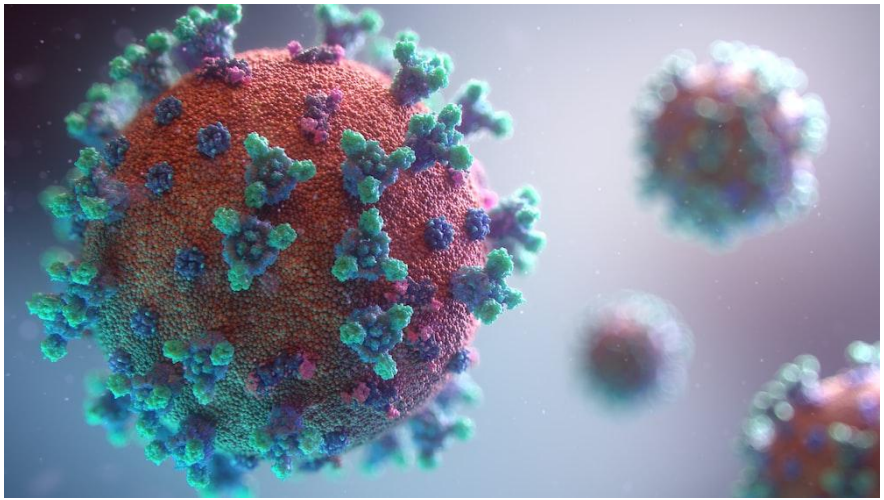
COVID



ŠPARK
(Špitalský, Pazera, Antal, Rohaľ, Kendereš)

Úvod: Zhrnutie dát

Ako boli dáta prvotne získané?

- dáta z UK

Ako vyzerajú dáta, ktoré sme dostali?

- dáta z troch sekvenčných behov
- každý beh obsahuje údaje pre vzorky pacientov SARC-CoV-2.
- pre každý beh tri tabuľky: *results*, *reads*, *match*
- dokopy 9 tabuliek

Súbory spracované do *merged* foriem pre jednoduchšie narábanie

Úvod: Zhrnutie dát

- dáta nahrané na DaVinci cluster
- v notebookoch z neho načítavame dáta
- rýchlejšie

Súbory vizualizácia dát

Pôvodné súbory

- batch10-results
- batch10-reads
- batch10-match
- batch11-results
- batch11-reads
- batch11-match
- batch12-results
- batch12-reads
- batch12-match

Spracované súbory

- batch10-merged
- batch11-merged
- batch12-merged
- batchall-merged

Úvod: Zhrnutie dát

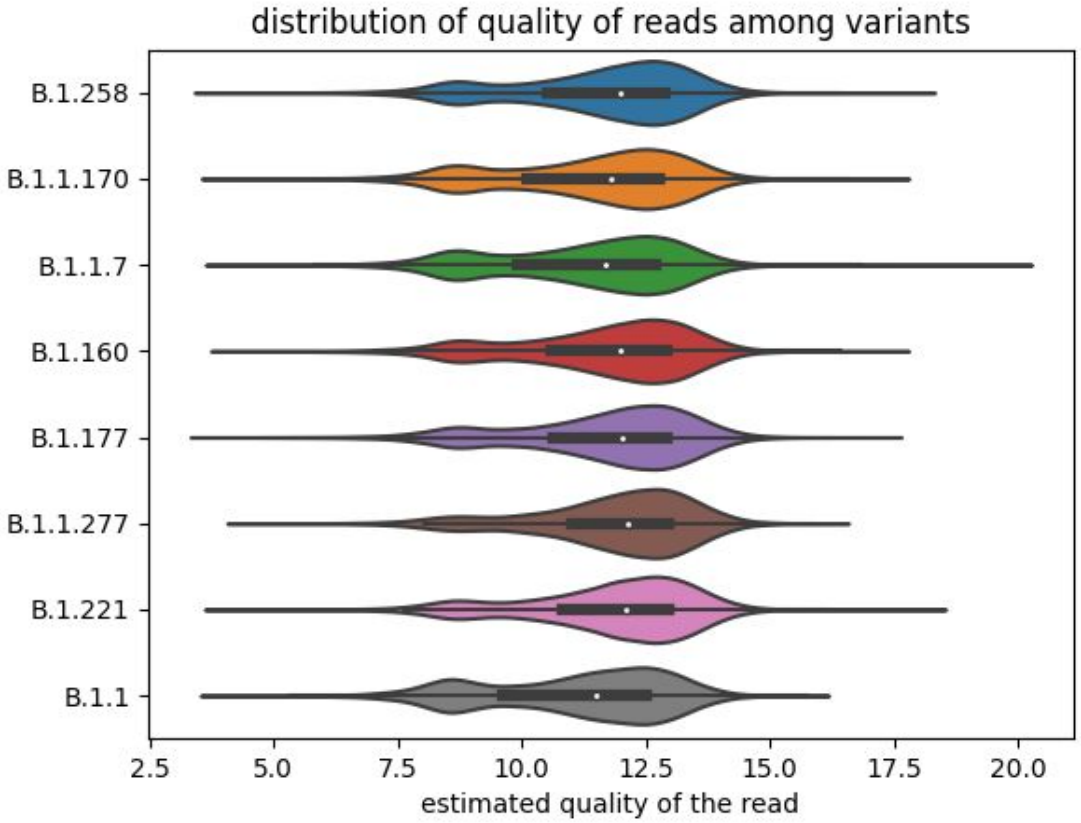
Niektoré dôležité stĺpce v tabuľkách:

- **barcode** - unikátny identifikátor
- **pango** - variant Covidu
- **estQuality** - odhadovaná kvalita vzorky
- **organism** - organizmus, z ktorého vzorka pochádza (Covid, virus, none)
- **notDetermined** - počet báz, ktoré neboli úspešne prečítané

Najviac nás zaujímala *estQuality*

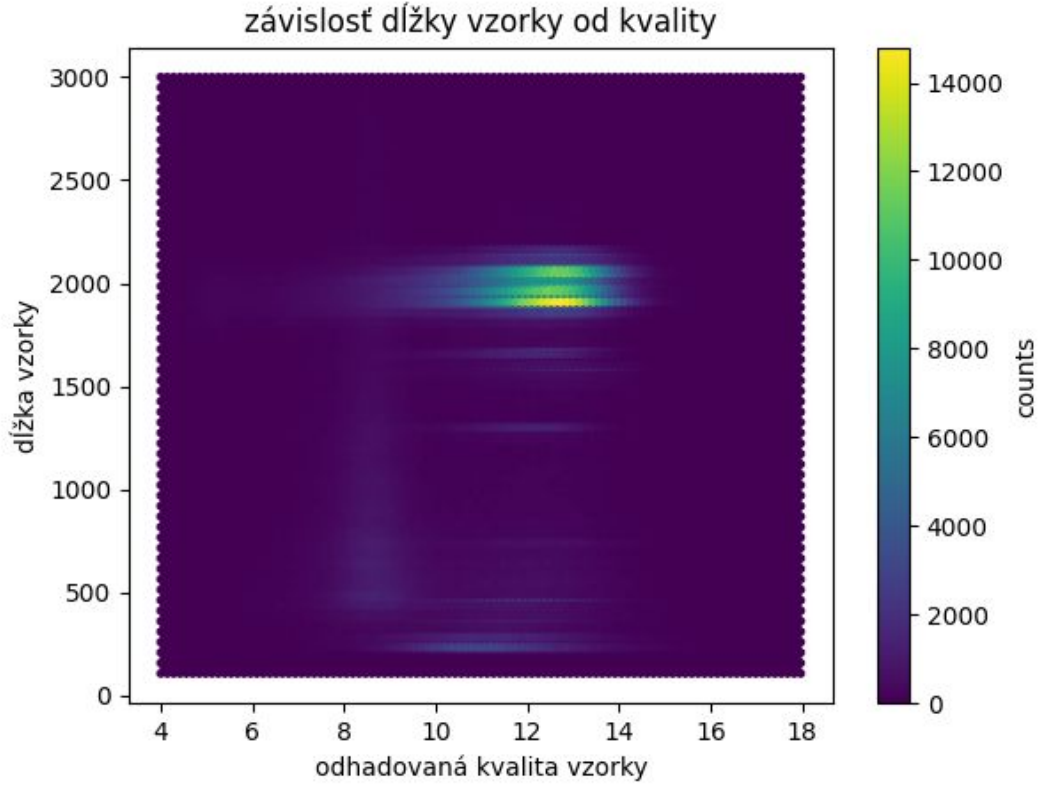
Korelácie medzi kvalitou a inými vlastnosťami readov

- na prvý pohľad sme predpokladali, že by sa dal predpovedať variant a dĺžka vzorky
- **variant:**

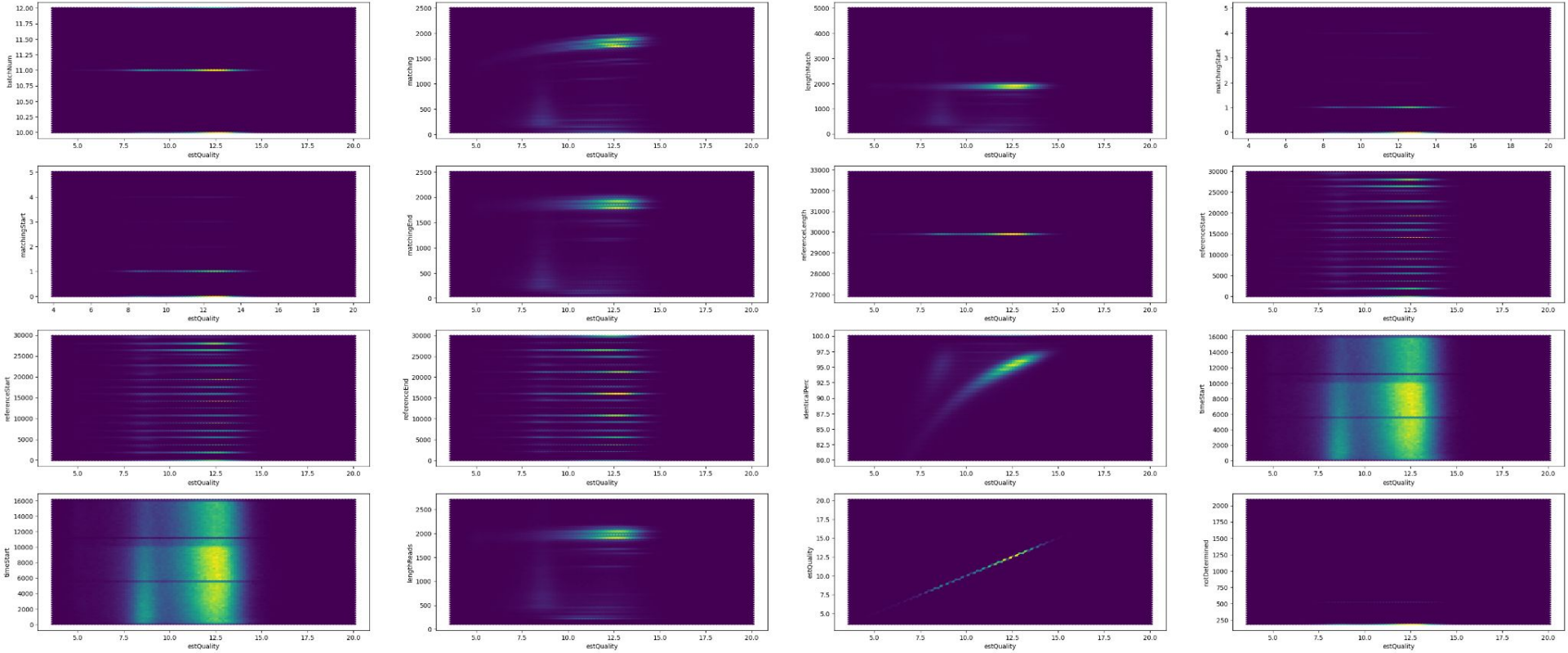


Korelácie medzi kvalitou a inými vlastnosťami readov

- **dĺžky vzoriek:**
- korelácia:
  0.2804
- slabá závislosť



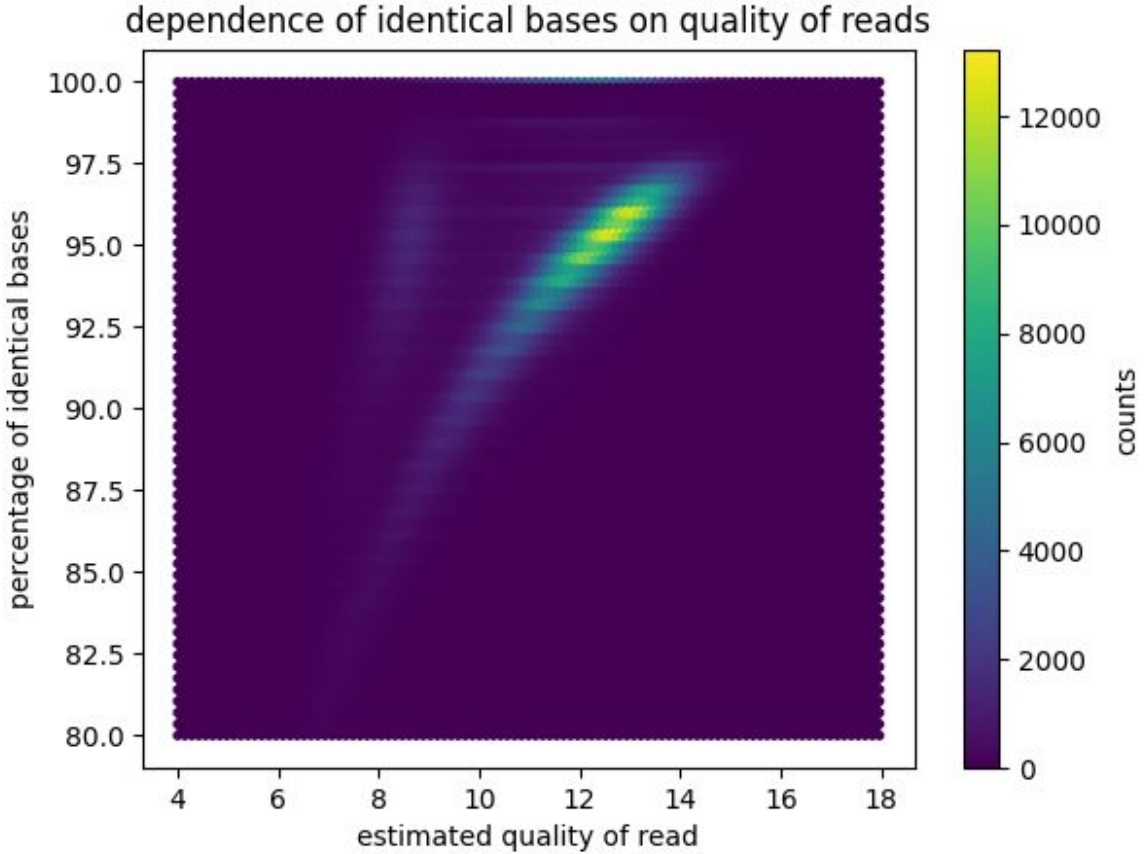
Dá sa podľa kvality odhadovať čokoľvek iné?



Kvalita a percento báz zhodných s referenčnou vzorkou

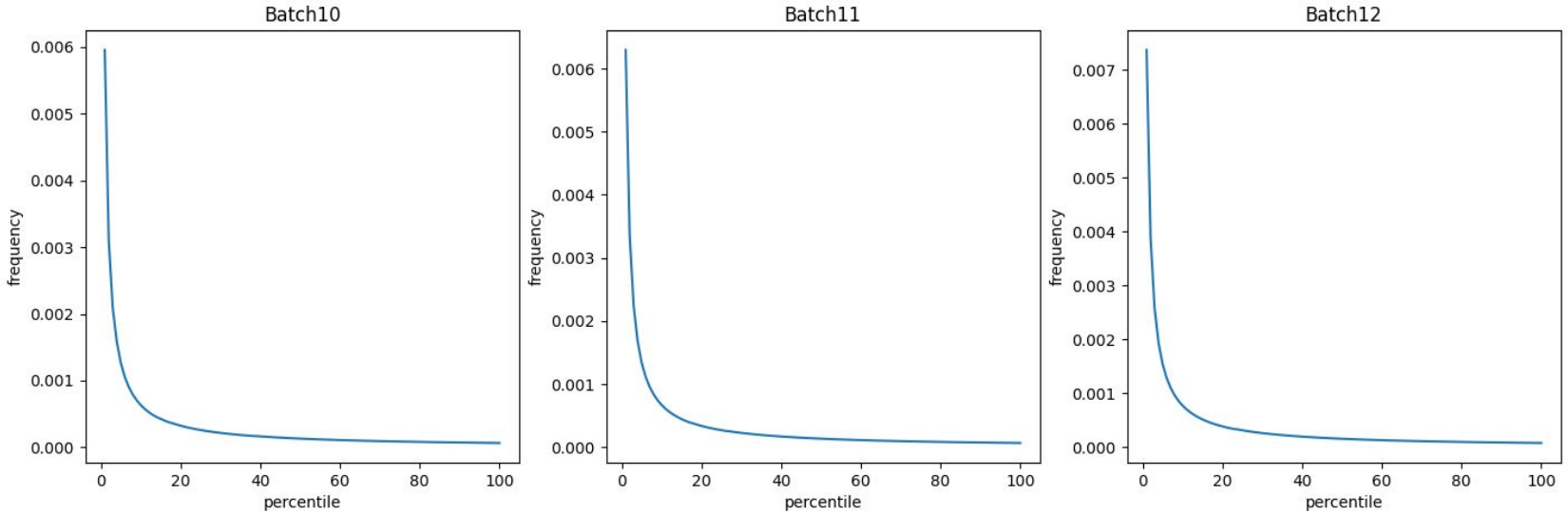
- korelácia: 0.62718
- jasná závislosť
- čím je read kvalitnejší, tým viac by sa mal podobať referenčnej sekvenácii

- decision tree regressor (sklearn)



Výskyt organizmov klasifikovaných ako “non-target”

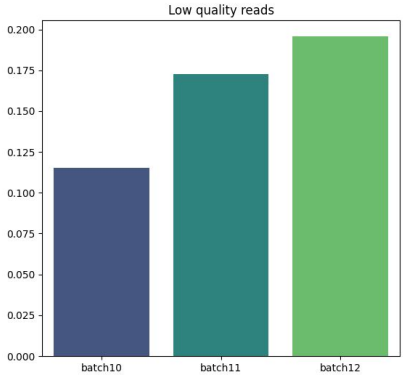
- percentil stĺpca “estimated quality”
- očividná klesajúca tendencia

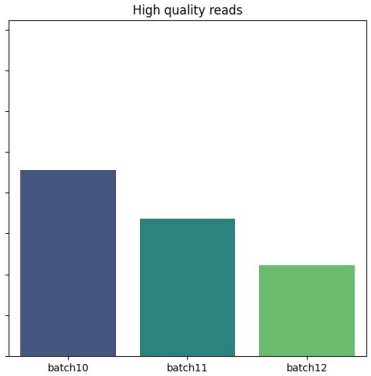


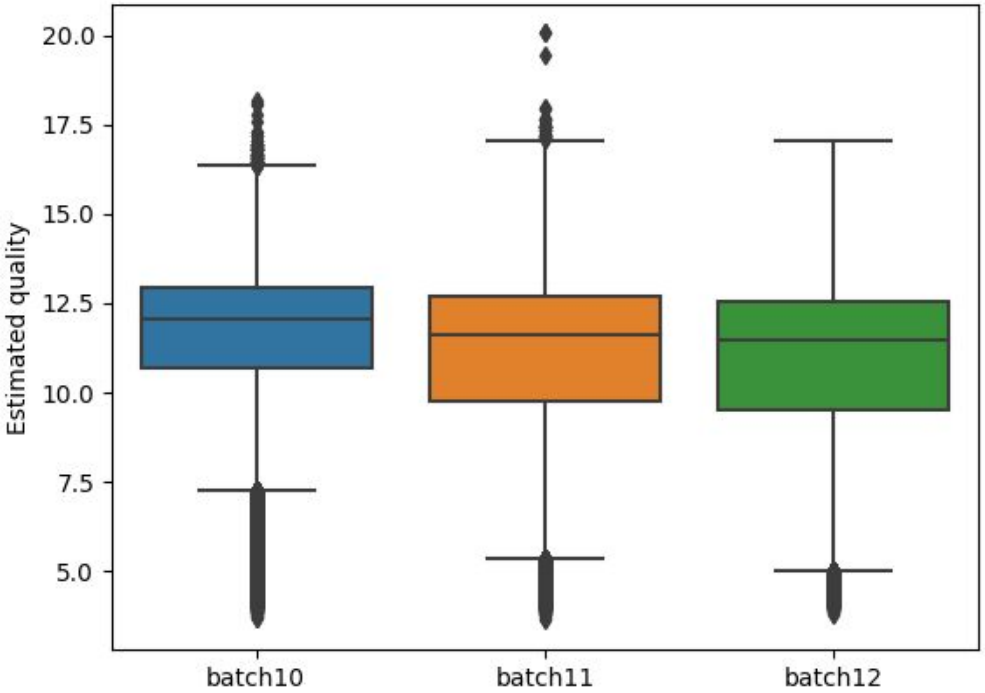
Otázka: Je rozdiel medzi “batch” súbormi?

odpoveď: ÁNO (aspoň si myslíme)

- prvotný pohľad na veľa aspektov
- estimated quality

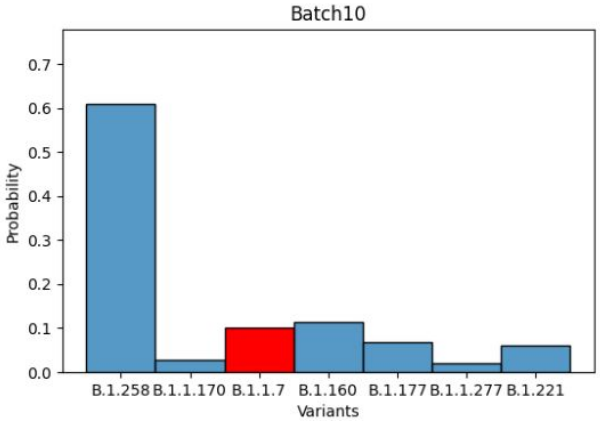


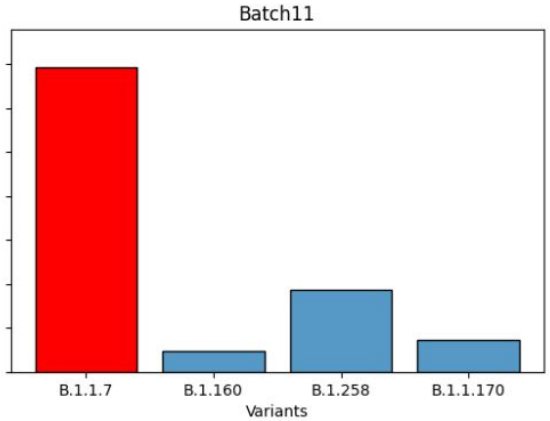


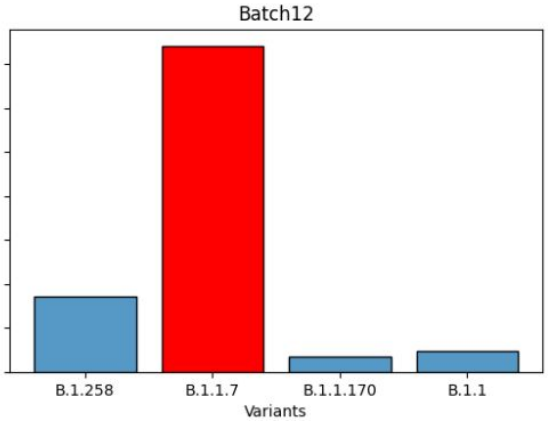


Hypotéza

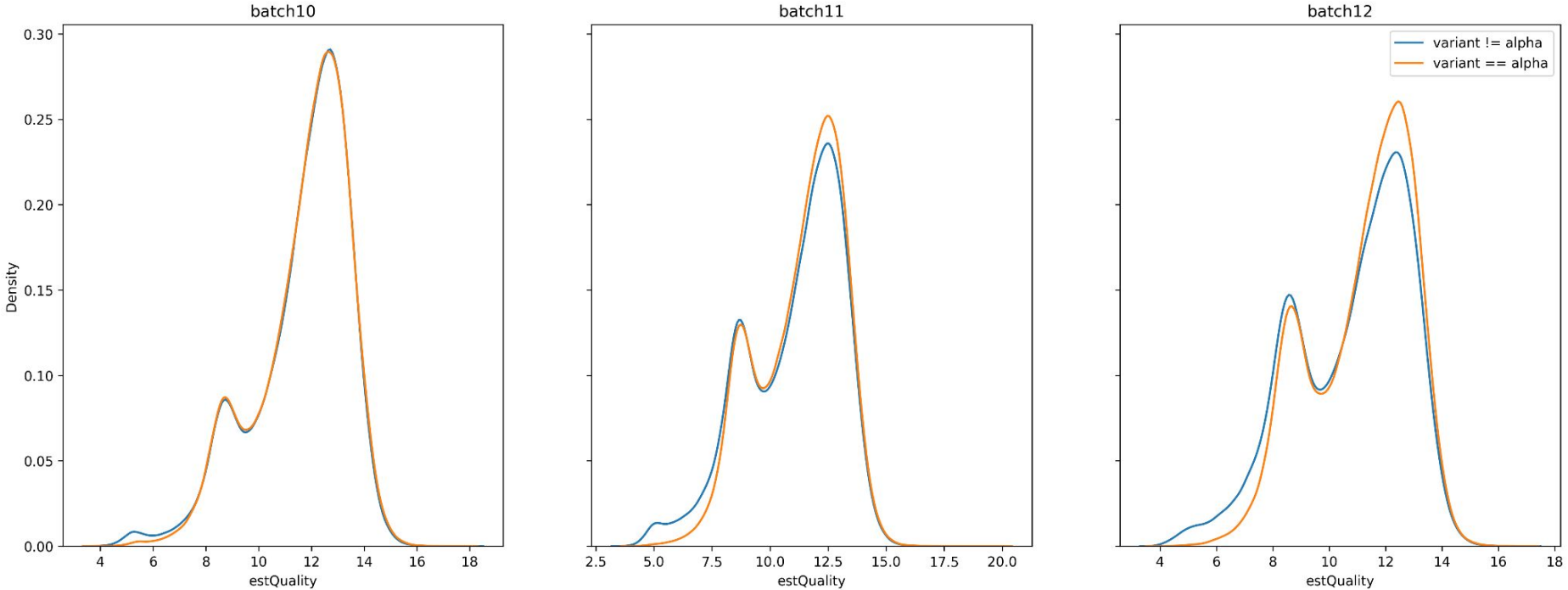
- Zníženie kvality kvôli prevalencii alfy variantu?





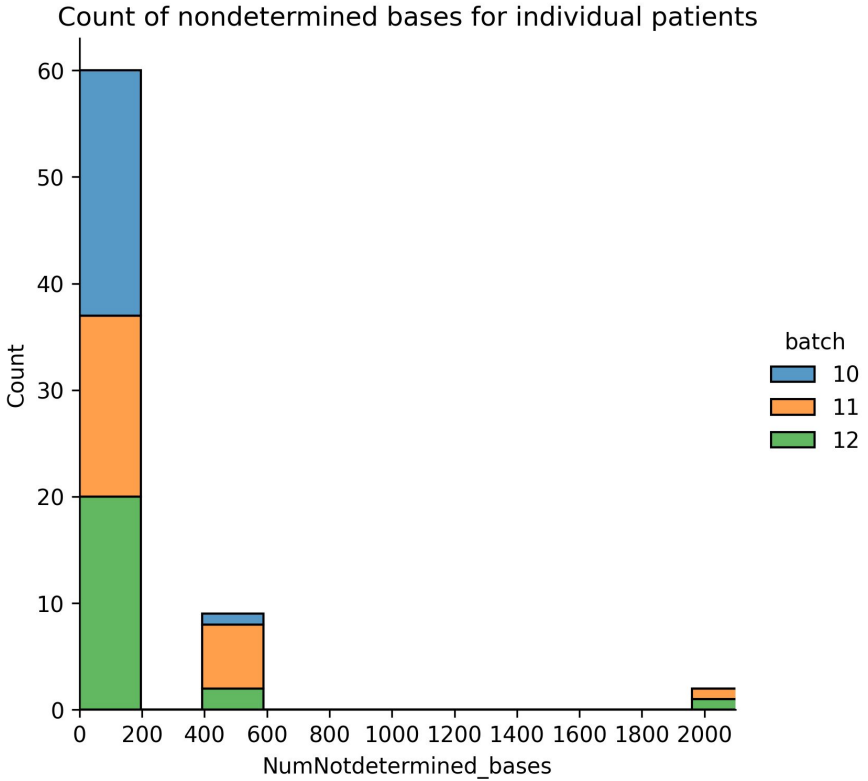


Hypotéza bola vyvrátená



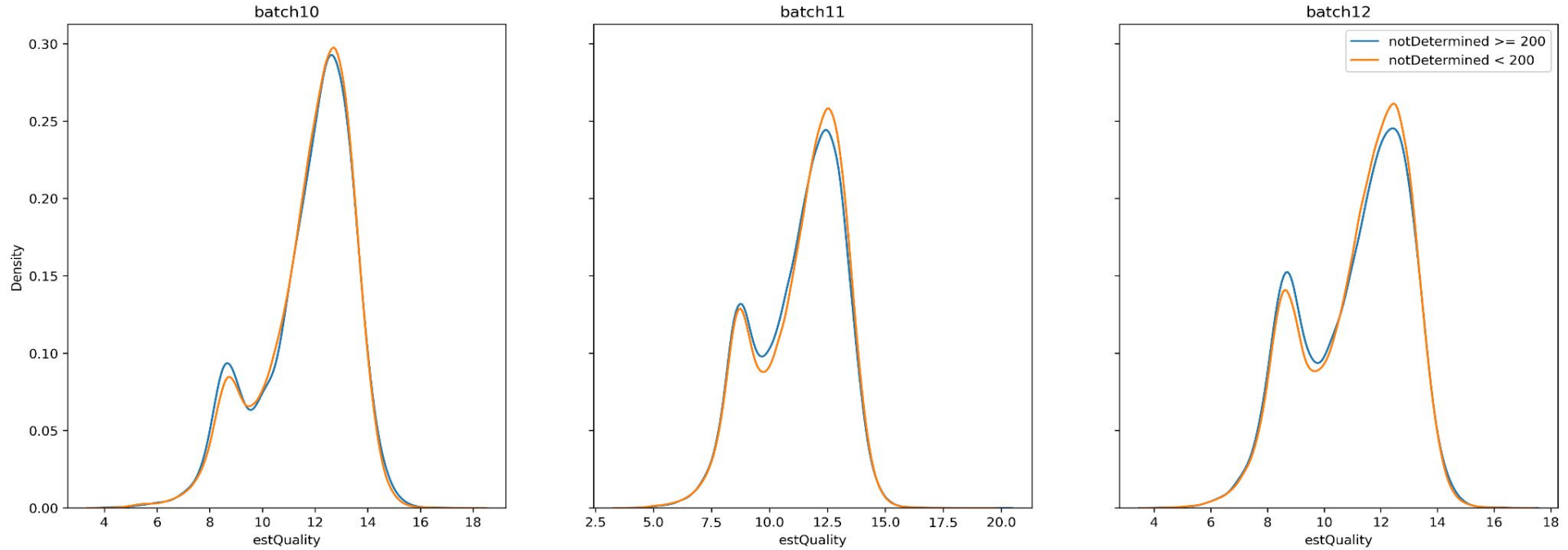
Čo to teda spôsobilo?

- všimli sme si väčší vyskyt pacientov s 200 > undetermined bázami v batchoch 11 a 12
- neurčenosť báz ➡ menšie pokrytie ➡ nižšia kvalita



Hypotéza bola pravdivá

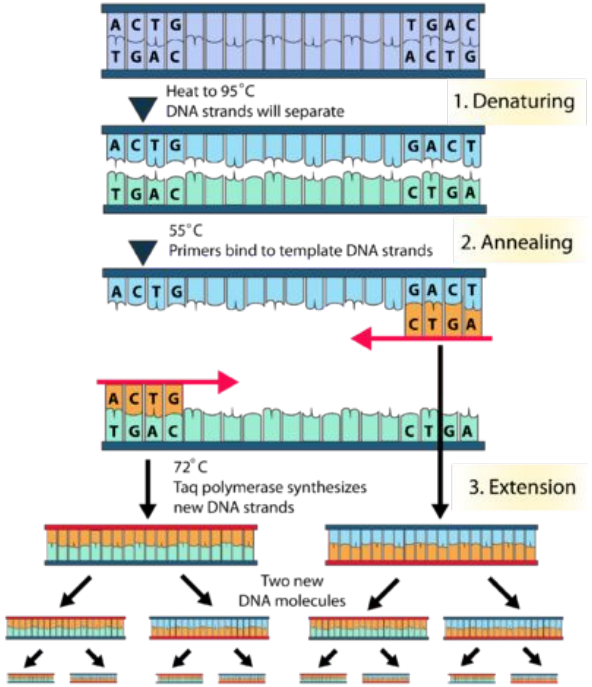
- Nárast notDetermined báz u pacientov s vysokou pravdepodobnosťou spôsobil zníženie kvality medzi batchmi



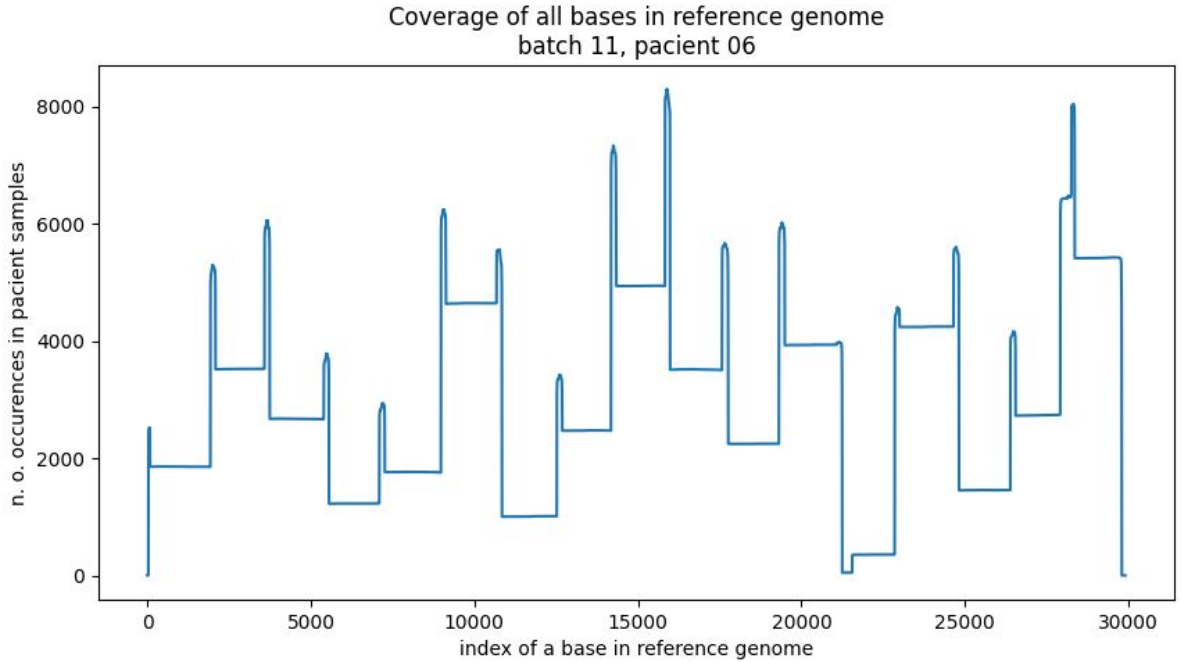
Pokrytie referenčného genómu

- https://www.ncbi.nlm.nih.gov/nuccore/MN908947
- PCR - nekonzistné kopírovanie úsekov
- spracovanie iba kvalitných vzoriek - dĺžka okolo 2000, vyššia odhadovaná kvalita

```
LOCUS       MN908947               29903 bp ss-RNA     linear   VRL 18-MAR-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   MN908947
VERSION     MN908947.3
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29903)
  AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y.,
            Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H.,
            Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z.
  TITLE     A new coronavirus associated with human respiratory disease in
            China
```

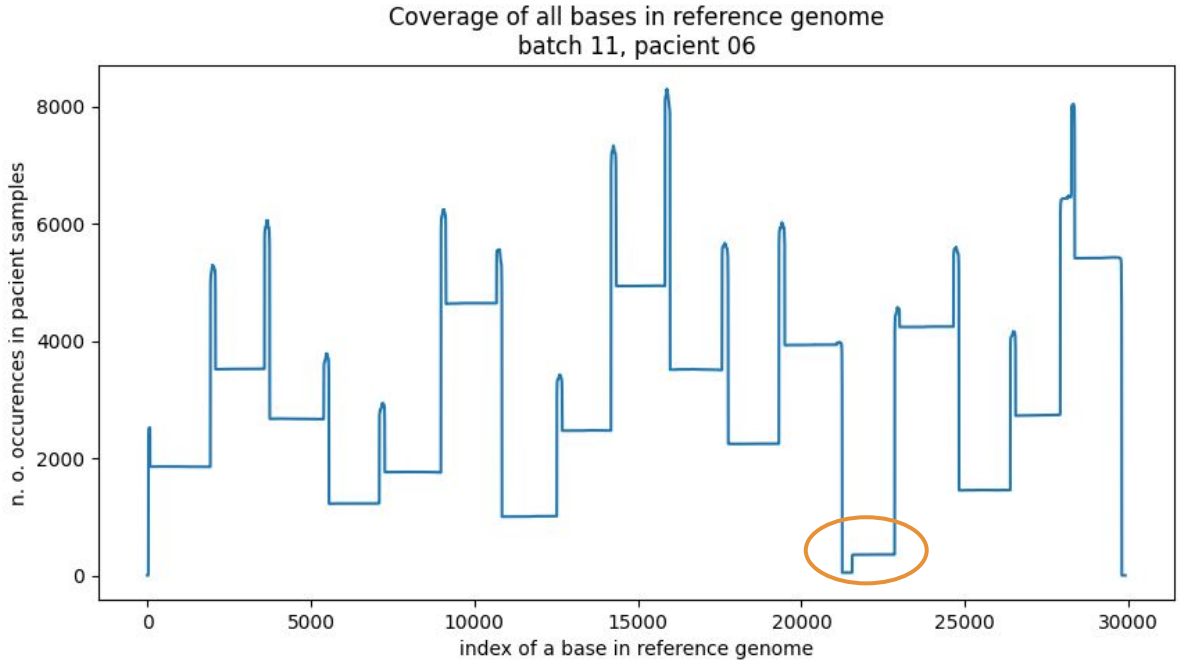


Pokrytie referenčného genómu

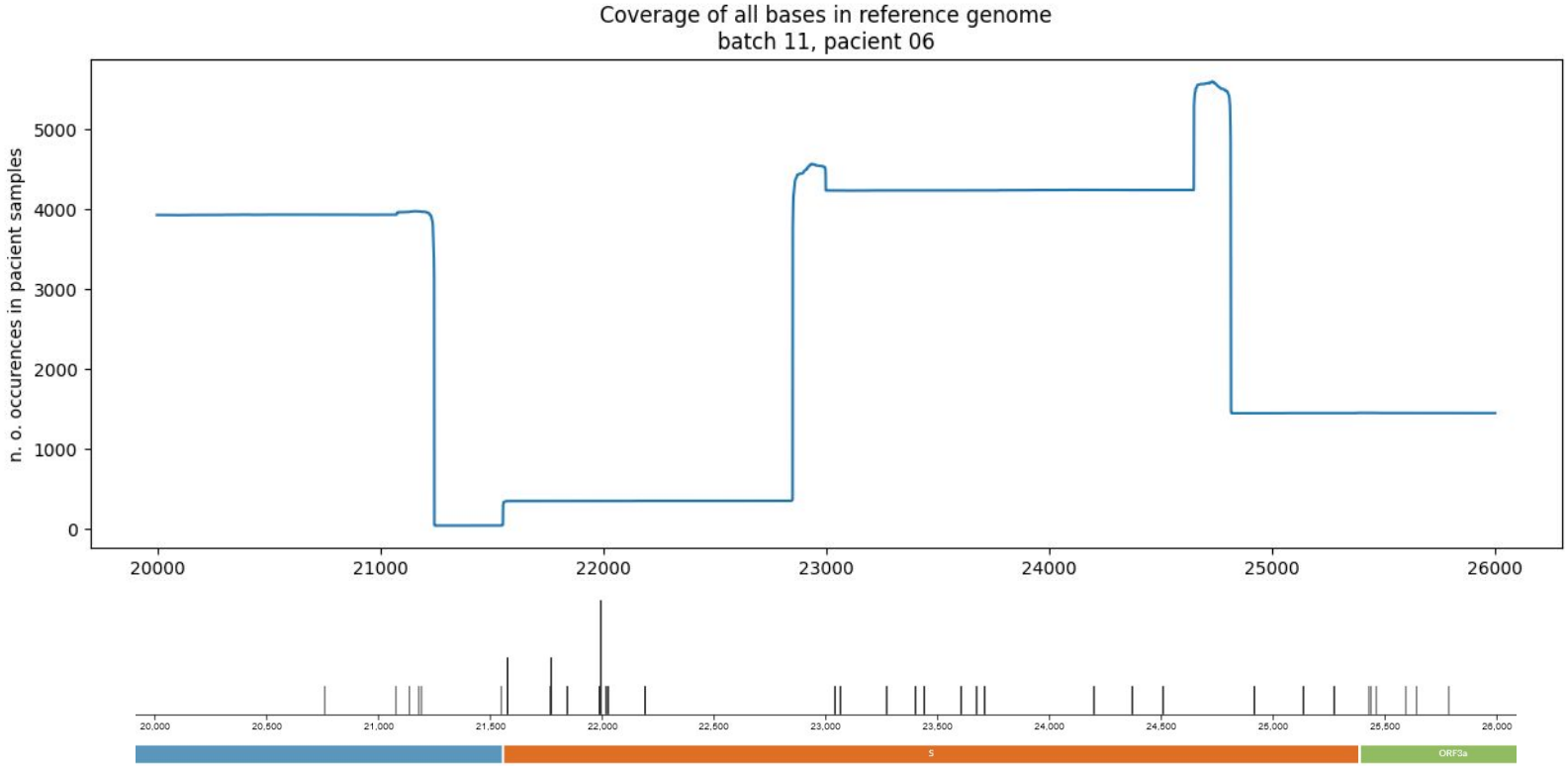


Pokrytie referenčného genómu

- hypotéza - mutácia



Mutácie spike proteínu



Záver

- najväčšie problémy pri analýze dát
  - konštrukcia relevantných otázok
  - nedostatok dát na uzavretie odpovede/nedostatok znalosti o dátach

- pozitívne výsledky
  - dokázali sme potvrdiť/vyvrátiť tendencie čo sme našli

Ďakujeme za pozornosť