

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2022

Tereza Švestková



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

KOMPARATIVNÍ A FYLOGENETICKÁ ANALÝZA VIRŮ VE FN BRNO

COMPARATIVE AND PHYLOGENETIC ANALYSIS OF VIRUSES IN THE UNIVERSITY HOSPITAL BRNO

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

Tereza Švestková

VEDOUCÍ PRÁCE

SUPERVISOR

Ing. Markéta Nykrýnová

BRNO 2022



Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Tereza Švestková

ID: 221532

Ročník: 3

Akademický rok: 2021/22

NÁZEV TÉMATU:

Komparativní a fylogenetická analýza virů ve FN Brno

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma SARS-CoV-2 a jeho sekvenování. 2) Sestavte osekvenované genomy a vyhodnoťte jejich kvalitu. 3) Sestavená data anotujte. 4) Vyhodnoťte variabilitu v anotovaných datech v rámci daných časových úseků. 5) Proveďte fylogenetickou analýzu. 6) Výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] YIN, Changchuan. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*. 2020, 112(5), 3588-3596. DOI: 10.1016/j.ygeno.2020.04.016.

[2] KOYAMA, Takahiko, Daniel PLATT a Laxmi PARIDA. Variant analysis of SARS-CoV-2 genomes. *Bulletin of the World Health Organization*. 2020, 98(7), 495-504. DOI: 10.2471/BLT.20.253591.

Termín zadání: 7.2.2022

Termín odevzdání: 27.5.2022

Vedoucí práce: Ing. Markéta Nykrýnová

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

UPOZORNĚNÍ:

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

ABSTRAKT

Práce je věnovaná koronaviru SARS-CoV-2 souvisejícím s těžkým akutním respiračním syndromem, který byl poprvé prokázán v roce 2019. Tento koronavirus zapříčinil pandemii, která ovlivnila takřka celý svět. Znalost genetické informace je potřebná pro vývoj vakcíny, zjištění nakažlivosti a pro predikci vývoje variant SARS-CoV-2. Pro získání genetických informací je třeba osekvenovat RNA a tyto genomové sekvence sestavit. Při porovnání sestavených genomů lze zjistit, v které části daný organismus mutoval. Na základě souhlasnosti či rozdílnosti v sestavených genomech je provedena fylogenetická analýza, která poukazuje na vývoj daného organismu a znázorňuje evoluční příbuznost s ostatními organismy. Praktická část je zaměřena na sestavení genomů ze vzorků od pacientů z FN Brno a vyhodnocení kvality sestavení. Po sestavení genomů je dalším cílem vyhodnocení variability a následná fylogenetická analýza.

KLÍČOVÁ SLOVA

koronaviry, SARS-CoV-2, mutace, sekvenování, genom

ABSTRACT

This thesis is focused on the SARS-CoV-2 coronavirus associated with severe acute respiratory syndrome, which was first identified in 2019. This coronavirus caused a pandemic that affected almost the entire world. Knowledge of the genetic information is needed for vaccine development, to determine infectivity and to predict the evolution of SARS-CoV-2 variants. To obtain genetic information, RNA must be sequenced and these genomic sequences must be assembled. By comparing the assembled genomes, it is possible to find out which part of the organism has mutated. Phylogenetic analysis is performed on the basis of the concordance or divergence in the assembled genomes, which indicates the evolution of the organism and shows the evolutionary relationship with other organisms. The practical part is focused on the assembly of genomes from samples from patients in the University Hospital Brno and evaluation of the quality of the assembly. After the genomes are assembled, the next goal is to evaluate the variability and subsequent phylogenetic analysis.

KEYWORDS

coronaviruses, SARS-CoV-2, mutations, sequencing, genome

ŠVESTKOVÁ, Tereza. *Komparativní a fylogenetická analýza virů ve FN Brno*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 69 s. Bakalářská práce. Vedoucí práce: Ing. Markéta Nykrýnová

Prohlášení autora o původnosti díla

Jméno a příjmení autora: Tereza Švestková
VUT ID autora: 221532
Typ práce: Bakalářská práce
Akademický rok: 2021/22
Téma závěrečné práce: Komparativní a fylogenetická analýza virů
ve FN Brno

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno

.....

podpis autora*

* Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí bakalářské práce paní Ing. Markétě Nykrýnové za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Úvod	11
1 Viry	12
1.1 RNA viry	12
1.2 Koronaviry	12
2 SARs-CoV-2	14
2.1 Srovnání SARs-CoV-2 s ostatními lidskými CoVs	14
2.2 Přenos SARs-CoV-2	14
2.2.1 Průběh virové infekce a jeho replikace	15
3 Fylogenetický vývoj a genetická informace SARs-CoV-2	16
3.1 Fylogeneze	16
3.2 Genom SARs-CoV-2	16
3.2.1 Genotypizace SARs-CoV-2	17
3.3 Prokázání výskytu SARs-CoV-2	17
3.3.1 Metoda RT - qPCR	18
3.3.2 Multiplex Real - time PCR	18
3.4 Metody sekvenování SARs-CoV-2	19
3.4.1 Sangerovo sekvenování	20
3.4.2 Illumina MiSeq	20
3.4.3 Nanopórové sekvenování Oxford Nanopore	20
3.5 Mutace SARs-CoV-2	21
3.5.1 Alfa mutace	21
3.5.2 Beta mutace	21
3.5.3 Gama mutace	21
3.5.4 Delta mutace	22
3.5.5 Omikron mutace	22
4 Sestavení vzorků SARs-CoV-2	23
4.1 Kontrola kvality sekvenačních dat	23
4.1.1 Kvalita čtení	23
4.1.2 Průměrné před skóre pro sekvenci	24
4.1.3 Obsah jednotlivých bází	25
4.1.4 Obsah páru G - C	26
4.1.5 Obsah neurčené báze N	27
4.1.6 Obsah adaptérů	28
4.1.7 Selekcce dat	29

4.2	Mapování genomů k referenční sekvenci	29
4.3	Sestavení genomů <i>de novo</i>	30
5	Analýza vzorků v závislosti na změnách v genomu v určitých časových úsecích	31
5.1	Anotování dat	31
5.2	Vyhodnocení variability	32
6	Fylogenetická analýza	38
	Závěr	42
	Literatura	43
	Seznam symbolů a zkratk	48
	Seznam příloh	49
A	Blokové schéma mapování sekvenačních dat k referenční sekvenci	50
B	Výsledky mapování k referenční sekvenci	51
C	Výsledky sestavení <i>de novo</i>	54
D	Počet mutací v daných genech u jednotlivých vzorků	55
E	Fylogramy různých metod konstrukce fylogenetických stromů	60
F	Výsledky shlukové analýzy z NEWICK formátů uvedených metod	62

Seznam obrázků

1.1	Mikroskopický snímek koronaviru, převzato z [3]	13
3.1	Schéma viru SARS-CoV-2. Znázornění čtyř strukturních proteinů: S, M, N a E, převzato z [11]	17
3.2	Polymerázová řetězová reakce - PCR, převzato z [14]	19
3.3	Lokace změn u jednotlivých variant, převzato z [24], [25]	22
4.1	Phred skóre kvality čtení na daných pozicích pro vzorek CSQ0254	23
4.2	Průměrné phred skóre kvality sekvence pro vzorek CSQ162	24
4.3	Průměrné procentuální zasoupení bazí pro vzorek CSQ0454	25
4.4	Průměrné procentuální zasoupení G - C párů pro vzorek CSQ0919	26
4.5	Průměrné procentuální zasoupení N báze pro vzorek CSQ210	27
4.6	Průměrné procentuální zasoupení adaptérů pro vzorek CSQ0529	28
4.7	Průměrné phred skóre pro kvalitní sekvenační data	29
5.1	Mutace ve vzorcích podle data odběru v genech ORF1a a ORF1b referenční sekvence	33
5.2	Mutace ve vzorcích podle data odběru v genech S, ORF3a, M, ORF7a, ORF7b, ORF8 a N referenční sekvence	33
5.3	Výskyt mutací v jednotlivých genech pro vzorky odebrané 7.7.2021, 3.8.2021 a 6.8.2021	34
5.4	Výskyt mutací v jednotlivých genech pro vzorky odebrané 13.8.2021, 20.8.2021 a 25.8.2021	35
5.5	Výskyt mutací v jednotlivých genech pro vzorky odebrané 10.9.2021, 15.9.2021	36
5.6	Výskyt mutací v jednotlivých genech pro vzorky odebrané 24.9.2021	37
6.1	Fylogenetický strom metodou Maximální pravděpodobnosti	39
6.2	Seskupení vzorků z fylogenetického stromu metodou Maximální pravděpodobnosti	40
A.1	Blokové schéma úpravy sekvenačních dat, mapování vzorků k referenční sekvenci a následné převedení do FASTA formátu	50
E.1	Fylogenetický strom metodou UPGMA	60
E.2	Fylogenetický strom metodou Spojování nejbližších sousedů	61
E.3	Fylogenetický strom metodou Minimální evoluce	61
F.1	Seskupení vzorků z fylogenetického stromu metodou Spojování nejbližších sousedů	62
F.2	Seskupení vzorků z fylogenetického stromu metodou UPGMA	63
F.3	Seskupení vzorků z fylogenetického stromu metodou Minimální evoluce	63

Úvod

Tato bakalářská práce se věnuje onemocnění covid-19 způsobené koronavirem SARS-CoV-2, které v posledních třech letech ovlivnilo život všech lidí na světě. Cílem této práce je seznámit čtenáře se základními informacemi o SARS-CoV-2, o příbuznosti s ostatními koronaviry a možnosti sekvenace odebraných vzorků.

Teoretická část nejprve pojednává o popisu SARS-CoV-2, jeho přenosu a porovnání s ostatními lidskými koronaviry. Následně je čtenář seznámen s genomem SARS-CoV-2, s metodami detekce viru metodou polymerázové řetězové reakce a dále jsou objasněny principy sekvenování genomu, jako například Sangerovo sekvenování a nanopórové sekvenování Oxford Nanopore. Tyto principy se používají pro získání genetické informace viru, ta je podstatná pro vývoj vakcíny či predikci vývoje pandemie. Závěr teoretické části je věnován doposud známým mutacím.

Praktická část se věnuje vyhodnocení kvality sekvenačních dat odebraných ve Fakultní nemocnici Brno. Kvalitní data byla sestavena namapováním k referenční sekvenci a sestavením *de novo*. Dalším krokem praktické části bylo vyhodnocení variability vzorků v rámci časových úseků odběrů vzorků a v rámci jednotlivých genů. Následuje fylogenetická analýza, která srovnává vývoj vzorků v čase a jejich příbuznost. Na základě této analýzy bylo provedeno hierarchické shlukování s předpokladem vytvoření shluků podle data odběru.

1 Viry

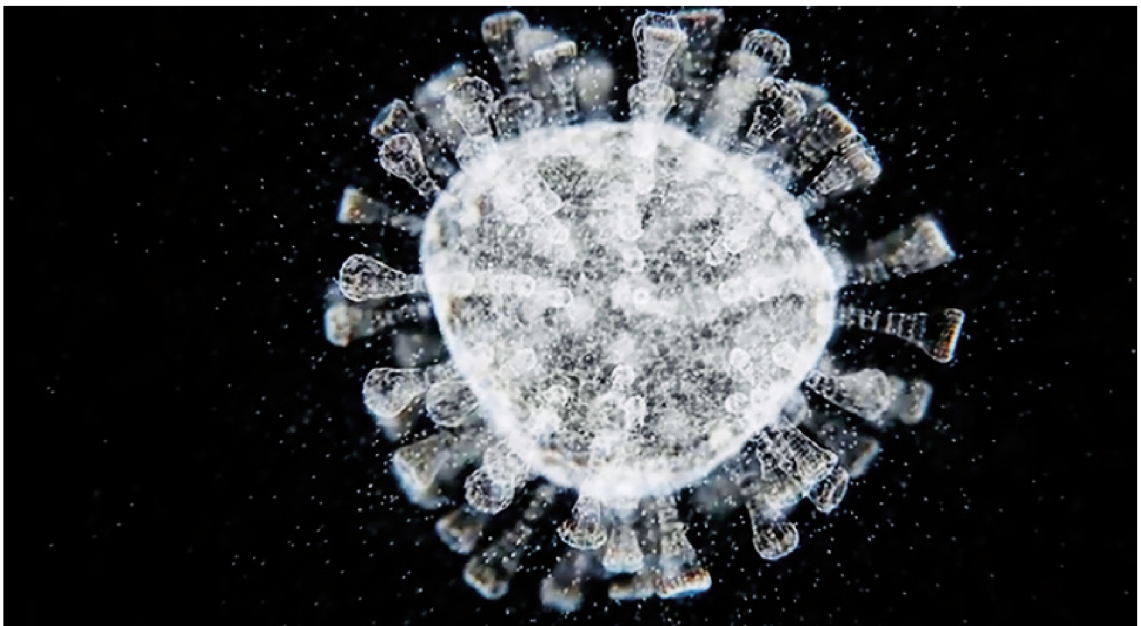
Pro pochopení genomu SARs-CoV-2 je vhodné popsat vyšší skupinu, do které se řadí a tou jsou viry. Jedná se o nebuněčné organismy, které nemají svůj vlastní proteosyntetický aparát, a proto potřebují pro svoji existenci hostitelskou buňku. Skládají se z nukleové kyseliny a bílkovinného obalu (neboli kapsidu) kolem nukleové kyseliny. V některých případech mohou obsahovat i enzymy. [1]

1.1 RNA viry

Obecně mají viry obsahující RNA - ribonukleovou kyselinu mnohokrát kratší délku genomu než viry s DNA - kyselinou deoxyribonukleovou. Díky tomu mají vyšší pravděpodobnost mutagenity. Jak již název vypovídá, genetická informace virů, ke kterým se řadí i koronaviry, je obsažena v jednovláknové RNA. Tato kyselina může sloužit přímo jako mediátorová RNA pro přepis do proteinů nebo bývá replikována za vzniku komplementárního vlákna. [2]

1.2 Koronaviry

Koronaviry, zkráceně CoVs, patří do čeledi *Coronaviridae*, pod kterou se řadí čtyři rody *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, *Deltacoronavirus*. Tyto viry dostaly název podle svého tvaru, jelikož struktura jejich obalu připomíná sluneční korunu, jak je vidět na obrázku 1.1. Genom CoV se skládá z šesti až jedenácti otevřených čtecích rámců (ORF) neboli sekvencí ribonukleotidů v mediátorové RNA, která je ohraničená iniciačním (start) a terminačním (stop) kodonem. Jejich genetická informace dosahuje délky 26 až 32 kilobází, což z nich dělá nejdélší známé RNA viry. I přes to zde však dochází k neustálému vývoji a četným mutacím. CoVs jsou sice typické pro onemocnění zvířat, ale snadno při určitých podmínkách infikují lidi, proto jsou důvodem několika náhlých, nejen lokálních epidemií po celém světě. [2]



Obr. 1.1: Mikroskopický snímek koronaviru, převzato z [3]

2 SARs-CoV-2

Tento koronavirus podle svého názvu souvisí s akutním respiračním syndromem. První výskyt byl prokázán v prosinci 2019 ve Wuhanu v Číně. V následujících kapitolách je popsán taxonomický vztah s ostatními koronaviry, přenos tohoto viru, možné sekvenování genomu a jeho prozatím známé mutace. [2]

2.1 Srovnání SARs-CoV-2 s ostatními lidskými CoVs

SARs-CoV-2 je osmý známý koronavirus vyskytující se u lidí a spadá do rodu *Betacoronavirus*. Získal označení SARs-CoV-2 kvůli taxonomickému a genomickému vztahu s druhem koronaviru označeným jako SARs-CoV, který zapříčiňuje akutní respirační syndrom. Fylogenetická analýza ukázala, že SARs-CoV-2 používá pro vstup do buněk stejný receptor jako SARs-CoV a to ACE2 - angiotensin-konvertující enzym 2. K tomuto enzymu vykazuje SARs-CoV-2 až dvacetkrát vyšší afinitu než SARs-CoV. Genomová analýza mezi těmito dvěma koronaviry prokazuje 58 % souhlasnost v oblasti nsp - protein kódovaný virem, který je využit pro virovou replikaci a 79,5 % podobnost na úrovni celého genomu. Co se týče patologických nálezů, z velké části se podobají příznakům, které se projevovaly v případě nákazy viry SARs-CoV a MERs-CoV. [4], [5]

2.2 Přenos SARs-CoV-2

K přenosu viru dochází z nosu či ústní dutiny v malých kapénkách nejen při kašlání, ale i při kýchní, mluvení nebo zpívání. Největší pravděpodobnost přenosu onemocnění je zejména v těsném kontaktu s infikovanou osobou (vzdálenost do 1 metru). Nákaza probíhá vdechnutím nebo stykem sliznic (oči, nos, ústa) s aerosolem obsahujícím virus. Dalším způsobem je nákaza v uzavřených prostorách, kde vydechovaný vzduch od infikovaných osob zůstává velmi dlouhou dobu. Posledním častým způsobem je nákaza z infikovaných povrchů a následné zanesení do sliznic při špatné hygieně, například v hromadné dopravě při dotyku madel a následné konzumaci potravin bez předešlého umytí rukou či promnutí očí. Jedním z důvodů, proč covid-19 přerost v pandemii, je jeho vysoká infekčnost i několik dní před projevením příznaků u lidí s nákazou. V některých případech se příznaky vůbec projevit nemusí a infikované osoby jsou tak dokonalými přenašeči. [2]

2.2.1 Průběh virové infekce a jeho replikace

Pokud dojde k přenesení viru od nakažené osoby či z infikovaného povrchu, lze průběh popsat v několika krocích. Dojde k připojení S - spike glykoproteinu na receptor ACE2 na buněčné membráně buněk hostitele. Tím jsou zahájeny změny konformace v S proteinu, které umožní průnik a uvolnění virové RNA. Po průniku do buňky se vir zbavuje kapsidu, dochází k fúzi RNA, aby byla umožněna transkripce genů. V podstatě ihned dochází k translaci ORF1a a ORF1b, které slouží pro vytvoření replikačního a transkripčního procesu. Nyní je připraveno prostředí pro replikaci virové informace a sestavení virionu. Při expresi dochází k transkripci, která je ukončena v regulačních sekvencích nacházejících se mezi ORF. Následuje translace za účelem kódování virových proteinů. Po ukončení replikace dochází k přemístění virové RNA do endoplazmatického retikula a Golgiho aparátu, kde se zabuduje do kapsidu. Celý průběh množení viru končí uvolněním z hostitelské buňky do okolí exocytózou. Z medicínského hlediska je průběh infekce hůře popsateľný, jelikož každý pacient může mít odlišné příznaky. Mezi nejčastější patologické nálezy patří horečky, bolesti hlavy, dušnost, kašel. Z krevních vzorků byl zjištěn úbytek bílých krvinek zodpovědných za imunitu, konkrétně T-lymfocytů. Na rentgenových snímcích hrudníku se často objevuje rychle postupující pneumonie čili zápal plic. Není zde pochyb, že SARs-CoV-2 má patologický vliv i na jiné tkáně jako například srdce či játra. [2]

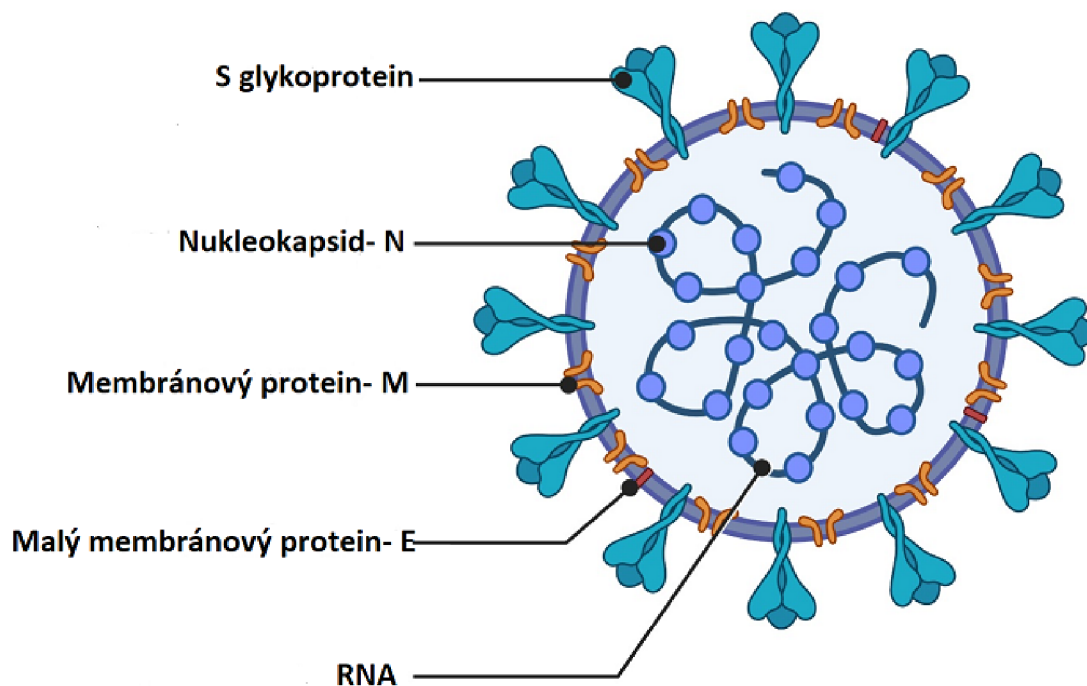
3 Fylogenetický vývoj a genetická informace SARs-CoV-2

3.1 Fylogeneze

Z fylogenetického hlediska je SARs-CoV-2 zařazen do skupiny se SARs-CoV a jinými koronaviry objevenými u netopýřů, cibetek a jiných savců. Jde o takzvané zoologické přelévání, kdy dochází k rekombinacím a následnému přenosu na jiné savce. Oproti netopýřímu genomu je genom SARs-CoV-2 vyskytující se u lidí o více jak tisíc bází delší. Fylogenetické výzkumy naznačují vysokou frekvenci rekombinací u netopýřů, z tohoto důvodu se vědci domnívají, že právě netopýři jsou původci CoVs. [2]

3.2 Genom SARs-CoV-2

Genom se skládá z jednovláknové RNA ve směru od 5' do 3' konce. První 2/3 kóduje 16 nsp, zbývající informace zahrnuje 4 strukturní proteiny: spike glykoprotein - špice (S), membránový (M), obalový (E) a nukleokapsidový protein (N). Struktura SARs-CoV-2 je znázorněna na obrázku 3.1. Glykoprotein S existuje v metastabilním stavu, který prochází při průniku viru do buňky reorganizací. Dochází k aktivaci serinovou proteázou na podjednotky S1 a S2. Podjednotka S1, která je ze 70 % identická se SARs-CoV, obsahuje vazebnou doménu pro receptor ACE2. Podjednotka S2, která má souhlasnost se SARs-CoV až z 99 %, je zodpovědná za průnik viru přes membránu hostitelské buňky. Protein M existuje v podobě dimeru, což znamená, že je složen ze dvou menších podjednotek transmembránové domény a dlouhé C-terminální domény. Jedná se o hlavní protein membrány viru, který určuje její strukturu. Mimo jiné M protein reaguje i s jinými strukturními proteiny v endoplazmatickém retikulu a Golgiho aparátu. Při této interakci dochází ke vzniku nových virových částic. Konkrétně vazba mezi M a S proteiny se jeví jako úzce spjatá s patogenitou SARs-CoV-2 a považuje se za podstatnou v udržení rovnováhy virionu. Nejmenší ze strukturních proteinů je obalový nebo také malý membránový protein E. Stejně jako u M proteinu je i jeho část přítomna v Golgiho aparátu při produkci nových virových buněk. Kromě této úlohy se E protein spojuje do větších molekul za vzniku iontových kanálů pro modifikaci membrány hostitelské buňky a ulehčuje tak jeho fúzi do extracelulárního prostředí. Jak je již z označení patrné, N protein plní funkci strukturovaného obalu RNA. Je složen ze dvou domén, které obsahují vazebná místa pro nukleové kyseliny. [2], [6], [7], [8], [9]



Obr. 3.1: Schéma viru SARS-CoV-2. Znázornění čtyř strukturálních proteinů: S, M, N a E, převzato z [11]

3.2.1 Genotypizace SARs-CoV-2

Genotypizace je metoda identifikace druhu pomocí DNA/ RNA sekvencí, která je důležitá pro pochopení vývoje a přenosu virů i jiných organismů. Sekvence SARS-CoV-2 jsou izolovány z epitelových buněk dýchacích cest infikovaných osob. Virové kontigy jsou sestaveny buď úplně nebo částečně, přičemž sestavení do úplného genomu se využívá pro genotypizaci pomocí vícenásobného zarovnávání sekvencí izolátů s referenčním genomem. Tato analýza srovnává více sekvencí a odhaluje tak schopnost mutagenity podle počtu změn v genomu. Nejvíce náchylné k mutacím jsou S proteiny, RNA primáza, nukleoprotein a RNA polymeráza, jelikož na nich bylo prokázáno nejvíce změn za nejkratší dobu. [10]

3.3 Prokázání výskytu SARs-CoV-2

Důležitým krokem k získání genomu SARs-CoV-2 pro další analýzy je zjištění pozitivitu vzorků. Prokázání virové infekce se provádí pomocí polymerázové řetězové reakce s reverzní transkripcí (RT - qPCR) nebo metodou polymerázové řetězové reakce v reálném čase (Real - time PCR).

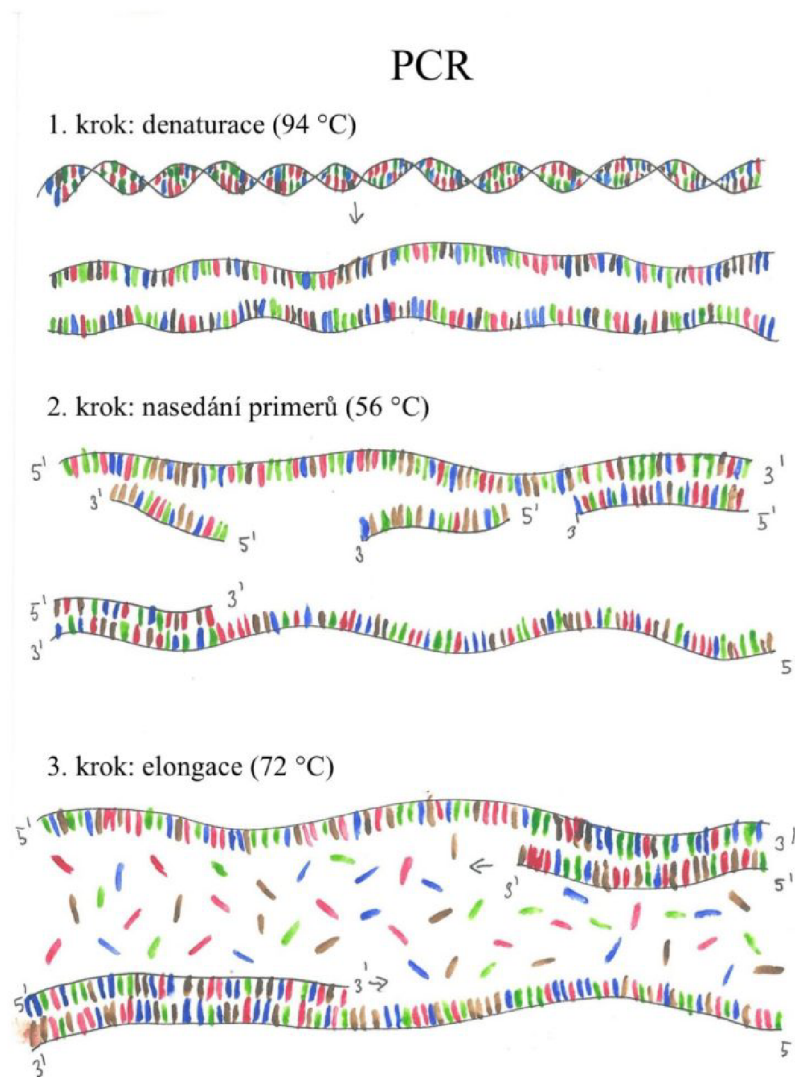
3.3.1 Metoda RT - qPCR

Jedná se o využívaný nástroj k prokázání onemocnění Covid-19. Jde o namnožení dostatečného počtu kopií pro detekci genetické informace SARs-CoV-2. Umožňuje použít RNA jako templát pro přepis do komplementárního vlákna DNA (cDNA) pomocí reverzních transkriptáz, což jsou enzymy, které tento proces umožňují. Prvním krokem je právě vznik cDNA, která je následně amplifikována běžným postupem PCR. Následuje příprava směsi, dochází ke smíchání templátové DNA, enzymu polymerázy, primerů a pufru. Templátová DNA slouží jako vzor pro syntetizování dostatečného počtu kopií za pomoci polymerázy. Primery jsou krátké syntetické komplementární sekvence, které označují oblast zájmu. Pufř je roztok, který vytváří vhodné prostředí pro tyto reakce. Výše zmíněná směs prochází třemi kroky, které se opakují do požadovaného počtu amplifikovaných vzorků. Nejprve dochází k zahřátí na 15 sekund na 94 °C. Při této teplotě řetězec denaturuje a rozplete se. V dalším kroku dochází při teplotách v rozmezí 50 - 60 °C k navázání primerů. Poslední krok probíhá při teplotách okolo 72 °C, kdy DNA-polymeráza v úseku za primery syntetizuje nové vlákno podle templátu. Tyto tři kroky se opakují, dokud nedosáhneme požadovaného stupně amplifikace. Schéma PCR je znázorněno na obrázku 3.2.

V případě použití k prokázání onemocnění se určuje přítomnost genetické informace SARs-CoV-2 pomocí produkce fluorescenčního světla. Za přítomnosti genomu je toto světlo detekováno přístrojem provádějícím PCR. [12], [13]

3.3.2 Multiplex Real - time PCR

Tato metoda je další z variant PCR pro potvrzení výskytu viru v odebraném vzorku. Amplifikace vzorku pomocí polymerázové řetězové reakce byla již popsána výše, proto v této podkapitole je stručný popis odlišností této metody. Při Multiplex Real-time PCR dochází současně k amplifikaci a detekci genetické informace viru. Během cyklu dochází k dvacetiminutové reverzní transkripci při 50 °C, inaktivaci reverzní transkriptázy a následné aktivaci polymerázy při 95 °C. Tato reakce se provádí po dobu 15 min (45 cyklů po 15 sekundách při 94 °C a 1 minutu při 60 °C). Při tomto tepelném cyklování se používají fluorescenční barviva, která se v PCR přístroji kumulují s přítomností pozitivních vzorků a jsou následně detekovány jako fluorescenční signál, který je úměrný množství PCR amplikonu. [15]



Obr. 3.2: Polymerázová řetězová reakce - PCR, převzato z [14]

3.4 Metody sekvenování SARs-CoV-2

Pokud máme vyselektované pouze pozitivní vzorky, přichází na řadu sekvenování, ke kterému se využívá několik metod. Patří sem například Sangerovo sekvenování, sekvenování druhé generace, kde se nejčastěji využívají platformy od firmy Illumina a třetí generace se zástupcem Oxford Nanopore. V následujících kapitolách jsou vybrány a podrobněji popsány nejčastější způsoby sekvenování.

3.4.1 Sangerovo sekvenování

Pro tuto metodu se používají pozitivní vzorky potvrzené metodou PCR, při které bylo nasyntetizováno mnoho oligonukleotidových kopií. Tyto oligonukleotidy jsou prodlužovány směsí deoxynukleotidtrifosfátů (dNTP) s jednotlivými nukleotidy. Dále se při Sangerově sekvenování ve směsi používají dideoxynukleotidové trifosfáty, které jsou značeny fluorescenčním markerem. Podle daného nukleotidu v sekvenci se projeví jiné fluorescenční zabarvení (A - zelená, T - červená, G - černá, C - modrá). V dalším kroku se značené oligonukleotidy roztřídí podle velikosti pomocí gelové elektroforézy. Gelová elektroforéza funguje na principu vzorků v gelové matrici, u které je na opačnou stranu přiveden elektrický proud. Jelikož je DNA záporně nabitá, jsou oligonukleotidy taženy ke katodě. Rychlost pohybu je určena velikostí, jelikož na jednotku hmotnosti mají všechny stejný elektrický náboj. Posledním krokem je čtení gelu za použité laseru. Laser excituje fluorescenční značky a dochází k detekci emitovaného světla. Výstupem Sangerova sekvenování je chromatogram, který značí fluorescenční pík každého nukleotidu. [16]

3.4.2 Illumina MiSeq

Postup je založen na sekvenování ze vzorků potvrzených metodou PCR. Dochází k syntéze komplementárního vlákna DNA k RNA reverzní transkriptázou pomocí hexamerů. Hexamery jsou molekuly složeny ze šesti menších podjednotek nebo-li monomerů. DNA je následně amplifikována a fragmentována. Dochází k navázání adaptérů a označení nukleotidů fluoroforem. V sekvenátoru směs prochází přes průtokovou buňku, kde se emituje fluorofor. Každý nukleotid je označen jinou barvou fluoroforu, což umožňuje softwarové převedení emise jednotlivých barev na sekvenci nukleotidů. [17]

3.4.3 Nanopórové sekvenování Oxford Nanopore

Tato metoda využívá pozitivní vzorky prokázané některou z metod PCR. Následným zmnožením genetické informace jsou získány amplifikační produkty, které jsou smíchány s dalšími reakčními složkami. Tato směs se používá pro vytvoření sekvenční knihovny. Příprava knihovny probíhá nativním spojením barkodu, sekvenčního adaptéru s částí genomu, který sekvenujeme. Molekula DNA prochází nanopórem, kde dochází k měření elektrického proudu. Podle intenzity elektrického proudu lze určit, která ze čtyř bází se vyskytuje na dané pozici. [18], [19], [20]

3.5 Mutace SARs-CoV-2

Jednotlivé mutace koronaviru se odlišují podle linií, prozatím známé spadají do linie A a linie B. První mutace se vyskytla ve Velké Británii v polovině prosince 2020 a WHO ji označila jako alfa, následné mutace byly nazývány podle řecké abecedy. Pro pochopení některých principů je zde vhodné vysvětlit pojem nsSNP, jedná se o nukleotidovou substituci v proteinové sekvenci, která vede ke změně aminokyseliny a tudíž k mutacím. Nejvíce mutací je detekováno v S proteinu. [21]

Lokace změn u dosavadních známých mutací je znázorněna na obrázku 3.3.

3.5.1 Alfa mutace

Vyvolala paniku nejen ve Velké Británii, ale ve většině zemí, protože pacienti s mutací alfa vykazovaly vyšší procento úmrtnosti. Tato mutace patří do linie B.1.1.7, která je charakterizována velkým počtem nsSNP v S proteinu, nejdůležitější změny jsou v N501Y, delece AH69/AV70, P681H a ORF8 Q27stop. Změna v N501Y znamená vyšší afinitu k ACE2, tudíž dochází k lehčímu prostupu do cílových buněk. Delece AH69/AV70 ovlivňuje schopnost vazby S proteinu. P681H má vliv na okolí míst, kde se uplatňuje štěpení furinem, který aktivuje některé proteiny při syntéze. Q27stop způsobuje nefunkčnost otevřeného čtecího rámce 8. [21]

3.5.2 Beta mutace

Byla nejdříve zachycena v Jižní Africe. Obdobně jako mutace alfa má početné mutace v S proteinu zahrnující i tři substituce v RDB u K417N, E484K a N501Y. Změna v N501Y se projevila i u britské mutace a to vyšší afinitou k ACE2, záměna v E484K podle dosavadních výzkumů má také tento efekt a tyto dvě záměny se vyskytují v kombinaci. Výše zmíněné polymorfismy K417N, E484K snižují schopnost neutralizačních látek úspěšně bojovat s virem, tzn. imunitní odpověď po očkování či prodělané nemoci je slabší. [21]

3.5.3 Gama mutace

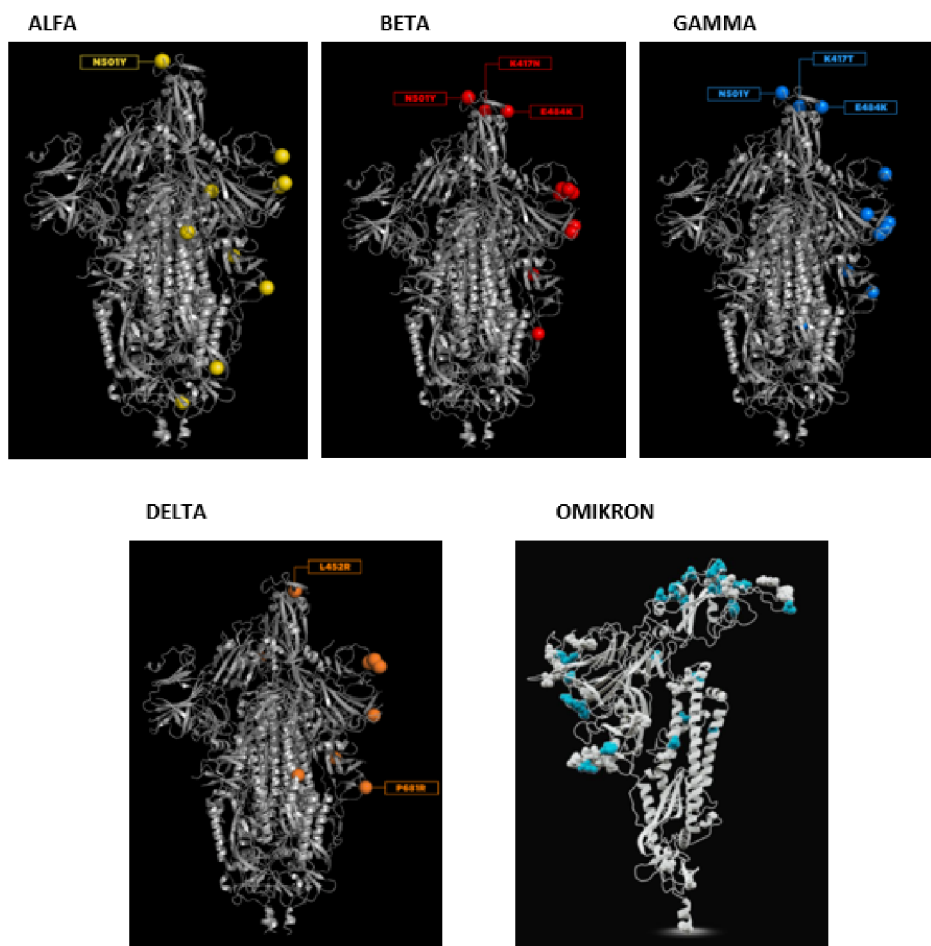
Poprvé se vyskytla u cestovatelů z Brazílie pocházející s linie B.1.1.28. Mutace opět obsahuje řadu změn v S proteinu, má také společné tři substituce jako předchozí varianta beta a to K417N, E484K a N501Y. Podle rozšíření se uvažuje o vyšší přenositelnosti, vyšší odolnost vůči obranným prostředkům, k těmto znakům se přidává i eskalující virulence - vyšší patogenita. [21]

3.5.4 Delta mutace

Tato mutace se vyskytla na počátku roku 2021 v Indii a vykazovala oproti předešlým mutacím vysokou přenositelnost. I delta varianta vykazuje změny v S proteinu a to delece L452R, T478K, D614G, P681R, D950N. Substituce P681R má pravděpodobně nejvyšší podíl na vysoké míře infekčnosti, jelikož právě tato záměna ovlivňuje štěpení S proteinu. [21]

3.5.5 Omikron mutace

Tento nový kmen B.1.1.529. byl poprvé objeven v Jižní Africe v listopadu 2021. WHO v případě omikronu poukázala na četné mutace v S proteinu, tudíž jsou zne-pokojivé vyhlídky v účinnostech vakcín, protože většina z nich využívá právě S protein. Hrozbou mohou být i náznaky, že se tato varianta vyhýbá naší imunitě dané T bílými krvinkami. [22]



Obr. 3.3: Lokace změn u jednotlivých variant, převzato z [24], [25]

4 Sestavení vzorků SARs-CoV-2

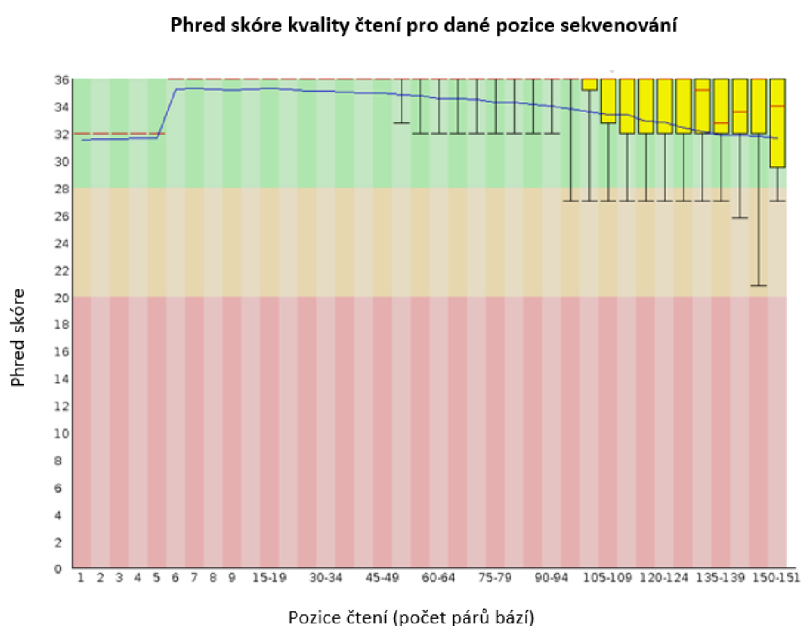
Cílem praktické části bylo sestavit osekvenované genomy SARs-CoV-2 ze vzorků od pacientů z FN Brno a vyhodnotit jejich kvalitu. Vzorky byly osekvenovány na platformě Illumina NextSeq metodou paired-end, která čte sekvenovaný fragment z obou konců, proto je možné získat přesnější zarovnání s detekovanými změnami typu indel. Celkově bylo použito 297 vzorků s délkou čtení 150 párů bází.

4.1 Kontrola kvality sekvenačních dat

U získaných dat bylo potřeba vyhodnotit jejich kvalitu, aby bylo možné pro další kroky využít pouze validní data. Kontrola kvality proběhla přes nástroj FastQC (v0.11.8, [26]). FastQC report poskytuje základní informace o sekvenačních datech, některé z nich jsou dále vypsané i s grafickým znázorněním.

4.1.1 Kvalita čtení

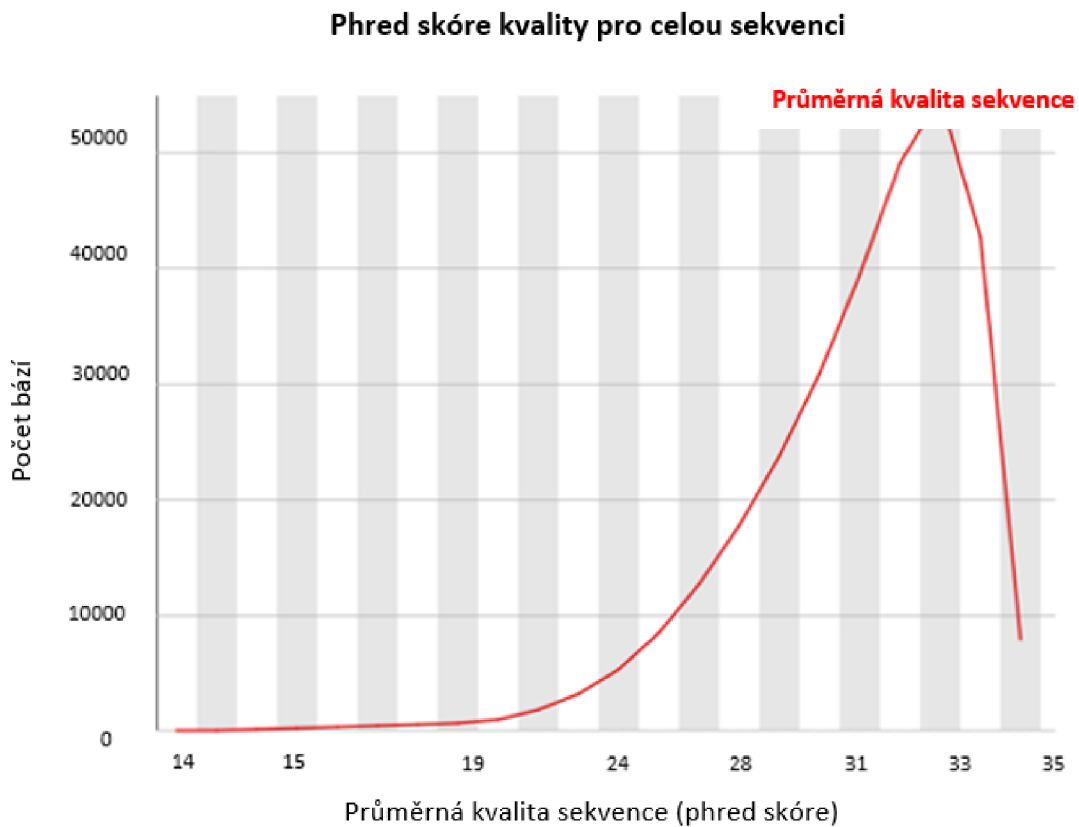
Kvalita čtení sekvence ukazuje kvalitu přečtení jednotlivých nukleotidů v rámci čtení, příklad vidíme na obrázku 4.1. Modrá čára představuje průměrné skóre na daných pozicích a červená čára odpovídá mediánu skóre kvality. Žluté boxploty vyznačují mezikvartilové rozmezí (25 - 75 %) průměrného skóre daných pozic vyznačených na ose x. Ideálně by měly tyto boxploty být vyznačeny pouze v zeleném pozadí, které značí vysoké skóre kvality.



Obr. 4.1: Phred skóre kvality čtení na daných pozicích pro vzorek CSQ0254

4.1.2 Průměrné phred skóre pro sekvenci

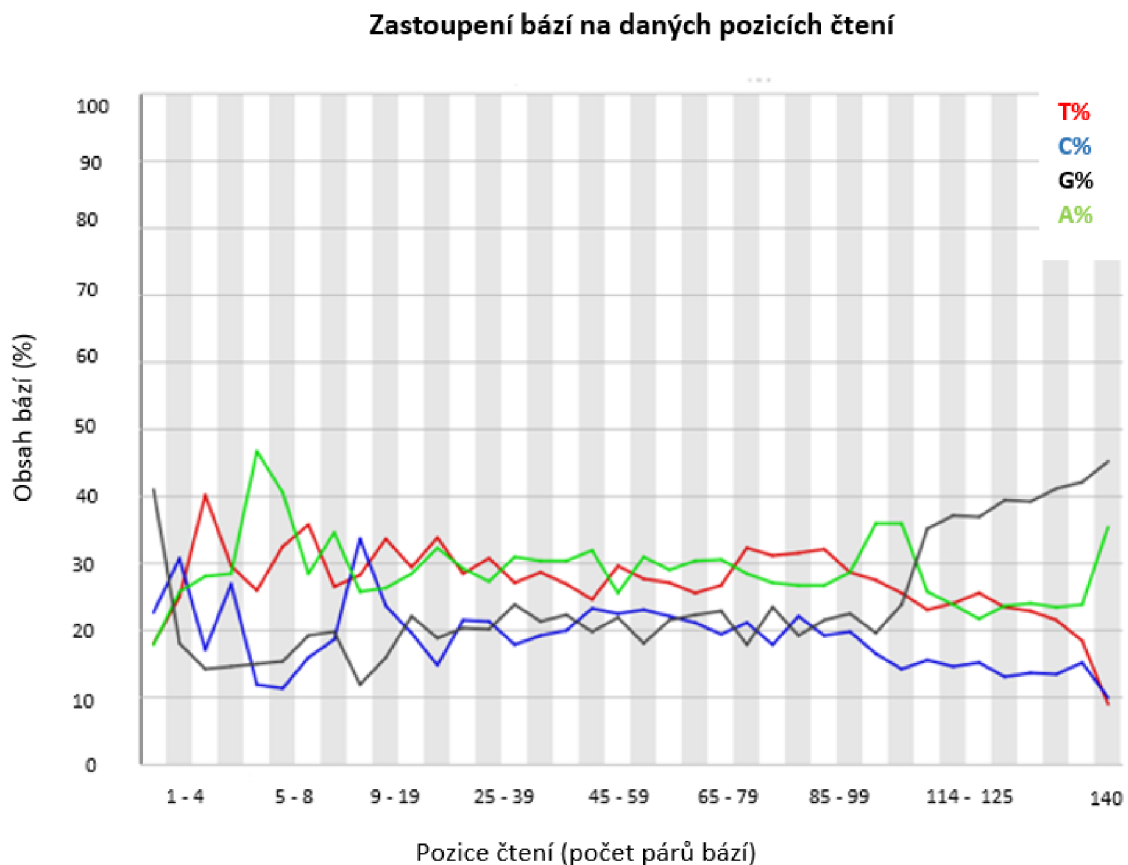
Skóre kvality celé sekvence vyjadřuje počet bází (osa y) s daným průměrným skórem (osa x). Kvalitní vzorky by měly mít vysoké průměrné skóre bez větších výkyvů v nízkých hodnotách skóre, jak je znázorněno na obrázku 4.2.



Obr. 4.2: Průměrné phred skóre kvality sekvence pro vzorek CSQ162

4.1.3 Obsah jednotlivých bází

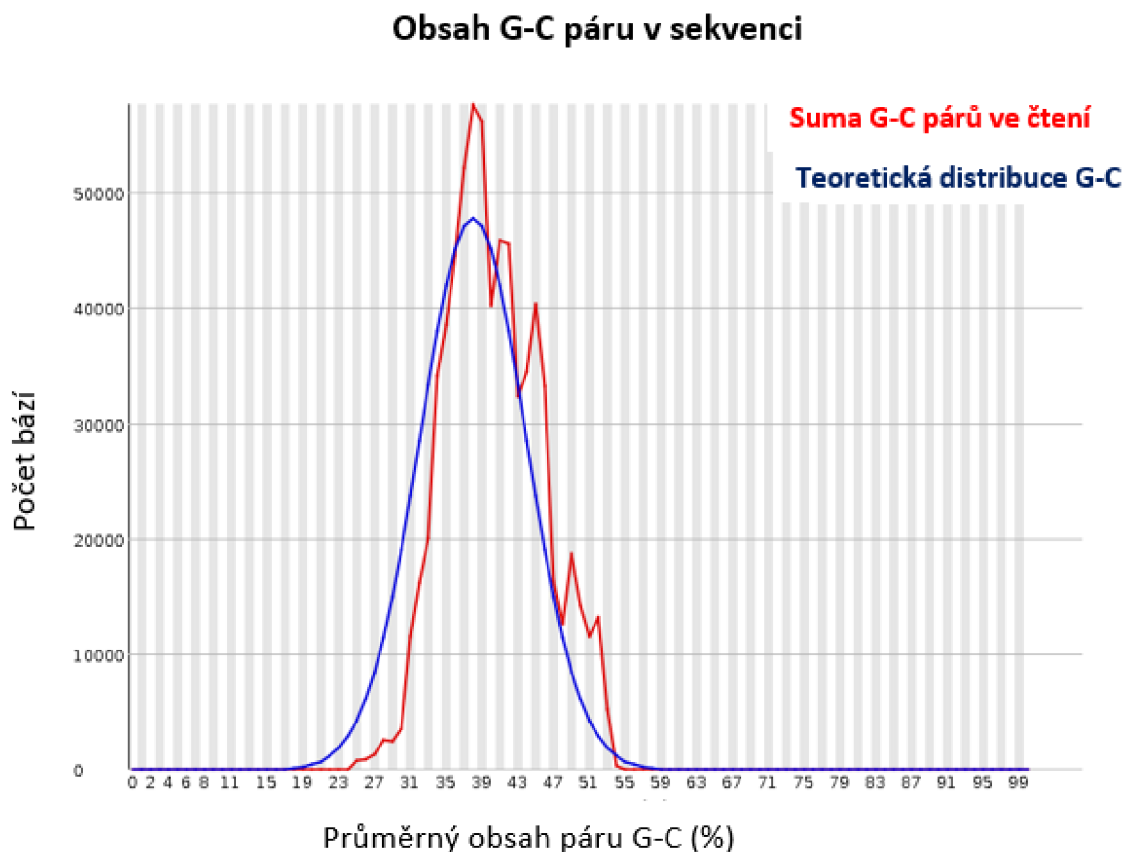
Graf zastoupení bází vyjadřuje zastoupení jednotlivých bází v určitých lokalizacích čtení. Požadujeme ideálně rovnoměrné zastoupení bází bez větších odchylek v poměru A - T ku C - G. Příklad méně optimálního rozložení bází je na obrázku 4.3.



Obr. 4.3: Průměrné procentuální zastoupení bází pro vzorek CSQ0454

4.1.4 Obsah páru G - C

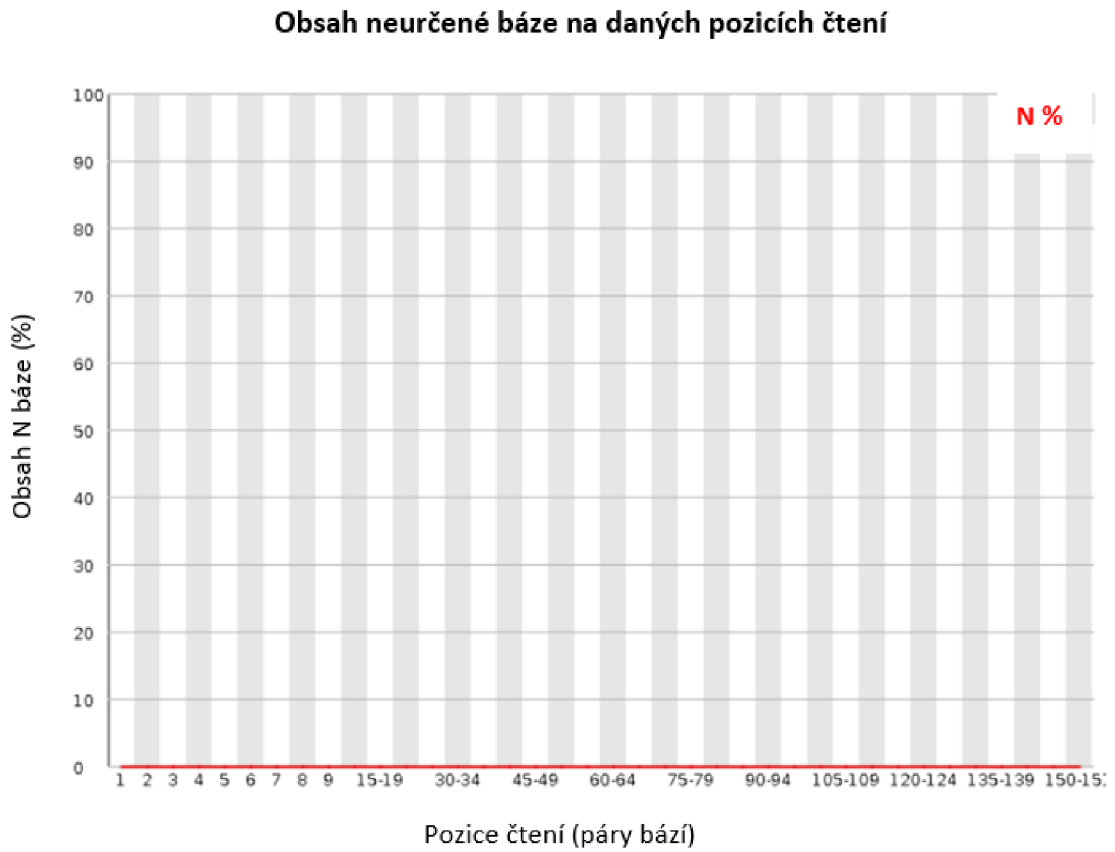
Další statistikou je procento G - C párů po celé délce sekvence. Vyjadřuje procentuální obsah G - C porovnaný s normální distribucí (Gaussovské rozložení), jak je vidět na obrázku 4.4. U kvalitní sekvence součet odchylek od teoretického rozložení není vyšší než 15 % celkového obsahu G - C v sekvenci.



Obr. 4.4: Průměrné procentuální zasoupení G - C párů pro vzorek CSQ0919

4.1.5 Obsah neurčené báze N

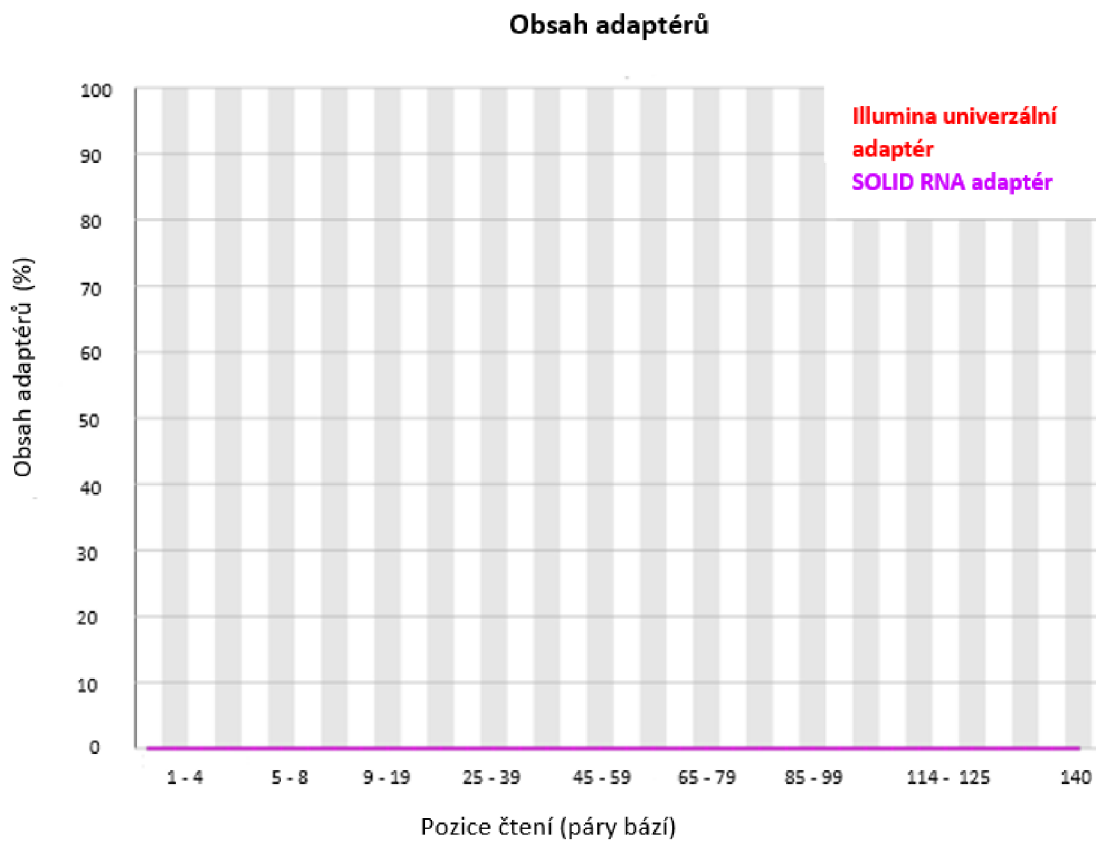
Tato statistika vyjadřuje počet bází, které sekvenátor nebyl schopen určit. V ideálním případě požadujeme nulový výskyt N báze jako je ukázáno na obrázku 4.5.



Obr. 4.5: Průměrné procentuální zasoupení N báze pro vzorek CSQ210

4.1.6 Obsah adaptérů

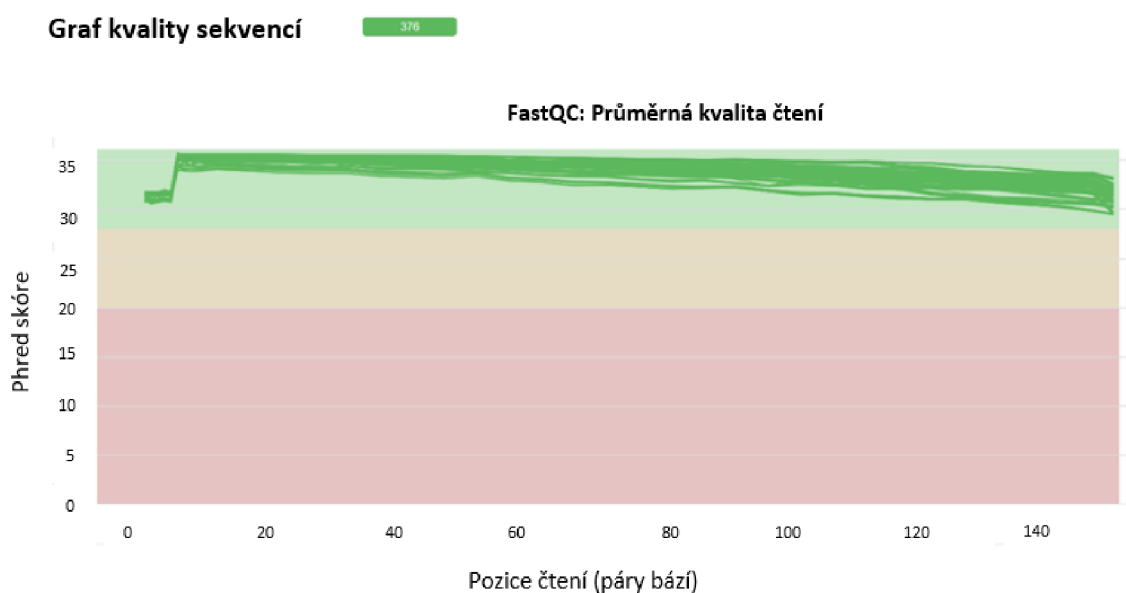
Obsah adaptérů vyjadřuje procentuální zastoupení chemicky syntetizovaných oligonukleotidů navázaných na konce molekul DNA. Ideálně požadujeme jejich obsah nulový. Na obrázku 4.6 vidíme sekvenační vzorek po odstrižení adaptérů.



Obr. 4.6: Průměrné procentuální zasoupení adaptérů pro vzorek CSQ0529

4.1.7 Selekce dat

Pro výběr vzorků k dalším praktickým postupům byl využit nástroj MultiQC (v1.6, [27]), který vytváří souhrn z jednotlivých reportů FastQC. V MultiQC reportu byl zjištěn vysoký obsah adaptérů pro všechny sekvence, proto bylo potřeba data vyčistit nástrojem Trimmomatic (v0.36, [28]). Po odstrižení adaptérů došlo opět k použití FastQC a MultiQC, na základě výstupů z obou programů byla vybrána pouze kvalitní data pro další postup. Data byla vybrána na základě obsahu adaptérů (méně jak 0.1 %) a průměrné phred skóre kvality čtení (vyšší než 27). Odpovídající parametry kvality mělo 188 vzorků, na obrázku 4.7 je znázorněna průměrná kvalita vybraných dat.



Obr. 4.7: Průměrné phred skóre pro kvalitní sekvenační data

4.2 Mapování genomů k referenční sekvenci

Sekvenační data SARs-CoV-2 s dostatečnou kvalitou bylo následně potřeba složit a k tomu účelu bylo vybráno skládání k referenční sekvenci. Jako referenční genom byl použit kompletní genom SARs-CoVs-2 izolovaný ve Wuhanu NC_045512.2 [29]. Pro mapování genomů k referenční sekvenci byl použit nástroj BWA - Burrows-Wheeler Alignment Tool (v0.7.17, [30]). Konkrétněji BWA - MEM, který zarovnává sekvence s maximální možnou shodou. Výsledkem je soubor ve formátu SAM (Sequence Alignment Map). Jedná se o soubor textového formátu, který obsahuje namapované sekvence, informace o referenční sekvenci, délce sekvence, kvalitě mapování. Nástrojem SAMtools (v1.14, [31]) byl soubor ve formátu SAM převeden na jeho binární

(komprimovanou) formu BAM. S využitím dalších funkcí algoritmu SAMtools bylo dosaženo finálního namapování k referenční sekvenci. Následně bylo možné použít Qualimap (v2.2.2, [32]) k vyhodnocení kvality mapování. Z celkových 188 složených vzorků bylo vyfiltrováno 17 sestavených sekvencí, které nesplňovaly kritérium kvality mapování minimálně před skóre 20 či průměrné pokrytí 3000, což je dostatečná hodnota k analýze SNP - jednonukleotidový polymorfismus. Hodnoty pokrytí a kvality mapování pro jednotlivé genomy jsou uvedeny v příloze B. V souhrnném Qualimap reportu vyšlo průměrné pokrytí ze všech namapovaných sekvencí 9232.91 a průměrná kvalita mapování vyšla 59.65. Pro další postup bylo potřeba převést BAM soubor do formátu FASTA. K tomu byl použit nástroj VCFtools (v0.1.16, [33]) a jeho binární forma BCFtools (v1.15, [34]). Použité příkazy z nástroje SAMtools a BCFtools jsou zobrazeny v blokovém schématu A.1.

4.3 Sestavení genomů *de novo*

Tato varianta sestavování nevyužívá referenční sekvenci. Pro sestavení genomu z otrimovaných dat byl použit assembler SPAdes - St. Petersburg genome assembler (v3.15.3, [35]). Pro nástroj SPAdes jako vstup byla použita párová čtení R1/ R2 s výchozím nastavením orientace čtení vpřed-vzad (forward - reverse). Vyhodnocení sestavených genomů proběhlo přes software QUASt - QQuality ASsessment Tool (v5.0.2, [36]). Pomocí softwaru QUASt byly získány základní informace o sestavených genomech mimo jiné i délky kontigů. Tyto kontigy byly velice krátké, proto výsledné genomy sestavené *de novo* nebyly využity pro další kroky praktické části. Výsledky pro některé genomy jsou uvedeny v příloze C.1.

5 Analýza vzorků v závislosti na změnách v genomu v určitých časových úsecích

Pro znalost mutací a možnost odhadu vývoje pandemie je třeba znát časový průběh vývoje SARs-CoV-2. Aby bylo možné provést vyhledání variability v jednotlivých genech, bylo potřeba data anotovat.

5.1 Anotování dat

Anotování dat znamená přesnou definici pozice počátku a konce určitého úseku genetické informace, která kóduje gen. Jelikož data byla namapována k referenční sekvenci, bylo potřeba pouze vyselektovat dané pozice genů. Genom SARs-CoV-2 obsahuje 11 genů. V tabulce 5.1 jsou uvedeny délky a pozice jednotlivých genů s počtem vzorků, které mají kompletně či částečně namapovaný daný gen. Znalost lokací jednotlivých genů je potřeba pro následný překlad do proteinové sekvence a vyhledání mutovaných pozic. Při ponechání dat v nukleotidech by se projevíly i synonymní mutace, které kódují stejnou aminokyselinu. Tyto mutace nemění chování daného organismu, proto nejsou z pohledu variability genu tak zajímavé.

Tab. 5.1: Tabulka genů s počtem kompletně či částečně namapovaných vzorků v daných genech

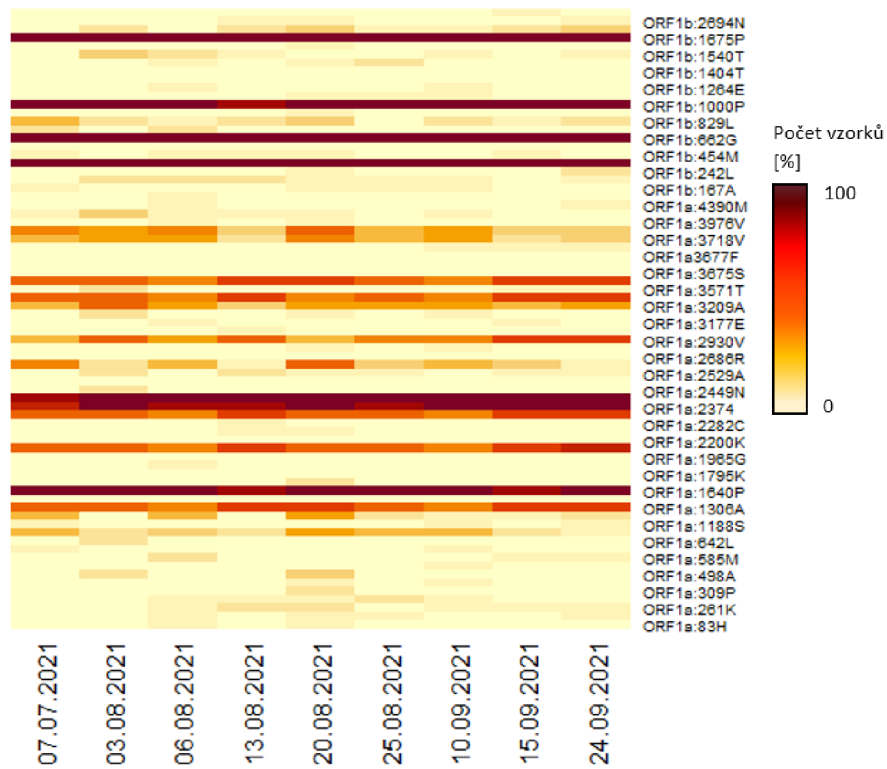
Gen	Začátek	Konec	Délka	Počet kompletně namapovaných vzorků	Počet částečně namapovaných vzorků
ORF1ab	266	21555	21289	0	171
S	21563	25384	3821	0	171
ORF3a	25393	26220	827	0	171
E	26245	26472	227	39	132
M	26523	27191	668	0	171
ORF6	27202	27387	185	0	171
ORF7a	27394	27759	365	0	171
ORF7b	27756	27887	131	15	156
ORF8	27894	28259	365	0	171
N	28274	29533	1259	0	171
ORF10	29558	29674	116	92	79

5.2 Vyhodnocení variability

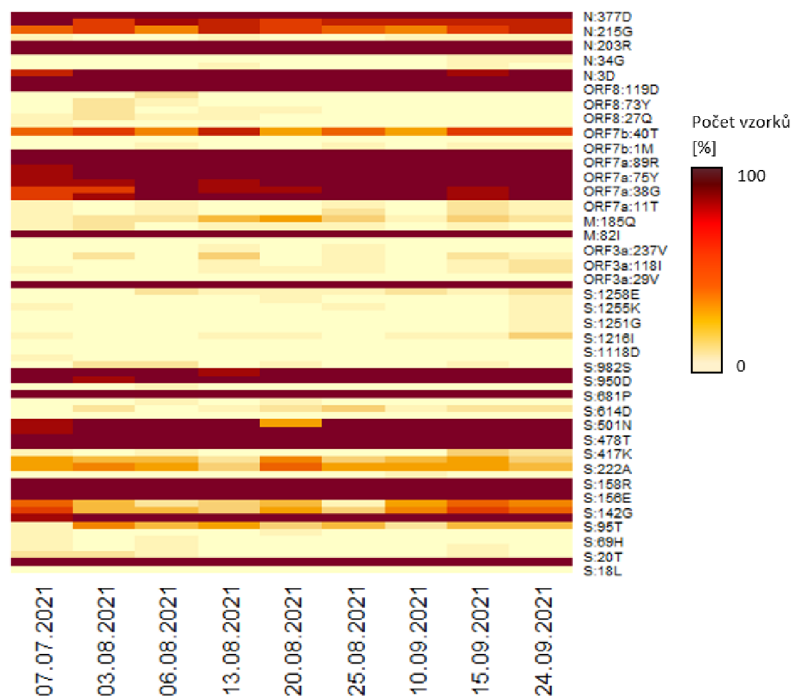
Anotovaná data byla přeložena do aminokyselinových sekvencí podle standardního genetického kódu, u kterého je brán jako iniciační kodon AUG. Přeložené geny jednotlivých vzorků byly porovnány s proteiny v přeložených genech z reference, v kterých se v odpovídajících týdnech odběrů vzorků vyskytly mutace. Tyto pozice byly převzaty od iniciativy COG-CZ (<https://virus.img.cas.cz/>), která mapuje mutace SARs-CoV-2 od počátku výskytu. Pozice kde se nacházely nesynonymní mutace byly zaznamenány a vykresleny formou heatmapy. Vzorky byly seskupeny podle data odběru, jejich počet v odpovídajících datech ukazuje tabulka 5.2. Jednotlivé mutace jsou vykresleny v heatmapách 5.1, 5.2 a označeny ve formátu `gen_v_referenční_sekvenci:pozice:aminokyselina_v_referenční_sekvenci` jako například ORF1a:642L. Nejčtenější počet mutací oproti referenční sekvenci se nachází ve vzorcích odebraných 15.09.2021 a nejméně 10.09.2021. Z heatmap lze určit, že nejvíce mutují pozice 1640 a 2374 v genu ORF1a, 454, 662, 1000 a 1675 v genu ORF1b, 19, 138, 156, 157, 158, 478, 484, 681 a 950 v genu S, 26 v genu ORF3a, 82 v genu M, 26, 38, 61, 75, 82, 89 a 120 v genu ORF7a, 119 a 120 v genu ORF8, 63, 203, 235 a 377 v genu N. Vykreslení počtu mutací v celých genech v rámci každého vzorku je znázorněno na heatmapách 5.3, 5.4, 5.5, 5.6. Celkový počet všech mutací v rámci každého vzorku je v tabulce D. Vizualně lze odvodit, že gen ORF7a je nejčteněji mutován napříč vzorky, narozdíl od toho jsou geny ORF1a a ORF1b nejodolnější vůči mutabilitě. V rámci genu S je možné během časových úseků pozorovat zpětnou mutaci do původní aminokyseliny.

Tab. 5.2: Tabulka počtu vzorků v určitých datech odběru

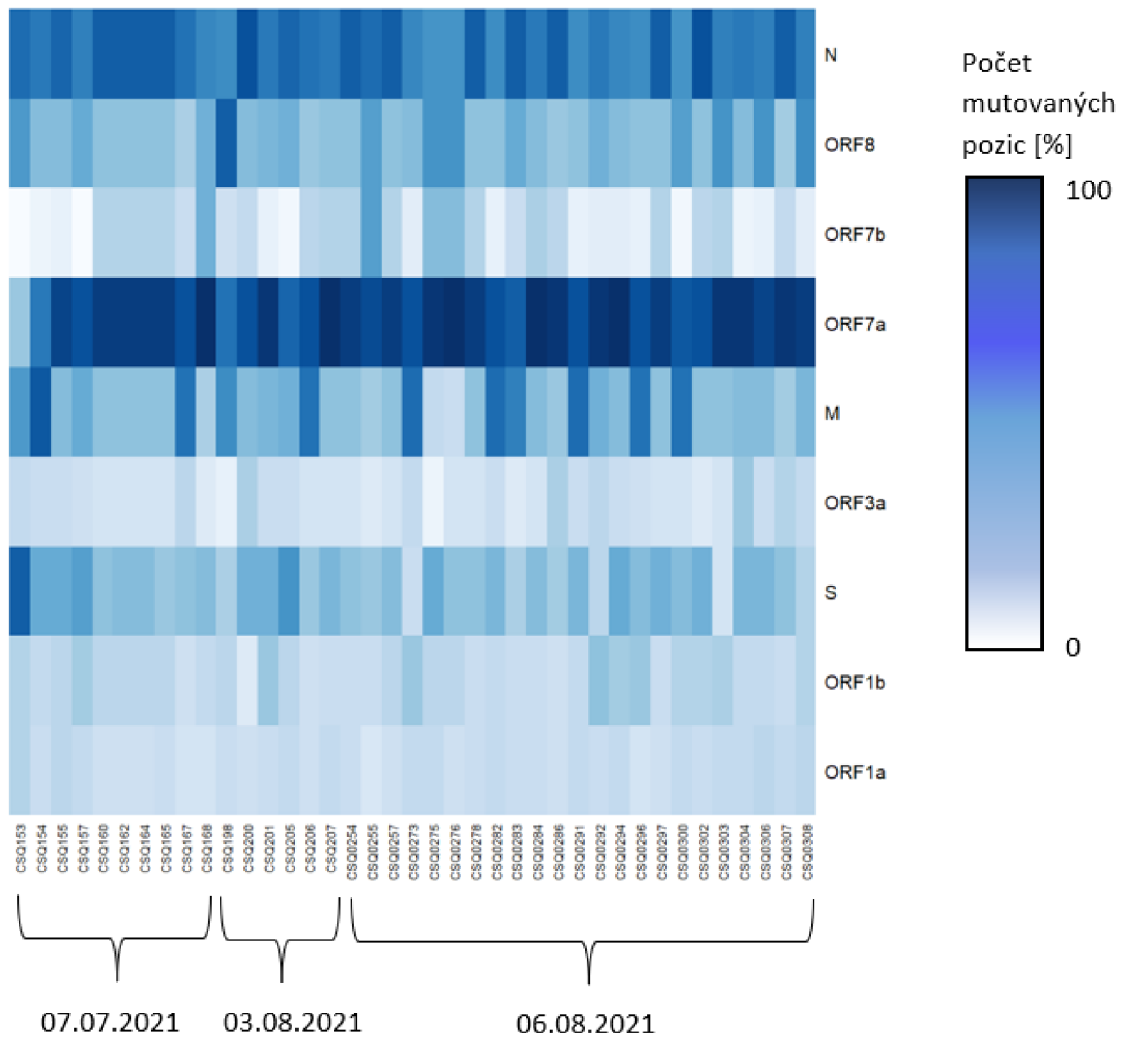
Datum	Počet vzorků
07.07.2021	10
03.08.2021	6
06.08.2021	23
13.08.2021	23
20.08.2021	11
25.08.2021	18
10.09.2021	16
15.09.2021	21
24.09.2021	43



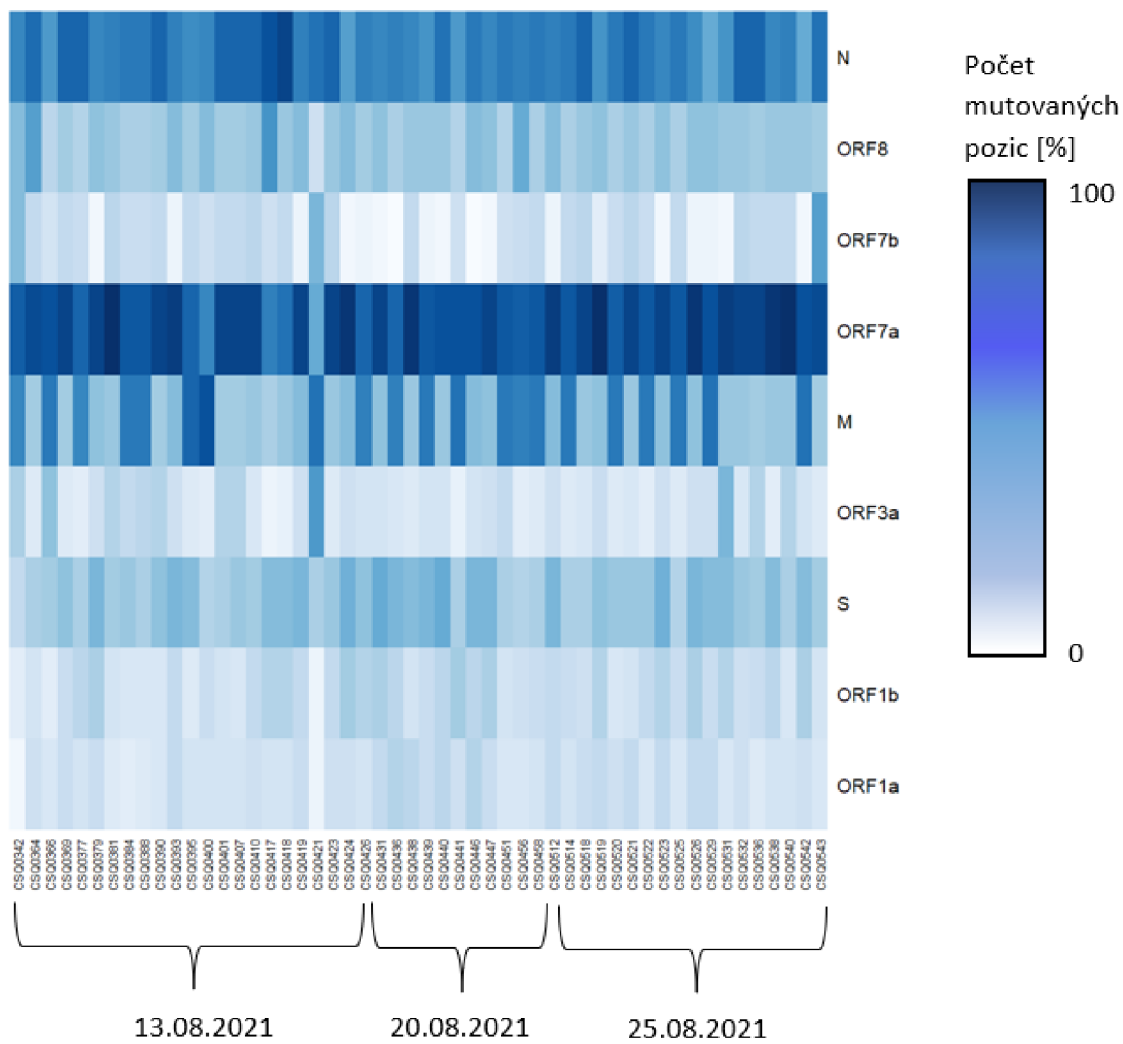
Obr. 5.1: Mutace ve vzorcích podle data odběru v genech ORF1a a ORF1b referenční sekvence



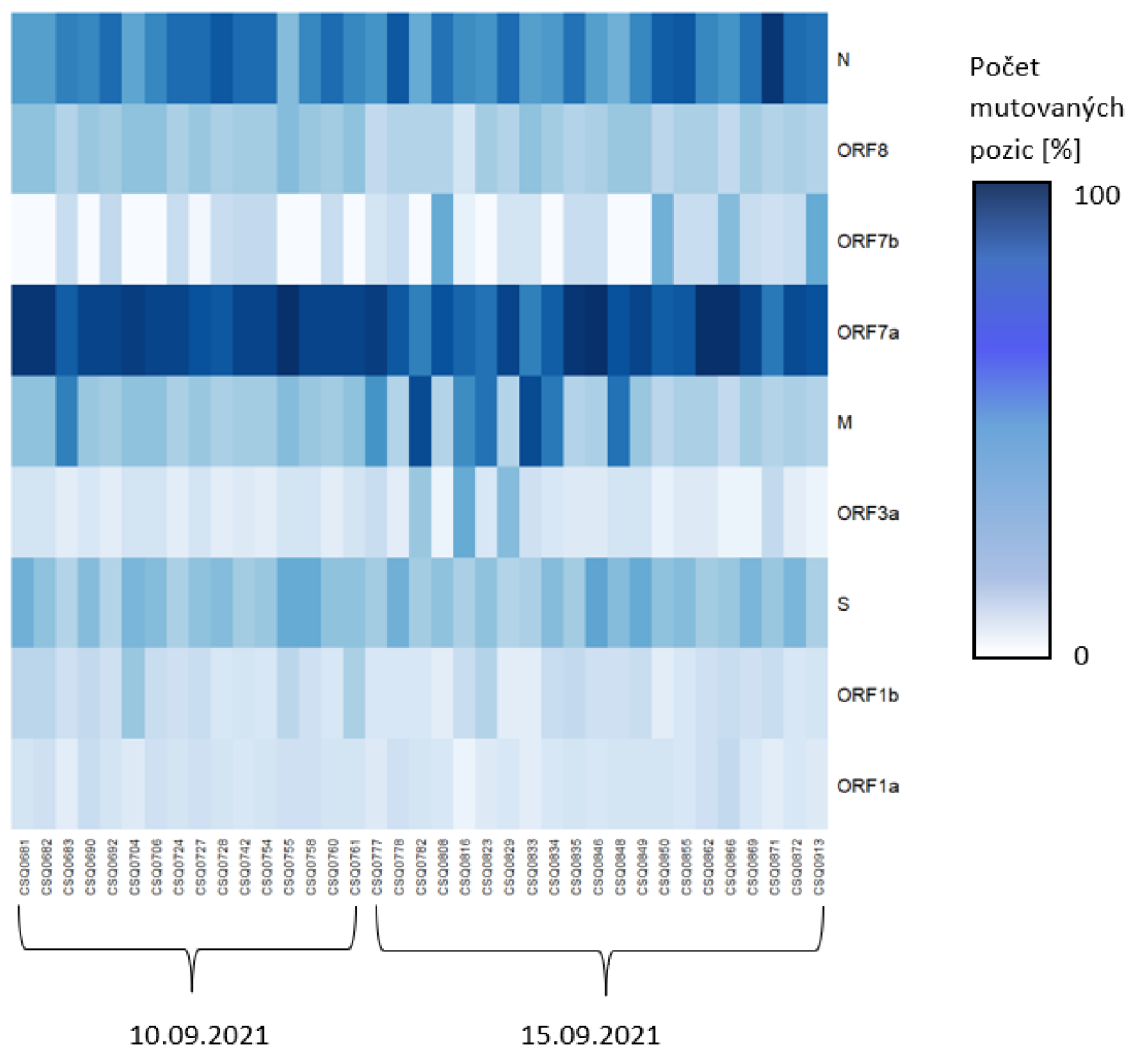
Obr. 5.2: Mutace ve vzorcích podle data odběru v genech S, ORF3a, M, ORF7a, ORF7b, ORF8 a N referenční sekvence



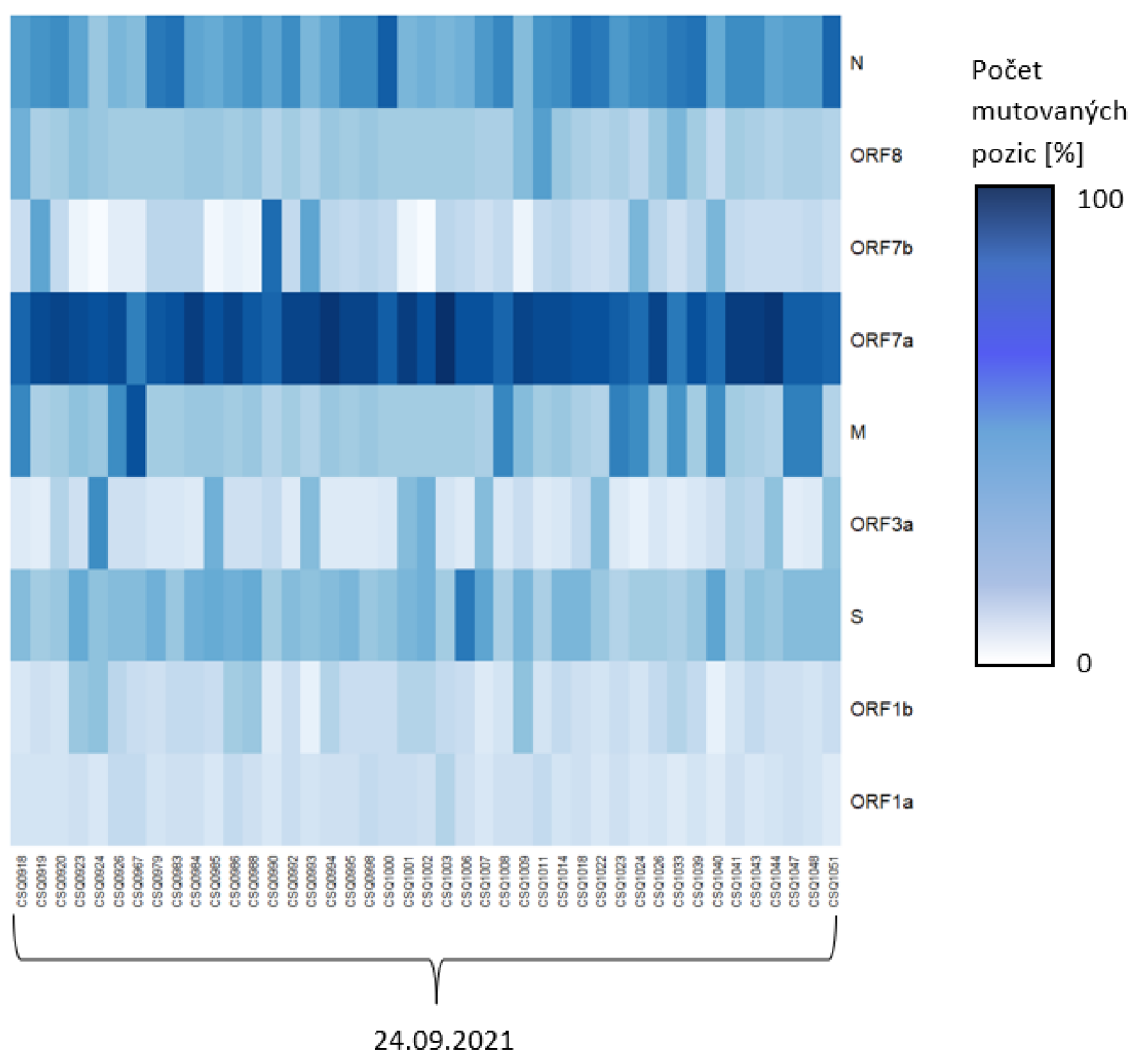
Obr. 5.3: Výskyt mutací v jednotlivých genech pro vzorky odebrané 7.7.2021, 3.8.2021 a 6.8.2021



Obr. 5.4: Výskyt mutací v jednotlivých genech pro vzorky odebrané 13.8.2021, 20.8.2021 a 25.8.2021



Obr. 5.5: Výskyt mutací v jednotlivých genech pro vzorky odebrané 10.9.2021, 15.9.2021

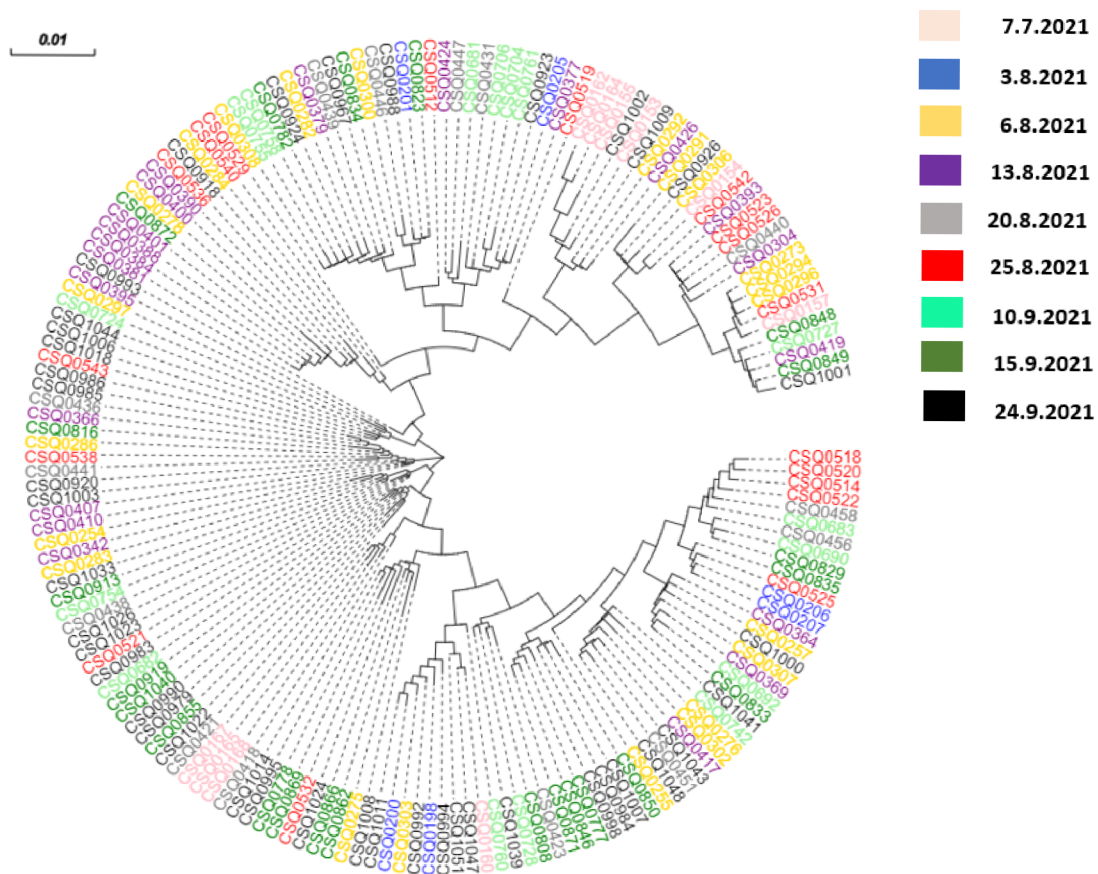


Obr. 5.6: Výskyt mutací v jednotlivých genech pro vzorky odebrané 24.9.2021

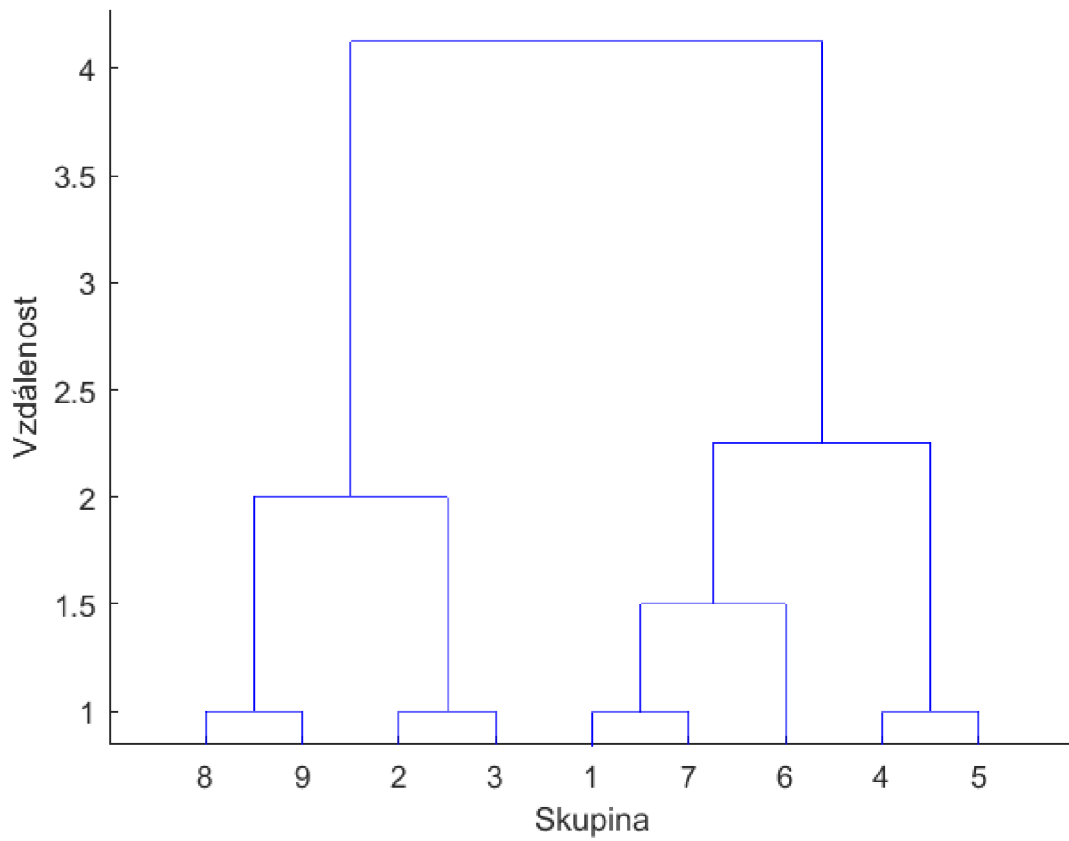
6 **Fylogenetická analýza**

Geny všech vzorků byly zarovnány nástrojem pro vícenásobné zarovnání ClustalW (v2.0., [37]) s výchozím nastavením parametrů penalizace otevření mezery 15 a penalizace prodloužení mezery 6,66. Ze zarovnání byl vytvořen fylogenetický strom v softwaru MEGA X (v1.0., [38]) ve formátu NEWICK. Pro vytvoření fylogenetických stromů byly využity čtyři metody. Metoda Maximální pravděpodobnosti, která porovnává všechny možné topologie fylogenetických stromů z daných sekvencí/ vzorků a určuje věrohodnost na základě délky větví a počtu genetických změn. Metoda Spojování nejbližších sousedů, je založena na matici vzdáleností mezi každým párem koncových větví, pár s nejmenší vzdáleností je spojen do jednoho uzlu a následuje přepočítání vzdáleností všech koncových větví ke vzniklým uzlům, algoritmus pokračuje dokud nejsou spojeny všechny koncové větve. Metoda UPGMA (metoda neváženého párování s aritmetickým průměrem) pracuje s distanční maticí průměrných vzdáleností mezi jednotlivými vzorky, ve které je vyhledána minimální vzdálenost a následně je matice přepočítána vůči spojeným vzorkům dokud nejsou spojeny všechny koncové větve. Metoda minimální evoluce je založena na porovnání všech možných kombinací fylogenetických stromů na základě minimálního součtu vzdáleností mezi koncovými větvemi. Z NEWICK formátů jednotlivých algoritmů byly v online softwaru Evolview (v3.0., [39]) vyobrazeny fylogramy, které vyznačují příbuznost mezi vzorky společnými uzly a různou délkou větví podle evolučního vztahu. Fylogram vytvořený metodou Maximální pravděpodobnosti je na obrázku 6.1, fylogramy získané pomocí jiných metod konstrukce jsou vykresleny v příloze E.1, E.2, E.3. Jak ukazuje fylogram metody Maximální pravděpodobnosti, vzorky odebrané ve stejný datum nevykazují nejbližší vzájemnou podobnost a nevytvářejí nejbližší společné shluky, jak bylo předpokládáno. Tento výsledek je pravděpodobně způsoben malým časovým rozptylem mezi jednotlivými daty odběrů.

Následně bylo z NEWICK formátu v Matlabu (vR2022a, [40]) provedeno hierarchické shlukování do 9 skupin podle počtu časových úseků odběrů vzorků. Výsledky shlukování z fylogenetického stromu metodou Maximální pravděpodobnosti jsou na obrázku 6.2 s vyhodnocením četnosti vzorků podle data v příslušných skupinách ve formě matice záměn v tabulce 6.1. Z kontingenční tabulky je patrné, že jednotlivé skupiny neodpovídají seskupení vzorků na základě stejných dat odběrů. V ideální případě by měla nejvyšší četnosti obsahovat diagonála, avšak jak vidíme např. u skupiny 4, vzorky odebrané 24.09.2021 jsou se stejným zastoupením i ve skupině 1 a 5. Dendrogramy z konstrukcí fylogenetických stromů metodami Minimální evoluce, Nejbližší spojování sousedů a UPGMA jsou vykresleny v příloze F.1, F.2, F.3.



Obr. 6.1: Fylogenetický strom metodou Maximální pravděpodobnosti



Obr. 6.2: Seskupení vzorků z fylogenetického stromu metodou Maximální pravděpodobnosti

Tab. 6.1: Matice záměn skupin hierarchického shlukování z fylogenetického stromu metodou Maximální pravděpodobnosti

	Skupina	1	2	3	4	5	6	7	8	9
Skupina	Datum	25.8. 2021	3.8. 2021	15.9. 2021	24.9. 2021	20.8. 2021	13.8. 2021	10.9. 2021	7.7. 2021	6.8. 2021
1	25.8. 2021	5	0	1	1	2	2	2	1	4
2	3.8. 2021	2	2	0	0	0	0	2	0	0
3	15.9. 2021	8	0	4	3	1	1	3	0	2
4	24.9. 2021	8	5	4	8	8	1	4	2	2
5	20.8. 2021	3	0	0	1	2	0	3	0	1
6	13.8. 2021	4	0	1	2	4	6	4	1	2
7	10.9. 2021	5	1	0	2	1	0	6	0	1
8	7.7. 2021	0	1	3	0	0	0	0	4	2
9	6.8. 2021	5	1	1	1	3	2	3	1	6

Závěr

Náplní předkládané bakalářské práce je vypracování literární rešerše o koronaviru SARS-CoV-2, vyhodnocení kvality sekvenčních dat od pacientů z FN Brno, sestavení osekvenovaných genomů *de novo*, mapováním genomu k referenční sekvenci, vyhledání variability v sestavených datech a jejich následná fylogenetická analýza.

Teoretická část obsahuje základní popis viru SARS-CoV-2, srovnání s ostatními koronaviry a metody sekvenování. V práci je také objasněn způsob jeho přenosu a replikace viru. Teoretická část je uzavřena popisem prozatím známých mutací.

Praktická část je zaměřena na práci se sekvenčními daty, která byla osekvenována na platformě Illumina NextSeq. Prvním úkolem bylo vyhodnotit kvalitu dat poskytnutých z FN Brno. Přes nástroj FastQC byl zjištěn vysoký obsah adaptérů, proto bylo potřeba data upravit pomocí nástroje Trimmomatic, aby došlo k jejich odstrižení. Následně byla data vybrána podle kritéria obsahu adaptérů a před skóre kvality. Sestavení vyselektovaných dat probíhalo dvěma způsoby a to mapováním k referenční sekvenci a sestavení *de novo*.

Pro první způsob jako referenční sekvence byl využit kompletní genom SARS-CoVs-2 izolovaný ve Wuhanu NC_045512.2. Mapování probíhalo pomocí nástroje BWA a k vyhodnocení namapovaných genomů ve formátu BAM byl využit Qualimap. Ze souhrného reportu Qualimap byla zjištěna například průměrná kvalita mapování 59.65 a průměrné pokrytí 9236. Pro sestavení *de novo* byl využit assembler SPAdes a následné vyhodnocení sestavených genomů proběhlo nástrojem Quast. Z Quast analýzy byla zjištěna velice krátká délka kontigů, proto sestavené genomy tímto způsobem nebyly využity pro další kroky této práce.

Kvalitní namapovaná data byla anotována a přeložena do proteinových sekvencí s následným vyhledáním změn v genech oproti referenční sekvenci. Nejvyšší celková mutabilita byla nalezena u vzorků odebraných 15.9.2021 a největší náchylnost k mutacím vykazuje gen ORF7a a N.

Pro fylogenetickou analýzu bylo potřeba data zarovnat nástrojem pro vícenásobné zarovnání ClustalW. Z těchto dat byly sestaveny fylogenetické stromy čtyřmi metodami - UPGMA, Maximální pravděpodobnost, Spojování nejbližších sousedů a Minimální evoluce. Následně byla z fylogenetických stromů provedeno hierarchické shlukování, kde byly vzorky seskupeny do 9 skupin podle počtu dat odběrů. Obě analýzy poukázaly na nepodobnost vzorků v rámci daných skupin podle data odběru, což může být způsobeno nízkým časovým rozptylem mezi jednotlivými daty odběru vzorů.

Literatura

- [1] MURRAY, Patrick R. *Medical microbiology*. 4th ed. St. Louis: Mosby, c2002.
- [2] SAXENA, Shailendra K., ed. *Coronavirus Disease 2019 (COVID-19)*. Singapore: Springer Singapore, 2020. Medical Virology: From Pathogenesis to Disease Control. ISBN 978-981-15-4813-0. DOI 10.1007/978-981-15-4814-7
- [3] Mikroskopický snímek koronaviru SARS-CoV-2. <https://www.vtei.cz/>. Dostupné z: <https://www.vtei.cz/wp-content/uploads/2020/06/Mlejnkoval1.jpg>
- [4] CHEN, Yu, Qianyun LIU a Deyin GUO. *Emerging coronaviruses: Genome structure, replication, and pathogenesis* Journal of Medical Virology. 2020, 92(4), 418-423. ISSN 0146-6615. DOI 10.1002/jmv.25681
- [5] KOYAMA, Takahiko, Daniel PLATT a Laxmi PARIDA. *Variant analysis of SARS-CoV-2 genomes*. Bulletin of the World Health Organization. 2020, 98(7), 495-504. ISSN 0042-9686. DOI 10.2471/BLT.20.253591
- [6] GUO, Sheng, Jing YANG, Yu LEI, Bin LIU, Wei ZHANG, Li ZHANG a Zhili ZUO. *Which species does the virus like most: Binding modes study between SARS-CoV-2 S protein and ACE2 receptor*. Journal of Molecular Graphics and Modelling. 2021, 105. ISSN 10933263. DOI 10.1016/j.jmgm.2021.107893
- [7] CAO, Yipeng, Rui YANG, Wei WANG, et al. *Probing the Formation, Structure and Free Energy Relationships of M Protein Dimers of SARS-CoV-2*. Computational and Structural Biotechnology Journal. 2022. ISSN 20010370. DOI 10.1016/j.csbj.2022.01.007
- [8] DUART, Gerard, Maria J. GARCÍA-MURRIA a Ismael MINGARRO. *The SARS-CoV-2 envelope (E) protein has evolved towards membrane topology robustness*. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2021, 1863(7). ISSN 00052736. DOI 10.1016/j.bbamem.2021.183608
- [9] ZHAO, Huaying, Di WU, Ai NGUYEN, et al. *Energetic and structural features of SARS-CoV-2 N-protein co-assemblies with nucleic acids*. IScience. 2021, 24(6). ISSN 25890042. DOI 10.1016/j.isci.2021.102523
- [10] GUPTA, Ruchir, Jacob CHARRON, Cynthia L. STENGER, et al. *SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome: Insights into functional evolution and human genomics*. Journal of Biological Chemistry. 2020, 295(33), 11742-11753. ISSN 00219258. DOI 10.1074/jbc.RA120.014873

- [11] *Schéma viru SARS-CoV-2. Znázornění čtyř strukturních proteinů: S, M, N a E.* [Http://scielo.sld.cu/](http://scielo.sld.cu/). Dostupné z: <http://scielo.sld.cu/img/revistas/ibi/v39n3//1561-3011-ibi-39-03-e867-gf1.jpg>
- [12] FARRELL, Robert E. *RT-PCR*. RNA Methodologies. Elsevier, 2010, 2010, s. 385-448. ISBN 9780123747273. DOI 10.1016/B978-0-12-374727-3.00018-8
- [13] SCHADENDORF, Dirk. a Beate M. CZARNETZKI. *Gene Amplification by Polymerase Chain Reaction in Dermatology*. Journal of Investigative Dermatology. 1991, 97(5), 751-755. ISSN 0022202X. DOI 10.1111/1523-1747.ep12485015
- [14] *Metoda PCR.* <https://doba-genova.cz>. Dostupné z: https://doba-genova.cz/wp-content/uploads/2019/04/PCR_popisky_10apr2019-768x1024.jpg
- [15] GOMES, Laksiri, Chandima JEEWANDARA, Tibutius Pramanayagam JAYADAS, et al. *Surveillance of SARS-CoV-2 variants of concern by identification of single nucleotide polymorphisms in the spike protein by a multiplex real-time PCR* Journal of Virological Methods. 2022, 300. ISSN 01660934. DOI 10.1016/j.jviromet.2021.114374
- [16] GOMES, Alicia a Bruce KORF. Genetic Testing Techniques. *Pediatric Cancer Genetics*. Elsevier, 2018, 2018, s. 47-64. ISBN 9780323485555. DOI 10.1016/B978-0-323-48555-5.00005-3
- [17] BHOYAR, Rahul C., Vigneshwar SENTHIVEL, Bani JOLLY, Mohamed IMRAN, Abhinav JAIN, Mohit Kumar DIVAKAR, Vinod SCARIA a Sridhar SIVASUBBU. *An optimized, amplicon-based approach for sequencing of SARS-CoV-2 from patient samples using COVIDSeq assay on Illumina MiSeq sequencing platforms* STAR Protocols. 2021, 2(3). ISSN 26661667. DOI 10.1016/j.xpro.2021.100755
- [18] YAN, Yi, Ke WU, Jun CHEN, et al. *Rapid Acquisition of High-Quality SARS-CoV-2 Genome via Amplicon-Oxford Nanopore Sequencing* Virologica Sinica. 2021, 36(5), 901-912. ISSN 1674-0769. DOI 10.1007/s12250-021-00378-8
- [19] LU, Hengyun, Francesca GIORDANO a Zemin NING. *Oxford Nanopore MinION Sequencing and Genome Assembly*. Genomics, Proteomics Bioinformatics 2016, 14(5), 265-279. ISSN 16720229. DOI 10.1016/j.gpb.2016.05.004

- [20] BRINKMANN, Annika, Sophie-Luisa ULM, Steven UDDIN, et al. *AmpliCoV: Rapid Whole-Genome Sequencing Using Multiplex PCR Amplification and Real-Time Oxford Nanopore MinION Sequencing Enables Rapid Variant Identification of SARS-CoV-2* *Frontiers in Microbiology*. 2021, 12. ISSN 1664-302X. DOI 10.3389/fmicb.2021.651151
- [21] PEACOCK, Sharon J. *SARS-CoV-2 Variants: Past, Present and Future* YANO, Makoto, Fumihiko MATSUDA, Anavaj SAKUNTABHAI a Shigeru HIROTA, ed. *Socio-Life Science and the COVID-19 Outbreak*. Singapore: Springer Singapore, 2022, 2022-12-01, s. 3-23. *Economics, Law, and Institutions in Asia Pacific*. ISBN 978-981-16-5726-9. DOI 10.1007/978-981-16-5727-6_1
- [22] CHOUDHARY, Om Prakash, Manish DHAWAN a PRIYANKA. *Omicron variant (B.1.1.529) of SARS-CoV-2: Threat assessment and plan of action* *International Journal of Surgery*. 2022, 97. ISSN 17439191. DOI 10.1016/j.ijisu.2021.106187
- [23] KANNAN, Saathvik R., Austin N. SPRATT, Kalicharan SHARMA, Hitendra S. CHAND, Siddappa N. BYRAREDDY a Kamal SINGH. *Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies* *Journal of Autoimmunity*.. 2022, 126. ISSN 08968411. DOI 10.1016/j.jaut.2021.102779
- [24] HOWES, Laura. *COVID-19: WHAT YOU NEED TO KNOW ABOUT SARS-COV-2 VARIANTS*. Dostupné z: <https://cen.acs.org/sections/new-covid-19-variants-vaccines-effective.html>
- [25] *Omicron, the New Covid-19 Variant: What to Know*. Dostupné z: <https://www.wsj.com/articles/omicron-coronavirus-variant-what-to-know-11637935500>
- [26] Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Dostupné z: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- [27] EWELS, Philip, Måns MAGNUSSON, Sverker LUNDIN a Max KÄLLER. *MultiQC: summarize analysis results for multiple tools and samples in a single report*. *Bioinformatics*. 2016, 32(19), 3047-3048. ISSN 1367-4803. DOI 10.1093/bioinformatics/btw354
- [28] BOLGER, Anthony M., Marc LOHSE a Bjoern USADEL. *Trimmomatic: a flexible trimmer for Illumina sequence data*. *Bioinformatics*.. 2014, 30(15), 2114-2120. ISSN 1460-2059. DOI 10.1093/bioinformatics/btu170

- [29] WU, Fan, Su ZHAO, Bin YU, et al. *A new coronavirus associated with human respiratory disease in China*. *Nature*. 2020, 579(7798), 265-269. ISSN 0028-0836. DOI 10.1038/s41586-020-2008-3
- [30] LI, H. a R. DURBIN. *Fast and accurate short read alignment with Burrows-Wheeler transform*. *Bioinformatics*. 2009, 25(14), 1754-1760. ISSN 1367-4803. DOI 10.1093/bioinformatics/btp324
- [31] LI, H., B. HANDSAKER, A. WYSOKER, et al. *The Sequence Alignment/Map format and SAMtools*. *Bioinformatics*. 2009, 25(16), 2078-2079. ISSN 1367-4803. DOI 10.1093/bioinformatics/btp352
- [32] GARCÍA-ALCALDE, Fernando, Konstantin OKONECHNIKOV, José CARBONELL, et al. *Qualimap: evaluating next-generation sequencing alignment data*. *Bioinformatics*. 2012, 28(20), 2678-2679. ISSN 1460-2059. DOI 10.1093/bioinformatics/bts503
- [33] DANECEK, P., A. AUTON, G. ABECASIS, et al. *The variant call format and VCFtools*. *Bioinformatics*. 2011, 27(15), 2156-2158. ISSN 1367-4803. DOI:10.1093/bioinformatics/btr330
- [34] DANECEK, Petr, James K BONFIELD, Jennifer LIDDLE, et al. *Twelve years of SAMtools and BCFtools*. *GigaScience*. 2021, 10(2). ISSN 2047-217X. DOI 10.1093/gigascience/giab008
- [35] PRJIBELSKI, Andrey, Dmitry ANTIPOV, Dmitry MELESHKO, Alla LAPIDUS a Anton KOROBAYNIKOV. *Using SPAdes De Novo Assembler*. *Current Protocols in Bioinformatics*. 2020, 70(1). ISSN 1934-3396. DOI 10.1002/cpbi.102
- [36] GUREVICH, Alexey, Vladislav SAHELIEV, Nikolay VYAHHI a Glenn TESLER. *QUAST: quality assessment tool for genome assemblies*. *Bioinformatics*. 2013, 29(8), 1072-1075. ISSN 1460-2059. DOI 10.1093/bioinformatics/btt086
- [37] LARKIN, M.A., G. BLACKSHIELDS, N.P. BROWN, et al. *Clustal W and Clustal X version 2.0*. *Bioinformatics*. 2007, 23(21), 2947-2948. ISSN 1367-4803. DOI 10.1093/bioinformatics/btm404
- [38] KUMAR, Sudhir, Glen STECHER, Michael LI, Christina KNYAZ, Koichiro TAMURA a Fabia Ursula BATTISTUZZI. *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms*. *Molecular Biology and Evolution*. 2018, 35(6), 1547-1549. ISSN 0737-4038. DOI 10.1093/molbev/msy096

- [39] SUBRAMANIAN, Balakrishnan, Shenghan GAO, Martin J LERCHER, Songnian HU a Wei-Hua CHEN. *Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees*. Nucleic Acids Research. 2019, 47(W1), W270-W275. ISSN 0305-1048. DOI 10.1093/nar/gkz357
- [40] SOBIE, Eric A. *An Introduction to MATLAB*. Science Signaling. 2011, 4(191). ISSN 1945-0877. DOI 10.1126/scisignal.2001984

Seznam symbolů a zkratek

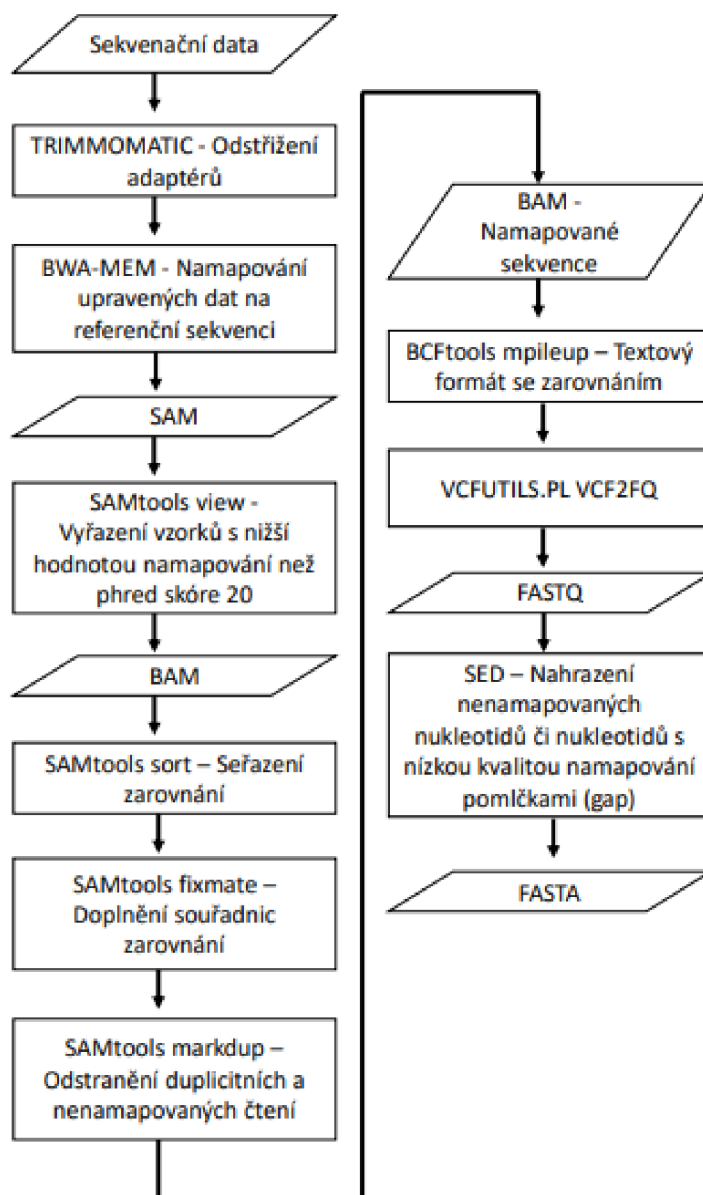
ACE2	angiotensin-konvertující enzym ²
CoVs	koronaviry
DNA	deoxyribonukleová kyselina
MSA	Multiple Sequence Alignment – vícenásobné zarovnání
nsp	nestrukturní protein kódovaný virem, ale není jeho součástí, využit pro virovou replikaci
nsSNP	nesynonymní polymorfismy
ORF	otevřený čtecí rámec
PCR	polymerázová řetězová reakce
RDB	receptorová vazebná doména
RNA	ribonukleová kyselina
RTC	komplex replikace-transkripce
RT-PCR	real-time polymerázová řetězová reakce
RT-qPCR	polymerázová řetězová reakce s reverzní transkripcí
SARs-CoV	koronavirus související s těžkým akutním respiračním syndromem
SNP	single nucleotide polymorphisms/ jednonukleotidový polymorfismus
WHO	světová zdravotnická organizace

Seznam příloh

A	Blokové schéma mapování sekvenačních dat k referenční sekvenci	50
B	Výsledky mapování k referenční sekvenci	51
C	Výsledky sestavení <i>de novo</i>	54
D	Počet mutací v daných genech u jednotlivých vzorků	55
E	Fylogramy různých metod konstrukce fylogenetických stromů	60
F	Výsledky shlukové analýzy z NEWICK formátů uvedených metod	62

A Blokové schéma mapování sekvenačních dat k referenční sekvenci

Níže je uvedené blokové schéma vyjadřující postup při úpravě sekvenačních dat, jejich následnému namapování k referenční sekvenci a převedení namapovaných BAM souborů do souborů FASTA.



Obr. A.1: Blokové schéma úpravy sekvenačních dat, mapování vzorků k referenční sekvenci a následné převedení do FASTA formátu

B Výsledky mapování k referenční sekvenci

Níže jsou uvedeny tabulky výsledků z vyhodnocení kvality sestavených genomů mapováním k referenční sekvenci.

Tab. B.1: Tabulka výsledků všech mapovaných genomů

Vzorek	Pokrytí	Kvalita mapování	Vzorek	Pokrytí	Kvalita mapování
CSQ0254	15693,680	59,967	CSQ0431	10833,151	59,967
CSQ0255	12275,656	59,969	CSQ0436	9679,528	59,967
CSQ0257	15015,310	59,971	CSQ0438	9424,207	59,968
CSQ0273	14151,483	59,967	CSQ0439	10906,098	59,966
CSQ0275	11310,649	59,814	CSQ0440	10586,263	59,964
CSQ0276	10850,957	59,814	CSQ0441	8096,072	59,971
CSQ0278	12643,919	59,872	CSQ0446	12849,523	59,969
CSQ0282	12559,486	59,969	CSQ0447	11516,780	59,965
CSQ0283	13215,079	59,965	CSQ0451	11213,474	59,964
CSQ0284	14690,235	59,969	CSQ0456	9692,396	59,970
CSQ0286	13732,297	59,969	CSQ0458	9525,181	59,969
CSQ0291	14029,398	59,969	CSQ0512	12482,802	59,962
CSQ0292	8433,114	59,966	CSQ0514	10159,715	59,965
CSQ0294	13034,143	59,965	CSQ0518	10508,337	59,970
CSQ0296	13736,564	59,967	CSQ0519	9679,085	59,942
CSQ0297	12732,416	59,967	CSQ0520	11039,540	59,967
CSQ0300	9320,829	59,965	CSQ0521	10019,159	59,968
CSQ0302	12842,321	59,969	CSQ0522	10449,087	59,968
CSQ0303	13865,662	59,967	CSQ0523	10042,735	59,964
CSQ0304	13572,019	59,965	CSQ0525	10183,644	59,966
CSQ0306	13356,782	59,963	CSQ0526	10969,396	59,965
CSQ0307	13428,306	59,968	CSQ0529	11045,992	59,966
CSQ0308	13513,908	59,964	CSQ0531	9947,227	59,967

Vzorek	Pokrytí	Kvalita mapování	Vzorek	Pokrytí	Kvalita mapování
CSQ0364	12273,753	59,964	CSQ0532	8840,017	59,966
CSQ0366	10436,403	59,970	CSQ0536	10184,887	59,967
CSQ0369	11380,924	59,964	CSQ0538	9611,533	59,967
CSQ0374	2,341	43,144	CSQ0540	10437,233	59,967
CSQ0377	11008,144	59,963	CSQ0542	10729,617	59,968
CSQ0379	11462,023	59,964	CSQ0543	11915,363	59,966
CSQ0380	0,605	18,795	CSQ0681	13140,792	59,966
CSQ0381	10470,767	59,965	CSQ0682	17598,421	59,964
CSQ0384	9524,092	59,964	CSQ0683	18004,843	59,966
CSQ0388	11540,637	59,964	CSQ0690	17183,844	59,968
CSQ0390	10872,014	59,961	CSQ0692	19472,757	59,966
CSQ0393	9591,133	59,964	CSQ0704	16133,573	59,968
CSQ0395	9839,718	59,966	CSQ0706	15583,872	59,967
CSQ0400	10582,062	59,969	CSQ0724	13767,541	59,968
CSQ0401	11838,838	59,964	CSQ0727	14311,341	59,967
CSQ0407	11416,141	59,967	CSQ0728	14174,024	59,969
CSQ0410	11310,679	59,966	CSQ0742	13283,956	59,968
CSQ0417	11246,156	59,969	CSQ0754	13615,051	59,970
CSQ0418	10209,541	59,968	CSQ0755	13498,928	59,970
CSQ0419	9396,897	59,967	CSQ0758	14405,651	59,969
CSQ0421	8945,940	59,954	CSQ0760	14913,967	59,969
CSQ0423	10186,005	59,967	CSQ0761	15033,001	59,970
CSQ0424	14427,831	59,961	CSQ0777	3244,644	59,928
CSQ0426	9584,006	59,967	CSQ0778	3004,986	59,963
CSQ0782	16381,930	59,965	CSQ1000	3815,326	59,972
CSQ0787	2920,257	59,960	CSQ1001	3739,196	59,971
CSQ0800	2927,763	59,965	CSQ1002	3711,828	59,970
CSQ0801	2695,529	59,970	CSQ1003	3359,866	59,972
CSQ0804	2754,851	59,964	CSQ1006	3304,392	59,972
CSQ0808	3939,235	59,964	CSQ1007	4146,720	59,973
CSQ0816	3835,654	59,965	CSQ1008	3466,928	59,969
CSQ0823	6537,788	59,969	CSQ1009	3604,946	59,969
CSQ0829	6041,321	59,963	CSQ1011	3869,243	59,971
CSQ0833	4969,954	59,970	CSQ1012	1809,077	59,968
CSQ0834	4963,090	59,964	CSQ1013	2928,606	59,970
CSQ0835	5316,785	59,970	CSQ1014	6846,509	59,969
CSQ0846	5946,005	59,970	CSQ1016	2617,057	59,973

Vzorek	Pokrytí	Kvalita mapování	Vzorek	Pokrytí	Kvalita mapování
CSQ0848	5544,316	59,970	CSQ1017	2936,248	59,969
CSQ0849	4901,859	59,969	CSQ1018	3218,726	59,968
CSQ0850	3729,952	59,810	CSQ1022	3567,069	59,972
CSQ0855	5265,584	59,970	CSQ1023	3389,047	59,968
CSQ0862	5277,543	59,967	CSQ1024	3553,591	59,970
CSQ0866	4506,677	59,809	CSQ1026	5581,422	59,967
CSQ0869	4008,958	59,967	CSQ1031	2360,391	59,965
CSQ0871	5320,191	59,973	CSQ1033	4116,393	59,971
CSQ0872	5377,496	59,971	CSQ1039	4104,907	59,970
CSQ0913	3693,220	59,973	CSQ1040	3097,667	59,964
CSQ0918	4022,282	59,965	CSQ1041	3606,537	59,968
CSQ0919	5550,017	59,961	CSQ1043	3565,557	59,970
CSQ0920	3948,337	59,972	CSQ1044	4321,115	59,970
CSQ0922	2612,143	59,971	CSQ1045	2051,839	59,968
CSQ0923	3911,671	59,974	CSQ1047	4480,739	59,969
CSQ0924	4380,736	59,973	CSQ1048	4181,334	59,967
CSQ0926	4288,626	59,967	CSQ1051	4805,201	59,969
CSQ0967	3657,907	59,968	CSQ153	27033,218	59,964
CSQ0968	2868,961	59,967	CSQ154	29482,380	59,966
CSQ0979	3882,617	59,968	CSQ155	26801,731	59,968
CSQ0983	3408,672	59,974	CSQ157	26035,751	59,966
CSQ0984	3285,067	59,970	CSQ160	25722,458	59,962
CSQ0985	7031,230	59,972	CSQ162	22586,008	59,963
CSQ0986	3355,878	59,970	CSQ164	16846,027	59,964
CSQ0987	2623,149	59,961	CSQ165	27174,935	59,964
CSQ0988	4363,049	59,970	CSQ167	28332,570	59,963
CSQ0990	4245,111	59,964	CSQ168	25292,112	59,945
CSQ0991	2658,139	59,969	CSQ198	9458,833	59,976
CSQ0992	3705,969	59,974	CSQ200	9658,160	59,968
CSQ0993	3167,475	59,964	CSQ201	10680,583	59,968
CSQ0994	3459,016	59,972	CSQ205	11608,128	59,970
CSQ0995	4272,845	59,973	CSQ206	13278,426	59,968
CSQ0996	2091,075	59,971	CSQ207	11157,486	59,971
CSQ0998	3514,520	59,969	CSQ0342	9103,694	59,968

C Výsledky sestavení *de novo*

Níže je uvedena tabulka výsledků pro některé genomy z vyhodnocení kvality sestavených genomů *de novo*.

Tab. C.1: Přehled vybraných výsledků sestavených genomů *de novo*

Vzorek	Celková délka kontigů	Nejdelší kontig	N50	L50
CSQ0255	1447	920	920	1
CSQ0306	1903	753	630	2
CSQ0369	2370	917	785	2
CSQ0458	3349	1002	723	2
CSQ0531	5692	1002	673	4
CSQ0816	26808	1233	340	23
CSQ0920	136438	769	102	448
CSQ0990	206811	717	99	752
CSQ1006	37607	1063	247	40
CSQ1044	672	687	100	672

D Počet mutací v daných genech u jednotlivých vzorků

Níže jsou uvedeny tabulky počtu mutací v daných genech u jednotlivých vzorků.

Tab. D.1: Počet mutací ve vzorcích v jednotlivých genech

Vzorek	ORF1a	ORF1b	S	ORF3a	M	ORF7a	ORF7b	ORF8	N
CSQ0153	9	4	18	1	1	2	0	2	4
CSQ0154	8	4	15	1	2	5	0	2	5
CSQ0155	9	4	16	1	1	7	0	2	6
CSQ0157	9	5	14	1	1	5	0	2	4
CSQ0160	9	5	13	1	1	7	1	2	6
CSQ0162	9	5	14	1	1	7	1	2	6
CSQ0164	9	5	14	1	1	7	1	2	6
CSQ0165	9	5	12	1	1	7	1	2	6
CSQ0167	10	5	15	2	2	7	1	2	6
CSQ0168	9	6	17	1	1	9	2	3	6
CSQ0198	15	7	15	1	2	7	1	5	6
CSQ0200	12	4	14	2	1	5	1	2	5
CSQ0201	7	6	14	1	1	7	0	2	5
CSQ0205	8	4	16	1	1	5	0	2	5
CSQ0206	9	4	13	1	2	7	1	2	6
CSQ0207	11	4	14	1	1	7	1	2	5
CSQ0254	10	4	13	1	1	7	1	2	6
CSQ0255	9	5	14	1	1	7	2	3	6
CSQ0257	9	5	14	1	1	7	1	2	6
CSQ0273	8	6	13	1	2	7	0	2	5
CSQ0275	18	8	22	1	1	9	2	4	6
CSQ0276	16	8	19	2	1	9	2	4	6
CSQ0278	10	4	13	1	1	7	1	2	6
CSQ0282	9	4	15	1	2	7	0	2	5

Vzorek	ORF1a	ORF1b	S	ORF3a	M	ORF7a	ORF7b	ORF8	N
CSQ0283	13	5	13	1	2	7	1	3	7
CSQ0284	9	4	13	1	1	7	1	2	5
CSQ0286	10	4	12	2	1	7	1	2	6
CSQ0291	9	4	15	1	2	7	0	2	5
CSQ0292	6	6	14	1	1	7	0	2	5
CSQ0294	8	6	16	1	1	8	0	2	5
CSQ0296	6	7	14	1	2	7	0	2	5
CSQ0297	9	4	16	1	1	7	1	2	6
CSQ0300	11	6	15	1	2	7	0	3	5
CSQ0302	11	6	15	1	1	6	1	2	6
CSQ0303	11	6	14	1	1	7	1	3	5
CSQ0304	9	4	14	2	1	7	0	2	5
CSQ0306	10	4	14	1	1	7	0	3	5
CSQ0307	13	5	14	2	1	7	1	2	6
CSQ0308	9	4	18	1	1	7	0	3	5
CSQ0342	11	6	15	3	2	7	2	3	6
CSQ0364	10	4	12	1	1	7	1	3	6
CSQ0366	12	4	15	3	2	7	1	2	5
CSQ0369	12	5	14	1	1	7	1	2	6
CSQ0377	11	7	14	1	2	7	1	2	7
CSQ0379	9	6	15	1	1	7	0	2	5
CSQ0381	8	4	12	2	1	7	1	2	5
CSQ0384	8	4	15	2	2	7	1	2	6
CSQ0388	9	4	13	2	2	7	1	2	6
CSQ0390	9	4	15	2	1	7	1	2	6
CSQ0393	9	4	14	1	1	7	0	2	5
CSQ0395	9	3	15	1	2	6	1	2	5
CSQ0400	12	6	14	1	3	7	1	3	7
CSQ0401	10	4	12	2	1	7	1	2	6
CSQ0407	10	4	14	2	1	7	1	2	6
CSQ0410	10	5	12	1	1	7	1	2	6
CSQ0417	12	7	15	1	1	5	1	3	6
CSQ0418	10	6	14	1	1	5	1	2	6
CSQ0419	8	4	14	1	1	7	0	2	5
CSQ0421	19	8	24	6	3	7	3	3	9
CSQ0423	10	5	13	1	1	7	1	2	6
CSQ0424	8	6	15	1	1	7	0	2	4
CSQ0426	7	6	15	1	2	7	0	2	6
CSQ0431	11	6	17	1	1	7	0	2	5

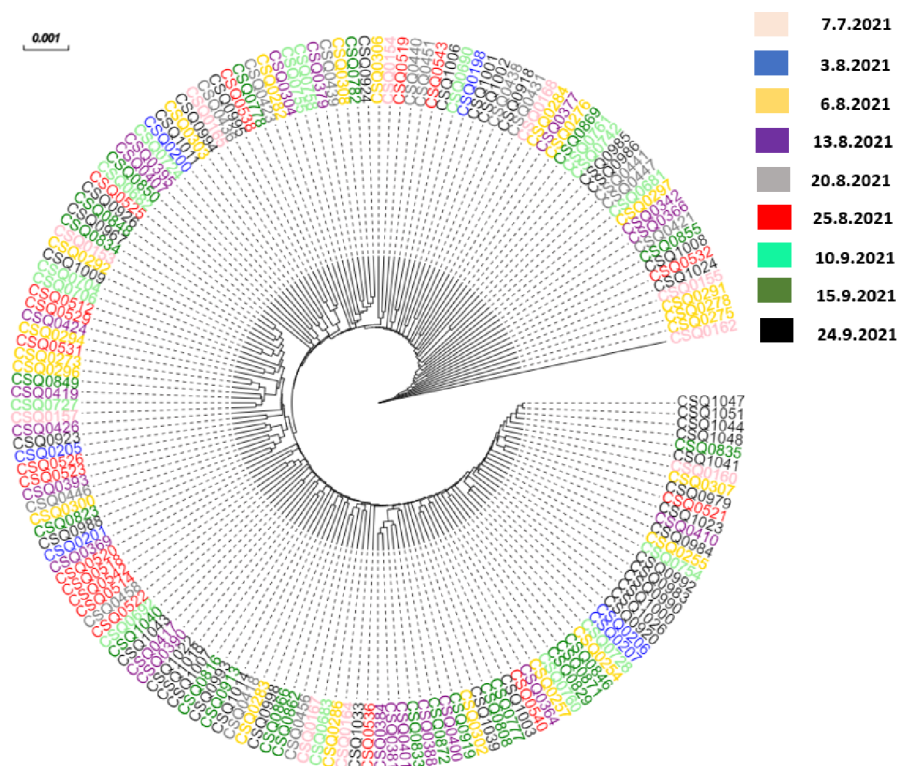
Vzorek	ORF1a	ORF1b	S	ORF3a	M	ORF7a	ORF7b	ORF8	N
CSQ0436	15	6	17	1	2	7	0	2	6
CSQ0438	14	4	15	1	1	7	1	2	5
CSQ0439	9	4	16	1	2	7	0	2	5
CSQ0440	12	5	17	1	1	7	0	2	6
CSQ0441	13	8	14	1	2	7	1	2	5
CSQ0446	13	5	14	1	1	6	0	2	5
CSQ0447	9	6	15	1	1	7	0	2	5
CSQ0451	11	5	14	2	2	7	1	2	6
CSQ0456	10	5	13	1	2	7	1	3	6
CSQ0458	10	5	13	1	2	7	1	2	6
CSQ0512	9	4	14	1	1	7	0	2	5
CSQ0514	10	5	13	1	2	7	1	2	6
CSQ0518	10	4	11	1	1	7	1	2	6
CSQ0519	9	5	14	1	1	9	0	2	5
CSQ0520	11	4	15	1	2	7	1	2	6
CSQ0521	12	4	13	1	1	7	1	2	6
CSQ0522	10	6	15	1	2	7	1	2	6
CSQ0523	10	5	15	1	1	7	0	2	5
CSQ0525	9	5	12	1	2	7	1	2	6
CSQ0526	10	6	15	1	1	8	0	2	5
CSQ0529	10	4	14	1	2	7	0	2	4
CSQ0531	7	6	15	3	1	8	0	2	5
CSQ0532	9	4	13	1	1	7	1	2	6
CSQ0536	8	5	13	2	1	7	1	2	6
CSQ0538	10	6	15	1	1	7	1	2	5
CSQ0540	10	4	12	2	1	7	1	2	5
CSQ0542	8	6	15	1	2	7	0	2	4
CSQ0543	9	4	13	1	1	7	2	2	6
CSQ0681	8	5	15	1	1	7	0	2	4
CSQ0682	9	5	13	1	1	7	0	2	4
CSQ0683	8	5	13	1	2	7	1	2	6
CSQ0690	10	5	14	1	1	7	0	2	5
CSQ0692	10	5	12	1	1	7	1	2	6
CSQ0704	6	7	15	1	1	7	0	2	4
CSQ0706	9	4	14	1	1	7	0	2	5
CSQ0724	11	5	13	1	1	7	1	2	6
CSQ0727	8	4	14	1	1	7	0	2	6

Vzorek	ORF1a	ORF1b	S	ORF3a	M	ORF7a	ORF7b	ORF8	N
CSQ0728	10	4	16	1	1	7	1	2	7
CSQ0742	9	4	13	1	1	7	1	2	6
CSQ0754	10	4	14	1	1	7	1	2	6
CSQ0755	8	5	16	1	1	7	0	2	3
CSQ0758	9	4	17	1	1	7	0	2	5
CSQ0760	10	4	15	1	1	7	1	2	6
CSQ0761	8	6	13	1	1	7	0	2	5
CSQ0777	10	5	16	2	2	9	1	2	6
CSQ0778	12	4	19	1	1	7	1	2	7
CSQ0782	10	4	15	3	3	7	0	2	5
CSQ0808	11	4	16	1	1	7	2	2	6
CSQ0816	12	8	17	4	2	7	1	2	6
CSQ0823	6	6	14	1	2	6	0	2	5
CSQ0829	12	4	13	3	1	7	1	2	6
CSQ0833	10	4	16	2	3	7	1	3	6
CSQ0834	8	5	16	1	2	7	0	2	5
CSQ0835	11	6	15	1	1	9	1	2	7
CSQ0846	9	5	21	1	1	9	1	2	5
CSQ0848	8	4	15	1	2	7	0	2	4
CSQ0849	8	4	17	1	1	7	0	2	5
CSQ0850	12	4	17	1	1	7	2	2	7
CSQ0855	9	4	16	1	1	7	1	2	7
CSQ0862	11	5	14	1	1	9	1	2	6
CSQ0866	17	7	18	1	1	9	2	2	6
CSQ0869	11	5	15	1	1	6	1	2	5
CSQ0871	9	6	16	2	1	6	1	2	8
CSQ0872	10	4	17	1	1	7	1	2	6
CSQ0913	10	5	14	1	1	7	2	2	6
CSQ0918	10	4	17	1	2	7	1	3	5
CSQ0919	11	5	14	1	1	7	2	2	5
CSQ0920	10	4	14	2	1	7	1	2	5
CSQ0923	7	7	17	1	1	7	0	2	4
CSQ0924	7	8	14	4	1	7	0	2	3
CSQ0926	10	5	17	1	2	9	0	2	4
CSQ0967	11	4	17	1	3	7	0	2	4
CSQ0979	9	4	17	1	1	7	1	2	6
CSQ0983	10	5	14	1	1	7	1	2	6

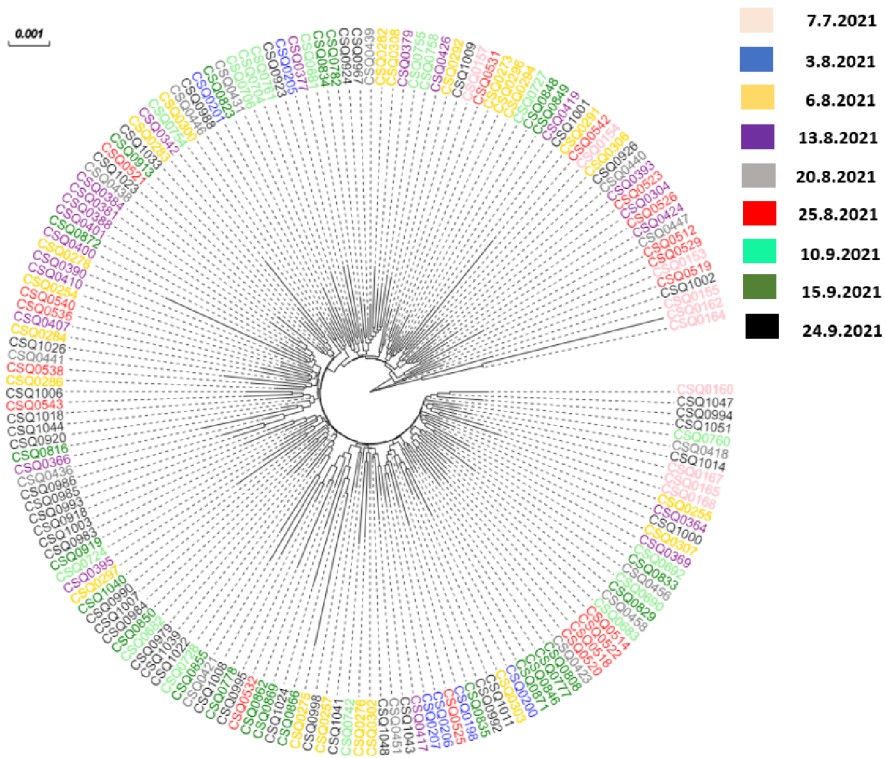
Vzorek	ORF1a	ORF1b	S	ORF3a	M	ORF7a	ORF7b	ORF8	N
CSQ0984	10	5	16	1	1	7	1	2	4
CSQ0985	6	4	17	3	1	7	0	2	4
CSQ0986	11	7	18	1	1	9	0	2	5
CSQ0988	8	7	17	1	1	7	0	2	5
CSQ0990	11	4	15	2	1	7	3	2	5
CSQ0992	11	5	15	1	1	7	1	2	5
CSQ0993	13	4	16	3	1	7	2	2	4
CSQ0994	10	6	15	1	1	7	1	2	4
CSQ0995	10	5	16	1	1	7	1	2	5
CSQ0998	12	5	13	1	1	7	1	2	5
CSQ1000	11	5	15	1	1	7	1	2	7
CSQ1001	10	6	18	3	1	9	0	2	4
CSQ1002	10	6	17	3	1	7	0	2	4
CSQ1003	14	5	13	1	1	9	1	2	4
CSQ1006	10	5	25	1	1	7	1	2	4
CSQ1007	10	4	20	3	1	7	1	2	5
CSQ1008	11	4	13	1	2	7	1	2	6
CSQ1009	7	7	14	1	1	7	0	2	3
CSQ1011	13	4	12	1	1	7	1	3	5
CSQ1014	9	5	16	1	1	7	1	2	5
CSQ1018	12	5	17	2	1	7	1	2	6
CSQ1022	11	5	15	3	1	7	1	2	6
CSQ1023	11	5	12	1	2	7	1	2	5
CSQ1024	11	5	16	1	2	7	2	2	6
CSQ1026	9	5	12	1	1	7	1	2	5
CSQ1033	9	8	15	1	2	7	1	3	7
CSQ1039	9	5	14	1	1	7	1	2	6
CSQ1040	11	4	22	2	2	7	2	2	5
CSQ1041	11	4	12	2	1	7	1	2	5
CSQ1043	10	6	15	2	1	7	1	2	5
CSQ1044	10	5	16	3	1	9	1	2	5
CSQ1047	12	5	17	1	2	7	1	2	5
CSQ1048	10	4	17	1	2	7	1	2	5
CSQ1051	10	6	18	3	1	7	1	2	7

E Fylogramy různých metod konstrukce fylogenetických stromů

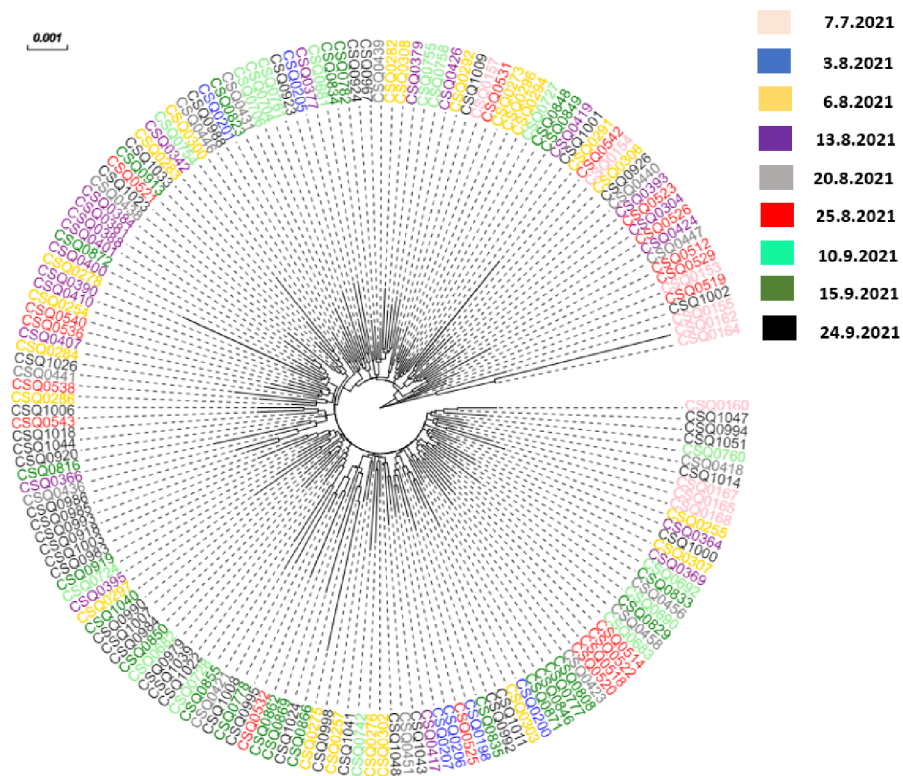
Níže jsou uvedeny fylogramy konstrukce fylogenetických stromů metodami Maximální pravděpodobnosti, Spojování nejbližších sousedů a Minimální evoluce.



Obr. E.1: Fylogenetický strom metodou UPGMA



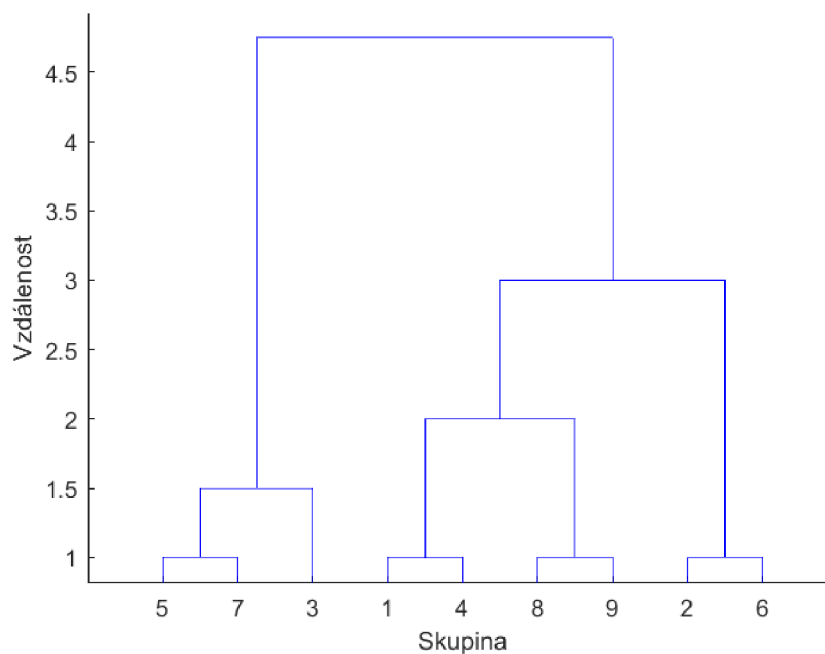
Obr. E.2: Fylogenetický strom metodou Spojování nejbližších sousedů



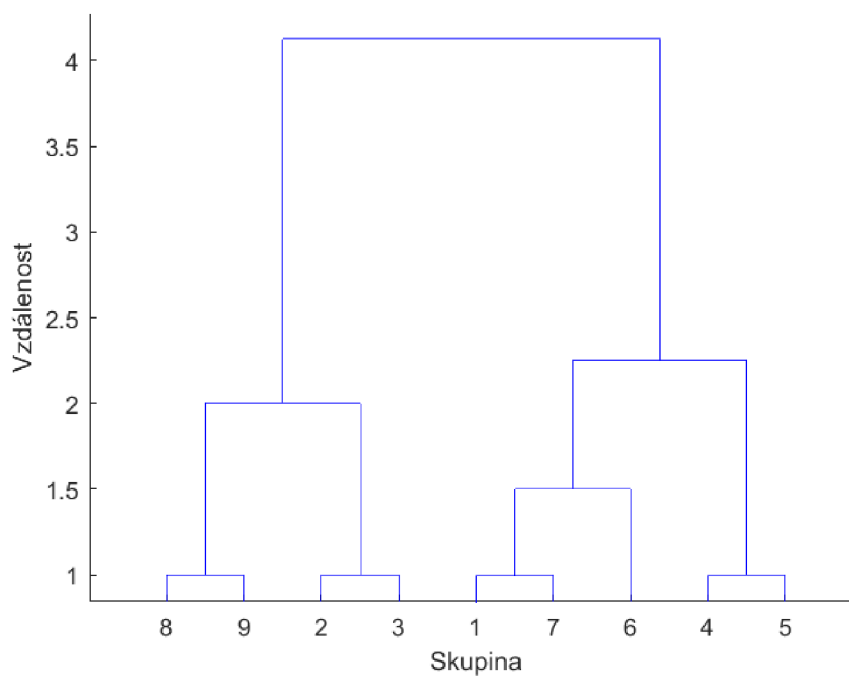
Obr. E.3: Fylogenetický strom metodou Minimální evoluce

F Výsledky shlukové analýzy z NEWICK formátů uvedených metod

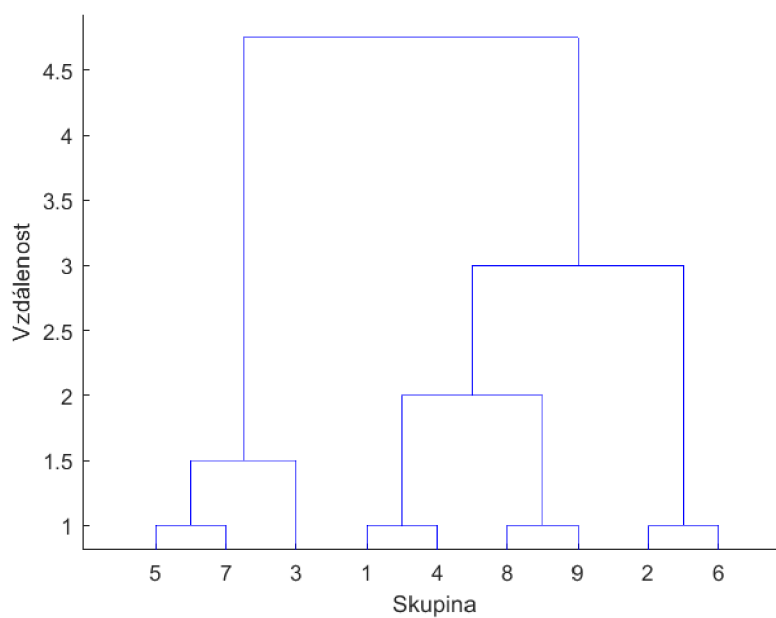
Níže jsou uvedeny tabulky a vykreslené dendrogramy zařazení vzorků do skupin shlukové analýzy z NEWICK formátů uvedených metod.



Obr. F.1: Seskupení vzorků z fylogenetického stromu metodou Spojování nejbližších sousedů



Obr. F.2: Seskupení vzorků z fylogenetického stromu metodou UPGMA



Obr. F.3: Seskupení vzorků z fylogenetického stromu metodou Minimální evoluce

Tab. F.1: Zařazení vzorků do shluků

Minimální evoluce		Spojování nejbližších sousedů		Maximální pravdě- podobnost		UPGMA	
Vzorek	Shluk	Vzorek	Shluk	Vzorek	Shluky	Vzorek	Shluk
CSQ0160	1	CSQ0160	1	CSQ0518	1	CSQ1047	1
CSQ1047	1	CSQ1047	1	CSQ0520	1	CSQ1051	1
CSQ0994	1	CSQ0994	1	CSQ0514	1	CSQ1044	1
CSQ1051	1	CSQ1051	1	CSQ0522	1	CSQ1048	1
CSQ0760	1	CSQ0760	1	CSQ0458	1	CSQ0835	1
CSQ0418	1	CSQ0418	1	CSQ0683	1	CSQ1041	1
CSQ1014	1	CSQ1014	1	CSQ0456	1	CSQ0160	1
CSQ0167	1	CSQ0167	1	CSQ0690	1	CSQ0307	1
CSQ0165	1	CSQ0165	1	CSQ0829	1	CSQ0979	1
CSQ0168	1	CSQ0168	1	CSQ0835	1	CSQ0521	1
CSQ0255	1	CSQ0255	1	CSQ0525	1	CSQ1023	1
CSQ0364	1	CSQ0364	1	CSQ0206	1	CSQ0410	1
CSQ1000	1	CSQ1000	1	CSQ0207	1	CSQ0984	1
CSQ0307	1	CSQ0307	1	CSQ0364	1	CSQ0255	1
CSQ0369	1	CSQ0369	1	CSQ0257	1	CSQ0754	1
CSQ0692	1	CSQ0692	1	CSQ1000	1	CSQ0992	1
CSQ0833	1	CSQ0833	1	CSQ0307	1	CSQ0983	1
CSQ0456	1	CSQ0456	1	CSQ0369	1	CSQ0990	1
CSQ0690	1	CSQ0690	1	CSQ0692	1	CSQ1000	1
CSQ0829	1	CSQ0829	1	CSQ0833	1	CSQ1026	1
CSQ0458	1	CSQ0458	1	CSQ1041	1	CSQ0920	1
CSQ0683	1	CSQ0683	1	CSQ0742	1	CSQ0206	1
CSQ0514	1	CSQ0514	1	CSQ0276	1	CSQ0207	1

Minimální evoluce		Spojování nejbližších sousedů		Maximální pravdě- podobnost		UPGMA	
Vzorek	Shluk	Vzorek	Shluk	Vzorek	Shluky	Vzorek	Shluk
CSQ0522	1	CSQ0522	1	CSQ0302	1	CSQ0728	1
CSQ0518	1	CSQ0518	1	CSQ0417	1	CSQ0254	1
CSQ0520	1	CSQ0520	1	CSQ1043	1	CSQ0846	1
CSQ0423	1	CSQ0423	1	CSQ0451	1	CSQ0871	1
CSQ0808	1	CSQ0808	1	CSQ1048	1	CSQ0829	1
CSQ0777	1	CSQ0777	1	CSQ0255	1	CSQ0760	1
CSQ0846	1	CSQ0846	1	CSQ0850	1	CSQ0257	1
CSQ0871	1	CSQ0871	1	CSQ1007	1	CSQ0364	1
CSQ0200	1	CSQ0200	1	CSQ0984	1	CSQ0540	1
CSQ0303	1	CSQ0303	1	CSQ0998	1	CSQ1003	1
CSQ1011	1	CSQ1011	1	CSQ0777	1	CSQ0777	1
CSQ0992	1	CSQ0992	1	CSQ0846	1	CSQ0808	1
CSQ0835	1	CSQ0835	1	CSQ0871	1	CSQ1039	1
CSQ0198	1	CSQ0198	1	CSQ0423	1	CSQ0302	1
CSQ0525	1	CSQ0525	1	CSQ0808	1	CSQ0919	1
CSQ0206	1	CSQ0206	1	CSQ0728	1	CSQ0400	1
CSQ0207	1	CSQ0207	1	CSQ1039	1	CSQ0872	1
CSQ0417	1	CSQ0417	1	CSQ0760	2	CSQ0388	1
CSQ1043	1	CSQ1043	1	CSQ0160	2	CSQ0833	1
CSQ0451	1	CSQ0451	1	CSQ1047	2	CSQ0401	1
CSQ1048	1	CSQ1048	1	CSQ1051	2	CSQ0381	1
CSQ0302	1	CSQ0302	1	CSQ0994	2	CSQ0384	1
CSQ0276	1	CSQ0276	1	CSQ0198	2	CSQ0536	1
CSQ0742	1	CSQ0742	1	CSQ0992	2	CSQ1033	1
CSQ1041	1	CSQ1041	1	CSQ0303	2	CSQ0164	1
CSQ0257	1	CSQ0257	1	CSQ0200	2	CSQ0286	1
CSQ0998	1	CSQ0998	1	CSQ1011	2	CSQ0683	1
CSQ0866	1	CSQ0866	1	CSQ0275	3	CSQ0456	1
CSQ1024	1	CSQ1024	1	CSQ0862	3	CSQ0862	1
CSQ0869	1	CSQ0869	1	CSQ0866	3	CSQ0866	1
CSQ0862	1	CSQ0862	1	CSQ1024	3	CSQ0998	1
CSQ0532	1	CSQ0532	1	CSQ0532	3	CSQ0283	1
CSQ0995	1	CSQ0995	1	CSQ0869	3	CSQ0418	1
CSQ0778	1	CSQ0778	1	CSQ0778	3	CSQ1014	1
CSQ1008	1	CSQ1008	1	CSQ0995	3	CSQ0913	1

Minimální evoluce		Spojování nejbližších sousedů		Maximální pravdě- podobnost		UPGMA	
Vzorek	Shluk	Vzorek	Shluk	Vzorek	Shluky	Vzorek	Shluk
CSQ0275	2	CSQ0275	2	CSQ1008	3	CSQ0167	1
CSQ0421	3	CSQ0421	3	CSQ1014	3	CSQ0993	1
CSQ0855	3	CSQ0855	3	CSQ0418	3	CSQ0816	1
CSQ0728	3	CSQ0728	3	CSQ0168	3	CSQ1007	1
CSQ1022	3	CSQ1022	3	CSQ0165	3	CSQ1018	1
CSQ1039	3	CSQ1039	3	CSQ0167	3	CSQ1022	1
CSQ0979	3	CSQ0979	3	CSQ0421	4	CSQ0390	1
CSQ0682	3	CSQ0682	3	CSQ1022	4	CSQ0417	1
CSQ0850	3	CSQ0850	3	CSQ0855	4	CSQ1043	1
CSQ0984	3	CSQ0984	3	CSQ0979	4	CSQ1040	1
CSQ1007	3	CSQ1007	3	CSQ0990	4	CSQ0692	1
CSQ0990	3	CSQ0990	3	CSQ1040	4	CSQ0522	1
CSQ1040	3	CSQ1040	3	CSQ0919	4	CSQ0458	1
CSQ0297	3	CSQ0297	3	CSQ0682	4	CSQ0514	1
CSQ0395	3	CSQ0395	3	CSQ0983	4	CSQ0542	1
CSQ0724	3	CSQ0724	3	CSQ0521	4	CSQ0518	1
CSQ0919	3	CSQ0919	3	CSQ1023	4	CSQ0520	1
CSQ0983	3	CSQ0983	3	CSQ1026	4	CSQ0369	1
CSQ1003	3	CSQ1003	3	CSQ0438	4	CSQ0201	1
CSQ0918	3	CSQ0918	3	CSQ0754	4	CSQ0988	1
CSQ0993	3	CSQ0993	3	CSQ0913	4	CSQ0823	1
CSQ0436	3	CSQ0436	3	CSQ0342	4	CSQ0393	1
CSQ0366	3	CSQ0366	3	CSQ0254	5	CSQ0523	1
CSQ0816	3	CSQ0816	3	CSQ0410	5	CSQ0526	1
CSQ0920	3	CSQ0920	3	CSQ0407	5	CSQ0205	1
CSQ1044	3	CSQ1044	3	CSQ1003	5	CSQ0923	1
CSQ1018	3	CSQ1018	3	CSQ0920	5	CSQ0426	1
CSQ0543	3	CSQ0543	3	CSQ0441	5	CSQ0157	1
CSQ1006	3	CSQ1006	3	CSQ0538	5	CSQ0727	1
CSQ0286	3	CSQ0286	3	CSQ0286	5	CSQ0419	1
CSQ0538	3	CSQ0538	3	CSQ0816	5	CSQ0849	1
CSQ0441	3	CSQ0441	3	CSQ0366	5	CSQ0296	1
CSQ1026	3	CSQ1026	3	CSQ0436	5	CSQ0273	1
CSQ0284	3	CSQ0284	3	CSQ0985	5	CSQ0531	1
CSQ0407	3	CSQ0407	3	CSQ0986	5	CSQ0294	1

Minimální evoluce		Spojování nejbližších sousedů		Maximální pravdě- podobnost		UPGMA	
Vzorek	Shluk	Vzorek	Shluk	Vzorek	Shluky	Vzorek	Shluk
CSQ0536	3	CSQ0536	3	CSQ0543	5	CSQ0424	1
CSQ0540	3	CSQ0540	3	CSQ1018	5	CSQ0529	1
CSQ0254	3	CSQ0254	3	CSQ1006	5	CSQ0512	1
CSQ0410	3	CSQ0410	3	CSQ1044	5	CSQ0706	1
CSQ0390	3	CSQ0390	3	CSQ0724	5	CSQ0704	1
CSQ0278	3	CSQ0278	3	CSQ0297	5	CSQ1009	1
CSQ0400	3	CSQ0400	3	CSQ0395	5	CSQ0292	1
CSQ0872	3	CSQ0872	3	CSQ0993	5	CSQ0153	1
CSQ0401	3	CSQ0401	3	CSQ0381	6	CSQ0834	1
CSQ0388	3	CSQ0388	3	CSQ0384	6	CSQ0967	1
CSQ0381	3	CSQ0381	3	CSQ0388	6	CSQ0848	1
CSQ0384	3	CSQ0384	3	CSQ0401	6	CSQ0926	1
CSQ0438	3	CSQ0438	3	CSQ0872	6	CSQ0525	1
CSQ1023	3	CSQ1023	3	CSQ0278	6	CSQ0682	1
CSQ0521	3	CSQ0521	3	CSQ0400	6	CSQ0850	1
CSQ0913	3	CSQ0913	3	CSQ0390	6	CSQ0407	1
CSQ1033	3	CSQ1033	3	CSQ0536	6	CSQ0395	1
CSQ0283	3	CSQ0283	3	CSQ0918	6	CSQ0724	1
CSQ0342	3	CSQ0342	3	CSQ0284	6	CSQ0200	2
CSQ0985	4	CSQ0985	4	CSQ1033	4	CSQ0300	1
CSQ0986	4	CSQ0986	4	CSQ0283	4	CSQ0446	1
CSQ0754	5	CSQ0754	5	CSQ0540	6	CSQ1011	2
CSQ0704	5	CSQ0704	5	CSQ0379	7	CSQ0423	1
CSQ0439	5	CSQ0439	5	CSQ0424	7	CSQ0306	1
CSQ0292	5	CSQ0292	5	CSQ0923	7	CSQ0690	1
CSQ1009	5	CSQ1009	5	CSQ0205	7	CSQ0198	1
CSQ0157	5	CSQ0157	5	CSQ0377	7	CSQ1001	1
CSQ0300	5	CSQ0300	5	CSQ0529	7	CSQ0303	2
CSQ0446	5	CSQ0446	5	CSQ0308	7	CSQ0994	2
CSQ0988	5	CSQ0988	5	CSQ0755	7	CSQ0165	3
CSQ0201	5	CSQ0201	5	CSQ0758	7	CSQ0436	3
CSQ0823	5	CSQ0823	5	CSQ0782	7	CSQ0995	3
CSQ0431	5	CSQ0431	5	CSQ0924	7	CSQ0538	3
CSQ0706	5	CSQ0706	5	CSQ0282	7	CSQ0778	3

Minimální evoluce		Spojování nejbližších sousedů		Maximální pravdě- podobnost		UPGMA	
Vzorek	Shluk	Vzorek	Shluk	Vzorek	Shluky	Vzorek	Shluk
CSQ0761	5	CSQ0761	5	CSQ0439	7	CSQ0282	4
CSQ0923	5	CSQ0923	5	CSQ0967	7	CSQ0304	4
CSQ0205	5	CSQ0205	5	CSQ0834	7	CSQ0755	4
CSQ0377	5	CSQ0377	5	CSQ0300	7	CSQ0758	4
CSQ0834	5	CSQ0834	5	CSQ0988	7	CSQ0439	4
CSQ0782	5	CSQ0782	5	CSQ0201	7	CSQ0308	4
CSQ0924	5	CSQ0924	5	CSQ0823	7	CSQ0782	4
CSQ0967	5	CSQ0967	5	CSQ0512	7	CSQ0924	4
CSQ0426	5	CSQ0426	5	CSQ0761	7	CSQ1006	5
CSQ0531	5	CSQ0531	5	CSQ0519	8	CSQ1002	1
CSQ0294	5	CSQ0294	5	CSQ0155	8	CSQ0431	1
CSQ0727	5	CSQ0727	5	CSQ1002	8	CSQ0168	1
CSQ0848	5	CSQ0848	5	CSQ0153	8	CSQ0284	1
CSQ0849	5	CSQ0849	5	CSQ1009	8	CSQ0377	1
CSQ0419	5	CSQ0419	5	CSQ0292	8	CSQ0276	1
CSQ1001	5	CSQ1001	5	CSQ0426	8	CSQ0869	1
CSQ0273	5	CSQ0273	5	CSQ0162	8	CSQ0438	6
CSQ0296	5	CSQ0296	5	CSQ0164	8	CSQ0918	6
CSQ0542	5	CSQ0542	5	CSQ0926	9	CSQ0742	1
CSQ0926	5	CSQ0926	5	CSQ0542	9	CSQ0441	1
CSQ0440	5	CSQ0440	5	CSQ0393	9	CSQ0447	1
CSQ0393	5	CSQ0393	5	CSQ0523	9	CSQ0681	1
CSQ0523	5	CSQ0523	5	CSQ0526	9	CSQ0297	1

Minimální evoluce		Spojování nejbližších sousedů		Maximální pravdě- podobnost		UPGMA	
Vzorek	Shluk	Vzorek	Shluk	Vzorek	Shluky	Vzorek	Shluk
CSQ0304	5	CSQ0304	5	CSQ0440	9	CSQ0342	1
CSQ0526	5	CSQ0526	5	CSQ0304	9	CSQ0366	1
CSQ0424	5	CSQ0424	5	CSQ0273	9	CSQ0421	1
CSQ0447	5	CSQ0447	5	CSQ0294	9	CSQ0855	1
CSQ0512	5	CSQ0512	5	CSQ0296	9	CSQ1008	1
CSQ0529	5	CSQ0529	5	CSQ0531	9	CSQ0532	1
CSQ0153	5	CSQ0153	5	CSQ0157	9	CSQ1024	1
CSQ0519	5	CSQ0519	5	CSQ0848	9	CSQ0155	1
CSQ1002	5	CSQ1002	5	CSQ0727	9	CSQ0291	1
CSQ0155	5	CSQ0155	5	CSQ0419	9	CSQ0278	1
CSQ0154	5	CSQ0154	5	CSQ0306	9	CSQ0985	7
CSQ0306	5	CSQ0306	5	CSQ0154	9	CSQ0986	7
CSQ0164	5	CSQ0164	5	CSQ1001	9	CSQ0162	9
CSQ0681	6	CSQ0681	6	CSQ0446	7	CSQ0379	4
CSQ0282	7	CSQ0282	7	CSQ0447	7	CSQ0154	1
CSQ0308	7	CSQ0308	7	CSQ0681	7	CSQ0519	1
CSQ0379	7	CSQ0379	7	CSQ0431	7	CSQ0440	1
CSQ0755	7	CSQ0755	7	CSQ0706	7	CSQ0451	1
CSQ0758	7	CSQ0758	7	CSQ0704	7	CSQ0543	5
CSQ0291	8	CSQ0291	8	CSQ0291	9	CSQ0761	1
CSQ0162	9	CSQ0162	9	CSQ0849	9	CSQ0275	8