

Referat

Fællesmøde for arbejdsgrupperne for Fortolkning, National Genomdatabase og Tools og Workflows

Dato: 06-02-2024
Enhed: NGC
Sagsbeh.: GTH.NGC
Sagsnr.: 2314076
Dok.nr.: 2912646

Dato: 15. januar 2023 kl. 13.00 – 15.00
Sted: Virtuelt (teams), link i mødeindkaldelse
Mødeleder: Bettina Lundgren
Sekretær: Gitte Tofterup Hansen og Kristina Schwenn Madsen

Dagsorden

Punkt	Aktivitet
1/24	Velkomst v/ Bettina Lundgren Referat: Bettina Lundgren bød velkommen og orienterede om, at hun har overtaget formandsposten for arbejdsgruppen for fortolkning efter Camilla Borchorst. Bettina Lundgren orienterede endvidere om, at regionerne har inviteret følgende personer med til at give input til pkt. 3 og 4: Region Sjælland: Thomas Lau Lindestrand-Hansen, Eskild Morten Landt Region Midtjylland: Michelle Nygaard Thomsen og Rikke Kærgaard-Robinson Region Syddanmark: Dorthe Ørnskov og Rasmus Højrup Pausgaard.
2/24	Status for NGC (O) v/ Bettina Lundgren, Majbrit Fonnesbæk Hansen, Nicolas Rapin Referat: Bettina Lundgren gav en status på NGC, herunder: KPI oversigt for NGC. Der er pr. 31. december 2023 i alt 23.897 genomer i den Nationale Genomdatabase. Samlet antal genomer offentliggøres på www.ngc.dk og opdateres månedligt. American Society of Human Genetics (ASHG) 75 års jubilæum. Danmark og Nationalt Genom Center udvalgt som "Thought Leader" ved American Society of Human Genetics (ASHG) 75 års jubilæumskonference i Washington DC.

Punkt	Aktivitet
	Filmen "The treatment of the future is tailored to the individual patient" præsenterer det danske initiativ. Filmen kan ses her. Ekspertermøde om personlig medicin på Christiansborg 30. november. NGC deltog ved ekspertermødet med en præsentation af erfaringerne fra de første strategiperioder. Der er stor politisk vilje til det videre arbejde med den ny strategi, og ved mødet var der input, som er i tråd med regionernes input til den ny strategi, herunder forslag om fokus på ATMP'er. Status på NGC's internationale samarbejder. Ansøgningen til EU-projektet Genome of Europe; en fælles europæisk kohorte af referencegenomer til brug i klinik og forskning er sendt ind. Der er 30 lande med i ansøgningen. Danmarks bidrag bliver et dansk referencegenom bestående af 1500 WGS'er, som nævnt ved 15. styregruppemøde 9/10-23. Data fra de 1500 danskere udstilles i genomdatabasen, og kan efterfølgende tilgås af klinikere og forskere. NGC har overtaget ledelsen af en arbejdsgruppe om Health Economics and Outcome Research i regi af 1+ Million Genomes. Arbejdsgruppen har fokus på at udvikle modeller for økonomisk impact af WGS'er og skal i første omgang udforme udkast til muligt kommende EU-opslag. Status på afrapportering på patientgrupper. Ultimo 2023 vil der være ni færdige statusrapporter for patientgrupper med arvelige sygdomme. Udarbejdelse af rapporter for kræftgrupperne påbegyndes januar 2024. Ved 17. møde i styregruppen 8/2-24 forelægges foreløbig afsluttende rapport (for arvelige sygdomme) til drøftelse. Kort øvrig status: • Komplekse/udiagnosticerede patienter – NGC har fået spørgsmål fra kliniske miljøer vedr. mulighed for rekvirering af WGS til komplekse/udiagnosticerede patienter. Indikationen indgår i patientgruppen sjældne sygdomme, og Bettina Lundgren orienterede om, at klinikere skal rette henvendelse til GM og MOMA mph. at sikre, at patienterne bliver undersøgt/får tilbudt analysen. • NGC kan nu modtage indberetning af regionale WGS'er. • Udstilling af genomdata til forskere (POC). Der er en proces i gang med proof of concept. NGC åbner forventeligt op for udstilling af genomdata til forskere Q1 2024. • NGC har forlænget licenser til fortolkningsværktøjer (QCI og VarSeq) 2024 ud. • NGC evaluerer governance setup i løbet af Q1 2024. • NGC tilbyder facilitering af ny erfaringsdeling for fortolkere Q1 2024. • Forventet forlængelse af aftale med DTU til medio 2025. • Aftale på plads mellem Danske Regioner og Indenrigs- og Sundhedsministeriet om finansiering af infrastrukturen for personlig medicin for 2024. Der var spørgsmål til, om NGC arbejder med integration til andre databaser. Majbrit Fønnesbæk Hansen bekræftede, at NGC skal til at kigge på integration til andre databaser i løbet af 2024, og at NGC er opmærksom på at

Punkt	Aktivitet
	fortolkningsvar skal indberettes til Regionernes Bio- og GenomBank, hvorfor det vil være det oplagt sted at starte med at integrere til og dermed undgå dobbelt indberetning. I den forbindelse blev det nævnt, at det også kan være en fordel at kunne se patologisvar. NGC tog bemærkningen til efterretning. Majbrit Fonnesbæk Hansen orienterede om status for indberetningsløsningen. Pt. er 7.399 genomer i 2. generation af genomdatabasen. Der kan nu indberettes følgende andre scenarier end germline: Somatic, non-merged, merged (resubmission), single sample, multiple sample, corrections, og snart også føtal prøver. I 2024 arbejdes der på også at kunne modtage forskerindberetning og indberetning af andre analysetyper end WGS og RNAseq. Nicolas Rapin orienterede om, at midlertidige filer, der ikke skal tages backup af, flyttes til denne folder på HPC'en: /ngc/nfsscratch.
3/24	Tiltag til minimering af storage-forbrug i den nationale genomdatabase (O) v/ Christian Dubois Referat: Toppen af storagekapaciteten på den nationale infrastruktur er ved at være nået, hvilket giver udfordringer ift. performance på HPC'en. Christian Dubois orienterede om, at NGC på den baggrund har set på, hvordan storage pt. bliver brugt hhv. i genomdatabasen, hvor NGC er dataansvarlige, og i det regionale produktionsflow, hvor regionerne er dataansvarlige. NGC har rådført sig med internationale eksperter via governance (advisory boards og arbejdsgrupper) ift. til tiltag til minimering af storageforbruget i den nationale genomdatabase. På den baggrund har NGC besluttet fremover at gemme CRAM-filer, da det er i tråd med internationale standarder. Beslutningen om alene at gemme CRAM-filer i genomdatabasen reducerer filstørrelsen pr. WGS fra 90 til 27 GB (gennemsnitlig beregning) og vil potentielt kunne medføre en samlet besparelse på ca. 6 mio. kr. årligt. WGS-faciliteterne bemærkede, at man skal være opmærksom på forholdet mellem compute-udgifter og storage-besparelse. Bilag 2 viser, at det vil koste 1,3 mio. kr. (til compute) at aligne og gemme data som CRAM med NGC's værktøj hertil (Sentieon). Besparelsen i storage første år er 5,9 mio. Det vil altså være en positiv business case allerede 1. år. WGS-faciliteterne bemærkede endvidere, at sporbarhed er vigtig i den kliniske produktion.
4/24	Drøftelse af mulig reduktion af storageforbrug i det regionale produktionsflow gennem reduktion af filtyper og komprimering (D) v/Christian Dubois Referat:

Punkt	Aktivitet
	Der er mange elementer i storageforbruget, herunder er det en central omkostningsdriver, hvorfor det er vigtigt med konsensus omkring, hvordan vi i Danmark gemmer analysedata fra helgenomsekventeringer. Ift. det regionale produktionsflow, hvor regionerne er dataansvarlige, har NGC set på forskellige måder at reducere storageforbruget på mhp. at illustrere potentialet for dataminimering. Ved mødet blev emnet drøftet ud fra to temaer: Tema 1: Hvilke filtyper bør gemmes/kan slettes mhp. storageminimering? Tema 2: Hvordan sikrer vi, at fortolkningsgrundlaget bevares ifm. komprimering af FASTQ-filer til CRAM? På tværs af de to temaer blev også et tredje tema berørt: Tema 3: Forskning i lyset af storageminimering (Hvad vil være forskeres forventning og krav til data i genomdatabasen storageminimering taget i betragtning?). <u>Ad. tema 1</u> WGS-faciliteterne tilkendegav, at filtyperne BAM og gVCF produceres ikke længere i den nationale pipeline. Reduktion af storage af disse filtyper vedrører derfor den historiske produktion. <i>BAM-filer</i> WGS-faciliteterne tilkendegav, at BAM-filer (hvor der er genereret CRAM-filer) kan slettes uden problemer (dvs. fortolkningsgrundlaget vil fortsat kunne genskabes), der kræves blot en instruks fra de afdelinger/regioner, som er dataansvarlige, hvorefter sletning kan iværksættes. Der var konsensus herom blandt arbejdsgruppernes medlemmer. For prøver hvor der ikke er genereret CRAM-filer, men kun BAM, skal der bruges compute til at generere en CRAM-fil. Samlet set vil det dog stadig være en betydelig besparelse. <i>BCL-filer</i> WGS-facilitet Vest tilkendegav, at BCL-filer skal gemmes, indtil faciliteten er færdig med at bruge dem. Herefter kan de slettes. <i>gVCF-filer</i> Der var blandt medlemmerne ønske om, at gVCF fortsat gemmes. Det vil være en fordel for forskning, og det vil også være de data, der kan bruges til at lave frekvensdata til klinisk brug. Mere generelt ift. drøftelsen af, hvilke filtyper der bør gemmes/kan slettes, blev det understreget, at CNV-, MSI- og repeatkald skal gemmes, samt at VCF-filer skal gemmes af hensyn til forskning, variant og frekvensdatabaser. <u>Ad. 2</u> Næstformændene bakkede op om, at data komprimeres til CRAM formatet, blot blev det pointeret, at det er vigtigt, at rådata kan genskabes ud fra CRAM-filen. Blandt de øvrige medlemmer var der overordnet set også opbakning til komprimering til CRAM-formatet.

Punkt	Aktivitet
	Komprimering af FASTQ til CRAM var ikke WGS-faciliteternes ønske. De foretrækker SRING-formatet med den begrundelse, at CRAM er et alignmentformat, hvorfor de mener, at SPRING er mere velegnet til at komprimere ikke-alignet data. Hertil blev det fra et medlem bemærket, at der ikke er alignment indenfor SPRING, og at det vurderes vigtigt at gemme alignment-data, hvis man skal tilbage og se, hvad data er kaldt på. Der blev gjort opmærksom på, at RNAseq opererer på FASTQ-data, hvorfor man her ikke anbefaler at komprimere til CRAM, alternativt anbefaler at der komprimeres til en ikke-alignet CRAM. <u>Behov for yderligere afklaring ifm. komprimering til CRAM-formatet</u> Ved mødet blev det drøftet, hvilken CRAM-fil, der skal gemmes i genomdatabasen – den der er brugt i klinisk produktion eller en, NGC komprimerer på baggrund af den indberettede FASTQ-fil. Der er ønske om at nå en konsensus herom mhp. at sikre optimal brug af genomdatabasen for hhv. forskning og klinik. Ift. at nå konsensus.
5/24	Arbejdsgrupperne frem til nu og efter medio 2024 (D) v/Klaus Brusgaard, Martin Larsen, Charles Vesteghem og Bettina Lundgren Referat: Arbejdsgrupperne drøftede erfaringerne med at arbejde rådgivende i governancestrukturen mhp. at der sikres fælles offentligt ansvar for og samarbejde om udvikling, implementering og drift af den nationale infrastruktur for personlig medicin. Drøftelsen fokuserede på hhv. hvor vi er pt. og hvad medlemmerne mener, kan blive vigtigt fremadrettet. Der var enighed om, at der med arbejdsgrupperne er skabt et godt fundament for samarbejde på tværs af regionerne. Der er opbygget fælles viden, og både indenfor og på tværs af regionerne har afdelingerne lært af hinanden. Den del finder medlemmerne det vigtigt at bygge videre på. Fremadrettet ser medlemmerne det som vigtigt, at ambition ift. genomdatabasen defineres tydeligere, herunder blev en nærmere beskrivelse af, hvad snitfladerne mellem regionerne og NGC skal være, efterspurgt ift. at medlemmerne kan rådgive målrettet. Medlemmerne bidrager gerne med input hertil. Fremadrettet vil det fortsat være vigtigt at se på sammensætning af grupperne ift. de emner, der skal rådgives om. Herunder ses det også som vigtigt, at det sikres at information når ud til alle relevante afdelinger/parter. Det kan enten være gennem bredere repræsentation i arbejdsgrupperne eller styrket koordination og kommunikation til relevante afdelinger/parter.
5/24	Evt. Referat: ingen bemærkninger.

Deltagere

Arbejdsgruppen for Fortolkning

Bettina Lundgren (formand), NGC
Klaus Brusgaard (næstformand), indstillet af Region Syddanmark
Charlotte Brasch Andersen, indstillet af Dansk Selskab for Kliniske Akademikere
Lise Barlebo Ahlborn, indstillet af Dansk Selskab for Kliniske Akademikere
Camilla Darum Sørensen, indstillet af Dansk Selskab for Kliniske Akademikere
Linea Cecilie Melchior, udpeget af Dansk Selskab for Kliniske Akademikere
Inge Søkilde Pedersen, indstillet af Region Nordjylland
Lotte Andreasen, indstillet af Region Midtjylland
Lasse Kjær, indstillet af Region Sjælland
Karen Grønskov, indstillet af Region Hovedstaden
Maria Rossing, supplerende ekspert fra NGC's WGS-faciliteter (øst)
Dorte Launholt Lildballe, supplerende ekspert fra NGC's WGS-faciliteter (vest)

Arbejdsgruppen for Tools og Workflows

Nicolas Rapin (formand), NGC
Martin Larsen (næstformand), indstillet af Region Syddanmark
Frederik Otzen Bagger, indstillet af NGC's WGS-faciliteter (øst)
Søren Vang, indstillet af NGC's WGS-faciliteter (vest)
Mads Sønderkær, indstillet af Region Nordjylland
Michael Knudsen, indstillet af Region Midtjylland
Vibe Skov, indstillet af Region Sjælland
Mads Bak, indstillet af Region Hovedstaden
Rasmus Brøndum, indstillet af Aalborg Universitet
Nicolai Birkbak, indstillet af Aarhus Universitet
Lars Andersen, indstillet af Syddansk Universitet
Karina Banasik, indstillet af Københavns Universitet
Lars Rønn Olsen, indstillet af Danmarks Tekniske Universitet
Rasmus Marvig, supplerende ekspert fra NGC's WGS-faciliteter (øst)
Filipe Vieira, supplerende ekspert fra Genomisk Medicin, Rigshospitalet

Arbejdsgruppen for den Nationale Genomdatabase

Majbrit Fonnesbæk Hansen (formand), NGC
Charles Vesteghem (næstformand), indstillet af Region Nordjylland
Christen Andersen, indstillet af Københavns Universitet (afbud)
Ditte Demontis, indstillet af Aarhus Universitet
Dorte Lildballe, indstillet af Region Midtjylland
Frederik Otzen Bagger, indstillet af NGC WGS Øst
Inge Søkilde Pedersen, indstillet af Aalborg Universitet
Mads Jørgensen, indstillet af Region Syddanmark
Martin Jakob Larsen, indstillet af Syddansk Universitet
Mette Christiansen, indstillet af NGC WGS Vest
Morten Dunø, indstillet af Region Hovedstaden

Øvrige deltagere, som er inviteret af regionerne mhp. at sikre bred inddragelse af regionernes faglige miljøer:

Region Sjælland: Thomas Lau Lindestrand-Hansen, Eskild Morten Landt

Region Midtjylland: Michelle Nygaard Thomsen og Rikke Kærgaard-Robinson

Region Syddanmark: Dorthe Ørnskov og Rasmus Højrup Pausgaard

Fra NGC deltager: Christian Dubois, Peter Johansen, Christoffer Egeberg Hother og
Mikael Kronborg

Pkt. 3/24 Tiltag til minimering af storage-forbrug i den nationale genomdatabase (O) v/Christian Dubois

Indstilling

Det indstilles, at arbejdsgrupperne tager orientering om tiltag til minimering af storage-forbrug i den nationale genomdatabase til efterretning.

Baggrund

NGC har konstateret, at der er udfordringer omkring storagekapacitet. For at udnytte den eksisterende hardware optimalt er det NGC's vurdering, at den samlede mængde af data med fordel kan reduceres gennem hhv. komprimering og sletning af visse filtyper. At filer ikke opbevares flere steder vil samtidig understøtte efterlevelse af pligt til dataminimering efter databeskyttelsesforordningen.

På den baggrund har NGC set på, hvordan den samlede mængde af data i den nationale genomdatabase kan reduceres gennem komprimering af visse filtyper.

Når der i dag produceres en helgenomsekventering (WGS 30X) i NGC, genereres der en række filer. Kliniske genomdata på NGC's HPC kan overordnet opdeles i to kategorier:

1. **Data i den nationale genomdatabase.**
Dette data modtager NGC via regionernes indberetning, og NGC er dataansvarlig.
2. **Data der produceres i forbindelse med WGS produktionsflowet**
Dette data opbevares i de regionale foldere på HPC'en, og regionerne er dataansvarlige.

Bilag 1 beskriver tiltag, som NGC vil iværksætte som dataansvarlig mhp. minimering af storage-forbrug i den nationale genomdatabase.

Bilag 2 omfatter forslag til, hvordan regionerne kan understøtte en minimering af mængden af data i den regionale del af NNF produktionsflowet på HPC'en og herunder samtidig understøtte minimering af omkostninger hertil.

Bilagene blev præsenteret ved 16. møde i styregruppen for implementering af personlig medicin den 8/12-2023. Ved mødet var der opbakning til at se på området, og styregruppen bakkede op om NGC's plan om inddragelse af de faglige miljøer, blandt andet via dette fællesmøde med de tekniske arbejdsgrupper.

Som optakt til drøftelserne orienterer NGC kort om de tiltag til minimering af storage-forbrug i den nationale genomdatabase, der igangsættes, hvor NGC er dataansvarlig.

Tiltagene vil potentielt kunne medføre en samlet besparelse på ca. 6 mio. kr. årligt beregnet ved opbevaring af 60.000 WGS med en dybde på 30X (baseret på en reduktion i filstørrelsen pr. WGS fra 90 til 27 GB, der opnås ved at komprimere datafiler fra FASTQ til CRAM-formater).

NGC's valg af tiltag/komprimeringsmetode tager afsæt i rådgivning fra NGC's internationale advisory board. Det har videre været en forudsætning for NGC's valg af komprimeringsmetode, at komprimeringen skal ske på baggrund af internationalt anerkendte og åbne standarder og med værktøjer, som ikke er underlagt licensomkostninger eller binding til bestemte leverandører. Dette er gjort for at sikre, at indholdet i den nationale genomdatabase er let tilgængeligt, har høj driftsstabilitet og omkostningerne i forbindelse med komprimeringen holdes på et minimum. Endeligt har det vægtet positivt, at det valgte format forventes at give en effektiv samarbejdsplatform internationalt.

Bilag

Bilag 1: Notat vedr. tiltag til minimering af storageforbrug i den nationale genomdatabase

Pkt. 4/24 Drøftelse af mulig reduktion af storageforbrug i det regionale produktions-flow gennem reduktion af filtyper og komprimering (D) v/ Nicolas Rapin

Indstilling

Det indstilles, at arbejdsgrupperne drøfter NGC's oplæg til mulig reduktion af storageforbrug i det regionale produktions-flow gennem reduktion af filtyper og komprimering.

Problemstilling

Dette punkt er en forlængelse af pkt. 3/24. Punktet vedrører, hvordan man i den del af HPC'en, hvor regionerne er dataansvarlige, kan minimere omfanget af data som genereres i produktionsflowet. Dette for sikre en hensigtsmæssig udnyttelse af den tilgængelige storagekapacitet på HPC'en samt for at sikre dataminimering efter databeskyttelsesforordningen.

NGC har set på, hvordan den samlede mængde af data kan reduceres gennem hhv. komprimering og sletning af visse filtyper. Vedlagte notat (bilag 2) omfatter forslag til, hvordan regionerne kan understøtte en minimering af mængden af data i den regionale del af NNF produktionsflowet på HPC'en og herunder samtidig understøtte minimering af omkostninger hertil.

Regionerne har pt. ikke storage-omkostninger forbundet med opbevaring af data, der genereres ifm. sekventering og processering af NNF-prøver. Disse omkostninger oppebæres for nærværende af NGC via NNF-bevillingen, der udløber medio 2024.

Det er op til regionerne – som dataansvarlige for den data, som genereres i produktionsflowet – at afgøre, hvilket data de ønsker at gemme.

Løsning

Der lægges op til, at arbejdsgrupperne drøfter NGC's oplæg til minimering af datamængden, der produceres i forbindelse med WGS-produktionsflowet, hvor regionerne er dataansvarlig.

Ifm. udarbejdelse af oplægget har NGC afdækket, hvilke filtyper der gemmes i det regionale workflow, hvad de fylder og eventuelle muligheder for at slette visse af filerne samt eventuelt foretage komprimering af visse af filerne (bilag 2, tabel 2). NGC har set på mulighederne for at reducere datamængden pr. WGS, ved at:

A: Reducere antallet af filtyper, der gemmes pr. WGS

B: Gemme rådata (FASTQ) i et komprimeret format (CRAM)

Ved scenarie A (Bilag 2, tabel 2) vurderer NGC, at der – forudsat der fremover ikke gemmes BCL, BAM og GVCF-filer på den enkelte prøve – potentielt kan realiseres

en besparelse på op til 215 GB pr. WGS, svarende til 64 %. Denne indsats vil kunne iværksættes umiddelbart og vil kunne realisere minimeringen af data løbende i takt med at filerne slettes.

Da regionerne er dataansvarlige, vil det kræve, enten at regionerne selv sletter filerne, eller at NGC får en instruks om at gøre det som databehandler.

Ved scenarie B vurderer NGC, at der potentielt kan realiseres en samlet reduktion af omfanget af data pr WGS (forudsat at der reduceres i fil-formaterne jf. scenarie A) på op til 305 GB eller 91%.

Vurderingen er baseret på et scenarie, hvor regionerne vælger at opbevare rådata i samme tabsfri CRAM-format, som NGC og øvrige europæiske genominitativer (England, Norge, Holland, Finland og Frankrig) bruger af filformat til opbevaring af rådata.

Der er yderligere potentiale for reduktion af omfanget af opbevaring, hvis regionerne bruger den nationale genomdatabase til opbevaring og ikke gemmer kopier på den regionale del af produktionsflowet på HPC'en. Dette vil i givet fald kræve en videre undersøgelse af fortolkningen af forarbejderne til det eksisterende lovgrundlag, således at der er en klinisk relevant adgang til at hente kopier fra databasen til brug for patientbehandling. Dette vil medføre en samlet reduktion af omfanget af data pr WGS på op til 355 GB eller 99 %.

Bilag

Bilag 2: Notat vedr. mulige reduktioner af storageforbrug gennem minimering af omfang af data