ŠPARK - COVID

ŠPARK - Andrej Špitalský, Teo Pazera, Tomáš Antal,
Rafael Rohaľ, Róbert Kendereš
15.5.2023

Obsah

Obsah **1**
Úvod **2**
Použité dáta **3**
Spracovanie dát 3
Popis stĺpcov v merged tabuľkách 3
Výsledky **5**
Hodnoty závislé od kvality 5
Výskyt organizmov klasifikovaných ako “non-target” 7
Rozdiely medzi “batch” súbormi 7
Analýza referenčného genómu 11
Príprava dát 11
Pokrytie referenčného genómu 12
Priemerná odhadovaná kvalita referenčného genómu 14
Záver **16**
Zoznam literatúry **17**

Úvod

V našom projekte sme sa venovali dátam Univerzity Komenského o sekvenácii genómu koronavírusu. Dáta boli z obdobia prelomu januára a februára, kedy začínalo obdobie vysokej prevalencie variantu Alfa. Sledovali sme, aké trendy v dátach tento variant spôsobuje, čo vieme predikovať o dátach na základe odhadovanej kvality sekvenácie a ako vplývajú mutácie genómu na podobnosť vzoriek od pacientov s referenčnou sekvenciou.

Použité dáta

Dáta sme získali od pani docentky Brejovej (1). Jednalo sa o sekvenáciu vzoriek genetického materiálu (hlavne) koronavírusu v troch behoch po 24 pacientov. Genetická informácia koronavírusu bola amplifikovaná PCR metódou a vzorky jednotlivých pacientov boli označené na oboch koncoch DNA vlákna označené tzv. *barcodmi*, aby sa umožnili spoločné sekvenovanie vzoriek od viacerých pacientov naraz.

Pre každý sekvenčný beh sme mali k dispozícii tri súbory - *reads, match* a *results*. Tabuľka *reads* opisovala dáta o sekvenčných behoch, okrem iného odhadovanú kvalitu čítania či čas, kedy začala sekvenácia danej vzorky. Jednotlivé riadky tabuľky reprezentujú dáta o sekvenácii jednej vzorky DNA od pacienta. Tabuľka *match* porovnáva vzorky od pacientov s referenčným genómom SARS-CoV-2 (6). Obsahuje údaje, kde sa daná vzorka najlepšie namapovala na nejaký úsek referenčného genómu - tento úsek budeme nazývať zodpovedajúci región. Tabuľka *reads* zhŕňa výsledky analýz a okrem iného priradí danému pacientovi aj variant koronavírusu, ktorý bol u neho nájdený.

Naše spracované dáta a iné súbory sme nahrali na stránku, ktorú sme spojazdnili cez Davinci cluster Univerzity Komenského (5).

Spracovanie dát

Keďže pre každý z troch batchov máme tri tabuľky - *reads, match* a *results*, spojíme ich pre lepšiu manipuláciu do jednej tabuľky, ktorú budeme nazývať *merged*. Takisto, pre potreby analýzy všetkých batchov naraz, spojíme tieto tri *merged* tabuľky do jednej, budeme nazývať *merged_all*. Tieto tabuľky sú dostupné na linkoch:

- batch 10 - http://www.st.fmph.uniba.sk/~spitalsky3/viz-projekt/batch10-merged.gz
- batch 11 - http://www.st.fmph.uniba.sk/~spitalsky3/viz-projekt/batch11-merged.gz
- batch 12 - http://www.st.fmph.uniba.sk/~spitalsky3/viz-projekt/batch12-merged.gz
- batch all- http://www.st.fmph.uniba.sk/~spitalsky3/viz-projekt/batchall-merged.gz

Popis stĺpcov v *merged* tabuľkách

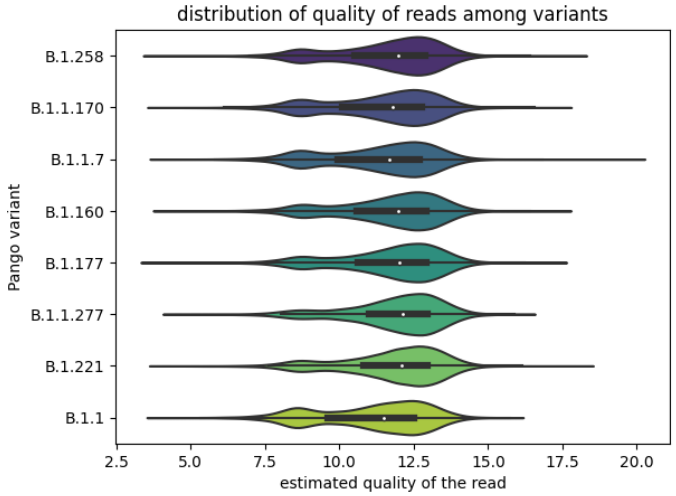
- ***ID*** - identifikátor konkrétnej vzorky
- ***batchNum*** (len pri *merged_all*) - poradové číslo batchu, z ktorého je daná vzorka
- *strand* - ktoré vlákno DNA bolo použité pri sekvenácii
  - \+ pre 5’-3’ vlákno
  - \- pre 3’-5’ vlákno
- ***matching*** - koľko báz bolo rovnakých v zodpovedajúcom regióne
- ***lengthMatch*** - dĺžka zodpovedajúceho regiónu na vzorke
- ***matchingStart*** - na koľkej báze vzorky začína zodpovedajúci región
- ***matchingEnd*** - na koľkej báze vzorky končí zodpovedajúci región
- ***referenceID*** - ID referenčného genómu, s ktorým sa porovnávajú dáta
  - v našich dátach iba MN908947.3
- ***referenceLength*** - dĺžka použitého referenčného genómu
  - v našich dátach vždy 29 903
- ***referenceStart*** - na koľkej báze referenčného genómu začína zodpovedajúci región
- ***referenceEnd*** - na koľkej báze referenčného genómu končí zodpovedajúci región
- ***identicalPerc*** - koľko percent báz sa zhoduje v zodpovedajúcom regióne

- ***timeStart*** - interný čas prístroja, kedy začal sekvenovať danú vzorku
- ***lengthReads*** - koľko báz mala prečítaná vzorka
- ***estQuality*** - odhadovaná kvalita sekvenácia vzorky
    - čím vyššie číslo, tým menej chýb by sa vo vzorke malo vyskytovať
- ***barcode1*** - či aspoň na jednom konci vzorky bol nájdený barcode nejakého pacienta
    - 1 až 24 - ak bol nájdený barcode (pacient má pridelené vlastné číslo)
    - 1 až 72 - pri *merged_all*
    - unclassified - nebol nájdený barcode ani na jednom konci vzorky
- ***barcode2*** - či na oboch koncoch vzorky bol nájdený barcode toho istého pacienta
    - 1 až 24 - ak bol nájdený barcode (pacient má pridelené vlastné číslo)
    - 1 až 72 - pri *merged_all*
    - unclassified - nebol nájdený barcode na oboch koncoch vzorky
- ***barcode3*** - rovnaké ako barcode2, len vzorka musí spĺňať kritéria navyše (dostatočná dĺžka, kvalita a musí sa jednať od DNA SARS-CoV-2)
    - 1 až 24 - ak bol nájdený rovnaký barcode na oboch koncoch a spĺňa kritériá
    - 1 až 72 - pri *merged_all*
    - none - nebol nájdený barcode na oboch koncoch vzorky alebo nespĺňa kritériá
- ***organism*** - identifikácia organizmu, z ktorého pochádzala vzorka
    - target - SARS-CoV-2
    - virus - iná virálna DNA
    - none - nezhoduje sa s ničím v databáze
- ***statusCode*** - klasifikácia vzorky skratkou
    - G.ok - skratka pre kvalitné vzorky, ktoré boli použité pre hľadanie mutácií
- ***notDetermined*** - počet báz, ktoré sa pri danom pacientovi nepodarilo určiť
- ***numMutations*** - počet mutácií DNA Covidu získané z konkrétneho pacienta
- ***Pango*** - klasifikácia variantu Covidu, ktorý sa našiel u pacienta
- ***probPango*** - pravdepodobnosť, že bol variant určený správne

Výsledky

Hodnoty závislé od kvality

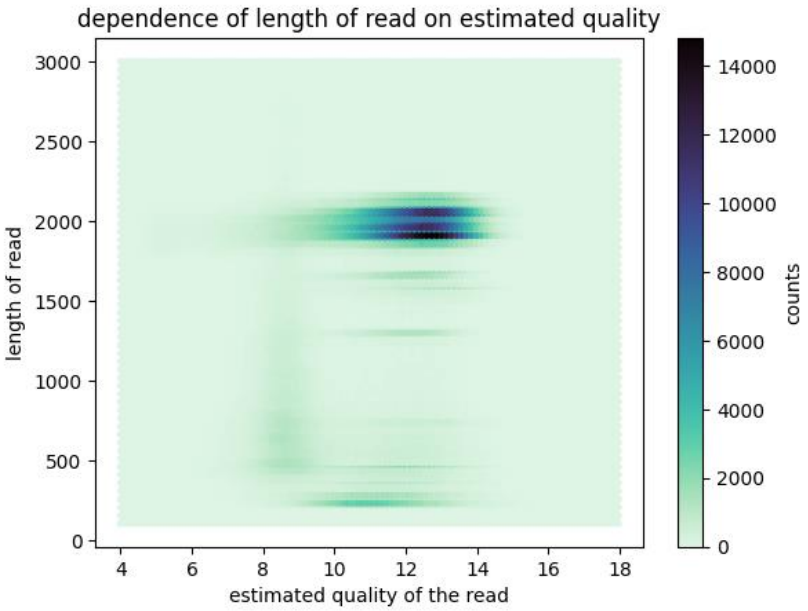
Najviac nás zaujala hodnota odhadovanej kvality vzorky (estQuality). Mali sme o nej málo informácií, ale nejaké predpoklady. Chceli sme sa pozrieť na to, ktoré hodnoty najviac ovplyvňujú hodnotu estQuality. Mali sme podozrenie, že by sa estQuality menila s variantom a možno aj s dĺžkami jednotlivých čítaní. Vykreslili sme si pre to nejaké ploty a pozreli sme sa, či tieto hodnoty majú koreláciu s estQuality.



obr. 1 : Distribúcia odhadovanej kvality čítaní medzi variantmi

Pozreli sme sa na distribúciu kvality v jednotlivých variantoch. Na obr.1 je očividné, že medzi variantom a kvalitou nie je korelácia, a teda kvalita nevyplýva z variantu.

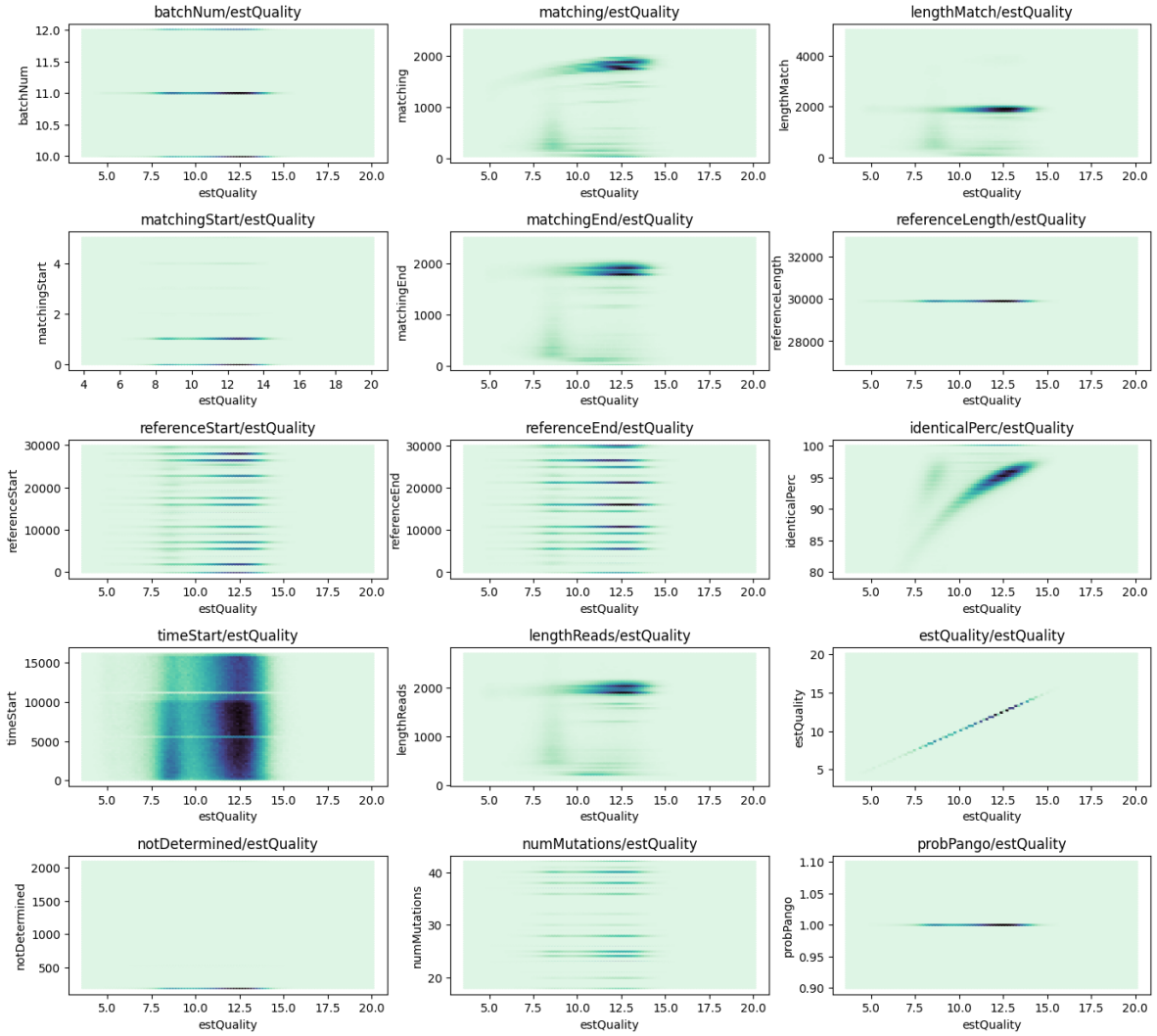
Ďalej sme sa pozreli na lengthReads, teda dĺžku čítania.



obr. 2 : Závislosť dĺžky čítaní od odhadovanej kvality čítania

Na prvý pohľad na obr.2 nevieme povedať, či tam je nejaká korelácia. Zároveň, na grafe môžeme vidieť zhluk dát okolo hodnoty 2000 a vo viacerých hodnotách nižšie. Ak tam teda nejaká korelácia je, je veľmi slabá na to, aby sme vedeli dobre predpovedať dĺžku readu čisto podľa kvality.

Variant a dĺžka čítania teda nemajú vplyv na hodnotu estQuality. Pozreli sme sa, či existuje korelácia medzi estQuality a inými číselnými hodnotami z tabuľky.



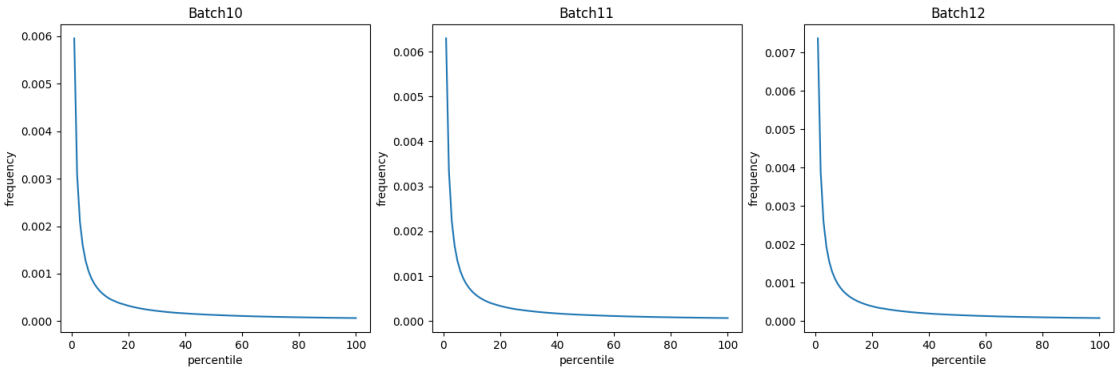
obr. 3 : Závislosť všetkých numerických hodnôt od odhadovanej kvality čítania

Všimli sme si očividné korelácie medzi identicalPerc a trochu pre matching. Pozreli sme sa teda na ich numerické hodnoty. Korelácia medzi matching a estQuality bola ~0.40. Korelácia medzi identicalPerc a estQuality je ~0.63. identicalPerc je percento báz, ktoré sa zhodujú s referenčnou sekvenciou. Matching je počet týchto zhodných báz, preto je korelácia podobná, no nie je rovnaká, vzhľadom na to, že matching závisí aj na externom faktore, a to je samotná dĺžka readu. identicalPerc touto hodnotou ovplyvnené nie je.

IdenticalPerc s koreláciou 0.63 môžeme teda klasifikovať ako najvplyvnejší faktor pre kalkuláciu estQuality a teda podľa danej kvality vieme spätne odhadnúť percento zhodných báz. Túto teóriu sme si potvrdili pomocou decision-tree regresie(2) s pravdepodobnosťou správneho odhadu 0.769, a priemerná štvorcová chyba iba 14.547.

Výskyt organizmov klasifikovaných ako “non-target”

Ďalej sme sa pozreli na percentuálny výskyt čítaní označených ako non-target s meniacim sa percentilom premennej odhadovaná kvalita.

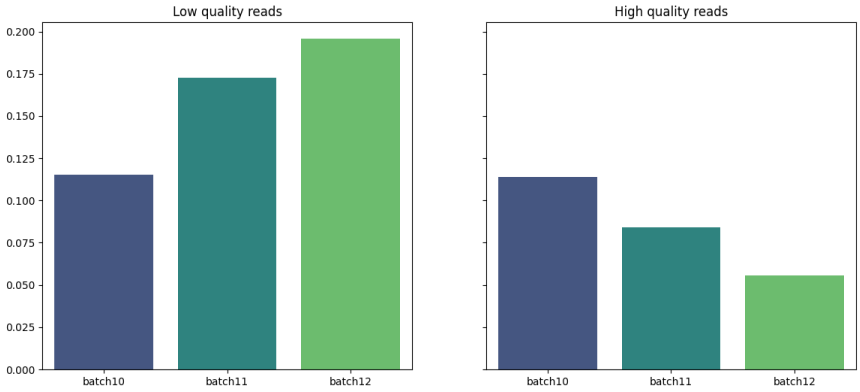


obr. 4 : Výskyt organizmov iných ako target podľa percentilu kvality

A teda podľa grafu vidíme klesajúcu tendenciu tohto výskytu čiže v čítaniach s vysokou kvalitou je týchto organizmov málo. V čítaniach s vysokou kvalitou sa skôr vyskytovali organizmy označené ako target (covid).

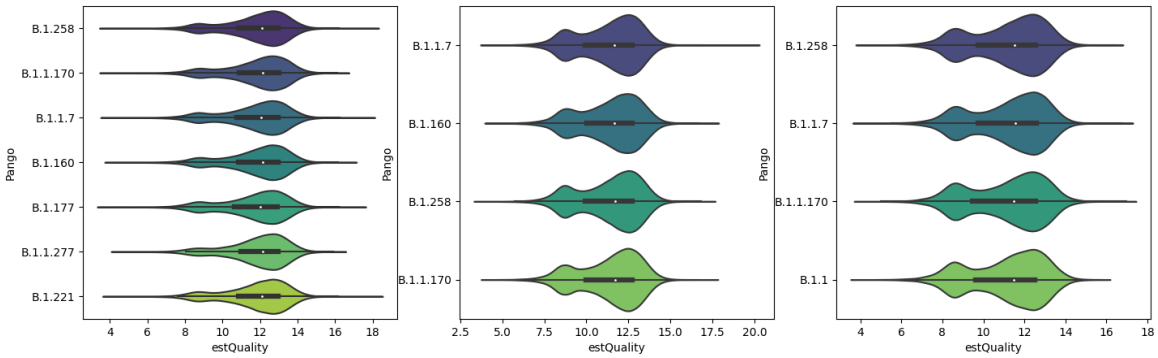
Rozdiely medzi “batch” súbormi

Ďalšia veľká otázka bola porovnať “batche” - 3 súbory ktoré sme dostali. Najskôr sme sa pozreli na viacero aspektov jednotlivých súborov, ako napríklad počet organizmov označených inak ako “target”, počet čítaní s nízkou kvalitou, počet čítaní s vysokou kvalitou, a počet dlhých a krátkych čítaní. Z tohto pôvodného porovnania sme si všimli hlavne tendenciou stúpajúceho počtu čítaní s nízkou kvalitou a klesajúceho počtu čítaní s vysokou kvalitou.



obr. 5 : Počet výskytov čítaní s nízkou a vysokou kvalitou

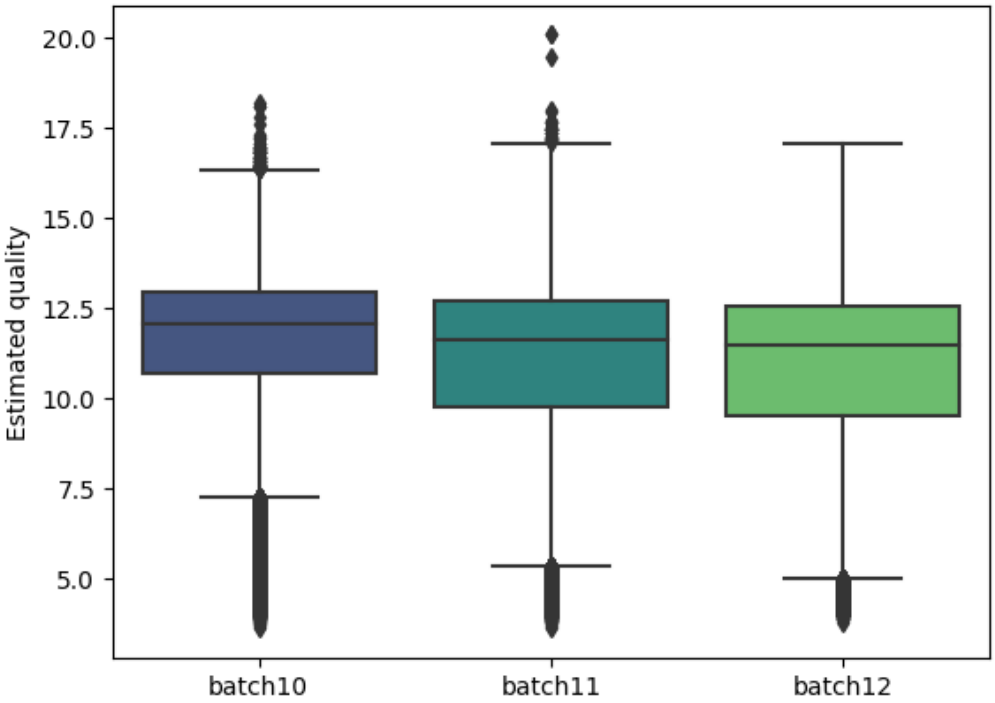
Vo väčšine prípadov nás teda zaujala premenná “estimated quality” a snažili sme sa nájsť závislosti medzi touto a dalšími premennými.



obr. 6 : Kvalita podľa variantu

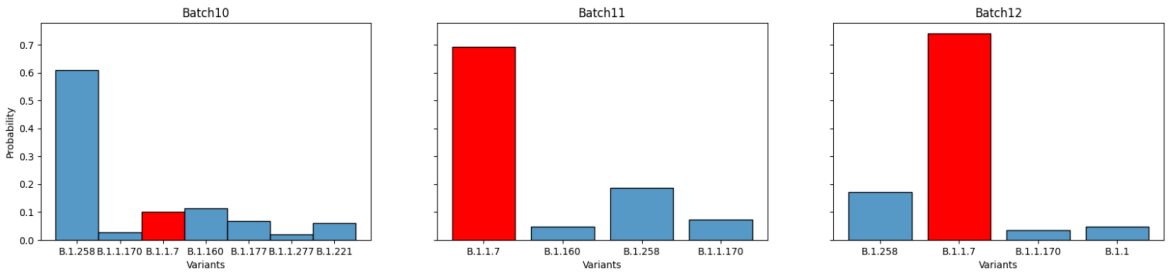
Pozreli sme sa na rozdelenie kvality podľa variantu covidu vo všetkých troch súboroch. Podľa grafu vidíme, že rozdelenie kvality je rovnomerné pre všetky druhy covidu a teda podľa premennej estimated quality by sme nevedeli odhadnúť o aký variant jednotlivého čítania sa jedná.

Ďalej sme sa zamerali na priemerné rozdelenie odhadovanej kvality medzi súbormi. Použili sme boxplot, aby sme sa mohli v jednom grafe zamerať aj na jednotlivé kvantily.



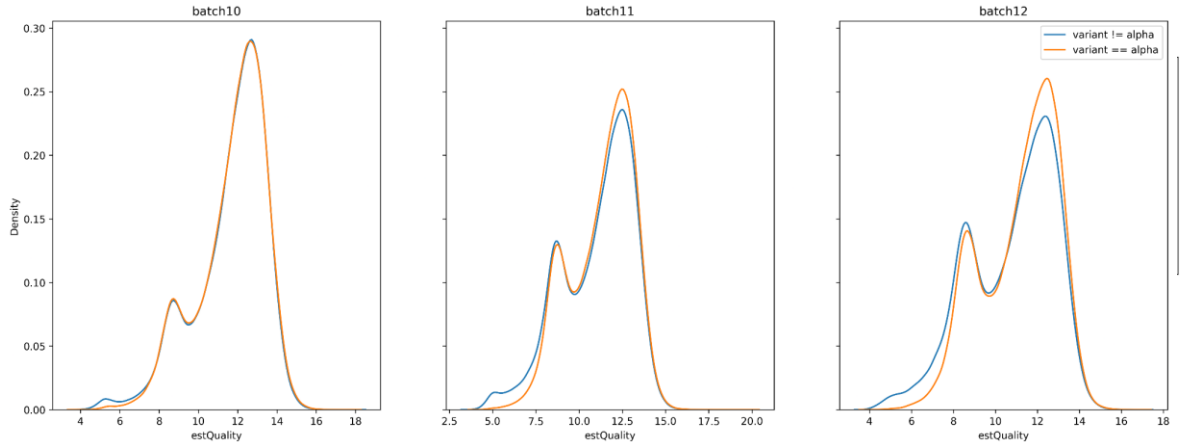
obr. 7 : Boxploty priemernej kvality pre súbory

Z grafu je zreteľné, že priemerná kvality klesá. Otázka bola, prečo tomu tak je. Ďalšou exploratívnou analýzou sme prišli k tomuto grafu.



Obr. 8 : Percentuálny počet výskytov variantov v súboroch

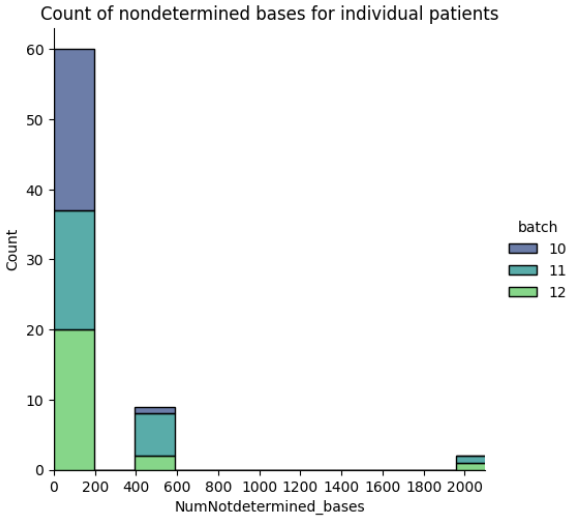
Je priamo vidieť prevalenciu alfa variantu v batch-och 11 a 12. Dospeli sme teda k hypotéze, že zníženie kvality spôsobuje práve prevládajúci výskyt alfa variantu v obidvoch súboroch. Snažili sme sa túto hypotézu potvrdiť nejakým spôsobom a preto sme sa pozreli na čitania, ktoré obsahovali len variant alfa a všetky ostatné.



Obr. 9 :Hustota výskytu čítaní alfa variantu a všetkých ostatných

Zistili sme, že alfa variant má v nižších kvalitách menšie percentuálne zastúpenie a naopak vo vyšších kvalitách vyššie percentuálne zastúpenie. Z čoho vyplýva pravý opak našej pôvodnej hypotézy, a teda alfa variant by mal celkovú kvalitu súboru zvyšovať. Preto sme si túto hypotézu vyvrátili.

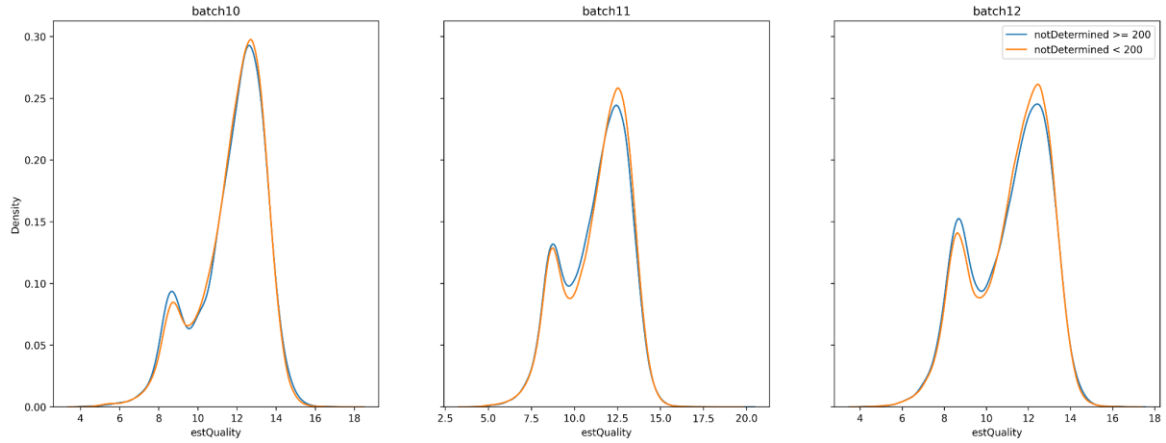
Museli sme teda hľadať ďalej čo spôsobilo pokles kvality. Pozreli sme sa na tabuľku results kde nám stĺpec “notDetermined” ukazuje koľko báz u určitého pacienta nebolo rozoznateľných.



obr. 10 :Počet neurčených báz u pacientov medzi batchmi

Väčšina pacientov vo všetkých batchoch mala okolo 190 až 200 neurčených báz. Ale v batchoch 11 a 12 bol počet týchto neurčených báz pri viacerých viac ako 200 okolo 560 a este aj 2000. O pacientoch ktorým neurčíme bázy vieme menej a preto nám napadlo že by to mohlo mať aj vplyv na kvalitu.

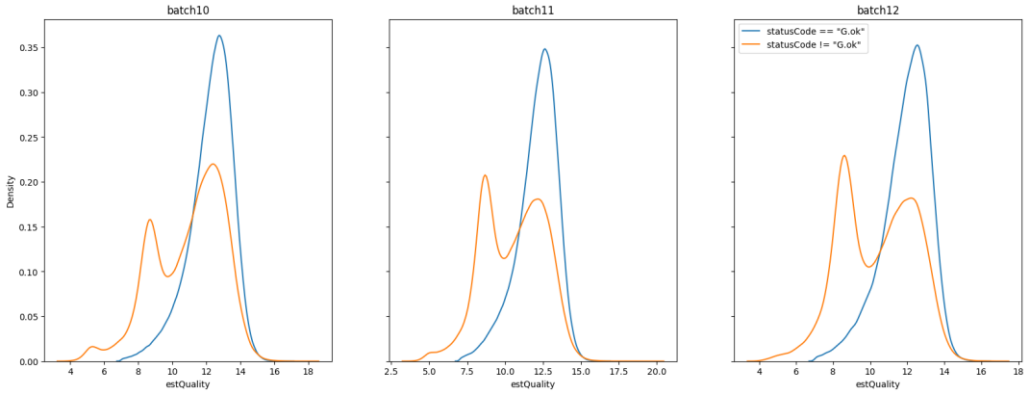
Našu hypotézu sme si overili pomocou grafu



obr. 11 : Hustota výskytov čitaní s neurčenými bázami

Z ktorého je vidieť, že čítania s počtom neurčených báz väčším ako 200 majú v nižších kvalitách väčšie percentuálne zastúpenie a vo vyšších kvalitách nižšie percentuálne zastúpenie a teda sme dospeli k záveru že práve tieto čítania by mohli spôsobovať priemerné zníženie kvality v daných súboroch.

Ešte sme sa pozreli na čitania ktoré boli zohľadnené teda použité v študií z ktorej naše dáta pochádzajú. Čítania klasifikované ako "G.ok" sú tie ktore splňajú podmienky: obidva konce readu musia mať rozoznateľný barcode + dostatočnú dĺžku čítania a tiež dostatočnú kvalitu. Tieto teda boli použité na vyvodenie záverov v štúdii. Ukázalo sa nejaké rozdelenie hodnôt estQuality okolo 12.5.



obr. 12 :Hustota výskytov čítaní ktoré boli klasifikované ako

Analýza referenčného genómu

Signifikantná časť dát sa venovala porovnávaniu vzoriek s referenčným genómom (viac informácii o referenčnom genóme je na adrese 6). Vieme, že genetická informácia vírusu SARS-CoV-2 bola amplifikovaná vo vzorke metódou polymerázovej reťazovej reakcie - PCR. Tá využíva tzv. PCR primery, ktoré sú nadizajnové tak, aby sa naviazali na špecifický úsek DNA a spustili replikáciu niekoľko báz za týmto úsekom. Toto viazanie primerov a následná replikácia prebehne vo vzorke viackrát, čím vieme zvýšiť koncentráciu konkrétnej genetickej informácie. na ktorú boli PCR primery nadizajnované. Táto metóda však nezabezpečí

zvýšenie koncentrácie celej genetickej informácie rovnomerne, keďže každý primer je zodpovedný za replikáciu iba časti DNA a vieme, že ich viazanie nemusí byť konzistentné.

Príprava dát

Rozdelili sme dáta jednotlivých batchov podľa pacientov. Pre každú vzorku máme k dispozícii údaje, že ktorá časť referenčného genómu je jej najpodobnejšia (stĺpce *referenceStart* a *referenceEnd*). Vytvorili sme dataframe, ktorý má riadok pre každú bázu referenčného genómu (čiže 29903 riadkov) a stĺpce, ktoré opisujú niekoľko charakteristík pre danú bázu referenčného genómu pre konkrétneho pacienta:

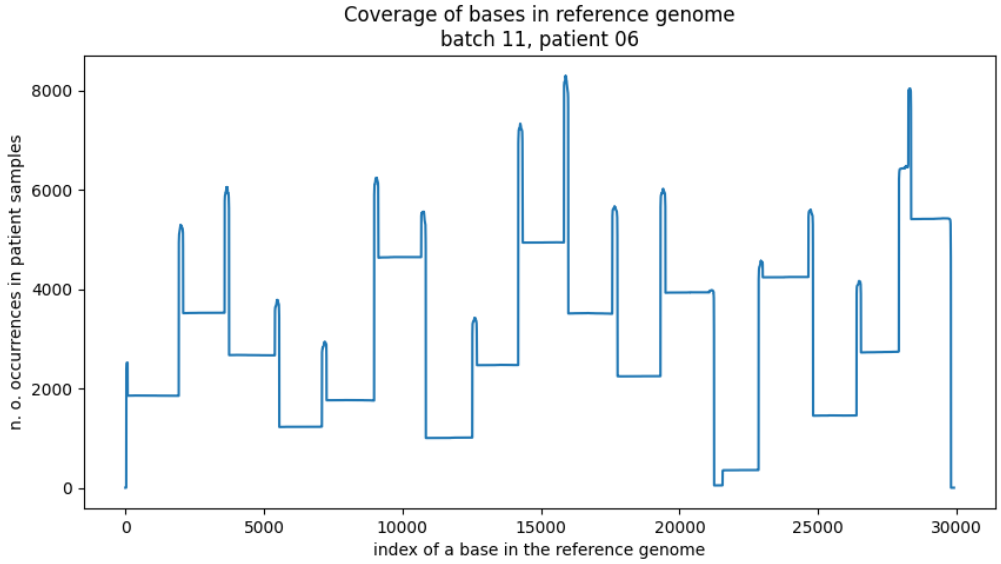
- ***base*** - index bázy referenčného genómu pre daný riadok (začíname číslovať od 1)
    - pre prehľadnejšie spracovávanie
- ***occs*** - koľkokrát bola daná báza v zodpovedajúcom regióne pre vzorku
    - všetky nasledovné štatistiky hovoria o nejakej charakteristike danej bázy referenčného genómu iba pre tie vzorky, v ktorých bola táto báza v zodpovedajúcom regióne
- ***sumStatusOK*** - koľko vzoriek malo *statusCode G.ok*
- ***sumStatusNotOK*** - koľko vzoriek malo *statusCode* iný ako *G.ok*
- ***avgEstQuality*** - priemerná odhadovaná kvalita čítania na danej báze
- ***avgIdenticalPercentage*** - priemerné percento zhodných báz v zodpovedajúcom regióne, v ktorom bola táto báza
- ***avgLengthReads*** - priemerná dĺžka čítania vzorke
- ***avgLengthMatch*** - priemerná dĺžka zodpovedajúceho regiónu na vzorku
- ***dominantStrand*** - ktoré vlákno DNA bolo viac zastúpené vo vzorkách

Pri analyzovaní týchto dát sme nakoniec použili iba tie vzorky, ktoré mali *G.ok statusCode*, aby pri vyvodení potenciálnych záverov sme mohli rátať s tým, že sa nejedná o chybu pri sekvenácii. Následne sme sa pozreli najmä na štatistiky *occs* a *avgEstQuality*, ktoré ako jediné viedli k záverom.

Predspracované dáta sú dostupné na stránke pod názvom *batchall_coverage*. Keďže sa jednalo o 71 tabuliek (pre každého pacienta jedna), spojili sme ich do jednej a pridali sme stĺpce *batchNum* a *patientNum* pre spätnú identifikáciu pacientov.

Pokrytie referenčného genómu

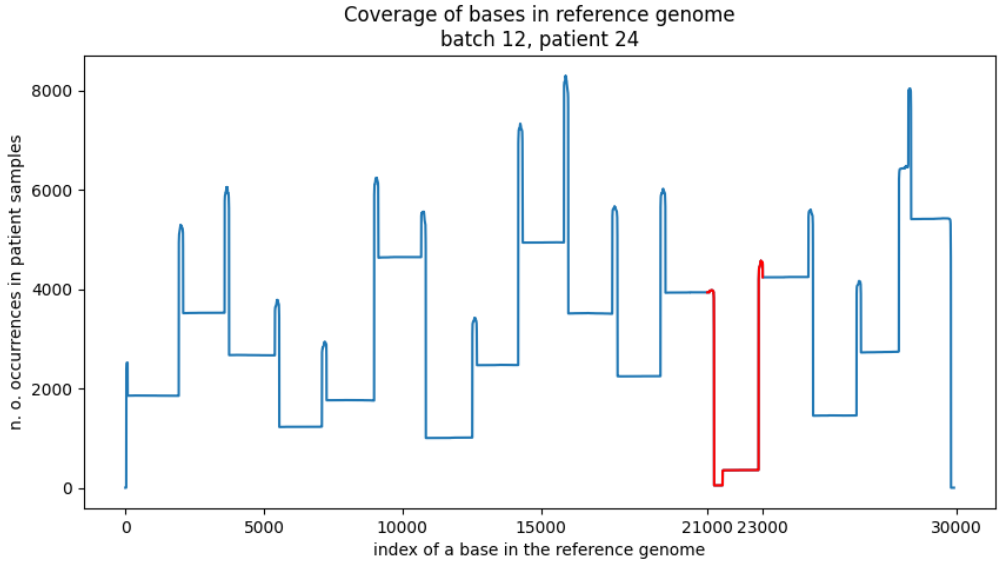
Stĺpec *occs* ukazuje, že koľkokrát bola daná báza referenčného genómu v regióne, ktorý najlepšie zodpovedá nejakej vzorke od pacienta. Túto štatistiku budeme volať pokrytie referenčného genómu, lebo hovorí o tom, ako jednotlivé vzorky a ich zodpovedajúce regióny pokrývajú referenčný genóm. Pozreli sme sa na to, či si vieme všimnúť nejaké trendy (napríklad, či nejaká časť referenčného genómu nebola v zodpovedajúcom regióne vôbec alebo naopak, príliš veľakrát).



obr. 13: Pokrytie referenčného genómu vzorkami od pacienta 6 z batchu 11

Prvý očividný trend, ktorý vidíme na *obr. 13* (ktorý bol konzistentný pre všetkých pacientov), sú horizontálnych úseky na grafe. To znamená, že napríklad všetky bázy v regióne od 1 do 2000 boli v zodpovedajúcom regióne podobne veľakrát. To je ale pravdepodobne výsledok PCR metódy a toho, že sme vyfiltrovali iba dáta, ktoré boli označené *G.ok*, čo mimoiného znamená, že majú dĺžku okolo 2000. PCR metóda pravdepodobne prispieva k tomuto trendu tým, že má preddefinované miesta, kde sa naviaže PCR primer a potom replikuje niekoľko báz v regióne za naviazaným úsekom. Keď to spojíme s faktom, že sme brali do úvahy iba tie vzorky, ktoré mali dĺžku 2000, dostávame z toho, že naozaj by sme mali vidieť iba horizontálne úseky na grafe. Takisto z toho môžeme usúdiť, že blízko okrajov týchto horizontálnych úsekov sa viazali PCR primery.

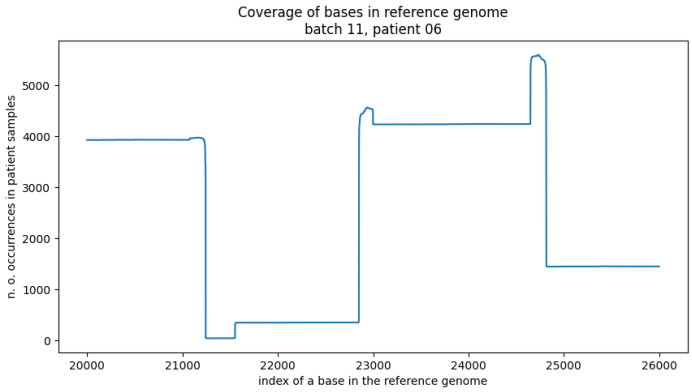
Väčšinou, síce rozmiestnenie horizontálnych úsekov bolo zhodné, ale ich poloha vo vertikálnom smere nie. Avšak úsek približne medzi bázami 21000 až 23000 mal konzistentne nízke hodnoty pre všetkých pacientov, ako vidno na *obr. 14*.



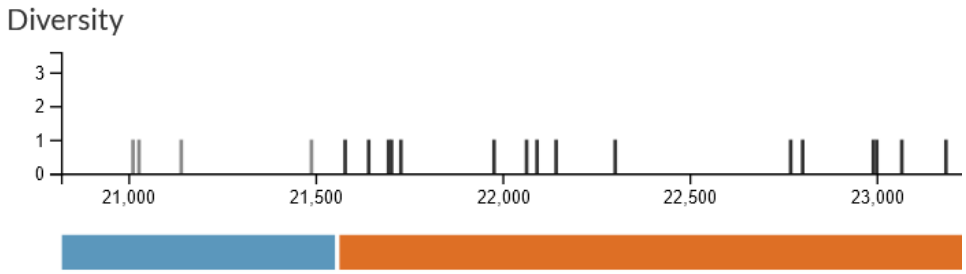
obr. 14: Zvýraznený úsek (červenou), ktorý mal konzistentne nízke hodnoty pri všetkých pacientoch

Keďže sa tento trend nachádzal vo všetkých 71 pokrytiach referenčného genómu, hľadali sme čo to môže spôsobovať. Naše dáta neobsahujú konkrétne sekvencie, z ktorých by sme mohli niečo zistiť alebo iné údaje, ktoré by nám pomohli zistiť príčinu takéhoto trendu, preto sme hľadali nejaké vysvetlenie v iných analýzach.

Skonštruovali sme hypotézu, že takýto prepad pokrytia môže byť spôsobený mutáciou v úseku DNA, na ktorý sa mal PCR primer naviazať. Zmena tohto úseku by spôsobila jeho nenadviazanie a teda následné nízke amplifikovanie úseku, za ktorý bol „zodpovedný" tento primer. Túto hypotézu posilňoval fakt, že v regióne medzi 21000. a 23000. bázou začína gén tzv. spike proteínu. Tento proteín slúži vírusu na vnikanie do buniek, ale zároveň sa naše ľudské imunitné bunky vedia naučiť, ako tento spike proteín vyzerá a pri ďalšom stretnutí vedia adekvátnejšie reagovať. Mysleli sme si teda, že práve toto bude región, ktorý bude často mutovať, aby bolo ťažšie pre naše telo rozpoznať tento koronavírus.



obr. 15: Skúmaný úsek s konzistentne nízkym pokrytím



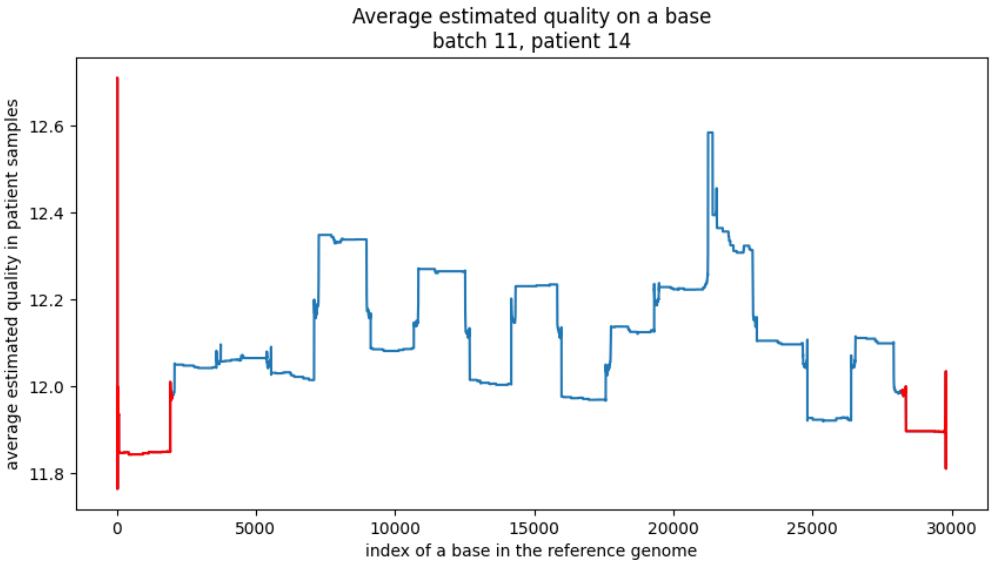
obr. 16: Dáta z iniciatívy Nextstrain ukazujúce mutácie Alfa (B.1.1.7) variantu koronavírusu medzi 21000. a 23000. bázou spolu s označením génov ORF1b (modrá) a génu pre spike proteín (oranžová)

Z *obr. 15* vidíme, že klesajúci trend začína za 21000. bázou. Z analýzy dát z iniciatívny Nextstrain (4) sme zistili, že sa v regióne okolo 21000. bázy nachádzajú mutácie (*obr. 16*), aj keď to nie sú mutácie génu pre spike proteín, ako sme pôvodne predpokladali. Z toho môžeme vyvodiť záver, že skúmaný trend by mohol byť spôsobený mutáciou v regióne potrebnom pre naviazanie PCR primera, čo znižuje replikáciu daného regiónu genetickej informácie. Pre úplné potvrdenie tejto hypotézy by však bolo treba rozsiahlejší výskum.

Ak by sme predpokladali, že hypotéza je pravdivá, môžeme po objavení sa takéhoto trendu v podobných dátach predpokladať, že PCR primer sa nedokázal naviazať na svoj komplementárny úsek kvôli mutácii genetickej informácie. To by mohlo implikovať potrebu redizajnovania tohto PCR primera.

Priemerná odhadovaná kvalita referenčného genómu

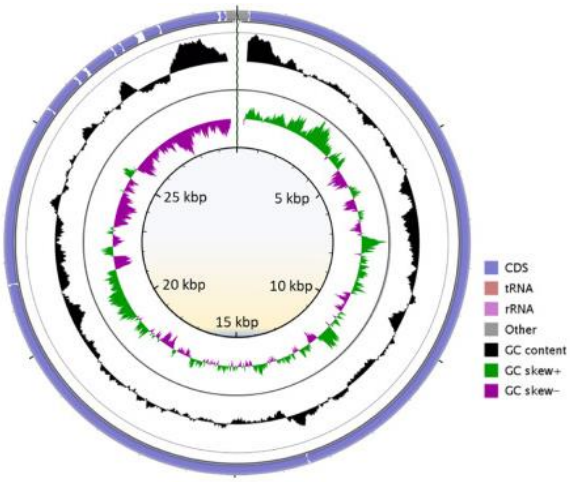
Ďalej sme sa pozreli na to, aká bola priemerná odhadovaná kvalita pre jednotlivé bázy referenčného genómu, spomedzi všetkých zodpovedajúcich regiónov, ktorým patrili.



obr. 17: Graf priemernej odhadovanej kvality pre konkrétne bázy referenčného genómu spolu so zvýraznenými úsekmi (červenou), na ktorých bolo vidno konzistentný trend nižšej kvality oproti okoliu

Podobne ako predtým, aj tu sme si všimli jeden konzistentný trend - konkrétne nižšia odhadovaná kvalita na prvých a posledných približne 2000 bázach, oproti ich okoliu, ako vidno na *obr. 17*. Sformulovali sme hypotézu, že tento trend je taktiež spôsobený horším naviazaním PCR primerov na začiatku a konci DNA, resp. horšou replikáciou začiatočných a koncových regiónov. Bohužiaľ, naše dáta nepostačovali na to, aby sme hypotézu potvrdili a taktiež sme nenarazili na žiadnu štúdiu, ktorá by sa vyjadrovala o takejto problematike.

Vieme však, že jedno z kritérií pre kvalitnejšiu a spoľahlivejšiu sekvenáciu DNA, je aby pomer báz adenínu a tymínu (AT podiel) ku guanínu a cytozínu (GC podiel) vo vzorke bol 50:50. Ahmad Abu Turab Naqvi s tímom v článku o štruktúre genómu Sars-CoV-2 (3) okrem iného sledovali aj práve GC podiel v regiónoch DNA koronavírusu.



obr. 18: Graf z článku od profesora Naqviho, kde čierny prstenec zobrazuje GC podiel pre jednotlivé časti genetickej informácie koronavírusu

Na *obr. 18* čierny prstenec zobrazuje GC podiel v DNA koronavírusu. Môžeme si všimnúť, že začiatočný a koncový región (na prstenci hore vpravo a hore vľavo) majú tento podiel zvýšený. Z toho by sme mohli usúdiť, že pozorovaný trend v našich dátach by mohol byť spôsobený práve zvýšeným GC podielom.

Záver

Počas našej analýzy sme sa zo začiatku hlavne venovali opisu premennej odhadovanej kvality, chceli sme predikovať z nej nejaké iné premenné alebo naopak odhadnúť kvalitu na základe nejakej premennej. Našli sme Spearmanovu koreláciu 0.6 medzi odhadovanou kvalitou a percentom zhodných báz, z čoho sme usúdili, že kvalita bude závislá od tohto percenta. Tiež sme si potvrdili, že výskyt nontarget organizmov (všetko okrem covidu) má výrazný vplyv na kvalitu, keďže sa takmer nevyskytujú vo výšších kvantiloch kvality.

Pri porovnaní batchov sme sa zamerali na pokles kvality, tu sme odhalili, že s najvyššou pravdepodobnosť tento pokles spôsobil väčší počet pacientov s neurčenými bázami v batchoch 11,12.

Pri analýze referenčného genómu sme poznamenali trendy spôsobené pravdepodobnými mutáciami na regiónoch potrebných pre PCR primer. Následne sme vyvodili záver, že pri pozorovaní potrebných dát je možno potrebné predizajnovať tento primer.

Zoznam literatúry

1. Brejová B., http://compbio.fmph.uniba.sk/~bbrejova/tmp/covid-viz/, 15.5.2023
2. dokumentácie knižnice scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
3. Naqvi A. A. T and team, 2020, *Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7293463/, 13.5.2023
4. Nextstrain, *Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months*, https://nextstrain.org/ncov/gisaid/global/6m, 13.5.2023
5. tím ŠPARK, http://www.st.fmph.uniba.sk/~spitalsky3/viz-projekt/viz-projekt.html, 15.5.2023
6. Wu F. and team, 2020, *Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome*, https://www.ncbi.nlm.nih.gov/nuccore/MN908947, 13.5.2023