Izglītības un zinātnes ministrijas finansētas valsts pētījumu programmas

"Covid-19 seku mazināšanai"

**Projekta Nr. VPP-COVID-2020/1-0016**

**„COVID-19 saistīto paraugu biobankas un asociēto datu integrētās platformas izveide Latvijā"**

Līguma numurs, noslēgšanas datums: Nr. 6-1/2, 10.07.2020.

Projekta īstenošanas termiņš: 01.07.2020.-31.12.2020.

Valsts pētījuma programmas finansējuma saņēmējs:

**APP Latvijas Biomedicīnas pētījumu un studiju centrs**

# ZINĀTNISKĀ PĒTĪJUMA REZULTĀTS

## *Vienota COVID-19 pētniecības datu informācijas sistēma*

# Satura rādītājs

# Izmantotie saīsinājumi

**LBMC** – Latvijas Biomedicīnas pētījumu un studiju centrs

**LMT** – Latvijas Mobilais Telefons

**LU** – Latvijas Universitāte

**RSU** – Rīgas Stradiņa unviersitāte

**RTU** – Rīgas Tehniskā universitāte

**HPC** – no angļu valodas *High Performance Computing* jeb augstas veiktspējas skaitļošana

**VIGDB** – Valsts iedzīvotāju genoma datubāze

**VPP** – Valsts pētījumu programma

# Izmantotie termini

**Asociētie dati** - ar biobankā uzglabātajiem bioloģiskajiem paraugiem saistītie dati, kas tiek ievākti anketu formā, iesaistot konkrēto pacientu pētījumā (tajā skaitā pētījuma dalībnieka antopometriskie rādītāji, iedzimtības faktori).

**Augstas pieejamības datu analīzes risinājums** - risinājums, kas nodrošina pieejamību datiem un to analīzes rīkiem, ņemot vērā faktisko noslodzi.

**Azure** - Microsoft Azure ir mākoņskaitļošanas pakalpojums, kas nodrošina iespēju izmantot lielu klāstu servisu un tehnoloģisko risinājumu.

**Datu ezera infrastruktūras instance** - datu ezera infrastruktūras instance ir LU īstenotā tehnoloģiskās pārneses projekta "Uz genoma un veselības datiem balstītas vēža prognozēšanas infrastruktūras izveide " ietvaros izstrādātās infrastruktūras kopija, kurā iespējams veikt nepieciešamos pielāgojumus konkrētā projekta realizācijai.

**Datu platforma** - dokumenta ietvaros ar terminu tiek saprasts programmatūras un aparatūras komplekss datu iegūšanai, glabāšanai, apstrādei, vizualizācijai un atvērtai piekļuvei pētniecībai.

**Dokers platforma** - platforma, kas ļauj darbināt konteinera risinājumu datu ezera infrastruktūrā.

**Galaxy platforma** - grafiska tiešsaistes vide, kas satur rīkus biomedicīnas datu apstrādei, koplietošanai un izgūšanai.

**HPC datu glabātuve** – augstas veiktspējas skaitļošanas jeb HPC infrastruktūrā ietilpstošas datu glabāšanas iekārtas, kas savienotas ar skaitļošanas klasteri.

**HPC klasteris** – augstas veiktspējas skaitļošanas jeb HPC infrastruktūras skaitļošanas mezgli, uz kuriem tiek izpildīti datu apstrādes uzdevumi.

**Latvijas Akadēmiskais datortīkls** – Izglītības un Zinātnes ministrijas (IZM) pārraudzītais datortīkls, kas savieno Latvijas akadēmiskās organizācijas un nodrošina pieslēgumu Eiropas akadmēmiskajam datortīklam GEANT. Ar šo jēdzienu tekstā tiek apzīmēta arī datortīkla daļa, kas tapusi Genomikas datu tīkla iniciatīvas ietvaros (https://gdn.lv) un nav tiešā IZM pārraudzībā.

**Konteineru risinājums** - risinājums kas ļauj programmatūras vienumus uzglabāt standartizētā veidā to pārvietošanai, izvēršanai un izpildei.

**Smilškastes risinājums** - datu un programmatūras kopums, kas paredzēts datu analīzes programmatūras izstrādei un noskaņošanai.

# Ievads

Covid-19 pandēmija un tās sekas ir radījušas virkni izaicinājumu Latvijas veselības aprūpē un sabiedrībā kopumā, radot nepieciešamību pēc tūlītējas rīcības vīrusa infekcijas izpētē.

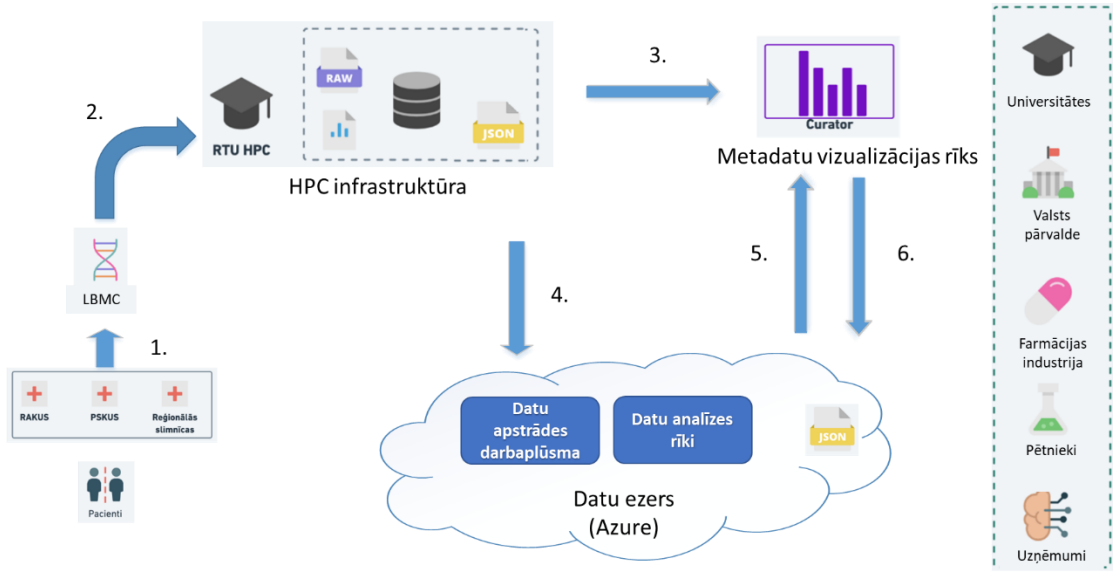
Valsts pētījuma programmas "*Covid-19 seku mazināšanai*" ietvaros veselības aprūpes un sabiedrības veselības jomā tika apstiprināts projekts "*COVID-19 saistīto paraugu biobankas un asociēto datu integrētās platformas izveide Latvijā*", ko kopīgi īstenoja Latvijas Biomedicīnas pētījumu un studiju centra (LBMC) pētnieki sadarbībā ar Rīgas Stradiņa universitātes (RSU), Rīgas Tehniskās universitātes (RTU) un Latvijas Universitātes (LU) pētniekiem. Projekta "*COVID-19 saistīto paraugu biobankas un asociēto datu integrētās platformas izveide Latvijā*" galvenais mērķis ir izveidot centralizētu biobanku un datu apmaiņas platformu, veicinot vīrusa izplatības ierobežošanas aktivitātes, jaunu ārstēšanas metožu izstrādi un starptautisku sadarbību. Viens no šī projekta galvenajiem uzdevumiem bija izveidot datu platformu un nodrošināt nepieciešamos rīkus, lai nodrošinātu piekļuvi visiem ar COVID-19 saistītajiem klīniskajiem un analītiskajiem datiem pētniecībai un izmantošanai ārstniecībā.

Šis ziņojums atspoguļo galvenos rezultātus un apraksta datu platformas un rīku izveides pētījuma rezultātu. Konkrētā uzdevuma galvenais mērķis bija ne tikai nodrošināt visu ar COVID-19 saistīto klīnisko un analītisko datu pieejamību, bet arī funkcionālos risinājumus turpmākai analīzei un interpretācijai. Projekts paredzēja izveidot funkcionālu platformu, izmantojot savstarpēji cieši saistītas komponentus: (1) datu uzglabāšanas komponents, kas paredzēts neapstrādātiem datiem, kas iegūti analizējot audu paraugus ar datu iegūšanas saskarni; (2) neapstrādāto datu sekundārās analīzes komponents; (3) klīnisko, fenotipisko un ģenētisko (sekundārās analīzes rezultātu) datu uzglabāšanas komponents ar kombināciju un datu transformācijas funkcionalitāti; (4) komponenti, kas ļaus veikt statistisko analīzi, kā arī izmantot mašīnmācīšanās algoritmus; (5) datu vizualizācijas komponents, kas ļaus pētniekiem un projekta dalībniekiem apskatīt metadatus, datu kvantitatīvos marķierus un izpildīt datu pieprasījumus (informācijas panelis). Pētījums balstās uz pētnieku esošajām zināšanām un pieredzi, kā arī uz jauno iegūto pieredzi, izveidojot vienotu COVID-19 pētniecības atbalsta informācijas sistēmu, kas nodrošina pilnvērtīgu attālinātu pētnieka darba vietu klīnisko un analītisko datu analīzei – datu izgūšanai, integrācijai ar citām datu kopām, vizualizācijai, kā arī primārajai un sekundārajai analīzei.

Pētījuma periods: no 2020.gada 1.jūlija līdz 2020.gada 31.decembrim. Šajā laikā projekta komandai ir izdevies izveidot Latvijā pirmo integrēto lielapjoma medicīnisko un pētniecības datu analīzes platformu, kas var tikt izmantota citiem medicīnas pētījumiem, veicinot uz zināšanām balstītu veselības aprūpes risinājumu izstrādi un ieviešanu.

# 1. Informācijas sistēmas vispārējā arhitektūra un piekļuves scenāriji

COVID-19 pētniecības atbalsta informācijas sistēmu veido savstarpēji integrēta pētniecības datu infrastruktūra un datu transformācijas un analīzes rīku komplekts, kas vispārējā arhitektūrā iekļauj augstas veiktspējas skaitļošanas (HPC, no angļu valodas *High Performance Computing*) infrastruktūrā ietilpstošās datu glabāšanas iekārtas (turpmāk tekstā HPC datu glabātuvi), datu apstrādi augstas veiktspējas infrastruktūrā, datu apstrādes darbaplūsmu un analīzes infrastruktūru, kā arī metadatu vizualizēšanas un atlases risinājumu (skat. 1. attēlu). HPC datu glabātuve un augstas veiktspējas apstrāde darbojas Latvijas Akadēmiskajā datortīklā, izmantojot akadēmiskos skaitļošanas resursus, datu kombinācija un transformācija izmanto datu ezera mākoņskaitļošanas resursus, savukārt metadatu vizualizēšanas un atlases rīks *Curator* izmanto publisko tīkla infrastruktūru.




**1. attēls. Pētniecības datu infrastruktūra**

Atbilstoši projekta realizācijas plānam, pētījumam tika piesaistīti pacienti un iegūts bioloģiskais materiāls. Detalizētāka informācija par veiktajiem procesiem un datu apstrāde legālajiem nosacījumiem ir pieejams šī projekta Zinātniskā pētījuma rezultātu ziņojumā "Covid-19 pacientu un pārslimojušo cilvēku bioloģisko paraugu biobanka". Iegūto medicīnas datu un

analīžu rezultātu izmantošanu datu platformas vajadzībām nosaka pētījuma protokols, Covid-19 pētījumspecifiskā piekrišana un Valsts Iedzīvotāju genoma datu bāzes informētās piekrišanas forma, kas apstiprināti Centrālās medicīnas ētikas komitejā (atļaujas biobankas veidošana: Nr. 01-29.1/2429, Nr. 1/19-04-05 un Covid-19 pacientu kohortas veidošana Nr. 01-29.1/5034). Pētījums tika realizēts saskaņā ar Helsinku deklarāciju un Konvenciju par cilvēktiesību un cieņas aizsardzību bioloģijā un medicīnā - Konvenciju par cilvēktiesībām un biomedicīnu. Tāpat visas pētījumā veiktās darbības tiek veiktas saskaņā ar Pasaules medicīnas asociācijas izdoto Taipejas Deklarāciju par Ētiskiem apsvērumiem attiecībā uz ar veselību saistītām datubāzēm un biobankām (*WMA Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks*), šīs deklarācijas atjaunināto 67. versiju, kas pieņemta 2016. gada oktobrī.

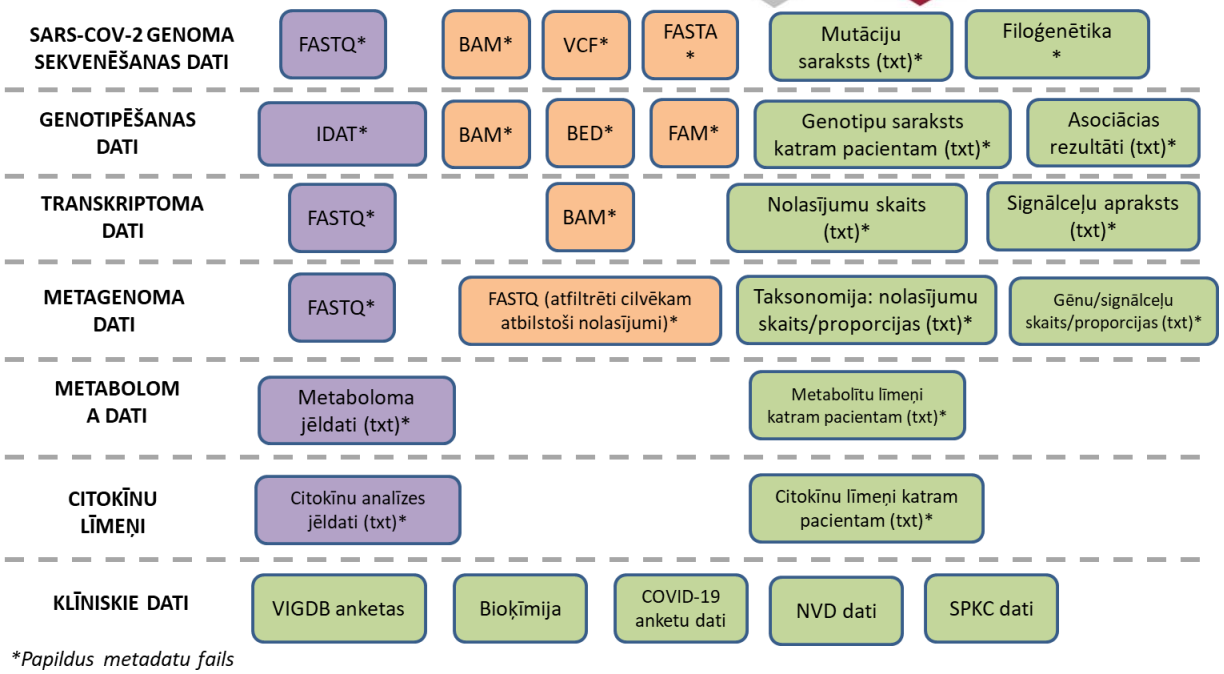
Lai varētu veikt pētījuma datu analīzi un iegūt rezultātus, kas veicinātu vīrusa izplatības ierobežošanu vai jaunu ārstēšanas metožu izstrādi, būtiska ir atbilstoša datu sagatavošana, kā arī metadatu pieejamības nodrošināšana.

HPC infrastruktūra ir neatņemama datu platformas sastāvdaļa, jo primāro datu (tai skaitā genoma sekvencēšanas rezultātā izveidoto FASTQ formāta jēldatu) apstrāde prasa ievērojamas skaitļošanas un glabāšanas jaudas. Piemēram, vienam metagenoma paraugam (cilvēka zarnu mikrobioma dati, kas tika uzkrāti un analizēti papildus vīrusa genoma datiem) sekvencēšanas dati FASTQ formātā aizņem līdz 8 GBaiti, bet tā apstrādes rezultātā ģenerētie starprezultāti līdz pat 100 GBaiti. Projekta ietvaros iegūto 500 metagenomu paraugu apstrādei ir nepieciešami līdz pat 54 TBaitiem glabāšanas vietas, un līdzīgas prasības ir arī citiem projektā iegūtajiem datu veidiem. Savukārt viena metagenoma parauga datu analīzei tiek izmantotas aptuveni 750 procesora kodolstundas. Šādu datu glabāšana un apstrāde nav iespējama ar esošajiem LBMC informācijas tehnoloģiju resursiem, tādēļ projekta vajadzībām tiek izmantota Latvijā lielākā specializētā HPC infrastruktūra.

Papildus molekulāro datu uzkrāšanai un analīzei, tika iegūti arī pacientu klīniskie un sociālekonomiskie dati. Datu transformācijai un apvienošanai tiek izmantota datu ezera infrastruktūra (detalizētāk aprakstīta 3. nodaļā “Datu kombinācija un transformācija datu ezerā”), kuras galvenais uzdevums ir nodrošināt dažādu primāro datu (piemēram, metagenoma gadījumā iegūto sākotnējo DNS secību) un arī sekundāro datu (piemēram, metagenoma taksonomisko vienību skaits un proporcijas, gēnu un signālceļu apraksti) pieejamību, saglabājot aprakstus

(metadatu līmenī) par izmantotajām datu analīzes metodēm un pielietotajiem parametriem. Šāda pieeja nodrošina iespēju atkārtoti veikt standartizētus, automatizētus un mērogojamus datu apstrādes uzdevumus. Pēc datu transformācijas pieejamo datu metadati tiek nodoti metadatu vizualizēšanas un atlases rīkam *Curator*, kas nodrošina metadatu pieejamību citām (tajā skaitā citu valstu) pētniecības grupām. Šādā veidā tiek nodrošināta informācijas pieejamība lielākam zinātnieku un medicīnas darbinieku skaitam, veicinot iegūto datu izmantošanu, vienlaikus panākot individualizēto datu aizsardzību. Metadatus šī projekta kontekstā veido gan katras analīzes un datu kvalitātes kritēriju aprakstošā informācija, gan dati par katrā platformas komponentā uzkrāto datu veidiem un to apjomiem. Metadatu izveides laikā tiek nodrošināts, ka tajos ietvertā informācija nesatur ar konkrētiem (individualizētiem) pacientiem vai paraugiem saistītus datus. Tādā veidā tiek panākta iespēja izmantot metadatus gan sabiedrības informēšanai, gan potenciālo pētījumu dizainam un izmantošanai, plašā apjomā izmantojot agregētos datus, kas ir īpaši svarīgi COVID-19 datu pieejamības nodrošināšanai gan nacionālā, gan starptautiskā līmenī, vienlaikus nodrošinot personas un saistīto medicīnas datu drošību.

Šajā pētījumā apstrādātie molekulārie, bioķīmiskie un metaboloma dati, kā arī transformētie klīniskie un sociālekonomiskie dati pētniekiem tālākai analīzei – hipotēžu pārbaudei un modeļu veidošanai, ir pieejami gan HPC infrastruktūrā, gan datu ezera infrastruktūrā. Datu analīzei izmantojamā infrastruktūras izvēle ir atkarīga no datu analīzes uzdevuma specifikas un pētniecības grupas kompetencēm. Projekta ietvaros radīto datu kategorijas un piemēri primāro un sekundāro datu veidiem ir redzami 2. attēlā.




**2. attēls. Projekta ietvaros radīto datu kategorijas un piemēri datu veidiem un formātiem.**

Ar violeto krāsu ir atzīmēti primārie molekulārie, bioķīmiskie un metaboloma dati un to formāti; ar smilškrāsu ir norādīti dati, kas atbilst sekundāro datu pirmajam līmenim vairākām nākošās paaudzes sekvencēšanas un genotipēšanas kategorijām; ar zaļo krāsu apzīmēti sekundārie datu veidi, kas izmantojami tālākajā datu analīzē un interpretācijā; ar zvaigznīti norādīti datu veidi, kam pieejami papildus metadatu apraksti.

Pētnieciskās grupas var piekļūt iespējai veikt reāllaika datu identificēšanu, izmantojot vizualizētos metadatus, kas nodrošina datu privātumu un drošību, metadatiem atrodoties atsevišķajā metadatu metadatu vizualizēšanas un atlases rīkā *Curator* (*Curator* metadatu vizualizācijas un atlases rīks detalizēti aprakstīts 4. nodaļā "Datu vizualizācija"), vienlaikus sniedzot pētniekiem iespēju veikt meklēšanu vairākās datu kolekcijās un veikt atbilstošus datu pieprasījumus. Piekļuve metadatu visualizācijas platformai tiek nodrošināta tīmekļa vietnē www.longenesis.com/curator (reģistrējoties vietnē, pētniekiem ir iespēja atlasīt metadatu kopas un attiecīgi pieprasīt tās).

Pieprasījums datu piekļuvei tiek nosūtīts atbildīgām personām (konkrētajā gadījumā Latvijas Biomedicīnas centram), kas var lemt par datu izsniegšanu. Datu iegūšanas pirmais solis ir atrisināt visus administratīvos jautājumus datu izsniegšanai, kas sevī ietver ētisko un juridiskos aspektus. Pēc administratīvo jautājumu atrisināšanas, tehniskais datu atlases komplekts tiek nodots

datu ezera infrastruktūrai, kas atbilstoši pieprasījumam sagatavo datu kopu komplektus nodošanai datu prasītājam (skat. 3. attēlu).


**3. attēls. Pētnieku piekļuve datiem**

Datu ezera infrastruktūras nodrošina iespēju pieņemt datu atlases nosacījumus un sagatavot atbilstošo datu kopu. Pēc datu kopas sagatavošanas, Latvijas Biomedicīnas centra pārstāvji var nodot to atbilstošajam pētniecības grupas pārstāvim.

Pētniekiem, kas darbā izmanto sekvencēšanas primāros datus, tiek nodrošināta piekļuve šiem datiem HPC datu glabātuvē. *Curator* metadatu vizualizēšanas un atlases rīkā tiek nodrošināta sekvencēto datu raksturojošie lielumi (metadati), iekļaujot identifikatorus un atsauces uz Galaxy platformu HPC infrastruktūrā. Veicot reģistrāciju platformas rīkā, pētniekiem ir iespēja atlasīt metadatu kopas un attiecīgi pieprasīt tās. Iegūstot atļauju piekļūt pie sekvencēšanas primārajiem datiem, prasītājam tiek nodrošināts lietotājvārds un parole piekļuvei Galaxy platformai https://galaxy.hpc.rtu.lv, kur būs iespēja veikt atlasīto datu lejupielādi vai apstrādi ar platformā iebūvētiem bioinformātikas rīkiem. Primāro datu izsniegšanas soļi ir parādīti 4. attēlā.


**4. attēls. Pētnieka piekļuve sekvencēšanas primārajiem datiem**

Savukārt pieteikšanās tikai Galaxy platformas vai HPC resursu izmantošanai ar saviem datiem iespējama tiešā veidā reģistrējoties https://hpc.rtu.lv.

Piekļuvi izvēlētajiem datiem nodrošina LBMC, kā Valsts Iedzīvotāju genoma datubāzes (VIGDB) galvenais apstrādātājs atbilstoši Cilvēka genoma izpētes likuma un saistīto Ministru kabineta nosacījumiem. Izstrādātā informācijas sistēma uzglabā pētījumu rezultātā iegūtos datus par VIGDB iesaistītajiem gēnu donoriem. Līdz ar to nodrošinot pētnieku piekļuvi Cilvēka genoma izpētes likuma 15. panta pirmo daļā minētajiem paraugiem un datiem atbilstoši MK noteikumiem

Nr.695 “Noteikumi par genoma datu bāzē uzglabāto kodēto audu paraugu, kodēto DNS aprakstu, kodēto veselības stāvokļa aprakstu un kodēto ģenealoģiju uzglabāšanas un izsniegšanas kārtību un izsniegšanas akta veidlapas paraugu un tās aizpildīšanas un uzglabāšanas kārtību”, pētnieks var izmantot iespēju iegūt valsts pētījumu programmas (VPP) rezultātā iegūtos datus par atbilstošajiem gēnu donoriem. Lai nodrošinātu datu drošību informācijas sasaistei tiek izmantots unikāls VIGDB kods. Atbilstoši MK noteikumu Nr.695 “Noteikumi par genoma datu bāzē uzglabāto kodēto audu paraugu, kodēto DNS aprakstu, kodēto veselības stāvokļa aprakstu un kodēto ģenealoģiju uzglabāšanas un izsniegšanas kārtību un izsniegšanas akta veidlapas paraugu un tās aizpildīšanas un uzglabāšanas kārtību” prasībām, piekļuve pētījumos iekļauto gēnu donoru VIGDB datiem un ar tiem saistīto VPP pētījumos radītajiem rezultātiem tiek nodrošināta balstoties uz MK noteikumu pielikumā esošas iesnieguma veidlapas pamata, iesniedzot Centrālās medicīnas ētikas komitejas; citu ētikas komisiju un Genoma izpētes padomes atzinumus atbilstoši konkrēto pētījumu specifikai. Piekļuves kārtība datiem tiek realizēta atbilstoši LBMC 2017. gada 3. aprīlī apstiprinātajiem “Valsts Iedzīvotāju genoma datubāzes iekšējie informācijas sistēmas drošības noteikumiem”. Saņemot piekrišanu no pētniekiem, kas izmantojuši VIGDB un informācijas sistēmas datus, tiek plānots informāciju par pieprasījumu datu piekļuvei ievietot arī VIGDB mājaslapā.

LBMC sadarbībā ar iesaistītajām institūcijām nodrošina izveidotās infrastruktūras uzturēšanu ilgtermiņā.

Lai optimizētu COVID-19 saistīto paraugu biobankas un asociēto datu integrētās platformas veiksmīgu izmantošanu pētniecības un veselības aprūpes vajadzībām, ierosinām steidzami veikt izmaiņas Cilvēka genoma izpētes likumā un visos ar šo likumu saistītajos normatīvajos aktos. Īpašu uzmanību nepieciešams pievērst:

1) novecojošajām tehniskajām un drošības prasībām attiecība uz datu uzglabāšanu, pārnesi un izmantošanu, kas neatbilst mūsdienās esošajiem datu apjomiem un informācijas tehnoloģiju attīstības pakāpei,
2) nosacījumiem, kas nodrošinātu VIGDB un Iedzīvotāju genoma valsts reģistra darbības optimizēšanai, lai nodrošinātu iespēju otrreizējo medicīnas datu sasaisti ar VIGDB uzglabātajiem un pētījumos iegūtajiem datiem.

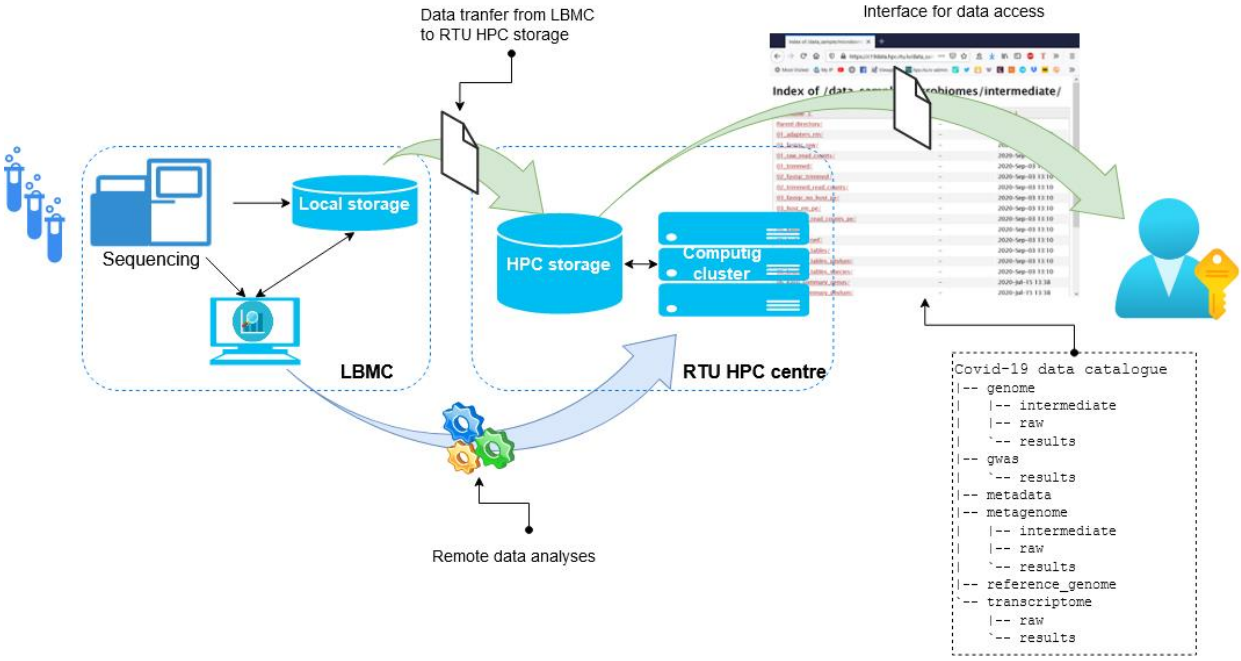
Visoptimālākais risinājums ir jauna Biobanku darbību regulējoša likuma izstrāde un apstiprināšana.

# 2. Datu glabāšanas un apstrādes HPC infrastruktūra

Datu apstrāde HPC infrastruktūrā notiek pēc šādas shēmas (parādīts 5. attēlā): primārie dati no LBMC lokālās datu glabātuves tiek pārsūtīti uz RTU HPC centra infrastruktūru, kur tie tiek saglabāti augstas veiktspējas datu glabātuvē. Datu pārraidei tiek izmantots Latvijas Akadēmiskais datortīkls, kas ar MikroTik un LMT atbalstu 2020. gadā tika pilnveidots Genomikas datu tīkla izveides projekta ietvaros. Datu glabātuve ir savienota ar skaitļošanas klasteri – daudzprocesoru sistēmu, kas spēj izpildīt lielu skaitu apstrādes uzdevumu paralēli, tādējādi paātrinot rezultātu iegūšanu. Iespēja ātri apstrādāt datus ir kritiski svarīga, ja no rezultāta ir atkarīga lēmumu pieņemšana, piemēram, par epidemioloģiskajiem drošības pasākumiem.

Genoma vai metagenomu paraugu primārie dati parasti tiek analizēti nevis pa vienam paraugam, bet komplektos. Tipiskais vienlaicīgi analizējamo SARS-Cov-2 genoma paraugu skaits, lai pārbaudītu, piemēram, noteiktas mutācijas esamību, ir līdz 100. Katram genoma paraugam tiek izmantoti divi līdz trīs procesora kodoli, attiecīgi kopējais nepieciešamās infrastruktūras apjoms ir līdz 300 procesora kodoliem. Savukārt viena metagenoma parauga analīzei parasti nepieciešami astoņi līdz 16 kodoli, kas skaitļošanas infrastruktūras prasības palielina vēl vairāk. Pie nepietiekama resursu apjoma apstrādes uzdevumi tiks veikti secīgi viens pēc otra, pagarinot kopējo apstrādes laiku. Arī datu glabātuves veiktspēja ir atkarīga no vienlaicīgi veicamo uzdevumu skaita, jo tai jāspēj nodrošināt datu nolasīšanu un rakstīšanu uzdevumu izpildes laikā. Vēlamā veiktspēja var tikt sasniegta ar slodzes balansēšanu uz vairākiem datu glabāšanas serveriem/disku masīviem, kā arī konfigurācijā izmantojot *Solid State Drives* (SSD) diskus.

Datu analīzes rezultāti tiek saglabāti HPC infrastruktūrā, organizējot tos datu katalogā, kam var piekļūt autorizēti lietotāji, izmantojot speciāli šim nolūkam izveidotu saskarni . Primāro datu un starprezultātu atstāšana HPC datu glabātuvē ļauj gan projekta grupas, gan citiem pētniekiem vajadzības gadījumā veikt atkārtotu šo datu analīzi pie citiem ievades parametriem.




**5. attēls. Datu glabāšanas un apstrādes risinājuma galvenie elementi**

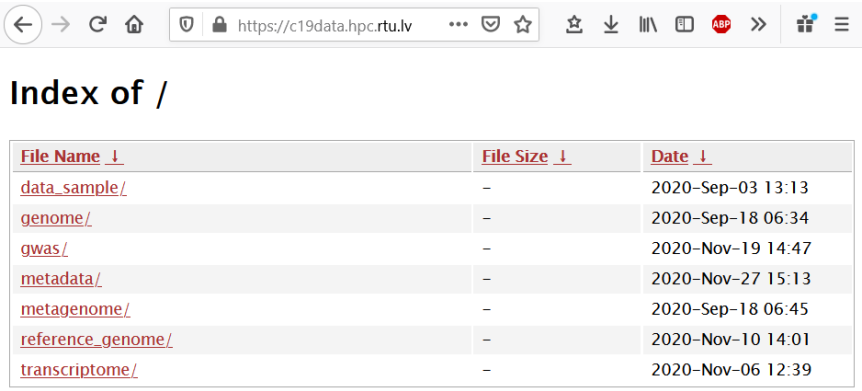
## 2.1. Risinājums datu glabāšanai un pieejamības nodrošināšanai

Pirms projekta realizācijas Latvijā nebija piemērotas saskarnes HPC infrastruktūrā izvietoto failu piekļuvei. Dati bija pieejami tikai HPC klastera lietotājiem, attālināti ar SSH protokolu pieslēdzoties piekļuves serverim, kas savienots ar tīkla datu glabātuvi (*Network Attached Storage*). Tāpēc, kā viens no projekta uzdevumiem tika izvirzīts viegli lietojamas un drošas saskarnes izveide HPC datu glabātuvē izvietoto COVID-19 datu pieejamībai, kas nodrošinātu:

- datu piekļuvi autorizētiem pētniekiem, kas nav esošie HPC infrastruktūras lietotāji un nepārzina HPC vides specifiku;
- iespēju veikt automatizētu datu apmaiņu ar ārējam platformām, tostarp Microsoft Azure datu ezera infrastruktūru.

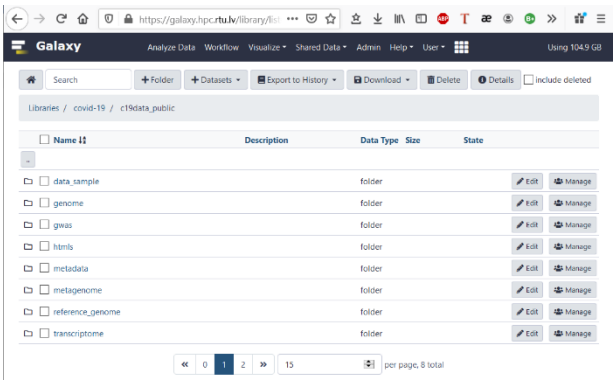
Šo mērķu sasniegšanai ir izveidotas divas saskarnes ar dažādu funkcionalitāti:

- Lai jau pašā projekta sākumā nodrošinātu datu pieejamību projekta partneriem un integrācijai ar datu ezeru, tika izveidota vienkārša Nginx [https://www.nginx.com/] bāzēta saskarne, kas servē failus no HPC datu glabātuves. Piekļuvei tiek izmantots HTTPS protokols ar WebDaw funkcionalitāti (autorizētiem lietotājiem pieejams https://c19data.hpc.rtu.lv). Saskarnes piemērs interneta pārlūkprogrammā parādīts 6. attēlā.

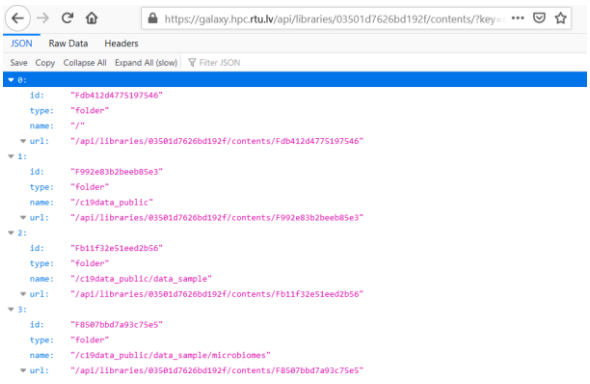



**6. attēls. Nginx bāzeta HTTPS saskarne**

- Funkcionālā ziņā plašāks risinājums – Galaxy datu apstrādes platforma, kas tiks izmantots kā primārā saskarne lietotājiem. Papildu datu izgūšanai, platforma dod iespēju veikt arī atlasīto datu analīzi ar jau instalētiem bioinformātikas rīkiem un iepriekš sagatavotām apstrādes darbplūsmām grafiskā vidē, kā arī programmatisku datu piekļuvi un mijiedarbi ar platformu caur Galaxy API (*Applications Programming Interface*). Galaxy saskarnes piemērs interneta pārlūkprogrammā paradīts 7. attēlā.




7.a. Lietotāju grafiskā saskarne




7.b. Galaxy API

**7. attēls. Galaxy platformas nodrošinātās saskarnes datu piekļuvei**

Lietotāju autentifikāciju un piekļuves tiesību pārvaldību abām saskarnēm nodrošina savienojums ar HPC infrastruktūras centrālo *Lightweight Directory Access Protocol* (LDAP) serveri, bet sistēma ir konfigurējama ar citiem ārējiem LDAP/aktīvās direktorijas serveriem. Papildu piekļuves tiesību pārvaldību atsevišķu mapju vai failu līmenī nodrošina Galaxy iebūvētie autorizācijas mehānismi. Lietotāju reģistrācija datu piekļuvei notiek caur metadatu vizualizācijas un atlases rīku *Curator* (skat. 4. nodaļā). Kā tika aprakstīts 1. nodaļā, metadatu vizualizācijas un atlases rīks *Curator* nodrošina datu identificēšanu un pieprasījumu, nosūtot attiecīgā lietotāja atlases parametrus datu publicētājam. Atlases parametrus datu publicētājam sistēma potenciāli spēj padot ne tikai paziņojuma veidā, bet arī automātiski aizsūtīt strukturētā objekta (JSON) lai padotu tikai atlasītos datus uz datu apstrādes platformu (kā, piemēram, Galaxy).

HPC datu glabātuvē izvietotie Latvijas COVID-19 pacientu dati ir organizēti katalogā, kas ir strukturēts atbilstoši projekta īstenošanas gaitā iegūtajiem datu tipiem. Iegūtie datu tipi ir šādi:

- vīrusa pilna genoma sekvencēšanas dati (direktorija *genome*);
- mikrobioma sekvencēšanas dati (direktorija *metagenome*);
- COVID-19 pacientu genotipēšanas dati (direktorija *gwas*);
- perifēro asins šūnu transkriptoma sekvencēšanas dati (direktorija *transcriptome*);
- paraugu metadati, kas satur citokīnu mērījumus, bioķīmiskos datus, aptaujas (direktorija *metadata*).

Ja attiecināms, katram datu tipam izveidotas trīs apakšdirektorijas, kas satur:

- ievades datus ar skaidrojošu metadatu failu (direktorija *raw*);
- starpposmu failus, kas tiek saglabāti, lai ļautu atkāpes no standarta datu analīzes darbplūsmas (direktorija *intermediate*);
- galvenos rezultātus (direktorija *results*).

Kopējais katalogā pieejamo datu apjoms ir 2.7 TBaiti un 1644 datu vienības (2021. gada 31. marts), kas ir tikai neliela publiskā daļa no kopējā ar primāro datu apstrādi saistītā datu apjoma. Projekta īstenošanai laikā lielākais nepieciešamais datu apjoms sasniedza 112 TBaitus, kas ietver primāro datu, datu analīzes rezultāta radītos starprezultātu, rezultātu datu apjomu. Noslēdzot projektu, datu glabātuvē novietoti visi būtiskie projekta otrās darba pakas "Veikt standartizētu paraugu analīzi, lai noskaidrotu bioķīmiskos, ģenētiskos, citus molekulāros un imunoloģiskos faktorus" rezultātā iegūtie dati, kuru apjoms ir 35 TBaiti. Ņemot vērā uz datu platformu attiecināmo konceptu par atvērtu piekļuvi pētniecībai, kopējais datu apjoms projekta pēcuzraudzības periodā būs atkarīgs no pētnieku intereses par datu platformā deponētajiem datiem, pētījumu mērķiem, pētniecības rezultātā radušos sekundāro datu apjoma un to piemērotības ilgtermiņa izmantošanai.

Datu glabāšanas risinājuma saskarnes izvietošana IZM akadēmiskajā datu centrā uz RTU rīcībā esošās *OpenStack* mākoņskaitļošanas platformas nodrošina pakalpojumam augstu pieejamību un tiešu savienojumu ar Latvijas Akadēmisko datortīklu. Projekta ietvaros tika pilnveidots *OpenStack* platformas savienojums ar datortīklu 10 Gb/s pārraides ātruma atbalstam.

Augsta tīkla caurlaidspēja un efektīvi pārraides protokoli ir kritiski svarīgi, lai ātri pārsūtītu datus no vietas, kur sekvencēšanas primārie dati tiek iegūti, uz HPC datu glabātuvi, kā arī lai nodrošinātu datu piekļuvi caur saskarni gala lietotājiem. Projekta ietvaros tika testēts faktiskais pārraides ātrums 10 Gb/s tīkla savienojumam starp LBMC un RTU HPC centru, izmantojot dažādus plaši izmantotus datu pārraides rīkus. Testēšanā iegūtie rezultāti ir pielietoti, lai uzlabotu Latvijas Akadēmisko datortīklu un datu glabāšanas infrastruktūru genoma datiem Genomikas datu tīkla iniciatīvas ietvaros.

## 2.2. HPC datu apstrādes rīki

HPC klasteris nodrošina vidi skaitļošanas uzdevumu/datu apstrādes veikšanai. Lai sasniegtu projekta izvirzītos mērķus un dotu iespēju pētniekiem veikt COVID-19 sekvencēšanas datu analīzi, klastera programmatūras vide ir sagatavota ar aktuālajiem atvērtā koda bioinformātikas rīkiem:

- Sekvenču pielīdzināšanas rīki (*Bowtie2, MAFFT, STAR, DIAMOND, VSEARCH*);
- Variāciju un mutāciju noteikšana un anotēšana (*BCFtools, DeepVariant, LoFreq, Strelka2, VarScan, VEP*);
- Nolasījumu pielīdzinājumu un variantu failu manipulācija (*SAMtools, VCFtools*);
- Nolasījumu taksonomiskā klasifikācija (*Kraken2*);
- Taksonomiskās kompozīcijas noteikšana (*Bracken, MetaPhlAn2, MetaPhlAn3*);
- Metagenomisko datu funkcionālā profilēšana (*HUMAnN2*);
- Patogēnu genomisko datu potenciāla pielietošana zinātnei un sabiedrības veselībai (*Nextstrain*);
- Sekvencēšanas datu kvalitātes kontrole (*FastQC, Fastp, Trimmomatics*).

Atsevišķas programmatūras paketes (*GATK4, DeepVariant, DIAMOND, HUMAnN, VEP*) ir ieviestas kā Singularity konteineri.

Programmatūra ir pielāgota izpildei uz HPC klastera, kas ietver:

- sagatavotus programmatūru/rīku ielādēšanas moduļus;

```
[[redacted]@ui-1 ~]$ module avail bio

---------------------------------------- /opt/exp_soft/modulefiles ----------------------------
   bio/bcftools/1.10.2              bio/humann2/2.8.1-sg             bio/samtools-1.9
   bio/bowtie2/2.4.1                bio/humann2/2.8.1          (D)   bio/samtools/1.10
   bio/bracken/2.5                  bio/kraken2/2.0.9-beta           bio/star/star-2.7.5c-test
   bio/deepvariant/1.0.0-test       bio/lofreq/2.1.5-test            bio/strelka/2.9.10
   bio/deepvariant/1.0.0      (D)   bio/mafft/7.471                  bio/trimmomatic/0.39
   bio/diamond/2.0.3-sg             bio/metaphlan2/2.6.0-test        bio/varscan/2.4.4
   bio/fastp/0.21.0                 bio/metaphlan2/2.7.7-conda (D)   bio/vcftools/0.1.17
   bio/fastqc/0.11.9                bio/metaphlan3/3.0.2-conda       bio/vep/101-sg
   bio/gatk/4.1.8.1-sg              bio/samtools-0.1.18              bio/vsearch/2.15.0
```

- programmatūras palaišanas skriptus saskaņā ar lietotāju vajadzībām, piemēram, BCFtools;

```
#!/bin/bash
#PBS -N testBcftools      ###  any convinient job name for TORQUE
#PBS -q batch
#PBS -l walltime=1:00:00  ###   wall clock time (execution time): hh:mm:ss
#PBS -l nodes=1:ppn=6
#PBS -j oe
#PBS -o testBcftools.out  ###   name of output file

module load  bio/bcftools/1.10.2  ### change to currentmodule name

###   Switch to the working directory; by default TORQUE launches processes
###   from your home directory.

cd $PBS_O_WORKDIR

time bcftools query -f '%CHROM  %POS  %REF  %ALT{0}\n'  ALL.chr16.phase1_release
_v3.20101123.snps_indels_svs.genotypes.vcf.gz
```

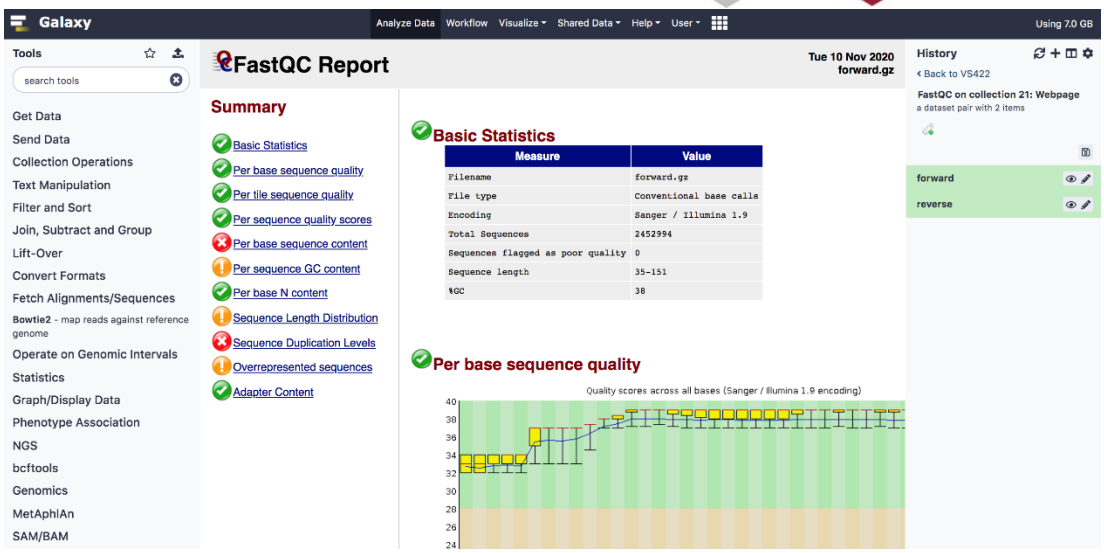
- papildus skriptus atsevišķu programmatūru ievades failu tīrīšanai;
- optimālus ievades parametrus virknes vai paralēlam izpildes režīmam klasterī.

HPC klastera piekļuve tiek nodrošināta reģistrējoties https://hpc.rtu.lv

## 2.3. Datu apstrāde *Galaxy* platformā

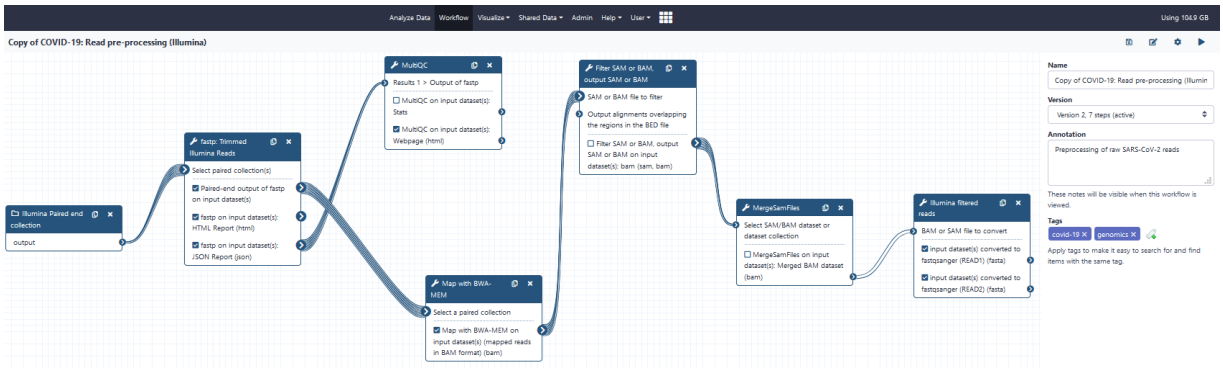
Standarta HPC vide prasa no lietotāja priekšzināšanas HPC klastera un Linux komandrindas izmantošanā. Lai dotu arī iespēju pētniekiem bez šādām priekšzināšanām veikt vienkāršas datu apstrādes operācijas, projekta ietvaros tika ieviesta Galaxy datu apstrādes platforma [https://galaxyproject.org/], kas tika integrēta arī ar iepriekš aprakstīto COVID-19 datu glabāšanas risinājumu. Galaxy vidē var veikt SARS-Cov-2 filoģenētisko analīzi, izmantojot viegli lietojamus grafiskus rīkus, tādejādi ļaujot plašākam pētnieku lokam analizēt primāros datus no projektā izveidotā COVID-19 datu kataloga. Šo vidi var izmantot, piemēram, Nacionālā references laboratorija.

RTU HPC infrastruktūrā tika izveidota lokāla Galaxy platformas instance (pieejama šeit ar autorizāciju: https://galaxy.hpc.rtu.lv), kas tika savienota ar RTU HPC klasteri, ļaujot uz to novirzīt intensīvus apstrādes uzdevumus. Platformā ir ieviesti tie paši bioinformātikas rīki un darbplūsmas, kas ir sagatavoti RTU HPC klasterī. Piemēram, 8. attēlā var redzēt grafisku kvalitāte kontroles atskaiti, kas iegūta ar FastQC rīku Galaxy platformā.




**8. attēls. Kvalitātes kontroles atskaite Galaxy platformā**

Papildus ir ievestas Galaxy darbplūsmas SARS-Cov-2 analīzei, balstoties uz "*Best practices for the analysis of SARS-CoV-2 data: Genomics, Proteomics, Evolution, and Cheminformatics* [https://covid19.galaxyproject.org/]" saskaņā ar ieteikumiem[1]. 9. attēlā parādīts vizuāls darbplūsmas piemērs Illumina ģenerētu SARS-Cov-2 primāro datu priekšapstrādei, kas ir pieejams uz projektā ieviestās Galaxy platformas.



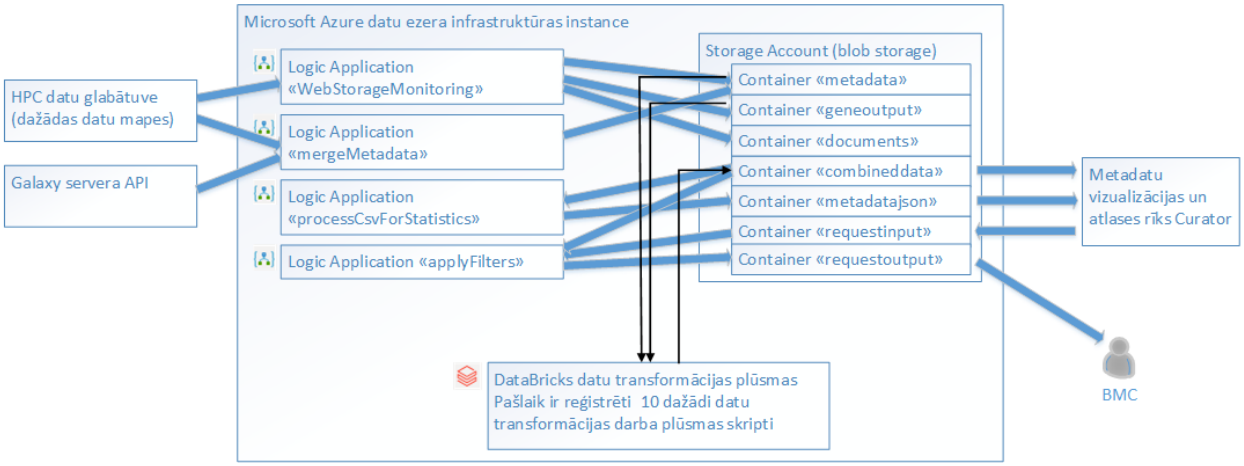

**9. attēls. SARS-Cov-2 primāro datu priekšapstrādes darbplūsma**

[1] Baker D, van den Beek M, Blankenberg D, et al. No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. PLoS Pathog. 2020 Aug 13;16(8):e1008643. doi: 10.1371/journal.ppat.1008643. PMID: 32790776; PMCID: PMC7425854.

# 3. Datu kombinācija un transformācija datu ezerā

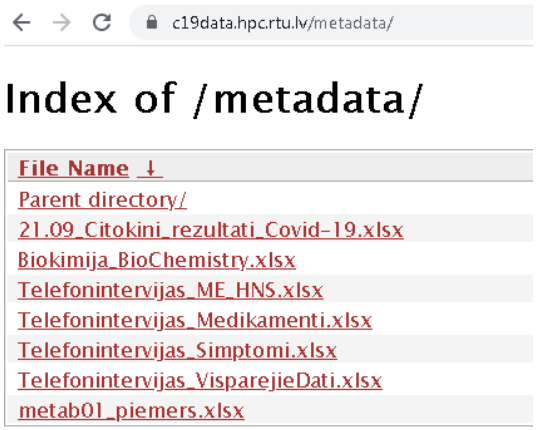
Datu kombinācijas, transformācijas un analīzes platforma (Microsoft Azure datu ezera projekts, Nr. KC-PI-2017/102, "Uz genoma un veselības datiem balstītas vēža prognozēšanas infrastruktūras izveide") nodrošina klīnisko, fenotipisko un ģenētisko datu (pēc sekundārās analīzes) datu krātuvi un datu transformācijas, kombinācijas darba plūsmas, kas ļauj sagatavot datus tālākiem datu analīzes soļiem. Infrastruktūra nodrošina integrētus datu analīzes rīkus nodrošinot iespēju veikt, piemēram, mašīnmācīšanās datu analīzi, kā arī ļaujot veikt datu analīzi izmantojot alternatīvos datu analīzes risinājumus, ko nodrošina HPC infrastruktūra.

Microsoft Azure datu ezera infrastruktūras instance ir uzstādīta un pielāgota tieši šī projekta vajadzībām (instances shematiskai attēlojums parādīt 10. attēlu), izmantojot loģiskās lietotnes tiek nodrošināta datu integrēšana starp infrastruktūras elementiem. Datu transformācijas un kombinācijas uzdevuma nodrošināšanas kontekstā ir jāizceļ klīniskie un fenotipiskie dati, ģenētisko datu sekundārās analīzes rezultāti un primārie ģenētiskie dati. Katram no minētajiem datu veidiem tiek nodrošināta sava datu apstrādes un transformācijas darba plūsma.




**10. attēls. Microsoft Azure datu ezera infrastruktūras instances shematisks attēlojums**

Klīnisko un fenotipisko dati sākotnēji tiek novietoti HPC datu glabātuvē atsevišķā darba mapē (skat. 11. attēlu), un projekta izstrādes laikā ir saskaņots, ka klīnisko un fenotipisko datu formāts ir XLSX, kur katrs XLSX fails reprezentē vienu datu tabulu.




**11. attēls. Klīnisko un fenotipisko datu mape HPC datu glabātuvē**

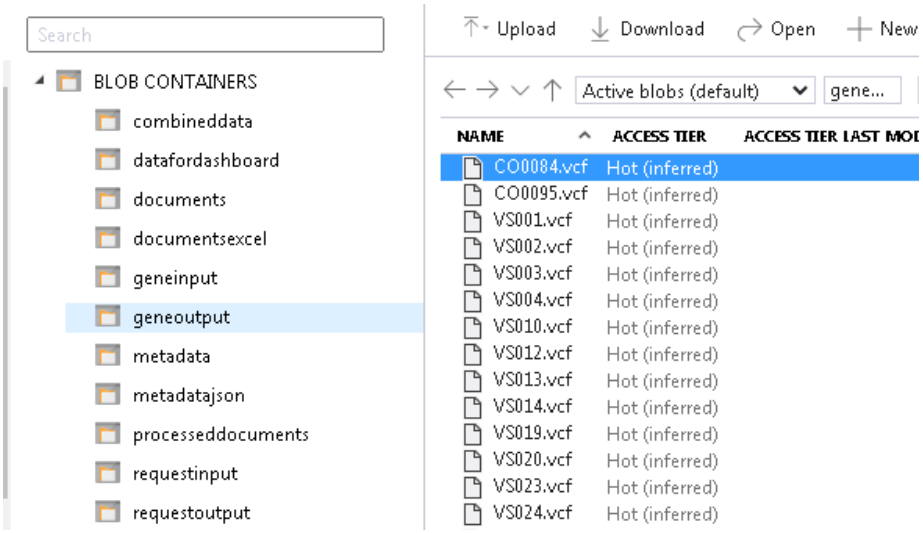
Datu ezera infrastruktūras loģiskā lietotne "*WebStorageMonitoring*" uzrauga HPC datu glabātuves atsevišķās darba mapēs saturu ar klīniskajiem un fenotipiskajiem datiem. Identificējot datu korekcijas, tiek iniciēts klīnisko un fenotipisko datu transports uz datu ezera infrastruktūras pagaidu datu krātuvi (konteineris "*metadata*"), lai veiktu datu transformācijas un kombinācijas darba plūsmas un iegūtu turpmākai datu analīzei derīgus datu masīvus.

Ģenētisko datu sekundārās analīzes dati arī tiek glabāti HPC datu glabātuvē atsevišķā darba mapē, un katra parauga sekundārās analīzes datu komplekts tiek noformēts kā atsevišķs VCF formāta fails. Datu ezera infrastruktūras loģiskā lietotne "*WebStorageMonitoring*" uzrauga arī ģenētisko datu sekundārās analīzes mapes saturu un, identificējot jaunu vai labotu failu, iniciē datu failu transportu uz datu ezera infrastruktūras pagaidu datu krātuvi (konteineris "*geneoutput*" un "*documents*"), lai veiktu datu transformācijas un kombinācijas darba plūsmas un iegūtu turpmākai datu analīzei derīgus datu masīvus.

Ģenētiskie primārie dati netiek transportēti uz datu ezera infrastruktūru, tā vietā, izmantojot datu ezera infrastruktūras loģisko lietotni "*mergeMetadata*", no HPC datu glabātuves tiek nolasīti primāro datu raksturojošie dati. Kā arī par primārajiem datiem tiek veikti pieprasījumi uz HPC infrastruktūras Galaxy serveri, lai iegūtu papildu raksturojošo un piekļuves informāciju. Iegūtā ģenētisko primāro datu raksturojošā informācija tiek apstrādāta līdzīgi kā klīnisko un fenotipisko datu komplekti (novietoti konteinerī "*metadata*"), lai nodrošinātu šo datu masīvu nodošanu metadatu vizualizācijas un atlases rīkam *Curator*.

Datu transformēšanas un kombinēšanas darba plūsmas tiek iniciētas secīgi. Kā viens no piemēriem datu transformēšanai ir bioķīmisko citokīnu līmeņu datu apvienošana ar ģenētisko datu

sekundārās analīzes rezultātu. Datu transformēšana un apvienošana tiek nodrošināta, izmantojot skriptus un datu ezera "*Data Bricks*" komponenti, kas ir izstrādāti tieši šī projekta datu apstrādei. Šāda pieeja nodrošina minimālu cilvēka iejaukšanos datu transportēšanā, transformācijā un kombinācijā, nodrošinot atkārtotu uzdevuma izpildi, ja ir pieejami papildinātie datu masīvi. Datu pagaidu krātuves, transformācijas starpposmi un gala transformācijas rezultāti tiek novietoti dažādos datu ezera krātuves konteineros (skat. 12. attēlu).

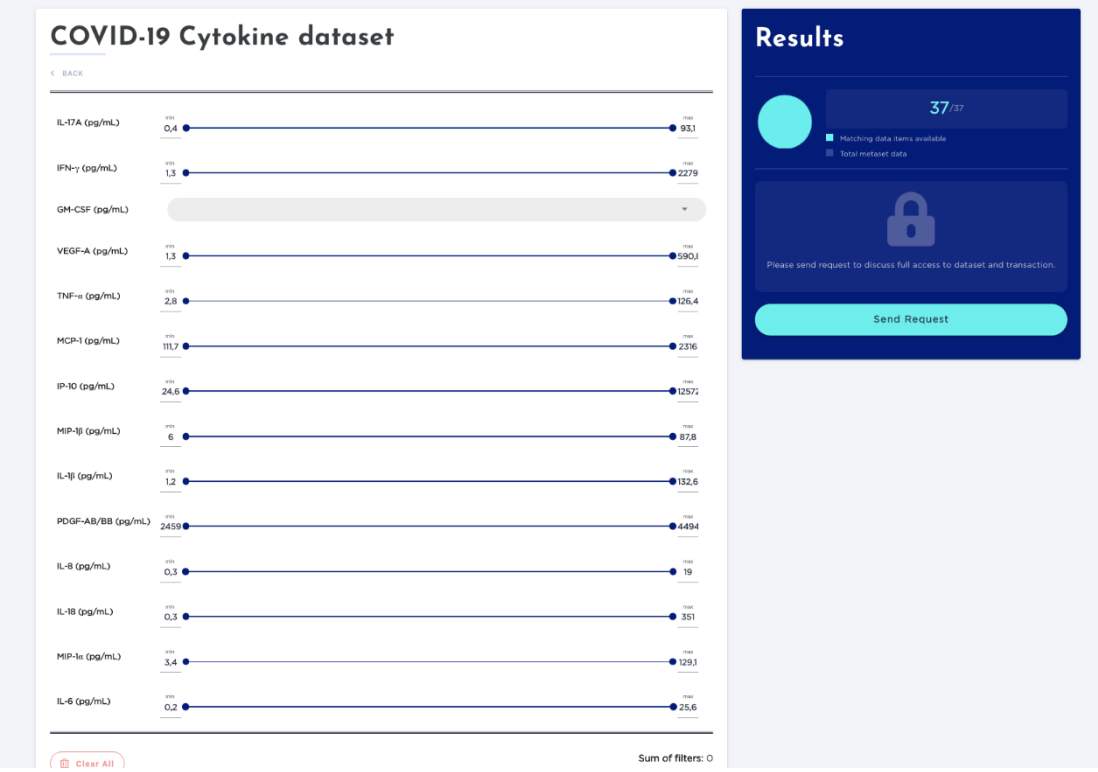



**12. attēls. Datu ezera infrastruktūras datu konteineri**

Datu transformēšanas un kombinēšanas darba plūsmu rezultātā tiek radītas datu kopas (CSV faili), kas reprezentē analizējamās datu kopas (konteineris "*combineddata*"). Lai nodrošinātu sagatavoto datu vizualizāciju, ir izstrādāta loģiskā lietotne metadatu ģenerēšanai "*processCsvForStatistics*". Uzģenerētie metadati JSON formātā tiek novietoti datu ezera konteinerī "*metadatajson*", un tiek nodrošināta piekļuve šiem datiem no metadatu vizualizācijas un atlases rīka *Curator*. No metadatu vizualizācijas un atlases rīka *Curator* tiek sagaidīti datu vaicājumu pieprasījumi JSON formātā (konteinerī "*requestinput*"). Datu ezera infrastruktūras loģiskā lietotne "*applyFilters*" nodrošina datu vaicājumu pieprasījumu izpildi datu konteinerī "*combineddata*" reģistrētajām datu kopām un sagatavo apakškopas atbilstoši definētajiem vaicājumiem (lai tos var tālāk nodot datu analīzes uzdevumu veikšanai). Pašreizējais loģiskās lietotnes "*applyFilters*" risinājums sagatavotās apakškopas reģistrē konteinerī "*requestoutput*", kam ir piekļuve LBMC atbildīgajām personām.

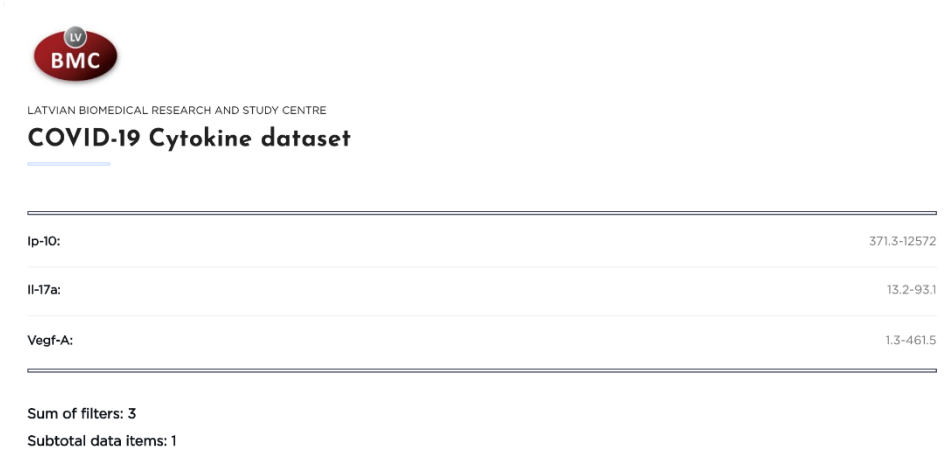
# 4. Datu vizualizācija

Metadatu vizualizācijas un atlases rīks *Curator* projekta ietvaros tiek izmantots savākto datu vizualizācijai un iespējai veikt pieejamo datu atlasi, tādējādi nodrošinot reāllaika datu apakškopu identifikāciju, kas ļaus pētniekiem pēc pieprasījuma veikt reāllaika datu identificēšanu un pieprasījumu turpmākās izmantošanas nolūkiem. Metadati, kas tiek periodiski ģenerēti no savāktajiem datiem, ar speciāli izveidotā skripta palīdzību tiek periodiski nolasīti no datu ezera un aizsūtīti uz *Curator* rīku. Pēc saņemšanas, risinājums veic metadatu transformāciju uz lietotājam draudzīgo, vizuālo attēlojumu (sk. 13. attēlu), sniedzot iespēju pētniekiem veikt vaicājuma darbības, lai identificētu nepieciešamās datu apakškopas.




**14. attēls. Ekrānuzņēmums no automātiski ģenerētās metadatu vizualizācijas, ko izmanto pētnieki datu atlasei un pieprasījumu veikšanai**

Kad attiecīgais pieprasījums ir izveidots, tas tiek attēlots vizuālā formā, kā arī tiek nosūtīts kā paziņojums datu publicētājiem gan platformas saskarnē, gan kā strukturēts objekts (sk. 14. attēlu).

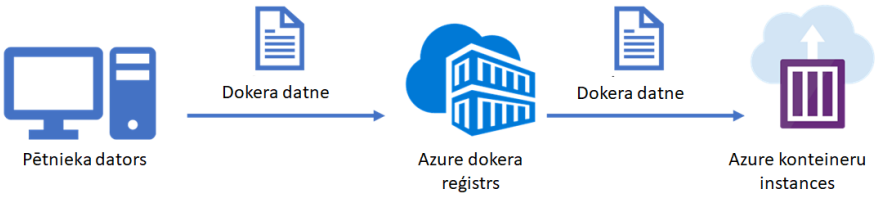



**14. attēls. Saņemtā pieprasījuma attēlojums ar atlases kritērijiem metadatu vizualizācijas un atlases rīkā Curator**

Metadatu vizualizācijas un atlases rīks *Curator* tiek nodrošināts sadarbībā ar Latvijā izvietotu digitālās veselības jaunuzņēmumu "Longenesis", nodrošinot metadatu vizualizāciju un iespēju veikt datu identificēšanu un atlasi pētniekiem - potenciāliem pētniecības sadarbības partneriem, kā arī pētījuma sponsoriem (https://www.longenesis.com/curator).

# 5. Sekundāro datu analīze

Lai nodrošinātu pamatu turpmākiem pētniecības pasākumiem ar ilgtermiņa ietekmi, pētījumā izmantoti vairāki mašīnmācīšanās algoritmi, kas nodrošina sekundārajai analīzei sniegto datu analīzi un vizualizāciju. Testa gadījumā grupa izmantoja klasterizāciju normalizētu datu kopā (vispārēja pieeja praktiskajā mašīnmācīšanās procesā, lai skaitliskās vērtība izteiktu vienotā vērtību skalā). Izmantotais testa gadījums sakņojas Python skriptu lietošanā, kas ļauj mums izmantot konteineru risinājumus, piemēram, Dokers platformu (skat. 15.attēlu), lai nodrošinātu horizontālu mērogojamību. Šī pieeja atbilst augstas pieejamības datu analīzes risinājuma ilgtermiņa mērķiem - šajā kontekstā augsta pieejamība ir sistēmas nodrošināt īpašība pētnieku piekļūšanai datiem, ņemot vērā faktisko noslodzi.



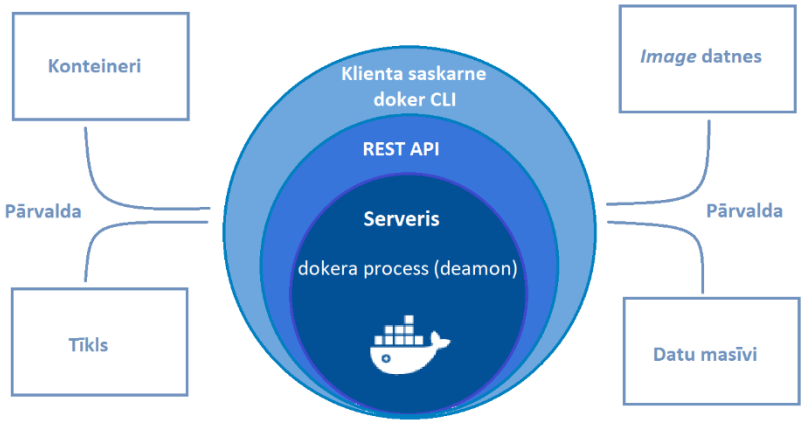

**15. attēls. Azure konteineru izmantošanas shēma**

Pieeja ļauj integrēt dažādas konfigurācijas izpildāmos skriptus (un ne tikai), izolējot no citiem, kā arī ļaujot kopā vienotā konteinera failā pievienot izpildei nepieciešamās specifiskās bibliotēkas, kā arī citus papildinājumus, kas ir specifiski konkrētam skriptam. Tomēr būtiskākā priekšrocība ir spēja balansēt kopējo slodzi uz sistēmu, ļaujot pielāgot sistēmas mērogu konkrēti risināmam uzdevumam.

Dokera infrastruktūras, kas nodrošina konteineru izmantošanu, ideja attēlota 16. attēlā.

Izveidotais risinājums kopumā demonstrē iespēju izveidot sarežģītus datu analīzes un apstrādes procesus, kas ir neatkarīgi no programmatūras un aparatūras risinājumiem. Tas nozīmē, ka izstrādāto datu apstrādes un analīzes darbplūsmu ir iespējams darbināt, kā esošajos Azure sakņotos datu ezeru risinājumos, tā arī specifiskos datu apstrādes risinājumos, kāds ir RTU HPC risinājums. Tādejādi pētniekam tiek atvieglota datu analīzes modeļu izstrāde un faktiski prasa tikai spēju sagatavot konteineru (īpaši izveidotu datni, kas satur datu analīzes programmatūra un nepieciešamās trešo pušu bibliotēkas). Atbilstoši risinājumam, konteiners pēc tam var tikt izpildīts atbilstošajā programmatūras infrastruktūrā. Lai arī pilnīgs dokera risinājums nav implementēts, ir

izveidotas būtiskākās risinājuma komponentes un jau šobrīd tiek nodrošināta abstrakcija no konkrētas programmatūras infrastruktūras pakalpojumiem, jo tiek izmantoti datu analīzes skripti konkrētu specifisku analīzes programmatūras vietā.




16. **Attēls Dokera infrastruktūras konceptuālā shēma**

CLI (*Command line interface*), REST API (*Representational state transfere Application programming interface*)

Pašreizējā implementācija nodrošina smilškastes (angl. *sandbox*) risinājumu projekta komandas pētniekiem, kā arī ārējiem pētniekiem, kas strādā pie genomu datu analīzes. Smilškastē izmantotie dati tika izveidoti un modificēti, lai varētu izmantot vispārējas nozīmes analīzes sistēmas vai pielāgotus risinājumus (piemēram, šajā gadījumā Python skriptos īstenotus klasterizācijas algoritmus), tādējādi atverot datus un visu infrastruktūru plašākam pētnieku klāstam.

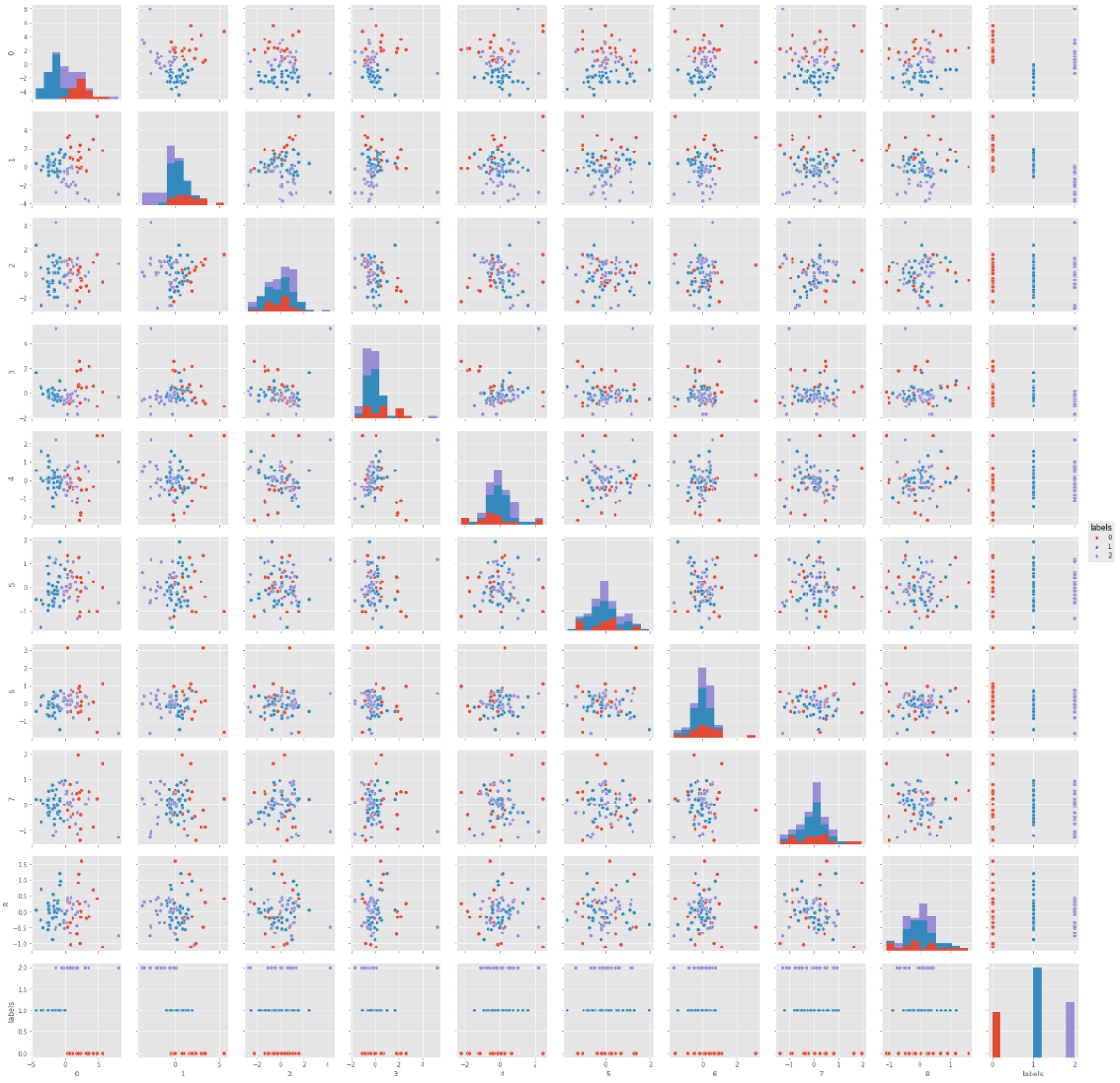
Demonstrācijas nolūkam tika izmantoti trīs klasterizācijas paņēmieni (viens no tiem attēlots 17. attēlā): skenējošie klasterizācijas algoritmi (piemēram, DBScan), uz klāsteru centru izveidi orientēti (piemēram, KMeans – k-vidējie centri) un hierarhiskas klasterizācijas algoritmi. Demonstrācijai tika izvēlēts KMeans tipa algoritms, kas sniedza labāku rezultātu, nekā minētās citas metodes. Pētījuma ietvaros galīgās demonstrācijas algoritmu komplekta iegūšanai tika izmantoti arī cita veida klasterizācijas algoritmi, bet to sniegtie rezultāti bija salīdzinoši vāji, tādēļ tie nav iekļauti galīgajā atskaitē. Tādejādi tika pārbaudīta iespēja izmantot dažādus algoritmus vienas un tās pašas darbplūsmas ietvaros.

Datu glabāšana, integrējot to ar Galaxy vidi, ievieš un demonstrē datu apstrādes darbplūsmu, kas nav atkarīga no konkrēta datu analīzes algoritma vai tā izstrādes platformas, atvieglojot tā izmantošanu pētniekam no dažādiem IKT, genomu, medicīnas vai citiem ar datiem saistītiem pētījumiem. Ir jāuzsver, ka lai nodrošinātu pētnieku ērtu piekļuvi datiem, kā arī ērtu pētījumu soļu atkārtošanu vienotā darbplūsmā, tika izvēlēts rīks Galaxy, kas līdz šim Latvijā nav lietots, kā arī citur tas parasti tiek lietots citā kontekstā. Tādēļ secinājums par vides piemērojamību Latvijas pētnieku vajadzībām ir būtisks.

Smilškaste ievieš divas datu apstrādes darbplūsmas – citokīnu un bioķīmisko analīzi. Abos gadījumos darbplūsmā tiek īstenoti šādi galvenie soļi:

- Lieko datu kolonnu atlase un izslēgšana no datu kopas;
- Katram atribūtam tiek novērtēts un attēlots vērtību sadalījums;
- Lai izvairītos no rezultātu nepareizas interpretācijas, tiek veikta vērtību mērogošana;
- Datu kopas reducēšana, izmantojot PCA (no angļu val. *Principal Component Analysis*) algoritmus;
- Klasterizācija (demonstrācijas nolūkos tiek izmantota K-Vidējie (*K-Means*) metode);
- Vērtību sadalījuma un klasteru vizualizācija (piemēri aplūkojami 17.attēlā).

17. attēlā redzamā vizualizācija ļauj identificēt datos novērojamos klasterus, kā arī vērtību sadalījumu klasteros.



**17. attēls. Klasterizācijas un vērtību sadalījuma piemērs, kas ir daļa no datu apstrādes darbplūsmas**

Tādējādi izveidotā vienotā COVID-19 pētniecības atbalsta informācijas sistēma nodrošina pilnvērtīgu attālinātu pētnieka darba vietu COVID datu analīzei – datu izgūšanai, integrācijai ar citām datu kopām, vizualizācijai, primārajai un sekundārajai analīzei. Piedāvātais tehniskais risinājums nodrošina datu uzglabāšanas un apstrādes slodžu balansēšanu datu centros vairākās institūcijās (RTU, LU un LBMC). Tāpat izstrādātā pieeja ir ērti mērogojama un, papildinot to ar jaunām darbplūsmām, izmantojama ne tikai COVID-19, bet arī citu pētījumu veikšanai.