ŠPARK - denník

**Názov skupiny: ŠPARK**

**Téma projektu: COVID**

**Členovia skupiny: Antal, Špitalský, Pazera, Rohaľ, Kendereš**

# Obsah pre rýchlu navigáciu

[**Obsah pre rýchlu navigáciu 1**](#_c7qsir6sgcp7)

[**Kontrolné stretnutia 2**](#_wzrubvfy0074)

[Stretnutie 5.4. 2](#_s48u2m738fny)

[Stretnutie 12.4. 2](#_23p55d2570yh)

[Stretnutie 26.4. 2](#_x3nk9vqce3cv)

[Stretnutie 3.5. 2](#_uozx3zrqrmvy)

[**Práca na projekte 3**](#_tl9hxjsgxx6c)

[2023-03-28 stretnutie 1 3](#_clxl7sla8djh)

[2023-04-03 stretnutie 2 3](#_9nm8fuz4kx7r)

[2023-04-05 stretnutie 3 3](#_4lju21y3hxge)

[2023-04-06 až 2023–04-11 samostatné analýzy 4](#_ehushc5h6xhc)

[2023-04-12 potenciálne veľké otázky na skúmanie 4](#_56jnsovfl0oh)

[2023-04-23 stretnutie x 4](#_x99xkvohpe5p)

[2023-4-28 vypracované otázky 5](#_lbd080acxkrs)

[2023-5-9 stretnutie pred prezentáciou 5](#_nom15elhcjhk)

[2023-5-15 dokončovanie správy a notebooku 5](#_3n27i97rv7rm)

[**Zhrnutie práce jednotlivých členov skupiny 6**](#_35bcerl26988)

[**Povolenie na zverejnenie projektu 6**](#_wcboxdig4nt2)

# 

# Kontrolné stretnutia

## Stretnutie 5.4.

1. Načítali sme dáta, prvotné exploratívne analýzy
2. Hľadáme, na čo sa zamerať

## Stretnutie 12.4.

1. 4 štádiá = dáta, analýza, vizualizacia, interpretácia - storytelling, vedieť povedať babke o 2 týždne
2. Budúci týždeň - mať 99% hotové, hotová vizualizácia vymýšlať ako interpretovať
3. Naštudujeme ArticProtocol k pokrytiu genómu
4. Vypracujeme odpovede na položené veľké otázky

## Stretnutie 26.4.

1. Spýtať sa na legalitu použitia decision-tree
2. Alpha variant quest line concluded
3. Task s referenčnými genómami dokončený
4. Hexplot namiesto scatterplotu pri regresii
5. Formulácia záveru do budúcna - otázka a odpoveď -> hypotéza a conclusion

## Stretnutie 3.5.

1. Uzavreté otázky
2. Naformulovať do zmysluplných viet
3. Spracovat do prezentácie

# 

# Práca na projekte

## 2023-03-28 stretnutie 1

Dokončené úlohy

* spojené tabuľky pre match, read
* k barcodom bol pridaný variant

Úlohy na preskúmanie

* kvalita readov v porovnaní s dĺžkou readov
* barcode 14 - prečo tak veľa neurčených báz
* úspešnosť readu voči variantu
* mutácie voči genómu
* zistiť koreláciu medzi short/long readmi a kvalitou
* ploty všetkého
* pozrieť sa na kvalitu testov over time
* znormalizovať dáta
* prečo je pri inom subjekte ako covide nižšia kvalita readu
* plotovať estQuality s rôznymi subjektmi
* ako je určená chyba čítania v results

## 2023-04-03 stretnutie 2

* Files s dátami - davinci server
* Vieme načítať dáta
* problém s nacitaj\_data funkciou (fixed)
* Problematic reads sú: nontarget, nízka estQuality
* describe() na dataset (na spojený reads a match)
* vyfiltrovali sme barcode1 unclassified
* podľa barcodu sme priradili každému riadku pango (variant) - máme spojené všetky tri tabuľky, z `rslts` nás zaujímal variant
* zadefinovanie low quality reads
* porovnat kolko je nontargetov v low quality tabuľke oproti celej - possible korelácia
* zistili sme, že frequency nontargetov klesá so stúpajúcou kvalitou readov (!)

## 2023-04-05 stretnutie 3

* Potvrdili sme vzťah non-target samplov s est. kvalitou readov
* Zistili sme, že v tabuľkách *‘match’* sú ID-čka duplicitné, nevieme prečo
* Spravili sme si tabuľku na paralelizovanie analýz
* Skúsili sme heatmap a scatter plot PCA, ale nevedeli sme vyvodiť závery

## 2023-04-06 až 2023–04-11 samostatné analýzy

* Mergenuté dataframes sú nahrané na stránke, pracuje sa už iba s nimi
* Pozorovali sme rozdielne aspekty medzi batchmi: väčšina štatistík veľmi podobná
  + odhadovaná kvalita medzi batchmi klesá - možno súvis s prevalenciou alfa variantu?)
* Pozerali sme sa, aké charakteristiky sa menia s časom - inconclusive
* Pre každého pacienta sme vykreslili pokrytie referenčného genómu jednotlivými reads - potenciálny záver, že sa reads mapujú len na určité regióny, ktoré sa prekrývajú na hraniciach
* Hľadali sme súvis medzi počtom nezistených báz pre pacientov a zisteným variantom vírusu u nich - variant B.1.1.170 má neprirodzene veľké hodnoty

## 2023-04-12 potenciálne veľké otázky na skúmanie

1. Dá sa na základe rozdelenie estimated kvality odhadovať rozdelenie dát ako rozdelenie variantov, dĺžky sekvenácií či počet non-target organizmom?
2. Koľko čítaní pokrýva jednotlivé časti referenčného genómu (zo súboru C)?  
   Súvisí počet báz s nízkym pokrytím skutočne s počtom báz, ktoré nebolo možné určiť v súbore A? Je priemerná dĺžka alebo kvalita čítaní v rôznych častiach referenčného genómu odlišná?
3. Čo spôsobilo zníženie kvality medzi batchmi a či to je ovplyvnené prevalenciou alfa variantu?

## 2023-04-23 stretnutie x

* Očakávali sme, že alpha bude mať nižší estQuality ako ostatné varianty, ale ukázalo sa, že vo všetkých batchoch bol rozdiel medzi alphou a ostatnými zanedbateľný (boli priblizne rovnake) alebo alpha mala vyššiu kvalitu
* Zníženie kvality medzi batchmi je s najväčšou pravdepodobnosťou spôsobené výskytom viacerých pacientov, ktorých vzorky mali počet notDetermined báz > 200
* Keď sme sa pozerali na pokrytie referenčného genómu jednotlivými samples pre konkrétneho pacienta, zistili sme, že región medzi 20000\*25000 je konzistentne omnoho menej pokrytý, ako zvyšok. To môže napovedať, že v tomto regióne bola mutácia na mieste, kde sa mal naviazať PCR primer (je to región génu Spike proteínu - práve ten mutuje najviac).
* Zároveň, keď sa pozrieme na priemernú estimated kvalitu na danej báze referenčného genómu, zistíme, že prvých +-2000 báz má nižšiu kvalitu, ako zvyšok, čo korešponduje s vysokým GC zastúpením v genómu covidu, čo spôsobuje menej spoľahlivú sekvenáciu. Podobný trend len v nižšej miere je aj na koncovom regióne, kde by malo byť GC zastúpenie tiež vyššie

2023-4-27 prezentácia

* Diskusia o prezentácii, čo tam dať, čo vynechať
* Šuppova prednáška, watch session - inšpirácia
* Osnova prezentácie spísaná
* Prezentujúci rozdelení
* Úvodný slide done
* Template dohodnutý

## 2023-4-28 vypracované otázky

* Čo spôsobilo zníženie kvality medzi batchmi a či to je ovplyvnené prevalenciou alfa variantu? (Teo,Tomáš)
* Koľko čítaní pokrýva jednotlivé časti referenčného genómu? (Andrej)
* Či sa niečo mení s variantom - popisné štatistiky (Rafko)
* Dá sa na základe rozdelenie estimated kvality odhadovať rozdelenie dát ako rozdelenie variantov, dĺžky sekvenácií či počet non-target organizmom? (Robo, Rafko)

## 2023-5-9 stretnutie pred prezentáciou

* Dokončovanie prezentácie
* Dohodnutie nadväznosti kapitol
* Nacvičenie prezentácie

## 2023-5-15 dokončovanie správy a notebooku

* Finalizovanie všetkých súborov - notebook, správa, denník, stránka

# Zhrnutie práce jednotlivých členov skupiny

* Špitalský - spracovanie *merged* tabuliek, spracovanie stránky, pokrytie referenčného genómu
* Antal, Pazera - analýza kvalitných čítaní (označených G.ok), porovnávanie batchov, závislosť odhadovanej kvality a výskytu organizmov klasifikovaných inak ako SARS-CoV-2
* Rohaľ, Kendereš - porovnávanie variantov, hodnoty závislé od odhadovanej kvality, štatistiky vzhľadom na čas

# Povolenie na zverejnenie projektu

**Zverejnenie finálneho projektu (správy a notebooku) - nechajte iba tú z troch možností, s ktorou súhlasíte**

* Všetci členovia skupiny súhlasia s úplne voľným zverejnením projektu, napr. na stránke predmetu.