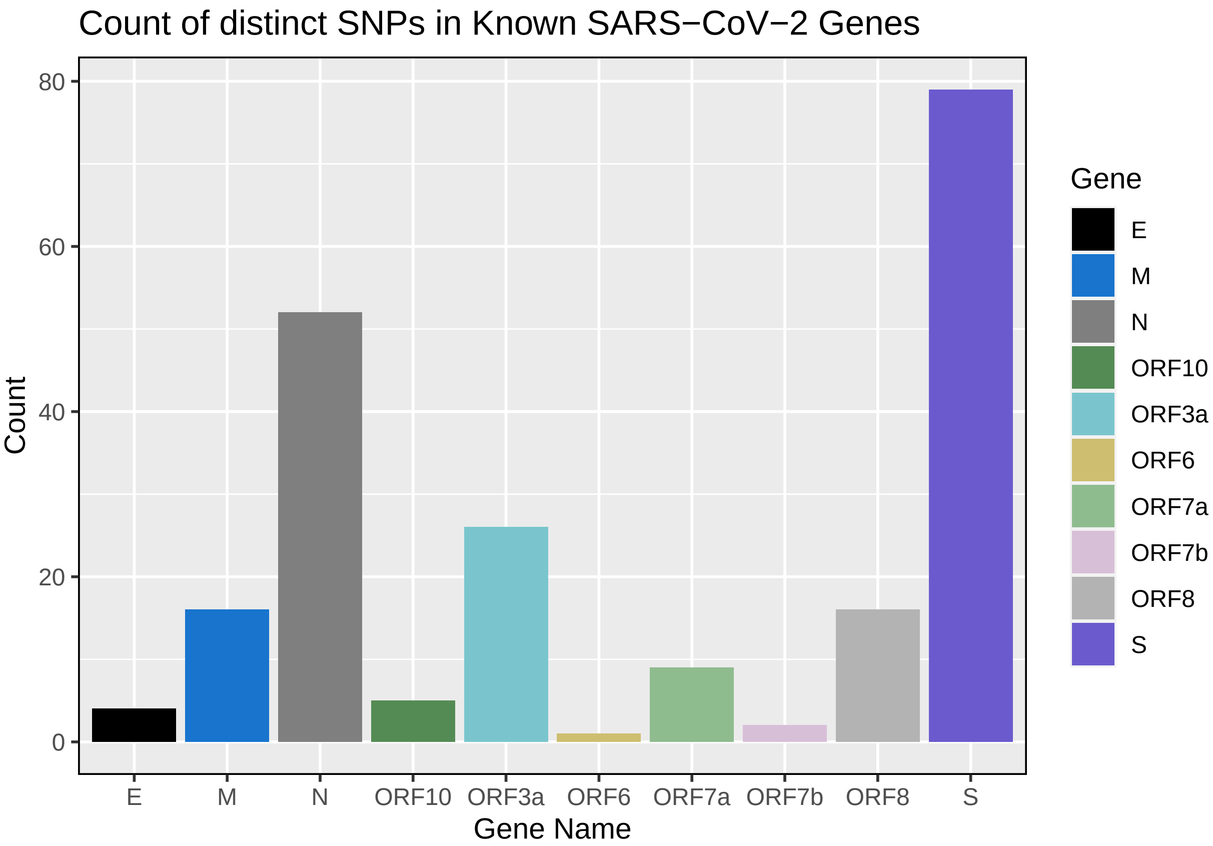
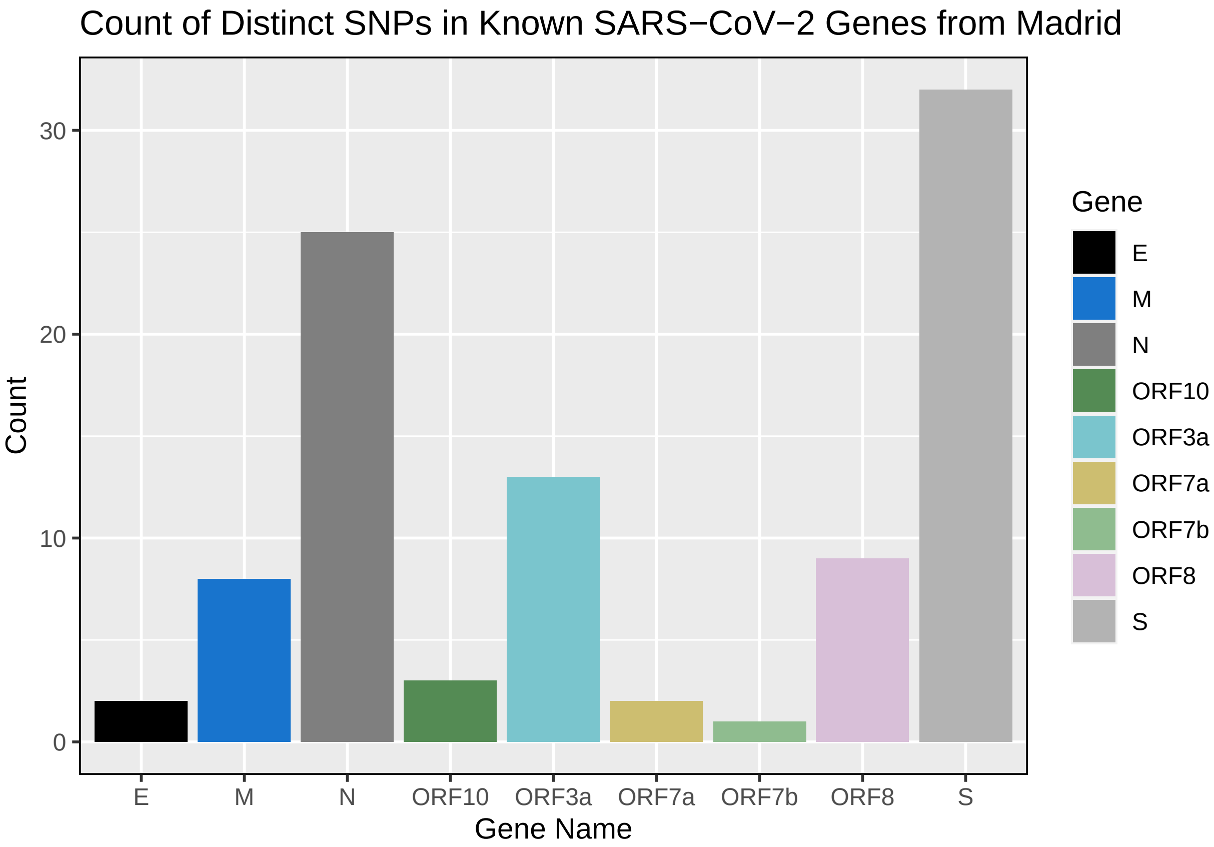
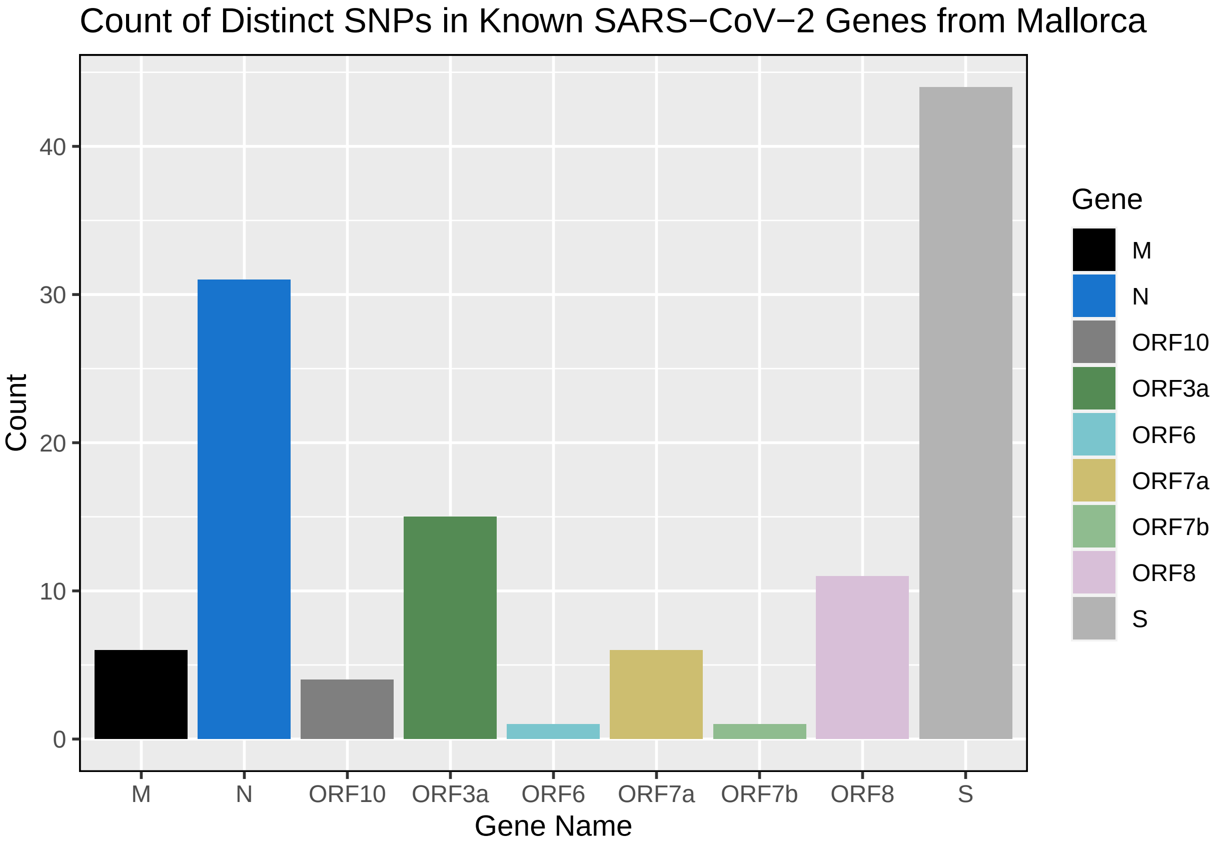
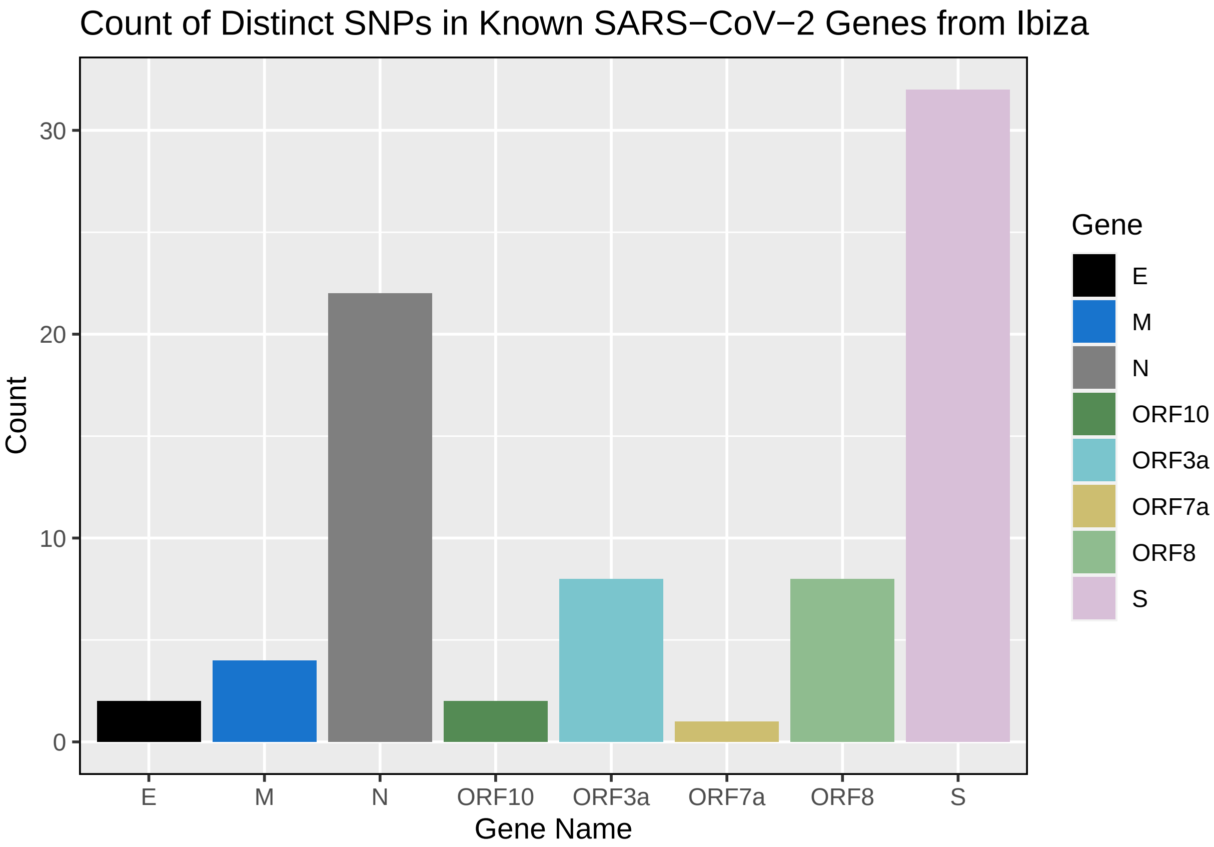


Figure 1: N and S genes have more unique SNPs in the set of samples analyzed.


Count of Distinct SNPs in Known SARS−CoV−2 Genes from Madrid


Count of Distinct SNPs in Known SARS−CoV−2 Genes from Mallorca

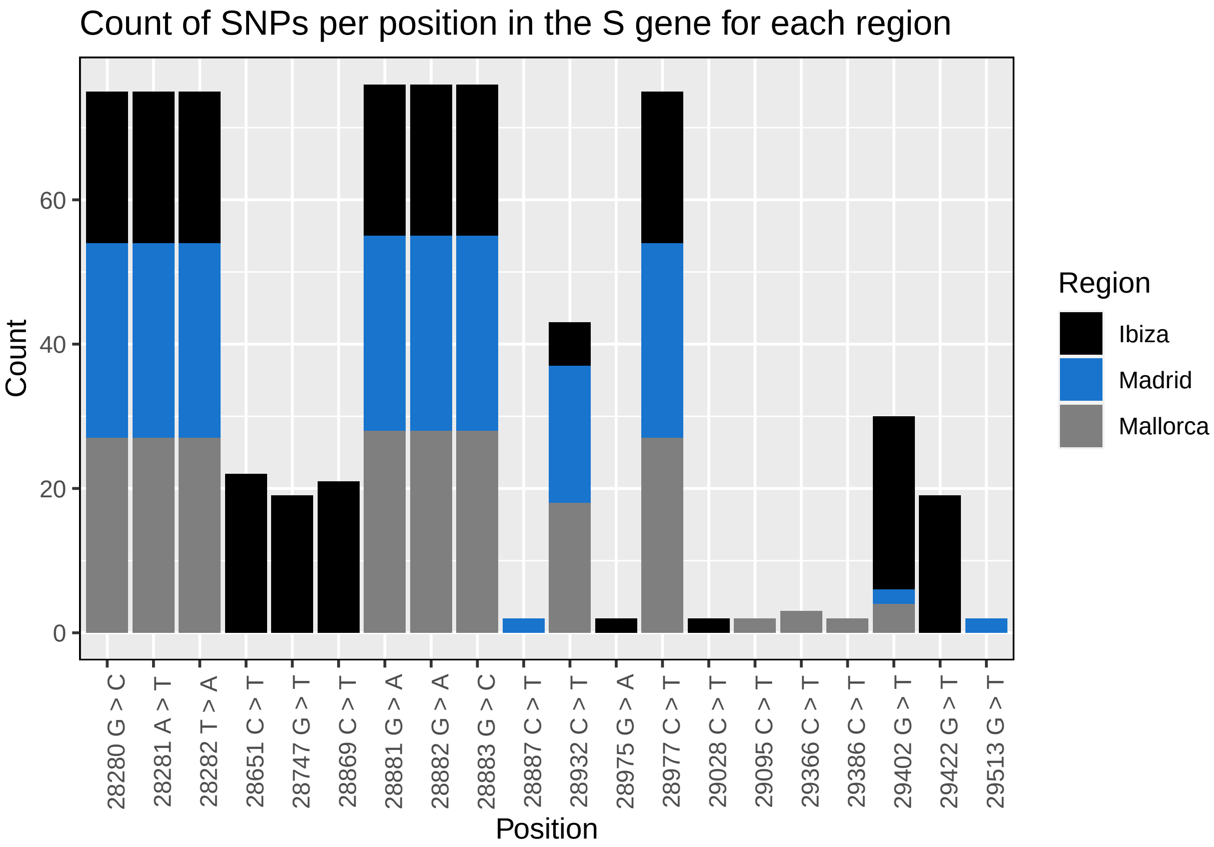


Tables

Reference	Alternate	Position	Gene	Region	Count
A	C	22005	S	Ibiza	1
A	G	22169	S	Palma	1
A	G	23403	S	Ibiza	51
A	G	23403	S	Inca	8
A	G	23403	S	Manacor	8
A	G	23403	S	Palma	32
A	G	23403	S	Santa Ponca	1
A	G	23588	S	Ibiza	6
A	G	24002	S	Palma	1
A	G	24002	S	Santa Ponca	1
A	G	24103	S	Palma	6
A	G	24454	S	Inca	1
A	T	21631	S	Inca	1
A	T	23063	S	Ibiza	23
A	T	23063	S	Inca	2
A	T	23063	S	Manacor	6
A	T	23063	S	Palma	18
A	T	23541	S	Palma	1
A	T	24774	S	Ibiza	1
A	T	24774	S	Palma	1
ATACATGT	AT	21764	S	Ibiza	4

Reference	Alternate	Position	Gene	Region	Count
ATACATGT	AT	21764	S	Manacor	2
ATACATGT	AT	21764	S	Palma	2
C	A	23271	S	Ibiza	21
C	A	23271	S	Inca	2
C	A	23271	S	Manacor	7
C	A	23271	S	Palma	18
C	A	23604	S	Ibiza	23
C	A	23604	S	Inca	2
C	A	23604	S	Manacor	7
C	A	23604	S	Palma	18
C	T	21614	S	Ibiza	1
C	T	21614	S	Palma	6
C	T	21762	S	Manacor	1
C	T	21846	S	Ibiza	1
C	T	21846	S	Palma	1
C	T	21855	S	Ibiza	21
C	T	21855	S	Palma	1
C	T	21859	S	Palma	1
C	T	22227	S	Ibiza	6
C	T	22227	S	Inca	5
C	T	22227	S	Manacor	1
C	T	22227	S	Palma	11
C	T	22227	S	Santa Ponca	1
C	T	22432	S	Palma	1
C	T	22530	S	Ibiza	3
C	T	22858	S	Inca	1
C	T	23613	S	Palma	1
C	T	23625	S	Ibiza	1
C	T	23709	S	Ibiza	21
C	T	23709	S	Inca	2
C	T	23709	S	Manacor	7
C	T	23709	S	Palma	18
C	T	24054	S	Ibiza	1
C	T	24370	S	Ibiza	5
C	T	24370	S	Inca	2
C	T	24370	S	Manacor	1
C	T	24374	S	Ibiza	2
C	T	24418	S	Ibiza	1
C	T	24642	S	Ibiza	1
G	A	22302	S	Palma	1
G	A	23867	S	Ibiza	1
G	A	24893	S	Ibiza	2
G	C	21770	S	Palma	1
G	C	23915	S	Palma	1
G	C	24914	S	Ibiza	21
G	C	24914	S	Inca	2
G	C	24914	S	Manacor	7
G	C	24914	S	Palma	18
G	T	21724	S	Palma	1
G	T	21786	S	Palma	1
G	T	21850	S	Ibiza	6
G	T	21898	S	Palma	1

Reference	Alternate	Position	Gene	Region	Count
G	T	22205	S	Palma	1
G	T	22346	S	Ibiza	1
G	T	23224	S	Inca	1
G	T	23593	S	Palma	1
G	T	25049	S	Palma	1
G	T	25088	S	Inca	1
G	T	25116	S	Palma	1
G	T	25116	S	Santa Ponca	1
G	T	25273	S	Inca	1
G	T	25314	S	Ibiza	1
T	A	23599	S	Palma	6
T	C	21628	S	Ibiza	1
T	C	21771	S	Palma	1
T	C	22828	S	Palma	1
T	C	22909	S	Ibiza	16
T	C	23042	S	Ibiza	2
T	C	24152	S	Ibiza	1
T	C	24847	S	Palma	1
T	G	24307	S	Inca	1
T	G	24506	S	Ibiza	21
T	G	24506	S	Inca	2
T	G	24506	S	Manacor	7
T	G	24506	S	Palma	18
TTTATTA	TTTA	21990	S	Ibiza	21
TTTATTA	TTTA	21990	S	Inca	2
TTTATTA	TTTA	21990	S	Manacor	7
TTTATTA	TTTA	21990	S	Palma	18



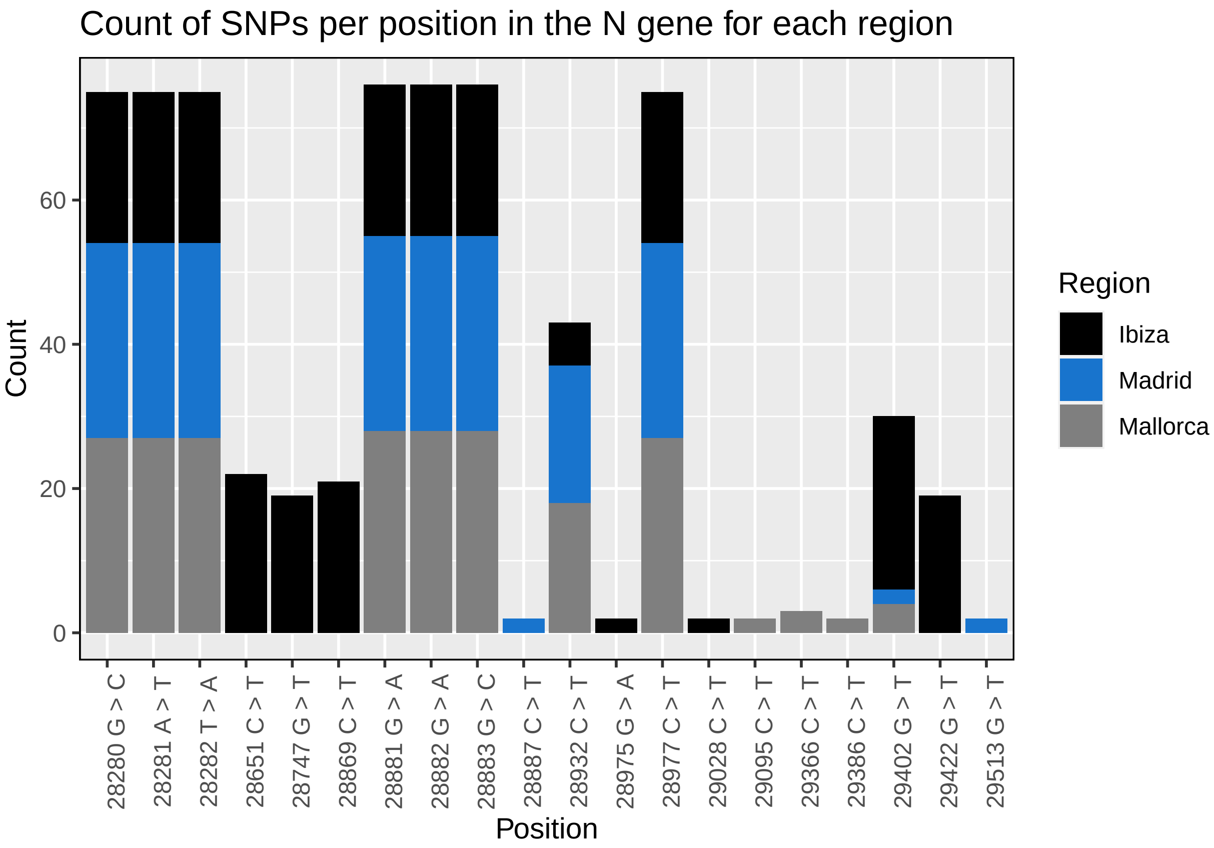
Reference	Alternate	Position	Gene	Region	Count
A	G	22169	S	Mallorca	1
A	G	23403	S	Mallorca	49
A	G	24002	S	Mallorca	2
A	G	24103	S	Mallorca	6
A	G	24454	S	Mallorca	1
A	T	21631	S	Mallorca	1
A	T	23063	S	Mallorca	26
A	T	23541	S	Mallorca	1
A	T	24774	S	Mallorca	1
ATACATGT	AT	21764	S	Mallorca	4
C	A	23271	S	Mallorca	27
C	A	23604	S	Mallorca	27
C	T	21614	S	Mallorca	6
C	T	21762	S	Mallorca	1
C	T	21846	S	Mallorca	1
C	T	21855	S	Mallorca	1
C	T	21859	S	Mallorca	1
C	T	22227	S	Mallorca	18
C	T	22432	S	Mallorca	1
C	T	22858	S	Mallorca	1
C	T	23613	S	Mallorca	1
C	T	23709	S	Mallorca	27
C	T	24370	S	Mallorca	3
G	A	22302	S	Mallorca	1
G	C	21770	S	Mallorca	1
G	C	23915	S	Mallorca	1

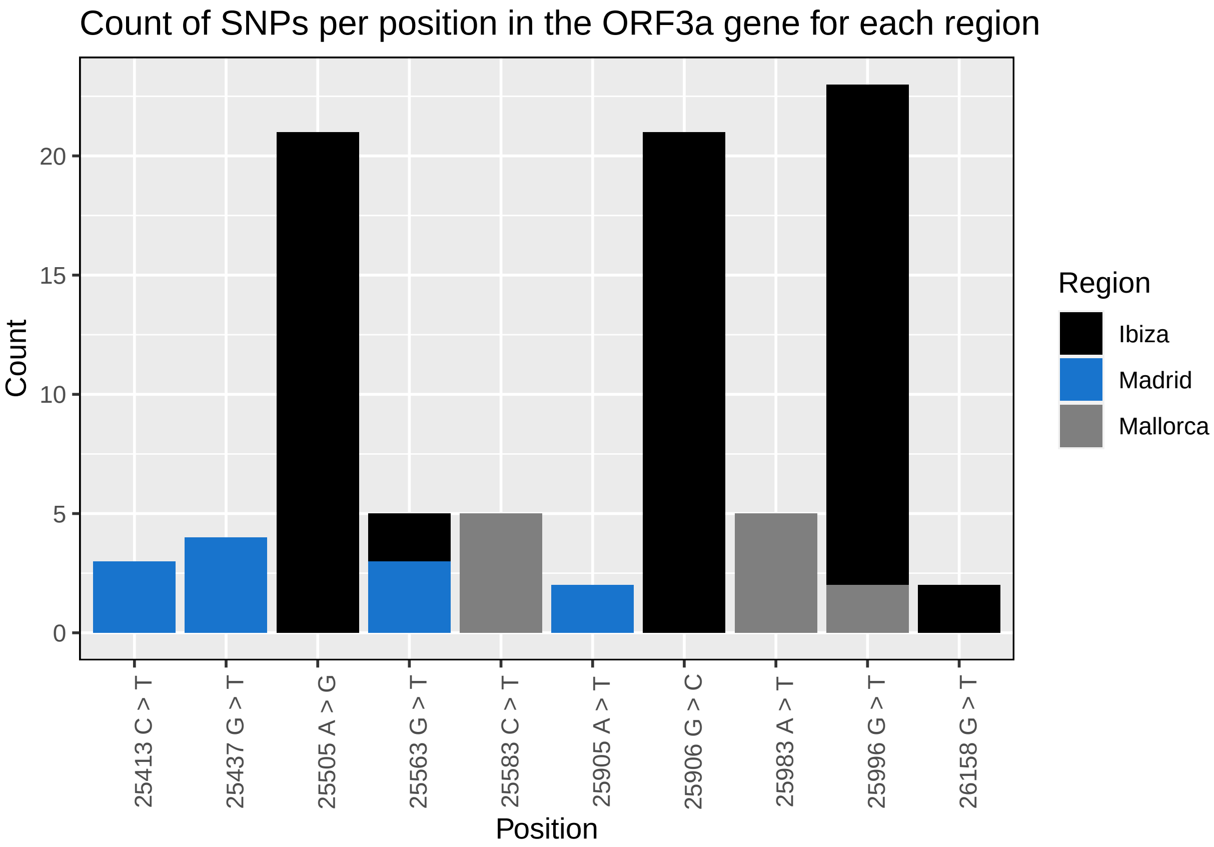
Reference	Alternate	Position	Gene	Region	Count
G	C	24914	S	Mallorca	27
G	T	21724	S	Mallorca	1
G	T	21786	S	Mallorca	1
G	T	21898	S	Mallorca	1
G	T	22205	S	Mallorca	1
G	T	23224	S	Mallorca	1
G	T	23593	S	Mallorca	1
G	T	25049	S	Mallorca	1
G	T	25088	S	Mallorca	1
G	T	25116	S	Mallorca	2
G	T	25273	S	Mallorca	1
T	A	23599	S	Mallorca	6
T	C	21771	S	Mallorca	1
T	C	22828	S	Mallorca	1
T	C	24847	S	Mallorca	1
T	G	24307	S	Mallorca	1
T	G	24506	S	Mallorca	27
TTTATTA	TTTA	21990	S	Mallorca	27

Reference	Alternate	Position	Gene	Region	Count
A	C	22005	S	Ibiza	1
A	G	23403	S	Ibiza	51
A	G	23588	S	Ibiza	6
A	T	23063	S	Ibiza	23
A	T	24774	S	Ibiza	1
ATACATGT	AT	21764	S	Ibiza	4
C	A	23271	S	Ibiza	21
C	A	23604	S	Ibiza	23
C	T	21614	S	Ibiza	1
C	T	21846	S	Ibiza	1
C	T	21855	S	Ibiza	21
C	T	22227	S	Ibiza	6
C	T	22530	S	Ibiza	3
C	T	23625	S	Ibiza	1
C	T	23709	S	Ibiza	21
C	T	24054	S	Ibiza	1
C	T	24370	S	Ibiza	5
C	T	24374	S	Ibiza	2
C	T	24418	S	Ibiza	1
C	T	24642	S	Ibiza	1
G	A	23867	S	Ibiza	1
G	A	24893	S	Ibiza	2
G	C	24914	S	Ibiza	21
G	T	21850	S	Ibiza	6
G	T	22346	S	Ibiza	1
G	T	25314	S	Ibiza	1
T	C	21628	S	Ibiza	1
T	C	22909	S	Ibiza	16
T	C	23042	S	Ibiza	2
T	C	24152	S	Ibiza	1
T	G	24506	S	Ibiza	21

Reference	Alternate	Position	Gene	Region	Count
TTTATTA	TTTA	21990	S	Ibiza	21

```
## # A tibble: 150 x 2
##    sample          n
##    <chr>       <int>
##  1 ERR5530587     34
##  2 ERR5530588     25
##  3 ERR5530589     26
##  4 ERR5530590     20
##  5 ERR5530591     33
##  6 ERR5530593     38
##  7 ERR5530594     23
##  8 ERR5530595     35
##  9 ERR5530596     25
## 10 ERR5530597     39
## # ... with 140 more rows
```





Gene Name	Start	End	Length
S	21563	25384	3821
ORF3a	25393	26220	827
E	26245	26472	227
M	26523	27191	668
ORF6	27202	27387	185
ORF7a	27394	27759	365
ORF7b	27756	27887	131
ORF8	27894	28259	365
N	28274	29533	1259
ORF10	29558	29674	116

Table 2: Gene names, locations, and lengths in the SARS-CoV-2 genome. Higher SNP counts in the S and N genes may be related to the larger size of these genes.

Sources Cited

Koyama,T. *et al.* (2020) Variant analysis of sars-cov-2 genomes. *Bulletin of the World Health Organization*, **98**, 495.