# Assemblies of putative SARS-CoV2-spike-encoding mRNA sequences for vaccines BNT-162b2 and mRNA-1273.

### (version 0.1.1Beta 03/23/21)

Dae Eun Jeong, Matthew McCoy, Karen Artiles, OrkanIlbay, Andrew Fire*, Kari Nadeau, Helen Park, Brooke Betts, Scott Boyd, Ramona Hoh, and Massa Shoura*

Departments of Pathology, Genetics, Pediatrics, and Medicine, Stanford University School of Medicine and Veterans Affairs Palo Alto Medical Center \*Correspondence: <afire@stanford.edu> and/or <massa86@stanford.edu>

RNA vaccines have become a key tool in moving forward through the challenges raised both in the current pandemic and in numerous other publichealth and medical challenges. With the rollout of vaccines for COVID-19, these synthetic mRNAs have become broadly distributed RNA species in numerous human populations. Despite their ubiquity, sequences are not always available for such RNAs. Standard methods facilitate such sequencing. In this note, we provide experimental sequence informationfor the RNA components of the initial [Moderna](https://pubmed.ncbi.nlm.nih.gov/32756549/)and [Pfizer/BioNTech](https://pubmed.ncbi.nlm.nih.gov/33301246/) COVID-19 vaccines, allowing a working assembly of the former and a confirmation of previously reported sequence information for the latter RNA.

Sharing of sequence information for broadly used therapeutics has the benefit of allowing any researchers or clinicians using sequencing approaches to rapidly identify such sequences as therapeutic-derived rather than host or infectious in origin.

For this work, RNAs were obtained as discards from the small portions of vaccine doses that remained in vials after immunization; such portions would have been required to be otherwise discarded and were analyzed under FDA authorization for research use. To obtain the small amounts of RNA needed for characterization,vaccine remnants were phenol-chloroform extracted using TRIzol Reagent (Invitrogen), with intactness assessed by Agilent 2100 Bioanalyzer before and after extraction.

Although our analysis mainly focused on RNAs obtained as soon as possible following discard, we also analyzed samples which had been refrigerated (~4 °C) for up to 42 days with and without the addition of EDTA. Interestingly a substantial fraction of the RNA remained intact in these preparations. We note that the formulation of the vaccines includes numerous key chemical components which are quite possibly unstable under these conditions– so these data certainly do not suggest that the vaccine as a biological agent is stable. But it is of interest that chemical stability of RNA itself is not sufficient to preclude eventual development of vaccines with a much less involved cold-chain storage and transportation.

For further analysis, the initial RNAs were fragmentedby heating to 94°C, primed with a random hexamer-tailed adaptor, amplified through atemplate-switch protocol (Takara SMARTerer Stranded RNA-seq kit), and sequenced usinga MiSeq instrument (Illumina) with paired end 78-per end sequencing. As a reference material in specific assays, we included RNA of known concentration and sequence (from bacteriophage MS2).

From these data, we obtained partial information on strandedness and a set of segments that could be used for assembly. This was particularly useful for the Moderna vaccine, for which the original vaccine RNA sequence was not available at the time our study was carried out. Contigs encoding full-length spikes were assembled from the Moderna and Pfizer datasets. The Pfizer/BioNTech data [Figure 1] verified the [reported sequence for that vaccine](https://berthub.eu/articles/posts/reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/), while the Moderna sequence [Figure 2] could not be checked against a published reference.

RNA preparations lacking dsRNA are desirable in generating vaccine formulations as these will minimize an otherwise dramatic biological (and non-specific) response that vertebrates have to double stranded character in RNA (https://www.nature.com/articles/nrd.2017.243). In the sequence data that we analyzed, we found that the vast majority of reads were from the expected sense strand. In addition, the minorityof antisense reads appeared different from sense reads in lacking the characteristic extensions expected from the template switching protocol. Examining only the reads with an evident template switch (as an indicator for strand-of-origin), we observed that both vaccines overwhelmingly yielded sense reads (>99.99%). Independent sequencing assays and otherexperimental measurements are ongoing and will be needed to determine whether this template-switched sense read fraction in the SmarterSeq protocol indeed represents the actual dsRNA content in the original material.

This work provides an initial assessment of two RNAs that are now a part of the human ecosystem and that are likely to appear in numerous other high throughput RNA-seq studies in which a fraction of the individuals may have previously been vaccinated.

ProtoAcknowledgements: Thanks to our colleagues for help and suggestions (Nimit Jain, Emily Greenwald, Lamia Wahba, William Wang, Amisha Kumar, Sameer Sundrani, David Lipman, Marc Salit, Bijoyita Roy).

### Figure 1: Spike-encoding contig assembled from BioNTech/PfizerBNT-162b vaccine.

Although the full coding region is included,the nature of the methodology used for sequencing and assembly is such that the assembled contig could lack some sequence from the ends of the RNA. Within the assembled sequence, this hypothetical sequence shows a perfect match to the corresponding sequence from documents available online derived from manufacturer communications with the World Health Organization [as reported by https://berthub.eu/articles/posts/reverse-engineering-source-code-of-the-biontech-pfizer-vaccine/]. The 5’ end for the assembly matches the start site noted in these documents, while the read-based assembly lacks an interrupted polyA tail (A30(GCATATGACT)A70) that is expected to be present in the mRNA.

### Figure 2: Spike-encoding contig assembled from Moderna mRNA-1273 vaccine.\*\*

This is a partial sequence of the vaccine RNA. Although the full coding region is included, the assembled contig could lack some sequence from the ends of the RNA.

**Figure 1: Spike-encoding contig assembled from BioNTech/PfizerBNT-162b vaccine.**

GAGAATAAACTAGTATTCTTCTGGTCCCCACAGACTCAGAGAGAACCCGCCACCATGTTCGTGTTCCTGGTGCTGCTGCCTCTGGTGTCCA GCCAGTGTGTGAACCTGACCACCAGAACACAGCTGCCTCCAGCCTACACCAACAGCTTTACCAGAGGCGTGTACTACCCCGACAAGGTGTT CAGATCCAGCGTGCTGCACTCTACCCAGGACCTGTTCCTGCCTTTCTTCAGCAACGTGACCTGGTTCCACGCCATCCACGTGTCCGGCACC AATGGCACCAAGAGATTCGACAACCCCGTGCTGCCCTTCAACGACGGGGTGTACTTTGCCAGCACCGAGAAGTCCAACATCATCAGAGGCT GGATCTTCGGCACCACACTGGACAGCAAGACCCAGAGCCTGCTGATCGTGAACAACGCCACCAACGTGGTCATCAAAGTGTGCGAGTTCCA GTTCTGCAACGACCCCTTCCTGGGCGTCTACTACCACAAGAACAACAAGAGCTGGATGGAAAGCGAGTTCCGGGTGTACAGCAGCGCCAAC AACTGCACCTTCGAGTACGTGTCCCAGCCTTTCCTGATGGACCTGGAAGGCAAGCAGGGCAACTTCAAGAACCTGCGCGAGTTCGTGTTTA AGAACATCGACGGCTACTTCAAGATCTACAGCAAGCACACCCCTATCAACCTCGTGCGGGATCTGCCTCAGGGCTTCTCTGCTCTGGAACC CCTGGTGGATCTGCCCATCGGCATCAACATCACCCGGTTTCAGACACTGCTGGCCCTGCACAGAAGCTACCTGACACCTGGCGATAGCAGC AGCGGATGGACAGCTGGTGCCGCCGCTTACTATGTGGGCTACCTGCAGCCTAGAACCTTCCTGCTGAAGTACAACGAGAACGGCACCATCA CCGACGCCGTGGATTGTGCTCTGGATCCTCTGAGCGAGACAAAGTGCACCCTGAAGTCCTTCACCGTGGAAAAGGGCATCTACCAGACCAG CAACTTCCGGGTGCAGCCCACCGAATCCATCGTGCGGTTCCCCAATATCACCAATCTGTGCCCCTTCGGCGAGGTGTTCAATGCCACCAGA TTCGCCTCTGTGTACGCCTGGAACCGGAAGCGGATCAGCAATTGCGTGGCCGACTACTCCGTGCTGTACAACTCCGCCAGCTTCAGCACCT TCAAGTGCTACGGCGTGTCCCCTACCAAGCTGAACGACCTGTGCTTCACAAACGTGTACGCCGACAGCTTCGTGATCCGGGGAGATGAAGT GCGGCAGATTGCCCCTGGACAGACAGGCAAGATCGCCGACTACAACTACAAGCTGCCCGACGACTTCACCGGCTGTGTGATTGCCTGGAAC AGCAACAACCTGGACTCCAAAGTCGGCGGCAACTACAATTACCTGTACCGGCTGTTCCGGAAGTCCAATCTGAAGCCCTTCGAGCGGGACA TCTCCACCGAGATCTATCAGGCCGGCAGCACCCCTTGTAACGGCGTGGAAGGCTTCAACTGCTACTTCCCACTGCAGTCCTACGGCTTTCA GCCCACAAATGGCGTGGGCTATCAGCCCTACAGAGTGGTGGTGCTGAGCTTCGAACTGCTGCATGCCCCTGCCACAGTGTGCGGCCCTAAG AAAAGCACCAATCTCGTGAAGAACAAATGCGTGAACTTCAACTTCAACGGCCTGACCGGCACCGGCGTGCTGACAGAGAGCAACAAGAAGT TCCTGCCATTCCAGCAGTTTGGCCGGGATATCGCCGATACCACAGACGCCGTTAGAGATCCCCAGACACTGGAAATCCTGGACATCACCCC TTGCAGCTTCGGCGGAGTGTCTGTGATCACCCCTGGCACCAACACCAGCAATCAGGTGGCAGTGCTGTACCAGGACGTGAACTGTACCGAA GTGCCCGTGGCCATTCACGCCGATCAGCTGACACCTACATGGCGGGTGTACTCCACCGGCAGCAATGTGTTTCAGACCAGAGCCGGCTGTC TGATCGGAGCCGAGCACGTGAACAATAGCTACGAGTGCGACATCCCCATCGGCGCTGGAATCTGCGCCAGCTACCAGACACAGACAAACAG CCCTCGGAGAGCCAGAAGCGTGGCCAGCCAGAGCATCATTGCCTACACAATGTCTCTGGGCGCCGAGAACAGCGTGGCCTACTCCAACAAC TCTATCGCTATCCCCACCAACTTCACCATCAGCGTGACCACAGAGATCCTGCCTGTGTCCATGACCAAGACCAGCGTGGACTGCACCATGT ACATCTGCGGCGATTCCACCGAGTGCTCCAACCTGCTGCTGCAGTACGGCAGCTTCTGCACCCAGCTGAATAGAGCCCTGACAGGGATCGC CGTGGAACAGGACAAGAACACCCAAGAGGTGTTCGCCCAAGTGAAGCAGATCTACAAGACCCCTCCTATCAAGGACTTCGGCGGCTTCAAT TTCAGCCAGATTCTGCCCGATCCTAGCAAGCCCAGCAAGCGGAGCTTCATCGAGGACCTGCTGTTCAACAAAGTGACACTGGCCGACGCCG GCTTCATCAAGCAGTATGGCGATTGTCTGGGCGACATTGCCGCCAGGGATCTGATTTGCGCCCAGAAGTTTAACGGACTGACAGTGCTGCC TCCTCTGCTGACCGATGAGATGATCGCCCAGTACACATCTGCCCTGCTGGCCGGCACAATCACAAGCGGCTGGACATTTGGAGCAGGCGCC GCTCTGCAGATCCCCTTTGCTATGCAGATGGCCTACCGGTTCAACGGCATCGGAGTGACCCAGAATGTGCTGTACGAGAACCAGAAGCTGA TCGCCAACCAGTTCAACAGCGCCATCGGCAAGATCCAGGACAGCCTGAGCAGCACAGCAAGCGCCCTGGGAAAGCTGCAGGACGTGGTCAA CCAGAATGCCCAGGCACTGAACACCCTGGTCAAGCAGCTGTCCTCCAACTTCGGCGCCATCAGCTCTGTGCTGAACGATATCCTGAGCAGA CTGGACCCTCCTGAGGCCGAGGTGCAGATCGACAGACTGATCACAGGCAGACTGCAGAGCCTCCAGACATACGTGACCCAGCAGCTGATCA GAGCCGCCGAGATTAGAGCCTCTGCCAATCTGGCCGCCACCAAGATGTCTGAGTGTGTGCTGGGCCAGAGCAAGAGAGTGGACTTTTGCGG CAAGGGCTACCACCTGATGACTCGAGCTGGTACTGCATGCACGCAATGCTAGCTGCCCCTTTCCCGTCC TGGGTACCCCGAGTCTCCCCCGAC CTCGGGTCCCAGGTATGCTCCCACCTCCACCTGCCCCACTCACCACCTCTGCTAGTTCCAGACACCTCCCAAGCACGCAGCAATGCAGCTCAA AACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAAACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACACCCTGGAGCTAGCA

Cyan: Putative 5’ UTR

Green: Start Codon

Yellow: Signal Peptide

Orange: Spike encoding region

Red: Stop codon(s)

Purple: 3’ UTR

Blue: Start of polyA region (incomplete)

**Figure 2: Spike-encoding contig assembled from Moderna mRNA-1273 vaccine.**

GGGAAATAAGAGAGAAAAGAAGAGTAAGAAGAAATATAAGACCCCGGCGCCGCCACCATGTTCGTGTTCCTGGTGCTGCTGCCCCTGGTGA GCAGCCAGTGCGTGAACCTGACCACCCGGACCCAGCTGCCACCAGCCTACACCAACAGCTTCACCCGGGGCGTCTACTACCCCGACAAGGT GTTCCGGAGCAGCGTCCTGCACAGCACCCAGGACCTGTTCCTGCCCTTCTTCAGCAACGTGACCTGGTTCCACGCCATCCACGTGAGCGGC ACCAACGGCACCAAGCGGTTCGACAACCCCGTGCTGCCCTTCAACGACGGCGTGTACTTCGCCAGCACCGAGAAGAGCAACATCATCCGGG GCTGGATCTTCGGCACCACCCTGGACAGCAAGACCCAGAGCCTGCTGATCGTGAATAACGCCACCAACGTGGTGATCAAGGTGTGCGAGTT CCAGTTCTGCAACGACCCCTTCCTGGGCGTGTACTACCACAAGAACAACAAGAGCTGGATGGAGAGCGAGTTCCGGGTGTACAGCAGCGCC AACAACTGCACCTTCGAGTACGTGAGCCAGCCCTTCCTGATGGACCTGGAGGGCAAGCAGGGCAACTTCAAGAACCTGCGGGAGTTCGTGT TCAAGAACATCGACGGCTACTTCAAGATCTACAGCAAGCACACCCCAATCAACCTGGTGCGGGATCTGCCCCAGGGCTTCTCAGCCCTGGA GCCCCTGGTGGACCTGCCCATCGGCATCAACATCACCCGGTTCCAGACCCTGCTGGCCCTGCACCGGAGCTACCTGACCCCAGGCGACAGC AGCAGCGGGTGGACAGCAGGCGCGGCTGCTTACTACGTGGGCTACCTGCAGCCCCGGACCTTCCTGCTGAAGTACAACGAGAACGGCACCA TCACCGACGCCGTGGACTGCGCCCTGGACCCTCTGAGCGAGACCAAGTGCACCCTGAAGAGCTTCACCGTGGAGAAGGGCATCTACCAGAC CAGCAACTTCCGGGTGCAGCCCACCGAGAGCATCGTGCGGTTCCCCAACATCACCAACCTGTGCCCCTTCGGCGAGGTGTTCAACGCCACC CGGTTCGCCAGCGTGTACGCCTGGAACCGGAAGCGGATCAGCAACTGCGTGGCCGACTACAGCGTGCTGTACAACAGCGCCAGCTTCAGCA CCTTCAAGTGCTACGGCGTGAGCCCCACCAAGCTGAACGACCTGTGCTTCACCAACGTGTACGCCGACAGCTTCGTGATCCGTGGCGACGA GGTGCGGCAGATCGCACCCGGCCAGACAGGCAAGATCGCCGACTACAACTACAAGCTGCCCGACGACTTCACCGGCTGCGTGATCGCCTGG AACAGCAACAACCTCGACAGCAAGGTGGGCGGCAACTACAACTACCTGTACCGGCTGTTCCGGAAGAGCAACCTGAAGCCCTTCGAGCGGG ACATCAGCACCGAGATCTACCAAGCCGGCTCCACCCCTTGCAACGGCGTGGAGGGCTTCAACTGCTACTTCCCTCTGCAGAGCTACGGCTT CCAGCCCACCAACGGCGTGGGCTACCAGCCCTACCGGGTGGTGGTGCTGAGCTTCGAGCTGCTGCACGCCCCAGCCACCGTGTGTGGCCCC AAGAAGAGCACCAACCTGGTGAAGAACAAGTGCGTGAACTTCAACTTCAACGGCCTTACCGGCACCGGCGTGCTGACCGAGAGCAACAAGA AATTCCTGCCCTTTCAGCAGTTCGGCCGGGACATCGCCGACACCACCGACGCTGTGCGGGATCCCCAGACCCTGGAGATCCTGGACATCAC CCCTTGCAGCTTCGGCGGCGTGAGCGTGATCACCCCAGGCACCAACACCAGCAACCAGGTGGCCGTGCTGTACCAGGACGTGAACTGCACC GAGGTGCCCGTGGCCATCCACGCCGACCAGCTGACACCCACCTGGCGGGTCTACAGCACCGGCAGCAACGTGTTCCAGACCCGGGCCGGTT GCCTGATCGGCGCCGAGCACGTGAACAACAGCTACGAGTGCGACATCCCCATCGGCGCCGGCATCTGTGCCAGCTACCAGACCCAGACCAA TTCACCCCGGAGGGCAAGGAGCGTGGCCAGCCAGAGCATCATCGCCTACACCATGAGCCTGGGCGCCGAGAACAGCGTGGCCTACAGCAAC AACAGCATCGCCATCCCCACCAACTTCACCATCAGCGTGACCACCGAGATTCTGCCCGTGAGCATGACCAAGACCAGCGTGGACTGCACCA TGTACATCTGCGGCGACAGCACCGAGTGCAGCAACCTGCTGCTGCAGTACGGCAGCTTCTGCACCCAGCTGAACCGGGCCCTGACCGGCAT CGCCGTGGAGCAGGACAAGAACACCCAGGAGGTGTTCGCCCAGGTGAAGCAGATCTACAAGACCCCTCCCATCAAGGACTTCGGCGGCTTC AACTTCAGCCAGATCCTGCCCGACCCCAGCAAGCCCAGCAAGCGGAGCTTCATCGAGGACCTGCTGTTCAACAAGGTGACCCTAGCCGACG CCGGCTTCATCAAGCAGTACGGCGACTGCCTCGGCGACATAGCCGCCCGGGACCTGATCTGCGCCCAGAAGTTCAACGGCCTGACCGTGCT GCCTCCCCTGCTGACCGACGAGATGATCGCCCAGTACACCAGCGCCCTGTTAGCCGGAACCATCACCAGCGGCTGGACTTTCGGCGCTGGA GCCGCTCTGCAGATCCCCTTCGCCATGCAGATGGCCTACCGGTTCAACGGCATCGGCGTGACCCAGAACGTGCTGTACGAGAACCAGAAGC TGATCGCCAACCAGTTCAACAGCGCCATCGGCAAGATCCAGGACAGCCTGAGCAGCACCGCTAGCGCCCTGGGCAAGCTGCAGGACGTGGT GAACCAGAACGCCCAGGCCCTGAACACCCTGGTGAAGCAGCTGAGCAGCAACTTCGGCGCCATCAGCAGCGTGCTGAACGACATCCTGAGC CGGCTGGACCCTCCCGAGGCCGAGGTGCAGATCGACCGGCTGATCACTGGCCGGCTGCAGAGCCTGCAGACCTACGTGACCCAGCAGCTGA TCCGGGCCGCCGAGATTCGGGCCAGCGCCAACCTGGCCGCCACCAAGATGAGCGAGTGCGTGCTGGGCCAGAGCAAGCGGGTGGACTTCTG CGGCAAGGGCTACCACCTGATGAGCTTTCCCCAGAGCGCACCCCACGGAGTGGTGTTCCTGCACGTGACCTACGTGCCCGCCCAGGAGAAG AACTTCACCACCGCCCCAGCCATCTGCCACGACGGCAAGGCCCACTTTCCCCGGGAGGGCGTGTTCGTGAGCAACGGCACCCACTGGTTCG TGACCCAGCGGAACTTCTACGAGCCCCAGATCATCACCACCGACAACACCTTCGTGAGCGGCAACTGCGACGTGGTGATCGGCATCGTGAA CAACACCGTGTACGATCCCCTGCAGCCCGAGCTGGACAGCTTCAAGGAGGAGCTGGACAAGTACTTCAAGAATCACACCAGCCCCGACGTG GACCTGGGCGACATCAGCGGCATCAACGCCAGCGTGGTGAACATCCAGAAGGAGATCGATCGGCTGAACGAGGTGGCCAAGAACCTGAACG AGAGCCTGATCGACCTGCAGGAGCTGGGCAAGTACGAGCAGTACATCAAGTGGCCCTGGTACATCTGGCTGGGCTTCATCGCCGGCCTGAT CGCCATCGTGATGGTGACCATCATGCTGTGCTGCATGACCAGCTGCTGCAGCTGCCTGAAGGGCTGTTGCAGCTGCGGCAGCTGCTGCAAG TTCGACGAGGACGACAGCGAGCCCGTGCTGAAGGGCGTGAAGCTGCACTACACCTGATAATAGGCTGGAGCCTCGGTGGCCTAGCTTCTTGCCC CTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCAAAAAAAAA

Cyan: Putative 5’ UTR

Green: Start Codon

Yellow: Signal Peptide

Orange: Spike encoding region

Red: Stop codon(s)

Purple: 3’ UTR

Blue: Start of polyA region (incomplete)