



Projektrapport:

Behov av standardisering inom genomikområdet

Projekt: SWElife Skalbara informatiklösningar

Målgrupp: AU Informatik

Datum: 2022-12-22

Version: 1.0

Sammanfattning

Denna rapport beskriver vilka behov kring standardisering som finns inom genomikområdet inklusive information som är viktig att hantera på ett standardiserat sätt. Rapporten fokuserar på standardiseringsbehov i den del av next generation sequencing-processen (NGS) där avancerade bioinformatiska dataanalyser utförs samt i viss mån i de delar av NGS-processen som kommer före och efter bioinformatikanalysen. Vid NGS-processen skapas och används stora mängder information. För att kunna nyttja denna information på ett effektivt sätt i olika steg av NGS-processen samt möjliggöra senare återanvändning och ge en hög sökbarhet, så krävs standardisering av informationen.

Kartläggningen har genomförts för flera av Genomic Medicine Sweden:s (GMS) sjukdomsområden: barncancer, hematologi, solida tumörer och sällsynta diagnoser. Behoven har uppdelats i sådana som är gemensamma för de kartlagda sjukdomsområdena respektive behov som är specifika för ett enskilt sjukdomsområde.

Rapporten beskriver allmänna rekommendationer kring forskning samt något om internationella standardiseringsinitiativ som pågår i ett av de samverkansområden som GMS deltar i.

Vi föreslår att GMS arbetar vidare på de initiala informationsspecifikationer som har påbörjats i detta projekt, inkluderar GMS övriga sjukdomsområden (farmakogenomik, infektionssjukdomar och komplexa sjukdomar) samt harmoniserar dessa med informationsspecifikationer som håller på att tas fram inom andra samverkansområden.

Denna rapport bör hanteras som ett levande dokument inom GMS och nya versioner av den kan skapas allteftersom GMS får djupare kunskaper om olika standardiseringsbehov. (En Word-version av rapporten kommer att levereras till beställaren.)

Innehåll

1. INLEDNING	6
1.1. BAKGRUND.....	6
1.2. SYFTE	6
1.2.1. <i>Bidra till måluppfyllnad</i>	6
1.2.2. <i>Mål för Genomic Medicine Sweden</i>	7
1.2.2.1. Etablerad nationell genomikplattform och kunskapsdatabas	7
1.2.2.2. Ökat nyttjande av genomik- och hälsodata för forskning och innovation	8
1.2.3. <i>Syften med standardisering</i>	8
1.3. MÅL FÖR DELPROJEKTET SKALBARA INFORMATIKLÖSNINGAR	10
1.4. AVGRÄNSNINGAR	10
1.5. ANGREPPSSÄTT OCH GENOMFÖRANDE	11
1.5.1. <i>Angreppssätt behovskartläggning</i>	11
1.5.2. <i>Genomförande av behovskartläggning</i>	12
1.6. INTRESSETER	12
1.7. REKOMMENDATIONER.....	14
2. KARTLÄGGNING AV BEHOV SOM ÄR GEMENSAMMA FÖR FLERA SJUKDOMSOMRÅDEN	15
2.1. VY.....	15
2.2. AKTIVITETER	17
2.2.1. <i>Ta prov</i>	17
2.2.2. <i>Sekvensera prov</i>	17
2.2.3. <i>Bearbeta sekvens med bioinformatik</i>	17
2.2.3.1. Alignment	18
2.2.3.2. Variant calling.....	18
2.2.3.3. Variantannotering.....	18
2.2.4. <i>Tolka och visualisera varianter</i>	19
2.3. INFORMATIONSMÄNGDER	19
2.3.1. <i>Remiss</i>	24
2.3.2. <i>Patientdata samt fenotypdata och annan klinisk data</i>	24
2.3.3. <i>Provdata</i>	27
2.3.4. <i>Samtycke</i>	27
2.3.5. <i>Genomiknod</i>	28
2.3.6. <i>Bibliotekspreparationsdata</i>	28
2.3.7. <i>Körningsdata</i>	29
2.3.8. <i>Initial sekvens</i>	29
2.3.9. <i>Sekvens</i>	29
2.3.10. <i>QC-data (från sekvensering)</i>	31
2.3.11. <i>QC-data (avser sekvensens kvalitet)</i>	32
2.3.12. <i>Samlad QC-data</i>	32
2.3.13. <i>Analysdata</i>	33
2.3.14. <i>Alignad sekvens</i>	33
2.3.15. <i>Referensgenom</i>	34
2.3.16. <i>Varianter</i>	35
2.3.17. <i>Konfiguration för bioinformatik</i>	37
2.3.18. <i>Bioinformatisk pipelinespecifikation</i>	37
2.3.19. <i>Körningslogg</i>	39
2.3.20. <i>Remissvar</i>	39
2.3.21. <i>Genlista</i>	40
2.3.22. <i>Tolkade varianter</i>	40
2.4. INFORMATIONSBÄRARE	42
2.4.1. <i>Patientjournal och laboratoriesystem</i>	42
2.4.2. <i>Internationella annoteringsdatabaser</i>	42
2.4.3. <i>Nationella variantdatabaser</i>	42

2.4.4.	Tolkningsverktyg	44
3.	KARTLÄGGNING AV BEHOV SOM ÄR SPECIFIKA FÖR ENSKILDA SJUKDOMSOMRÅDEN	45
3.1.	BARNCANCER	45
3.1.1.	Scenario	45
3.1.2.	Vy	45
3.1.3.	Aktiviteter	45
3.1.4.	Informationsmängder	45
3.1.4.1.	Patientdata samt fenotypdata och annan klinisk data	46
3.1.4.2.	Bibliotekspreparationsdata	46
3.1.4.3.	Panel av normaler	46
3.1.4.4.	Tolkade varianter	46
3.1.4.5.	Remissvar	46
3.1.5.	Informationsbärare	46
3.2.	HEMATOLOGI	47
3.2.1.	Scenario	47
3.2.2.	Vy	48
3.2.3.	Aktiviteter	49
3.2.3.1.	Ta prov	49
3.2.3.2.	Sekvensera prov	49
3.2.3.3.	Bearbeta sekvens med bioinformatik	49
3.2.4.	Informationsmängder	49
3.2.4.1.	Patientdata samt fenotypdata och annan klinisk data	49
3.2.4.2.	Provdata	50
3.2.4.3.	QC-data (från sekvensering)	50
3.2.4.4.	Tolkade varianter	50
3.2.5.	Informationsbärare	50
3.3.	SOLIDA TUMÖRER	51
3.3.1.	Scenario	51
3.3.2.	Vy	52
3.3.3.	Aktiviteter	53
3.3.3.1.	Ta prov	53
3.3.3.2.	Sekvensera prov	53
3.3.3.3.	Bearbeta DNA-sekvens med bioinformatik	53
3.3.3.4.	Bearbeta RNA-sekvens med bioinformatik	54
3.3.4.	Informationsmängder	54
3.3.4.1.	Projektbeskrivning/forskningsinitiativ	54
3.3.4.2.	Provdata	54
3.3.4.3.	Samtycke	54
3.3.4.4.	Sample sheet	55
3.3.4.5.	Sekvens	55
3.3.4.6.	Kopietal	55
3.3.4.7.	Varianter	56
3.3.4.8.	Fusionsgener	56
3.3.4.9.	Identitets-SNP	56
3.3.4.10.	Biomarkörer	56
3.3.4.11.	Paneler av normaler	57
3.3.4.12.	QC-data (avser sekvensens kvalitet)	57
3.3.4.13.	Samlad QC-data	58
3.3.4.14.	Tolkade varianter	58
3.3.5.	Informationsbärare	58
3.3.5.1.	Variantdatabas	58
3.4.	SÄLLSYNTA DIAGNOSER	59
3.4.1.	Scenario	59
3.4.2.	Vy	60
3.4.3.	Aktiviteter	61
3.4.3.1.	Ta prov	61
3.4.3.2.	Sekvensera prov	61

3.4.3.3.	Bearbeta sekvens med bioinformatik	61
3.4.3.4.	Tolka och visualisera varianter.....	61
3.4.3.5.	Rapportera varianter till databaser	62
3.4.4.	<i>Informationsmängder</i>	62
3.4.4.1.	Remiss	62
3.4.4.2.	Provdata.....	62
3.4.4.3.	Analysdata.....	62
3.4.4.4.	Släkträd	62
3.4.4.5.	Varianter.....	63
3.4.4.6.	Remissvar	63
3.4.4.7.	Tolkade varianter	63
3.4.5.	<i>Informationsbärare</i>	64
3.4.5.1.	Internationella annoteringsdatabaser	64
3.4.5.2.	Beacon	64
3.4.5.3.	Matchmaker Exchange.....	64
4.	INFORMATIONSSPECIFIKATION.....	66
4.1.	BAKGRUND.....	66
4.2.	METOD	66
4.3.	RESULTAT.....	67
5.	REKOMMENDATIONER AVSEENDE FORSKNING.....	70
6.	SAMVERKAN.....	73
6.1.	1+MG OCH B1MG-INITIATIVET	73
6.1.1.	<i>Kvalitetsmått för NGS-processen</i>	73
6.1.2.	<i>Bästa praxis vid delning och länkning av fenotyp- och genetisk data</i>	74
6.1.3.	<i>Ramverk för fenotyp- och klinisk metadata</i>	74
BILAGA A.	ORDLISTA.....	76
BILAGA B.	DELTAGARLISTA	78
BILAGA C.	TECKENFÖRKLARING	80
BILAGA D.	INFORMATIONSSPECIFIKATION.....	81
	FORMULÄREXEMPEL RESULTATRAPPORT SOLIDA TUMÖRER	81
	TEKNISKA FORMAT OCH VERKTYGSSTÖD FÖR INFORMATIONSSPECIFIKATIONER (OPENEHR-TEMPLATES) ..	89

EXTERN BILAGA

BILAGA E. Omvärldsanalys internationella projekt och databaser.pdf

1. Inledning

1.1. Bakgrund

Genomic Medicine Sweden (GMS) är ett projekt som under en 10-årsperiod kommer att arbeta för att införa precisionsmedicin brett inom hälso- och sjukvård. En av uppgifterna är att etablera en gemensam nationell infrastruktur för lagring av genomikdata och bearbetning av bioinformatiska analyser samt även etablera visualiseringar av den information som finns lagrad i plattformen. Under projektet kommer arbetsnamnet nationell genomikplattform (NGP) att användas som benämning för infrastrukturen.

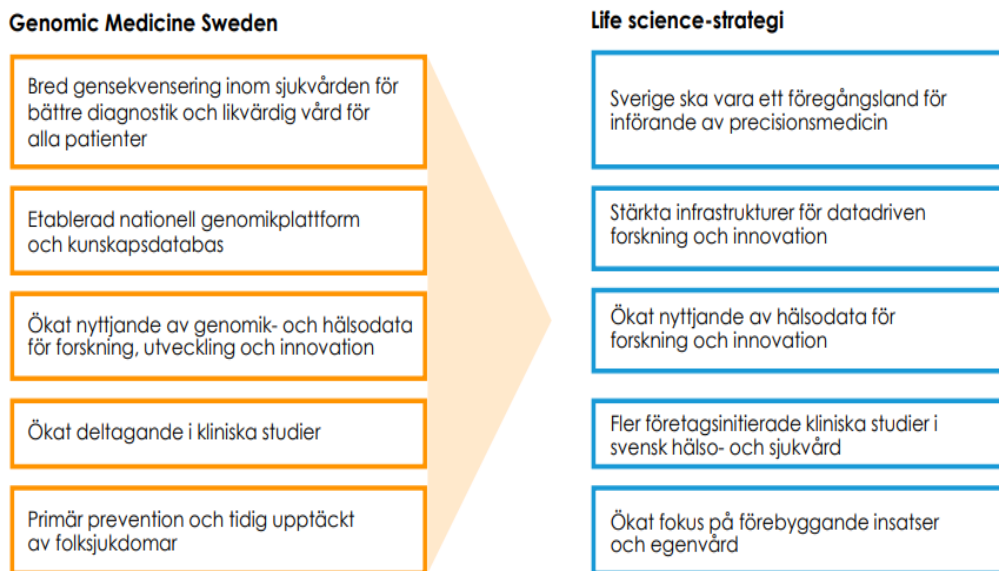
Vid arbete med NGP kommer frågor kring standardisering att aktualiseras. Regioner och universitetssjukhus lagrar idag genomikinformation och metadata på olika sätt, men för en effektiv hantering i en nationell infrastruktur bör informatiken harmoniseras.

SWElife-projektet 'Skalbara informatiklösningar inom GMS: nationell standardisering, avancerade analyser och visualisering av genomik- och hälsodata' är ett delprojekt inom GMS som löper under perioden 2020-01-01 till och med 2022-12-31. Delprojektet är uppdelat i tre arbetspaket och denna rapport avser arbete inom arbetspaket 1, 'Standardisering av metadata och hälsodata'.

1.2. Syfte

1.2.1. Bidra till måluppfyllnad

GMS utgör en del av Sveriges nationella strategi för livsvetenskaper. Denna rapport bidrar till att uppfylla delar av de fem övergripande målen för GMS¹.

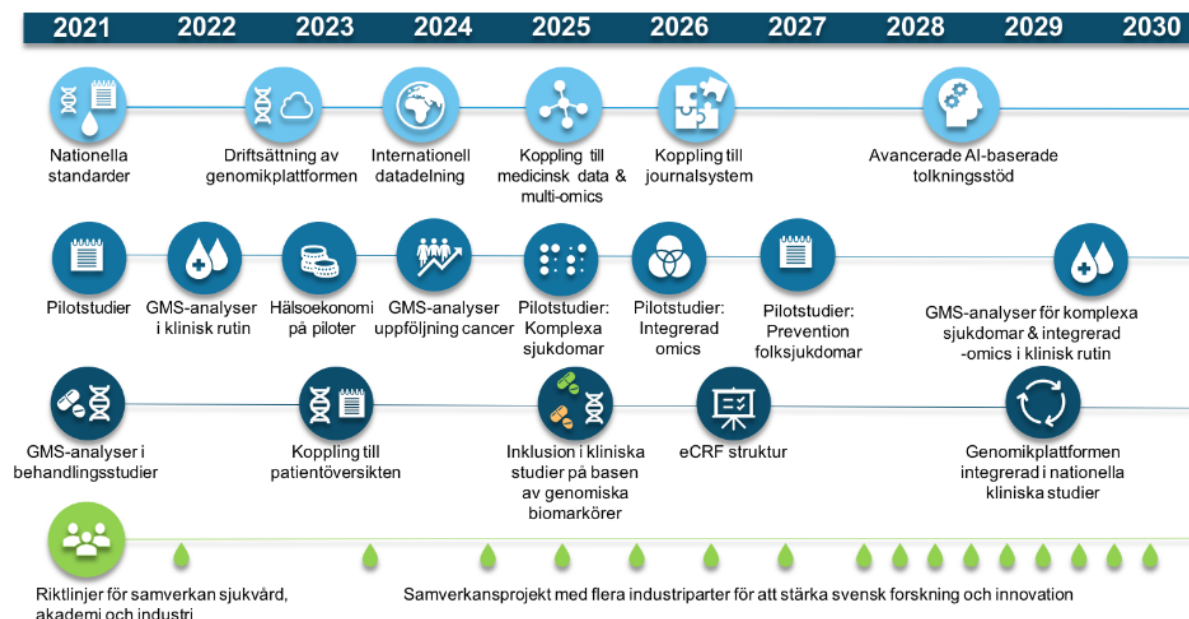


Figur 1. GMS övergripande mål kopplade till målen i den nationella strategin för livsvetenskaper

¹ Informationsblad om GMS strategiplan 2021-2030: <https://genomicmedicine.se/wp-content/uploads/2021/01/GMS-10-a-%CC%8Arsplan.pdf>

1.2.2. Mål för Genomic Medicine Sweden

Formuleringarna i detta avsnitt är hämtade från GMS projektplan 2021-2030².



Figur 2. GMS milstolpar för perioden 2021-2030

1.2.2.1. Etablerad nationell genomikplattform och kunskapsdatabas

En grundläggande förutsättning för ett lyckat införande av precisionsmedicin i svensk sjukvård är strukturering och samordning av genetiska data på en nationell nivå. Det samlade informationsunderlaget kan sedan ligga till grund för diagnostik av individen, nationell statistik och utfallsanalyser. Vidare är aggregerad och strukturerad data en nödvändighet i utvecklingen av nästa generations diagnostiska beslutsstöd och för medicinsk forskning och utveckling. En nationell IT-infrastruktur för lagring och beräkning av genomikdata i hälso- och sjukvården byggs upp. Den nationella genomikplattformen och kunskapsdatabasen kommer under projektets gång att expandera för att inkludera, integrera eller interagera med ytterligare datakällor inom sjukvården.

GMS milstolpar för att nå målet är:

- Nationella standarder för prov- och patientdata kopplade till genomikdata tas fram (år 1)
- Fullskalig pilot för nationell IT-infrastruktur för lagring och beräkning av genomikdata (år 2)
- Driftsättning av nationell genomikplattform för lagring och beräkning av genomikdata (år 3)
- Anslutning till projekt för grundläggande internationell datadelning (år 3)
- Koppling av infrastrukturen till andra typer av medicinska data och multi-omics data (år 3-5)
- Integrering av infrastruktur i journalsystem (år 5)
- Driftsättning av AI-baserade supportsystem och tolkningsstöd/visualisering (år 7-9)

² GMS projektplan 2021-2030: <https://genomicmedicine.se/wp-content/uploads/2021/01/GMS-projektplan-2021-2030.pdf>

1.2.2.2. Ökat nyttjande av genomik- och hälsodata för forskning och innovation

För att stärka svensk forskning och innovation inom precisionsmedicin avser GMS att utveckla en struktur och ett ramverk för samverkan mellan sjukvård, akademi och industri. Detta görs tillsammans med universitetens och regionernas innovationskontor och med representation från industrin. GMS kan som nationell infrastruktur erbjuda en unik resurs för enhetlig populationsbaserad molekyllär profilering av patientgrupper inom hälso- och sjukvården. Tillsammans med landets registerstrukturer kan GMS bidra med data för världsledande klinisk forskning som bygger på enhetliga "real-world" data avseende genomik kopplad till behandlingsrespons, biverkningar, livskvalitet, läkemedelsförskrivning och hälsoekonomi. Genom att genomlysna finansiella och legala aspekter, immaterialrättsliga frågor och utarbeta riktlinjer för samverkan skapas unika förutsättningar för ett ökat nyttjande av genomik- och hälsodata för forskning, innovation och samverkan med industrin, med bibehållet skydd för den personliga integriteten. Samtidigt initieras pilotprojekt inom GMS i syfte att stärka svensk forskning och innovation inom precisionsmedicin.

GMS milstolpar för att nå målet är:

- A. Riktlinjer för samverkan mellan sjukvården, akademien och industrin utarbetade (år 1-2)
- B. Etablering av ett nationellt förankrat ramverk för innovation och företagssamverkan samt modell för uppföljning (år 1-3)
- C. Initiering av samverkansprojekt inom GMS med deltagande av flera industriparter i syfte att stärka svensk forskning och innovation (år 1-10)
- D. Workshops och hearings för att informera om GMS och få återkoppling om pågående forsknings- och innovationsaktiviteter (årligen)
- E. Tillgängliggörande av genomik- och hälsodata till akademi och näringsliv (år 3-10)

1.2.3. Syften med standardisering

För att smidigt och resurseffektivt kunna dela data och mjukvara som använder sig av denna data måste likadana saker beskrivas på samma sätt hos alla verksamheter som vill samarbeta.

Information som skapas i vården kan vara utformad på olika sätt. Exempelvis narrativ, det vill säga ostrukturerad textmassa, eller textmassa som följer rubriker i en mall, det kan vara listor av diagnoser, resultat av laboratorieanalyser etc. För att kunna använda dessa olika typer av information automatiskt behöver texten ofta bearbetas och omtolkas, något som är tidskrävande och kan leda till kvalitetsproblem.

Även strukturerad information som skickas mellan två system kan behöva omtolkas och "mappas om" utifrån den inte är baserad på samma struktur eller standard. Det finns olika typer av omtolkningsproblem som kan uppstå, se tabellen nedan som beskriver typ I, II och III.

För typ I är data från två skilda system, som vi kallar A och B, såpass lika att den går att omtolka (översätta) med en algoritm och när algoritmen är införd i systemintegrationen så sker alla omtolkningar sedan automatiskt.

För typ II finns det flera utmaningar i att skapa en omtolkningsalgoritm, utmaningar som riskerar patientsäkerheten. I de flesta fall av kategori II görs därför omtolkningen manuellt av en sakkunnig person för varje patient/överföringstillfälle, vilket medför tidsödande s.k. dubbeldokumentation. Exempel på detta är överföring via fax eller PDF som läses av sjuksköterskor eller läkare och manuellt matas in strukturerat i mottagande system.

För typ III är det omöjligt att omtolka informationen, det går varken manuellt eller med algoritmer. Formatet av insamlad data kan då inte otvetydigt översättas till ett annat format.

Tabell 1. Exempel på olika typer av skillnader i dokumentationsstrukturer (tabellrader med Typ I - III) i två system (kolumner med exempelsystem A och B)

Exempel	System A	System B
Typ I A <-- --> B Lösbart med algoritm	Födelsevikt: 3300g Datum: 1954-03-13	Kroppsvikt: 3,3 kg Tidpunkt: 13 Mar 1954
Typ II A --> B Semantiska förluster och förvrängningar i omtolkningen. <i>Algoritmiskt svåröslbart (ofta omöjligt med dagens teknik) samt patientsäkerhetsmässigt riskfyllt...</i> <i>...men görs ganska ofta manuellt av vårdpersonal, vid varje enskild informationsöverföring</i> B --> A Saknad information kan ibland efterfrågas, kompletteras och omtolkas manuellt av medicinskt kunnig person. <i>Algoritmiskt löslbart!</i>	Opereras senast: 2018-01-30 Preliminär operationstid: 2018-01-20 15:30 Huvuddiagnos*: 323291000119108 osteoartrit i vänster höftled Övriga diagnoser*: 25343008 sekundär lokaliseradosteoartros i bäckenregion 299308007 smärta i höftled vid rörelse Åtgärd*: 33788003 insättning av total protes eller protetisk utrustning i höft med metylmetakrylat Operationstyp**: Lubinus SP II med klack Önskad anestesi*: 18946005 epiduralanestesi NEWS2-score vid inskrivning: 1 Anestesibedömning: - Kondition: klarar lättare fysisk träning - Hjärta/kärl: u.a. (u.a. = utan anmärkning) - Lungor: u.a. - Svalg: u.a. - Mag/tarm*: 162030005 halsbränna med sur eller vattnig uppstötning	Operationsdatum: 2018-01-20 Diagnoskod: M167; Annan sekundär koxartros Operationskod: NFB49; Primär total höftledsplastik med cement Anestesikod: ZXH50; Epiduralanestesi ASA-klass: 1 – Frisk patient *) Koder från SNOMED CT **) en särskild typ av höftledsplastik med cement
Typ III Omöjligt att veta rätt omtolkning om data t.ex. aggregerats (förvanskats) på olika sätt.	Antal rökta cigaretter per vecka: 6-10 ...angivna i ett system med alternativen: 0, 1-5, 6-10, 11-15, 16-30, 31-50, 51-100, 101+	Antal rökta cigaretter per vecka: ? ...angivna i ett system med alternativen: 0, 1-3, 4-7, 8-14, 15-28, 29-69, 70+

I nuläget i Sverige används ett antal oförenliga/inkompatibla/olikartade informationsstrukturer och format, vilket försvårar både primär och sekundär användning av information. Data kan vara låst till det system där den produceras och åtkomst kan kräva kostsamma integrationer. Om information i stället skulle definieras och struktureras på ett likartat sätt nationellt och internationellt samt på ett sätt som fungerar bra både för journalföring och sekundäranvändning (t.ex. uppföljning och forskning), så skulle den inte behöva omtolkas eller mappas om.

Ett mål för standardiseringsarbetet i GMS är att bit för bit eliminera skillnader i dataformat (geografiskt och mellan system) eller åtminstone minska dem så att de blir av typ I och kan konverteras automatiskt så att manuell omtolkning inte ska behövas för varje överföring/patientfall.

I denna rapport beskrivs främst *data*-standardisering, Hur det hänger ihop med standardisering av *arbetssätt och processer* kan vara lite komplext. Vissa delprocesser (kemiska, bioinformatiska etc.) kan behöva vara väldigt lika för att resultaten ska bli tillräckligt jämförbara, medan andra delprocesser kan göras lite olika eller i olika ordning utan att förstöra jämförbarheten i den slutliga datan. Å andra sidan, om man faktiskt använder samma process men dokumenterar/lagrar resultat på olika sätt (t.ex. olika filformat, struktur och variabelnamn) så kan data ändå inte jämföras och delas smidigt och automatiskt.

1.3. Mål för delprojektet Skalbara informatiklösningar

Delprojektet Skalbara informatiklösningar ska som första steg standardisera de informationsmängder som är vanligast förekommande inom sekvensering, bioinformatiska analyser samt ytterligare tolkningar som utförs på patientprover. I första hand vill vi hänvisa till hur befintliga (helst internationella) standarder kan tillämpas i de olika användningsfallen.

Ur projektbeställningen för delprojektet Skalbara informatiklösningar:

- utveckla nationella specifikationer för lagring av genomikdata inom svensk sjukvård
- ramverk för länkning av genomikdata och metadata till andra hälsodata för analys och visualiseringsändamål
- belysa de lagändringar som måste ske för att möjliggöra säker datadelning på nationell och internationell nivå

Nationell standardisering av genomisk metadata är nödvändig för vidare visualisering av data för sjukvård, akademi, näringsliv och myndigheter samt för utveckling av avancerade realtidsanalyser såsom AI-baserade applikationer och maskininlärning.

Viktiga krav/mål är:

- att genomikdata associeras med en stor mängd metadata i form av annan patient- och provinformation
- att kunna länka genomikdata till övriga datakällor inom sjukvården (journaldata, biobanksdata, kvalitetsregister m.m.) på ett strukturerat och systematiskt, samt för individen säkert och integritetsbevarande sätt, det vill säga inom ramen för GDPR, PDL etc.
- en väl fungerande struktur för att organisera, aggregera och länka data
- kunna samköra och aggregera data från ett stort antal källor (inklusive genomikdata), centrerade runt enskilda patienter eller hela patientgrupper för att avgöra behandlingars effektivitet och tillförlitlighet, information som är central för såväl sjukvården, akademien, myndigheter och näringslivet.
- länka ihop/samverka och nyttja den kunskap och erfarenhet som genererats genom dessa projekt (som pågår i Sverige kring data och precisionsmedicin)

1.4. Avgränsningar

Arbetet ska fokusera på genomikinformation, det vill säga sådan information som genereras vid breda genanalyser där NGS-metoder används (t.ex. breda genpaneler samt helgenom- och helexomsekvensering). Vi har därför inte tagit hänsyn till genetisk information som genereras vid "mindre breda" genanalyser såsom t.ex. fluoroscent in situ hybridisering (FISH)- och arrayanalyser.

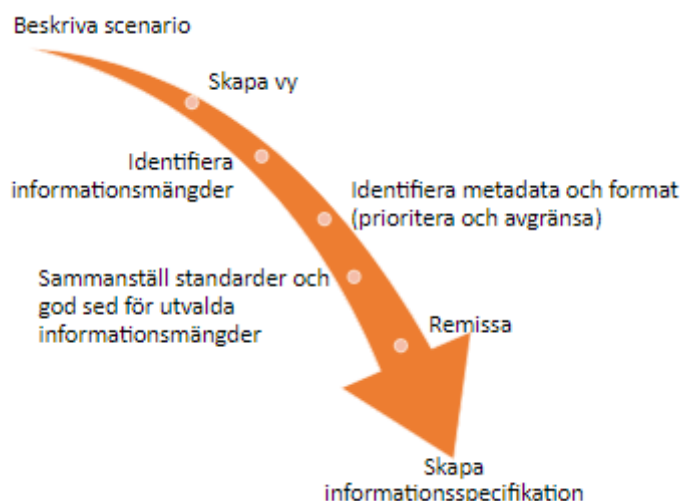
Arbetet ska grundas på att NGP används i nationell kontext. Vi har därför förutsatt att genomikinformation ska bearbetas samt lagras på NGP och att det därför är speciellt viktigt att den informationen standardiseras.

Projektet ska samarbeta med andra grupper inom GMS där aktiviteter pågår och identifiera standardiseringsbehov som är relevanta för dem. Vid projektet har aktiviteter pågått med utveckling av bioinformatikflöden för barncancer, hematologi, infektionssjukdomar, solida tumörer och sällsynta diagnoser. Dessutom har aktiviteter avseende NGP, nationellt tolkningsverktyg och nationella variantdatabaser pågått. Vi har avgränsat arbetet till ovan nämnda sjukdomsområden och aktiviteter. För sjukdomsområdena beskrivs flödena från remiss till det att remissvar skickas tillbaka till beställaren samt att relevant information har lagrats i NGP.

1.5. Angreppssätt och genomförande

1.5.1. Angreppssätt behovskartläggning

Nedan beskrivs övergripande den metod som har använts i samband med kartläggningen av behov av standardisering inom de olika områdena.



Figur 3. Steg i processen med behovskartläggning

Beskriva scenario

Sammanfatta en beskrivning av området och övergripande behov, med fokus för det som är specifikt för just detta område.

Vad behövs nu?

Vad behövs om tre år?

Skapa vy

Skapa en översiktlig vy (diagram) över flödet med aktiviteter och identifierade informationsmängder. Notationen som använts för detta är en internationell standard för att beskriva processer: Business Process Model Notation (BPMN) 2.0³. Den valdes för att på ett övergripande vis kunna visualisera kopplingen mellan de aktiviteter som genomförs och de informationsmängder som skapas eller används under förloppet.

³ Business Processing Modelling Notation: <https://www.bpmn.org/>

Identifiera informationsmängder

Skapa en tydlig bild över vilka informationsmängder som skapas/används under arbetets gång. Specificera vilka av dessa informationsmängder som är aktuella för standardisering utifrån upprepat nyttjande i flödet och/eller lagring i NGP.

Identifiera metadata och format (prioritera och avgränsa)

I detta steg specificeras vilken metadata för respektive informationsmängd som behöver finnas mellan flera aktiviteter i flödet och därmed är viktiga att standardisera.

Specificera även format som behöver finnas mellan flera aktiviteter i flödet och därmed är viktiga att standardisera.

Sammanställ standarder och god sed för utvalda informationsmängder

I många sammanhang finns det redan ett stort antal standarder att förhålla sig till, t.ex. inom bioinformatikflödet. I andra fall finns det olika standarder som används på olika laboratorier och sjukhus.

Att sammanställa relevanta delar av dessa standarder och rekommendera vilka som bör användas.

Remissa

Genomförs iterativt vid flera tillfällen. Remissinstanserna kan vara t.ex. representanter från GMS arbetsutskott för informatik respektive arbetsutskott för kliniska grupperingar, co-chairs samt kollegor som arbetar med angränsande områden.

Skapa informationsspecifikation

Sammanställ och specificera information så att den blir användbar för rapportens intressenter, för såväl de som arbetar med informationen som de som bygger upp infrastrukturen och integrationer.

1.5.2. Genomförande av behovskartläggning

Ovanstående angreppssätt har i hög grad använts vid genomförandet av arbetet, men i en del fall har det varit svårt att få tag på kontaktpersoner.

Vid genomförandet har standardiseringsgruppen utfört intervjuer digitalt via Microsoft Teams eller Zoom (se [Bilaga B. Deltagarlista](#)). En del uppföljande frågor har ställts via e-post. Vi har även tagit del av information om GMS som finns på GMS Teams arbetsyta respektive GMS hemsida⁴ samt inhämtat information från Internet om olika genomiksamarbeten och standarder.

En remissrunda har genomförts (se [Bilaga B. Deltagarlista](#)), men de flesta deltagare har inte fått möjlighet att granska rapporten och lämna förtydliganden. Resultatet av intervjun för infektionssjukdomar har inte inkluderats i denna rapport, men ett utkast kan lämnas till intressenter inom GMS på begäran.

1.6. Intressenter

Rapporten har beställts av GMS och riktar sig till intressenter inom GMS som arbetar för att införa och utveckla precisionsmedicin. Tabellen nedan beskriver de intressenter som rapporten primärt vänder sig till och hur de berörs av standardisering inom genomikområdet.

⁴ GMS hemsida: <https://genomicmedicine.se/>

Tabell 2. Primära intressenter som rapporten riktar sig till samt beskrivning av hur de berörs.

Intressent	Beskrivning
IT-personal	<p>Sätter upp och förvaltar infrastruktur för NGP samt installerar bioinformatiska analysflöden.</p> <p>Intresserade av att de gränssytor som interagerar med infrastrukturen sänder in information på ett enhetligt och entydigt sätt, att information som genereras vid bioinformatikanalyser lagras i enhetliga format och med entydig metadata som kan användas för indexering, att lagrad information kan kopplas till/överföras mellan andra IT-mjukvaror som innehåller hälsoinformation (både nationellt och internationellt) genom att använda standarder samt att lagrad information kan tillgängliggöras för forskning genom att använda standarder.</p>
Bioinformatiker	<p>Utvecklar bioinformatiska analysflöden och utför dataanalys.</p> <p>Intresserade av att enhetliga format och entydig metadata används vid bioinformatiska analyser (åtminstone för information som ska användas nationellt eller internationellt).</p>
Sjukhusgenetiker	<p>Tolkar varianters betydelse för sjukdom och levererar svar från genetiska analyser.</p> <p>Intresserade av att enhetliga format, entydig metadata och nationella/internationella standarder används vid delning av genomikdata med nationella och internationella initiativ för att kunna hitta matchande patientfall samt leverera korrekta svar.</p>
Läkare	<p>Olika typer av läkare är involverade i ett NGS-flöde och samarbetar i multidisciplinära konferenser eller specialistteam. De subgrupperar patienter, rapporterar till kvalitetsregister, ställer diagnos samt bestämmer prognos, behandling och uppföljning utifrån svaret från den genetiska analysen.</p> <p>Intresserade av att entydig metadata och standarder används då olika IT-mjukvaror utbyter information så att svarshantering och rapportering till register förenklas.</p>
Forskare	<p>Genomför forskningsstudier där kompetenser av olika slag är involverade, t.ex. en forskande läkare, forskande bioinformatiker, forskningssjuksköterskor etc.</p> <p>Intresserade av att få tillgång till genomikinformation för att studera patientfall där vården konstaterat att varianter har oklar signifikans, hitta nya sjukdomsorsakande varianter samt inkludera patienter att delta i studier avseende individanpassade behandlingar. För detta behöver metadata vara entydig och överföring av information ska vara standardiserad.</p>

Det finns fler intressenter som i en förlängning skulle vara aktuella att påverkas av standardisering av genomikdata, men rapporten är inte riktad till dem i första hand. Här följer ett urval:

- patient/person (förväntas i framtiden vilja få ut mer genomikinformation från vården än vad som sker idag)
- beslutsfattare (vill att hälsoekonomiska rapporter tas fram)
- gränssytor såsom kvalitetsregister, internationella nätverk och läkemedelsföretag
- personer som arbetar för att införa standarder på ett mer övergripande plan, t.ex. standardiserade patientjournalssystem och standardiserade svar till inremitterande läkare
- intressenter i de sjukdomsområden inom GMS som inte har inkluderats i denna rapport, det vill säga farmakogenomik, infektionssjukdomar och komplexa sjukdomar

- de som skapar och använder (framtida) kliniska beslutsstöd som delvis kan basera sig på den enskilda patientens gener (som kanske har sekvenserats tidigare i någon annan klinisk process)

1.7. Rekommendationer

Standardiseringsgruppen har inte mandat att bestämma vilka krav som ska gälla avseende standardisering för NGP. Vi lämnar därför istället rekommendationer. För att förtydliga vilka rekommendationer som standardiseringsgruppen har funnit viktigare än andra att uppfylla, används begreppen MÅSTE och BÖR i vissa fall. Innebörden av dessa begrepp beskrivs nedan.

MÅSTE: Obligatoriskt att uppfylla.

BÖR: Rekommenderat men ej obligatoriskt att uppfylla.

2. Kartläggning av behov som är gemensamma för flera sjukdomsområden

Detta kapitel sammanfattar vilka behov av standardisering som är gemensamma för flera av GMS sjukdomsområden: barncancer, hematologi, solida tumörer och sällsynta diagnoser. Kapitlet beskriver vilka aktiviteter som genomförs, de informationsmängder som skapas eller används då aktiviteterna utförs samt de informationsbärare som utgör lagringsplatser för informationsmängder eller som via API:er utbyter informationsmängder. Flödet illustreras även i en vy. Motsvarande avsnitt finns sedan i kapitlen för respektive sjukdomsområde med mer områdesspecifika detaljer.

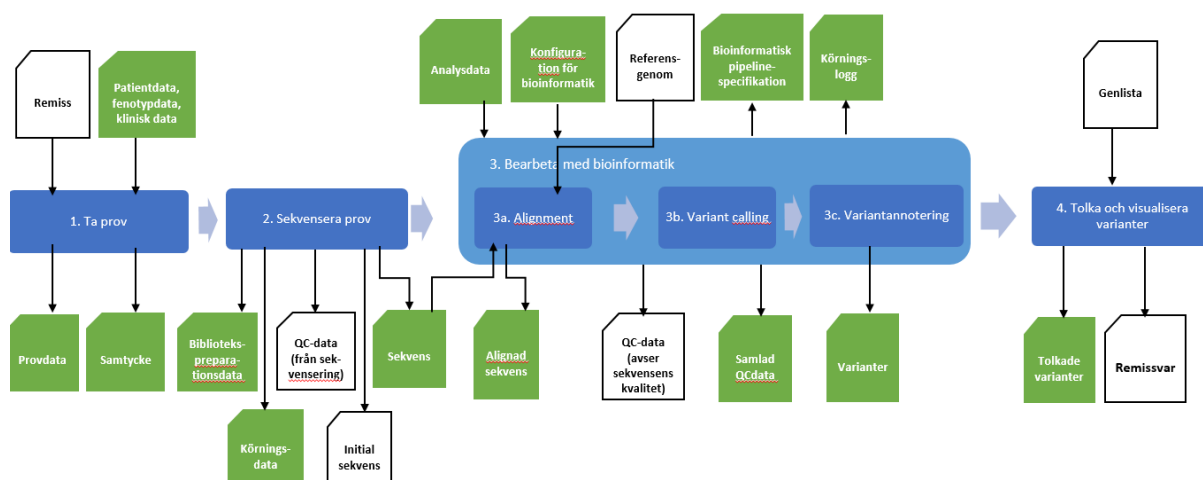
De aktiviteter som utförs är generella och utförs oavsett vilket sjukdomsområde som flödet avser. För att kunna kartlägga och visa på information som kan komma att behöva standardiseras så underlättar det att beskriva den kontext av aktiviteter som utförs oavsett vilket sjukdomsområde som avses.

2.1. Vy

I figuren nedan ges en visuell översikt (diagram) över de aktiviteter och informationsmängder som är involverade i ett generiskt flöde. (Informationsbärare illustreras inte i vyn.) För en beskrivning av vad symbolerna i vyerna representerar, se [Bilaga C Teckenförklaring](#).

Flödet startar med att det inkommer en remiss avseende analys av prov med hjälp av en NGS-metod och slutar med att analys har genomförts, resultat har visualiserats samt att information har lagrats. I kliniska flöden innebär det även att remissvar har levererats till remitterande läkare.

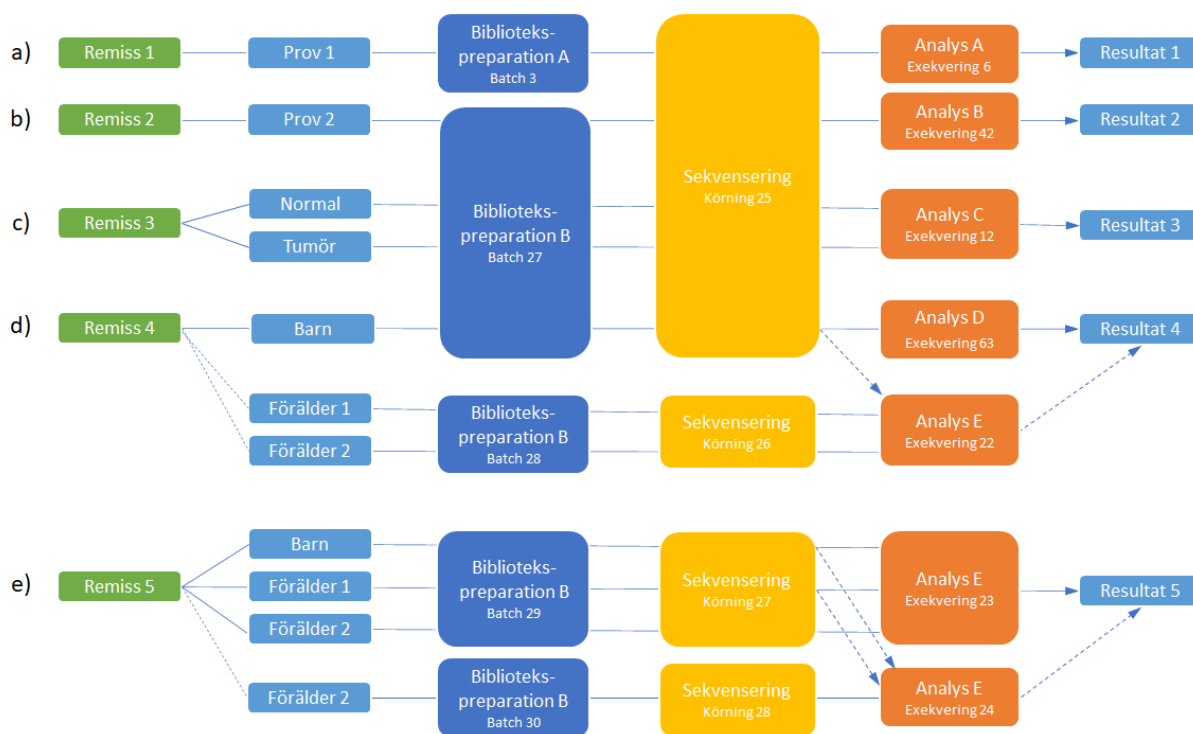
De prioriterade (i figuren grönfärgade) *informationsmängder* som bör lagras i NGP (eller som på annat sätt bör finnas lättillgängliga för NGP-relaterad användning) är (del)resultat av olika processteg eller data som behövs om man senare vill kunna utvärdera resultaten eller köra om bioinformatikflödet. I intervjuerna har det framkommit att även viss data som kan återgenereras kan vara så tidskrävande att återgenerera att den är rimlig att ändå välja att lagra. Även viss data från före och efter bioinformatikflödet är intressant att spara då den tillför en viktig kontext till den bioinformatiska datan i form av exempelvis fenotypdata, provdata och tolkade varianter. De prioriterade informationsmängder som (helt eller delvis) kan behöva lagras i NGP beskrivs detaljerat i avsnittet [Informationsmängder](#).



Figur 4. Förenklad vy avseende generella aktiviteter (i blått) samt inkommande respektive resulterande informationsmängder, som förekommer i flödet för flera av de olika sjukdomsområdena. (Grönt = prioriterat för lagring i NGP)

Som framgår av figuren nedan så kan viss data, metadata och loggar från olika steg vara gemensam för flera olika prov och/eller resultat och sambanden bli lite trassliga. I de fallen är det onödigt att dubbellagra gemensam data tillsammans med varje prov eller resultat, däremot bör varje resultat innehålla referenser (t.ex. identifierare) som kan följas för att hitta filerna.

Exempel: Konfigurations-, QC- och loggfiler för exekvering 12 av analys C i figuren lagras en gång gemensamt och resultaten ska då innehålla referenser till dessa - eller referenser till en unik identifierare för körningen om det räcker för att nå filernas innehåll.



Figur 5. Förenklad vy för provanalys genom sekvensering. Stordriftsfördelar uppnås ofta genom att för flera prov samordna bibliotekspreparationer, sekvenseringskörningar och bioinformatiska analysflöden.

a) Sekvensering av ett prov som har preparerats med en annan biblioteksmetod för sekvensering jämfört med resten av de prover som ingår i samma sekvenseringskörning.
b) Sekvensering av ett prov som har preparerats i ett bibliotek som är gemensamt för många prover.

c) Sekvensering av parade normal- och tumörprover, vilka jämförs i analysen.

d) Sekvensering av ett prov från ett barn samt sekvensering av prover från barnets biologiska föräldrar i en annan sekvenseringskörning vid en senare tidpunkt (se streckade linjer)

e) Sekvensering av prover från ett barn samt biologiska föräldrar (trioanalys) vid samma tidpunkt, inklusive ett exempel på resekvensering av ett prov från en av föräldrarna (se streckade linjer).

För tydlighets skull har processen som presenteras i figuren förenklats. I realiteten finns det många aspekter som kan påverka flödet:

i) relationen mellan en remiss och ett prov beror på laboratorieinformationssystemet samt det kan dessutom finnas ett eller flera prover i samma remiss, men ett prov kan också vara kopplat till många remisser

- ii) *pre-analys* som skapar DNA för bibliotekspreparation visas inte i figuren men kan skilja sig åt mellan olika prover
- iii) prover kan tas från olika vävnader (vilket påverkar *pre-analys*) eller kan vara subprover till samma prov
- iv) En analys kan representera ett eller flera arbetsflöden som behöver exekveras för att skapa information (rutan Resultat 5) som behövs för tolkningen

2.2. Aktiviteter

I detta avsnitt beskrivs kortfattat de aktiviteter som genomförs i ett generiskt flöde.

2.2.1. Ta prov

Biologiskt material, exempelvis ett blodprov, inhämtas från en person (s.k. humandata).

Inom sjukvården tas prov från patient i enlighet med vårdens rutiner. Data om provtagningen och provhanteringen lagras i system för laboratorieinformation (LIS/LIMS), remiss och remissvar, biobank, patientjournal m.m. Vid kliniska studier återanvänds ofta provinformation som inhämtas inom sjukvårdens regi (s.k. sekundär användning av informationen).

2.2.2. Sekvensera prov

Idag finns det flera sekvenseringsteknologier. Laboratorierna kan använda olika teknologier och sekvenseringsmaskiner från olika leverantörer såsom Illumina, Thermo-Fisher, PacBio och Oxford Nanopore. Innan sekvensering förbereds provet på olika sätt, beroende på sekvenseringsteknologi och tillämpning.

Först renas DNA eller RNA fram från provmaterialet. Om provmaterialet är RNA så omvandlas RNA till DNA innan sekvensering med NGS-metoder. Sedan prepareras ett *sekvenseringsbibliotek* av DNA-fragment så att de kan sekvenseras. Processen skiljer sig åt beroende på vilken sekvenseringsmaskin som används, men ofta klipps det extraherade materialet isär till fragment av önskad storlek och extrasekvenser (adaptrar) läggs till för att DNA-fragment ska kunna fästa till flödescellen i sekvenseringsmaskinen. Om flera prover sekvenseras samtidigt kan ytterligare DNA av känd sekvens som kallar *barcodes* fästas till DNA-fragmenten för att märka dem från vilken prov de kommer ifrån. I detta steg används olika typer av *kit* (*library preparation kit* med flera), vilkas namn, version och batchnummer behöver sparas för eventuell senare referens, exempelvis för att underlätta felsökning eller kvalitetsbedömning.

Under sekvenseringen av sekvenseringsbibliotek läser sekvenseringsmaskinen av de olika kvävebaserna i DNA- eller RNA-molekyl och råsignal omvandlas till digitala sekvenser, s.k. *reads*, det steget kallas *basecalling*.

2.2.3. Bearbeta sekvens med bioinformatik

Bioinformatik är avancerad dataanalys av genetikdata. Datan från sekvenseringen bearbetas med ett antal mjukvaror som ska köras i en specifik ordning. Resultatet från en mjukvara ska kunna användas som indata till en annan mjukvara (t.ex. nästa steg av processen). Ordningen för hur mjukvarorna ska köras och vilka filer som ska läsas och skrivas, konfigureras i ett *bioinformatiskt analysflöde*, även kallat *pipeline* eller *workflow*.

De vanligaste aktiviteterna i ett bioinformatikflöde är *alignment*, *variant calling* och *variantannotering* (se förklaringar i separata avsnitt nedan). I verkligheten kan det dock vara flera parallella processflöden som kombineras, se exempel för [sällsynta diagnoser](#)⁵.

I vissa bioinformatikflöden görs jämförelser mellan flera olika prover (t.ex. från olika familjemedlemmar eller mellan normal- och tumörprover) som en del av flödet.

2.2.3.1. Alignment

Sekvensinformation (reads) matchas mot ett referensgenom från en person, eller mot ett standardiserat "medelgenom", t.ex. GRCh38 som är baserat på många personer. Bearbetningar med olika verktyg kan vara involverade, men den slutgiltiga alignment-filen är normalt i BAM-format (binary alignment map). Resultatet kan vara att baser i olika reads har inpassats mot samma position i referensgenomet. Dessa ska sedan kontrolleras noggrannare vid variant calling. Viss filtrering kan ske, t.ex. filtreras duplicerade reads bort.

2.2.3.2. Variant calling

Variant calling innebär att mjukvaror (variant callers) identifierar baser i provets sekvensinformation som skiljer sig från ett referensgenom. De varianter som upptäcks skrivs till en VCF-fil (variant call format). Mjukvarorna är anpassade för att upptäcka en specifik typ av genetisk förändring, t.ex. är några mjukvaror bra på att identifiera "små" varianter, medan andra är bättre på att detektera strukturella varianter, biomarkörer etc. I variant calling ingår vanligtvis även kvalitetskontroll.

Viss filtrering utförs också. Tekniska fel vid sekvenseringen (artefakter) kan medföra att sekvenser felaktigt påvisas som varianter vid variant calling och de behöver filtreras ut från processen.

2.2.3.3. Variantannotering

Bioinformatisk annotering är processen att lägga till extra information till råa (obearbetade) variant calls, t.ex. information om allelfrekvenser i befolkningen, involvering i sjukdomar och fenotyper, prediktion om en variant påverkar proteinets funktion, zygositet, gensymbol, varianters placering i genomet (icke-kodande, kodande, splitsningsregioner) och funktionella annoteringar om hur proteinsyntesen påverkas (synonymous, missense, frameshift). Annoteringarna läggs till per automatik av olika annoteringsmjukvaror och mjukvarorna kontaktar internationella databaser för att kontrollera olika uppgifter, t.ex. hur sällsynt en variant är i en viss befolkningsgrupp.

Annoteringarna lagras ofta i INFO-kolumnen (ett key-value dictionary) i VCF-filen (se avsnittet [Varianter](#)).

⁵ Exempel på bioinformatikflöde med parallella processer: https://www.researchgate.net/figure/Schematic-illustration-of-the-different-components-in-our-current-bioinformatic-pipeline_fig3_350108801

2.2.4. Tolka och visualisera varianter

Klinisk tolkning utförs på kliniska laboratorier, ofta av en sjukhusgenetiker. Sjukhusgenetikern granskar de varianter och andra genetiska förändringar som är resultatet av bioinformatikflödet och bedömer om de har någon betydelse för en patients sjukdom och om de kan vägleda behandlingsbeslut. Till stöd för arbetet används ofta mjukvaror som kan visualisera information om varianter. Resultatet från tolkningen är en rapport, det vill säga ett svar från den genetiska utredningen.

En tolkning görs för varje enskild variant, en ytterligare analys kan göras på helheten. Tolkning och analys genomförs exempelvis med hjälp av mappning mot databaser.

Aktiviteten kan även beskrivas som en bedömning eller en klassificering.

2.3. Informationsmängder

I tabellen på nästa sida sammanfattas de informationsmängder (data, metadata etc.) som vi har funnit vara prioriterade för standardisering utifrån att de antingen nyttjas i flera steg av flödet och/eller att de så småningom ska komma att lagras i NGP och därför är viktiga att standardisera. Sammanställningen pekar på såväl befintliga standarder samt visar på behov av ytterligare standardisering av metadata och/eller format.

Efter tabellen följer avsnitt (med mer detaljer och referenser) om varje enskild informationsmängd som skapas eller används då aktiviteterna i flödet utförs.

Tabell 3. Viktiga informationsmängder. (Länkarna i den första kolumnen leder till motsvarande rapportavsnitt med mer information.)

Informations-mängd	Beskrivning och viktiga dataelement	Standarder nuläge	Viktiga metadata, relationer etc.	Lagras i NGP + föreslaget format
Patientdata samt fenotypdata och annan klinisk data	<p>Personlig information om en person. Observerbara egenskaper, symptom och tecken hos en person. Andra typer av klinisk data såsom diagnoser, histopatologisk undersökning av vävnader, röntgenbilder, fotografier, släkträd, behandlingsdata etc.</p> <p>Viktiga dataelement att standardisera och lagra i NGP (särskilt när de går att koda):</p> <ul style="list-style-type: none"> - frågeställning från remittent - kön - ålder - initial misstanke om vilken sjukdom patienten drabbats av - diagnos - behandlingsbeslut - behandlingsutfall - (vissa) fenotypegenskaper, särskilt de som är kodade och avsiktligt skickas med i samband med remiss. 	SNOMED CT, HPO, Phenopackets, nationella tjänstekontrakt, openEHR, HL7 FHIR m.m. men det mesta ligger idag i leverantörsspecifika format/databaser eller skickas med som datafält eller fritext i remisser.	Relationer mellan individer är viktiga att lagra ner t.ex. i samband med trioanalyser och andra släktskapsbaserade utredningar.	Dataelementen i kolumn 2 kan lagras i NGP delvis kodade med hjälp av HPO, SNOMED CT etc. inuti informations-specifikation Patientdata (openEHR JSON) - som bör tas fram i framtida arbete.
Provdata	<p>Information om provet, hur det har hanterats och samlats in (ursprung, teknik).</p> <p>Viktiga dataelement att standardisera och lagra:</p> <ul style="list-style-type: none"> - ID på provet - provtyp, t.ex. DNA eller RNA - provtagningsdatum/tidpunkt - hur provet har tagits - hur provet har förvarats/hanterats, - vävnad och mängd - tumörtyp - tumörcellhalt - information om att provet förvaras i biobank 	HL7 v2, nationella tjänstekontrakt, openEHR, HL7 FHIR.		Dataelementen i kolumn 2 kan lagras i NGP delvis kodade med hjälp av SNOMED CT etc. inuti informations-specifikation Provdata (openEHR JSON) - som bör tas fram i framtida arbete.
Samtycke	Ej utrett.		Att samtycke getts och <i>till vad</i> samt <i>var man kan hitta</i> samtyckesdokumentet. Bör sparas (strukturerat/indexerbart) i NGP.	Ej färdigutrett var själva samtyckesfilerna ska förvaras och i vilka format, se "Malläge" i avsnitt Samtycke
Genomiknod	Information som förtydligar vilken organisatorisk enhet som genomförde olika delar av flödet samt vem som äger datan.			Sparas för flera steg eftersom stegen kan utföras på olika platser:

	Viktiga dataelement att standardisera och lagra: - laboratorium som analyserade provet - kontaktinformation till dataägaren			- i informationsspecifikation Provdata - Sekvens (metadata) - Körningslogg (metadata) - i informationsspecifikation Resultatrapport
Bibliotekspreparationsdata	Information om hur nukleinsyra från provet har förberetts så att sekvensering ska kunna genomföras (bibliotekspreparation). Viktiga dataelement att standardisera och lagra: - adapter och index - kit - inputmängd och/eller inputkoncentrationer av DNA och RNA - measured insert size - intended insert size - sekvenseringsmaskin - läslängd			Informationsspecifikation Provdata (open-EHR JSON) - som bör tas fram i framtida arbete. En del av denna data bör eventuellt kopieras även till informationsspecifikation Resultatrapport om vi vill följa tyska HiGHmed:s exempel.
Sekvens	Genetisk information som samlas in vid processen att sekvensera DNA eller RNA. Information om nukleotider, med kvalitetsmått för varje läst kvävebas (FASTQ) eller utan kvalitetsmått (FASTA).	FASTA, FASTQ	Viktiga dataelement att standardisera och lagra: - version av FASTQ-standard. - namn på sekvenseringsmaskin - phred score encoding - vilken genomiknod som gjorde sekvenseringen Koppling till namn på provet (finns oftast i namnet på FASTA/FASTQ-filen), koppling till beskrivning av provet, information om kvalitet på datan (FASTQ-rapport), information om metoden som producerade sekvensen: bibliotekspreparation, typ av sekvensering.	FASTQ
Samlad QC-data	Teknisk kvalitet på sekvenser efter sekvensering.	FastQC och MultiQC. Utdata är HTML-format.	Provdata.	HTML från MultiQC.
Analysdata	Analysinformation som behövs för att utföra bioinformatisk analys efter sekvenseringen av provet. Viktiga dataelement att standardisera och lagra:	Provdata samlas ofta i CSV-filen (sample sheet). Typ av analys kan ses i laboratorieinformations-		

	- körningstyp (kan potentiellt fångas i ett annat skede än då analysdata genereras)	system eller beställas genom personlig kommunikation.		
Alignad sekvens	Sekvenser matchade till ett referens(genom), för att ge information om vilken del av referensen de kommer ifrån samt poäng (scores) som indikerar hur sannolikt det är att de kommer från den delen av referensen.	SAM, BAM, CRAM.	Referensgenom och dess version.	CRAM.
Varianter	Information om genetiska varianter, det vill säga de delar i sekvenserna som skiljer sig från referensen. Varianter kan antingen annoteras eller inte annoteras och annoteringen kan härröra från olika källor. Viktiga dataelement att standardisera och lagra: - anledningen till att reanalys gjordes - vilken typ av fil som var startpunkt för reanalys (t.ex. FASTQ, BAM, CRAM, initial VCF)	VCF.	Olika mjukvaror används för att prediktera och annotera varianterna, t.ex. med information om vilket referensgenom som har använts samt hur varianterna har filtrerats (filtrering utförd/filtrering ej utförd).	VCF och även VRS om möjligt.
Konfiguration för bioinformatik	Avser information om vilka inställningar och regler som används för en bioinformatikkörning och sparas för att den innehåller uppgifter som behövs för att felsöka eller upprepa en analys.	JSON eller YAML (fil med parametrar).	- namn/identifierare och versionsnummer för Bioinformatisk pipelinespecifikation - namn/identifierare och versionsnummer för den workflow-hanterare som kan köra konfigurationen - namn/identifierare och versionsnummer för den/de containrar som ska finnas tillgängliga	Sparas i befintligt format på NGP tillsammans med uppgifter om motsvarande workflow-hanterare, bioinformatisk pipelinespecifikation och containrar som används för analysen. Sökvägar till filer och hänvisningar till containrar och andra resurser i konfigurationsfilen hanteras på ett sätt så att körningen kan upprepas med samma resultat.
Bioinformatisk pipelinespecifikation	Instruktioner för vilka mjukvaror som ska köras i bioinformatikflödet samt vilka parametrar som ska användas (ändrade från default-värden eller nya). Viktiga dataelement att standardisera och lagra: - namn/identifierare och versionsnummer för de mjukvaror/containrar som används	.nf (nextflow script) med flera. Specifikationer för Nextflow eller Snakemake.	- namn/identifierare och versionsnummer för specifikationen - uppgifter om användardokumentation, supportnätverk och artiklar/metodbeskrivningar - uppgifter om processer för vidareutveckling och förvaltning av specifikationen - föreslagen citering för användning i rapporter och publikationer - licensinformation för specifikationen - namn/identifierare och	Specifikationer för Nextflow eller Snakemake.

			versionsnummer för den workflow-hanterare som kan köra specifikationen	
Körningslogg	<p>Logginformation om hur en bioinformatikkörnings olika steg exekverades, vad resultatet var samt andra detaljer. Den sparas för att den kan innehålla kompletterande uppgifter som behövs för att felsöka eller upprepa en analys.</p> <p>Viktiga dataelement att standardisera och lagra:</p> <ul style="list-style-type: none"> - namn på bioinformatikflöde - version på bioinformatikflöde - run command - program (plus versioner, indataparametrar, seeds) - checksummor, datumstämplar och/eller versionsnummer för databaser och datafiler som laddas under installation/körning - random seeds (det man sätter igång slumpgeneratorer med) ska också vara dokumenterade så att man verkligen kan reproducera körningen och därmed inte datan 	Logginformation sparas i en textfil.	Vilken genomiknod som exekverade bioinformatikflödet.	
Tolkade varianter	Varianter som är sannolika att påverka sjukdom eller behandling.	BAM, CRAM, VCF m.m. används i mjukvaror. Till remittent skickas dock ofta text eller PDF.	Verktyg och databaser som används i tolkningssteget samt namn på den person som tolkade varianter.	<p>Informationsspecifikation Resultatrapport (openEHR JSON), se påbörjat exempel i kapitlet Informationsspecifikation och i Bilaga D. Informationsspecifikation - Formulärexempel)</p> <p>+ JSON eller en VCF-fil med ytterligare information (anmärkingar) från tolkning.</p>

Specifikationer för vissa standardformat (SAM, BAM, CRAM, VCF, BED) beskrivs på Samtools webbsida⁶.

⁶ Samtools webbsida: <https://samtools.github.io/hts-specs/>

2.3.1. Remiss

Remiss är en handling som utgör en beställning av en tjänst inom sjukvården. Prover som ska analyseras av laboratorium beställs via en remiss. (Vid forskningsstudier är en remiss inte alltid involverad, men för kliniska studier behöver forskningspersoner ofta rekryteras från vården eller information om patienter återanvändas.)

Nuläge

Information från remiss fångas inte alltid på ett strukturerat sätt.

Målbild

Remiss inom respektive sjukdomsområde bör följa en struktur som kan återanvändas i efterföljande aktivitet så att de fäkt från remissen som behövs vid senare forskning och samarbeten kan lagras strukturerat i NGP. Det vore värdefullt att dela på struktureringsarbetet och tillsammans standardisera mycket av remissunderlag.

Vid remisstillfället efterfrågas ofta en hel del bakgrundsinformation (se exempel i avsnittet [Patientdata samt fenotypdata och annan klinisk data](#)). För att skapa goda förutsättningar för tolkningsaktiviteten (se avsnittet [Tolka och visualisera varianter](#)) samt lagring i NGP bör denna information fångas med hjälp av strukturerade formulär.

Utöver patient- eller fenotypdata efterfrågas ofta även sammanfattande journalutdrag, foton, tillväxtkurvor, släktanamnes m.m. (se exempel från [Uppsala](#)⁷ och [Västra Götalandsregionen](#)⁸). Det finns risk för att mycket av denna extra information om patienten inte får eller bör lagras i NGP och att den dessutom är väldigt olika strukturerad i källsystemen. Om man ändå vill fånga informationen på ett flexibelt men ändå någorlunda standardiserat sätt vid remisstillfället, t.ex. för att underlätta inför tolkning, så finns det i flera standarder sätt att skicka med dokument eller "citater" ur länkar till olika källor, exempelvis i openEHR⁹ och FHIR¹⁰. (Valda delar av denna information kan givetvis också struktureras och standardiseras om tydliga behov finns.)

2.3.2. Patientdata samt fenotypdata och annan klinisk data

Patientdata är personlig information om en person. Fenotypdata kan betraktas som en typ av klinisk data som avser observerbara egenskaper, symptom och tecken hos en person. Annan typ av klinisk data är diagnoser, histopatologisk undersökning av vävnader, röntgenbilder, fotografier, släkträd, behandlingsdata etc. Ibland benämns även vissa provsvar från andra discipliner som fenotyp- eller klinisk data, t.ex. provsvar från Klinisk Kemi.

Nuläge

En stor del av den kliniska datan registreras i patientjournalssystem. En del kommer in till bioinformatikflödet i form av fritext via remissen och genom att personal själv hämtar in data från patient och remittent samt genom manuell läsning av patientens journal etc. Vissa regioner använder strukturerade remissmallar som den i bilden nedan (källa [Region Skåne](#)¹¹). Sådana strukturerade mallar är innehållsmässigt intressanta och utgör en grund som GMS kan bygga vidare på och standardisera. Tekniskt sett är de dock idag ofta i form av PDF och därmed sällan integrerade med efterföljande flöde på ett automatiskt sätt.

⁷ Exempel på patient- eller fenotypdata (Uppsala):

<https://publikdocplus.regionuppsala.se/Home/GetDocument?containerName=e0c73411-be4b-4fee-ac09-640f9e2c5d83&reference=DocPlusSTYR-24556&docId=DocPlusSTYR-24556&filename=Information%20till%20inremitterande%20ang%C3%A5ende%20helexomsekvansanalys.pdf>

⁸ Exempel på patient- eller fenotypdata (Västra Götalandsregionen): <https://www.sahlgrenska.se/for-dig-som-ar/vardgivare/laboratoriemedicin/analyslistan/helexom/>

⁹ "Citation"-arketyp: <https://ckm.openehr.org/ckm/archetypes/1013.1.721>

¹⁰ "DocumentReference"-resurs: <https://build.fhir.org/documentreference.html>

¹¹ Strukturerad remissmall (Region Skåne): <https://vardgivare.skane.se/siteassets/1.-vardrictlinjer/laboratoriemedicin/remisser/genetik---helgenomsekvansering-wgs---konstitutionell-utredning>

Remiss vid genomisk utredning med helgenomsekvensering (WGS)

Namn: _____ Personnummer: _____

Symptomatologi/dysmorfologi	Ja	Ge förtydligande beskrivning	Nej	?
Avvikelse från tillväxtkurva	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
CNS avvikelser på MRT/DT	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Utvecklingsstörning HP:0001249, HP:0001263	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Mild	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Måttlig	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Grav	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Epilepsi HP:0001250/Myoklonier HP:0001336	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Muskelsvaghet/hypotoni HP:0001252	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Muskelhypertoni HP:0001276	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Perifer neuropati HP:0009830	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Andningsinsufficiens HP:0002093	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Ataxi HP:0001251/Spasticitet HP:0001257	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Artrogrypos HP:0002804/Dystoni HP:0001332	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Autism HP:0000717	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Autismspektrumstörning HP:0000729	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
ADHD HP:0007018	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Tvångssyndrom OCD HP:0000722	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Trotssyndrom ODD HP:0010865	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Inlärningsproblematik HP:0001328	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Finmotorisk försening HP:0010862	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Grovmotorisk försening HP:0002194	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Försenad språkutveckling HP:0000750	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Hörselnedsättning HP:0000365	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Synnedsättning HP:0000505	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Kraniell avvikelse HP:0000929	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Skelettavvikelser HP:0000924	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Avvikande händer, fingrar HP:0001155	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Avvikande fötter, tår HP:0001760	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Avvikande ögon HP:0000478	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Avvikande öron HP:0000598	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>
Avvikande näsa HP:0005105	<input type="radio"/>		<input type="radio"/>	<input type="radio"/>

Figur 6. Exempel på remiss vid genomisk utredning med helgenomsekvensering

Standardiserade datastrukturer och ontologier/terminologisystem används idag i varierande omfattning i källsystemen men användningen kan komma att utökas i takt med att journalsystem etc. moderniseras och standardiseras.

För att avgöra vilken behandling som ska användas behöver läkare ibland höra hur andra läkare valt att behandla och vilket utfallet av behandlingen var. De får då ringa och fråga. Vid kliniska studier behöver man också ibland ta kontakt med behandlande läkare. Information om behandling och utfall av behandlingen borde lagras i ett system som många kan komma åt. Det finns alltså en önskan om att koppla samman genetisk information med information om behandling och utfall.

Målbild

Vid informationsöverföring behöver terminologi/ontologi vara standardiserad. För fenotypdata ska HPO¹² användas (i figuren ovan syns koder från HPO, t.ex. HP:0001336). För annan klinisk data är koder från SNOMED CT¹³, OMIM¹⁴, ORPHAcodes¹⁵ och ICD¹⁶ intressanta.

¹² Human Phenotype Ontology (HPO): <https://hpo.jax.org/app/>

¹³ SNOMED Clinical Terminology (SNOMED CT): <https://www.snomed.org/>

¹⁴ Online Catalog of Human Genes and Genetic Disorders (OMIM): <https://www.omim.org/>

¹⁵ ORPHAcodes: <http://www.rd-code.eu/introduction/>

¹⁶ International Classification of Diseases (ICD): <https://www.who.int/standards/classifications/classification-of-diseases>

Även datastrukturerna (t.ex. liknande formuläret i bilden) som dessa koder stoppas in i behöver standardiseras (inklusive svarsalternativen) om vi ska få ett sammanhängande mer automatiserat flöde med minskad dubbeldokumentation.

För kommunikation med laboratorieinformationssystem och journalsystem är datastrukturer baserade på bland annat openEHR och HL7 (HL7 FHIR, HL7 v2 m.m.) vanliga. För utbyte av fenotypdata i forskning används t.ex. Phenopackets¹⁷ som datastruktur, men openEHR och HL7 innehåller många företeelser och detaljer som inte Phenopackets täcker. Lyckligtvis har det redan gjorts internationellt arbete för att representera Phenopackets-data i openEHR respektive i FHIR på ett sätt som underlättar kombinationer med övriga system och data. Inom projektet experimenterade vi litet med detta i juni 2021 och tror att de flesta behov för delning och lagring av strukturerad fenotypdata för sällsynta diagnoser inom GMS kan täckas med openEHR och FHIR (för detaljer se Github-biblioteket för [sällsynta diagnoser](https://github.com/modellbibliotek/GMS_informatics_tests)¹⁸). Ett API för att publicera lämpliga delar av denna data även enligt Phenopackets egen specifikation skulle lämpligen också kopplas till en databasvy av innehållet i NGP.

Många provsvar från andra discipliner, t.ex. från Klinisk Kemi, är redan strukturerade på sätt som relativt lätt kan översättas till format som lämpar sig att lagra i NGP.

Viktig metadata från remiss

- frågeställning från remittent

Viktig metadata om patient

- kön
- ålder
- relationer mellan individer

Relationer mellan individer: Relationer är viktiga att lagra ner t.ex. i samband med trioanalyser och andra släktskapsbaserade utredningar. (Vid experimentet i juni 2021 provade vi att använda identifierare och relationstyper från Phenopackets-standarden för att på ett pseudonymiserat sätt och i en openEHR-baserad lagringsstruktur kunna ange relationer mellan personer som kan ingå i flera separata släktskapsbaserade utredningar).

Viktig metadata om fenotyp

- fenotypegenskap

Viktig metadata om diagnos

- initial misstanke om vilken sjukdom patienten drabbats av
- diagnos

Viktig metadata om behandling

- behandlingsbeslut (hur läkare har valt att behandla en person tidigare)
- behandlingsutfall (vilket utfallet var på tidigare behandling)

Flera av ovanstående metadata hämtas ur journalsystem eller remiss. Många andra detaljer från journal etc. som ibland också kallas fenotypdata bör sannolikt (av bland annat legala skäl) stanna i vårdgivarnas egna datalager eller journalsystem och vid behov efterfrågas istället för att dubbellagras i NGP.

¹⁷ Phenopackets: <http://phenopackets.org/>

¹⁸ Github-bibliotek för sällsynta diagnoser: https://github.com/modellbibliotek/GMS_informatics_tests

2.3.3. Provdata

Provdata avser information om provet och hur det har hanterats.

Nuläge

I respektive laboratorieinformationssystem finns en del ostrukturerad och en del strukturerad data enligt respektive systems struktur.

Målbild

JSON-format som bygger på standarder som redan har strukturer för provdata, t.ex. openEHR¹⁹ och/eller HL7 FHIR.

Viktig metadata

- ID på provet
- provtyp, t.ex. DNA eller RNA
- provtagningsdatum/tidpunkt
- hur provet har tagits
- hur provet har förvarats/hanterats, t.ex. om det är färskt, fryst eller formalinbehandlat
- vävnad och mängd, t.ex. biopsi, cytologi, operationsmaterial, blodprov, benmärg
- tumörtyp
- tumörcellhalt
- information om att provet förvaras i biobank

ID på provet (providentifierare): Laboratorierna bygger upp providentifierare på sina egna sätt. De providentifierare som bioinformatiker på universitet får se är ofta avidentifierade och består av t.ex. årtal + provsekvensnummer (2021-157, 2021-158, 2021-159) och kompletteras ibland med system-ID. Det finns ingen garanti för att dagens lokala providentifierare är unika mellan laboratorier, t.ex. kan koden 2021-157 användas parallellt för olika provtagningar i olika regioner. I vissa laboratorieinformationssystem med begränsad längd på ID-fältet återanvänds dessutom gamla providentifierare flera gånger över den långa tid som GMS data kan förväntas vara relevant för forskning etc.

Information om provet förvaras i biobank: Det finns sätt att beskriva biobanksförvaring i modeller från openEHR²⁰ och HL7 FHIR²¹.

Metadata om ID på provet, provtagningsdatum, vävnad, tumörtyp och tumörcellhalt är viktiga för variantdatabaser.

2.3.4. Samtycke

Information om vad patienten har samtyckt till att provet får eller inte får användas för. Data avseende samtycke hanteras i samband med provtagningen.

Nuläge

I nuläget tyder allt på att GMS, på grund av legala begränsningar, i första hand kommer att samla in data när forskningsstudier som har godkända etikprövningsbeslut genomförs. I samband därmed behöver informerat samtycke signeras av patient och nära släktingar eller vårdnadshavare. Det är oklart vilken information om samtycke som behöver vidarebefordras till NGP.

¹⁹ Laboratoriemedicinska arketyper: <https://ckm.openehr.org/ckm/projects/1013.30.26> (generellt) och <https://ckm.openehr.org/ckm/projects/1013.30.50> (genomik)

²⁰ Arketyper för biobanksreferens <https://ckm.openehr.org/ckm/archetypes/1013.1.5885>

²¹ Tyskt nationellt projekt: <https://simplifier.net/medizininformatikinitiative-modulbiobank>

Målbild

Metadata om samtycket bör lagras standardiserat och om möjligt på maskinläsbara sätt. Det finns strukturer i openEHR²² och FHIR²³ för detta men även Global Alliance for Genomics and Health:s (GA4GH) arbete inom området som innefattar "Machine Readable Consent"²⁴ bör tittas närmare på för GMS del. Om vi t.ex. väljer att använda openEHR/FHIR-format bör vi se till att detta görs på ett sätt som smidigt kan automatöversättas till GA4GH:s format.

Viktig metadata

Vid intervju med bioinformatiker har inga behov avseende metadata identifierats, utan personer med klinisk expertis bör tillfrågas. (Potentiellt kan det röra sig om vilka ändamål personen samtycker till/inte samtycker till att provet får användas för samt information om en person har dragit tillbaka ett tidigare lämnat samtycke.)

2.3.5. Genomiknod

Information som förtydligar vilken organisatorisk enhet som analyserade provet samt vem som äger datan.

Nuläge

Oklart huruvida denna information fångas överhuvudtaget idag.

Målbild

Lagring: Informationen behöver lagras på NGP.

Format: JSON

Viktig metadata

- laboratorium som analyserade provet
- kontaktinformation till dataägaren

Metadata om laboratorium och kontaktinformation till dataägaren är intressant för variantdatabaser.

2.3.6. Bibliotekspreparationsdata

Information om hur nukleinsyra från provet har förberetts så att sekvensering ska kunna genomföras, s.k. bibliotekspreparation. Informationen sparas i form av ett laboratorieprotokoll.

Nuläge

Laboratorierna använder olika metoder för bibliotekspreparation, så det är många parametrar som varierar mellan dem.

Målbild

Information om preparationen ska fångas så att det i efterhand går att kontrollera detaljer om hur preparationen gjordes på ett specifikt laboratorium. Inom exempelvis barncancer använder vissa laboratorier kit som har förmåga att sekvensera väldigt små mängder av DNA, medan andra inte gör det. Sådana skillnader ska vara möjliga att kontrollera i efterhand.

Lagring: Information om bibliotekspreparation behöver lagras på NGP.

Format: JSON

²² Arketyper som kombineras: <https://ckm.openehr.org/ckm/archetypes/1013.1.1302> (informed consent request), <https://ckm.openehr.org/ckm/archetypes/1013.1.1303> (informed consent), <https://ckm.openehr.org/ckm/archetypes/1013.1.1307> (consent details)

²³ FHIR-resurs Consent: <https://www.hl7.org/fhir/consent.html>

²⁴ Machine Readable Consent m.m: <https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/>

Viktig metadata

- detaljer om hur provet konverterades till s.k. sekvenseringsbibliotek, t.ex. vilken adapter och vilka index som användes
- vilket kit användes
- inputmängd och/eller inputkoncentrationer av DNA och RNA
- genomsnittlig storlek på DNA-molekyler, det vill säga "measured insert size" eller "intended insert size", (denna information kan även tas fram under bioinformatikflödet och blir då mer exakt)
- typ av sekvenseringsmaskin
- läslängd, t.ex. paired-end 150 baspar

2.3.7. Körningsdata

Körningsdata avser själva utförandet av sekvenseringen av provet. Genereras efter det att sekvenseringskörningen har genomförts. Utgörs av en fil med metadata som innehåller information om själva körningen, inklusive det som är instrumentspecifikt för sekvenseringsmaskinen samt moment för förädling av provet inför sekvenseringen.

Nuläge

Format: JSON

Viktig metadata

- sekvenseringsmaskin
- datum
- körnings-ID
- kemi/batchchar/kit/panel
- batch på sekvenseringschip
- batch av assay
- vilken version m.m.

2.3.8. Initial sekvens

Information om de sekvenser av nukleotider som sekvenseringsmaskinen har identifierat. Den avser initial utdata som sekvenseringsmaskiner genererar, t.ex. bildfiler (image-data) i filformatet BCL för Illumina eller FAST5-filer för Nanopore.

Nuläge

Lagring: Bildfilerna från Illuminas första steg är stora och sparas sällan, möjligen en kortare tid på laboratoriet. De olika laboratorierna har olika lagringsregler och hittills har det varit billigare att analysera om provet än att lagra alltför mycket data på hårddiskar under lång tid. För andra typer av sekvenseringsmaskiner, t.ex. Nanopore, så är rådatan FAST5-filer och inte en bildfil. Mjukvaran för att göra om FAST5 till FASTQ förbättras fortfarande, därför kan det finnas värde i att spara FAST5 filer en längre tid.

Format: BCL (för Illumina), FAST5 (för Nanopore).

2.3.9. Sekvens

Information om de sekvenser av kvävebaser som sekvenseringsmaskinen har identifierat från nukleinsyrafragmenten (s.k. reads) samt tillhörande kvalitetsmått för varje bas.

Nuläge

FASTQ är ett väldigt välanvänt format för sekvensinformation och många bioinformatikmjukvaror som används senare i flödet kräver FASTQ som indataformat. De genereras från t.ex. BCL eller FAST5 (se avsnittet [Initial sekvens](#)). Det kan vara en eller flera FASTQ-filer per prov som går in i bioinformatikflödet. Om det finns flera FASTQ-filer per prov beror på om man har sekvenserat DNA-fragment från båda ändarna (paired-end-

sekvenseringsprotokoll för NGS) eller/och delat upp analysen i flera sekvenserings-“lanes” i en flödescell eller flera flödesceller i sekvenseringsmaskinen.

Format: FASTQ. Filformatet har fyra rader där den första är ett sequence ID, den andra är själva sekvensen, den tredje ett plustecken och den fjärde avser kvaliteten på läsningen för varje kvävebas. Valfri information kan läggas till. Se mer i FASTQ-specifikationen²⁵.

Olika tillämpning: Olika leverantörer kan ha olika sätt att tillämpa FASTQ-standarden, t.ex. kan filnamn namnges olika och de kan använda olika phred-score encoding.

Filnamn: FASTQ-filernas filnamn genereras på lite olika sätt av olika sekvenseringsmaskiner och innehåller information om provet och körningen. En del av denna information (t.ex. lane number) anges även inuti själva filen. Filnamnen kan även följa konventioner som gör att mjukvara (mappers) kan para ihop filer som innehåller olika läsriktningar (forward/reverse) på rätt sätt. I ett exempel från Illuminas dokumentation kan ett filnamn vara t.ex. Data\Intensities\BaseCalls\SampleName_S1_L001_R1_001.fastq.gz²⁶.

Tabell 4. Exempel på ett filnamns beståndsdelar samt förklaring av respektive beståndsdel.

Beståndsdelar	Förklaring (på engelska)
SampleName	The sample name provided in the sample sheet. If a sample name is not provided, the file name includes the sample ID, which is a required field in the sample sheet and must be unique.
S1	The sample number based on the order that samples are listed in the sample sheet starting with 1. In this example, S1 indicates that this sample is the first sample listed in the sample sheet.
L001	The lane number.
R1	The read. In this example, R1 means Read 1. For a paired-end run, there is at least one file with R2 in the file name for Read 2. When generated, index reads are I1 or I2.
001	The last segment is always 001.

Målbild

Lagring: FASTQ är en viktig ingång för vidare analys, så därför är det viktigt att lagra dem. Problem som kan uppstå vid lagring av stora mängder helgenomsekvenseringsdata och som måste åtgärdas, är begränsad lagringskapacitet och stora kostnader.

Prov-identifierare: Vid lagring i NGP behöver prov indexeras på något annat, mer unikt än vissa av dagens lokalt använda prov-id (SampleName) och filnamn, se avsnittet [Provdata](#). Metadata som ger en unik identifiering av provet behöver läggas till antingen i FASTQ-filernas filnamn eller överföras som separat metadata.

Relationer mellan prov (sample) och FASTQ-fil: Ofta är provets identifierare (SampleName) det enda som kan användas för att koppla FASTQ-filen till mer metadata om själva provet och den behandling av det som gjordes innan sekvenseringen. Det behövs alltså en uppslagsfunktion (via en fil, databastabell, ett datafält i laboratorieinformationssystemet eller liknande) för att koppla detta. Uppslagsfunktioner och vilken sorts metadata som önskas kopplade varierar mellan GMS olika sjukdomsområden. För sällsynta diagnoser vill man t.ex. koppla släktingars prov till varandra innan variant calling görs och för cancer kan man vilja koppla tumörprover till prover på normal vävnad hos samma patient.

²⁵ FASTQ-specifikationen: https://en.wikipedia.org/wiki/FASTQ_format

²⁶ Exempel filnamn (Illumina):

https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/NamingConvention_FASTQ-files-swBS.htm

Namn på sekvenseringsmaskin: I FASTQ-filens första rad ska ett sequence ID anges som innehåller instrumentnamn, t.ex. "MM123" för CASAVA version 1.8 Illumina. Det är oklart om det instrumentnamn som anges på FASTQ-filens första rad (sequence id) är tillräckligt för att NGP ska kunna avgöra vilket instrument som använts eller om informationen bör överföras som ett specifikt metadata för att skapa tydlighet.

Phred score encoding: Leverantörer kan använda olika encoding för att specificera kvaliteten på varje läst kvävebas. Vilken encoding som använts kan inte utläsas av FASTQ-filen, men om en bioinformatiker känner till vilken version av sekvenseringsmaskin som används, går det att lista ut det genom att läsa leverantörens dokumentation. Används t.ex. Illumina 1.3, så innebär det att phred-kvalitetspoäng mellan 0 till 62 skapas (som kodas med ASCII 64-126). För att slippa slå upp detta borde NGP- och bioinformatikgrupperna överväga att överföra encoding som ett separat metadata för att skapa tydlighet. Det kan då även behövas införas en uppslagstabell som matchar leverantör och version av sekvenseringsmaskin mot encoding, vid indexering i NGP.

Viktig metadata

Metadata om FASTQ-filer som bör följa med till NGP och indexeras.

- version av FASTQ-standard.
- namn på sekvenseringsmaskin
- phred score encoding

Hur FASTQ används, t.ex. vad olika delar av sekvenserna representerar, kan variera mellan olika typer av experiment. En del av sekvensen kan t.ex. representera en identifierare för ett prov när flera prover sekvenserats tillsammans och / eller en individuell cell i single-cell experiment.

För att kunna utvinna informationen som lagrats i FASTQ-filerna behövs normalt sett kontextuell information som beskriver experimentet. Utöver den information som representerar sekvensen med kvalitetsmått för varje position finns det flera tillverkarspecifika konventioner för att lagra ytterligare information, t.ex. om vilket instrument som använts och hur konfigurationen sett ut under körning. Ibland är informationen som härrör från samma körning också fördelad över flera FASTQ-filer (och andra filer) som då ofta sammanlänkas med hjälp av konventioner för hur filerna namnges och organiseras i mappar, t.ex. forward + reverse reads, multiplexing, etc.

2.3.10. QC-data (från sekvensering)

Avser kvalitetskontrolldata (QC-data) som sekvenseringsmaskinen producerar. Den används för att säkerställa att genererad sekvensinformation är av god kvalitet och kvantitet. QC-data kan användas både för att se om hela körningen med flera prover gick bra och för att lista ut om ett enskilt prov har tillräcklig kvalitet för vidare bioinformatisk analys.

Nuläge

Vid sekvensering och i det efterföljande bioinformatiska flödet sker många kvalitetskontroller med hjälp av olika mjukvaror. Efter varje QC-körning försöker man rätta till kvalitetsproblem. Om viss sekvensinformation inte är av god kvalitet och detta upptäcks då kvalitetskontroll körs efter sekvenseringen, tas den bort och skickas inte vidare till det bioinformatiska flödet. Man försöker hela tiden att leverera data av så god kvalitet som möjligt till nästa aktivitet.

Format: HTML, annat. Formaten varierar beroende på vilken mjukvara som används. Om verktyg FastQC används produceras HTML-filer.

Målbild

Lagring: Det är oklart om delar av denna kvalitetskontrolldata bör lagras på NGP. Trots att data av god kvalitet hela tiden levereras vidare, kan det ändå vara så att kvaliteten på indatan ifrågasätts i det bioinformatiska flödet. Det första som sker då är att gå tillbaka till den kvalitetskontrolldata som genererades vid sekvenseringen och efter bibliotekspreparation. Av den anledningen kan det bli aktuellt att vissa kvalitetsmått från denna aktivitet bör överföras till NGP och lagras där. Om data ska överföras bör den samlas in i en MultiQC-rapport, se avsnittet [Samlad QC-data](#).

Viktig data

All typ av kvalitetskontrolldata från körningen. Exempel på kvalitetsmått:

- andel baser som överstiger ett visst phred-värde
- andelen obestämda reads
- kontroll att reads är så långa som önskat
- mängden sekvenser

2.3.11. QC-data (avser sekvensens kvalitet)

Information avseende kvaliteten på datan från remiss/patient/prov. Kvalitet på sekvenserna (reads:en) från provet, om dessa är tillräckligt bra. (Denna information avser inte alls själva sekvenseringskörningen.)

Nuläge

Vid sekvensering och under det bioinformatiska flödet sker olika kvalitetskontroller (QC) med olika mjukvaror, t.ex. FastQC eller Picard HsMetrics²⁷ (kräver BAM- eller SAM-filer som indata för att kunna beräkna värden för kvalitetsmått). De algoritmer som mjukvarorna använder ändrar sig över tid. Därför behöver version av QC-mjukvara lagras.

Format: Ospecificerat (HTML-format produceras av FastQC.)

Målbild

Lagring: Denna typ av kvalitetskontrolldata kan genereras om vid behov. Samla in viktiga kvalitetsmått från denna informationsmängd i en MultiQC-rapport och lagra istället den på NGP, se avsnittet [Samlad QC-data](#).

Viktig data

- all typ av QC-data av sekvenserna
- version av QC-mjukvara
- provdata

2.3.12. Samlad QC-data

Avser kvalitetsinformation som är sammanställd från andra kvalitetskontrollrapporter (QC).

Nuläge

Vid sekvensering och i det bioinformatiska flödet skapas det olika kvalitetskontrollrapporter med hjälp av mjukvaror såsom FastQC, Picard HsMetrics och Samtools. Multi-QC-mjukvara kan söka igenom katalogerna där rapporterna har sparats och kompilera ihop utvalda, relevanta kvalitetsmått och deras värden i en enda rapport. Innehållet i MultiQC-rapporten samt de tröskelvärden laboratorerna har bestämt ska uppnås för varje kvalitetsmått,

²⁷ Picard HsMetrics: <https://broadinstitute.github.io/picard/javadoc/picard/picard/analysis/directed/HsMetrics.html>

specificeras i konfigurationen för bioinformatikflödet. MultiQC-rapporten innehåller värden för samtliga prover i en viss körning. (Se mer information om MultiQC²⁸.)

Format: HTML, JSON. Filerna har ofta filnamn som innehåller ordet multiqc, t.ex. multiqc_data.json, multiqc_DNA.html eller multiqc_RNA.html.

Målbild

I nationell kontext MÅSTE det definieras minimikrav för vilka kvalitetsmått MultiQC-rapporten ska innehålla. Specificera vilka övergripande kvalitetsmått som ska fångas från samtliga bioinformatikflöden. Utöver dem kan det finnas ytterligare kvalitetsmått som är specifika för ett visst flöde.

Eftersom GMS önskar delta i nationella nätverk MÅSTE de även gå igenom de kvalitetsmått som Beyond One Million Genomes (B1MG) har specificerat hittills och se om det finns några som bör införas redan nu. Se mer i kapitlet [1+MG- och B1MG-initiativet](#).

Lagring: Kvalitetskontrolldata kan genereras om vid behov, men åtminstone multiQC-rapporten vore bra att spara på NGP.

Viktig metadata

Se exempel på viktig metadata i avsnitten [QC-data \(från sekvensering\)](#) och [QC-data \(avser sekvensens kvalitet\)](#).

2.3.13. Analysdata

Analysdata som behövs för att utföra bioinformatisk analys efter sekvenseringen av provet.

Innehåller exempelvis data om:

- vad som ska analyseras
- vilka prover som hör ihop, exempelvis för sällsynta diagnoser vid trioanalyser
- vilka gener man vill titta på (helgenom eller delar)
- varför analysen körs

Nuläge

Format: JSON

Målbild

Lagring: Metadata om körningstyp (information om vilka gener som undersökts, t.ex. genpanel, helgenomsekvensering eller heltranskriptomsekvensering) är viktig att spara på NGP, men det är oklart i vilket läge den bör fångas. Laboratorierna utför analyser lokalt eller skickar uppdrag om analys till extern leverantör. Detta påverkar i vilket läge som metadatan bör fångas. Eventuellt kan det ske i ett tidigare steg än då analysdata genereras.

Viktig metadata

- körningstyp, det vill säga information om vilka gener som undersökts, t.ex. genpanel, helgenomsekvensering eller heltranskriptomsekvensering
- prov-ID:s som ingår i analys, t.ex. ID på prover vid jämförelse (trioanalys) eller två olika prover där det ena är normal-prov och det andra från en tumör
- germline/Somatisk (det vill säga medfödd eller förvärvad variant)

2.3.14. Alignad sekvens

I aktivitet [Alignment](#) sker bearbetningar där olika verktyg kan vara involverade. Den slutgiltiga alignment-filen är i BAM- eller CRAM-format.

²⁸ MultiQC: <https://multiqc.info/>, <https://multiqc.info/docs/#downstream-analysis>

I en SAM-fil sparas alla sekvensbitar från FASTQ-filen/filerna och information om hur väl de matchade referensgenomet och vid vilken position. Flera kopior av samma avsnitt av genomet kan ligga bredvid varandra i SAM-filen. En effektivare, binär variant av SAM kallas BAM.

Resultatet kan vara att baser i olika reads har inpassats mot samma position i referensgenomet. Dessa ska sedan kontrolleras noggrannare vid variant calling.

Format: SAM/BAM-specifikationen²⁹, version 1.6 vid rapportens författande, innehåller i avsnitt 1.3 ett icke obligatoriskt avsnitt @HD "File-level metadata" som för användning i NGP MÅSTE inkluderas. I @HD-avsnittet finns fältet, VN, "format version", med som ett obligatoriskt fält vilket således också MÅSTE inkluderas. Detta för att underlätta uppgraderingar.

Vi har inga tvingande rekommendationer om övriga frivilliga avsnitt och fält, men standardens inkluderade rekommendationer i avsnitt 2 "Recommended Practice for the SAM Format" BÖR följas. Givetvis MÅSTE standardens obligatoriska villkor alltid följas.

BAM-filer är stora, i storleksordningen 100 GB. CRAM-formatet³⁰ är en komprimerad variant av BAM. Eftersom de upptar mindre lagringsyta BÖR CRAM (inte BAM), version 3.0 eller senare (inställd på "lossless" komprimering), användas om alignade sekvenser ska överföras till eller lagras i NGP.

Lagring och indexering: Tolkningsverktyg använder sig ofta av BAM-/CRAM-filer och utifall tolkningsverktyg ska köras mot NGP så behöver de tillgängliggöras där - alternativt tillgängliggöras på något sätt från lokal/regional lagring. Standardiseringsmässigt drar denna rapport inga slutsatser om var dessa filer lagras, det beror snarare på arkitekturella val och användningsfall. (Vid vårt arbete har det framförts åsikter om att BAM/CRAM troligen bör lagras på NGP. Om både FASTQ samt BAM/CRAM lagras, så dubblas s.k. footprint. Fördelen med BAM/CRAM är att filerna kan användas direkt av ett visualiseringsverktyg såsom Integrative Genomics Viewer (IGV) och Archer. Av den anledningen vore det bra att även lagra BAM/CRAM.

(Det är inte ovanligt att den bioinformatiska pipelinen lägger in beskrivning av referensgenoms-version i BAM/CRAM-filens filnamn.)

2.3.15. Referensgenom

Vid analys jämförs en persons arvs massa med ett standardiserat referensgenom, för att se i vilka delar personen har förändringar.

Format och exempel: Referensgenom är vanligen FASTA-filer som publiceras i och hämtas från en internationell databas. Historiska och senaste publiceringar beskrivs i information från GATK³¹. Vilka versioner av referensgenom som används anges i konfigurationen för bioinformatikflödet.

GRCh37/b37, hg19 eller senare version MÅSTE användas om alignade sekvenser ska överföras till eller lagras i NGP. Om möjligt BÖR dock GRCh38/hg38 hellre användas.

²⁹ SAM/BAM-specifikationen: <https://samtools.github.io/hts-specs/SAMv1.pdf>

³⁰ CRAM-specifikationen: <https://samtools.github.io/hts-specs/CRAMv3.pdf>

³¹ Historiska och senaste publiceringar avseende referensgenom: <https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19>

Lagring och indexering: Om analyser ska köras mot NGP behöver själva referensgenomet lagras där, i övriga fall räcker det att i analysresultatet ange vilket referensgenom som har använts, se rapportens avsnitt om detta.

(Det är inte ovanligt att bioinformatikflödet lägger in beskrivning av referensgenoms-version i BAM/CRAM-filens filnamn.)

2.3.16. Varianter

Information om genetiska varianter, det vill säga de delar av sekvensinformationen som skiljer sig från referensgenomet. Varianter kan antingen annoteras eller inte annoteras och annoteringen kan härröra från olika källor. Lagras oftast i VCF-format.

Nuläge och målbild

Format: VCF. Även om formatet är standardiserat i viss omfattning, kan dess innehåll variera. Detta försvårar standardisering av formatet även om behovet av standardisering är stort för att underlätta utbyte av information samt göra information i databaser transparent. Specifikationen för VCF³² förvaltas av organisationen Global Alliance for Genomics and Health (GA4GH). (Se även information avseende VCF från GATK³³).

Generellt sett skapas det en VCF-fil per prov (single sample VCF), men det kan vara flera prover i samma VCF-fil (multi sample VCF), t.ex. prover från föräldrar och barn vid trioanalys eller parade tumör- och normalprover vid canceranalys. Vid variant calling samkörs alla prover samtidigt och det produceras då en kolumn (eller en fil) per prov, med varianter i förhållande till referensgenomet. En single sample-VCF och en multi sample-VCF är alltså ganska lika, men har olika många kolumner.

Flera VCF-filer skapas: Eftersom olika mjukvaror (variant callers) är anpassade för att identifiera specifika typer av genetiska förändringar skapas det ofta flera VCF-filer vid en bioinformatikkörning.

Exempel på en bioinformatikkörning som genererar tre olika VCF-filer:

- punktmutationer (single nucleotide polymorphism) + mitokondriella förändringar
- strukturell variant
- single tandem repeat (STR)

Kombinera VCF-filer: Variant callers som är anpassade för att identifiera samma specifika typ av genetiska förändring, är olika bra på att identifiera egenskaper i den data som de ska analysera. För att få en mer heltäckande bild av den genetiska förändringen, används därför flera callers i dessa fall. Varje enskilt verktyg genererar en VCF-fil och dessa behöver sedan kombineras ihop till en enda VCF-fil, den s.k. merge-VCF-filen. För t.ex. strukturella varianter kan tre olika variant callers användas, vars VCF-filer kombineras till en merge-VCF. En svaghet med VCF-formatet är då att de fält som förs in i de tre olika VCF-filerna kan se olika ut (de är inte standardiserade fullt ut). Detta medför problem när de tre olika VCF-filerna ska kombineras ihop. Här finns alltså ett behov av standardisering.

Annotering och filtrering: Efter variant calling görs vanligen automatiserad annotering genom att t.ex. slå upp varianter i internationella databaser. Annoteringarna lagras ofta i VCF-filens INFO-kolumn (ett key-value dictionary), exempelvis "denna variant finns i dessa gener och

³² VCF-specifikationen: <http://samtools.github.io/hts-specs/VCFv4.3.pdf>

³³ Information om VCF från GATK: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format>

dess specifika transkript”. Inför den efterföljande tolkningsaktiviteten (se avsnitt [Tolka och visualisera varianter](#)) utförs ofta extra annoteringar, t.ex. avseende familjeinformation.

Laboratorierna använder olika strategier för filtrering, dels på vilken nivå de görs och dels i vilka steg av flödet som de utförs. Om laboratoriet tillämpar hård filtrering³⁴ tas varianter som inte uppfyller kraven bort från VCF-filen. Om de använder mjuk filtrering behålls varianterna och flaggas upp på ett sådant sätt att de ignoreras senare i flödet (annoteringen sker alltså additivt i flera steg så att ingen information raderas). En del laboratorier överlämnar relativt ofiltrerade VCF-filer till tolkningsaktiviteten, medan andra utför nästan all filtrering som är nödvändig. Det är sannolikt den slutliga filen som gått igenom flest steg som skall lagras i NGP.

Vi ser flera fördelar med att tillämpa mjuk (oförstörande) filtrering i filer som ska lagras i NGP, men dessa behöver vägas mot behovet av att reducera antal varianter som ska tolkas. En strategi med mjuk filtrering medför följande sannolika konsekvenser:

- Vi kan förhoppningsvis klara oss med att bara spara en VCF-fil per prov i NGP och därmed göra sökningar i variantinformation mer konsekventa/likartade.
- I de fall där det finns verktyg/applikationer/processteg som behöver hårt filtrerad indata eller som genererar hårt filtrerad utdata behöver man efterbearbeta utdata (inklusive den berikning/annotering som gjorts i steget/verktyget) så att den istället kan representeras som en mjuk filtrering i den VCF-fil som går vidare till nästa verktyg/steg (samt så småningom till lagring i NGP).

Varje annoteringsmjukvara lämnar i VCF-filen en notering om mjukvaran (eller kommandot) som körts, antingen i en mjukvaruspecifik rad (t.ex. `##VEP=v100`) eller på annat sätt. All information som behövs för att identifiera verktyg och verktygsversioner skall finnas i körningslogg och bioinformatisk pipelinespecifikation (se avsnitten [Körningslogg](#) och [Bioinformatisk pipelinespecifikation](#)). I `##INFO`-raderna i VCF-filens header finns förklaringar av de datafält som mjukvaran lägger in i INFO-kolumnerna.

Normalisering: VCF-formatet är inte så strikt formulerat, vilket medför att olika callers representerar varianter i VCF-filerna på olika sätt. Verktyg, såsom VT normalize³⁵, kan användas för att normalisera varianter, men då kan samtidigt information försvinna. I en nationell kontext bör normalisering hanteras på ett enhetligt sätt. Vi rekommenderar att en workshop hålls där personer som har djup kunskap inom bioinformatik diskuterar frågan.

Reanalys: Reanalyser utgår ofta från sekvensinformationen (FASTQ), trots att det egentligen skulle räcka att utgå från alignment-data (BAM/CRAM) eller variantdata (VCF) i vissa fall. Angreppssättet medför ökade beräkningskostnader och i framtiden bör eventuellt mer flexibla reanalys-ramverk byggas. I så fall behöver vissa metadata sparas (se viktig metadata nedan). För VCF-filerna kommer alltid den initiala VCF-filen, det vill säga den ofiltrerade och oannoterade versionen, vara startpunkten för reanalyser.

Kompletterande specifikation: För att utbyta variationsdata (inklusive genomiska varianter) på ett standardiserat sätt rekommenderar GA4GH att specifikationen Variation Representation Specification (VRS) används. Den är en standard för att datormässigt (computable) representera variation och kompletterar de humanläsbara standarder och flatfiles-standarder som finns för att representera genomiska varianter. (Se mer i extern Bilaga E Omvärldsanalys

³⁴ Hård filtrering: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering>

³⁵ VT normalize: <https://pubmed.ncbi.nlm.nih.gov/25701572/>

internationella projekt och databaser.pdf, kapitel 6.1.) Vi rekommenderar att VRS BÖR användas i den mån det är möjligt.

Viktig metadata

- anledningen till att reanalys gjordes
- vilken typ av fil som var startpunkt för reanalys (t.ex. FASTQ, BAM, CRAM, initial VCF)

2.3.17. Konfiguration för bioinformatik

Avser information om vilka inställningar och regler som används för en bioinformatikkörning och sparas för att den innehåller uppgifter som behövs för att felsöka eller upprepa en analys.

Konfigurationen kan bestämma vilken väg som ska tas genom flödet i en bioinformatisk pipelinespecifikation, vilka resurser och versioner som ska användas i olika steg (t.ex. version av referensgenom) och vilken utfil en mjukvara ska skriva till, tröskelvärden, filtreringskriterier m.m. Konfigurationen kan bestå av en eller flera filer och hur filerna ser ut varierar mellan workflow-hanterare, pipelinespecifikationer, den mjukvara som används och dess versioner.

För att en analys ska kunna upprepas måste alla resurser, specifikationer och mjukvaror som konfigurationen avser finnas tillgängliga. En vanlig lösning är att paketera och versionshantera resurser, specifikationer och mjukvaror tillsammans med alla andra element som behövs i s.k. containrar som kan köras på samma sätt i olika datormiljöer.

Nuläge

Format: JSON, YAML. En textfil i olika filformat som normalt innehåller ordet "config", t.ex. "config.yaml". Konfigurationsfiler som används för GMS olika bioinformatikflöden framgår av GMS kodbibliotek i Github³⁶.

Målbild

Lagring: Konfigurationsfiler sparas på NGP tillsammans med uppgifter om motsvarande workflow-hanterare, bioinformatisk pipelinespecifikation och containrar som används för analysen. Sökvägar till filer och hänvisningar till containrar och andra resurser i konfigurationsfilen hanteras på ett sätt så att körningen kan upprepas med samma resultat.

Viktig metadata

- namn/identifikator och versionsnummer för Bioinformatisk pipelinespecifikation
- namn/identifikator och versionsnummer för den workflow-hanterare som kan köra konfigurationen
- namn/identifikator och versionsnummer för den/de containrar som ska finnas tillgängliga

2.3.18. Bioinformatisk pipelinespecifikation

Avser maskinläsbar information som beskriver hur sekvensinformationen ska bearbetas med olika mjukvaror, vilka resurser som krävs och vilka resultat som genereras samt inställningar som kan variera mellan olika analyser och bioinformatikkörningar. Pipelinespecifikationerna utarbetas och sparas gemensamt inom GMS för att återanvändas mellan olika analyser.

Pipelinespecifikationen är en fil eller en uppsättning filer som med hjälp av en workflow-hanterare gör det möjligt att genomföra en komplett bioinformatikkörning med minimal konfiguration. En del av specifikationen utgörs av ett programspråk som är särskilt avsett för att beskriva stegen i informatikflöden och ofta specifikt för den workflow-hanterare som används, t.ex. Nextflow och Snakemake. I stegen beskrivs hur mjukvaran som används konfigureras, vilka filer som ska läsas och skrivas och hur mjukvarans loggar ska hanteras. En

³⁶ GMS kodbibliotek i Github: <https://github.com/genomic-medicine-sweden>

annan del av specifikationen är en default-konfiguration som bestämmer en uppsättning inställningar som ger bra resultat i de flesta analyser där specifikationen används. Den beskriver även vilka filer som ingår och hur de varierar mellan workflow-hanterare och den mjukvara som används.

Specifikationen har stor betydelse för vad konfigurationen och körningsloggen innehåller och hur resultaten produceras. Det är därför viktigt att de är väl dokumenterade, noggrant granskade av flera experter, utförligt testade och tillgängliga för inspektion tillsammans med resultaten. För att en analys ska kunna upprepas med samma resultat vid olika tillfällen är det viktigt att definiera vilka versioner av mjukvaror och andra resurser som används. Det kan också vara bra att exponera utgångsvärden för slumpvalsgenerering och default-värden som används i de olika mjukvarorna.

För att göra pipelinespecifikationerna tillgängliga för såväl återanvändning inom GMS, som för granskning av experter och för inspektion i anslutning till forskningspublikationer, kan de publiceras på plattform för öppen utveckling av programvara, t.ex. Github, och registreras i internationella databaser för utbyte av specifikationer, t.ex. nf-core, snakemake workflow catalogue, och WorkflowHub. Oavsett vilken/vilka lösningar som används har varje specifikation ett namn och ett versionsnummer som kan identifiera den.

Nuläge

Vi har sett exempel på att internationellt erkända specifikationer har vidareutvecklats för användning inom GMS och tillgängliggjorts via GMS Github-organisation³⁷ samt att andra har utvecklats i samverkan med internationella nätverk och tillgängliggjorts i gemensamma databaser som nf-core.

Format: .nf med flera format. Specifikationer för Nextflow eller Snakemake med containrar för Singularity och Docker.

Målbild

Specifikationer utarbetas och används gemensamt på ett systematiskt sätt inom GMS och i samverkan med internationella nätverk så långt det är möjligt. Specifikationerna anpassas för att acceptera konfiguration (ovan) och producera körningsloggar (nedan) enligt motsvarande målbilder.

Format: Specifikationer för Nextflow eller Snakemake med containrar för Singularity och Docker.

Viktig metadata

- namn/identifierare och versionsnummer för specifikationen
- uppgifter om användardokumentation, supportnätverk och artiklar/metodbeskrivningar
- uppgifter om processer för vidareutveckling och förvaltning av specifikationen
- föreslagen citering för användning i rapporter och publikationer
- licensinformation för specifikationen
- namn/identifierare och versionsnummer för den workflow-hanterare som kan köra specifikationen
- namn/identifierare och versionsnummer för de mjukvaror/containrar som används

³⁷ GMS Github-organisation: <https://github.com/genomic-medicine-sweden>

2.3.19. Körningslogg

Logginformation om hur en bioinformatikkörnings olika steg exekverades, vad resultatet var samt andra detaljer. Den sparas för att den kan innehålla kompletterande uppgifter som behövs för att felsöka eller upprepa en analys. (Kallas även pipeline run log.)

Körningsloggen skapas varje gång bioinformatikflödet körs, kan bestå av en eller flera filer och innehåller ofta uppgifter om parametrar som av olika anledningar kan variera mellan körningar, t.ex. slumptal, parametrar som inte definierats i konfigurationsfilen, versioner av resurser som laddats ned etc. Hur filerna ser ut varierar mellan workflow-hanterare, pipelinespecifikationer, och den mjukvara som används.

Nuläge

Format: Olika workflow-hanterare använder troligen sina egna filformat för körningsloggen, det vill säga Nextflow har ett och Snakemake ett annat, men vanligen är det en textfil.

Målbild

Format: I möjligaste mån ska väletablerade standardiserade format användas för loggfilen. Vanligen finns sådana redan definierade för workflow-hanterare och kan återanvändas. Om något ovanligt/egenkonstruerat bioinformatikflöde inte har inbyggt stöd för sådan loggning så behöver motsvarande data lagras i maskinläsbara, indexerbara och väldokumenterade format.

Viktig metadata

- uppgifter som beskriver filformatet och loggens innehåll
- run command
- namn/identifierare och versionsnummer för Bioinformatisk pipelinespecifikation
- namn/identifierare och versionsnummer för den workflow-hanterare
- program (plus versioner, indataparametrar och seeds)
- checksummor, datumstämplar och/eller versionsnummer för databaser och datafiler som laddas under installation/körning
- random seeds (det man sätter igång slumpgeneratorer med) ska också vara dokumenterade så att man verkligen kan reproducera körningen och därmed inte datan

2.3.20. Remissvar

Information som kan finnas i ett svar, t.ex. från den genetiska analysen eller det samlade svaret till inremitterande läkare.

Nuläge

Laboratoriernas rapporter från den genetiska analysen skiljer sig åt. En del är väldigt detaljerade och har tillägg med information om väldigt många gener och varianter. Andra lagrar huvuddelen av informationen i egna databaser och laboratorieinformationssystem (LIMS). Sedan destilleras denna information ner till något ytterst kort i svaret så att mottagaren inte ska behöva läsa igenom alltför mycket information.

Format: I nuläget avges rapporten från genetisk analys i olika typer av format, t.ex. PDF.

Målbild

Det vore värdefullt att skapa mer enhetliga och strukturerade rapporter från den genetiska analysen. I en förlängning är det även intressant att strukturera det samlade svaret till inremitterande läkare, men det är svaret från den genetiska analysen som har högst prioritet för GMS del.

Som beskrivits i avsnittet [Remiss](#) ska relevant information i remissen samt information utöver denna (t.ex. sammanfattande journalutdrag, foton, tillväxtkurvor, släktanamnes) helst redan vara strukturerad när det är dags att lämna remissvar. Detta skulle göra det enklare att jämföra

vilken information hos olika laboratorier som avser exakt samma sak. Även själva remissvaret bör fångas på ett strukturerat sätt.

Ur ett rent NGP-lagringsperspektiv så är det inte självklart att hela remissvaret behöver standardiseras, utan det kanske räcker med att det som beskrivs i avsnittet [Tolkade varianter](#) standardiseras och lagras i NGP. Däremot bör standardiseringsarbetet av dessa två informationsmängder (remissvar och tolkad variant) göras samtidigt eftersom de delvis överlappar varandra och ofta med fördel skapas i samma verktyg eller process av samma personer. Många av dagens patientjournalssystem saknar förmåga att ta emot remissvar strukturerat och lagrar det huvudsakligen som fritext och PDF. Förändring är dock på gång, flera regioner kommer kunna ta emot strukturerade remissvar baserade på openEHR och/eller HL7 FHIR.

I ett remissvar kan även länkar till kliniskt relevant referensinformation som gällde vid remisstillfället behöva skickas med så att man i mottagande system senare kan avgöra vad vårdpersonalen som läste remissen hade att utgå ifrån.

Vi lämnar inga specifika rekommendationer avseende själva längden på rapporten från den genetiska analysen. Personer med klinisk expertis bör besluta huruvida rapporten ska vara kort eller lång eller om båda versioner får förekomma.

2.3.21. Genlista

Information om gener som är kända för att vara associerade med en viss sjukdom, s.k. genlistor eller *in silico*-genpaneler.

Vid breda genanalyser i klinisk kontext, används genlistor som är specifika för en viss sjukdomskategori, t.ex. bröstcancer, intellektuell funktionsnedsättning eller skelettdysplasi. De begränsar mängden gener/varianter som sjukhusgenetikern behöver granska i tolkningsaktiviteten, vilket är viktigt för att snabba upp svarstiderna till patienten. Om en gen läggs till eller tas bort ur genlistan påverkar det resultatet av tolkningen ganska mycket.

Nuläge

Genlistorna kompileras ihop av expertteam inom sjukvården och uppdateras allteftersom nya upptäckter av sjukdomsassocierade gener görs inom forskning och ny vetenskaplig litteratur publiceras. De är inte standardiserade och kan skilja sig åt mellan olika laboratorier.

Lagring: Genlistorna lagras troligen i tolkningsverktyg och följaktligen blir det tolkningsverktygens uppgift att lagra information så att det går att slå upp vilken version av en genlista som använts vid en viss tidpunkt.

Målbild

Lagring: Vid reanalys kan det behövas information om vilken version av en genlista som har använts vid tidigare analyser, men det är oklart huruvida det räcker att tolkningsverktygen sparar information om sådant eller om även NGP bör göra det.

Viktig metadata

- genlistans namn
- genlistans versionsnummer
- generna i genlistan

2.3.22. Tolkade varianter

Klinisk tolkning är en process där tillgängliga evidens för huruvida en variant har betydelse för en patients sjukdom samlas in (t.ex. från olika databaser) och bedöms, varefter en klassificering sker utifrån evidensen. Historiskt sett har laboratorier använt egna metoder för

tolkning och klassificering, men de har alltmer övergått till att använda internationella klassificeringsscheman. Klassificeringsscheman ändras över tid, så det är viktigt att fånga information om vilket klassificeringsschema som användes då klassificeringen gjordes.

Vid arbetet används olika tolkningsverktyg, det vill säga mjukvaror som har olika stöd för annotering, prioritering, evidensbaserad klassificering och rapportering av varianter. Verktygen har integrationer till internationella databaser och har ibland byggt upp egna kunskapsdatabaser som deras användare kan nyttja.

Med hjälp av verktygen granskar användarna visuell information om varianter, lägger in information och gör tolkningar. Visualiseringsstödet för små varianter är i regel bra, men för vissa verktyg skulle visualisering av kopietalsförändringar och fusionsgener behöva förbättras. Ibland behöver användarna även visualisera alignad sekvens för att kunna utesluta artefakter (det händer att artefakter slinker igenom till tolkningsaktiviteten trots att de reduceras så mycket det går i bioinformatikflödet). Det är främst sjukhusgenetiker som är målgrupp för visualiseringar, men ibland ombeds bioinformatikerna att ge synpunkter på olika patientfall och då kan det vara bra att även de kan visualisera informationen.

Genlistan (se avsnitt [Genlista](#)) är av central betydelse vid klinisk tolkning. Den importeras in i tolkningsverktygen och används för att välja varianter inom den specifika genlistan.

Nuläge

Harmonisera vilka varianter som ska granskas: Laboratorierna har olika tillvägagångssätt för att bestämma vilka varianter som ska tolkas. Några analyserar enbart de gener som är med i genlistan, medan andra bedömer i princip alla varianter efter det att de har filtrerat mot blod-DNA-sekvensen. Det vore värdefullt att harmonisera hur laboratorierna gör tolkning, det vill säga hur laboratorierna bestämmer exakt vilka av varianterna som de ska granska.

Lagring: Information om tolkade varianter är intressant för flera mottagare t.ex. NGP, variantdatabaser, kvalitetsregister, framtida patientjournalssystem och beslutsstöd. Alla former av tolkningar är av intresse för NGP, det vill säga information om samtliga varianter som har tolkats (inte enbart information om de patogena varianter som svaras ut) samt historiska tolkningar (information om tolkningar som har gjorts bakåt i tiden).

Format: BAM, CRAM, VCF. Vi har inte inventerat vilka indataformat samtliga verktyg stödjer, men många bör ta VCF-filer som indata. Vissa verktyg tar BAM-/CRAM-filer som indataformat för att användarna ska kunna granska alignad sekvens visuellt.

Målbild

Format: JSON, VCF. När tolkade varianter överförs från laboratoriernas olika tolkningsverktyg till NGP ska JSON eller en VCF med information om tolkade varianter användas. (För indataformaten till tolkningsverktygen ska BAM, CRAM eller VCF användas.) Format behöver vara kompatibla mellan tolkningsverktyg, variantdatabaser och patientjournalssystem.

Viktig metadata

- tolkningsdatum/utsvarningsdatum
- vilka evidens som låg till grund för tolkningen inklusive referenser till publikationer för gener/varianter som är studerade tidigare
- information om tolkningen inkluderar experimentella evidens eller wet lab-evidens
- klassificeringsschema som användes vid tolkningen
- signatur och kontaktuppgifter till den som gjort tolkningen

2.4. Informationsbärare

I detta avsnitt beskrivs informationsbärare som utgör lagringsplatser för informationsmängder eller som via API:er utbyter informationsmängder som är generella för samtliga kartlagda sjukdomsområden.

2.4.1. Patientjournal och laboratoriesystem

Avser IT-system som är lagringsplatser för och således utgör källor till information om prov, patient, fenotyp m.m.

Det finns ett antal olika patientjournalssystem, laboratorieinformationssystem samt ibland särskilda remiss- och svarssystem bland regionerna. Dessa system är källa till mycket av det som i bioinformatiskt perspektiv kategoriseras som "metadata", exempelvis det som beskrivs mer detaljerat i avsnitt [Remiss](#), [Provdata](#) samt [Patientdata samt fenotypdata och annan klinisk data](#). Journalsystem, laboratoriesystem (eller särskilda remiss- och svarssystem) är även mottagare (eller vidareförmedlare) av remissvar (se avsnitt [Remissvar](#)) från den bioinformatiska processen. Oavsett lösning så behöver flera typer av data kunna överföras på standardiserat sätt i båda riktningar.

Nuläge

Idag är en stor del av denna information i fritext- eller PDF-form vid överföring mellan system vilket medför varierande form och kvalitet samt ofta behov av omtolkning och manuell återinmatning i nästa system i kedjan (dubbeldokumentation).

Målbild

Överenskommen återanvändbar struktur i utbytet så att data kan gå automatiskt mellan systemen utan manuell omtolkning och återinmatning. Kliniska beslutsstödssystem kopplade till journalsystem kommer också behöva utgå från strukturerad data från bioinformatiska analyser för att bidra till att precisionsmedicin kan införas effektivt.

2.4.2. Internationella annoteringsdatabaser

Avser databaser som är lagringsplatser för information om annoteringar relevanta inom genomikområdet. Databaserna används av olika annoteringsmjukvaror vid variantannotering i bioinformatikflödet (se aktivitet [Variantannotering](#)) samt av tolkningsverktyg vid tolkning (se informationsmängd [Tolkade varianter](#)). De databaser som mjukvarorna använder varierar, men några av dem beskrivs i extern Bilaga E Omvärldsanalys internationella projekt och databaser.pdf.

2.4.3. Nationella variantdatabaser

Avser databaser som är lagringsplatser för information om identifierade varianter, annoterade varianter, tolkade varianter samt även annan information. Databaserna kan utgöra grund för att hämta information om huruvida en variant har påträffats hos andra laboratorier och hur laboratoriet i så fall tolkade varianten, beräkna aggregerade variantfrekvenser samt generera diverse analyser.

Nuläge

I nuläget använder laboratorierna lokala variantdatabaser i viss mån. Exempelvis kan ett laboratorium ha en databas som håller ordning på vilka varianter som har påträffats hos laboratoriets patienter med sällsynta diagnoser. Dessa används sedan för att räkna fram med vilken frekvens en variant förekommer hos laboratoriets patienter. Laboratorierna använder även lokala artefaktdatabaser men det är oklart om dessa klassas som variantdatabaser.

Målbild

GMS vill skapa nationella variantdatabaser där information om varianter kan delas över hela Sverige. Därför har GMS variantdatabasgrupp tagit fram en specifikation kring vilka önskemål som finns. Initialt ska en variantdatabas avseende tolkade varianter utvecklas som innehåller information om samtliga varianter som har tolkats (inte enbart information om de sjukdomsorsakande varianter som svaras ut) samt information om tolkningar som gjorts bakåt i tiden (historisk tolkning). Därefter ska en databas med identifierade varianter utvecklas som avser tekniskt korrekta och biologiskt relevanta varianter kombinerade med fenotypdata. Båda typerna av databaser efterfrågas av sjukdomsområdena barncancer, hematologi, solida tumörer och sällsynta diagnoser.

Enligt GMS AU Informatik ska NGP användas som grund för den tekniska lösningen. Variantdatabaserna fungerar då som vyer av NGP som gör det möjligt att söka bland de tolkade och annoterade varianter som finns lagrade i NGP.

Information om tolkade varianter behöver överföras i någon sorts fil (t.ex. JSON eller VCF) från de olika tolkningsverktyg som laboratorierna använder till NGP på ett standardiserat sätt. All information som väljs ut som index ska vara kodad med internationella ontologier i möjligaste mån (men det är kanske inte realistiskt att kräva för historisk information). GMS ska alltså eftersträva att de olika laboratorierna överför entydig och enhetlig information.

Det behöver tydliggöras exakt vilken information som ska indexeras för varje sjukdomsområde. Vi ger endast allmänna riktlinjer kring den information som variantdatabasgruppen hittills har identifierat (se länkar till respektive informationsmängd nedan).

Variantdatabaserna ska versionshanteras på ett sätt som gör det möjligt att se förändringar mellan numrerade versioner eller efter/före ett visst brytdatum.

Viktig metadata i en variantdatabas med tolkade varianter

- kön, ålder, diagnos, på sikt ska mer klinisk data läggas till (se avsnittet [Patientdata samt fenotypdata och annan klinisk data](#))
- vilket laboratorium som analyserade provet (se avsnittet [Genomiknod](#))
- provtagningsdatum, vävnad, tumörtyp (se avsnittet [Provdata](#))
- tolkningsdatum, underliggande evidens och referenser till publikationer, signatur och kontaktuppgifter till den som gjort tolkningen, information om tolkningen inkluderar experimentella evidens eller wet lab-evidens, klassificeringsschema (se avsnittet [Tolkade varianter](#))

Viktig metadata i en variantdatabas med samtliga relevanta varianter

- kön, ålder, frågeställning från remittent, fenotyp, diagnos (se avsnittet [Patientdata samt fenotypdata och annan klinisk data](#))
- kontaktinformation till dataägaren (se avsnittet [Genomiknod](#))
- inputkoncentrationer av DNA/RNA (se avsnittet [Bibliotekspreparationsdata](#))
- ID på provet, vävnad, tumörtyp, tumörcellhalt (se avsnittet [Provdata](#))

Viktig metadata - övrigt

Variantdatabasgruppen har identifierat ytterligare metadata som är viktiga (ett urval listas nedan). För att vi ska förstå mer om dessa typer av information behövs ytterligare dialog, t.ex. med person som har klinisk expertis.

- processdatum
- bioinformatiska verktyg som användes vid analys inklusive versionsnummer
- gennamn
- ankomstdatum
- svarsdatum

- analysdatum
- analysmetod
- QC-data från sekvenseringen, t.ex. medeltäckning i provet, totalt antal reads, uniformity, on target

2.4.4. Tolkningsverktyg

Avser mjukvaror som tillför information om samt utgör lagringsplats för tolkade varianter.

Laboratorierna använder många olika typer av kommersiella och egenutvecklade tolkningsverktyg, t.ex. Alamut (Sophia Genetics), Alissa Interpret (Agilent), Coyote (egenutvecklat), QCI (Qiagen) och Scout (egenutvecklat). Även MS Excel används i vissa fall. Ett och samma laboratorium kan ha flera olika typer av tolkningsverktyg som är anpassade för olika sjukdomsområden. Verktygen kan ses som visualiseringsverktyg eller beslutsstöd som är begränsade till att ställa genetisk diagnos. Nedan ges några exempel på tolkningsverktyg som används.

QCI

Qiagen Clinical Insight (QCI) är ett kommersiellt tolkningsverktyg. Det finns i olika versioner som är anpassade för olika sjukdomsområden. QCI har en funktion för identifikation av mutationsmönster.

Qiagen har byggt upp en egen kunskapsdatabas, Qiagen Knowledge Base, som deras kunder kan nyttja. Kunskapsdatabasen är versionshanterad och det släpps nya versioner av den löpande.

Att köra en QCI-analys kostar pengar per exom, så därmed spelar komplexitet på analys roll. Kunden betalar för tillgång till den kurerade kunskapsdatabasen som hålls uppdaterad.

Information om alignad sekvens och varianter (BAM, VCF) kan skickas som indata till QCI.

Scout

Scout är ett egenutvecklat tolkningsverktyg. Det har valts som nationellt tolkningsverktyg inom GMS, vilket innebär att det kommer att vidareutvecklas så att det passar GMS behov. Scout har stöd för tolkning av sällsynta diagnoser samt cancer, men vidareutveckling pågår inom båda områdena.

Information om alignad sekvens och varianter (BAM, CRAM, VCF) ska skickas som indata. Variantfilerna används för att visualisera varianter, medan alignad sekvens används för att visuellt granska hur reads avseende en viss variant ser ut.

Visualiseringsverktygen GENS och Integrative Genomics Viewer (IGV) har integrerats i Scout. GENS används för visualisering av kopietalsförändringar för sällsynta diagnoser, medan IGV används för visualisering av alignad sekvens. Vid visning av släktträd används ett verktyg från leverantören Madeline för att rita upp själva bilderna.

Exempel på hur Scout fungerar: Scout tar den automatiskt annoterade VCF-filen och lägger in i sin databas. Sedan läggs tolkningar och patientinformation till manuellt och lagras då också i databasen. För fenotypbeskrivningar används termer ur HPO. Dessutom används OMIM-fenotyper och OMIM-gener. För matchning mellan fenotyp (klinisk bild) och genotyp används olika databaser såsom ClinVar, Matchmaker Exchange, ORPHAcodes och OMIM.

3. Kartläggning av behov som är specifika för enskilda sjukdomsområden

Detta kapitel sammanfattar behov av standardisering som är specifika för enskilda sjukdomsområden inom GMS. Nedan redogörs för aktiviteter (alternativt detaljer om aktiviteter) som genomförs för ett visst sjukdomsområde, informationsmängder (alternativt detaljer om informationsmängder) som skapas eller används då aktiviteterna utförs samt informationsbärare (alternativt detaljer om informationsbärare) som utgör lagringsplatser för informationsmängder eller som via API:er utbyter informationsmängder. För några av sjukdomsområdena illustreras flödet även i en vy.

Aktiviteter, informationsmängder och informationsbärare som är gemensamma för samtliga kartlagda sjukdomsområden har beskrivits i föregående kapitel (se [Kartläggning av behov som är gemensamma för flera sjukdomsområden](#)).

3.1. Barncancer

GMS arbetsutskott för barncancer genomför helgenom- och helexomsekvensering (samt DNA-metylering av hjärntumörer) för alla barn med cancer. Syftet är att ge en mer detaljerad diagnos och subgruppera patienter så att behandling kan styras bättre än tidigare.

Parallellt pågår även projekt för att identifiera medfödda genetiska förändringar som finns i icke-cancerceller. På så vis kan biverkningar eller terapiresistens undvikas. Dessutom kan nya gener och signalvägar upptäckas som kan användas för behandling och kontrollprogram.

3.1.1. Scenario

DNA-baserad analys kompletteras med RNA-baserad analys för att få information om fusionsgener. Dessa hjälper till att sätta diagnosen och det finns läkemedel som riktas mot fusionsgener.

En persons tumörprov samanalyseras med ett normalprov från samma person som jämförelsematerial (s.k. parade tumör- och normalprover).

Trioanalys sker i vissa fall.

Information kommer att delas med Barntumörbanken (en biobank inom barncancerområdet) och uttag av information ska kunna göras för forskning.

3.1.2. Vy

En vy har inte sammanställts för barncancer.

3.1.3. Aktiviteter

Några aktiviteter eller detaljer om aktiviteter som genomförs specifikt för barncancer har inte identifierats. (De aktiviteter som är gemensamma för barncancer samt de övriga kartlagda sjukdomsområdena sammanfattas i föregående kapitel, se avsnittet [Aktiviteter](#).)

3.1.4. Informationsmängder

Nedan beskrivs detaljer om informationsmängder som är specifika för barncancer. Informationsmängderna skapas eller används då aktiviteterna i flödet utförs. (De informationsmängder som är gemensamma för samtliga kartlagda sjukdomsområden sammanfattas i föregående kapitel, se avsnittet [Informationsmängder](#)).

3.1.4.1. Patientdata samt fenotypdata och annan klinisk data

Viktig metadata

- initialt antagande från patolog efter histopatologianalys, t.ex. "melanoma". Detta antagande kan komma att revideras efter det att man har fått fram resultaten efter den molekylärgenetiska analysen.

3.1.4.2. Bibliotekspreparationsdata

Hur bibliotekspreparation utförs skiljer sig mycket åt mellan laboratorier. Inom barncancer använder olika laboratorier samma metoder så långt som det är möjligt, men hantering av prover med liten mängd DNA sköts på olika sätt. Vissa laboratorier har ett speciellt kit som de använder för att kunna köra prover som endast innehåller små mängder av DNA, medan andra laboratorier låter bli att köra denna typ av prover. Dessa typer av skillnader är viktiga att fånga som metadata.

3.1.4.3. Panel av normaler

Mjukvaror (variant callers) behöver olika typer av indata. En del behöver bara konfigurationen för bioinformatikflödet samt alignad sekvens, men andra behöver en panel med normaler.

Inom barncancer är det vanligast att utföra parade tumör-normalanalyser, men för vissa hematologiska maligniteter går det inte att få fram ett normalprov vid primär diagnos eftersom tumörceller har migrerat till alla vävnader. En del laboratorier utför då tumor-onlyanalys, vilket innebär att de inte har tillgång till något normalprov att jämföra med. Istället används en panel av normaler för att fånga kopietalsförändringar bättre.

Panelen av normaler används vid variant calling (se föregående kapitel, aktivitet [Variant calling](#)).

3.1.4.4. Tolkade varianter

Tolkningen av medfödda varianter hos barn skiljer sig lite från de tolkningar av medfödda varianter som görs inom sjukdomsområdet sällsynta diagnoser. För barn granskas icke-sjukdomsgener, medan sällsynta diagnoser granskar i princip alla gener som är kända sedan tidigare för att vara associerade med sjukdom. De båda sjukdomsområdenas bioinformatikflöden är alltså väldigt lika men omfattningen av tolkningen skiljer sig åt - för barn granskas ett färre antal gener.

3.1.4.5. Remissvar

För solida tumörer hos barn skickas svaret från den genetiska utredningen tillbaka till patologen. Därefter går det ut som ett tillägg till det vanliga patologisvaret (som innehåller information om diagnos och subgrupp) till behandlande läkare. Svar som gäller eventuella medfödda faktorer kan följa senare.

3.1.5. Informationsbärare

Vi har inte identifierat någon informationsbärare som är specifik för barncancer, men de informationsbärare som används av samtliga kartlagda sjukdomsområden sammanfattas i föregående kapitel, se avsnittet [Informationsbärare](#).

3.2. Hematologi

GMS arbetsutskott för hematologi (blodcancersjukdomar) utvecklar två nationella breda genpaneler som ska användas för att analysera gener med NGS-metoder. Den ena omfattar 199 gener som har betydelse för flera olika typer av myeloiska maligniteter. Den andra omfattar 252 gener som har betydelse för olika lymfatiska maligniteter. Utvärdering ska ske för att se om de nya genpanelerna ger fördelar framför de olika, mindre genpaneler som används i nuläget inom svensk sjukvård.

Det pågår även studier för att undersöka om ännu bredare och mer heltäckande genanalyser, det vill säga helgenomsekvensering kombinerat med helexomsekvensering (RNA-sekvensering), kan identifiera nya sjukdomsorsakande varianter hos personer som insjuknar i akuta leukemier samt om de kan leda till förbättrad klinisk diagnostik, prognos och behandling.

Dessutom ska riktlinjer tas fram för hur personer, som ärver genetiska förändringar som orsakar blodcancersjukdomar, ska diagnosticeras och följas upp.

3.2.1. Scenario

Det skulle kunna vara intressant att visualisera hur sjukhusgenetiker svarar ut på olika laboratorier. Detta kan skilja sig åt.

Hematologi vill dela genomikinformation avseende tolkade varianter med Regionalt Cancercentrums individuella patientöversikter. Planer på samarbete med Blodcancerregistret diskuteras.

Bioinformatiskt flöde Hematologi

The diagram illustrates the bioinformatics workflow in hematology, starting from a referral and project description, through sample collection, sequencing, alignment, variant calling, and interpretation, leading to a final report and storage in various databases.

Key Components and Data Flow:

- Inputs:** Remiss (Referral), Projektbeskrivning/Forskningsinitiativ (Project description/Research initiative), Provdata (Sample data), Biobank, Genpanel (Gene panel), Initial sekvens (Initial sequence), Sekvens (Sequence), Konfig för pipeline (Pipeline configuration), Referensgenom (Reference genome), Kopietal (Copy number), Befintliga annoteringsdatabaser (Existing annotation databases), Annoterade varianter (Annotated variants), Variantdatabas (Variant database), QC-data (avser sekvensens kvalitet) (QC data (concerns sequence quality)), Genlista (Gene list), and Remissvar (Referral response).
- Processes:**
 - Ta prov (Take sample):** Receives Remiss, Projektbeskrivning/Forskningsinitiativ, Provdata, and Biobank.
 - Sekvensera prov (Sequence sample):** Receives Genpanel and Initial sekvens.
 - Alignment:** Receives Sekvens and Konfig för pipeline.
 - Variant calling:** Receives Alignad sekvens (Aligned sequence) and Referensgenom.
 - Variantannotering (Variant annotation):** Receives Kopietal and Befintliga annoteringsdatabaser.
 - Variantfiltrering (Variant filtering):** Receives Annoterade varianter and Variantdatabas.
 - Tolka och visualisera varianter (Interpret and visualize variants):** Receives Tolkade varianter (Interpreted variants) and QC-data (avser sekvensens kvalitet).
- Outputs and Storage:**
 - Analys genomförd och lagrad (Analysis completed and stored):** The final output of the interpretation process.
 - Scout / Coyote, QCI, Beacon, and NGP-Hematologi:** Databases where the results are stored.

Figur 7. Vy över uppfattat nuläge för hematologiflödet

3.2.3. Aktiviteter

Nedan beskrivs detaljer kring aktiviteter som genomförs specifikt för hematologi. (De aktiviteter som genomförs i ett generiskt flöde sammanfattas i föregående kapitel, se avsnittet [Aktiviteter](#).)

3.2.3.1. Ta prov

Provet är ofta ett blodprov eller benmärgsprov. Ibland tas ett vävnadsprov med hjälp av benmärgsbiopsi.

3.2.3.2. Sekvensera prov

För närvarande är det DNA som sekvenseras.

Olika sekvenseringsmaskiner används på olika orter, men det är inte så stor skillnad mellan dem. Beroende på sekvenseringsmaskin uppkommer dock olika artefakter.

3.2.3.3. Bearbeta sekvens med bioinformatik

Bioinformatikflödet som beskrevs vid intervjun med bioinformatiker avsåg myeloiska maligniteter, men flödet kommer inte skilja sig nämnvärt när det anpassas för lymfoida maligniteter senare. Det avser hantering av enbart tumörprover och inte parade tumör-normalprover. I nuläget finns det inget behov av att anpassa bioinformatikflödet för RNA-sekvensering.

Variant calling

Bioinformatikflödet utför flera typer av variant calling för olika syften. De typer av genetiska förändringar som identifieras är punktmutationer, indels och kopietalsförändringar.

3.2.4. Informationsmängder

Nedan beskrivs detaljer om informationsmängder som är specifika för hematologi. Informationsmängderna skapas eller används då aktiviteterna i flödet utförs. (De informationsmängder som hanteras i samtliga kartlagda sjukdomsområden sammanfattas i föregående kapitel, se avsnittet [Informationsmängder](#).)

3.2.4.1. Patientdata samt fenotypdata och annan klinisk data

För hematologi är frågeställning en viktig typ av metadata i en variantdatabas, eftersom de inte alltid har en diagnos klar när de molekyllärgenetiska analyserna görs.

Hematologi vill i samarbete med Blodcancerregistret undersöka om varianter är kopplade till kliniska parametrar såsom ålder, kön, klinisk diagnos, överlevnad, risk för återfall och behandlingsresultat.

Viktig metadata

- frågeställning från remittent
- kön
- ålder
- klinisk diagnos
- överlevnad
- risk för återfall
- behandlingsresultat

3.2.4.2. Provdatab

För personer som genomgår behandling behöver uppföljningsprover tas. Ett prov tas först vid diagnos och sedan fortsätter provtagning under hela den tid som personen får behandling, i t.ex. fem år. En del laboratorier har fler uppföljningsprover att hantera jämfört med andra.

3.2.4.3. QC-data (från sekvensering)

Förutom artefakter kan andra typer av tekniska problem uppstå vid sekvensering som behöver åtgärdas, s.k. bakomliggande brus. Brus stör mer helheten, medan en artefakt kan avse en falsk positiv träff. Vid kvalitetskontroll går det att se om ett visst värde har varit brusigt, eller om hela körningen har varit det.

3.2.4.4. Tolkade varianter

Något som skulle kunna vara intressant att visualisera är att jämföra hur sjukhusgenetiker svarar ut på olika laboratorier, eftersom detta kan skilja sig åt.

Det finns önskemål om att dela information om tolkade varianter kopplade till enskilda cancerdiagnoser med Regionalt Cancercentrums individuella patientöversikter. (Se även solida tumörers informationsmängd [Tolkade varianter](#).)

3.2.5. Informationsbärare

Vi har inte identifierat någon informationsbärare som är specifik för hematologi. (De informationsbärare som används av samtliga kartlagda sjukdomsområden sammanfattas i föregående kapitel, avsnittet [Informationsbärare](#).)

3.3. Solida tumörer

GMS arbetsutskott för solida tumörer implementerar bred och mer heltäckande molekylär karakterisering av tumörer. Tumörernas molekylära egenskaper kan sedan användas för att utveckla diagnostik, prognostik och behandling riktad mot en persons specifika tumöregenskaper. De inför även analysmetoder för cirkulerande tumör-DNA samt analysverktyg och multidisciplinära behandlingskonferenser som är enhetliga över landet.

Det arbete som har kommit längst är utveckling av en nationell bred genpanel för att analysera 560 cancerassocierade gener med NGS-metoder. Genpanelen ger information om genetiska förändringar som är viktiga för val av behandling och molekylär subtypning av tumörer samt förändringar som kan utgöra grund för studieinklusion. Den inkluderar även ett antal farmakogener som är viktiga för läkemedelsdosering. Genpanelen implementeras i klinisk rutin under hösten 2022 samt ska användas i två kliniska forskningsstudier som ska undersöka olika behandlingsalternativ, Megalit och Pluto.

Parallellt undersöker arbetsutskottet klinisk nytta av att använda ännu bredare och mer heltäckande analyser, det vill säga helgenom- och helexomsekvensering.

Arbete pågår även inom GMS bioinformatikgrupper för att ta fram ett nationellt bioinformatikflöde som ska passa solida tumörer (och farmakogenomik i viss mån).

3.3.1. Scenario

Inom solida tumörer kompletteras DNA-baserad analys med RNA-baserad analys för att ta fram förstärkt information om fusionsgener. Sådan information är viktig för behandlingsprediktion eftersom det har tagits fram läkemedel som riktas mot fusionsgener. Många av fusionsgenerna tros även vara viktiga för vilken diagnos som sätts och detta kommer att undersökas ytterligare.

Prover från solida tumörer är ofta formalinfixerade vilket gör att det bildas en särskild typ av artefakter vid sekvensering. Vid analys finns ofta bara ett sparsamt material att utgå ifrån.

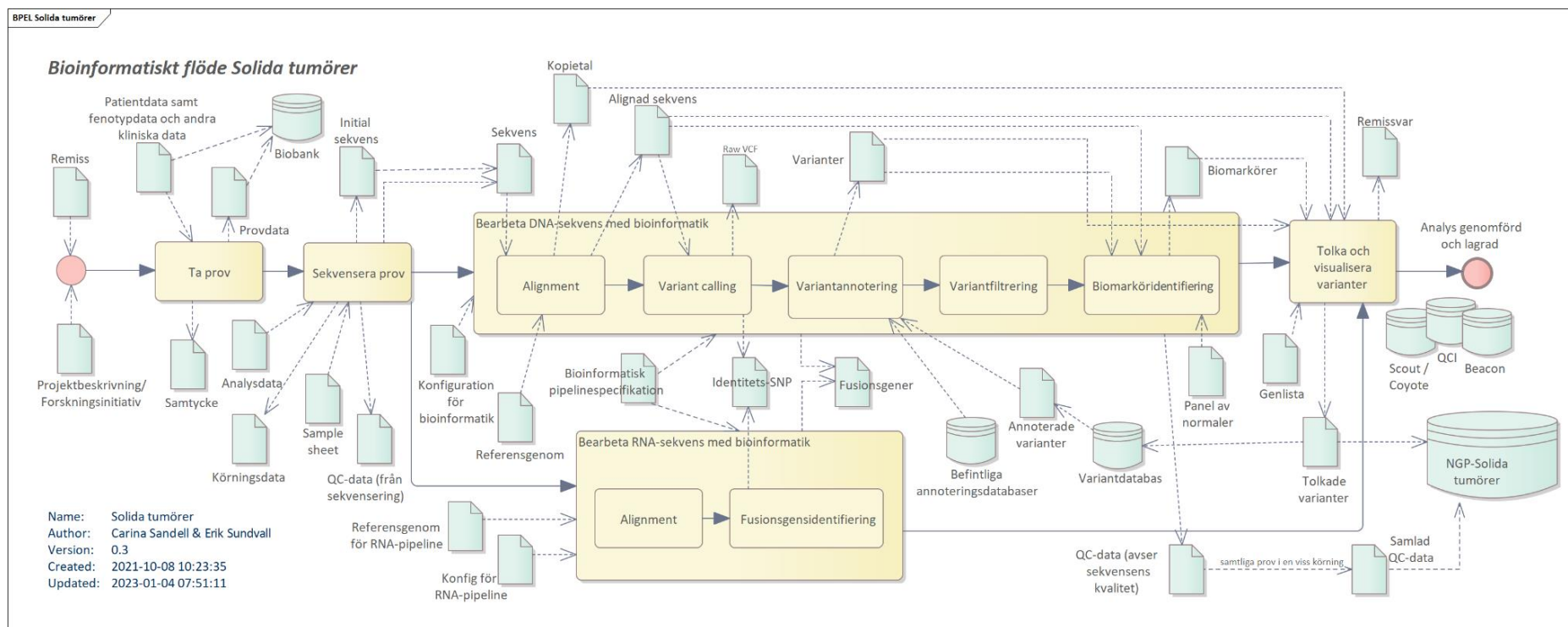
Ibland samanalyseras en persons tumörprov med ett normalprov från samma person som jämförelsematerial (s.k. parade tumör- och normalprover).

Förutom information om varianter (punktmutationer, indels, kopietalsförändringar) identifieras även information om fusionsgener och biomarkörer.

Solida tumörer vill dela genomikinformation avseende tolkade varianter med Regionalt Cancercentrum Västs individuella patientöversikter.

3.3.2. Vy

I detta avsnitt ges en visuell översikt (diagram) över de aktiviteter, informationsmängder och informationsbärare som är involverade i ett flöde för solida tumörer. För en beskrivning av vad symbolerna i vyerna representerar, se [Bilaga C. Teckenförklaring](#).



Figur 8. Vy över uppfattat nuläge i flöde för solida tumörer

3.3.3. Aktiviteter

Nedan beskrivs detaljer kring aktiviteter som genomförs specifikt för solida tumörer. (De aktiviteter som genomförs i ett generiskt flöde sammanfattas i föregående kapitel, avsnittet [Aktiviteter](#).)

3.3.3.1. Ta prov

Prover tas från en persons solida tumör vid olika typer av biopsier (vävnadsbiopsi, fin nålpunktion etc.) eller i samband med operation. Förutom ett tumörprov (sjuk vävnad), tas ibland ett normalprov (icke-sjuk vävnad) från samma person. Ett sådant parat normalprov kan användas som jämförelsematerial och underlätta den bioinformatiska analysen senare i processen. Provtagaren är försiktig vid sin provtagning för att inte orsaka några skador, vilket medför att det ofta bara finns ett sparsamt provmaterial att utgå ifrån vid analys. Efter biopsi hanteras provmaterialet på olika sätt. Det fixeras ofta i formalin och bäddas in i paraffin, men kan även sparas färskfruset.

Förutom vävnadsprover planerar GMS även att använda prov från flytande biopsier. I dessa fall används blodprover för att analysera cirkulerande tumör-DNA som avsöndrats från solida tumörer i blodplasma.

Proverna skickas tillsammans med remiss till en patologavdelning för att fastställa cancerdiagnos. Därefter utför en genetiker på ett molekylärpatologiskt laboratorium genetisk analys av provmaterialet.

3.3.3.2. Sekvensera prov

Kvaliteten på extraherad nukleinsyra varierar vilket i sin tur kan påverka sekvenseringsresultatet. Om provmaterialet är formalinfixerat så bryts nukleinsyra ner, vilket medför att det blir svårt att genomföra en heltäckande sekvensering av provmaterialet. Det bildas en speciell typ av artefakter som behöver plockas bort senare i processen. Provmaterial som sparas färskfruset är bättre lämpat för breda genanalyser, t.ex. helgenomsekvensering.

GMS sekvenserar prover på både DNA- och RNA-nivå, eftersom DNA-sekvensering ibland behöver kompletteras med RNA-sekvensering för att få förstärkt information om fusionsgener. (DNA-sekvensering ger viss information om fusionsgener, men för att få en fullständigare bild behöver även RNA sekvenseras.) Det ska vara möjligt att sekvensera antingen DNA eller RNA, eller att sekvensera båda två samtidigt. Hanteringen av DNA och RNA ska vara så lik som möjligt, så att de enkelt kan samköras. Huruvida det är RNA eller DNA som ska sekvenseras beror på vilken typ av analys som efterfrågas.

3.3.3.3. Bearbeta DNA-sekvens med bioinformatik

Bioinformatikflödet som beskrivs i denna rapport stämmer inte riktigt för parade tumör- och normalprover, eftersom det då blir fler indatafiler och andra algoritmer som används. När det nationella bioinformatikflödet för solida tumörer utvecklas kommer det att stödja analys av tumörprover initialt, men kommer sedan att vidareutvecklas så att det kan hantera parade tumör- och normalprover.

Variant calling

De genetiska förändringar som identifieras är huvudsakligen behandlingsstyrande, men kan även underlätta att fastställa en mer exakt diagnos. Sekvensering av tumör-DNA kan t.ex. hjälpa till att lokalisera i vilket organ en cancersjukdom ursprungligen uppstod (tumour of origin).

Det utförs flera typer av variant calling för olika syften, t.ex. för att identifiera punktmutationer, indels och kopietalsförändringar. Bioinformatikanalysen ska även kunna identifiera andra typer av genetiska förändringar, såsom fusionsgener och information om heterozygositet. Processerna för att identifiera dessa olika typer av förändring skiljer sig åt.

Variantfiltrering

I det generella avsnittet beskrivs variantfiltrering som en del av variant calling-aktiviteten. Filtrering kan dock lika gärna ses som en enskild aktivitet. För solida tumörer är det naturligare att definiera den separat och därför har den tagits upp som en specifik aktivitet i flödet för solida tumörer.

Biomarköridentifiering

Förutom de genetiska förändringar som identifieras vid variant calling, identifieras även olika behandlingsstyrande biomarkörer (mutationsmönster) som förekommer i vissa typer av cancersjukdomar.

3.3.3.4. Bearbeta RNA-sekvens med bioinformatik

Fusionsgensidentifiering

Vid variant calling av DNA identifieras viss information om fusionsgener. För att få en fullständigare bild om fusionsgener analyseras även RNA. För övrigt identifieras inga andra typer av genetiska förändringar i denna aktivitet.

3.3.4. Informationsmängder

Nedan beskrivs detaljer om informationsmängder som är specifika för solida tumörer. Informationsmängderna skapas eller används då aktiviteterna i flödet utförs. (De informationsmängder som hanteras i samtliga kartlagda sjukdomsområden sammanfattas i föregående kapitel, se avsnittet [Informationsmängder](#).)

3.3.4.1. Projektbeskrivning/forskningsinitiativ

Vid nya uppdrag finns ofta en projektbeskrivning med information om analysen avser hälso- och sjukvård eller forskning. Projektbeskrivningen specificerar även om några extra analyser behöver genomföras. Projektbeskrivningen avgör vart resultatet från bioinformatikanalysen ska levereras. Om den avser forskning ska information inte ligga kvar på sjukhusets servrar.

3.3.4.2. Provdata

För solida tumörer är tumörcellhalt ett viktigt metadata.

Viktig metadata

- tumörcellhalt: Avser patologens bedömning av andelen cancerceller i ett vävnadsprov. Visar renhet på prov (när proverna tas är de inte alltid så rena). Anges i procent. Används för att kunna identifiera kopietalsförändringar senare i bioinformatikflödet. Värdet skiljer sig åt mellan olika prover, men brukar ligga på minst 20 %. Information om tumörcellhalt för respektive prov skickas in till bioinformatikflödet via ett s.k. sample sheet.

3.3.4.3. Samtycke

Hantering av samtycke behövs troligen, ifall att det dras tillbaka i senare skede.

3.3.4.4. Sample sheet

Information som beskriver själva sekvenseringskörningen, men även associationer mellan prov och adaptrar (extra indexsekvenser som läggs till fragmenten vid bibliotekspreparation för att kunna hålla ordning på vilket prov fragmentet tillhör). Om flera prover sekvenserats tillsammans (multiplexing), behöver de delas upp i separata filer efter sekvenseringskörningen - en fil för varje prov (demultiplexing). För att kunna göra detta utgår mjukvaror från informationen i sample sheet

Nuläge

Laboratorierna använder olika typer av sample sheets som fylls i. De ser också olika ut beroende på vilket sjukdomsområde de avser. Ett och samma laboratorium kan alltså använda olika sample sheets beroende på om analysen avser t.ex. hematologi, solida tumörer eller sällsynta diagnoser.

Format: CSV (comma-separated values), annat. (För Illuminas produkter kallas filen normalt "SampleSheet.csv" och är en enkel, kommaseparerad fil.)

Målbild

Det finns behov av att harmonisera de skillnader som idag finns i sample sheet hos olika laboratorier (i de fall de ska användas i en nationell kontext). Ett ytterligare arbete behöver utföras för att utreda vilka de skillnaderna är.

Lagring: Det är oklart om sample sheet behöver lagras på NGP.

Viktig metadata

- flödescell-ID
- radnummer
- prov-ID
- indexsekvens

3.3.4.5. Sekvens

Bioinformatikflödet tar emot FASTQ-filer som indata.

Nuläge

För små genpaneler översänds (förutom inkommande FASTQ-filer) dessutom en fil som anger vilken cancertyp som ska analyseras, t.ex. lunga, kolon eller cancer. Denna behövs eftersom de små genpaneler som används lokalt i nuläget normalt är anpassade för en viss cancertyp. De bredare genpaneler som GMS håller på att ta fram ska kunna fånga information om flera cancertyper i en och samma analys, så då blir filen om cancertyp överflödig.

Lagring: Laboratorierna har olika lagringsrutiner. Eftersom data från helgenomsekvensering kräver mycket lagringsutrymme är det ibland billigare att spara det fysiska provet och köra om sekvenseringen vid behov, jämfört med att lagra FASTQ-filerna.

3.3.4.6. Kopietal

Information om identifierade kopietalsförändringar.

Nuläge

Hanteringen av kopietalsförändringar skiljer sig från identifiering av punktmutationer och indels. Identifieringen sker under variant calling-aktiviteten med flera olika mjukvaror (CNV callers). Resultaten från dessa kombineras (mergas) med hjälp av ytterligare en mjukvara. Viss annotering sker också.

Format: txt, VCF. Filer av olika format är inblandade i kopietalshanteringen. Identifierade kopietalsförändringar rapporteras i txt-format, men det sker även en hantering av filer i VCF-format för att kunna kombinera resultaten från de olika CNV caller-mjukvarorna.

Målbild

Lagring: Resultatfiler bör lagras på NGP, men det är oklart exakt vilka.

3.3.4.7. Varianter

Solida tumörer har fler artefakter att hantera jämfört med andra sjukdomsområden, eftersom de hanterar formalinfixerat material. Så många artefakter som möjligt tas bort under bioinformatikanalysen för att det ska bli enklare att tolka varianterna i den efterföljande tolkningsaktiviteten.

3.3.4.8. Fusionsgener

Information om identifierade fusionsgener.

Nuläge

Fusionsgener identifieras från DNA respektive RNA med hjälp av olika mjukvaror. Därefter genereras resultatrapporter, en från DNA- respektive en från RNA-identifieringen.

Format: tsv, txt. Filer av olika format är inblandade i hanteringen av fusionsgener. Fusionsgener som identifierats från DNA rapporteras i en txt-fil, medan de som identifierats från RNA samlas ihop i en tsv-fil.

Målbild

Lagring: Resultatfiler bör lagras på NGP, men det är oklart exakt vilka.

3.3.4.9. Identitets-SNP

Information för att förhindra provförväxling. Så kallad identitets-SNP används för providentifiering, det vill säga för att jämföra att DNA- och RNA-proverna avser samma prov.

3.3.4.10. Biomarkörer

Information om olika behandlingsstyrande biomarkörer som har identifierats.

Nuläge

I många fall är VCF-filer indata till biomarköranalyserna, men ibland är det BAM-filerna. Biomarkörer identifieras genom att analysera DNA-data och beräknas med hjälp av olika algoritmer eller mjukvaror:

- TMB (tumörmutationsbörda): Ett mått på hur stor andel förvärvade varianter det finns i en tumör. Om en person har hög tumörbörda (90%) kan immunterapi stimulera immunsystemet att bekämpa cancer bättre, jämfört med om personen har låg tumörbörda (5-10%). TMB anges i procent och beräknas med hjälp av en algoritm. Används framför allt vid behandling av lungcancer.
- MSI (mikrosatellitinstabilitet): Ett mått på hur stor andel av definierade mikrosatellitregioner som visar på instabilitet. (Varianter kan göra att mikrosatellitupprepningar av baser bildas och dessa ökar risk för cancer.) MSI anges i procent och beräknas med en mjukvara. Används framför allt vid behandling av koloncancer.

- HRD (homolog rekombinationseffekt): Ett mått på nedsatt reparationsförmåga av DNA. DNA skadas kontinuerligt men kan reparera sig självt. Varianter kan dock medföra att reparationsprocessen blir defekt. HRD anges som ett tal (HRD-score). Det beräknas utifrån kopietalsinformation och integrerar tre olika mått på genomisk instabilitet (loss of heterozygosity, telomeric allelic imbalance och large-scale state transitions). Används framför allt vid behandling av bröstcancer.

De olika TMB-, MSI- och HRD-måtten rapporteras ut i en eller flera rapporter. Hanteringen skiljer sig åt mellan olika laboratorier. I exempelvis Uppsala genereras en rapportfil per prov för varje biomarköranalys. Alla måtten kommer inte att användas i klinisk kontext initialt. De mått som ska användas kliniskt sammanställs sedan av sjukhusgenetiker i svaret.

Målbild

Det sker mycket utveckling på biomarkörområdet och många förändringar förväntas att ske, t.ex. kommer nya biomarkörer att läggas till. Allteftersom nya biomarkörer läggs till, blir det mer relevant att samla ihop de olika rapportfilerna i en enda resultatrapport.

Format: tsv, txt

Lagring: Resultatfilerna från biomarkörhanteringen bör troligen lagras på NGP.

Viktig metadata

- mått och värde för biomarkören
- vilken algoritm som användes för analysen
- vilket laboratorium värdet erhöles från

3.3.4.11. Paneler av normaler

Avser information om normalsekvenser. Innebär att sekvenser från ett antal DNA-prover från frisk vävnad väljs ut (t.ex. 20 stycken) och kompileras ihop till en referenspanel. Det skapas alltså panel med friska referenser att jämföra mot. Panelen kan sedan användas för att jämföra med tumörprover på olika sätt, t.ex. för att kunna filtrera bort artefakter.

Nuläge

Ett bioinformatikflöde kan behöva använda flera olika paneler av normaler som genereras på olika sätt, beroende på vilken kategori av variant som ska identifieras. Solida tumörer har t.ex. paneler med normaler för kopietalsförändringar och mikrosatellitinstabilitet.

Konfigurationsfilen för bioinformatikflödet behöver anpassas så att rätt paneler av normaler inkluderas i bioinformatikanalysen.

Format: tsv, txt. Filer och prover som ska användas i generering av paneler av normaler specificeras i olika typer av filer, t.ex. samples.tsv och units.tsv.

Målbild

Lagring: Dessa filer behöver troligen inte lagras på NGP.

3.3.4.12. QC-data (avser sekvensens kvalitet)

Kvalitetskontrollinformation används för felsökning och uppföljning över tid.

3.3.4.13. Samlad QC-data

Målbild

MultiQC-rapporter som genereras i nationell kontext kan behöva innehålla laboratoriespecifika kvalitetsmått. I Uppsala används t.ex. hotspot-mått som behöver användas av deras tolkningsverktyg för att granska varianter i tolkningsaktiviteten.

3.3.4.14. Tolkade varianter

Nuläge

Inom cancerområdet använder laboratorierna olika klassificeringsscheman, men en arbetsgrupp inom GMS arbetar för att ta fram standardiserade varianttolkningar. Nationellt används bland annat ACMG/AMP:s internationella klassificeringssystem för förvärvade varianter från 2017³⁸ samt belgiska riktlinjer.

Målbild

Det finns även önskemål om att dela information om tolkade varianter kopplade till enskilda cancerdiagnoser med Regionalt Cancercentrums individuella patientöversikter³⁹. Patientöversikterna är visualiseringsstöd med vilka läkare och patient kan få en översikt över en patients sjukdomsförlopp och behandling. Det vore värdefullt att dessutom visa upp genomikinformation avseende tolkade varianter där. Patientöversikterna utvecklas på INCA-plattformen (informationsnätverk för cancervården) som är en plattform för olika svenska kvalitetsregister.

3.3.5. Informationsbärare

I detta avsnitt beskrivs informationsbärare som utgör lagringsplats för genomikrelaterad information.

3.3.5.1. Variantdatabas

Uppsala planerar att skapa en lokal variantdatabas, men inte börjat utveckla någon ännu. Den kommer innehålla information om tolkade varianter.

³⁸ ACMG/AMP:s klassificeringssystem för förvärvade varianter: <https://pubmed.ncbi.nlm.nih.gov/27993330/>

³⁹ Individuell patientöversikt: <https://cancercentrum.se/samverkan/vara-uppdrag/kunskapsstyrning/patientoversikter>

3.4. Sällsynta diagnoser

GMS arbetsutskott för sällsynta diagnoser (sällsynta sjukdomar) har som målsättning att säkerställa att alla patienter i Sverige med en misstänkt ovanlig genetisk sjukdom får tillgång till den bästa och mest adekvata genetiska utredningen. Nya tekniker för gensekvensering möjliggör analys av hela arvsmassan och öppnar upp för bättre diagnostik och snabbare behandling.

3.4.1. Scenario

När helgenomsekvensering integreras i sjukvården kommer fler personer med sällsynta diagnoser få en genetisk förklaring till sina symptom. Det leder till att de snabbt får korrekt medicinsk behandling och uppföljning. Helgenomsekvensering förebygger redan idag för tidig död och svåra handikapp för personer med behandlingsbara sällsynta diagnoser.

För sällsynta diagnoser är det speciellt viktigt att kunna dela information både nationellt och internationellt, eftersom det kanske bara finns ett fåtal personer med samma genetiska diagnos och kliniska bild i landet. För att kunna hitta fler personer med varianter i samma gen och samma kliniska bild behöver fenotypegenskaper kombineras med genomikinformation samt dessa behöver kopplas till kliniska beskrivningar. Data behöver även kunna delas med andra genomikcentrum som ingår i en studie för att kunna utvärdera att alla centrum har samma kvalitet i bioinformatik och tolkning samt utvärdera bästa tillvägagångssätt vid breda genanalyser.

Arbetsutskottet kommer genomföra kliniska studier för att hitta nya sjukdomsgener och sjukdomsorsakande varianter hos personer där sällsynt diagnos misstänks men ingen sjukdomsorsakande variant har kunnat påvisas inom den kliniska utredningen. Dessutom vill man kunna identifiera patienter som ska erbjudas möjlighet att inkluderas i studier avseende nya typer av individanpassade behandlingar.

En vanligt förekommande typ av analys är trio-analys, där en persons prov samanalyseras med föräldrarnas prover.

Figur 9. Vy över uppfattat nuläge i flöde för sällsynta diagnoser

3.4.3. Aktiviteter

Nedan beskrivs detaljer kring aktiviteter som genomförs specifikt för sällsynta diagnoser. (De aktiviteter som genomförs i ett generiskt flöde sammanfattas i föregående kapitel, avsnittet [Aktiviteter](#).)

3.4.3.1. Ta prov

Prov tas i enlighet med vårdens rutiner på patienten samt ibland även på de biologiska föräldrarna, syskon eller andra släktingar för att möjliggöra jämförelse med patientens resultat. Provet är ofta ett blodprov men andra typer av prov kan också förekomma, t.ex. prov från hudbiopsi och muskelbiopsi samt vävnadsprov.

3.4.3.2. Sekvensera prov

Utförs ibland av regionen, men kan även utföras av annan leverantör. Exempel på leverantör som erbjuder sekvenseringstjänster för helgenomsekvensering och genpaneler är Blueprint Genetics och Centogene.

3.4.3.3. Bearbeta sekvens med bioinformatik

Variant calling

Vid identifiering av varianter så samkörs alla prover (t.ex. tre vid trioanalys) samtidigt och det produceras då en VCF-kolumn (eller en fil) per prov. I denna ser man varianter i förhållande till referensgenom och skillnader mellan de (t.ex. tre) jämförda proverna.

Variantannotering

Inför den manuella tolkningen i tolkningsaktiviteten görs ofta extra annoteringar avseende familjeinformation m.m.

Variantprioritering

Vid prioritering rankas varianter enligt olika rankningsmodeller (algoritmer) för att avgöra vilka varianter som enligt rankningen är mest troliga att orsaka sjukdom. Modellerna tar hänsyn till olika parametrar såsom nedärvningsmönster, sällsynthet och predikterad proteinpåverkan. Ibland utförs denna aktivitet som en del av bioinformatikflödet, men ibland utförs den i tolkningen. (Exempelvis kan tolkningsverktyget Scout utföra prioritering, åtminstone för en del av sjukdomen.)

För sällsynta diagnoser har vi tagit upp variantprioritering som en separat aktivitet eftersom det vid intervju framkommit att det är en viktig aktivitet. Aktiviteten beskrivs dock inte i rapportens generella avsnitt eftersom den bara utförs i vissa bioinformatikflöden.

3.4.3.4. Tolka och visualisera varianter

För att tolka varianter visualiseras de ofta i ett verktyg (t.ex. Alamut eller Scout) som ger en översikt av information om varianterna och kan innehålla länkar till brett använda publika databaser som exempelvis ClinVar och HGMD Professional från Qiagen. Vid tolkningen av resultatet är man intresserad av ifall en variant är känd och finns dokumenterad i dessa internationella databaser och även i lokala databaser som innehåller tolkade varianter och hur vanlig varianten är i en population (gnomAD). Information av intresse kan vara i vilken gen/transkript varianten finns, vilken biologisk process som kan påverkas på grund av detta

(UniProt), vilken typ av variant det är, hur troligt det är att det påverkar en funktion av proteinet och hur evolutionärt konserverad platsen är som påverkas av varianten. Poäng som genereras från olika verktyg (PolyPhen2, CADD, Align GVGD, SIFT och Mutation Taster) för att förutsäga funktionella effekter finns idag i Alamut. För bästa möjliga slutsats från tolkningen och för att avgöra om de funna varianterna kan vara en möjlig orsak till den observerade sjukdomsbilden behövs patientens andra information men även familjeinformation det vill säga om varianten eller sjukdom även finns hos biologiska familjemedlemmar.

3.4.3.5. Rapportera varianter till databaser

Det rekommenderas starkt att varianter ska rapporteras till databas(er). Detta möjliggör en snabbare och bättre tolkning av varianter i framtiden, både för det specifika laboratoriet men även för andra. Det är ett viktigt sätt att kunna dela data nationellt och internationellt. Databaser av intresse kan vara exempelvis ClinVar och den nationella variantdatabasen NGP.

3.4.4. Informationsmängder

Nedan beskrivs detaljer om de informationsmängder som är specifika för sällsynta diagnoser. Informationsmängderna skapas eller används då aktiviteterna i flödet utförs. (De informationsmängder som hanteras i samtliga kartlagda sjukdomsområden sammanfattas i föregående kapitel, se avsnittet [Informationsmängder](#).)

3.4.4.1. Remiss

Som beskrivits i avsnittet [Remiss](#), efterfrågas information om sammanfattande journalutdrag, foton, tillväxtkurvor, släktanamnes m.m. Av dessa är information om foton och släktanamnes speciellt viktigt för sällsynta diagnoser.

3.4.4.2. Provdataba

Målbild

Avseende familje-ID och andra informationsstrukturer för släktskap så finns intressanta strukturer i Phenopackets som även speglas i openEHR och FHIR, se referenser i avsnittet [Patientdata samt fenotypdata och annan klinisk data](#).

Viktig metadata

- familje-ID (ID som identifierar släktskap)

3.4.4.3. Analysdata

Analysdata innehåller bland annat information om vilka prover som hör ihop. Detta är speciellt viktigt för sällsynta diagnoser i fall då trioanalyser ska köras.

3.4.4.4. Släktträd

Information som beskriver familjerelationer.

Nuläge

Familjerelationer mellan prover representeras på ett strukturerat sätt i s.k. pedigree-filer, ofta i filformatet PED. Informationen används för flera syften och i olika steg av bioinformatikflödet, t.ex. för att matcha att kön stämmer överens mellan det förväntade och observerade samt vid variantprioritering för att väga släktskap (nedärvningsmönster) och sjukdomsstatus.

Format: PED, PLINK. PED-formatet kan representera grundläggande familjestrukturer (t.ex. förälder-barn-trio-relation) men inte komplexare strukturer (tvillingar, adoption, donatorer,

havandeskap, vital status, multipla fenotyper och dataproveniens) som är av nytta vid genetisk vägledning och riskbedömning. S.k. GATK-verktyg kan ta emot PED-filer som indata och använda dem för att beräkna pedigree-relaterade annoteringar, men de ska då baseras på PLINK-pedigree-filer⁴⁰.

PED är en enkel textfil med sex obligatoriska kolumner som beskriver relationer, kön och sjukdomsstatus:

- familje-ID – en unik identifierare för en familj, t.ex. F1
- individ-ID – unik identifierare för en person
- pappans ID – en unik identifierare för personen pappa
- mammans ID – en unik identifierare för personens mamma
- kön – en indikation av varje persons kön (man/kvinna/okänt)
- fenotyp – antingen en egenskap eller en sjukdomsstatus (missing, unaffected, affected)

Målbild

Vi har inte identifierat något behov av att representera stora, komplexa familjestrukturer, men om ett sådant behov skulle uppstå kan Global Alliance for Genomics and Health:s (GA4GH) standard Pedigree vara ett intressant alternativ som bör utvärderas. GA4GH utvärderar huruvida HL7 FHIR:s kärnmodeller ska utvidgas till att omfatta genomik-pedigrees.

3.4.4.5. Varianter

Vid variant calling samkörs alla prover (t.ex. tre vid trio-analys) samtidigt och det produceras då en kolumn (eller en fil) per prov i VCF-filen. I denna ser man varianter i förhållande till referensgenomet och skillnader mellan de (t.ex. tre) jämförda proverna.

Det finns standardisering för dokumentation för VCF-formatet.

3.4.4.6. Remissvar

Många av dagens patientjournalssystem saknar förmågan att ta emot remissvar från sällsynta diagnoser på ett strukturerat sätt och lagrar det huvudsakligen som fritext och PDF.

3.4.4.7. Tolkade varianter

Laboratorierna följer ACMG/AMP:s internationella klassificeringsschema för medfödda varianter från 2015⁴¹. ACMG/AMP rekommenderar att klassificera varianter utifrån hur starka evidens som finns för att en variant har sjukdomsorsakande (patogen) eller icke sjukdomsorsakande (benign) effekt. Varianterna klassificeras enligt en femgradig skala: benign, sannolikt benign, oklar konsekvens, sannolikt patogen eller patogen.

Om det föreligger evidens för att en variant har patogen eller sannolikt patogen effekt kan den rapporteras som sjukdomsorsakande i svaret från den genetiska utredningen, under förutsättning att patientens symtom stämmer överens med sjukdomsbilden. Varianter med oklar konsekvens behöver undersökas ytterligare och kan t.ex. inkluderas i en forskningsstudie för vidare analys.

Viktig metadata

- OMIM gener
- vid re-analys behövs information om vilken version av en in silico-genpanel samt vilken databas som användes vid den tidigare analysen.

⁴⁰ PLINK-pedigree-filer: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531972>

⁴¹ ACMG/AMP:s klassificeringsschema för medfödda varianter: <https://www.nature.com/articles/gim201530?ux=07df2189-4e01-4c08-8ef3-5619cff0ca61&ux2=3739b439-66b5-4bf5-921e-0916eef236a7&ux3=&uxconf=Y>

3.4.5. Informationsbärare

I detta avsnitt beskrivs informationsbärare som via API utbyter några av de beskrivna informationsmängderna samt även informationsbärare som utgör lagringsplats för genomikrelaterad information.

3.4.5.1. Internationella annoteringsdatabaser

Avser databaser som utgör lagringsplats för olika typer av annoteringar som är relevanta inom genomikområdet.

Vilka databaser som nyttjas beror på vilket verktyg som används, vissa verktyg använder samma databaser. Verktöget Alamut använder t.ex. databaserna ClinVar, Genome aggregation database (gnomAD) och Human Genome Mutation Database (HGMD) Professional vid annoteringsprocessen. Själva processen består av att:

- importera en BAM-fil till Alamut
- klassificera en variant i Alamut
- alla klassificerade varianter lagras i en lokal Alamut-databas

3.4.5.2. Beacon

Avser datakällor som har implementerat den internationella specifikationen Beacon för att utbyta information om genom och fenotyp samt annan klinisk information. Tillsammans bildar datakällorna ett nätverk i vilket deltagarorganisationerna kan ställa s.k. queries till varandra som dela information. (Se mer om Beacon version 1 och 2 i extern Bilaga E Omvärldsanalys internationella projekt och databaser.pdf, kapitel 6.1.)

Nuläge

Det nationella tolkningsverktyget Scout är anslutet till Beacon och har anpassats till version 2 av Beacon-specifikationen. Det finns förslag om att NGP borde tillgängliggöras som en internationell Beacon-resurs, men inga beslut har fattats.

Målbild

Som generell riktlinje gäller att information som tas emot från Beacon-nätverket (efter en utgående förfrågan) och som ska lagras i NGP eller vidarebefordras till t.ex. variantdatabaser, kvalitetsregister eller patientjournalssystem ska vara standardiserade. Vi har dock inte funnit att det finns något behov av att lagra eller vidarebefordra information som returneras från Beacon-nätverket.

Viktig metadata

Fenotyp och annan klinisk data (personens kön, geografiskt ursprung, etnicitet, BMI, vikt i kilogram, längd i centimeter), sjukdomskod och sjukdomsnamn, referensgenom (GRCh38, GRCh37), kromosomnummer (1-22, X, Y), muterad allel (A, C, G, T), position i genom, frekvens, variantCount, callCount, sampleCount.

3.4.5.3. Matchmaker Exchange

Avser datakällor som har implementerat Matchmaker Exchange-API:t för att utbyta information om patienter, fenotyper, varianter och gener. Datakällorna bildar ett nätverk som används för att hitta andra patienter med varianter i samma gen och samma eller liknande fenotyp (s.k. patientmatchning). Patientmatchning är speciellt viktigt för sällsynta diagnoser eftersom provstorlek är ett hinder för att upptäcka orsaker till sällsynta sjukdomar. En enda familj är ofta tillräcklig för att identifiera en eller flera kandidatvarianter, men för att fastställa vilka de sjukdomsorsakande varianterna är, behöver man utforska orelaterade fall som har en variant i samma gen och liknande fenotyper. (Se mer om Matchmaker Exchange i extern Bilaga E Omvärldsanalys internationella projekt och databaser.pdf, kapitel 6.1.)

Nuläge

I Sverige deltar Genomic Medicine Center Karolinska och Clinical Genomics i Stockholm i nätverket via noden 'PatientMatcher'⁴², det vill säga en server (kopplad till en MongoDB) som implementerat Matchmaker Exchange-API:t och således kan kommunicera med de övriga datakällorna i nätverket. Det nationella tolkningsverktyget Scout har integrerats med PatientMatcher. Via Scouts GUI kan behöriga användare rapportera in nya patientfall till nätverket samt kontrollera om andra i nätverket har upptäckt liknande patientfall.

Målbild

Det har inte framkommit något behov av att lagra eller vidarebefordra information som tas emot från Matchmaker Exchange-nätverket (efter en utgående förfrågan) till NGP, variantdatabaser, kvalitetsregister eller patientjournalssystem.

Viktig metadata

Patientens kön, HPO-termer (HPO-fenotypkod, HPO-fenotypnamn), diagnostermer (diagnoskod, diagnosnamn), gensymbol, variantattribut (kromosom, position, zygositet, de novo, loss of function).

⁴² PatientMatcher: <https://publications.scilifelab.se/publication/e517cf24bfc8442fb1064e821c3>

4. Informationsspecifikation

4.1. Bakgrund

En leverans i detta projekt är ett första utkast till en informationsspecifikation (Resultatrapport Solida Tumörer) som presenteras i detta kapitel och [Bilaga D. Informationsspecifikation - exempel](#) samt tillhörande Github-bibliotek för [solida tumörer](#)⁴³.

Syftet med att skapa informationsspecifikationer är att på (inter)nationell nivå samsas kring ett standardiserat format för att spara och hantera genomikrelaterad (meta)data strukturerat minst lika väldefinierat som sekvensdata (CRAM etc.) och varianter (VCF etc.). Detta för att öka interoperabilitet och systemoberoende data som kan delas över vårdgränser och mellan vårdinformationssystem. Det möjliggör även att smarta applikationer som exempelvis beslutsstöd kan användas där genomikdata kan kombineras med annan relevant klinisk patientdata. Standardisering av data underlättar även sekundär användning och förenklar och förbättrar möjligheterna till forskning.

Det finns data från flera steg i ett beställnings- och svarsflöde som bör standardiseras, exempel är remiss, remissvar, formulär för beställning av sekvensering (likt det exempel som presenteras i avsnittet [Patientdata samt fenotypdata och annan klinisk data](#)) samt labbsvar. I ett första steg har det inom detta projekt skapats ett utkast för att visa hur ett labbsvar (svarsrapport) för en genetisk utredning av solida tumörer kan se ut. För andra typer av resultatrapporter samt för två andra datamängder, *Provdata* respektive *Patientdata*, skulle liknande informationsspecifikationer också behöva tas fram i framtida arbete.

Som redan nämnts i rapporten påbörjades ett experiment med en standardiserad resultatrapport för sällsynta sjukdomar i juni 2021 som byggde på openEHR. Materialet som genererades under experimentet är inte färdigt för användning men finns att ta del av via Github-biblioteket för [sällsynta diagnoser](#)⁴⁴. Den informationsspecifikation som presenteras i detta kapitel kan ses som en fortsättning av detta experiment men där fokus istället har varit på solida tumörer och molekylär patologi.

4.2. Metod

I analysen för att skapa denna informationsspecifikation har projektet tagit del av exempel från Karolinska Universitetssjukhuset (nedan förkortat till "Karolinska") och flödet för genetiska utredningar av solida tumörer. Givet de processer och informationsflöden som idag gäller för remiss- och svarsflödet på Karolinska har olika exempel på standardiserade modeller utretts genom en omvärldsanalys. I denna analys har projektet funnit att det tyska informatikprojektet HiGHmed har tagit fram ett [formulärunderlag](#)⁴⁵ (en s.k. openEHR template) för en molekylär-rapport. Då HiGHmed:s rapport fångar många av de behov som har identifierats på Karolinska beslutades att den skulle användas som grund för att presentera ett exempel på hur en svarsrapport kan se ut i en svensk kontext. Det finns även många fördelar med att återanvända arbete som redan har gjorts internationellt då samsyn kring hur en standard implementeras är eftersträvänsvärt och ökar möjligheterna till framtida utbyte av data och internationella multicenterstudier.

OpenEHR stödjer flera språk och för att skapa detta utkast översattes de fåtal delar av HiGHmed:s formulärunderlag (arketyper + template) som inte redan fanns på engelska, från

⁴³ Github-bibliotek för solida tumörer: <https://github.com/genomic-medicine-sweden/Information-specifications/tree/main/result-report/solid-tumours>

⁴⁴ Github-bibliotek för sällsynta diagnoser: https://github.com/modellbibliotek/GMS_informatics_tests

⁴⁵ HiGHmed:s formulärunderlag: <https://ckm.highmed.org/ckm/templates/1246.169.236>

tyska till engelska och anpassades till behoven som identifierats hos Karolinska. I ett framtida arbete behöver arketyperna dock genomgå en mer standardiserad och djupgående översättningsprocess där engelska översättningar och termer valideras. För de informationsmängder som kommuniceras med journal- och laboratorieinformationssystem kommer även svenska översättningar behöva skapas och valideras.

En djupgående behovsanalys har inte genomförts på grund av begränsningar i arbetsresurser och tid. Angreppssättet har också varit att skapa en generisk och bred modell som fångar många behov. I dagens flöden på Karolinska skapas mycket data under processens olika steg. Den informationsmodell som presenteras här syftar delvis till att sammanställa viktiga delar av denna information samt att presentera resultatet av analysen. I dagens flöden sammanställs svaren i ett laboratorieinformationssystem till stor del i form av fritext som sedan skickas till journalsystemet där den beställande klinikern kan ta del av resultatet. Denna föreslagna informationsspecifikation tar höjd inte bara för klinikerns behov av att ta del av resultatet i beaktning utan även för möjlighet till uppföljningar, framtida beslutsstöd och sekundär användning som exempelvis forskning.

4.3. Resultat

Den informationsspecifikation som presenteras i detta kapitel och [Bilaga D. Informationsspecifikation - exempel](#) ska endast ses som ett exempel för att visa på hur openEHR kan användas för att skapa en standardiserad svarsrapport.

Arbetet kring informationsspecifikationen är ett pågående arbete. Här presenteras ett utkast med syfte att bättre visualisera hur en implementering av openEHR-baserade svarsrapporter för genomikutredningar skulle kunna se ut. Syftet är att ha en modell till utgång för diskussion och en framtida mer djupgående behovsanalys.

På grund av begränsningar av arbetstid saknar modellen idag begrepps- och kodbindningar, vilket innebär att vallistor och andra informationselement binds till standardiserade koddatabaser som exempelvis SNOMED CT. För att ytterligare standardisera formulären och öka interoperabilitet bör vallistor skapas där det är möjligt och man bör undvika att använda fritextfält där det inte är nödvändigt. Vallistor kan med fördel även begreppbindas med standarder som exempelvis SNOMED CT och/eller HGNC.

Att ta fram en nationell informationsspecifikation är en iterativ process där synpunkter, önskingar, behov och krav måste inhämtas brett. Därför ska informationsspecifikationen ses som resultatet för den första iterationen med syfte att inhämta synpunkter i ett första steg.

Informationsspecifikationen (openEHR-templat) presenteras i olika format för att illustrera dess innehåll/tolkning. I Bilaga D visas bland annat visuella presentationer av hur formulär baserade på templat skulle kunna se ut för en slutanvändare i ett användargränssnitt. Även olika varianter av scheman, beskrivningar (inklusive Excel) samt instans exempel visas.

Informationsspecifikationerna planeras, publiceras och arbetas med i GitHub-biblioteket (under [information-specifications](#)⁴⁶) och olika format av den påbörjade informationsspecifikationen för Resultatrapport Solida Tumörer samt en del data-instans exempel finns i Github-biblioteket för [solida tumörer](#)⁴⁷).

⁴⁶ Github-bibliotek för planering, publicering och arbete med informationsspecifikationerna: <https://github.com/genomic-medicine-sweden/Information-specifications>

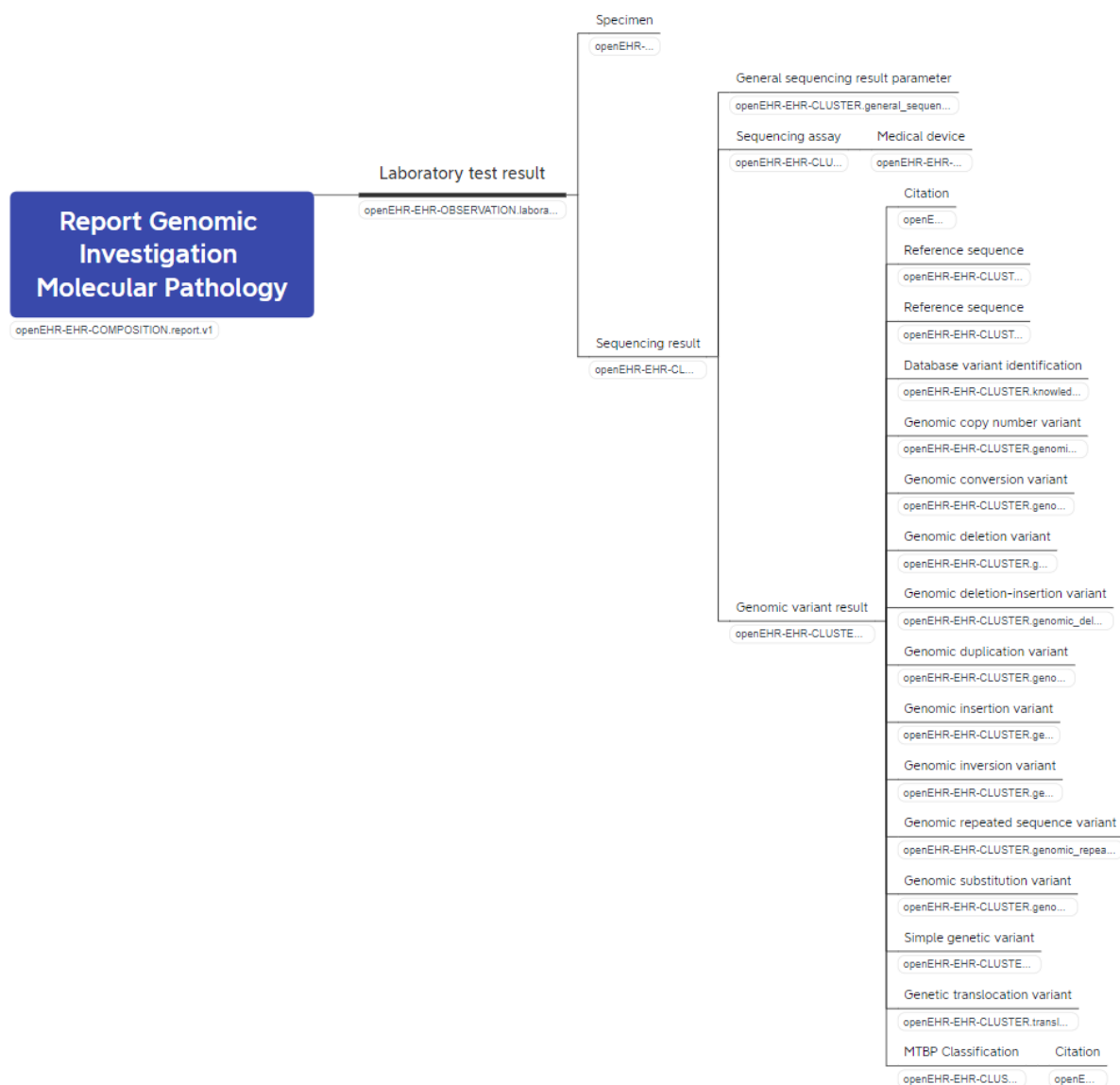
⁴⁷ Github-biblioteket för solida tumörer: <https://github.com/genomic-medicine-sweden/Information-specifications/tree/main/result-report/solid-tumours>

OpenEHR bygger på arketyper som kombineras för att skapa ett formulär (template). OpenEHR använder AQL som är ett språk för att kunna läsa/hämta informationen som sparas. Varje enskild arketyp i ett formulär har en unik sökväg vilket möjliggör att information enkelt kan hämtas utifrån ett systems eller en användares önskemål.

I formuläret har följande **arketyper** använts:

- openEHR-EHR-COMPOSITION.report.v1
 - openEHR-EHR-CLUSTER.case_identification.v0
 - openEHR-EHR-OBSERVATION.laboratory_test_result.v1
 - openEHR-EHR-CLUSTER.specimen.v1
 - openEHR-EHR-CLUSTER.sequencing_result.v0
 - openEHR-EHR-CLUSTER.general_sequencing_result_parameter.v0
 - openEHR-EHR-CLUSTER.sequencing_assay.v1
 - openEHR-EHR-CLUSTER.device.v1
 - openEHR-EHR-CLUSTER.genomic_variant_result.v1
 - openEHR-EHR-CLUSTER.reference_sequence.v1
 - openEHR-EHR-CLUSTER.knowledge_base_reference.v1
 - openEHR-EHR-CLUSTER.genomic_copy_number_variant.v1
 - openEHR-EHR-CLUSTER.genomic_conversion_variant.v1
 - openEHR-EHR-CLUSTER.genomic_deletion_variant.v1
 - openEHR-EHR-CLUSTER.genomic_deletion_insertion_variant.v1
 - openEHR-EHR-CLUSTER.genomic_duplication_variant.v1
 - openEHR-EHR-CLUSTER.genomic_insertion_variant.v1
 - openEHR-EHR-CLUSTER.genomic_inversion_variant.v1
 - openEHR-EHR-CLUSTER.genomic_repeated_sequence_variant.v1
 - openEHR-EHR-CLUSTER.genomic_substitution_variant.v1
 - openEHR-EHR-CLUSTER.simple_variant.v0
 - openEHR-EHR-CLUSTER.translocation_variant.v0
 - openEHR-EHR-CLUSTER.reference_sequence.v1
 - openEHR-EHR-CLUSTER.citation.v0
 - openEHR-EHR-CLUSTER.mtbp_classification.v0
 - openEHR-EHR-CLUSTER.citation.v0

De har kombinerats för att skapa ett formulär genom att sättas ihop enligt följande trädstruktur.



Figur 10. openEHR-arketyper som har kombinerats ihop för att skapa ett formulär (template).

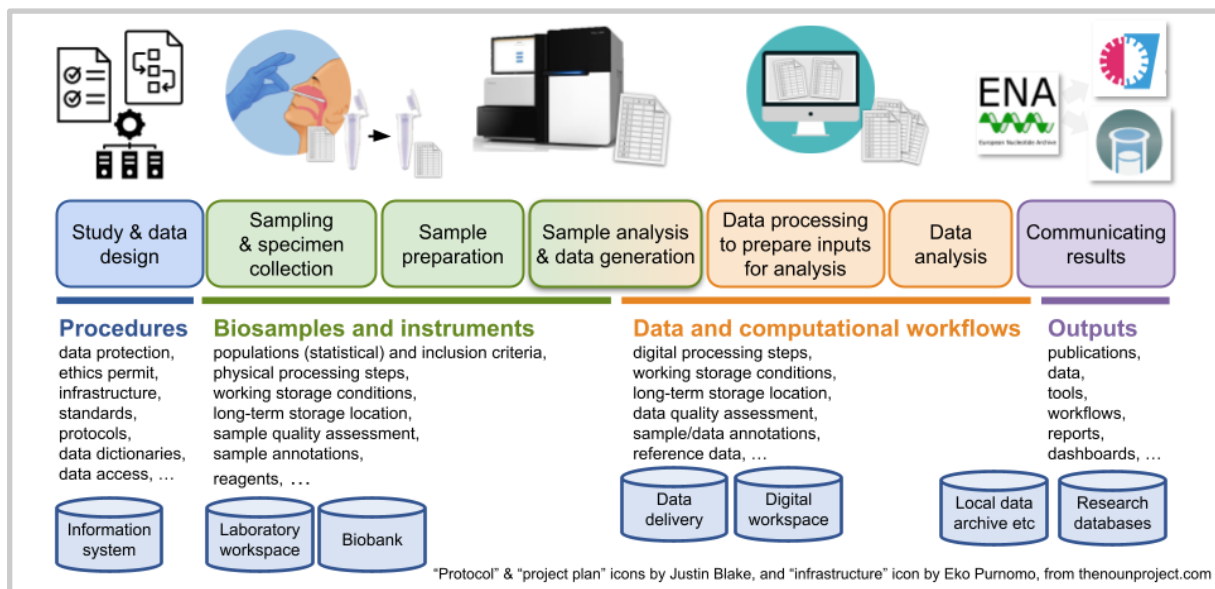
5. Rekommendationer avseende forskning

Svensk forskning genomgår en omställning till ett öppet vetenskapssystem som enligt de nationella målen ska vara genomförd fullt ut senast 2026. Omställningen omfattar bland annat en övergång mot öppen tillgång till forskningsdata för att gynna såväl forskning som ett kunskapsintensivt samhälle. Vetenskapsrådet koordinerar arbetet och har uttryckt en uppsättning vägledande principer som i det här sammanhanget kan sammanfattas till att 1) forskningsdata skapas enligt principen ”så öppet som möjligt, så stängt som nödvändigt”, 2) rutiner och incitament gör att öppen vetenskap är en naturlig del i arbetssättet, och 3) data som forskningen bygger på lever upp till FAIR-principerna för forskningsdata – det vill säga att forskningsdata är Findable, Accessible, Interoperable, och Reusable (FAIR).

Från finansiärer och vetenskapliga tidskrifter ställs också allt högre krav på god datahantering och transparens och viktiga aspekter lyfts fram i Science Europe:s så kallade baskrav. Baskraven är indelade i sex avsnitt:

1. Beskrivning av data – återanvändning av befintlig data och/eller produktion av ny data
2. Dokumentation och datakvalitet
3. Lagring och säkerhetskopiering
4. Rättsliga och etiska aspekter
5. Tillgängliggörande och långtidsbevarande
6. Ansvar och resurser

Ett praktiskt sätt att förhålla sig till de tre principerna och de sex baskraven är att ha dem i åtanke i samband med planeringen av informationsflöden så att data rutinmässigt produceras och bearbetas enligt FAIR-principerna och god datahanteringssed. Ett enkelt och generellt informationsflöde för forskningsdata kan illustreras med stadierna: Planera, Samla in, Bearbeta, Analysera, Bevara, Publicera, och Återanvända. Tillgång till dokumentation och metadata som beskriver hur och under vilka förhållanden prover och data hanteras i varje stadie är en förutsättning för att data ska kunna bli en del av ett transparent och öppet vetenskapssystem.



Figur 11. Exempel på information som beskriver hur och under vilka förhållanden prover och data hanteras i olika stadier som är en förutsättning för att data ska kunna bli en del av ett transparent och öppet vetenskapssystem.

Konventioner för hur dokumentation och metadata ska se ut och hur data ska göras tillgängliga i samband med publicering varierar mellan forskningsområden och tidskrifter. Som generellt viktiga informationsresurser finns Sveriges nationella register som är tillgängliga för forskning, Biobank Sverige som nationell infrastruktur för biobankning, och EMBL-EBI:s internationella databaser för biomolekylär forskning. Andra viktiga områdesspecifika informationsresurser beskrivs i extern Bilaga E Omvärldsanalys internationella projekt och databaser.pdf. Det är viktigt att ta de konventioner som utvecklats inom respektive område i beaktning för att förebygga och undvika de omtolkningsproblem som beskrivs i rapportens inledande avsnitt. Det är också önskvärt att data deponeras och indexerats i områdesspecifika databaser i så hög grad som möjligt för att synliggöra och tillgängliga dem för forskning.

För generell sekvensinformation, fenotypinformation och andra personuppgifter om levande eller döda personer finns Federated European Genome Phenome Archive (FEGA Sweden)⁴⁸ som erbjuder en modell där dataset med lämpliga etikgodkännande lagras inom landets gränser och delas först efter ansökan och godkännande av en för ändamålet tillsatt kommitté. För anonymiserad sekvens- och provinformation finns European Nucleotide Archive (ENA)⁴⁹ som t.ex. bör användas för information om virus, bakterier och andra mikroorganismer där spår av mänskliga genomsekvenser och personuppgifter maskeras på ett sätt som är förenligt med lagstiftning och god forskningssed. Gemensamt för båda exemplen är att sekvensinformationen kopplas till information som beskriver forskningsstudien, provet som använts och hur det förberetts, vilken utrustning och procedur som använts för sekvenseringen och hur den konfigurerats, och med vilka verktyg som data analyserats. För FEGA Sweden behövs även information om etikgodkännandet, processen för att ansöka om åtkomst, och kommittén som fattar beslut om huruvida åtkomst ska medges eller inte. De filer och dataformat som krävs för FEGA Sweden och ENA beskrivs under databasernas användardokumentation:

- [FEGA Sweden](#)⁵⁰
- [ENA](#)⁵¹

Eftersom omställningen till ett öppet vetenskapssystem är under aktiv implementering i såväl Sverige som Europa kommer det vara viktigt att följa utvecklingen i på området och särskilt de europeiska initiativ som verkar för normbildning och utveckling av god sed inom olika forskningsområden. Visioner och målbilder utvecklas i initiativ som [1+MG](#)⁵² och [EHDS](#)⁵³ samtidigt som lösningsförslag och implementationer arbetas fram i projekt som [Nordic Commons](#)⁵⁴, EJP-RD, B1MG, och GDI. Konventioner utvecklas även inom olika områden och arbetsutskott inom GMS.

Nuläge

Sekvensinformationen som produceras inom GMS synliggörs och tillgängliggörs som regel inte via EMBL-EBI:s internationella forskningsdatabaser Federated EGA och ENA. Och det kan finnas både tekniska och juridiska hinder för att dela informationen i den form som produceras idag.

⁴⁸ FEGA Sweden: <https://fega.nbis.se/>

⁴⁹ ENA: <https://www.ebi.ac.uk/ena/browser/home>

⁵⁰ Filer och dataformat som krävs för FEGA Sweden: <https://fega.nbis.se/submission/>

⁵¹ Filer och dataformat som krävs för ENA: <https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html>

⁵² European 1+Million Genomes | NBIS, <https://nbis.se/infrastructure/1plusMG-sv.html>

⁵³ EHDS | E-hälsomyndigheten, <https://www.ehalsomyndigheten.se/nyheter/2022/ehds/>

⁵⁴ Nordic Commons | NordForsk, <https://www.nordforsk.org/nordic-commons>

Målbild

De filer som krävs för att synliggöra och tillgängliga information via Federated EGA, ENA och andra områdesspecifika databaser skapas i samband med den bioinformatiska analysen så att de kan bli en del av ett transparent och öppet vetenskapssystem.

Källhänvisningar till aktuella versioner av filer och dokument som beskriver eller definierar procedurer för de olika aktiviteterna. Källhänvisningarna måste vara långsiktigt tillgängliga på ett sätt som gör att de kan användas som komplement i forskningspublikationer och i den information som publiceras tillsammans med sekvensinformationen.

En bedömning görs om vilken information som kan delas efter anonymisering, resonemanget och processen för anonymisering dokumenteras, och den anonymiserade informationen lagras på ett sätt som gör att den kan lämnas ut och göras tillgänglig via öppna forskningsresurser som ENA.

6. Samverkan

GMS samarbetar med andra nationella och internationella initiativ inom genomikområdet. Nedan beskrivs några delar av 1+ Million Genomes (1+MG) standardiseringsarbete som kan vara relevanta för GSM. Det finns även ytterligare initiativ som bör undersökas, t.ex. TK 620 Genomik och precisionsmedicin⁵⁵ och Heilsa Tryggvedottir⁵⁶, i vilka GSM deltar med representanter.

6.1. 1+MG och B1MG-initiativet

Europeiska unionen (EU) driver initiativet 1+ Million Genomes (1+MG) som ska tillgängliggöra genom för minst 1.000.000 europeiska medborgare. Det innebär att forskare och kliniker kommer få möjlighet att studera genotyp- och fenotypinformation från mer än en miljon människor.

Projektet Beyond 1 Million Genomes (B1MG) ska bidra till 1+MG genom att skapa infrastruktur, legala riktlinjer och bästa praxis för att möjliggöra åtkomst till relevant information. Det övergripande målet är att förenkla delning av human-hälsainformation inom Europa. Detta kan uppnås genom att utveckla nationella delningsnätverk och koppla ihop dem i ett internationellt nätverk där information förblir lagrad lokalt men är accessbar över Europa. Informationen kommer vara länkad så att en persons genetiska information kan matchas till personens fenotypgenskaper (t.ex. vikt, blodgrupp och medicinska historia).

Sverige är ett av 26 medlemsländer som deltar i 1+MG. GSM deltar med representanter för Sverige i ett antal 1+MG-arbetsgrupper som dessutom fungerar som expertgrupper till några av B1MG:s arbetspaket. B1MG är indelat i sex olika arbetspaket, varav 'WP3 Standards and quality guidelines' är ett. B1MG WP3 har påbörjat arbete med att ta fram gemensamma kvalitetsmått, bästa praxis vid delning och länkning av fenotyp och genetisk data samt ett ramverk för fenotyp- och klinisk metadata (se mer i avsnitten nedan). Arbetet är inte slutfört, utan nya leveranser kommer att göras.

6.1.1. Kvalitetsmått för NGS-processen

Nuläge

Det saknas riktlinjer inom Europa för vilka mått som bör användas för att säkerställa kvalitet vid helgenom- och helexomsekvensering. Kvalitetskontroll (QC) ska finnas i alla steg av NGS-processen, från provmaterial till sekvensering och bioinformatikanalys. Många laboratorier har infört NGS-metoder, men de tillämpar inte nomenklatur för kvalitetsmått på samma sätt. B1MG WP3 har identifierat vilka kvalitetsmått som används för närvarande bland medlemsländerna samt även vilka kvalitetsmått som är lämpliga att använda för att retrospektivt avgöra om historisk cancerinformation genererade från helgenomsekvensering har god kvalitet.

Framtidsbild

Medlemsländerna bör använda en gemensam minimiuppsättning av kvalitetsmått. En harmonisering av kvalitetsmått gör det enklare att säkerställa att information som utbyts inom 1+MG har sådan god kvalitet att det inte utgör någon risk för patienten att använda den. Dessutom blir det lättare att avgöra om information har tillräcklig kvalitet för att inkluderas i en studie. Laboratorier kan även förbättra sina interna processer genom att få kunskap om vilka

⁵⁵ TK620 Genomik och precisionsmedicin: <https://www.sis.se/standardutveckling/tksidor/tk600699/sistk-620/>

⁵⁶ Heilsa Tryggvedottir: <https://neic.no/heilsa/>

kvalitetsmått andra använder. Målet är att etablera kvalitetsstandarder som laboratorierna i medlemsländerna tillämpar på samma sätt.

Rekommendation: Eftersom GMS vill dela data internationellt, bör GMS gå igenom de kvalitetsmått som B1MG WP3 har identifierat hittills⁵⁷ och överväga om det är något som redan nu behöver läggas till. GMS bör även bevaka det fortsatta arbetet avseende kvalitetsmått inom B1MG WP3.

6.1.2. Bästa praxis vid delning och länkning av fenotyp- och genetisk data

Nuläge

I Europa finns genetik- och fenotypinformation i olika datakällor som använder olika taxonomi och terminologikoder för att beteckna samma sjukdomstillstånd. Det är dessutom svårt att identifiera och få åtkomst till relevant information.

B1MG WP3 har börjat kartlägga vilka beprövade tillvägagångssätt som finns bland medlemsländerna för att dela och länka fenotyp- och genetikinformation samt klassificerat dem som bästa, lovande eller innovativ praxis (lovande och innovativ praxis är exempel på tillvägagångssätt som skulle kunna utvecklas till att bli en bästa praxis)⁵⁸. I arbetet ingår även att undersöka vilka terminologier som finns implementerade i varje medlemsland. Dessa kommer beskrivas som bästa praxis för vilka terminologier som bör användas.

Framtidsbild

Medlemsländerna bör, så långt det går, använda praxis klassificerad som 'bästa' för att dela och länka fenotyp- och genetikdata. För att kunna hitta vilken genetisk och klinisk information som finns tillgänglig inom 1+MG ska det gå att söka i en datakatalog, det vill säga en webapplikation som kallas 'Accessible Genome Dashboard'. Formerna för hur intressenter ska få åtkomst till informationen är inte klara, men kan t.ex. innebära att man anmäler intresse att ta del av information till dess ägare.

Rekommendation: Det är svårt att göra konkreta rekommendationer i detta läge när all praxis inte är beskriven. GMS bör fortsätta bevaka B1MG:s arbete.

6.1.3. Ramverk för fenotyp- och klinisk metadata

Nuläge

Det saknas gemensamma riktlinjer inom Europa för att fånga och beskriva metadata. Fenotypmetadata och annan klinisk metadata beskrivs på olika sätt, är ofta ostrukturerad, spridd över olika system som inte är länkade, fångar inte en persons hela historia och registreras mestadels på det lokala språket.

B1MG WP3 håller på att ta fram ett ramverk som ska ge vägledning till vilka standarder, terminologier och verktyg som bör användas för att fånga och beskriva klinisk metadata. För att underlätta semantisk interoperabilitet ytterligare har 1+MG:s olika arbetsgrupper börjat definiera vilken klinisk metadata som minimum behöver kunna fångas och delas inom en europeisk infrastruktur. Metadatan ska avse klinisk information om humangenom/ sekvensinformation, livsstil, livskvalitet, förekomst av vissa sjukdomar och vital status. Ett första utkast finns framtaget för cancer som B1MG WP3 kommer att utgå ifrån för att genomföra en informationsmodellering. Dataelement, värdelistor, terminologibindningar till

⁵⁷ B1MG Quality metrics for sequencing: <https://zenodo.org/record/4889391#.YkPxEi3P0uV>

⁵⁸ B1MG Documented best practices in sharing and linking phenotypic and genetic data: <https://zenodo.org/record/5780228#.YkQCpy3P0uW>

koder från kodsystém etc. kommer att definieras vid modelleringen. Alltsammans kommer att mynna ut i en testbar prototyp.

Framtidsbild

Medlemslánderna bör använda en gemensam minimiuppsättning av fenotypmetadata och annan klinisk metadata. Metadata ska följa en entydig semantik och medlemslánderna ska använda gemensamma terminologistandarder så långt det är möjligt. I de fall ett medlemsland inte följer de standarder som rekommenderas av 1+MG, kan landet behöva utveckla mappningar från sin lokala standard till den internationellt rekommenderade standarden samt utvärdera om den lokala standarden stödjer översättning till lokalt språk.

B1MG WP3 kommer identifiera terminologistandarder som är lämpliga för varje domän och sedan mappa dem mot befintlig bästa praxis i medlemslánderna. (I kapitel 4.1.4 av den första versionen av ramverket⁵⁹ specificeras vilka terminologistandarder som rekommenderas initialt.)

Rekommendation: GMS bör fortsätta bevaka och invänta resultatet av B1MG:s arbete. Det är GMS avsikt att dela information med 1+MG, men det är inte klarlagt vilka IT-system som ska vara inblandade i delningen. Om det är NGP som kommer att vara Sveriges genomknod i 1+MG:s internationella nätverk, bör system som levererar information till NGP följa de terminologistandarder som rekommenderas av B1MG samt tillhandahålla ommappade värden i de fall det är relevant. (Eftersom NPU används av laboratorier på frivillig basis och olika regioner har implementerat HPO, SNOMED CT, ICD och ORPHAcodes i olika hög grad kommer ett ommappingsbehov sannolikt att uppstå.)

⁵⁹ B1MG Phenotypic and clinical metadata framework: <https://zenodo.org/record/6573854#.Y33MERSZNPY>

Bilaga A. Ordlista

I denna bilaga listas förkortningar och några av de ord som används i rapporten.

Rapporten riktar sig till intressenter som är verksamma inom genomikområdet. Vi lämnar därför endast beskrivningar för ett fåtal av de ord som förekommer i rapporten. Om en intressent inte är familjär med biologisk bakgrund, tekniker eller bioinformatik så kan utbildningsmaterial från European Bioinformatics Institute (EBI) användas för att få kunskap om ett ämne⁶⁰.

Ord, förkortning	Beskrivning
1+MG	1+ Million Genomes
ACMG	American Society of Clinical Oncology
ADL	Archetype Definition Language
AI	Artificiell intelligens
AMP	Association for Molecular Pathology
API	Application programming interface
AQL	Archetype Query Language
AU	Arbetsutskott
B1MG	Beyond One Million Genomes
BAM	Binary alignment map
BCL	Binary base call
BPMN	Business Process Model Notation
CRAM	Compressed Reference-oriented Alignment Map
CSV	Comma-Separated Values
DNA	Deoxyribonukleinsyra
EGA	European Genome Phenome Archive
EHDS	European Health Data Space (det europeiska hälsodataområdet)
EJP-RD	European Joint Programme on Rare Diseases
EMBL-EBI	European Molecular Biology Laboratory-European Bioinformatics Institute
ENA	European Nucleotide Archive
etc.	Et cetera
EU	Europeiska unionen
FHIR	Fast Healthcare Interoperability Resource
FISH	Fluorescent In Situ Hybridisering
Förvärvade varianter (somatiska)	Genetiska varianter som förvärvas under en persons livstid. De finns bara i cancerceller och ärvs inte vidare till efterföljande generationer. Kallas även somatiska varianter.
GA4GH	Global Alliance for Genomics and Health
GDI	Genomic Data Infrastructure
GDPR	General Data Protection Regulation
GMC	Genomic Medicine Center
GMS	Genomic Medicine Sweden
gnomAD	Genome aggregation database
HGNC	HUGO Gene Nomenclature Committee
HL7 FHIR	Health Level Seven Fast Healthcare Interoperability Resources
HPO	Human Phenotype Ontology
HRD	Homolog rekombinationseffekt
ICD	International Classification of Disease
IGV	Integrative Genomics Viewer
INCA	Informationsnätverk för cancervården

⁶⁰ Utbildningsmaterial EBI: <https://www.ebi.ac.uk/training/online-demand?facets=type:Online%20tutorial&page=2&query=>

JSON	Java-Script Object Notation
LIMS	Laboratory Information Management System
LIS	Laboratorieinformationssystem
Medfödd variant (germline)	Genetiska varianter som uppstår i kroppens germceller (könsceller) och som ärvs vidare till efterföljande generationer. Kallas även germline-varianter.
m.m.	Med mera
MSI	Mikrosatellitinstabilitet
NGP	Nationell genomikplattform
NGS	Next-generation sequencing
NPU	Nomenclature for Properties and Units
Nukleotid	De byggstenar som nukleinsyror är uppbyggda av
OMIM	Online Mendelian Inheritance in Man
openEHR	Open Electronic Health Record
PDF	Portable Document Format
PDL	Patientdatalagen
QC	Quality control
QCI	Qiagen Clinical Insight
RCC Väst	Regionalt Cancercentrum Väst
Read	Ett sekvenserat DNA-fragment, det vill säga en bit av en DNA-sekvens som har lästs av en sekvenseringsmaskin
RNA	Ribonukleinsyra
SAM	Sequence alignment/map format
SNOMED CT	SNOMED Clinical Terminology
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
STR	Single tandem repeat
t.ex.	Till exempel
TMB	Tumörmutationsbörda
Trioanalys	En analys där barnets (patientens) genom jämförs med föräldrarnas genom
TSV	Tab Separated Values
u.a.	Utan anmärkning
VCF	Variant Call Format
VRS	Variation Representation Specification
WGS	Whole Genome Sequencing
XML	Exstensible Markup Language

Bilaga B. Deltagarlista

I denna bilaga sammanfattas vilka personer som har deltagit vid intervjuer och i standardiseringsgrupp.

Klinisk expertis: Deltagare som har intervjuats avseende klinisk expertis.

Sjukdomsområde	Namn	Roll	Tidpunkt
Barncancer	David Gisselsson-Nord	Överläkare, Lunds universitet, GMS AU Barncancer	2022-06-21

Bioinformatik: Deltagare som har intervjuats avseende bioinformatik.

Sjukdomsområde	Namn	Roll	Tidpunkt
Barncancer	Valtteri Wirta	Avdelningschef Clinical Genomics, Karolinska Institutet, GMS AU Informatik	2022-06-07
Hematologi	Arielle Munters	Bioinformatiker, Uppsala universitet, GMS AU Informatik	2022-03-01
Infektionssjukdomar	Jyotirmoy Das	Bioinformatiker, Linköpings universitet, GMS AU Informatik	2021-10-11 2021-11-09
Infektionssjukdomar	Tanja Normark	Bioinformatiker, Karolinska Institutet, GMS AU Informatik	2021-10-11 2021-11-09
Solida tumörer	Jonas Almström	Bioinformatik, Uppsala universitet, GMS AU Informatik	2021-10-08 2021-11-12
Solida tumörer	Claes Ladenvall	Bioinformatiker och ledare för bioinformatikgruppen, Uppsala universitet, GMS AU Informatik	2021-11-12
Solida tumörer	Patrik Smeds	Bioinformatiker, Uppsala universitet, GMS AU Informatik	2021-10-08 2021-11-12
Sällsynta diagnoser	Nina Norgren	Bioinformatiker, Umeå universitet, GMS AU Informatik	Löpande under 2021

Tolkningsverktyg och variantdatabas: Deltagare som har intervjuats avseende tolkningsverktyg och variantdatabas.

Aktivitetssområde	Namn	Roll	Tidpunkt
Tolkningsverktyg	Henrik Stranneheim	Bioinformatiker, Karolinska Institutet, GMS AU Informatik	2021-12-03
Variantdatabas	Jari Häkkinen	Forskare, Lunds universitet, GMS AU Informatik	2021-08-19 2021-12-03

Informationsspecifikation: Deltagare som intervjuats vid framtagande av informationsspecifikation.

Sjukdomsområde	Namn	Roll	Tidpunkt
Sällsynta diagnoser	Nina Norgren	Bioinformatiker, Umeå universitet, GMS AU Informatik	Juni 2021


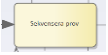



Standardiseringsgrupp: Deltagare i standardiseringsgruppen.

Namn	Roll
Ludvig Eek Hofmann	Informatiker, Region Stockholm
Ulrika Hermansson	IT-samordnare, Västra Götalandsregionen
Nina Norgren	Bioinformatiker, Umeå universitet
Wolmar Nyberg Åkerström	Data steward, Uppsala universitet
Carina Sandell	Hälsoinformatiker, Region Östergötland
Katja Stojkovic	Bioinformatiker, Umeå universitet
Erik Sundvall	Informationsarkitekt, Region Stockholm

Bilaga C. Teckenförklaring

Nedan beskrivs vad symbolerna i behovskartläggningens vyer representerar (se t.ex. avsnittet [Vy](#)).

Tabell 1. Förklaring av symboler i vyerna för flödena för sjukdomsområdena.

Symbol	Rubrik	Beskrivning
	Start och stopp	Avgränsning för beskrivning av flödet. Beskriver var det börjar och slutar. Stoppsymbolen har en kraftigare linje runt om än startsymbolen.
	Aktivitet	Aktivitet som genomförs manuellt eller automatiserat i det beskrivna flödet. Beskriver vad som genomförs.
	Prioriterad informationsmängd	Information som skapas eller används i samband med en viss aktivitet. Den lyfts fram som trolig kandidat till lagring i NGP. Det är därför av stor relevans att denna standardiseras.
	Informationsmängd	Information som skapas eller används i samband med en viss aktivitet. Beskriver det informationsinnehåll som skapas eller används.
	Lagringsyta för information eller API för informationsutbyte	Används för att visualisera att information sparas eller utbyts med ett system. Den fysiska lagringen kan ske på en eller flera platser (eller informationsutbyte ske via API till internationell tjänst). Symbolen ska endast ses som en presentation av att data samlas eller utbyts i en viss kontext.

Bilaga D. Informationsspecifikation

Bilagan innehåller formulärexempel samt information om tekniska format och verktygsstöd.

Formulärexempel Resultatrapport Solida Tumörer

På de närmast följande sidorna presenteras exempel på ett första utkast av formulär (skapat och redigerat i verktyget EHR Studio från leverantören Better) för att någorlunda pedagogiskt försöka visa vilka fält som kan ingå i informationsspecifikationen för Resultatrapport Solida Tumörer.

Skulle man vilja använda detta formulär för datainmatning bör man först arbeta vidare och optimera layouten och användbarheten, exempelvis genom formulärlogik som döljer/visar olika fält beroende på val i tidigare fält. I många fall kommer dock mycket av innehållet automatiskt registreras från filtrerade och annoterade VCF-filer m.m. snarare än fyllas i manuellt via formulär.

Report Genomic Investigation Molecular Pathology

context

Report ID

Status

Temporary

Final

Edited

Laboratory test result

Test name

Specimen - 1

+

Specimen type

Tumour

Blood

Buccal swab

Normal tissue

Liquid

Muscle

Fibroblasts

Other

Laboratory specimen identifier

Collection date/time

dd/MM/yyyy

HH : MM

Parent specimen identifier

Date/time received

dd/MM/yyyy

HH : MM

Comment

Clinical information provided

Sequencing result - 1

+

Analysis-ID/Labinternal Identifier

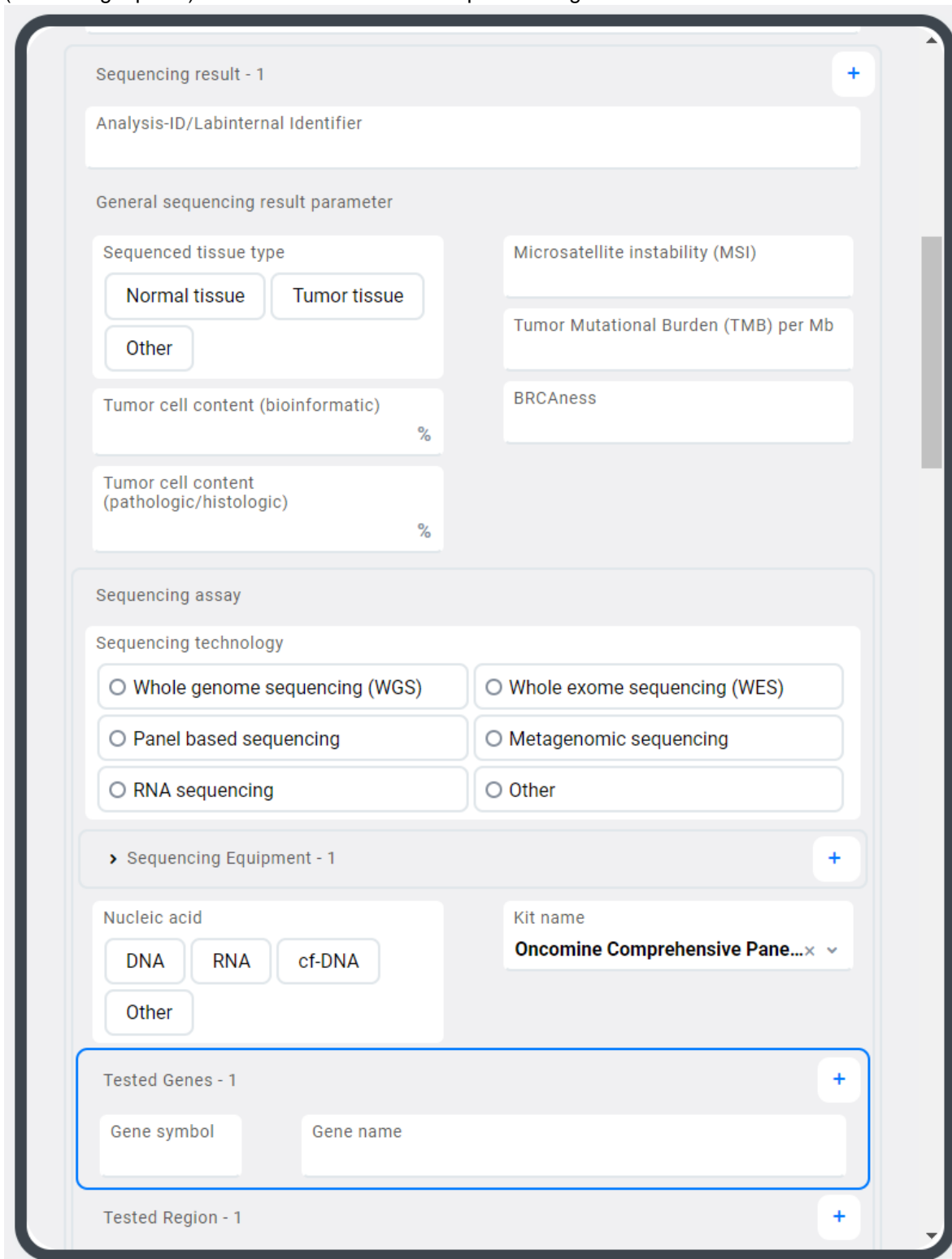
General sequencing result parameter

Sequenced tissue type

Microsatellite instability (MSI)

82

Knappen med blått plustecken innebär att denna del av informationsstrukturen kan upprepas valfritt antal gånger. Nedtill på skärmbild nedan har muspekaren förts över plustecknet för “tested genes” (i t.ex. en genpanel) och den sektion som kan repeteras ringas då in i blått.



The screenshot shows a web form for sequencing results. The form is organized into several sections, each with a title and a blue plus icon in the top right corner, indicating it can be repeated. The sections are:

- Sequencing result - 1**: Contains a text input for "Analysis-ID/Labinternal Identifier".
- General sequencing result parameter**: Contains several input fields:
 - Sequenced tissue type**: Radio buttons for "Normal tissue", "Tumor tissue", and "Other".
 - Microsatellite instability (MSI)**: A text input field.
 - Tumor Mutational Burden (TMB) per Mb**: A text input field.
 - BRCAness**: A text input field.
 - Tumor cell content (bioinformatic)**: A text input field with a percentage sign.
 - Tumor cell content (pathologic/histologic)**: A text input field with a percentage sign.
- Sequencing assay**: Contains radio buttons for "Whole genome sequencing (WGS)", "Whole exome sequencing (WES)", "Panel based sequencing", "Metagenomic sequencing", "RNA sequencing", and "Other".
- Sequencing Equipment - 1**: Contains:
 - Nucleic acid**: Radio buttons for "DNA", "RNA", "cf-DNA", and "Other".
 - Kit name**: A dropdown menu showing "Oncomine Comprehensive Pane...".
- Tested Genes - 1**: This section is highlighted with a blue border. It contains two text input fields: "Gene symbol" and "Gene name".
- Tested Region - 1**: A text input field at the bottom of the form.

Delen om “Sequencing equipment” (maskinens modell, serienummer etc.) är ihopfälld här för att spara skärmyta och kommer i regel vara förfylld från inkommande data.

Avsnittet "Genomic variant result" upprepas för varje bedömd variant som man vill ha med i rapporten. I skärmbilden visas (av pedagogiska skäl) alla varianttyper som har egna strukturerade informationsmängder. I en riktig applikation (t.ex. baserad på input från filtrerade varianter m.m.) så är det mer sannolikt att man bara visar en variant.

Tested Region - 1

Chromosomal location

Start

End

Genomic variant result - 1

Reference sequence

Reference genome assembly

Source name

URL

Database variant identification - 1

Source Name

Identification

Identification Version

Variant

Genomic

› conversion variant

Genomic

› copy number variant

Genomic

› deletion variant

Genomic

› repeated sequence variant

Genomic

› inversion variant

Genomic

› duplication variant

Genomic

› insertion variant

Genetic

› translocation variant

Genomic

› substitution variant

Genomic

› deletion-insertion variant

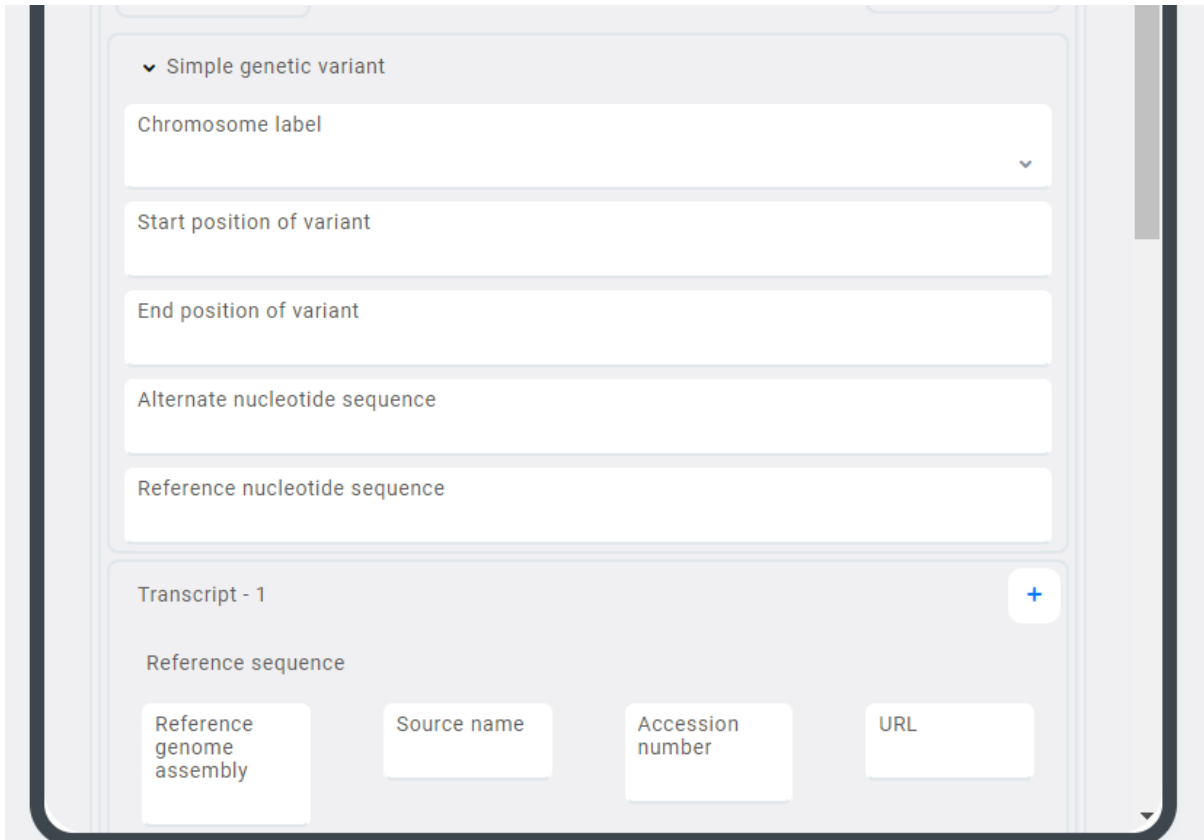
› Simple genetic variant

Transcript - 1

Om man expanderar varianttyperna från förra bilden så finns fler detaljer, här har fyra exempel expanderats även om man i en verklig applikation bara skulle visa en.

<p>Genomic ▼ conversion variant</p> <p>Start converted position</p> <p>End converted position</p> <p>Replacing sequence start position</p> <p>Replacing sequence end position</p>	<p>Genomic ▼ copy number variant</p> <p>Start</p> <p>End</p> <p>Total copy number</p> <p>Copy number change type</p> <p>Gain</p> <p>Loss</p>	<p>Genomic ► deletion variant</p> <p>Genomic ► insertion variant</p> <p>Genomic deletion- insertion variant ▼</p> <p>Start position</p> <p>End position</p> <p>Deleted sequence</p> <p>Inserted sequence</p>	<p>Genomic ▼ repeated sequence variant</p> <p>Start position</p> <p>End position</p> <p>Repeat unit - 1 +</p> <p>Repeat order</p> <p>Repeated sequence</p> <p>Copy number</p>
<p>Genomic ► inversion variant</p> <p>Genomic ► substitution variant</p>	<p>Genomic ► duplication variant</p>		<p>Genetic ► translocation variant</p>

“Simple genetic variant” är till för när man inte lyckas kategorisera in verkligheten i någon av de ovan listade strukturerade informationsstrukturena.



The screenshot shows a web form titled "Simple genetic variant" with a dropdown arrow. The form contains several input fields:

- Chromosome label (with a dropdown arrow)
- Start position of variant
- End position of variant
- Alternate nucleotide sequence
- Reference nucleotide sequence

Below these fields is a section titled "Transcript - 1" with a blue "+" button to its right. Under "Transcript - 1", there is a label "Reference sequence" and four input fields:

- Reference genome assembly
- Source name
- Accession number
- URL

DNA region name

Distance from splicing site

DNA change

RNA change

Amino Acid Change

Amino acid change type

☐ Wild type
 ☐ Deletion

☐ Duplication
 ☐ Frameshift

☐ Initiating methionine
 ☐ Insertion

☐ Insertion and deletion
 ☐ Missense

☐ Nonsense
 ☐ Silent

☐ Stop codon mutation
 ☐ Other

Predicted impact - 1

Score

1

Qualitative prediction

Functional impact - 1

Impact

Source - 1

Gene

Gen-Name (HGNC)

Full gene name

Copy number overlap

/

Part of fusion

First

Second

Fusion exon

ACMG classification

Pathogenic

Likely pathogenic

Uncertain significance

Likely benign

Benign

Part of fusion

FirstSecond

Fusion exon

Best transcript candidate

Median read depth

Allele depth

Allele frequency1

Population allele frequency1

VCF quality filter - 1

Filter passed☐

Filter name

Description

Strand bias ratio1

Strand bias p-value1

Genotype

Allelic state

Heteroplasmic

Homoplasmic

Homozygous

Heterozygous

Hemizygous

Other

Genotype quality

Genotype probability

MTBP Classification

Variant relevance

Putative functional relevant

Unknown/contradictory functional significance

Putative functionally neutral

► Evidence

Conclusion

Test diagnosis

Comment

Test request details - 1

Original test requested name

Requester order identifier

Receiver order identifier

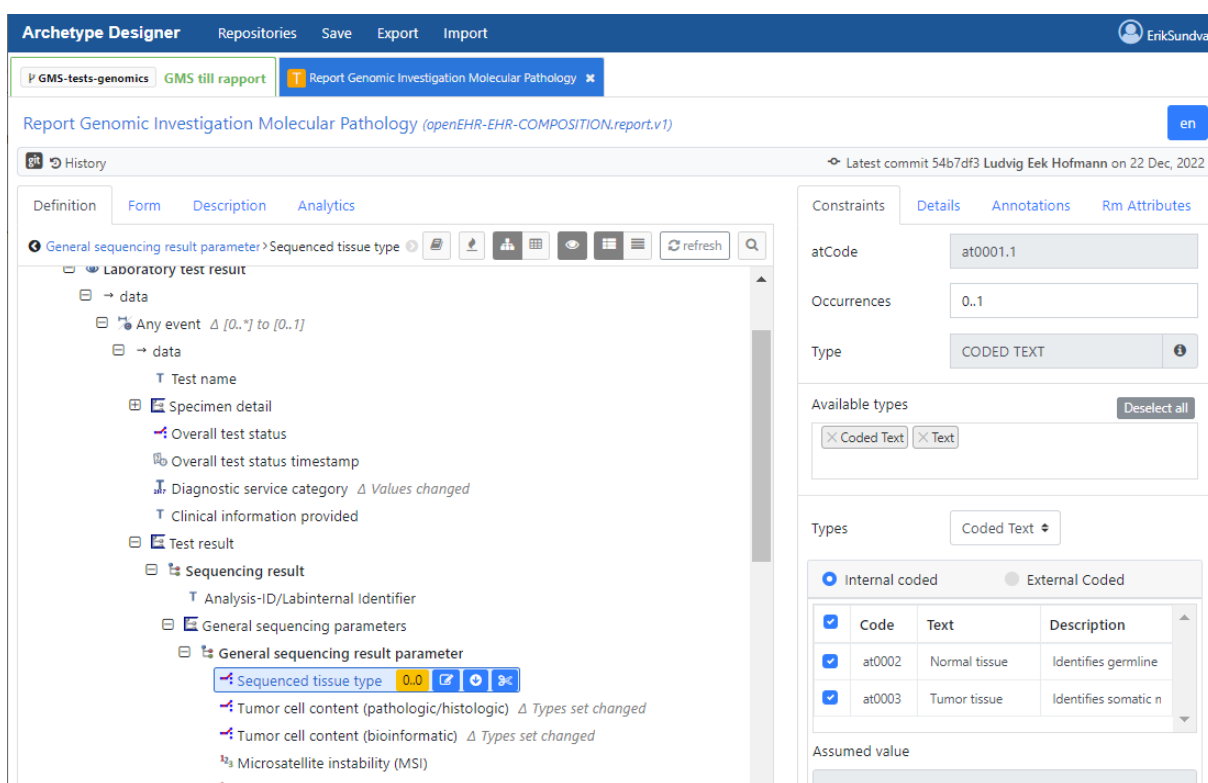
Laboratory internal identifier

Tekniska format och verktygsstöd för informationsspecifikationer (openEHR-templates)

Informationsspecifikationer i form av openEHR-arketyper och templates kan skapas i openEHR-baserade verktyg t.ex. det gratis tillgängliga [Archetype Designer](https://tools.openehr.org/designer/#/)⁶¹ och sedan exporteras i flera olika format, baserade på t.ex. XML, ADL och JSON (se [openEHR-specifikationer](https://specifications.openehr.org/)⁶²). Flera openEHR-baserade plattformar har även API för att konvertera både scheman/specifikationer och datainstanser mellan olika tekniska format.

Via <http://openehr.se/> omdirigeras man (när detta skrivs) till den svenska delen av openEHR:s diskussionsforum där det finns introduktionsinlägg som länkar till andra resurser, exempelvis en [guide för att komma igång med modelleringsverktyg](#) såsom Archetype Designer.

För att experimentera med den version av template och tillhörande arketyper som användes vid rapportskrivandet så kan man ladda ner filen '[2023-01-04-Report Genomic Investigation Molecular Pathology.zip](https://github.com/genomic-medicine-sweden/Information-specifications/raw/main/result-report/solid-tumours/2023-01-04-Report%20Genomic%20Investigation%20Molecular%20Pathology.zip)'⁶³ från Github-biblioteket för [solida tumörer](https://github.com/genomic-medicine-sweden/Information-specifications/tree/main/result-report/solid-tumours)⁶⁴ och sedan importera den i Archetype Designer. I det Github-bibliotekets readme.md-fil finns även instruktioner för att komma åt senaste version av templates m.m.



Code	Text	Description
at0002	Normal tissue	Identifies germline
at0003	Tumor tissue	Identifies somatic n

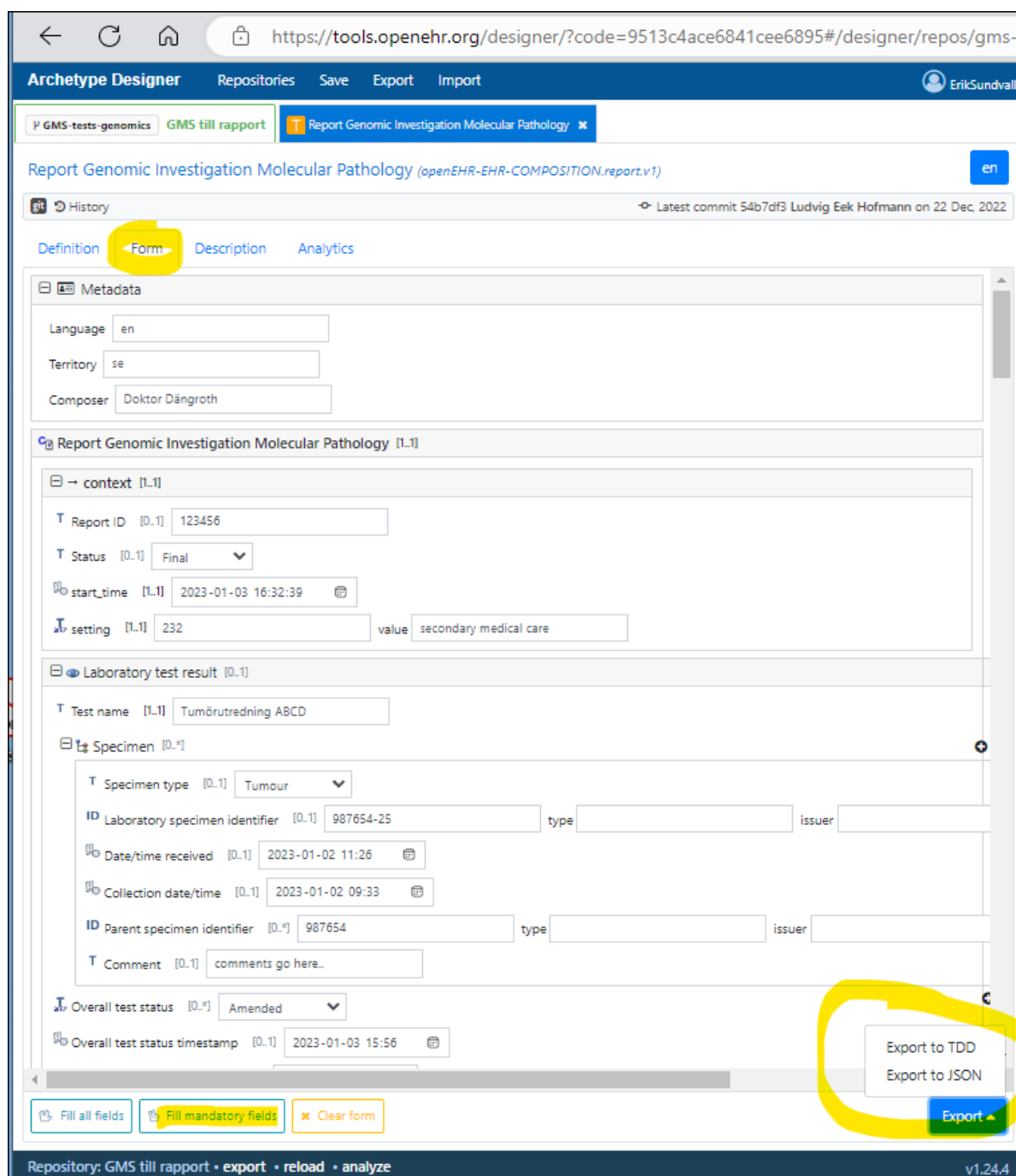
Templates kan detaljstuderas och redigeras med Archetype Designer

⁶¹ Archetype Designer: <https://tools.openehr.org/designer/#/>

⁶² openEHR-specifikationer: <https://specifications.openehr.org/>

⁶³ Report Genomic Investigation Molecular Pathology: <https://github.com/genomic-medicine-sweden/Information-specifications/raw/main/result-report/solid-tumours/2023-01-04-Report%20Genomic%20Investigation%20Molecular%20Pathology.zip>

⁶⁴ Github-bibliotek för solida tumörer: <https://github.com/genomic-medicine-sweden/Information-specifications/tree/main/result-report/solid-tumours>



I Archetype Designer finns även en inbyggd formulärvy för experimentella ändamål där man kan testa att fylla i data och exportera datainstanser i två förenklade format.

JSON-formatet man kan välja ger ett förenklat och relativt kompakt template-specifikt format som brukar kallas "structured JSON" och beskrivs mer på nästa sida.

Valet Export to TDD⁶⁵ ger instanser ett XML-format som inte är särskilt kompakt, men kan vara användbart i XML-baserade sammanhang. Båda exemplen finns [oförkortade på Github](#)⁶⁶.

⁶⁵ TDD-instanser följer ett template-specifikt TDS-schema. TDS och TDD finns beskrivna på: <https://openehr.atlassian.net/wiki/spaces/spec/pages/30408770/Template+Data+Schema+TDS+Specification+and+associated+Template+Data+Document+TDD>

⁶⁶ Github-bibliotek för solida tumörer: <https://github.com/genomic-medicine-sweden/Information-specifications/tree/main/result-report/solid-tumours>

Formulärexemplet på föregående sida kan exportera nedanstående i openEHR:s förenklade "structured" JSON-format. Formuläret och därmed JSON-filen innehåller dock en del tomma fält och fält som är fyllda med orealistisk platshållar-data. Grundstrukturen framgår dock ändå.

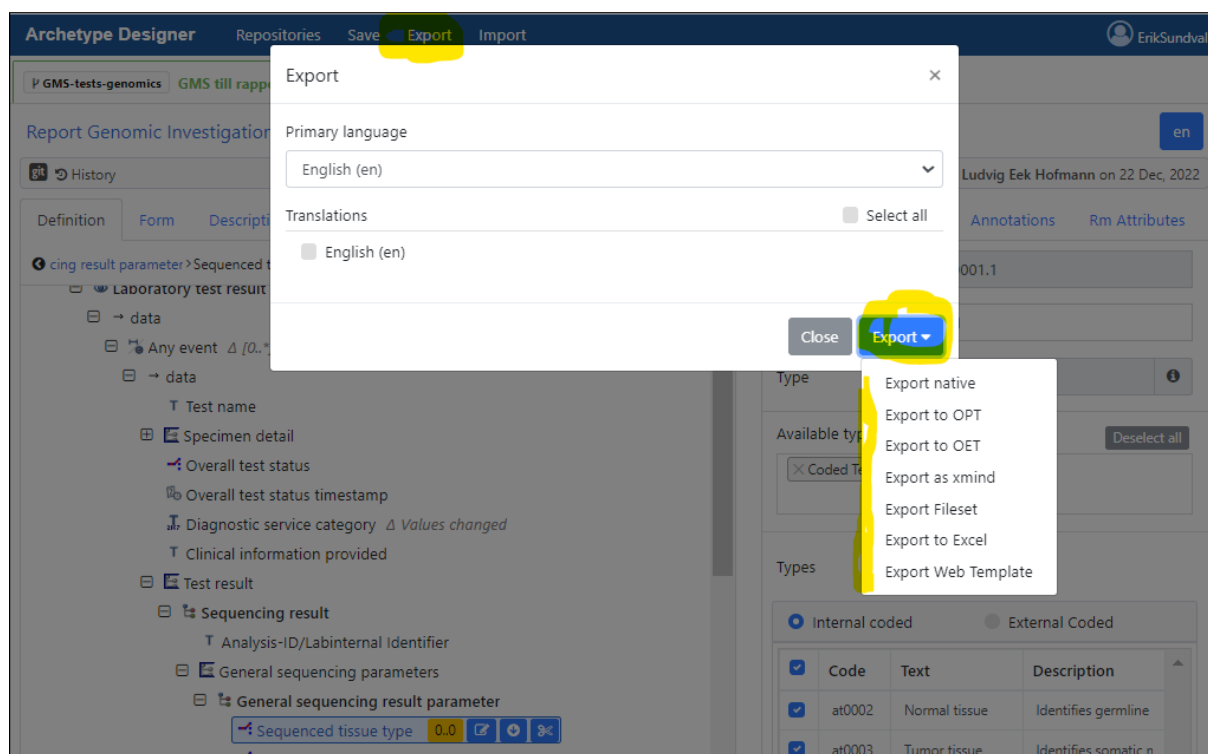
```
{
  "ctx": {
    "language": "en",
    "territory": "se",
    "composer_name": "Doktor Dängroth"
  },
  "report_genomic_investigation_molecular_pathology": [
    {
      "context": [
        {
          "report_id": [
            {
              "code": "123456"
            }
          ],
          "status": [
            {
              "code": "Final"
            }
          ],
          "start_time": [
            {
              "code": "2023-01-03T16:32:39"
            }
          ],
          "setting": [
            {
              "code": "232",
              "value": "secondary medical care"
            }
          ]
        }
      ],
      "laboratory_test_result": [
        {
          "test_name": [
            {
              "code": "Tumörutredning ABCD"
            }
          ],
          "specimen": [
            {
              "specimen_type": [
                {
                  "code": "Tumour"
                }
              ],
              "laboratory_specimen_identifier": [
                {
                  "id": "987654-25"
                }
              ],
              "date_time_received": [
                {
                  "code": "2023-01-02T11:26"
                }
              ],
              "collection_date_time": [
                {
                  "code": "2023-01-02T09:33"
                }
              ],
              "parent_specimen_identifier": [
                {
                  "id": "987654"
                }
              ],
              "comment": [
                {
                  "code": "comments go here.."
                }
              ]
            }
          ],
          "overall_test_status": [
            {
              "code": "at0040"
            }
          ],
          "overall_test_status_timestamp": [
            {
              "code": "2023-01-03T15:56"
            }
          ],
          "diagnostic_service_category": [
            {
              "code": "1236877003"
            }
          ]
        }
      ],
    }
  ],
}
```

FORTSÄTTER I NÄSTA SPALT

```
"sequencing_result": [
  {
    "analysis-id_labinternal_identifier": [
      {
        "code": "at0003"
      }
    ],
    "general_sequencing_result_parameter": [
      {
        "sequenced_tissue_type": [
          {
            "code": "at0003"
          }
        ],
        "tumor_cell_content_pathologic_histologic": [
          {
            "unit": "%"
          }
        ],
        "tumor_cell_content_bioinformatic": [
          {
            "unit": "%"
          }
        ]
      }
    ],
    "sequencing_assay": [
      {
        "sequencing_technology": [
          {
            "code": "at0012"
          }
        ],
        "kit_name": [
          {
            "code": "at0021"
          }
        ],
        "nucleic_acid": [
          {
            "code": "at0022"
          }
        ],
        "tested_genes": [
          {
            "gene_symbol": [
              {
                "code": "gene_symbol"
              }
            ]
          }
        ]
      }
    ],
    "genomic_variant_result": [
      {
        "database_variant_identification": [
          {
            "source_name": [
              {
                "code": "source_name"
              }
            ]
          }
        ],
        "genomic_deletion-insertion_variant": [
          {
            "inserted_sequence": [
              {
                "code": "inserted_sequence"
              }
            ]
          }
        ],
        "genomic_substitution_variant": [
          {
            "reference_nucleotide": [
              {
                "code": "reference_nucleotide"
              }
            ]
          }
        ]
      }
    ]
  }
],
```

...TRUNKERAT P.G.A. PLATSBRIST

Ovanstående var alltså en variant av datainstans, men även själva informationsspecifikationen eller "schemat" som definierar struktur och tillåtna värden för instanserna kan exporteras respektive dokumenteras i flera olika format. De exporterade exemplen finns [oförkortade på Github](#)⁶⁷.



- Native: Archetype Designers egna proprietära format
- OPT: Operational Template, standardiserat format som används för att importera kompletta informationsspecifikationer (inklusive ingående arketyptdetaljer) för att konfigurera journalsystem/CDR eller som grund för att bygga formulär i verktyg
- OET: ett standardiserat format som dock inte inkluderar själva arketyperna (som därför måste tillgängliggöras på annat vis)
- xmind: mindmap som dokumenterar templatens struktur
- fileset: zip.fil med ingående arketyper som separata filer plus OPT och Native
- Excel: Excel-ark som dokumenterar templatens struktur och AQL-sökvägar
- Web template är ett utvecklarvänligt exportformat för openEHR-templates (kan jämföras med scheman) som inte kräver djup kunskap om openEHR:s inre mekanismer, se exempel på nästa sida samt beskrivning av web-template och av det förenklade template-specifika data-instans-formatet "flat-json" på t.ex. https://ehrbase.readthedocs.io/en/latest/09_flat/00_description/index.html#web-template Från web.templates kan både "structured" JSON-formatet (se föregående sida) och "flat" JSON-formatet enkelt konstrueras eller valideras.

För intern lagring inuti NGP eller i en openEHR-CDR som skulle kunna bli en del av (eller komplement till) NGP kan även openEHR:s mer generella "kanoniska" template-oberoende format bli aktuella att användas men för överföring *till* NGP kommer sannolikt de förenklade template-specifika formaten ("structured" och/eller "flat") föredras av bioinformatiker och utvecklare.

⁶⁷ Github-bibliotek för solida tumörer: <https://github.com/genomic-medicine-sweden/Information-specifications/tree/main/result-report/solid-tumours>

Nedan finns början av web-templaten (strukturbeskrivning och valideringsregler som utgör dataschema) för samma template/informationsspecifikation som använts i bilagans övriga figurer och tabeller. **Rödfärgade rader** påverkar strukturnivå och **gulmarkerade id-rader** används som identifierare (keys) i t.ex. "structured" instans-exemplet i föregående tabell och skulle kunna användas för att även producera sökvägar i "flat"-json-formatet. **Lilafärgade rader** definierar tillåtna val i en flervälslista (i det fall avseende rapportens status)

```
{
  "templateId": "Report Genomic Investigation Molecular Pathology",
  "semVer": "0.1.2-alpha.5",
  "version": "2.3",
  "defaultLanguage": "en",
  "languages": [ "en" ],
  "tree": {
    "id": "report_genomic_investigation_molecular_pathology",
    "name": "Report Genomic Investigation Molecular Pathology",
    "localizedName": "Report Genomic Investigation Molecular Pathology",
    "rmType": "COMPOSITION",
    "nodeId": "openEHR-EHR-COMPOSITION.report.v1",
    "min": 1,
    "max": 1,
    "localizedNames": {
      "en": "Report Genomic Investigation Molecular Pathology"
    },
    "localizedDescriptions": {
      "en": "Document to communicate information to others, commonly in response to a request from another party."
    },
    "aqlPath": "",
    "children": [ {
      "id": "context",
      "rmType": "EVENT_CONTEXT",
      "nodeId": "",
      "min": 1,
      "max": 1,
      "aqlPath": "/context",
      "children": [ {
        "id": "report_id",
        "name": "Report ID",
        "localizedName": "Report ID",
        "rmType": "DV_TEXT",
        "nodeId": "at0002",
        "min": 0,
        "max": 1,
        "localizedNames": {
          "en": "Report ID"
        },
        "localizedDescriptions": {
          "en": "Identification information about the report."
        },
        "aqlPath": "/context/other_context[at0001]/items[at0002]/value",
        "inputs": [ {
          "type": "TEXT"
        } ]
      } ],
      "id": "status",
      "name": "Status",
      "localizedName": "Status",
      "rmType": "DV_TEXT",
      "nodeId": "at0005",
      "min": 0,
      "max": 1,
      "localizedNames": {
        "en": "Status"
      },
      "localizedDescriptions": {
        "en": "The status of the entire report. Note: This is not the status of any of the report components."
      }
    } ],
    "aqlPath": "/context/other_context[at0001]/items[at0005]/value",
    "inputs": [ {
      "type": "TEXT",
      "list": [ {
        "value": "Temporary",
        "label": "Temporary"
      }, {
        "value": "Final",
        "label": "Final"
      }, {
        "value": "Edited",
        "label": "Edited"
      } ],
      "listOpen": false
    } ],
    "id": "start_time",
    "name": "Start time",
    "rmType": "DV_DATE_TIME",
    "min": 1,
    "max": 1,
    "aqlPath": "/context/start_time",
    "inputs": [ {
      "type": "DATETIME"
    } ],
    "inContext": true
  }, {
    "id": "setting",
    "name": "Setting",
    "rmType": "DV_CODED_TEXT",
    "min": 1,
    "max": 1,
    "aqlPath": "/context/setting",
    "inputs": [ {
      "suffix": "code",
      "type": "TEXT"
    }, {
      "suffix": "value",
      "type": "TEXT"
    } ],
    "inContext": true
  } ],
  "id": "laboratory_test_result",
  "name": "Laboratory test result",
  "localizedName": "Laboratory test result",
  "rmType": "OBSERVATION",
  "nodeId": "openEHR-EHR-OBSERVATION.laboratory_test_result.v1",
  "min": 0,
  "max": 1,
  "localizedNames": {
    "en": "Laboratory test result"
  },
  "localizedDescriptions": {
    "en": "The result, including findings and the laboratory's interpretation, of an investigation performed on specimens collected from an individual or related to that individual."
  },
  "aqlPath": "/content[openEHR-EHR-OBSERVATION.laboratory_test_result.v1]",
  "children": [ {
    "id": "test_name",
    "name": "Test name",
```

FORTSÄTTER I NÄSTA SPALT

...TRUNKERAT P.G.A. PLATSBRIST

Nedan visas de första raderna (av 184) i Excel-baserade dokumentationen som automatgenereras från templatens av Archetype Designer.

Report Genomic Investigation Molecular Pathology							
Archetypes included (Container archetype)	Dep	Text	Description	Data Type	Occurrences	Comment / Suggestion	Path
openEHR-EHR-COMPOSITION.report.v1	1	Report Genomic Investigation Molecular Pathology	Document to communicate information to others, commonly in response to a request from another party.	ARCHETYPE / COMPOSITION	Mandatory		/
		Report ID	Identification information about the report.	TEXT	Optional		/context/other_context[at0001]/items[at0002]
		Status	The status of the entire report. Note: This is not the status of any of the report components.	TEXT Temporary Final Edited	Optional		/context/other_context[at0001]/items[at0005]
		Extension	Additional information required to capture local context or to align with other reference models/formalisms.	SLOT / CLUSTER	Optional, repeating	For example: local information requirements or additional metadata to align with FHIR or CIMI	/context/other_context[at0001]/items[at0006]
openEHR-EHR-CLUSTER.case_identification.v0 (openEHR-EHR-COMPOSITION.report.v1)	1.1	Case identification	To record case identification details for public health purposes.	ARCHETYPE / CLUSTER	Prohibited		/context/other_context[at0001]/items[openEHR-EHR-CLUSTER.case_identification.v0, 'Case identification']
		Case identifier	The identifier of this case.	TEXT	Mandatory		/context/other_context[at0001]/items[openEHR-EHR-CLUSTER.case_identification.v0, 'Case identification']/items[at0001]
openEHR-EHR-OBSERVATION.laboratory_test_result.v1 (openEHR-EHR-COMPOSITION.report.v1)	1.1	Laboratory test result	The result, including findings and the laboratory's interpretation, of an investigation performed on specimens collected from an individual or related to	ARCHETYPE / OBSERVATION	Optional		/content[openEHR-EHR-OBSERVATION.laboratory_test_result.v1, 'Laboratory test result']
		Test name	Name of the laboratory investigation performed on the specimen(s).	TEXT	Mandatory	A test result may be for a single analyte, or a group of items, including panel tests. It is strongly recommended that 'Test name' be coded with a terminology, for example LOINC or SNOMED CT. For example: 'Glucose', 'Urea and Electrolytes', 'Swab', 'Cortisol (am)', 'Potassium in perspiration' or 'Melanoma histopathology'. The name may sometimes include specimen type and patient state. For	/content[openEHR-EHR-OBSERVATION.laboratory_test_result.v1, 'Laboratory test result']/data[at0001]/events[at0002]/data[at0003]/items[at0005]
		Specimen detail	Details about the physical substance that has been analysed.	SLOT / CLUSTER	Optional, repeating	If the specimen type is sufficiently specified with a code in the Test name, then this additional data is not required. Linking results to specific specimens may be recorded using 'Specimen identifier' elements in both the CLUSTER.specimen and the	/content[openEHR-EHR-OBSERVATION.laboratory_test_result.v1, 'Laboratory test result']/data[at0001]/events[at0002]/data[at0003]/items[at0065]
		Overall test status	The status of the laboratory test result as a whole.	CODED TEXT: local::at0107::Registered - The existence of the test is registered	Optional, repeating	The values have been specifically chosen to match those in the HL7 FHIR Diagnostic report, historically derived from HL7v2 practice. Other local codes/terms can be used via the Text 'choice'.	/content[openEHR-EHR-OBSERVATION.laboratory_test_result.v1, 'Laboratory test result']/data[at0001]/events[at0002]/data[at0003]/items[at0073]