Background: The exact origin of the cause of the Severe Acute Respiratory Syndrome (SARS) is still an open question. The genomic sequence relationship of SARS-CoV with 30 different single-stranded RNA (ssRNA) viruses of various families was studied using two non-standard approaches. Both approaches began with the vectorial profiling of the tetra-nucleotide usage pattern $V$ for each virus. In approach one, a distance measure of a vector $V$, based on correlation coefficient was devised to construct a relationship tree by the neighbor-joining algorithm. In approach two, a multivariate factor analysis was performed to derive the embedded tetra-nucleotide usage patterns. These patterns were subsequently used to classify the selected viruses.

Results: Both approaches yielded relationship outcomes that are consistent with the known virus classification. They also indicated that the genome of RNA viruses from the same family conform to a specific pattern of word usage. Based on the correlation of the overall tetra-nucleotide usage patterns, the Transmissible Gastroenteritis Virus (TGV) and the Feline CoronaVirus (FCoV) are closest to SARS-CoV. Surprisingly also, the RNA viruses that do not go through a DNA stage displayed a remarkable discrimination against the CpG and UpA di-nucleotide ($z = -77.31$, $-52.48$ respectively) and selection for UpG and CpA ($z = 65.79$, $49.99$ respectively). Potential factors influencing these biases are discussed.

Conclusion: The study of genomic word usage is a powerful method to classify RNA viruses. The congruence of the relationship outcomes with the known classification indicates that there exist phylogenetic signals in the tetra-nucleotide usage patterns, that is most prominent in the replicase open reading frames.
A common question is often asked when investigating viral evolution: what hallmark, in term of genome sequence or RNA word usage, could be used to trace back the emergence of a new pathogen in humans/animals? In particular, Coronavirus are prone to recombination [6,7] and like all other viruses they mutate at a high frequency [8]. This makes extremely hazardous to try to trace the origin of the virus. Nevertheless, this prompted us to investigate their relationships using the RNA word usage hoping to identify some RNA viruses that display similar word usage pattern. Such RNA viruses might hint about the origin of SARS-CoV. This study will contribute to our understanding of the RNA word usage of SARS-CoV and some other pathogenic RNA viruses. In the present study, we explored the relationships of 31 RNA viruses, which are known to cause diseases to their corresponding hosts with either similar symptoms or infectiousness, including SARS-CoV, based on their global tetra-nucleotide usage pattern.

Preliminary analysis of the sequence data indicated that there are 11–14 open reading frames in the SARS-CoV genome [9–11]. The overall gene order for this novel pathogen supported its placement in the family of Coronaviridae which includes the animal/human Coronavirus. It should be emphasized that the sequence similarity shown is attributed mainly to the large RNA-dependent RNA polymerase (replication enzyme or RdRp) residing in the first two open reading frames (ORFs). These two ORFs constitute more than 65% (>20 kb) of the total genome size and these regions are more conserved in their nucleotide sequences due to their specialized role for viral RNA replication. Therefore, the possible relationship based on the sequence of the replication enzyme alone was also investigated.

Results and Discussion

Mono-nucleotide bias

Table 1 presents the breakdown of the RNA sequence into mononucleotide frequencies for the 31 viral genomes in our dataset. Except for the Rabbit Hemorrhagic disease Virus (RHV) that shows a fair usage of the four nucleotides in approximately equal number, the other RNA viruses have a biased genome composition. Bovine Coronavirus (BCoV) and Human Coronavirus 229E (HCoV) favor the U nucleotide (35.5% and 34.6%) at the expense of the C nucleotide (15.3% and 16.7%). Relatively strong nucleotide biases are visible in the other genomes and we will mention a few of the extremes. The highest base count is 28.4% G in the Yellow Fever Virus (YFV), 38.9% A in the Respiratory syncytial Virus (RSV), 35.5% U in the Bovine Coronavirus (BCoV) and 28.5% C count in the Foot-and-Mouth disease Virus (FMV). The lowest base counts are 15.8% G in the Human Respiratory syncytial Virus (HRV), 21.2% A in the Equine arteritis Virus (EV1), 20.9% U in the Igbo Ora Virus (IOV) and 13.6% C in the Bovine ephemeral Fever Virus (BFV). The A nucleotide is the most popular base among RNA viruses (ranging from 21.2% to 38.9%), and C is the most variable nucleotide (ranging from 13.6% to 33.1%).

From the standpoint of the overall genomic composition analysis, the G+C content is an interesting property for a genome, in that the overall content often correlates with the organism pathogenicity [12]. Most of the pathogens genomes have a low G+C content, while some such as Mycobacterium tuberculosis has a relatively high G+C content. Therefore, as expected in Table 1, we noted that most of the pathogenic viruses are A+U-rich (>50%), except for Porcine reproductive and Respiratory syndrome Virus (PRV), Equine arteritis virus (EV1), Rabbit hemorrhagic disease virus (RHV), Simian hemorrhagic Fever Virus (SFV) and Foot-and-Mouth disease Virus C (FMV).

Di-nucleotide bias

The frequencies of occurrence for di-nucleotides were compared to the random RNA counterparts having the same base proportion in order to compute the z value that reflected their di-nucleotide bias (Table 2). Among the 31 virus sequences examined, the frequencies of occurrence for di-nucleotide were not randomly distributed, with only a few exceptional di-nucleotides starting with a purine residue present at the expected frequencies (ApC, ApG, GpC, |z| < 3). A remarkable deviation from the expected frequencies occurs for the di-nucleotide pairs CpG and UpA (suppression or under-representation, z < -50) as well as di-nucleotides pairs CpA and UpG (enhancement or over-representation, z > 40). These di-nucleotide biases, together with mono-nucleotide bias [13], have a direct impact on the codon usage of viruses. For example, in the codon usage for the 24 protein coding sequences in human Coronavirus 229E (Table 3), only 2.85% of codons contain the under-represented subword CpG di-nucleotide whereas 11.26% of the codons contain the over-represented CpA di-nucleotide (the aggregate codon usage containing each di-nucleotide subword without mono- and di-nucleotide bias is close to 6.25%).

In double stranded DNA genomes the deficiency in di-nucleotide CpG is often supposed to be due to the fact that they are the targets for methyltransferase activity that leads to cytosine deamination [14,15]. It is however unlikely that the mechanism of deamination that alters the genetic contents at the DNA level would affect the viral RNA content of most RNA viruses without a DNA stage. There might exist specific cytosine RNA methylases that...
could be responsible for this effect [16]. However it is more consistent to propose that, unlike the mechanism of cytosine deamination in the DNA realm, the dominating process is cytosine deamination in RNA viruses, converting cytosine to uracil (C\textastersiskoid) instead of thymine (T). As a consequence of this mechanism, di-nucleotide CpG changes to either di-nucleotide UpG or CpA in the direct/complementary strands of RNA viruses and causes the over-representation in di-nucleotide UpG and CpA (z > 19). Interestingly, there is experimental evidence in vitro that the rate of cytosine deamination is faster (>100 times) in the single stranded than in double-stranded state [17]. Apart from the under-representation in di-nucleotide CpG and over-representation in di-nucleotide UpA and CpA, the reason for the observed di-nucleotide UpA scarcity in RNA may be explained by its chemical lability [18]. The UpA dinucleotide is chemically the most unstable among the 16 dinucleotides. Furthermore, UpA

Table 1: RNA virus in current study.

| Virus Name                          | Type                | Accession Number | DNA Stage | Segment | Acronym | Size (nt) | G   | A   | U   | C   | A+U% |
|-------------------------------|-------------------|-----------------|---------|---------|---------|----------|-----|-----|-----|-----|------|
| Avian infectious bronchitis virus | ss-RNA             | NC_001451       | No      | I       | ABV     | 27608    | 21.7| 28.9| 33.2| 16.2| 62.1 |
| Bovine coronavirus             | ss-RNA             | NC_003045       | No      | I       | BCoV    | 31028    | 21.8| 27.4| 35.5| 15.3| 62.9 |
| Equine arteritis virus         | ss-RNA             | NC_002532       | No      | I       | EVI     | 12704    | 26.0| 21.2| 27.1| 25.6| 48.3 |
| Human coronavirus 229E         | ss-RNA             | NC_002645       | No      | I       | HCoV    | 27317    | 21.6| 27.2| 34.6| 16.7| 61.7 |
| Lacasse dehydrogenase-elevating virus | ss-RNA         | NC_002534       | No      | I       | LDV     | 14225    | 25.9| 23.1| 28.2| 22.6| 51.3 |
| Murine hepatitis virus         | ss-RNA             | NC_001846       | No      | I       | MHV     | 31357    | 23.9| 26.0| 32.3| 17.9| 58.2 |
| Porcine epidemic diarrhea virus| ss-RNA             | NC_003436       | No      | I       | PDV     | 28033    | 22.8| 24.7| 33.2| 19.2| 58.0 |
| Porcine reproductive and respiratory syndrome virus | ss-RNA | NC_001961 | No | I | PRV | 15428 | 26.2| 21.7| 25.3| 26.7| 47.0 |
| Feline coronavirus             | ss-RNA             | NC_002534       | No      | I       | LDV     | 14225    | 25.9| 23.1| 27.4| 25.6| 48.3 |
| Simian hemorrhagic fever virus  | ss-RNA             | NC_002534       | No      | I       | TGV     | 28586    | 20.6| 29.5| 32.9| 17.0| 62.4 |
| Transmissible gastroenteritis virus | ss-RNA         | NC_002534       | No      | I       | AEY     | 7055     | 25.7| 27.0| 28.3| 19.0| 55.3 |
| Avian encephalomyelitis virus   | ss-RNA             | NC_002534       | No      | I       | BDV     | 12355    | 25.2| 32.7| 22.3| 19.8| 54.9 |
| Bovine viral diarrhea virus genotype 2 | ss-RNA         | NC_002534       | No      | I       | PMV     | 8115     | 25.6| 24.8| 21.2| 28.5| 45.9 |
| Foot-and-mouth disease virus C  | ss-RNA             | NC_002534       | Yes     | 2       | HV1     | 11772    | 0.21| 0.33| 0.29| 0.17| 62.7 |
| Igbo Ora virus                 | ss-RNA             | NC_002534       | No      | I       | IOV     | 11821    | 24.1| 31.1| 20.9| 24.0| 51.9 |
| Poliovirus                     | ss-RNA             | NC_002534       | No      | I       | PV1     | 7440     | 23.0| 29.7| 24.0| 23.3| 53.7 |
| Rabbit hemorrhagic disease virus| ss-RNA             | NC_002534       | No      | I       | RHV     | 7437     | 25.5| 25.9| 23.9| 24.7| 49.8 |
| Tamana bat virus               | ss-RNA             | NC_002534       | No      | I       | TBY     | 10053    | 21.5| 33.2| 28.3| 16.9| 61.6 |
| Yellow fever virus             | ss-RNA             | NC_002534       | No      | I       | YFV     | 10862    | 0.28| 0.27| 0.23| 0.21| 0.50 |

The information about 31 RNA viruses being investigated in this study. Their accession number, abbreviation, genome size, number of segments and whether they undergo DNA stage are tabulated. The breakdown of the RNA nucleic acids and A+U contents are also shown.
appears to be a preferential target for ribonucleases [19]. This lability would create a selection pressure against di-nucleotide UpA in RNA viruses.

If we choose a critical value for \( z \) (|\( z \) = 3.29) that only allows a chance of 1 in 1000 error for classifying a word as biased (over/under-represented), all di-nucleotides show some kind of bias in their usage pattern across 31 different viruses (Table 4, derived from the complete form of Table 2 provided as the additional file 1). The causes for these biases await further investigation.

**Tetra-nucleotide bias**

Inspection of the tetra-nucleotide usage pattern for RNA viruses (additional file 2) reveals considerable differences. The frequencies of occurrence for tetra-nucleotides were compared to artificial chromosomes constructed as random RNA sequences having the same nucleotide succession up to order three to compute the \( z \) values that reflect their tetra-nucleotide bias in the corresponding virus (Table 5). If we choose a critical value for \( z \) (|\( z \) = 3.29) that only allows a chance of 1 in 1000 error for classifying a word as over/under-represented, 96% of the tetra-nucleotides show a strong bias in their usage pattern across 31 viruses (shown in Table 4, derived from the complete form of Table 5 provided as the additional file 1). This indicated strongly that tetra-nucleotides are being used in a different manner between different viruses, providing us with a tool to study the relationships between viruses based on the tetra-nucleotide bias exhibited in their genomes.

### Table 2: Di-nucleotide bias for six RNA viruses.

| Di-nucleotide | BCoV | MHV | SARS | ABV | HCoV | PDV | Average z value across 31 viruses |
|---------------|------|-----|------|-----|------|-----|-------------------------------|
| CG            | 497  | 1034| -103.8| 798 | 1342|     | 104.4                         |
| GC            | 1344 | 1037| -62.19| 1694| 1341|     | 36.12                         |
| AU            | 2845 | 3007| -26.44| 1968| 1941|     | -58.91                        |
| UA            | 3181 | 3000| -30.10| 2404| 2616|     | -35.12                        |
| AG            | 1824 | 1848| -4.25 | 1968| 1941|     | -4.77                         |
| GA            | 1629 | 1849| -39.08| 1745| 1941|     | -32.74                        |
| AC            | 1371 | 1303| 12.93 | 1384| 1458|     | -13.50                        |
| CA            | 1594 | 1297| 56.03 | 1705| 1453|     | 46.19                         |
| CU            | 1801 | 1674| 22.52 | 1874| 1806|     | 12.28                         |
| UC            | 1179 | 1867| -88.35| 1296| 1802|     | -94.30                        |
| GU            | 2449 | 2394| 9.10  | 2473| 2402|     | 11.92                         |
| UG            | 3101 | 2392| 120.25| 3146| 2408|     | 128.13                        |

The di-nucleotide bias in six RNA viruses. The \( z \) value quantifies the di-nucleotide bias as defined in equation 1. \( N(w) \) and \( E(w) \) are actual and expected frequency of occurrence for word \( w \). The last column is the average \( z \) value across 31 RNA viruses.

### Table 3: Codon usage for Human CoronaVirus 229E (HCoV).

| Amino Acid | Codon | Usage/% | Amino Acid | Codon | Usage/% |
|------------|-------|---------|------------|-------|---------|
| Arg        | CGU   | 1.04    | Ile        | AUU   | 3.34    |
|            | CGC   | 0.41    |            | AUC   | 0.74    |
|            | CGA   | 0.17    |            | AUA   | 1.35    |
|            | CGG   | 0.13    |            | Gly   | GGU     | 4.12 |
|            | AGA   | 1.23    |            | GCC   | 1.43    |
|            | AGG   | 0.36    |            | GAA   | 0.67    |
| Leu        | UUA   | 1.49    |            | GAG   | 0.22    |
|            | UUG   | 2.96    |            | Val   | GUU     | 6.00 |
|            | CUG   | 2.48    |            | GUC   | 1.23    |
|            | CUC   | 0.46    |            | GUA   | 1.09    |
|            | CUA   | 0.65    |            | GUG   | 1.90    |
|            | CUG   | 0.63    |            | Lys   | AAA     | 3.15 |
| Ser        | UCU   | 2.70    |            | Asn   | AAU     | 4.15 |
|            | UCC   | 0.66    |            | GAC   | 1.82    |
|            | UCA   | 1.37    |            | AAC   | 2.04    |
|            | UCG   | 0.20    |            | Gln   | CAA     | 1.17 |
|            | AGU   | 1.86    |            | GAG   | 0.46    |
|            | AGC   | 0.71    |            | his   | CAU     | 1.14 |
| Thr        | ACU   | 3.23    |            | CAC   | 0.46    |
|            | ACC   | 0.76    |            | Glu   | GAA     | 2.81 |
|            | ACA   | 2.21    |            | GAG   | 1.21    |
|            | ACG   | 0.29    |            | Asp   | GAU     | 3.09 |
| Pro        | CUC   | 1.65    |            | GAC   | 1.96    |
|            | CCC   | 0.35    |            | Tyr   | UAU     | 3.00 |
|            | CCA   | 1.07    |            | UAC   | 1.46    |
|            | CCG   | 0.19    |            | Cys   | UGU     | 2.26 |
| Ala        | GCU   | 3.58    |            | UGC   | 0.95    |
|            | GCC   | 0.83    |            | Phe   | UUU     | 4.59 |
|            | GCA   | 1.80    |            | UUC   | 1.10    |
|            | GCG   | 0.42    |            |        |         |

The relative usage of synonymous codons in the 24 known CDSs of Human Corona Virus 229E (HCoV).
Table 4: Overall statistics for biased di-nucleotides and tetra-nucleotides.

| Percentage of di-nucleotide that can be used to discriminate between viruses ($|z| > 3.29$) | Percentage of tetra-nucleotide that can be used to discriminate between viruses ($|z| > 3.29$) | Virus | Percentage of biased di-nucleotide ($|z| > 3.29$)/% | Percentage of biased tetra-nucleotide ($|z| > 3.29$)/% |
|---|---|---|---|---|---|
| 100% | 96.09% | BCoV | 93.8 | 29.7 |
| MHV | 93.8 | 28.1 |
| SARS | 81.3 | 34.4 |
| ABV | 81.3 | 27.3 |
| HCoV | 93.8 | 31.3 |
| HCoV | 93.8 | 31.3 |
| PDV | 81.3 | 28.5 |
| TGV | 87.5 | 31.6 |
| LDV | 93.8 | 19.5 |
| PRV | 93.8 | 15.6 |
| SFV | 93.8 | 16.0 |
| FCoV | 75.0 | 11.7 |
| EV1 | 87.5 | 14.5 |
| TBV | 75.0 | 21.9 |
| AEV | 93.8 | 11.7 |
| PV1 | 87.5 | 11.7 |
| YFV | 93.8 | 29.3 |
| BDV | 87.5 | 17.6 |
| HLV | 93.8 | 9.4 |
| FMV | 87.5 | 12.1 |
| IOV | 75.0 | 9.8 |
| HVI | 62.5 | 12.5 |
| RSV | 87.5 | 18.8 |
| HRV | 87.5 | 19.1 |
| BRV | 93.8 | 19.9 |
| TV2 | 81.3 | 15.2 |
| REV | 87.5 | 18.4 |
| MV1 | 81.3 | 15.2 |
| CDV | 75.0 | 16.0 |
| APV | 93.8 | 11.7 |
| BFV | 87.5 | 15.2 |
| IBV | 87.5 | 23.4 |

The percentage of biased di-nucleotides and tetra-nucleotides that shows strong biases ($|z| > 3.29$) in 31 RNA viruses (right). For di-nucleotides, all 16 (100%) of them show strong biases in part of or all 31 RNA viruses. For tetra-nucleotides, 246 (96%) of the tetra-nucleotides show strong biases in part of or all 31 RNA viruses.

Table 5: Tetra-nucleotide bias for three RNA viruses. The tetra-nucleotide bias in three viruses.

| Tetra-nucleotide | N(w) | E(w) | z | N(w) | E(w) | z | N(w) | E(w) | z | N(w) | E(w) | z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AAAA | 148 | 206.2 | -7.4 | 147 | 145.8 | 0.2 | 222 | 216.5 | 0.7 | 187 | 170.9 | 2.2 | 170 | 183.9 | -1.7 |
| AAAC | 110 | 103.7 | 1.1 | 98 | 91.2 | 1.3 | 154 | 148.1 | 0.9 | UAAC | 78 | 105.1 | -4.8 | 85 | 122.4 | -6.1 |
| AAAU | 217 | 320.9 | 0.2 | 210 | 206.6 | 1.6 | 230 | 200.6 | 1.6 | UAUU | 209 | 309.9 | 0.3 | 245 | 259.8 | -1.7 |
| AACA | 133 | 114.1 | 3.2 | 113 | 112.1 | 0.2 | 215 | 175.7 | 5.4 | UACA | 178 | 163.7 | 2.0 | 122 | 123.4 | 0.2 |
| AACG | 29 | 40.7 | -3.3 | 35 | 61.3 | -6.2 | 44 | 66.3 | -5.0 | UACG | 50 | 54.4 | -1.1 | 58 | 57.8 | -0.3 |
| AACU | 20 | 40.7 | -3.3 | 35 | 61.3 | -6.2 | 44 | 66.3 | -5.0 | UACU | 50 | 54.4 | -1.1 | 58 | 57.8 | -0.3 |
| AAUG | 246 | 229.3 | 2.0 | 234 | 232.1 | 0.2 | 230 | 201.5 | 3.7 | UAGG | 310 | 271.5 | -4.3 | 278 | 238.1 | 3.6 |
| AAUU | 265 | 265.5 | -0.1 | 212 | 207.8 | 0.5 | 211 | 212 | -0.1 | UAUU | 314 | 345 | -3.0 | 253 | 248.5 | 0.5 |

The z value quantifies the tetra-nucleotide bias, as defined in equation (1). N (w) and E (w) are actual and expected frequency of occurrence for word w.
| ACAA | 144 | 137.1 | 1.1 | 115 | 114.1 | 0.2 | 269 | 204.1 | 8.3 |
| ACAC | 84 | 66.4 | 3.9 | 88 | 75.4 | 2.6 | 168 | 142.2 | 3.9 |
| ACAG | 118 | 105.7 | 2.2 | 104 | 109.4 | 0.5 | 151 | 145.1 | 0.9 |
| ACAT | 128 | 123.5 | 0.7 | 106 | 122.7 | -2.7 | 186 | 172.9 | 0.0 |
| ACCT | 105 | 76.9 | 5.8 | 116 | 85.1 | 6.1 | 161 | 117.7 | 3.3 |
| ACCG | 56 | 37.3 | 5.6 | 84 | 57.6 | 6.3 | 54 | 60.7 | -1.6 |
| ACCU | 24 | 35.5 | -3.5 | 52 | 57.4 | -1.3 | 31 | 48.8 | -4.6 |
| AGCU | 83 | 77.7 | 1.1 | 97 | 97.1 | -0.1 | 119 | 132.2 | 1.2 |
| AGCU | 29 | 34.1 | -1.6 | 49 | 52.5 | -5.0 | 31 | 54.8 | -5.8 |
| ACUG | 32 | 44.5 | -3.1 | 34 | 56.5 | -4.2 | 40 | 50.4 | -2.2 |
| ACUG | 57 | 71.8 | 3.3 | 87 | 80.6 | 1.3 | 114 | 117.2 | 0.5 |
| AGGU | 107 | 88.9 | 3.5 | 115 | 103.4 | 2.1 | 116 | 112.3 | 0.5 |
| AGGU | 177 | 170.4 | 0.9 | 158 | 145.2 | 1.9 | 128 | 141.1 | -2.0 |
| AGGU | 113 | 105.8 | 1.3 | 102 | 112.1 | -1.7 | 105 | 118.8 | -2.3 |
| AGGC | 77 | 54.1 | 6.7 | 91 | 80.7 | 1.2 | 54 | 65.8 | -0.5 |
| AGGC | 44 | 44.8 | 0.9 | 62 | 71.2 | -0.4 | 32 | 48.4 | -3.8 |
| AGCA | 123 | 122.6 | 0.6 | 132 | 120.7 | 0.0 | 148 | 124.6 | 0.0 |
| AGCA | 116 | 96.4 | 3.6 | 114 | 117.8 | -0.6 | 138 | 99.5 | 3.0 |
| AGCC | 65 | 61.9 | 0.7 | 114 | 104.6 | 1.7 | 92 | 80.7 | 1.5 |
| AGCC | 55 | 59.1 | -0.7 | 88 | 75.9 | 1.7 | 50 | 60.7 | -4.2 |
| AGGU | 137 | 143.8 | -1.3 | 128 | 119.7 | -2.3 | 73 | 68.8 | -5.5 |
| AGGU | 137 | 159.3 | -2.2 | 124 | 155.6 | 4.6 | 115 | 116.9 | -0.3 |
| ACGG | 62 | 71.6 | -2.2 | 75 | 93.9 | -3.5 | 76 | 87.3 | -2.2 |
| AUGC | 77 | 54.1 | 7.1 | 91 | 80.7 | 2.1 | 65 | 65.8 | -0.5 |

Table 5: Tetra-nucleotide bias for three RNA viruses. The tetra-nucleotide bias in three viruses. z value quantifies the tetra-nucleotide bias, as defined in equation (1). (Continued)
Approach one – Sequence Relationship of Viruses based on The Correlation of Tetra-nucleotide Bias

Two relationship trees were derived, one from the entire genome and the other from the replication ensemble (Figure 1). The result based on the replication ensemble sequence was included because these regions in RNA viruses are submitted to a strong selective pressure to ensure successful replication of their own RNA in the host cell. The two distance trees can be clustered distinctly into two major groups of viruses. Interestingly, this clustering validates our approach, since these clusters are consistent with biological properties of the viruses: Group #1 corresponds to all positive strand sRNA viruses while Group #2 corresponds to negative strand sRNA viruses. Each group must undergo different evolutionary paths which lead to their distinct pattern in tetra-nucleotide usage. The classification for the two main groups of viruses (positive/negative strand sRNA viruses) demonstrate a level of congruence with the taxonomy of the viruses [20] and indicated that there exists a relationship signal in tetra-nucleotide usage patterns.

Inside both relationship trees, Avian Encephalomyelitis Virus (AEV), Lactate Dehydrogenase-elevating Virus (LDV), Porcine Reproductive and respiratory syndrome Virus (PRV), Equine arteritis Virus (EV1), Rabbit Hemorrhagic disease Virus (RHV), Yellow Fever Virus (YFV), are the most distinct viruses in the tetra-nucleotide usage pattern. From the family of positive strand sRNA viruses, CoronaViruses form the largest cluster. The SARS-CoV is found to be at the basal position of other Coronavirus types and remains closest..
Figure 1
Two Relationship trees based on the correlation coefficients of tetra-nucleotide usage bias. The distance tree for 31 RNA viruses based on tetra-nucleotide usage pattern for the entire genome (right) and the replication enzyme (left). The correlation distances are shown on top of each branch.
to the Transmissible Gastroenteritis Virus (TGV) and Feline CoronaVirus (FCoV). This placement is consistent with the findings from two seminal papers [9,10] where the SARS-CoV was classified in a separate group from the rest of the known CoronaViruses. In addition, both distance trees suggested that the Bovine CoronaVirus (BCoV) and the Mouse Hepatitis Virus (MHV) should be grouped together whereas the Human CoronaVirus 229E (HCoV) is the closest to the Porcine epidemic Diarrhea Virus (PDV). For the family of negative strand ssRNA viruses, there are two obvious classes that have evolved through different branches of word usage pattern. The first class covers Hantaan Virus (HV1), Reston Ebola Virus (REV), Bovine Ephemeral Fever Virus (BFV), Bovine Respiratory syncytial Virus (BRV), Respiratory Syncytial Virus (RSV) and Human Respiratory syncytial Virus (HRV). The second class covers the remaining negative strand ssRNA viruses.

**Approach two – Sequence Relationship of Viruses based on The Factors of the Tetra-nucleotide Usage Pattern [21–23]**

The overall tetra-nucleotide usage pattern (additional file 2) was decomposed into several eigen-vectors using a factor analysis algorithm. They are the uncorrelated components of the original usage pattern embedded within the overall tetra-nucleotide usage pattern. Three eigen-vectors, which carry 83.3% of the variance for the viral tetra-nucleotide usage patterns, were retained (Figure 2). From the three dimensional figures (Figure 3, Figure 4, Figure 5 and Figure 6) plotted against these retained eigen-vectors, the negative strand ssRNA viruses stemmed clearly out from the positive strand ssRNA viruses. This is most obvious when the axes of projection were the 1st and 3rd eigen-vector. However in Figure 5 and Figure 6, when we project the tetra-nucleotide usage patterns (entire genome) for each virus on the 1st, 2nd and 3rd eigen-vector axes, the separation between viruses showed a different outcome when V was derived from the entire genome. The two main families of viruses were grouped into three clusters, two being allocated to the positive strand ssRNA viruses. It is particularly interesting that all viruses in the upper left corner corresponded to the viruses originating from the Corona-Virus family. Unexpectedly, the Hantaan Virus (HV1) is the only negative strand ssRNA virus to have a high loading on the eigen-vector that corresponded to the tetra-nucleotide usage pattern for the positive strand ssRNA viruses.

It is important to realize what factor analysis will provide and how this analysis is different from the previous method of relationship tree generation using correlation coefficient. For the previous method that is based on correlation coefficient of word usage patterns, it treats the vectorial profiling V for each virus as a whole entity. However, the factor analysis considers the vectorial profiling V as a superposition of many patterns which can be separated into mutually uncorrelated patterns of word usage. Each eigen-vector represents the embedded component of RNA word usage patterns communalised by a group of viruses presumably under the same selection pressures. By projecting the overall usage patterns on these eigen-vectors, it is possible to determine a group of viruses that adopt a common strategy of word usage.

**Conclusion**

Using the two approaches to study the tetra-nucleotide usage pattern in RNA viruses, we reached the following conclusions:

1. Based on the correlation of the overall tetra-nucleotide usage patterns, the Transmissible Gastroenteritis Virus (TGV) and the Feline CoronaVirus (FCoV) are closest to SARS-CoV.
2. Based on the three most significant eigen-vectors, the genomes of the viruses from the same family conform to a similar tetra-nucleotide usage pattern, irrespective of their genome size.

3. The study of word usage is a powerful method to classify RNA viruses. The congruence of the relationship trees with the known classification indicates that there exist phylogenetic signals in tetra-nucleotide usage patterns, and this signal is most prominent in the replicase open reading frames.

**Methods**

**Dataset**

We focused our study on the genomic sequences (their translated strand) of ssRNA viruses (Table 1), which incorporated 20 species from the family of positive strand ssRNA viruses and 11 species from the family of negative strand ssRNA viruses. We are aware of the fact that these viruses constitute completely different species, most probably unrelated to one another. They are included in a common study in order to try to have means to identify relevant features from purely statistical background properties. The coverage included the viruses that are known to cause diseases to their corresponding hosts. The acronym for each virus is shown in the table and is referred to throughout this study. All sequences corresponding to
their translated strand were retrieved from GenBank, and the accession numbers and genomic size (in nucleotides) for individual virus were provided for reference. For the present study, two sets of data were generated from the complete sequence for each virus. Dataset 1 covered the entire genome and dataset 2 covered only their replicase open reading frame. The flowchart for studying the tetra-nucleotide usage pattern in 31 viruses is shown in Figure 7.

**Figure 4**

2-D plots for Figure 3 with different viewpoint specifications. The tetra-nucleotide usage patterns for the replicase open reading frame in each virus have been redisplayed on the (1st vs 2nd), (1st vs 3rd) and (2nd vs 3rd) eigen-vector axes ('o' represents positive strand ssRNA virus; 'x' represents negative strand ssRNA virus). For the top figure, the order for 'o' is [15,17,12,16,8,14,9,11,13,4,7,3,10,6,2,5,1] (left to right), whereas 'x' is [24,27,25,28,22,26,18,23,20,19,21] (left to right). For the middle figure, the order for 'o' is [15,17,12,16,8,14,9,11,13,4,7,3,10,6,2,5,1] (left to right), whereas 'x' is [24,27,25,28,22,26,18,23,20,19,21] (left to right). For the bottom figure, the order for 'o' is [15,17,12,16,8,14,9,11,13,4,7,3,10,6,2,5,1] (left to right), whereas 'x' is [24,27,25,28,22,26,18,23,20,19,21] (left to right). *The corresponded virus for each number follows Figure 3.
Computer hardware and software
Sun Fire 6800 Server with 24 CPUs (each running with a clock speed of 900 MHz) was employed throughout this study. The computation of correlation coefficient and factor analysis algorithm were implemented using Matlab Technical Programming language.

Method for counting the frequency of occurrence for RNA words
It is necessary to address the question of how we counted the number of time each tetra-nucleotide (for example 'GAGA' or any other tetra-nucleotide), appeared in a given genome. For this study, we adopted the convention of not counting overlapping words [24]. Take a sequence 'UAUGAGAGAGAUCCGAGA' as example. With second or higher overlapping words not counted, the tetra-nucleotide 'GAGA' is counted as occurring only twice, namely in position 4–7 and 13–16. Positions 6–9 are omitted because they overlap with 'GAGA' at position 4–7.

However, when we counted tetra-nucleotide 'UGAG', position 3–6 would also be registered as position 4–6 already recorded when counting tetra-nucleotide 'GAGA'. In short, all frequency counting of tetra-nucleotide were started anew when we changed from counting the frequency of one tetra-nucleotide to another; this was to preserve the correlation of tetra-nucleotides which have overlapping subword (e.g. 'UAGA' and 'GACA'). A table showing the frequencies of tetra-nucleotides is shown in the additional file 2.

Vectorial profiling (V) of the viral RNA genome word usage pattern
The nucleotide composition has being suggested to be a specific characteristic in different virus phylogeny [25].
Because most viral genomes are short, and because we lack a prior information on the tempo and modes of evolution of RNA viruses, we proceeded as follows. We created a vector, \( V = [C_1, C_2, ..., C_n] \), with each element representing the frequency for a specific RNA word of length \( n \). The number of components \( k \) in \( V \) increases exponentially with word size \( n \) - \( k = 4^n \). In order to use \( V \) for discrimination between viruses, two criteria must be met. First, \( V \) must contain sufficient components (di-

nucleotide \( k = 16 \); tri-nucleotide \( k = 64 \); tetra-nucleotide \( k = 256 \)); second, the frequencies for tetra-nucleotides must show a prominent bias (over/under-representation) that is unique for a family of viruses.

For the first criteria, there are pros and cons for choosing either longer or shorter words. When the shorter words are used, they inherit the problem of inadequate representation of the viral genome because the long motifs will be

---

**Figure 6**

2-D plots for Figure 5 with different viewpoint specifications. The tetra-nucleotide usage patterns table in the additional file 2 (entire genome) for each virus have been redisplayed on the \((1^{st} \text{ vs } 2^{nd}), (1^{st} \text{ vs } 3^{rd})\) and \((2^{nd} \text{ vs } 3^{rd})\) eigen-vector axes (‘o’ represents positive strand ssRNA virus, ‘x’ represents negative strand ssRNA virus). For the top figure, the order for ‘o’ is [3,7,1,4,2,5,6,17,13,20,10,16,9,8,11,15,12,14,18,19] (left to right), whereas ‘x’ is [26,22,30,23,24,28,31,27,25,29,21] (left to right). For the middle figure, the order for ‘o’ is [3,7,1,4,2,5,6,17,13,20,10,16,9,8,11,15,12,14,18,19] (left to right), whereas ‘x’ is [26,22,30,23,24,28,31,27,25,29,21] (left to right). For the bottom figure, the order for ‘o’ is [3,7,1,4,2,5,6,17,13,20,10,16,9,8,11,15,12,14,18,19] (left to right), whereas ‘x’ is [26,22,30,23,24,28,31,27,25,29,21] (left to right). *The corresponded virus for each number follows Figure 5.
neglected. But the shorter words have an advantage of saving computational time. On the other hand, when the longer words are used, they cause a problem of computer tractability due to a larger word set to explore \((k = 4^n)\). However, the larger words have an advantage of accounting for the correlation of their sub-words. In contrast, the number of their occurrences falls down rapidly, preventing accurate statistical analysis. We chose tetra-nucleotides for our study because they provide 256 vector components (additional file 2) and account for correlation of sub-words up to the order three.

For the second criteria, the bias in RNA word usage was examined. The bias in word usage (of size \(n\)) is influenced by the bias of word with sizes less than \(n\) [26]. Therefore, in order to evaluate the true bias of word size \(m\), it is required to compare the frequencies of word usage in the original sequence to that of model chromosomes that take into account the biases of word size \(m-1\), \(m-2\) ... \(1\). These model chromosomes were generated by obeying the Markov model of the order \((m-1)\)th. This can be achieved by shuffling \(m-1\) viral nucleotides as one whole unit so that the nucleotide successions up to order \((m-1)\)th were being preserved. Several statistical approaches have been proposed for quantifying word biases [27,28]. In this study, we employed the \(z\) statistics (Equation 1) for di-nucleotide and tetra-nucleotide biases [27,28]. The \(z\) value is a measure of the bias of a word, with values close to zero meaning no bias, negative values meaning under-representation and positive values meaning over-representation of the word \(w\) in the RNA text.

\[
z_w = \frac{N(w) - E(w)}{\sqrt{Var(w)}}
\]  

(1)

where \(w\) is a word of size \(m\); \(N(w)\) is observed count in actual viral RNA; \(E(w)\) and \(Var(w)\) are expected count and variance for \(w\) derived from the 100 artificial chromosomes that preserved the nucleotide succession up to order \(m-1\).
**Approach one – sequence relationship of viruses based on the correlation of tetra-nucleotide bias**

A scale-invariant parameter, the correlation coefficient $r$, was employed to compare between word usage patterns of viruses. The correlation coefficient $r$ measures the degree of linear relationship between two vectors. Here, the two vectors are the tetra-nucleotide word usage pattern $V$ corresponding to each viral genome. The magnitude of $r$ would indicate how much of the change of pattern in the tetra-nucleotide word usage in one virus is explained by the change in another. The magnitude of $r$ is always between -1 and +1 and the relationship between the two variables will approach perfect linearity as the magnitude of correlation coefficient approaches to extreme values (+/-1). However, perfect positive correlation ($r = 1$) does not mean identity of the paired $V_i$, but, rather, identity up to positive linearity, that is, identity between the paired standardized values. This is a crucial property of $r$ (scale-invariant) that enables the comparison of viral genome despite their differences in genomic sizes. Positive magnitude of $r$ indicates positive association whereas negative magnitude of $r$ indicates negative association between two usage patterns. For this study, correlation coefficient, $r$, for let say virus 1 and virus 2, is defined as follow:

$$r_{12} = \frac{\sum_{i=1}^{256} (V_{1i} - \bar{V}_1)(V_{2i} - \bar{V}_2)}{(n-1)S_{V1}S_{V2}}$$  \hspace{1cm} (2)

where $V_{1i}$, $V_{2i}$ are vector representing the tetra-nucleotide usage pattern; $S_{V1}$ and $S_{V2}$ standard deviation of $V_{1i}$, $V_{2i}$; $\bar{V}_1$, $\bar{V}_2$ are the mean of $V_{1i}$, $V_{2i}$.

Then, the distance between the tetra-nucleotide usage patterns of two viruses is defined as follows:

Distance $D_{ij} = 1 - r_{ij}$ \hspace{1cm} (3)

where $D_{ij}$ is the distance between the tetra-nucleotide usage patterns of virus i and virus j; $r_{ij}$ is the correlation coefficient between the tetra-nucleotide usage patterns of virus i and virus j.

Prior to the construction of a relationship tree, the pairwise distance matrix $M$ of size 31 by 31 was constructed (see additional file 3). Pair-wise distance between two viral genomes is measured by the value of $(1 - r)$. Each row/column corresponds to a specific virus and an entry at the intersection of row X and column Y corresponds to the distance between virus X and virus Y. Such matrix has a diagonal entry of value 0. For the purpose of constructing a relationship tree, only the lower/upper triangular matrix of $M$ is required. After obtaining lower/upper triangular matrix of $M$, the neighbor-joining method (NJ) algorithm was used to construct the relationship tree (Figure 1). The neighbor-joining method is based on minimum-distance principle. Details of the NJ algorithm are available in [29].

**Approach two – sequence relationship of viruses based on the factors of the tetra-nucleotide usage pattern**

The factor analysis is a statistical method that reveals simpler patterns within a complex set of tetra-nucleotide usage patterns $V$ (additional file 2). It seeks to discover if the observed usage patterns can be explained in terms of a much smaller number of un-correlated pattern sets called factors (eigen-vectors). Suppose we take a simple case where there are 31 viruses each represented by two components $(x,y)$ in vector $V$ ($x,y$ represent the frequencies of occurrence for two specific tetra-nucleotides). Then, in a scatter-plot we can think of the regression line as the original X-axis, rotated so that it approximates the regression line. This type of rotation maximize the variance of the variables $(x,y)$ on the eigen-vector. The remaining variability around this the first eigen-vector was captured in the subsequent eigen-vectors. In this manner, consecutive eigen-vectors are extracted but with a diminishing importance. What each eigen-vector represents is the embedded RNA word usage patterns communalised by a group of viruses presumably under the same selection pressures.

We implemented the factor analysis algorithm [21–23] in Matlab Technical Programming Language and computed a set of eigen-vectors. Then, the original usage pattern $V$ was re-mapped for each virus onto the new coordinate system based on these derived eigen-vectors. The difference between approach two and approach one is discussed in the results and discussion section.

**Authors’ contributions**

YLY participated in the design and performed the statistical analysis.

AD participated in the design and overall coordination of this study.

XWZ participated in the design of the study.

All authors read and approved the final manuscript.
Additional material

Additional File 1
The RNA word biases of different sizes in RNA viruses. These tables show the di-nucleotide, tetra-nucleotide and penta-nucleotide biases for 31 RNA viruses. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-43-S1.xls]

Additional File 2
Vectorial profiling of tetra-nucleotide usage pattern in seven RNA viruses. The tetra-nucleotide frequencies of occurrence in seven viral genomes. Each column represents a tetra-nucleotide usage pattern V{ij} for a single virus. We derived correlation coefficient (r) by comparing any two columns simultaneously. This parameter r indicates the likeness of word usage patterns in any two viruses. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-43-S2.xls]

Additional File 3
The distance matrices. Each entry in matrix M is computed using Equation 3. The correlation coefficient (r) in equation 3 is obtained by comparing any two columns in the tetra-nucleotide usage patterns table in the additional file 2 simultaneously. Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-43-S3.xls]

Acknowledgements
Indispensable support was provided by the doctoral fellowship from The University of Hong Kong (HKU). We wish to thank the Hong Kong Innovation and Technology Fund for supporting work upstream of the present study, that made it possible at a time when the unexpected SARS outbreak reached Hong Kong. Finally, we wish to thank Dr Ralf Altmeyer for his critical interest for this work as he came at the head of the HKU-Pasteur Research Centre.

References
1. Cumulative Number of Reported Probable Cases of SARS [http://www.who.int/csr/sarscountry/en/]
2. Fouchier RA, Kuiken T, Schuurman R, Van Amerongen G, Van Boeijen GJ, Van Den Hoogen BG, Peiris M, Rim W, Stohr K and Osterhaus AD: Aetiology: Koch’s postulates fulfilled for SARS virus. Nature 2003, 423(6937):240.
3. Hoey J and Maskalik J: SARS update. CMAJ 2003, 168(10):1294-5.
4. James FJ: SARS Web information. AIDS Treat News 2003:6.
5. Situation Updates – SARS [http://www.who.int/csr/sars/archive/en/]
6. Drosten C, Gunther S, Preiser W, Van Der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, Berger A, Burghuer J, Cmatal J, Eisenmann M, Escriou N, Grywna K, Kramme S, Manuguerra JC, Muller S, Rickerts V, Sturmer M, Vieth S, Kliek HD, Osterhau AD, Schmitz H and Doerr HW: Identification of a Novel CoronaVirus in Patients with Severe Acute Respiratory Syndrome. N Engl J Med 2003, 348(20):1967-76.
7. Van Vugt J, Storgaard T, Oleksiewicz MB and Botner A: High frequency RNA recombination in porcine reproductive and respiratory syndrome virus occurs preferentially between parental sequences with high similarity. J Gen Virol 2001, 82(Pt 11):2615-20.
8. Lerner DL, Wagaman PC, Phillips TR, Prospero-Garcia O, Henricksen SJ, Fox HS, Bloom FE and Elder JH: Increased mutation frequency of feline immunodeficiency virus lacking functional deoxyuridine mono-phosphatase. Proc Natl Acad Sci USA 1992(12):7480-4. 1995 Aug 1
9. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattri J, Asano JK, Barber SA, Chan ST, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein E, Siddiqui A, Simatus AL, Scott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, BerNark K, Booth TF, Bowness D, Drebog M, FeRNando L, Fick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C and Roper RL: The Genome Sequence of the SARS-Associated Coronavirus. Science 2003, 300(5624):1399-404.
10. Oberste MS, Monroe SS, Nix WA, Campagnoli R, Incenngle JP, Panaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Kiaskie T, Rollin PE, Sanchez A, Siffick S, Holloway B, Limor J, McCausland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhau AD, Drolet B, Anderson LJ and Bellini WJ: Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. Science 2003, 300(5624):1394-9.
11. Kaissie T, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Conner JA, Lim W, Rollin PE, Dowell SF, Ling AE, Humphrey CD, Shieh WJ, Guerner J, Paddock CD, Rota P, Fields B, DeRisi J, Yang JY, Cox N, Hughes JM, LeDuc JW, Bellini WJ and Anderson LJ: A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. N Engl J Med 2003, 348(20):1953-66.
12. Hacker J and Carniel E: Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep 2001, 2(5):376-81.
13. van Hemert FJ and Berk hout B: The tendency of tentorial open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 1995, 41(2):132-40.
14. Hubacek J: Biological function of DNA methylation. Foka Microbiol (Pragho) 1993, 37(3):323-9.
15. Karlin S, Doerfler W and Cardon L: Why is CPG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 1994, 68(5):2889-97.
16. Gantz KR, Stromberg K and Montes de Oca F: Specific RNA methylation associated with avian myeloblastosis virus. Nature 1971, 234:35-37.
17. Frederico LA, Kunkel TA and Shaw BR: A sensitive genetic assay for the detection of cytosome deamination: determination of rate constants and the activation energy. Biochemistry 2001, 29(10):2532-7.
18. Bibillo A, Figierowicz M, Ziomek K and Kierzek R: The nonenzymatic hydrosis of oligoribonucleotides. VII. Structural elements affecting hydrosis. Nucleosides Nucleotides Nucleic Acids 2000, 19(5-6):977-94.
19. Beutler E, G UPDATED 2019, 19(5-6):977-94.
20. Regenmortel MHV, van Fauquet CM, Bishop DHL, Carstens EB, Estes MK, Lemon SM, Manliof J, Maynardigt J, Me A, Geoch DJ, Pringle CR and Wickner RB: Virus Taxonomy: Classification and Nomenclature of Viruses. Seventh Report of the International Committee on Taxonomy of Viruses Academic Press, San Diego; 2000.
21. Bartholomew DJ: Factor Analysis for Categorical Data, Journal of the Royal Statistical Society. Series B (Methodological). 1980, 42(3):289-321.
22. Kim J and Mueller Charles W: Introduction to factor analysis: What it is and how to do it Newbury Park, CA: Sage Publications; 1978.
23. Bartholomew DJ: Factor Analysis for Categorical Data, Journal of the Royal Statistical Society. Series B (Methodological). 1980, 42(3):289-321.
24. Ewens WJ and Grant G:R.: Statistical Methods in Bioinformatics Springer-Verlag New York, Inc, New York; 2001.
25. Bronson EC and Anderson JN: Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol 1994, 38(5):506-32.
26. Rocha EP, Viari A and Danchin A: Oligo-nucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. Nucleic Acids Res 1998, 26(12):2971-80.
27. Leung MY, Marsh GM and Speed TP: Over- and underrepresentation of short DNA words in herpesvirus genomes. J Comput Biol 1996, 3(3):345-60.
28. Schbath S, Prum B and de Turckheim E: Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J Comput Biol 1995, 2(3):417-37.
29. Saitou N and Nei M: The neighbor-joining method: A new method for reconstructing trees. Mol Biol and Evol 1987, 4:406-425.