An Analysis of Multiple Sequence Alignments from European COVID-19 patients

Francesco Porto (816042),
Francesco Stranieri (816551)
Ali Manan (817205)

May 2020

1 Abstract

In this report we analyze Multiple Sequence Alignments (MSAs) obtained from 18 samples of European COVID-19 patients and we compare against 3 samples from China, in order to figure out how similar or different they are. We also present our tool for detecting differences in a MSA and three ad-hoc formats for storing them. By analyzing the results of our tool on the provided samples, we make an attempt to either confirm or deny various claims about the origin of the outbreak in Europe.

2 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China, and has since spread globally, resulting in an ongoing pandemic. As of 20 May 2020, more than 4.89 million cases have been reported across 188 countries and territories, resulting in more than 323,000 deaths. More than 1.68 million people have recovered.

With regards to Europe, COVID-19 has affected over 1.755.620 people, resulting to over 159172 deaths as of May 21 [1]. Italy has been the third most affected country, totaling over 227364 cases and 32330 deaths [2].

There are conflicting reports about the spread of COVID-19 in Europe: one of Italy's leading news website suggested it could have started in Germany [3], but other sources [4][5] suggest it could have started in either France or Belgium.

The objective of this report is to either try to support or deny these claims, by using the tools of Bioinformatics, in particular Multiple Sequence Alignments, to understand how similar and/or different sequences from European COVID-19 patiens are to each other.

3 Reads Description

We consider a total of 21 reads, of which 18 from various European countries and 3 from China. While at first it may seem unnecessary to include reads from China in an European-focused analysis, we feel like it may lead to better insights on how the virus mutated from its conception in Wuhan. We will now provide the full list of reads that were used in our analysis, and for each one we report its collection date and place. Note that all reads come from the NCBI website [6].

- **REFERENCE**
  - NC045512 (xx/12)
- **CHINA-WUHAN**
  - LR757995 (26/12)
  - MT291826 (30/12)
- **GERMANY - BAVARIA**
  - MT270104 (xx/01)
  - MT270108 (xx/01)
- **FINLAND**
  - MT020781 (29/01)
- **ITALY**
  - MT066156 (30/01)
  - MT077125 (31/01)
- **GERMANY - BAVARIA**
  - MT270113 (xx/02)
  - MT358638 (xx/02)
- **FRANCE**
  - MT470100 (xx/03)
  - MT470104 (xx/03)
- **GREECE - ATHENS**
  - MT459898 (07/03)
  - MT459899 (07/03)

With xx we mean that only the month is known.

- **SPAIN**
  - MT292577 (08/03)
- **GREECE - ATHENS**
  - MT459836 (14/03)
  - MT459842 (23/03)
  - MT459841 (29/03)
- **NETHERLANDS**
  - MT457399 (28/04)
  - MT457392 (29/04)
  - MT457400 (06/05)

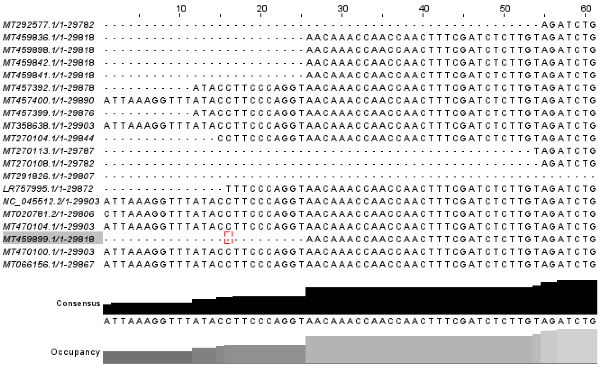
4 Aligners Used

We decided to compare the results of **Clustal-Omega** [7], **KALIGN** [8] and **MUSCLE** [9], three well-known multiple sequence aligners. Note that the tools were used via the freely available EMBL-EBI Web APIs [10]. We will now provide a brief description of each tool, taken from their respective documentations on the EMBL-EBI website:

- **CLUSTAL-Omega** is a multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences.
- **KALIGN** is a fast and accurate multiple sequence alignment algorithm for DNA and protein sequences. This tool uses the *Wu-Manber string-matching algorithm*, to improve both the accuracy and speed of multiple sequence alignment. It is a fast and robust alignment method, especially well suited for the increasingly important task of aligning large numbers of sequences.
- **MUSCLE** stands for MUltiple Sequence Comparison by Log-Expectation. MUSCLE enables high-throughput applications to achieve average accuracy comparable to the most accurate tools previously available, which is expected to be increasingly important in view of the continuing rapid growth in sequence data.

5 Preliminary Graphical Analysis

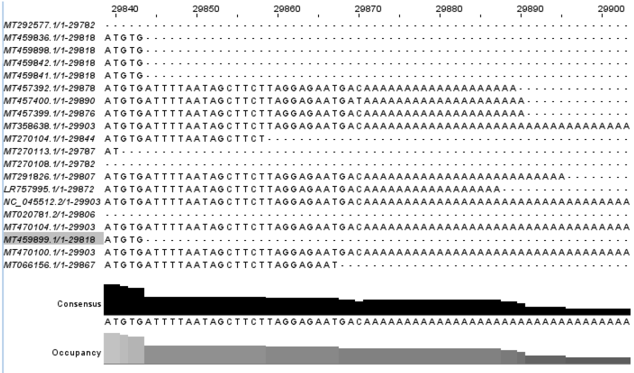
We decided to use **Jalview** [11], a well-known alignment visualization tool, to get a general idea of where differences could be. We found that large differences



```
                             10        20        30        40        50        60
MT292577.1/1-29782  ------------------------------------------------------AGATCTG
MT459836.1/1-29818  -------------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT459898.1/1-29818  -------------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT459842.1/1-29818  -------------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT459841.1/1-29818  -------------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT457392.1/1-29878  -----------ATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT457400.1/1-29890  ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT457399.1/1-29876  -----------ATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT358638.1/1-29903  ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT270104.1/1-29844  --------------CCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT270113.1/1-29787  -----------------------------------------------------TAGATCTG
MT270108.1/1-29782  ------------------------------------------------------AGATCTG
MT291826.1/1-29807  -------------------------------------------------------------
LR757995.1/1-29872  ---------------TTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
NC_045512.2/1-29903 ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT020781.2/1-29806  CTTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT470104.1/1-29903  ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT459899.1/1-29818  -------------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT470100.1/1-29903  ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
MT066156.1/1-29867  ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG

          Consensus
                    ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTG
          Occupancy
```

Figure 1: The left-hand side of an alignment in Jalview.



```
                    29840     29850     29860     29870     29880     29890     29900
MT292577.1/1-29782  -----------------------------------------------------------------
MT459836.1/1-29818  ATGTG------------------------------------------------------------
MT459898.1/1-29818  ATGTG------------------------------------------------------------
MT459842.1/1-29818  ATGTG------------------------------------------------------------
MT459841.1/1-29818  ATGTG------------------------------------------------------------
MT457392.1/1-29878  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAA--------------
MT457400.1/1-29890  ATGTGATTTTAATAGCTTCTTAGGAGAATGATAAAAAAAAAAAAAAAAAAAA-------------
MT457399.1/1-29876  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAA-------------
MT358638.1/1-29903  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
MT270104.1/1-29844  ATGTGATTTTAATAGCTTCT---------------------------------------------
MT270113.1/1-29787  AT---------------------------------------------------------------
MT270108.1/1-29782  -----------------------------------------------------------------
MT291826.1/1-29807  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAAAAAAA--------
LR757995.1/1-29872  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAA----------------
NC_045512.2/1-29903 ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
MT020781.2/1-29806  -----------------------------------------------------------------
MT470104.1/1-29903  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
MT459899.1/1-29818  ATGTG------------------------------------------------------------
MT470100.1/1-29903  ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
MT066156.1/1-29867  ATGTGATTTTAATAGCTTCTTAGGAGAAT------------------------------------

          Consensus
                    ATGTGATTTTAATAGCTTCTTAGGAGAATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
          Occupancy
```

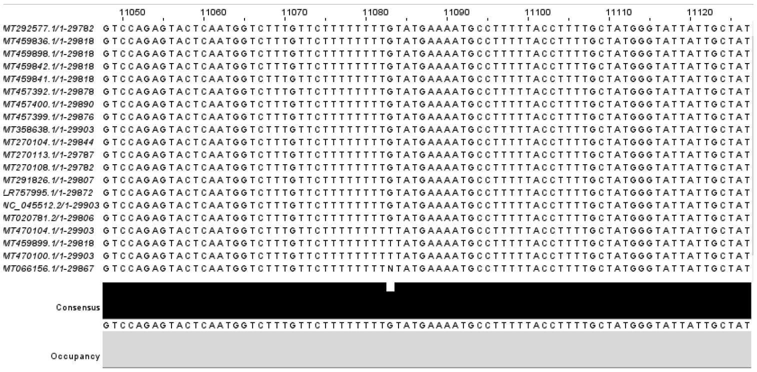
Figure 2: The right-hand side of an alignment in Jalview.

are mostly focused on the first and the last ~60 bases, as shown in the Figure 1 and 2.

Small differences happen somewhat consistently every ~50 to 150 bases, as presented in Figure 3. Note that these results apply consistently to all aligners, with no noticeable difference.

6 Project Description

In order to obtain the list of differences from the reference sequence, we decided to create our own tool, along with a new specific file formats to store and share the differences.



```
                      11050    11060    11070    11080    11090    11100    11110    11120
MT292577.1/1-29782   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT459836.1/1-29818   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT459898.1/1-29818   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT459842.1/1-29818   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT459841.1/1-29818   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT457392.1/1-29878   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT457400.1/1-29890   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT457399.1/1-29876   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT358638.1/1-29903   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT270104.1/1-29844   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT270113.1/1-29787   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT270108.1/1-29782   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT291826.1/1-29807   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
LR757995.1/1-29872   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
NC_045512.2/1-29903  GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT020781.2/1-29806   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT470104.1/1-29903   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTTTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT459899.1/1-29818   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTTTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT470100.1/1-29903   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTTTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT
MT066156.1/1-29867   GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTNTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT

Consensus
                     GTCCAGAGTACTCAATGGTCTTTGTTCTTTTTTTTGTATGAAAATGCCTTTTTACCTTTTGCTATGGGTATTATTGCTAT

Occupancy
```

Figure 3: An example of an isolated difference in the middle of an alignment in Jalview.

6.1 Tool Structure

We provide 2 main scripts:

- **findDifferencesPairwise.py**: a tool that provides differences in a pairwise manner, that is a list of differences between a (single) given sequence and the reference.
- **findDifferencesMSA.py**: a tool that provides differences between multiple sequences and the reference. The differences are also "compacted": if a difference has been found by multiple aligners, it will appear only once.

There's also another script, **DifferencesIO.py** which provides various I/O operations used in the two main scripts. The tool, along with the results (that will be analyzed later) can be found attached to this report.

6.2 Formats Description

We created multiple ad-hoc formats for storing and sharing the differences, which we called **MAD** ("*Multiple* [sequence] *Alignment Difference*"); it comes in 3 versions. More specifically, the first script uses MAD-1 as the output format (which highlights differences between each single sequence and the reference), while the second one uses the MAD-2 and MAD-3 formats (which highlight differences that span across multiple sequences).

The MAD format is inspired by the VCF format [12]. It is composed of two sections:

- A **HEADER** section, which contains information about the number of matches, mismatches and bases not available of a specific sequence.
- A **CONTENT** section, that for each difference contains:
  1. **START**: the starting position of the difference.

2. **LEN**: the length of the difference.
3. **TYPE**: the type of the difference. All the values this field can contain will be specified below.
4. **REF**: the bases of the reference where the difference happens. Contains * where value can be inferred from TYPE.
5. **SEQ**: the bases of the sequences where the difference happens. Contains * where value can be inferred from TYPE.
6. **WHERE**: the sequences where the difference appears (MAD-2 and MAD-3 only).
7. **ALIGNS**: the tools that found the difference (MAD-3 only).

6.3 Event Types

Given a specific position i on the alignment, the possible events that can be specified in the format are:

- **Insertions (INS)**: At position i, reference is '-' while sequence is different from '-'.

- **Deletions (DEL)**: At position i, sequence is '-' while reference is different from '-'.

- **Reference not available (NA1)**: At position i, reference is 'N' while sequence is different from 'N'.

- **Sequence not available (NA2)**: At position i, sequence is 'N' while reference is different from 'N'.

- **Reference and Sequence not available (NA3)**: At position i, both reference and sequence are equal to 'N'.

6.3.1 Differences between MAD-1, MAD-2 and MAD-3

At the beginning of the project, we planned to show differences in a pairwise manner, so we created a format capable of describing differences in such a way. This idea is represented by the **MAD-1** (or simply MAD) and can be used to compare a specific sequence to all the other ones.

As the project went on, we felt like such a format had two critical issues: it was hard to use to actually make comparisons between multiple sequences and/or multiple aligners. In order to solve the first problem, we created the **MAD-2** format, which allows us to display all sequences that present a specific difference in the same row.

To solve the second problem, we created the **MAD-3** format, which is exactly the same as the MAD-2 format but with the ALIGN field added, so that a single file containing all differences (across multiple aligners) is produced.

6.4 Approach

We will now describe our approach for finding differences. Consider the example shown in Table 1, where we have a reference (R) and 3 aligned sequences (S1, S2, S3).

	1	2	3	4	5
R:	A	A	C	G	A
S1:	A	C	G	T	C
S2:	A	C	G	T	A
S3:	A	A	G	T	G

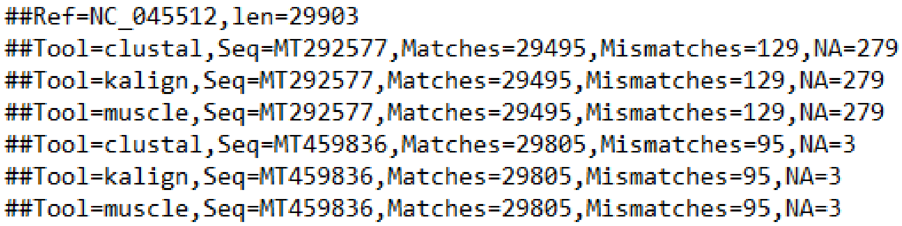
Table 1: The approach employed by our algorithm, visualized.

We will proceed column-wise:

- In column 1, the reference is equal to all sequences, so nothing is done.
- In column 2, S1 and S2 both differ from the reference; they also share the same base (C). We create a new difference that spans S1 and S2, and starts at pos=2. This difference is added to the "active" differences.
- In column 3, S1, S2 and S3 differ from the reference; they also share the same base (G). We create a new difference that spans S1,S2 and S3, and starts at pos=3. We ask ourselves: "Is there an active difference that could be expanded upon?" The answer is no (since there is no active difference that spans S1, s2 AND s3); so we remove the previous difference from the active ones, and we add the new one.
- In column 4, S1, S2 and S3 differ from the reference; they also share the same base (T). We create a new difference that spans S1,S2 and S3, and starts at pos=4. We ask ourselves: "Is there an active difference that could be expanded upon?" The answer is YES (since there is an active difference that spans S1, s2 AND s3), so we increase the length of the difference by 1.
- In column 5, both S1 and S3 differ from the reference, but in different ways (C in S1 and G in S3). We consider these as separate differences, and we try to extend the active difference. It is obvious to see that this cannot be done, so we remove the previous difference from the active ones, and we add the new ones. And so on.

7 Results Analysis

We will analyze the finalResult.MAD3 format which contains all differences found by our tool, grouped by sequences and aligners. Let us start by taking a look at (a small section of) its header from Figure 4.

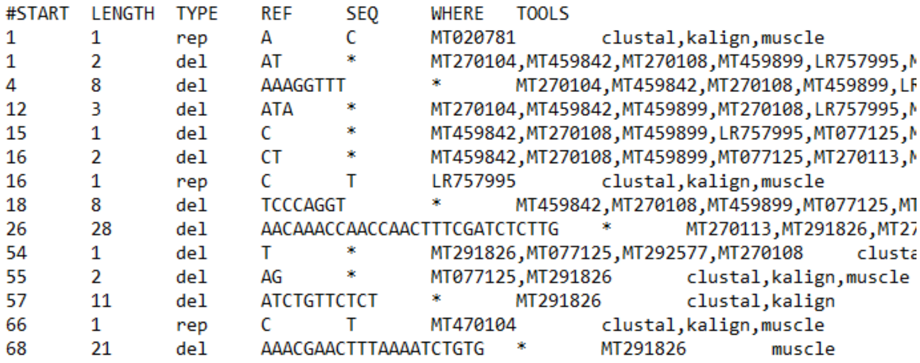


```
##Ref=NC_045512,len=29903
##Tool=clustal,Seq=MT292577,Matches=29495,Mismatches=129,NA=279
##Tool=kalign,Seq=MT292577,Matches=29495,Mismatches=129,NA=279
##Tool=muscle,Seq=MT292577,Matches=29495,Mismatches=129,NA=279
##Tool=clustal,Seq=MT459836,Matches=29805,Mismatches=95,NA=3
##Tool=kalign,Seq=MT459836,Matches=29805,Mismatches=95,NA=3
##Tool=muscle,Seq=MT459836,Matches=29805,Mismatches=95,NA=3
```

Figure 4: Part of the header of finalResult.MAD3.

As we can see, results are consistent across multiple aligners. We think that may be due to the sequences belonging to the same virus, and were taken over a (relatively) small time frame.

As we noticed from Jalview, there are many differences in the left-hand side of the alignments, as seen in Figure 5.



```
#START  LENGTH  TYPE    REF     SEQ     WHERE   TOOLS
1       1       rep     A       C       MT020781        clustal,kalign,muscle
1       2       del     AT      *       MT270104,MT459842,MT270108,MT459899,LR757995,M
4       8       del     AAAGGTTT        *       MT270104,MT459842,MT270108,MT459899,LR
12      3       del     ATA     *       MT270104,MT459842,MT459899,MT270108,LR757995,M
15      1       del     C       *       MT459842,MT270108,MT459899,LR757995,MT077125,M
16      2       del     CT      *       MT459842,MT270108,MT459899,MT077125,MT270113,M
16      1       rep     C       T       LR757995        clustal,kalign,muscle
18      8       del     TCCCAGGT        *       MT459842,MT270108,MT459899,MT077125,MT
26      28      del     AACAAACCAACCAACTTTCGATCTCTTG    *       MT270113,MT291826,MT27
54      1       del     T       *       MT291826,MT077125,MT292577,MT270108     clusta
55      2       del     AG      *       MT077125,MT291826       clustal,kalign,muscle
57      11      del     ATCTGTTCTCT     *       MT291826        clustal,kalign
66      1       rep     C       T       MT470104        clustal,kalign,muscle
68      21      del     AAACGAACTTTAAAATCTGTG   *       MT291826        muscle
```

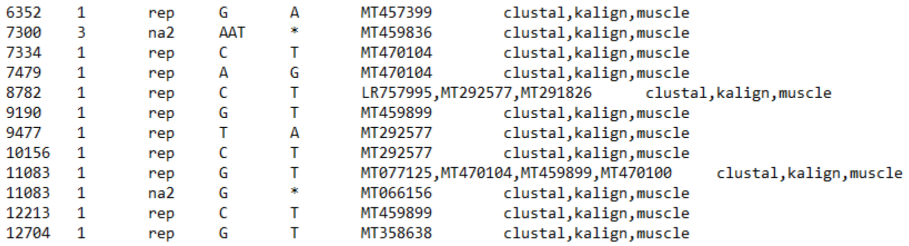
Figure 5: Differences on the left-hand side of the alignment, from finalResult.MAD3.

In the middle of the alignments, only small differences can be found as reflected in Figure 6.

The right-hand side is similar to the left and shown in Figure 7.

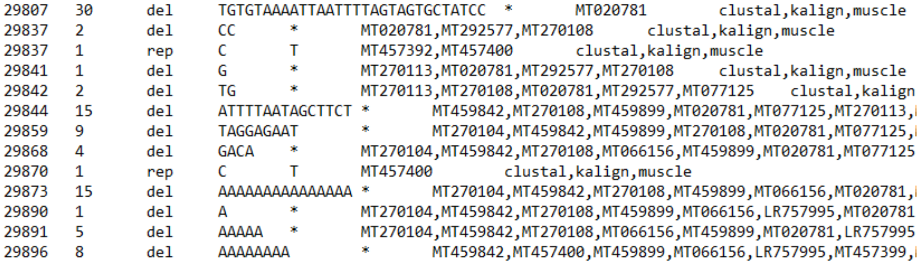
We also found an abnormal difference, illustrated in Figure 8, possibly due to read low quality (we can't be sure since the reads are provided in FASTA format instead of FASTQ) which led to a strange alignment.

Note that the full finalResult.MAD3 file is available in the provided files.



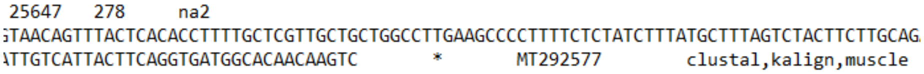
```
6352    1       rep     G       A       MT457399        clustal,kalign,muscle
7300    3       na2     AAT     *       MT459836        clustal,kalign,muscle
7334    1       rep     C       T       MT470104        clustal,kalign,muscle
7479    1       rep     A       G       MT470104        clustal,kalign,muscle
8782    1       rep     C       T       LR757995,MT292577,MT291826      clustal,kalign,muscle
9190    1       rep     G       T       MT459899        clustal,kalign,muscle
9477    1       rep     T       A       MT292577        clustal,kalign,muscle
10156   1       rep     C       T       MT292577        clustal,kalign,muscle
11083   1       rep     G       T       MT077125,MT470104,MT459899,MT470100     clustal,kalign,muscle
11083   1       na2     G       *       MT066156        clustal,kalign,muscle
12213   1       rep     C       T       MT459899        clustal,kalign,muscle
12704   1       rep     G       T       MT358638        clustal,kalign,muscle
```

Figure 6: Differences on the middle of the alignment, from finalResult.MAD3.



```
29807   30      del     TGTGTAAAATTAATTTTAGTAGTGCTATCC  *       MT020781        clustal,kalign,muscle
29837   2       del     CC      *       MT020781,MT292577,MT270108      clustal,kalign,muscle
29837   1       rep     C       T       MT457392,MT457400       clustal,kalign,muscle
29841   1       del     G       *       MT270113,MT020781,MT292577,MT270108     clustal,kalign,muscle
29842   2       del     TG      *       MT270113,MT270108,MT020781,MT292577,MT077125    clustal,kalign
29844   15      del     ATTTTAATAGCTTCT *       MT459842,MT270108,MT459899,MT020781,MT077125,MT270113,
29859   9       del     TAGGAGAAT       *       MT270104,MT459842,MT459899,MT270108,MT020781,MT077125,
29868   4       del     GACA    *       MT270104,MT459842,MT270108,MT066156,MT459899,MT020781,MT077125
29870   1       rep     C       T       MT457400        clustal,kalign,muscle
29873   15      del     AAAAAAAAAAAAAAA *       MT270104,MT459842,MT270108,MT459899,MT066156,MT020781,
29890   1       del     A       *       MT270104,MT459842,MT270108,MT459899,MT066156,LR757995,MT020781
29891   5       del     AAAAA   *       MT270104,MT459842,MT270108,MT066156,MT459899,MT020781,LR757995
29896   8       del     AAAAAAAA        *       MT459842,MT457400,MT459899,MT066156,LR757995,MT457399,
```

Figure 7: Differences on the right-hand side of the alignment, from finalResult.MAD3.



```
 25647   278     na2
GTAACAGTTTACTCACACCTTTTGCTCGTTGCTGCTGGCCTTGAAGCCCCTTTTCTCTATCTTTATGCTTTAGTCTACTTCTTGCAG
ATTGTCATTACTTCAGGTGATGGCACAACAAGTC       *       MT292577        clustal,kalign,muscle
```

Figure 8: An abnormal difference in the alignment.

7.1 Results Visualization

We will now visualize the results via different charts, in order to get a better idea of which sequences are more or less similar to the reference.

Consider the chart in Figure 9. As you can see, the most similar sequence in term of matches is MT358638 (Germany), followed by MT470100 and MT470100 (both from France). This seems to confirm the possible origin of the European strain is located in either of these countries.

Surprisingly, one of the sequences from China (LR757995) is only fourth, while the other is much lower. The same holds true for Italy: one of the reads (MT066156) is in the middle section of the chart, while the other (MT077125) is much lower. We think that may be due to different strains of the virus.

We then have a large group of sequences from Greece (such as MT459842 and MT459898), followed by one from Finland (MT020781).

Something we did not expect is that the sequences from the Netherlands

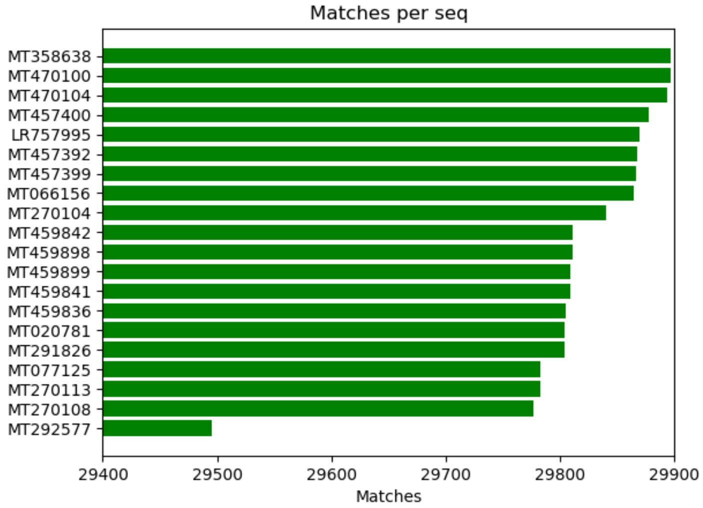


Figure 9: Number of matches with reference, by sequence.

(such as MT457400 and MT457399) are in the top half of the chart, perhaps indicating that the virus has recently arrived there.

With regards to the number of mismatches, according to Figure 10, we have the opposite situation from the chart above.

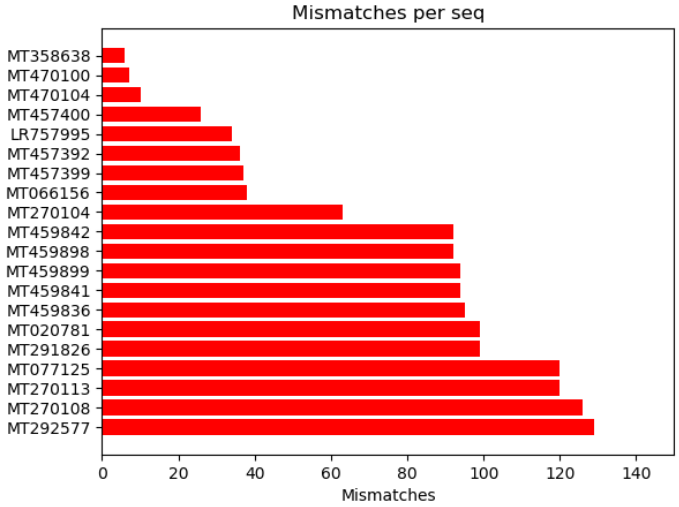


Figure 10: Number of mismatches with reference, by sequence.

We also decided to include the number of bases not available for each sequence, as detailed in Figure 11. The only noticeable result is the one from sequence MT292577 from Spain (the same ”anomaly” that we had found before). We feel like it could be happening due to sequencing errors, but we cannot be sure.

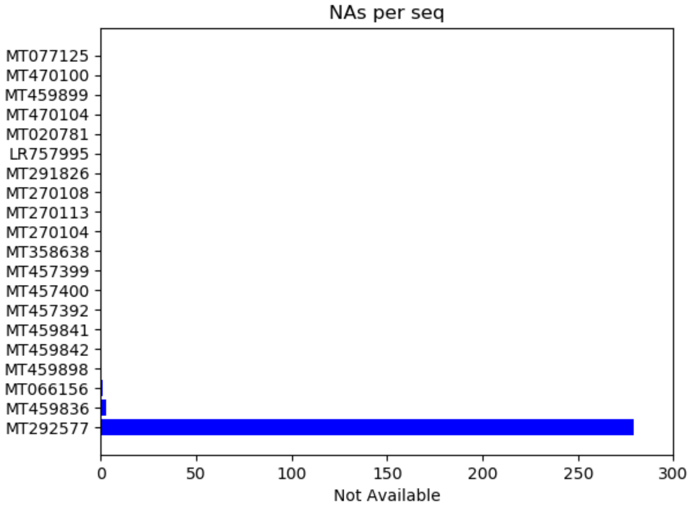


Figure 11: Number of bases not available, by sequence.

Finally, with the chart in Figure 12, we wanted to know what type of difference was the most common.

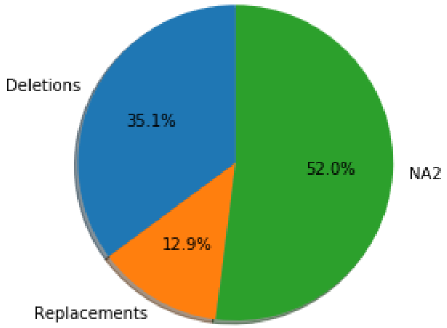


Figure 12: Differences found by type.

Sequence not available (NA2) happens to be the most common type of difference, mostly due to the MT292577 sequence. If we remove it from the data, we obtain the chart in Figure 13 which we think better represents the overall alignment. As we can see, deletions become the most common type of differences, followed by replacements. Now NAs only account for a small amount of the overall differences.

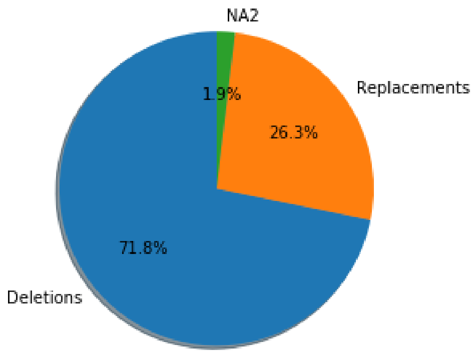


Figure 13: Differences found by type, with the MT292577 sequence removed.

8 Conclusion

In this report we have analyzed the Multiple Sequence Alignments from 18 samples of COVID-19 patients in Europe, and we compared them to 3 samples from China. We have discussed our tool for finding differences in such alignments, and the MAD file formats for storing them. Finally, we have analyzed the results of such tool in order to try to figure out the origin of the pandemic in Europe, and verified the claims that suggest Germany and France seem to be the countries where the European COVID-19 pandemic first started.

9 Work Distribution

We have cooperated throughout the development of this project. More specifically:

- Alì Manan mostly worked on getting the reads, and providing the alignments.
- Francesco Porto mostly worked on tool development and format definition.
- Francesco Stranieri mostly worked on results analysis and tool development.

References

[1] Covid-19 situation update worldwide, as of 21 may 2020. https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. Accessed: 2020-05-22.

[2] Covid-19 situation update for the eu/eea and the uk, as of 21 may 2020. https://www.ecdc.europa.eu/en/cases-2019-ncov-eueea. Accessed: 2020-05-22.

[3] Virus cinese, primo caso in germania. https://www.ansa.it/sito/notizie/topnews/2020/01/28/virus-cinese-primo-caso-in-germania_95c6826b-8f35-49b0-bfcb-04df772d51c4.html. Accessed: 2020-05-22.

[4] Gianfranco Spiteri, James Fielding, Michaela Diercke, Christine Campese, Vincent Enouf, Alexandre Gaymard, Antonino Bella, Paola Sognamiglio, Maria José Sierra Moros, Antonio Nicolau Riutort, Yulia V. Demina, Romain Mahieu, Markku Broas, Malin Bengnér, Silke Buda, Julia Schilling, Laurent Filleul, Agnès Lepoutre, Christine Saura, Alexandra Mailles, Daniel Levy-Bruhl, Bruno Coignard, Sibylle Bernard-Stoecklin, Sylvie Behillil, Sylvie van der Werf, Martine Valette, Bruno Lina, Flavia Riccardo, Emanuele Nicastri, Inmaculada Casas, Amparo Larrauri, Magdalena Salom Castell, Francisco Pozo, Rinat A. Maksyutov, Charlotte Martin, Marc Van Ranst, Nathalie Bossuyt, Lotta Siira, Jussi Sane, Karin Tegmark-Wisell, Maria Palmérus, Eeva K. Broberg, Julien Beauté, Pernille Jorgensen, Nick Bundle, Dmitriy Pereyaslov, Cornelia Adlhoch, Jukka Pukkila, Richard Pebody, Sonja Olsen, and Bruno Christian Ciancio. First cases of coronavirus disease 2019 (covid-19) in the who european region, 24 january to 21 february 2020. *Eurosurveillance*, 25(9), 2020.

[5] Coronavirus: France's first known case 'was in december'. https://www.bbc.com/news/world-europe-52526554. Accessed: 2020-05-22.

[6] Sars-cov-2 (severe acute respiratory syndrome coronavirus 2) sequences. https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/. Accessed: 2020-05-22.

[7] F. Sievers and D. G. Higgins. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.*, 1079:105–116, 2014.

[8] T. Lassmann. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics*, Oct 2019.

[9] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

[10] Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, and Rodrigo Lopez. The embl-ebi search and sequence

analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636—W641, July 2019.

[11] Andrew M. Waterhouse, James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 01 2009.

[12] The variant call format (vcf) version 4.2 specification. `https://samtools.github.io/hts-specs/VCFv4.2.pdf`. Accessed: 2020-05-22.