Calculation of B.1.1.7 (North American) substitutions

The constellation of substitutions carried by the B.1.1.7 lineage in comparison with Wuhan-Hu-1 were computed for isolates from North America using sequence records derived from Genbank so that the sequences can be released through ViPR and BV-BRC and so that they are relevant to circulation in the U.S, as follows:

1. Generate an S protein working set to use as a BLAST/Short Peptide Search database using query: SARS-related species, complete genome, S protein, sinceJUN2020, human, North America on 25JAN2021.

2. Use Short Peptide Search for GFQPTYGVGYQ to identify S proteins from North America since July 2020 carrying the N501Y substitution => 11 proteins

3. For quality control purposes, removed strains in which S protein sequences had any "X" ambiguous positions => 8 proteins from the following strains:

    a. SARS-CoV-2/human/USA/SEARCH-5574-SAN/2020,
    SARS-CoV-2/human/USA/CO-CDPHE-2100156850/2020,
    SARS-CoV-2/human/USA/CA-CDC-STM-P017/2020,
    SARS-CoV-2/human/USA/FL-BPHL-2779/2021,
    SARS-CoV-2/human/USA/FL-CDC-STM-P012/2020,
    SARS-CoV-2/human/USA/NY-Wadsworth-291673-01/2020,
    SARS-CoV-2/human/USA/NMDOH-2021003333/2021,
    SARS-CoV-2/human/USA/FL-BPHL-2776/2021

4. Use Quick Search to pull out corresponding whole genome sequences.

5. Use SNP tool to calculate consensus genome sequence => "B.1.1.7_North America_genome_consensus.fasta"

6. Combine with "Wuhan-Hu-1_genome.fasta" with "B.1.1.7_North America_genome_consensus.fasta" => "Wuhan+B.1.1.7(QC)_North America_consensus_genome.fasta"

7. Run SNP on "Wuhan+B.1.1.7(QC)_North America_consensus_genome.fasta" to generate SNP report of all consensus substitutions in comparison with Wuhan-Hu-1 => 48 SNPs, including 19 deletions in "Wuhan-Hu-1 + B.1.1.7_North America_consensus_genome_SNP.xlsx"

8. Annotate "B.1.1.7_North America_genome_consensus.fasta" with VIGOR4 => "B.1.1.7_North America_genome_consensus_VIGOR4"

9. Align "Wuhan-Hu-1 S protein.fasta" with "B.1.1.7_North America_S protein_consensus.fasta" to determine S protein substitutions => Spike: H69-, V70-, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H

10. Determine which SNPs were located within which ORF gene or mature peptide genomic region through Wuhan mapping table. Combine relevant B.1.1.7_genome_consensus_VIGOR4 protein with corresponding Wuhan-Hu-1 protein. Perform multiple sequence alignment and determine amino acid substitutions for all genomic SNPs => nsp3: T183I, A890D, I1412T; nsp6: S106-, G107-, F108-; RNA-dependent RNA polymerase: P323L; helicase: K460R; ORF8: Q27stop

11. BLAST consensus against "B1.1.7(QC)_North America_genome" working set => best representative (Identities = 29847/29861 (99%)) SARS-CoV-2/human/USA/FL-CDC-STM-P012/2020 -

https://www.viprbrc.org/brc/viprStrainDetails.spg?ncbiAccession=MW430974&decorator=corona&context=1611876836946

Calculation of B.1.351 substitutions

1. Generate an S protein working set to use as a BLAST/Short Peptide Search database using query: Coronaviridae, S protein, since 2021 on 11FEB2021 in ViPR.

2. Use Short Peptide Search for PGQTGNIADYN to identify S proteins in ViPR since 2021 carrying the K417N substitution => 4 proteins

3. For quality control purposes, removed strains in which S protein sequences had any "X" ambiguous positions => 3 proteins from the following strains:

    a. SARS-CoV-2/human/USA/SC-COVID21-0037/2021,
    SARS-CoV-2/human/USA/SC-CDC-LC0003421/2021,
    SARS-CoV-2/human/GHA/WACCBIP_nCoV_GS73/2021

4. Use Quick Search to pull out corresponding whole genome sequences – "B.1.351_genome_in VIPR_12FEB2021".

5. Use SNP tool to calculate consensus genome sequence => "B.1.351_genome_in VIPR_12FEB2021_consensus.fasta"

6. Combine "Wuhan-Hu-1_genome.fasta" with " B.1.351 consensus ViPR only_genome.fasta" => " Wuhan-Hu-1 + B.1.351 consensus ViPR only_genome.fasta"

7. Run SNP on " Wuhan-Hu-1 + B.1.351 consensus ViPR only_genome.fasta" to generate SNP report of all consensus substitutions in comparison with Wuhan-Hu-1 => 39 SNPs, including 18 deletions in " Wuhan-Hu-1 + B.1.351 consensus ViPR only_genome_SNP.xlsx"

8. Manually curate " B.1.351_genome_in VIPR_12FEB2021_consensus.fasta " to remove – characters and annotate with VIGOR4 => "B.1.351 consensus ViPR only_genome_VIGOR4 "

9. Align "Wuhan-Hu-1 S protein.fasta" with " B.1.351_S protein_consensus_ViPR only.fasta to determine S protein substitutions => Spike: D80A, D215G, L241-, L242-, A243-, K417N, E484K, N501Y, D614G, A701V

10. Determine which SNPs were located within which ORF gene or mature peptide genomic region through Wuhan mapping table. Combine relevant "B.1.351 consensus ViPR only_genome_VIGOR4" protein with corresponding Wuhan-Hu-1 protein. Perform multiple sequence alignment and determine amino acid substitutions for all genomic SNPs => nsp2: T85I; nsp3: K837N, A1775V; 3C-like proteinase: K90R; nsp6: S106-, G107-, F108-; RNA-dependent RNA polymerase: P323L; helicase: T588I; ORF3a: Q57H; S171L; envelope protein: P71L; nucleocapsid phosphoprotein: T205I

11. Review MSA and SNP reports to determine best representative sequence to be SARS-CoV-2/human/GHA/WACCBIP_nCoV_GS73/2021 which is 99% identical to the B.1.351 ViPR only consensus with two one-ambiguous (N), two two-ambiguous, two-nine ambiguous, and one 218-ambiguous regions - https://www.viprbrc.org/brc/viprStrainDetails.spg?ncbiAccession=MW571126&decorator=corona&context=1613244976238

Calculation of P.1 (North American) substitutions

The constellation of substitutions carried by the B.1.1.7 lineage were computed for the first US isolate (SARS-CoV-2/human/USA/MN-MDH-2399/2021) in comparison with Wuhan-Hu-1 using the sequence record derived from Genbank so that the sequence can be released through ViPR and BV-BRC and so that they are relevant to circulation in the U.S, as follows, as follows:

1. Use Quick Search to pull out SARS-CoV-2/human/USA/MN-MDH-2399/2021 whole genome sequence => "P.1_US_genome_consensus.fasta"

2. Combine with "Wuhan-Hu-1_genome.fasta" with "P.1_US_genome_consensus.fasta" => "Wuhan+P.1_US_genome_consensus.fasta"

3. Run SNP on "Wuhan+P.1_US_genome_consensus.fasta" to generate SNP report of all consensus substitutions in comparison with Wuhan-Hu-1 => 49 SNPs, including 9 deletions and 5 insertions in "Wuhan-Hu-1 + P.1_first US_genome_SNP.xlsx"

4. Annotate "P.1_US_genome_consensus.fasta" with VIGOR4 => "P.1_US_genome_consensus_VIGOR4"

5. Align "Wuhan-Hu-1 S protein.fasta" with "P.1_US_S protein_consensus.fasta" to determine S protein substitutions => Spike: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F

6. Determine which SNPs were located within which ORF gene or mature peptide genomic region through Wuhan mapping table. Combine relevant P.1_US_genome_consensus_VIGOR4 protein with corresponding Wuhan-Hu-1 protein. Perform multiple sequence alignment and determine amino acid substitutions for all genomic SNPs => nsp3: S370L, K977Q; nsp4: S184N; 3C-like proteinase: A260V; nsp6: S106-, G107-, F108-; RNA-dependent RNA polymerase: P323L; helicase: E341D; ORF3a protein: S253P; ORF8 protein: E92K; nucleocapsid phosphoprotein: P80R, R203K, G204R

7. Representative sequence: SARS-CoV-2/human/USA/MN-MDH-2399/2021 - https://www.viprbrc.org/brc/viprStrainDetails.spg?ncbiAccession=MW520923&decorator=corona&context=1611876882126

Calculation of CAL.20C substitutions

1. Generate an S protein working set to use as a BLAST/Short Peptide Search database using query: SARS-related species, complete genome, S protein, sinceJUN2020, human, North America on 25JAN2021.

2. Use Short Peptide Search for NYNYRYRLFR to identify S proteins from North America since July 2020 carrying the L452R substitution => 55 proteins

3. For quality control purposes, removed strains in which S protein sequences had any "X" ambiguous positions => 17 proteins from the following strains:

    a. SARS-CoV-2/human/USA/CA-CZB-12872/2020,
    SARS-CoV-2/human/USA/NMDOH-2021013232/2021,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00003/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00014/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012360496/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00041/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00074/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00077/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00169/2021,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00141/2021,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00055/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00030/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00031/2020,
    SARS-CoV-2/human/USA/WA-S3353/2020, SARS-CoV-2/human/USA/WA-S3355/2020,
    SARS-CoV-2/human/USA/CA-CDC-STM-A100056/2021,
    SARS-CoV-2/human/USA/KY-CDC-STM-A100059/2021

4. Use Quick Search to pull out corresponding whole genome sequences.

5. Use SNP tool to calculate consensus genome sequence => "CAL.20C_genome_consensus.fasta"

6. Combine with Wuhan-Hu-1_genome.fasta with CAL.20C_genome_consensus.fasta => Wuhan+CAL.20C consensus_genome_25JAN2021.fasta

7. Run SNP on "Wuhan+CAL.20C consensus_genome_25JAN2021.fasta" to generate SNP report of all consensus substitutions in comparison with Wuhan-Hu-1 => 21 SNPs in "Wuhan-Hu-1 + CAL.20C consensus_genome_SNP.xlsx"

8. Annotate "CAL.20C_genome_consensus.fasta" with VIGOR4 => "CAL.20C_genome_consensus _VIGOR4"

9. Align "Wuhan-Hu-1 S protein.fasta" with "CAL.20C_S protein_consensus.fasta" to determine S protein substitutions => Spike: S13I, W152C, L452R, D614G

10. Determine which SNPs were located within which ORF gene or mature peptide genomic region through Wuhan mapping table. Combine relevant CAL.20C_genome_consensus_VIGOR4 protein with corresponding Wuhan-Hu-1 protein. Perform multiple sequence alignment and determine amino acid substitutions for all genomic SNPs.

11. BLAST consensus against "genome_North America_CAL.20C_QC" working set => best representative (Identities = 29893/29901 (99%)) "SARS-CoV-2/human/USA/CA-LACPHL-AF00141/2021" -

https://www.viprbrc.org/brc/viprStrainDetails.spg?ncbiAccession=MW485882&decorator=corona&context=1611876924806

Calculation of B1.375 substitutions

1. Generate an S protein working set to use as a BLAST/Short Peptide Search database using query: SARS-related species, complete genome, S protein, sinceJUN2020, human, North America on 25JAN2021.

2. Use Short Peptide Search for NVTWFHAISGTNGTKRFD to identify S proteins from North America since July 2020 carrying the 69/70 deletion => 49 proteins

3. Use Short Peptide Search for YGFQPTNGVG to identify S proteins from North America since July 2020 that do not also carry the N501Y substitution of B1.1.7 => 38 proteins from the following strains:

    a. SARS-CoV-2/human/USA/FL-BPHL-2259/2020,
    SARS-CoV-2/human/USA/GA-CDC-STM-A100055/2021,
    SARS-CoV-2/human/USA/UT-UPHL-2012039803/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012566145/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012558654/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012218768/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012875547/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012522014/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012983373/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012045106/2020,
    SARS-CoV-2/human/USA/UT-UPHL-2012663924/2020,
    SARS-CoV-2/human/USA/WI-UW-2622/2020,
    SARS-CoV-2/human/USA/WI-UW-2590/2021,
    SARS-CoV-2/human/USA/WI-UW-2607/2021,
    SARS-CoV-2/human/USA/WI-UW-2634/2020,
    SARS-CoV-2/human/USA/WI-UW-2645/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P004/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P007/2020,
    SARS-CoV-2/human/USA/CA-CDC-STM-P002/2020,
    SARS-CoV-2/human/USA/CA-CDC-STM-P008/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P005/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P011/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P013/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P014/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P015/2020,
    SARS-CoV-2/human/USA/FL-CDC-STM-P058/2020,
    SARS-CoV-2/human/USA/MA-CDC-STM-P006/2020,
    SARS-CoV-2/human/USA/MA-CDC-STM-P016/2020,
    SARS-CoV-2/human/USA/MA-CDC-STM-P018/2020,
    SARS-CoV-2/human/USA/PA-CDC-STM-P010/2020,
    SARS-CoV-2/human/USA/PA-CDC-STM-P020/2020,
    SARS-CoV-2/human/USA/CA-LACPHL-AF00166/2021,
    SARS-CoV-2/human/USA/WI-UW-2513/2020,
    SARS-CoV-2/human/USA/WI-UW-2527/2020,
    SARS-CoV-2/human/USA/MA-CDC-STM-A100010/2021,
    SARS-CoV-2/human/USA/FL-CDC-STM-A100001/2021,
    SARS-CoV-2/human/USA/FL-CDC-STM-A100090/2021,
    SARS-CoV-2/human/USA/UT-UPHL-2012257709/2020

4. Use Quick Search to pull out corresponding whole genome sequences.

5. Use SNP tool to calculate consensus genome sequence => “B.1.375_North America_genome_consensus.fasta”

6. Combine with Wuhan-Hu-1_genome.fasta with B.1.375_North America_genome_consensus.fasta=> “Wuhan+ B.1.375_North America_genome_consensus.fasta”

7. Run SNP on “Wuhan+ B.1.375_North America_genome_consensus.fasta” to generate SNP report of all consensus substitutions in comparison with Wuhan-Hu-1 => 38 SNPs, including 22 deletions in “Wuhan-Hu-1 + B.1.375 consensus_genome_SNP.xlsx”

8. Annotate “B.1.375_North America_genome_consensus.fasta” with VIGOR4 => “B.1.375_North America_genome_consensus _VIGOR4”

9. Align “Wuhan-Hu-1 S protein.fasta” with “B.1.375_North America_S protein_consensus.fasta” to determine S protein substitutions => Spike: H69-, V70-, D614G

10. Determine which SNPs were located within which ORF gene or mature peptide genomic region through Wuhan mapping table. Combine relevant B.1.375_North America_genome_consensus _VIGOR4 protein with corresponding Wuhan-Hu-1 protein. Perform multiple sequence alignment and determine amino acid substitutions for all genomic SNPs => nsp2: T85I; nsp3: T1010A; RNA-dependent RNA polymerase: P323L; helicase: E341D; 2'-O-ribose methyltransferase: A258V; ORF3a: Q57H; membrane glycoprotein: I48V; nucleocapsid phosphoprotein: T205I

11. BLAST consensus against “B.1.375_North America_genome” => best representative (Identities = 29869/29876 (99%) SARS-CoV-2/human/USA/FL-CDC-STM-P015/2020 - https://www.viprbrc.org/brc/viprStrainDetails.spg?ncbiAccession=MW430977&decorator=corona