Projektrapport:

Omvärldsanalys internationella projekt och databaser

Projekt: SWElife Skalbara informatiklösningar

Målgrupp: AU Informatik

Datum: 2022-12-22

Version: 1.1

Sammanfattning

Denna omvärldsanalys ingår som underlag för kommande arbete med framtagande av variantdatabaser, relaterade verktyg och integrationer samt arbete med metadata inom Genomic Medicine Sweden (GMS).

Denna version av rapporten fokuserar på databaser och verktyg för sällsynta diagnoser samt för cancer (somatiskt respektive hereditärt). Ytterligare analys kan behövas för andra områden. Även mer generella internationella projekt, terminologisystem, ontologier, standarder och specifikationer av potentiellt intresse exemplifieras.

Fokus har varit att främst lista upp och grovt beskriva och kategorisera delar i respektive databas/verktyg/standard/specifikation/projekt och påvisa samband/länkar mellan flera av dem. För djupdykning hänvisas istället till respektive originalkälla. Förhoppningsvis räcker materialet som start för att hitta, nyttja och bidra till befintliga internationella samarbeten samt vid behov bygga nya nationella strukturer, databaser, verktyg och samarbeten inom GMS.

Ytterligare exempel och GMS-områden kan behöva läggas till i framtida rapportversioner. Vi föreslår att rapporten hålls som ett levande dokument i GMS och uppdateras av de som känner till eller som under projektets gång lär känna fler exempel eller har/får djupare kunskaper. (En Word-version av denna rapport kommer att levereras till beställaren.)

Ett tänkbart förslag till framtida arbete är att ur de listade källorna analysera och syntetisera en sammanställning över vanligt förekommande typer/kategorier av variabler och metadata för sällsynta diagnoser, cancer somatiskt respektive cancer hereditärt samt för sådant som kan vara gemensamt för alla områden utifall det behövs för relaterat arbete inom GMS.

Revisionshistorik

Dokumentet har ändrats enligt följande.

Datum	Version nr	Kommentar	Reviderad av
2021-03-02	1.0	Första version	Ulrika Hermansson
2022-12-22	1.1	Lagt till information om Beacon, Matchmaker Exchange, NPU och VRS. Uppdaterat länkar.	Ulrika Hermansson

Innehåll

1. Bakgrund

GMS kliniska arbetspaket har gett projektet Skalbara informatiklösningar i uppdrag att utföra en omvärldsanalys över vilken metadata som ingår i och vilka standarder som används av internationella projekt och kliniska variantdatabaser för sällsynta diagnoser samt cancer. Kontaktpersoner hos de kliniska arbetspaketen är Anders Edsjö (Region Skåne) och Thoas Fioretos (Lunds universitet).

Omvärldsanalysen ska utgöra ett underlag inför kommande arbete med framtagande av variantdatabaser och arbete med metadata inom GMS.

1.1. Genomförande

I arbetsgruppen deltar Ulrika Hermansson (Västra Götalandsregionen), Ingrid Jakobsen (Universitetssjukhuset Örebro), Nina Norgren (Umeå universitet, Region Västerbotten) och Erik Sundvall (Region Östergötland, Linköpings universitet).

Arbetsgruppen har genomfört uppdraget huvudsakligen genom att inhämta information från officiella hemsidor, vetenskapliga publikationer m.m.

2. Internationella databaser sällsynta diagnoser

De databaser som är av intresse inom sjukdomsområde sällsynta diagnoser framgår i detta kapitel. Ytterligare information om metadata och datastruktur finns i Bilaga A. Information i internationella databaser och projekt.

2.1. ClinVar

ClinVar är en databas som sammankopplar data mellan humanvarianter och fenotypdata. Den är ett fritt tillgängligt publikt arkiv där forskare/kliniker har möjlighet att ladda upp data om upptäckta varianter som hittas i patientprover, tillsammans med stödjande bevis för dess kliniska signifikans.

Då varje post i databasen består av ett variant/fenotyp-par, kan samma variant förekomma i flera poster, om den är rapporterad för flera olika fenotyper. Standardiserade inputformat underlättar för sökning av genomiska positioner, varianter, fenotyper etc. Kliniker och forskare kan använda denna information vid utvärdering av potentiellt kausala varianter de stöter på.

Standard/specifikation	Kommentar
SNOMED CT, GeneReviews, Genetic Home Reference, Office of Rare Disease, MeSH, OMIM	Medicinska tillstånd
HPO, OMIM och andra resurser	Associerade egenskaper
HGVS. Genomiska sekvenser representeras i RefSeqGene/LRG-koordinater samt även i ”locations” (som t.ex. GRCh37/hg19, GRCh38/hg38)	Varianter

2.2. DECIPHER

DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources (DECIPHER) är en interaktiv webbaserad databas som inkorporerar verktyg designade för att vara behjälpliga i tolkningen av sällsynta genvarianter. DECIPHER förstärker den kliniska genetiska diagnostiken av sällsynta ärftliga sjukdomar genom att inhämta information från olika bioinformatikresurser relevanta för en variant som hittats i patienten.

Bidragande till databasen är ett internationellt nätverk av akademiska institutioner för klinisk genetik och genomik vid sällsynta diagnoser, som nu inkluderar mer än 270 centrum med hittills mer än 36.000 uppladdade fall. Varje bidragande centrum har en ansvarig kliniker/genetiker som överser dataöverföring och datadelning. Varje centrum äger kontrollen över sin data och huruvida den vid medgivande delas med valda medlemmar i en samarbetsgrupp eller tillgängliggörs öppet i anonymiserad form via Ensembl och andra genombrowsers. Vid delning kan medlemmar få tillgång till patientrapporter och kontakta varandra för att diskutera patienter av gemensamt intresse.

Standard/specifikation	Kommentar
HGNC, HGVS, GRCh38	Används för information om gener/varianter. ● Primär assembly GRCh38 sedan december 2020 ● Sökning kan ske på GRCh37/hg19 och patientdata som registreras med GRCh37/hg19 konverteras till GRCh38 av DECIPHER med hjälp av UCSC LiftOver tool. Ursprungligt angiven såväl som konverterad position anges i browserfunktionerna i förekommande fall.
HPO, OMIM, OMIM Morbid, GeneReview, ClinGen och andra källor	Används för fenotyper och associerade egenskaper.
ACMG, ClinGen	Används för klinisk klassificering. Ej obligatorisk vid deponering av data, annotering av patogenicitet och bidragandegrad (contribution) görs av patientdataägaren. Följer i förekommande fall: ● ACMG-standarder och riktlinjer ● The ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) Bayesian classification framework

2.3. HGMD

Human Gene Mutation Database (HGMD) tillhandahålls av Cardiff University. Akademisk licens är gratis, professionell licens fås via Qiagen. Deras mål är att samla alla kända genmutationer som orsakar nedärvda sjukdomar i människa. Den här databasen innehåller de första exemplen på alla mutationer som orsakar eller är associerade med nedärvd sjukdom, plus sjukdomsassocierade/funktionella polymorfier rapporterade i litteraturen. Somatiska och mitokondriella varianter är ej inkluderade. Varianter som saknar uppenbara fenotypiska konsekvenser är normalt sett ej heller inkluderade.

2.4. gnomAD

The Genome Aggregation Database (gnomAD) innehåller sekvensdata från exom- samt helgenomsekvensering och all aggregerad data är fritt tillgänglig. Version 3.1 av databasen innehåller mer än 76.000 helgenom, som mappats mot referensgenomet GRCh38. Version 2 av databasen innehåller 125.000 exom samt 15.000 helgenom och är mappad mot GRCh37-referensen. Version 2.1 är en strukturell variantdatabas och innehåller data från 10.000 helgenom mappade mot GRCh37-referensen. Alla personer med känd allvarlig sjukdom har tagits bort från databasen. All data har blivit ombearbetad med deras egna pipelines och en joint-calling har gjorts. Sökningar i den här databasen kan göras baserat på variant eller genomisk position.

2.5. ClinGen

Clinical Genome Resource (ClinGen) är ett program finansierat av National Institute of Health (NIH). Det är dedikerat till att bygga en central resurs som definierar den kliniska relevansen hos gener och varianter, för användning inom precisionsmedicin och forskning.

ClinGen jobbar med kurering av gener och varianter som blivit inskickade av laboratorier och kliniker från hela världen. Det som bedöms är följande:

- Är denna gen associerad med sjukdom, och via vilka mekanismer orsakar den sjukdom?
- Är den här varianten kausal?
- Kommer denna information påverka den medicinska behandlingen?

När kureringen är färdig publiceras resultaten fritt för allmänheten. Kurering görs inom följande områden:

- Gen-Sjukdom validitet. Genetisk data utvärderas genom litteraturstudier för att identifiera gener där patogena varianter orsakar sjukdom.
- Doskänslighet. Data samlas gällande haploinsufficiens.
- Patogenicitet. Kombinerar klinisk data, genetisk data, populationsdata, och resultat från funktionella studier med expertutlåtanden, för att klassificera varianter enligt ACMG/AMP-riktlinjer.
- Påverkan på klinisk behandling. Utvärdera tillgänglighet på effektiva behandlingar.

ClinGen har ett nära samarbete med ClinVar och deras databas. ClinGen använder sig av ClinGen Data Exchange för lagring av all gen- och variantdata, som också är integrerad med ClinVar.

Standard/specifikation	Kommentar
Allele model	Används för att länka samman varianter från olika databaser med olika ID:n och ger dem ett gemensamt ID
SEPIO	Monarch Initiative Scientific Evidence Provenance Information Ontology
GA4GH, HL7	ClinGen arbetar aktivt med dessa initiativ

2.6. dbVar

Database of human genomic structural Variation (dbVar) är en databas över humana strukturella genomvarianter (SV) med mer än 50 baspars längd där användare kan söka, visualisera och ladda ned data från inkomna studier. Sedan tidigare finns även icke-human data tillgänglig via FTP, men från november 2017 registreras och supportas dock endast humandata.

dbVar utvecklas och underhålls av National Center for Biotechnology Information (NCBI). Den ger tillgång till rådata i förekommande fall såväl som länkar till ytterligare resurser från NCBI och andra källor. Möjlighet finns att via sökfunktion eller studie-/grafiska genombrowsern visualisera och ladda ned data från över 150 studier, exempelvis 1000 Genomes Phase 3, Genome in a Bottle, Clinical Structural Variants, gnomAD med flera. Bulkdata kan laddas ned via FTP. 2018 introducerades ett nytt omfattande dataset av non-redundant SV (NR set) bestående av unika insertioner, duplikationer och deletioner. Datasetet uppdateras månatligt. Dessa filer är lämpliga att använda som referenser vid analys av human strukturell variation.

Standard/specifikation	Kommentar
GRCh38, GRCh37, NCBI36, HGVS, HGNC, SO	Gener/varianter Genome assemblies (submitter provided); GRCh38, GRCh37/Hg19, NCBI36 (hg18). Ommappning till nuvarande assembly med hjälp av NCBI Remap Service i förekommande fall. HGVS, HGNC, Sequence Ontology (SO).
HPO, MeSH, OMIM med flera källor	Fenotyper och associerade egenskaper.
Endast indirekt via länk mot ClinVar och OMIM	Klinisk relevans/patogenicitet.

3. Internationella databaser cancer somatiskt

De databaser som är av intresse inom sjukdomsområde cancer somatiskt, framgår i detta kapitel. Information om metadata finns i Bilaga A. Information i internationella databaser och projekt.

3.1. TCGA

The Cancer Genome Atlas (TCGA) var ett program som syftade till att karakterisera över 20.000 cancerprover, fördelade över 33 olika cancertyper. Programmet började 2006 som ett samarbete mellan the National Cancer Institute (NCI) och the National Human Genome Research Institute. Numera finns all data samplat från detta projekt i GDC data portal, som beskrivs mer utförligt i senare kapitel.

3.2. COSMIC

Catalogue Of Somatic Mutations In Cancer (COSMIC) erbjuder baserat på omfattande databaser ett antal verktyg för att utforska effekterna av somatiska mutationer i cancer hos människor. Webbplatsen baserar sig på två separata men relaterade delar:

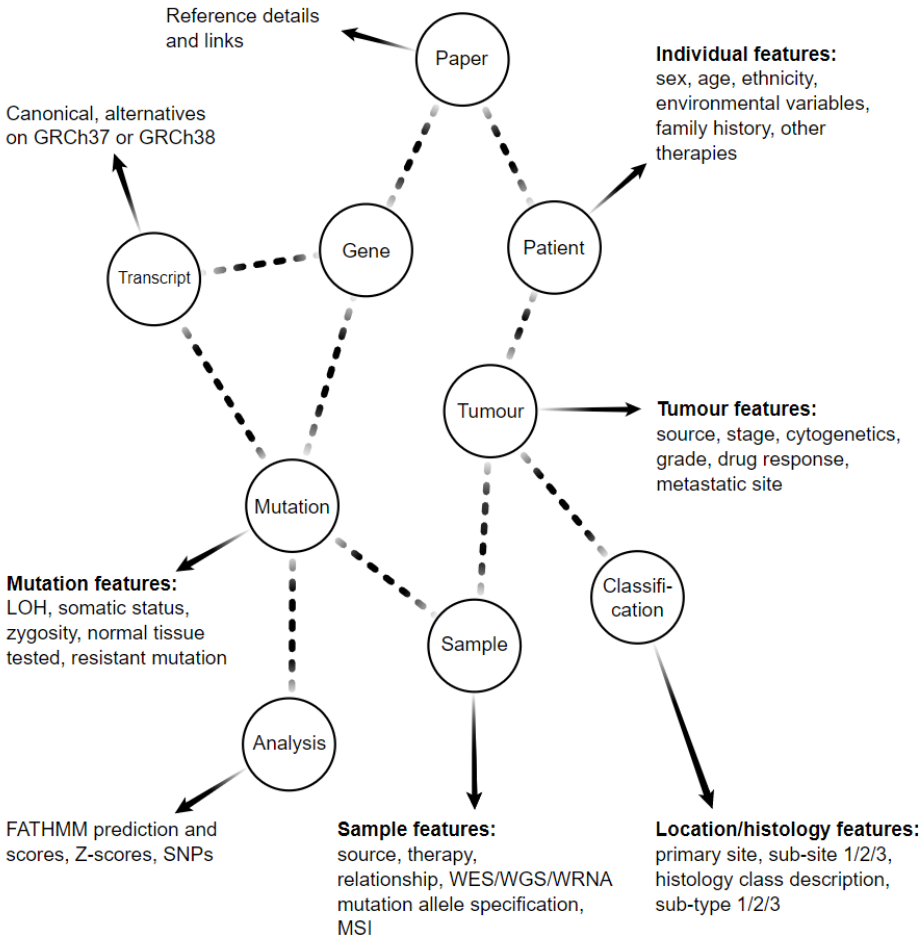
- COSMIC, den huvudsakliga databasen med somatiska mutationer. Uppdateras kvartalsvis. Datakällorna är både baserade på paneler (targeted gene-screening panels) via över 27.000 expertgranskade vetenskapliga artiklar och screeningdata (genome-wide screen data) från över 37.000 genom, inklusive från andra databaser som TCGA och ICGC.
- Cell Lines Project, mutationsprofiler (full exome sequencing) för över 1.000 cellinjer som används i cancerforskning.

Utöver att ladda ner filer med kategoriserad data så kan man via verktyg utforska innehållet, t.ex. visas bland annat histogram över hur vanliga olika typer av mutationer är. Användaren kan söka, navigera och zooma i hela genomet samt slå av/på olika spår och även söka fram detaljer för de enskilda varianter och prover som bygger upp histogrammen. Det finns även verktyg för att visa mutationer som kan ge resistens mot vissa behandlingar och visualiseringsverktyg som COSMIC-3D, som visar olika cancermutationers påverkan på proteiners 3D-struktur.

Standard/specifikation	Kommentar
NCIt, EFO	COSMIC hämtar data från många olika källor (artiklar m.m.) som använder olika kodsystem, ontologier etc. De konverterar källornas information om t.ex. tumörlokalisation, vävnad, histologi, fenotypdata, provbehandling och metadata till ett eget COSMIC-specifikt

	klassifikationssystem som i sin tur för vissa fält använder sig av t.ex. NCIt och EFO.
VCF, FASTA	Används som filformat för nedladdning av valda delar från databasen.
Beacon	Det går att söka efter träffar i COSMIC via GA4GH:s Beacon API.
GRCh37, GRCh38	Det går att välja vilket av dessa referensgenom man vill utgå ifrån.

***Förväxlingsvarning**: Ett av Sveriges vanligaste journalsystem, från Cambio Healthcare Systems AB, heter också Cosmic, så när man säger "Cosmic" till personal i de regioner som använder det journalsystemet så finns förväxlingsrisk.*



Figuren ger en översikt av innehåll och samband i COSMIC, bildkälla https://cancer.sanger.ac.uk/images/infographics/cosmic_graph.svg via https://cancer.sanger.ac.uk/cosmic/about

3.3. GDC data portal

Data i NCI:s Genomics Data Commons (GDC) data portal kommer huvudsakligen från större program, t.ex. TCGA och TARGET. Syftet är att fungera som en resurs för forskare inom cancerområdet, då det är en delad resurs som möjliggör datadelning över flera olika forskningsstudier. GDC möjliggör också för forskare att ladda upp egen data till databasen, där GDC bearbetar all data igen med hjälp av deras egna pipelines, och mappar resultaten till ett gemensamt referensgenom.

I Bilaga A. Information i internationella databaser och projekt listas vilken typ av data som kan laddas upp till GDC. Notera att inte alla program eller projekt har data för alla listade datatyper. Datan kan laddas upp via deras hemsida, eller ett API. Beroende på detta så accepteras olika filformat: XML, TSV eller JSON. När datan har laddats upp till GDC processas det genom deras pipelines och data genereras enligt tabell i bilaga A.

I Bilaga A. Information i internationella databaser och projekt finns även GDC:s datamodell representerad som en graf med noder som representerar vilken data som sparas för vardera post. Unika ID:n används för att identifiera kopplingar mellan projekt, prover, klinisk och molekylär data samt länkar också till den faktiska positionen av datan. Data sparas separat och i grafen sparas endast informationen om var datan är, inte själva datan. Grafen innehåller all information gällande en post, både vilken data som laddats upp och den data som genererats av olika pipelines.

3.4. CIViC

Clinical Interpretation of Variants in Cancer (CIViC) är en databas med information av betydelse för klinisk tolkning av varianter inom cancerområdet. Syftet är att främja precisionsmedicin genom att tillhandahålla ett webbforum där klinisk signifikans kan diskuteras och kunskap inom området kan spridas. Databasen kan nås via ett fritt åtkomligt webbgränssnitt (Browse section), ett publikt API eller mjukvaran CIViCpy Software Development Kit.

Datan tillhandahålls till största delen av nätverksmedlemmarna själva genom att manuellt registrera och editera information. Kravet är dock att informationen kommer från publicerad och granskad biomedicinsk litteratur som har ett associerat PubMed-ID eller ASCO-ID. Ett gransknings- och revideringsförfarande sker innan lagring.

Standard/specifikation	Kommentar
Entrez Gene (via MyGene.info)	När en gen skapas i CIViC importeras automatiskt information om gener (namn, synonymer, proteindomäner, pathways etc.) från Entrez Gene, via webbtjänsten MyGene.Info. MyGene.Info hämtar i sin tur data från Entrez Gene via veckovisa uppdateringar. Gennamn ska följa HGNC:s officiella gensymboler.
MyVariant.Info	Efter att variantdata har matats in manuellt i CIViC, hämtas utvald variantdata dynamiskt, via webbtjänsten MyVariant.Info. MyVariant.Info uppger på sin hemsida att de hämtar variantannoterings-data från många olika databaser (se dokumentation MyVariant.info).
TSV, VCF	CIViC producerar både nattliga och månatliga uppdateringar av sin databas i TSV- respektive VCF-format, uppdelat på olika entiteter (genes, variants, evidence, variant groups, assertions).
PubMed	Med hjälp av inmatat PubMed ID hämtar CIViC information som beskriver källan från PubMed-databasen.
GRCh37, Ensembl, Sequence Ontology, HGVS, ClinVar, gnomAD, Disease Ontology, NCIt, HPO, ACMG, NCCN	• Använd helst GRCh37 som referens vid framtagande av en variants koordinater. • Använd Ensembl archived version 75 (GRCh37) för transkript. • Använd Sequence Ontology-begrepp när varianttyp ska specificeras. • Använd bland annat HGVS-termer och ClinVar-ID då variantnamn väljs. • Utgå helst från gnomAD för uppskattningar avseende populationsfrekvens. • Använd Disease Ontology ID för sjukdom. • Ange NCI Thesaurus ID för läkemedel. • Ange associerade fenotyper i enlighet med HPO-databasen.

	• Använd ACMG-riktlinjernas patogenicitets-skala för klinisk signifikans. • Lista relevanta riktlinjer (t.ex. NCCN guidelines) som en variant och cancertyp förekommer i. • Ange FDA-godkända Companion Tests som är associerade med en Assertion.

3.5. Cancer Hotspots

Ett team hos Memorial Sloan Kettering Cancer Center (MSK) har tagit fram en statistisk algoritm som identifierar cancer hotspots. Med hjälp av metoden har de genomfört en analys vid två tillfällen där cirka 10.000 tumörprover analyserades den första gången och cirka 25.000 tumörprover analyserades den andra gången. Vid den andra analysen identifierades hundratals statistiskt signifikanta hotspot-mutationer.

Informationen finns lagrad i en databas, som har ett interaktivt webbgränssnitt fritt tillgängligt via Cancer Hotspots hemsida. Användare kan se alla gener som innehåller hotspots, de aminosyror som återkommande muteras samt hur fördelningen är mellan olika cancertyper. Datan som presenteras i webbgränssnittet kan laddas ner i en textfil. Dessutom går det att ladda ner själva resultatfilerna från analyserna.

MSK hittade ännu fler hotspots då de genomförde en analys för hur mutationer klustrar sig i tredimensionella proteinstrukturer, se 3D Hotspots.

Standard/specifikation	Kommentar
TCGA, the International Cancer Genome Consortium, oberoende publicerade sekvenseringsprojekt	Vid den andra analysen används dessa tre källor för inhämtande av data om mutationer.
ExAC, ClinVar	Information från ExAC och ClinVar användes för att exkludera irrelevanta varianter.
OncoKB	Mutationer annoterades avseende potentiell prognostisk och terapeutisk signifikans genom att använda OncoKB.

3.6. OncoKB

Oncology Knowledge Base (OncoKB) är en databas avseende precisionsonkologi som förvaltas av Memorial Sloan Kettering Cancer Center (MSK). Den innehåller information om effekter och behandlingskonsekvenser av variationer i cancergener. För närvarande finns information om somatiska varianter, men arbete pågår för att utöka databasen till att omfatta förvärvade mutationer.

Innehållet är manuellt inmatat av medlemmar i OncoKB Scientific Content Management Team (SCMT) under ledning av OncoKB Lead Scientist och genomgår ett granskningsförfarande innan publicering.

Behandlingsinformation klassificeras med hjälp av ett system av kriterier, kallat ’Levels of evidence’.

Det finns ett publikt webbgränssnitt där delar av datan i OncoKB är tillgänglig. Datan kan även nås programmatiskt via ett API samt verktyget OncoKB Annotator som kan användas för att annotera varianter. Uppdatering sker ungefär varje månad.

Standard/specifikation	Kommentar
cBioPortal, Cancer Hotspots, MSK-IMPACT, COSMIC, BRCA Exchange, IARC TP53 Database	Olika variantdatabaser används för information om genvarianter.
FDA, NCCN, NIH och andra guidelines	Riktlinjer om behandlingar associerade med varianter.
AACR, ESMO, ASCO, ASH	Konferenser. Medlemmar i SCMT deltar i konferenser för att inhämta användbar information.
PubMed	Information från vetenskaplig litteratur.
ClinicalTrials.gov	Information avseende kliniska prövningar. När läkemedel ska matas in, hämtas en lista med läkemedel att välja mellan via Clinical Trials API.
MSK best practices	Best practices hos Memorial Sloan Kettering Cancer Center.
HUGO, Ensembl, RefSeq, OncoTree	HUGO-gensymboler ska användas för gennamn. Ensembl och RefSeq transkript-ID används för transkript. OncoTree används för att ange tumörtyp. (OncoTree är i sin tur mappat till NCI Thesaurus och UMLS. UMLS inkluderar SNOMED CT som sin källa.)

3.7. cBioPortal for Cancer Genomics

cBioPortal for Cancer Genomics är en plattform som Memorial Sloan Kettering Cancer Center (MSK) har tagit fram för att integrera omiks-data från olika studier samt underlätta visualisering. På hemsidan finns ett interaktivt webbgränssnitt där användare kan utforska, analysera, visualisera och ladda ner data. Ett av målen är att enkelt kunna identifiera viktiga händelser i stora cancerdatasets, t.ex. vilka mutationer som kan orsaka maligna celltransformationer eller tumöröverlevnad (”driver mutations”).

Databasen innehåller information om olika varianter per gen, varianters effekt samt kliniska följder.

cBioPortal importerar data från stora studier som exempelvis TCGA och ICGC, men även publicerade sekvenseringsstudier från enskilda laboratorier. Innan en studie laddas in, körs variantdata igenom en standardpipeline för att skapa enhetliga annoteringar mellan olika studier. I begränsad omfattning finns även information om avidentifierad klinisk data. Vad som är tillgängligt skiljer sig åt från studie till studie.

cBioPortal-mjukvaran har öppen källkod tillgänglig på GitHub. Flera akademiska och kommersiella center har installerat privata instanser av mjukvaran och deras data är inte publikt tillgänglig.

Standarder, verktyg, databaser	Kommentar
PubMed, TCGA, ICGC, Count me in, Project GENIE, CCLE	Datakällor för omiks-data (t.ex. mutationer, fusioner, kopienummerförändringar, mRNA-uttryck, proteinnivåer, DNA-metylering) och avidentifierad klinisk data. TCGA legacy-data uppdateras varje kvartal genom att importera data via Broad Firehose. (Åtkomst till GENIE-data kräver registrering först.)
OncoKB, My Cancer Genome, CIViC	Integrerade resurser för att hämta annoteringar om varianters kliniska följder (t.ex. biologiska effekter, klinisk signifikans och terapieffekter).
COSMIC, Cancer Hotspots, 3D Cancer Hotspots	Integrerade resurser för att hämta annoteringar om hotspots och ”recurrence”.
MutationAssessor, PolyPhen-2, SIFT	Integrerade resurser för att hämta annoteringar om funktionspåverkan.
HUGO, Entrez Gene	För gen, använd HUGO-gensymboler, Entrez Gene-identifierare och genalias.

RefSeq ID, Ensembl ID, CCDS ID, Uniprot ID, gnomAD information, ClinVar ID, dbSNPID, Oncotree code	Olika typer av IDn från andra resurser finns tillgängligt.
Canonical UniProt transcript (via GenomeNexus)	Annoteringar för mutationer standardiseras genom att mutationsdata mappas till samma protein isoform. På så vis hanterar cBioPortal variantannoteringar på ett enhetligt sätt mellan olika studier. För detta används Genome Nexus (som använder VEP med ”the canonical UniProt transcript”).
hg19/GRCh37	Används av cBioPortal som referensgenom.
PFAM	Proteindomändefinitioner som visas i lollipop-diagram hämtas från PFAM.
MutSig, GISTIC	För vissa TCGA-studier finns resultat från MutSig- och GISTC-algoritmerna avseende statistisk signifikans och “variant recurrence” tillgängliga.

3.8. My Cancer Genome

My Cancer Genome är en kunskapsbas för precisionsmedicin. Den innehåller information om mutationer som påverkar cancer, effekter av terapi samt information om tillgängliga kliniska prövningar. Datan är tillgänglig via ett webbgränssnitt på My Cancer Genomes hemsida. Det finns även ett API.

Varje natt importeras dokument avseende kliniska prövningar in i kunskapsbasen från externa publika källor. Cancerrelaterade nyckelord väljer ut relevanta dokument och strukturerade fält i dokumenten söks igenom för att lagras som metadata i kunskapsbasen. Därefter flaggas lagrade poster upp för manuell genomgång.

Med hjälp av ett webbgränssnitt granskar sedan informationsansvariga datan manuellt och lägger till ytterligare information om sjukdomar, biomarkörer och läkemedel. Vid detta arbete följs kriterier framtagna av My Cancer Genome: Eligibility Criteria Assertions (ECA), Treatment Context Assertions (TCA) och Treatment Arm Assertions (TAA). Sedan publiceras datan och blir publikt tillgänglig via hemsidan.

Standard/specifikation	**Kommentar**
ClinicalTrials.gov, Cancer.gov, UMIN Clinical Trials Registry	Information avseende kliniska prövningar överförs från externa datakällor.
HGVS, HGNC, ISCN, RefSeq, WHO classification, OncoTree, SNOMED CT, NCIt	Ontologier och nomenklatursystem som används för biomarkörer, sjukdomar och läkemedel. Exempel: ● genomiska biomarkör-kriterier annoteras via RefSeqs databas (GRCh37, annotation release 105) ● HGVS används för variantnamn, HGNC för gennamn och ISCN för cytogenetiska biomarkörer. ● WHO:s klassifikationer för hematologi-sjukdomar, NCI Thesaurus, OncoTree och SNOMED CT används för sjukdomar. ● NCI Thesaurus och SNOMED CT används för läkemedel.
FDA, NCCN och behandlingsriktlinjer	Olika källor används för att ta fram prediktiv information om läkemedelsmottaglighet samt om ett läkemedel är godkänt av FDA.
UTA, UniProt, dbNSFP	Information om funktionspåverkan.
Project GENIE dataset	Biomarkör- och sjukdomsfrekvens hämtas från GENIE.

3.9. Mitelman Database

Mitelman Database of Chromosome Aberrations and Gene Fusions In Cancer (Mitelman Database) är en databas med information om cytogenetiska förändringar inom cancerområdet som har publicerats i vetenskaplig litteratur. Detta inkluderar information om studie, patient, tumör samt identifierade genfusioner.

Datan kan nås via ett publikt webbgränssnitt där användaren även kan ladda ner data samt få ytterligare information via internetlänkar till PubMed, NCBI Gene och OMIM Gene Map.

Innehållet sammanställs manuellt från publicerad litteratur av Felix Mitelman, Bertil Johansson och Fredrik Mertens. Databasen uppdateras varje kvartal i januari, april, juli och oktober.

Standard/specifikation	Kommentar
ISCN	Används för kromosomavvikelse och breakpoint.
Guidelines for Human Gene Nomenclature	För gener används HGNC-godkända gensymboler.
ICD-O, SNOMED, FAB propsals for the classification of acute leukemias and myeolodysplastic syndromes, WHO Classification of Tumours of Soft Tissue and Bone	Används för morfologi (tumörhistologi).
REAL/WHO classification Peripheral B-cell Neoplasma	Tumörer inom lymfatiska systemet (lymphoreticular system) konverteras till REAL/WHO-klassifikationen. Ospecificerade non-Hodgkin lymfom presenteras som Peripheral B-cell neoplasm, NOS.
TSV	Nedladdning av data i TSV-format.

4. Internationella databaser cancer hereditärt

De databaser som är av intresse inom sjukdomsområde cancer hereditärt, framgår i detta kapitel. Information om metadata finns i Bilaga A. Information i internationella databaser och projekt.

4.1. BRCA Exchange

BRCA Exchange är en databas som aggregerar information avseende varianter i generna BRCA1 och BRCA2. Syftet är att forskare och kliniker ska få information om vilken klinisk tolkning som gjorts för en variant, när tolkningen gjordes samt vilken data som låg till grund för tolkningen.

Resursen är fritt åtkomlig online via BRCA Exchanges hemsida och en smartphone app. Det finns även ett publikt API. Målet är att ge en heltäckande bild på ett enda ställe och därför överförs variantdata från många olika databaser över hela världen.

BRCA Exchange har utvecklat en dataanalys-pipeline som körs automatiskt en gång per månad. Den laddar ner aktuella variant- och populationsfrekvensdata från variantdatabaser, lokusspecifika databaser samt befolkningsgenetiska databaser.

Standard/specifikation	Kommentar
ClinVar, ENIGMA, LOVD, BIC, ExUV, gnomAD, ExAC, ESP, 1000 Genomes, Findlay et al	För automatiska överföringar av variant-, populationsfrekvens- och functional assay-data.

Guidelines for Human Gene Nomenclature	För gener används HGNC-godkända gensymboler.
HGVS, BIC, GRCh37, GRCh38, VRS, VCF	• Variantalias följer HGVS-nomenklaturen. • Variantalias visas även enligt BIC-nomenklaturen. • Variantalias används för att ta fram GRCh37 och GRCh38-genomkoordinater. • VRS-identifierare är en global unik identifierare för varianten. • I samband med överföring av variantdata från externa databaser konverteras data till VCF-filer. • En förkortning av HGVS-nomenklaturen för proteinalias används för aminosyror.

4.2. InSiGHT variantdatabaser

International Society for Gastrointestinal Hereditary Tumours (InSiGHT) är en internationell och multidisciplinär organisation. De har tagit fram databaser för nedärvda varianter som kan orsaka cancersjukdom i mag-tarmsystemet. För närvarande innehåller de information om varianter i generna APC, CDH1, EPCAM, GALNT12, MLH1, MLH3, MSH2, MSH6, MUTYH och PMS2. Dessutom finns information om de tolkningar av varianters patogenicitet som tas fram för MMR-gener (MLH1, MSH2, MSH6 och PMS2). Tolkningarna utförs av konsortiets VIC-kommittéer (Variant Interpretation Committes) i enlighet med VIC:s klassifikationskriterier.

Personer med kunskap inom området, kan rapportera in varianter till InSiGHT:s innehållsansvariga som tillsammans med en panel av experter inom området granskar att nomenklatur följs samt beslutar vilken tolkning som ska gälla innan datan publiceras.

Datan finns i 'Global Variome shared LOVD' vilket är en internationell, centraliserad och fritt åtkomlig LOVD-databas (version 3) med information om varianter. InSiGHT har dessutom ett eget webbgränssnitt på sin hemsida som länkar vidare till LOVD för mer detaljerad information. Webbgränssnittet innehåller även en länk till 'the Prospective Lynch Syndrome Database' som ger visualiserad information avseende risk för cancer.

Standard/specifikation	Kommentar
Guidelines for Human Gene Nomenclature	För gener, använd HGNC-godkända gensymboler.
ClinVar, dbSNP, ISCN	För varianter, använd ClinVar ID och dbSNP ID. Det går även att ange ISCN, men det verkar inte ha använts hittills. (Det står inte definierat vad ISCN är förkortning för.)
OMIM	För sjukdomar, ange OMIM ID.
NG, NC eller LRG accession-nummer	För genomreferenssekvens, använd NG, NC eller LRG-accessionnummer från NCBI/EBI.
NCBI, Ensembl, Uniprot	För transkriptreferenssekvens, använd NCBI ID från NCBI GenBank. Det går även att ange Ensembl ID och Uniprot ID.
Mutalyzer	När ett transkript ska läggas till, kontaktar LOVD automatiskt Mutalyzer för att hämta en lista med tillgängliga transkript att välja mellan. Mutalyzer hämtar i sin tur information om transkript från NCBI Map Viewer.

4.3. ENIGMA

Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) är ett internationellt konsortium som bedömer klinisk signifikans av varianter i BRCA1-, BRCA2- och andra bröstcancergener samt tillhandahåller expertutlåtanden till globala databaser och klassifikationsinitiativ.

ENIGMA tillhandahåller klassificeringar till BRCA Exchange, ClinVar och Global Variome shared LOVD BRCA1 samt Global Variome shared LOVD BRCA2. Till ClinVar rapporteras tolkningarna in manuellt, medan BRCA Exchange uppger att tolkningarna överförs från ENIGMA:s databas.

Denna analys har inte kunnat hitta mer information om exakt vilken databas som förvaltas av ENIGMA eller information om vilket webbgränssnitt som i så fall kan användas för att komma åt datan.

5. Internationella projekt

De internationella projekt som har hunnit undersökas framgår i detta kapitel. Information om relaterad metadata finns i Bilaga A. Information i internationella databaser och projekt.

5.1. Genomics England

Genomics England är en organisation som skapades för att leverera projektet 100.000 Genomes. Projektets mål var att utföra helgenomsekvensering av 100.000 genom hos patienter som har en cancersjukdom eller en sällsynt diagnos (inklusive deras familjer). Efter att ha uppnått målet år 2018, fortsätter arbetet genom att sekvensera ytterligare helgenom. Initiativet ska bidra till att utveckla hälso- och sjukvård, men även driva forskning inom genomikområdet framåt.

Arbetet inom Genomics England övervakas av en styrelse som leds av en ordförande. Styrelsen ska säkerställa att organisationen har processer för dokumentation, övervakning och rapportering. En oberoende kommitté (Access Review Committee) godkänner ansökningar om att få tillgång till data för forskningsändamål. Ett dataskyddsombud har utsetts, vilket stipuleras av lagstiftning, General Data Protection Regulation.

Inom den engelska hälso- och sjukvården, National Health Service (NHS), har ett antal nationella centrum bildats, benämnda NHS Genomic Medicine Center. De identifierar deltagare, inhämtar informerat samtycke, sköter insamling av prov och skickar proven för sekvensering till ett sekvenseringscentrum (utkontrakterat till Illumina). När sekvenserings-centrumet har utfört sekvensering skickas genomikdatan till Genomics England. De nationella centrumen tillhandahåller dessutom klinisk information, såsom fenotypdata, direkt till Genomics England. En IT-infrastruktur har skapats hos Genomics England för att lagra genomsekvenser och klinisk data. Datan analyseras inom infrastrukturen och viktiga fynd (såsom diagnos) skickas tillbaka till patientens läkare.

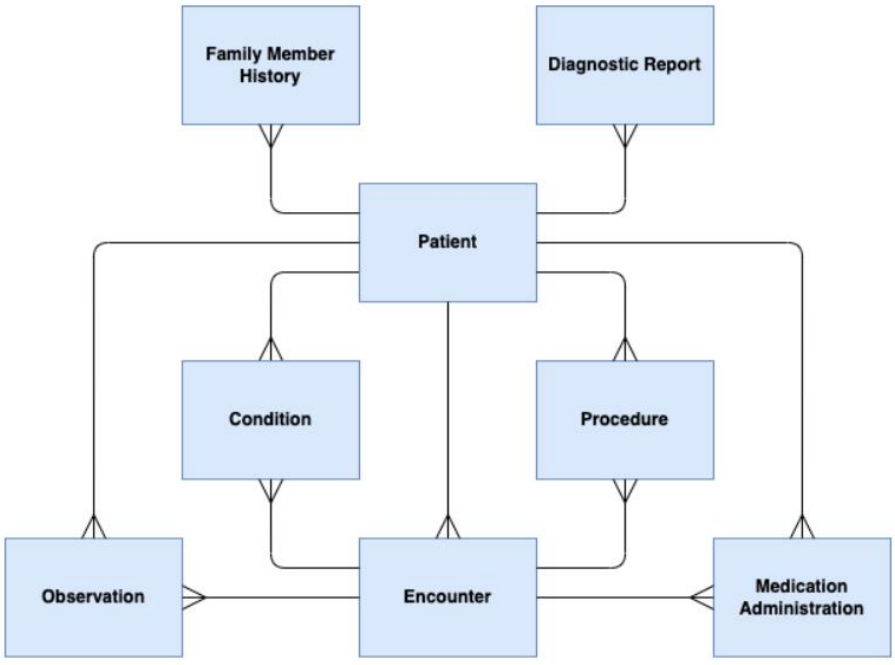
För forskningssyften har även en forskningsinfrastruktur etablerats som länkar ihop journaldata med genomikdata och klinisk data som samlas in inom Genomics England.

Standard/specifikation	Kommentar
BAM, VCF	Sekvenseringscentrumet (Illumina) tillhandahåller sekvenserad data inpassade till referensgenom i filformatet BAM. Dessutom identifierar de sekvensvarianter och rapporterar dem i filformatet VCF.
CRAM	Används som filformat för att lagra komprimerad genomikdata i IT-infrastrukturen.
HTSGET	Används för att tillgängliggöra genomikdatan i IT-infrastrukturen till olika laboratorier i England. Laboratorierna behöver åtkomst till genomikdatan för att visualisera den med verktyg såsom Integrative Genomics Viewer (IGV). Dessutom behöver de se över genvarianter, klinisk information och scoring av genvarianterna inför

	diagnosticering samt lagra genvarianter i sina kunskapsbaser. Projektet har utvecklat en open source-tjänst som implementerar specifikationen HTSGET och det är den tjänsten som laboratorierna kan använda.
WES, TES (i proof-of concept)	Processen att ta in nya pipelines i produktionsmiljön har varit relativt komplicerad. Därför utvärderar projektet om det går att använda WES och TES för att ta in arbetsflödena på ett effektivare sätt.
DRS (inplanerat för utvärdering)	Datan finns dels i IT-infrastrukturen och dels i molntjänster. Det utreds om DRS skulle kunna användas för att identifiera en fil oavsett vart den är lokaliserad.
VRS (inplanerat för utvärdering)	Projektet undersöker om genvarianterna kan representeras på detta sätt.
Beacon2, InterpredGenome (under övervägande)	Genomics England är involverade i GA4GH:s arbete med Beacon version 2 och InterpretedGenome. De vill kunna dela information internationellt, men vill då få mer detaljerad information om genvarianter än enbart ja/nej-svar på enkla frågor såsom ”finns denna variant”.
HPO	Används för närvarande för att definiera fenotypiska karaktäristika, men andra ontologier kan komma att användas framöver.
Guidance from NHS Digital and the Information Commissioner’s Office	Riktlinjer för hur data ska pseudonymiseras innan lagring i Research Environment.
SNOMED CT, ICD10, ICD-O, OPCS	För detaljer, se rubrik “Code Systems Overview” i dokumentet Data in Participant Explorer.

Exempel på användning av HL7 FHIR i Genomics England

FHIR används för dataåtkomst inom flera delar av Genomics England. Ett illustrativt exempel är för den data som används i Participant Explorer där flera FHIR-resurser samt FHIR:s terminologiserver-API nyttjas. Se samband i figur nedan.



Klinisk datamodell i “Participant Explorer” baserad på HL7 FHIR, bildkälla https://re-docs.genomicsengland.co.uk/pxa_data/

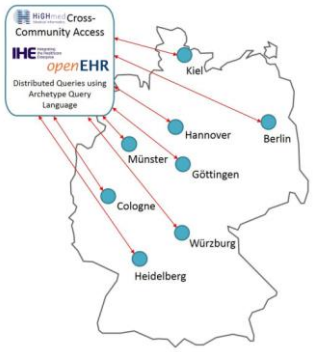
Exempel på användning av openEHR i Genomics England

Fenotypdata (journalutdrag m.m.) i 100 000 genomes hanteras för flera “NHS trusts” sedan 2017 med openEHR. I en video nämns att Genomics England behövde ca 13.000 olika typer

av datapunkter med fenotypdata etc. och videon beskriver (lite knapphändigt, cirka 13 minuter i filmen) en skiss över informationsflöden och IT-arkitektur som (lite likt tyska HiGHmed) har en openEHR-baserad nod för att standardisera innehåll från olika system (journal, laboratorier etc.) ute hos respektive vårdgivare.

5.2. HiGHmed och nationell informatiksatsning

Det tyska projektet HiGHmed är inte inriktat specifikt på genomik, men har strukturella likheter med GMS genom att syfta till att binda ihop flera regioner/universitetssjukhus och dela detaljerad klinisk data för både forskning och behandling.



HiGHmed-delprojekten om ”oncology” och ” infection control” har likheter med GMS motsvarande spår och hanterar datadelning, integrationer, gemensamma visualiseringar och verktyg m.m.

HiGHmed är ett av flera konsortier i den nationella tyska 160 M€-satsningen på medicinsk informatik. Satsningens mål är att skapa en miljö som gör det möjligt att använda forskningsresultat till patienternas direkta nytta - samtidigt som det säkerställer robust dataskydd och säkerhet. HiGHmed låter data ligga kvar hos respektive vårdgivare, fast smart standardiserat, så att man kan köra samma sökfrågor/datauttag mot alla vårdgivare (OHDSI och delar av GA4GH har liknande angreppssätt). I den tyska infrastrukturen för COVID-19-forskning (som kan vara intressant att titta på för GMS spår om infektionssjukdomar) samarbetar flera av dessa konsortier och använder standarder som IHE och HL7 FHIR för överföring samt openEHR för lagring och analys.

I denna rapports avsnitt om openEHR nedan länkas det bland annat till HiGHmed:s openEHR-baserade modeller för genomik och onkologi. Det finns en (aningen teknisk) presentation på YouTube där det från 4:55 och ca 10 minuter framåt ges en överblick av vad HiGHmed försöker lösa och hur (bland annat med standarder som openEHR och HL7 FHIR).

Vi har inte hittat detaljer online om hur HiGHmed jobbar med och integrerar variantdatabaser men skulle kunna fråga dem ifall angreppssättet för övrigt verkar intressant för GMS att undersöka närmare.

Hemsidor: https://www.highmed.org/ och https://www.medizininformatik-initiative.de/

6. Standarder, specifikationer och relaterade organisationer

6.1. Standarder i GA4GH

Global Alliance for Genomics and Health (GA4GH) är ett internationellt nätverk som tar fram ramverk och standarder avseende delning av genomik- och hälsorelaterad data. Arbetet är indelat i olika arbetspaket som bland annat hanterar datasäkerhet, legala aspekter och etik, klinisk data och fenotypdata, moln- och discovery-tjänster etc. GMS är en av medlemmarna.

Nedan beskrivs några av de standardiseringsinitiativ som GA4GH arbetar med.

SAM, BAM: Ett filformat som används för lagring av next generation DNA-sekvensdata, så kallade reads. SAM representerar datan i läsbar textform, medan BAM är dess komprimerade, binära motsvarighet. Har traditionellt använts för att representera så kallade alignments (DNA-

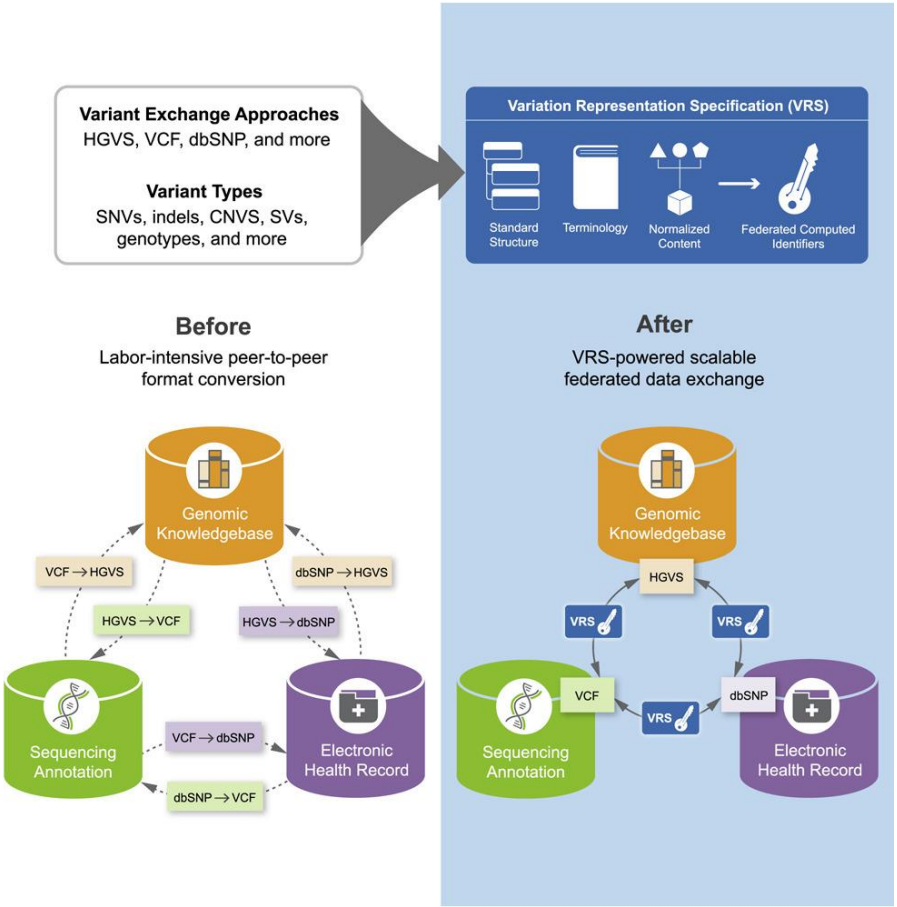
sekvensdata mappade till referenssekvenser), men GA4GH framhäver att formaten inte är begränsade enbart till alignment-data.

VCF: Ett filformat i textform som används inom bioinformatik för att lagra varianter. Istället för att lagra samtlig genetisk data, lagrar detta format enbart varianterna samt ett referensgenom. Har en binär motsvarighet, BCF.

HTSGET: Ett API för åtkomst av genomikdata. Gör att användare kan ladda ner read-data för de subsektioner av genomet som de är intresserade av, istället för att behöva ladda ner en mängd olika filer där datan finns.

WES, TES, DRS: Ett antal API-standarder som används för att skapa portabla verktyg i molnmiljöer. De gör det möjligt att exekvera individuella jobb i molnet (TES), exekvera och övervaka arbetsflöden i olika moln, plattformar och miljöer (WES) samt läsa/skriva dataobjekt mellan moln (DRS) utan att vara knuten till detaljer som gäller en viss miljö.

VRS: För att utbyta variationsdata (inklusive men inte begränsat till genomiska varianter) på ett standardiserat sätt föreslår GA4GH att specifikationen Variation Representation Specification (VRS)[1] används. Den utgör ett utbyggbart ramverk för att beräkningsmässigt (computable) representera en variant som kompletterar de humanläsbara standarder och flatfils-standarder som finns för att representera genomiska varianter.

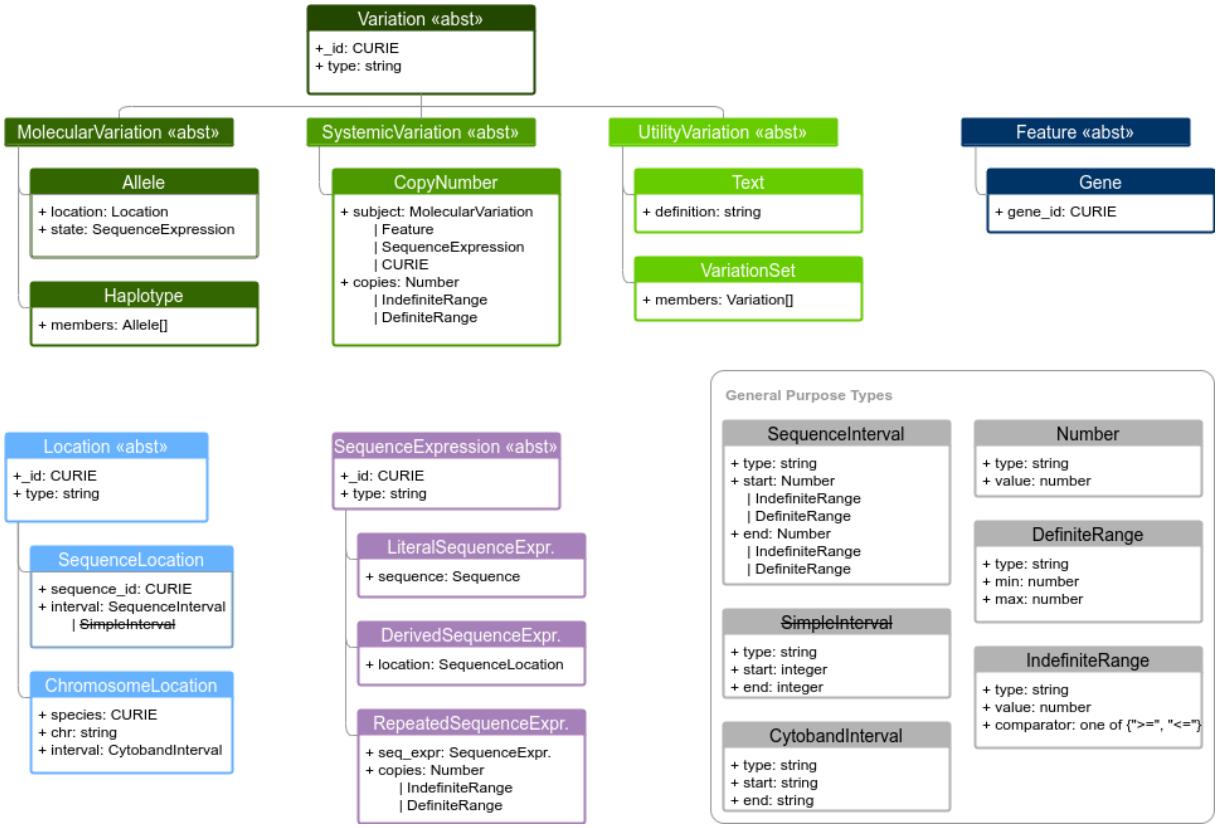


Figuren visar en grafisk sammanfattning av VRS (hämtat från en rapport om VRS[2])

[1] VRS-specifikationen: https://vrs.ga4gh.org
[2] Rapport om VRS: https://doi.org/10.1016/j.xgen.2021.100027

För att uppnå en exakt och beräkningsbar representation av varianter, utformades VRS med hjälp av flera ömsesidigt beroende komponenter vilka inkluderar en terminologi- och informationsmodell, maskinläsbara scheman, konventioner för delning samt globalt konsekventa och unika identifierare. Dessa komponenter gör det möjligt för specifikationen att hantera flera användningsfall och utvecklas för att passa flera behov. Specifikationen förvaltas av GA4GH.

Nedan beskrivs informationsmodellen.



Figuren visar VRS informationsmodell. Informationsmodellen består av flera dataklasser, inklusive både konkreta klasser och abstrakta superklasser (markerade med <<abst>>). Dessa klasser utgör konceptuella representationer av variationer (gröna rutor), funktion (blå rutor), plats (ljusblå rutor), sekvensuttryck (lila rutor) och allmänna ändamålstyper (grå rutor). De allmänna ändamålstyperna stöder de primära klasserna, inklusive intervall, räckvidd, nummer- och GA4GH-sekvenssträngar (visas inte i figuren). Se mer i rapport om VRS[3].

Den maskinläsbara VRS är skriven med JSON Schema[4]. De maskinläsbara schemadefinitionerna och exempelkoden är tillgängliga online i VRS-förrådet[5] (https://github.com/ga4gh/vrs).

Beacon: Beacon-projektet är ett GA4GH-initiativ för internationell delning av genomikdata och klinisk data. Organisationer (sjukhus och forskningsinstitut) som vill utbyta data systemmässigt behöver sätta upp en server som implementerar Beacon:s API. Därefter fungerar den som ett

[3] Rapport om VRS: https://doi.org/10.1016/j.xgen.2021.100027
[4] JSON Schema: https://json-schema.org/
[5] VRS-förrådet: https://github.com/ga4gh/vrs

slags sökverktyg för att kunna få information om t.ex. en viss variant har hittats hos de andra anslutna organisationerna eller få information om själva värdorganisationen (en beskrivning av organisationen och vilka typer av information de förvaltar). Beroende på vilka nätverk organisationen har anslutit sig till, kan sökning göras internationellt eller nationellt. Anslutna Beacons ska kunna hantera utgående förfrågningar till omvärlden och inkommande förfrågningar från omvärlden. Det är möjligt att sätta upp egna nätverk.

Version 1 av Beacon-specifikationen gav organisationer möjlighet att ställa frågor till andra anslutna organisationer, t.ex. ”har ni observerat kvävebas C på position 32926732 på kromosom 13 i genomet?”. Den andra organisationens Beacon svarade då ja eller nej. Det var även möjligt att returnera s.k. counts på vilka varianter, variant calls och prover som hittats för en viss variant, t.ex. ”i detta dataset hittas varianten 74 gånger, callas 74 gånger och finns i ett prov”.

2022-05-25 släpptes en ny version, Beacon version 2. I den versionen kan även annan information returneras, t.ex. allelfrekvenser, variantklassificeringar och en persons fenotyper som är associerade med den efterfrågade allelen. (Detta är möjligt eftersom version 2 stödjer en typ av autentisering som gör att information om t.ex. vilken fenotyp personen har, är säker att överföra.)

Åtkomst till Beacons säkerställs via organisationens eget säkerhetssystem, men GA4GH rekommenderar att använda tre nivåer av dataåtkomst (open, registered eller controlled). Organisationer kan välja att ha öppen dataåtkomst för anonyma användare avseende information om att en allel har observerats, men för att få ytterligare information om olika metadata (fenotyp, klinisk data) behöver användaren identifiera sig och ha samtyckt till ett antal villkor för att få använda informationen.

Phenopackets: Enligt GA4GH finns det ingen standard som har nått bred acceptans för hur fenotypegenskaper ska utbytas tillsammans med genomik- och annan relevant information. Det finns ontologier såsom HPO för att representera fenotypegenskaper, men ett ramverk för att utbyta denna information saknas. GA4GH har därför tagit fram Phenopackets som är en specifikation för att dokumentera och utbyta kliniska fenotyper. Reglerna i specifikationen definierar hur data ska organiseras för att skapa ett s.k. phenopacket, det vill säga en strukturerad representation av en persons medicinsk data. Specifikationen är fritt åtkomlig och drivs av GA4GH-nätverket samt ska kunna användas inom forskning samt hälso- och sjukvård.

Ett phenopacket karaktäriserar en person eller ett prov och länkar den/det till detaljerade fenotypbeskrivningar samt information om genom, diagnoser och behandling. Ett enkelt phenopacket kan innehålla en lista av de fenotypegenskaper som har observerats hos en person, medan ett mer komplext phenopacket kan innehålla information om fenotypegenskaper, prover och behandling som är associerade med en person. Den enkla typen av phenopacket kan t.ex. användas för diagnostik av sällsynta diagnoser, medan det komplexa är mer användbart i områden bortom diagnostik, t.ex. val av behandling.

Själva fenotypdelen av paketet kan innehålla parametrar kodade med exempelvis HPO, för att beskriva t.ex. fenotypegenskapen, när i tiden egenskapen först inträffade och om en lösning har hittats. Exempel: Fenotyp som observerats är ”allvarliga dagliga spädbarnsspasmer som inträffade i barndomen och löstes vid åldern 4 år och 2 månader hos ett barn med global utvecklingsstörning”. Andra delar av paketet kan representera information om personen (kön, vital status), prov, sjukdom, tolkning av varianter (ACMG/AMP-klassificering för en kausal variant eller parametrar för att representera en actionable somatisk variant), behandling och medicinsk åtgärd (medicinering/vaccinering/CAR T-cellterapi, dosering) etc.

Specifikationen är interoperabel med andra GA4GH-standarder t.ex. Beacon, Pedigree och VRS. Det är designat för att passa många olika typer av ontologier, men det finns ändå rekommendationer för vilka som helst bör användas. Exempelvis bör HPO användas för sällsynta diagnoser och NCI Thesaurus) för detaljerad information inom cancerområdet (patologisk staging eller detaljer om histologi och tumörmarkörer).

Matchmaker Exchange: Matchmaker Exchange är ett projekt som samarbetar med GA4GH och International Rare Diseases Research Consortium (IRDiRC). Matchmaker Exchange tillhandahåller ett API som implementeras av olika organisationer. Tillsammans bildar anslutna organisationer ett internationellt nätverk av datakällor som används för att hitta andra personer med varianter i samma gen och samma eller liknande fenotyp (s.k. patientmatchning). Om ett liknande fall hittas kan användare kontakta varandra och diskutera om fynden är tillräckliga för att bevisa att en variant är sjukdomsorsakande. Den typ av information som finns tillgänglig hos de olika organisationer som deltar i nätverket framgår av nätverkets hemsida. Forskningsinstitut och laboratorier kan sätta upp en egen Matchmaker Exchange-nod eller ansluta sig till en befintlig och på så vis delta i nätverket för att lösa patientfall.

6.2. Ontologier, klassifikationer, begrepps- och terminologisystem

HPO: Human Phenotype Ontology är en ontologi för att beskriva fenotyp-abnormaliteter hos människor. HPO är hierarkiskt indelad i olika HPO-grupper.

Guidelines for Human Gene Nomenclature: En nomenklatur som standardiserar namngivning av gener. Riktlinjerna avser bland annat proteinkodande, RNA- och pseudogener.

ICD, ICD-10, ICD-11: International Classification of Diseases är en sjukdomsklassifikation från WHO. ICD-10 (revision 10 av ICD) används idag ofta för diagnoskoder i journaler m.m. i Sverige. Svensk översättning, anpassning, utgivning och övrig förvaltning görs av Socialstyrelsen. Arbete med nästa generation, ICD-11, pågår och svensk översättning väntas finnas publicerad cirka 2024.

ICD-O: International Classification of Diseases for Oncology är en morfologi-klassifikation avseende tumörsjukdomar. Innehåller morfologikoder och används bland annat i cancerregister.

NPU: Nomenclature for Properties and Units används inom laboratoriemedicin för att man vid utbyte av information mellan system ska säkerställa att det är samma analyser som gjorts. NPU-koder används t.ex. i NPÖ (Nationella Patientöversikten) och patientjournalsystem där vårdpersonal och invånare kan se resultat på analyser utförda av anslutna regioner. NPU-koderna säkerställer att de resultat som visas i NPÖ är jämförbara, och kan visas tillsammans i en kumulativ översikt, oavsett vilket laboratorium som producerat svaret.

Den internationella NPU-terminologin är utarbetad av IUPAC (International Union of Pure and Applied Chemistry, Internationella kemiunionen) och IFCC (International Federation of Clinical Chemistry and Laboratory Medicine). Översatta versioner av NPU används som nationell terminologi i Danmark, Norge och Sverige.

Equalis är nationellt releasecentrum för NPU i Sverige, vilket innebär att de ansvarar för översättning av terminologin till svenska och för administration av den svenska NPU-databasen. Förutom de internationella NPU-koderna ingår även nationella SWE-koder i

databasen. De används när det saknas lämplig NPU-kod för laboratorieundersökningar som utförs i Sverige.

Vissa regioner och organisationer använder lokala koder som ett komplement, i de fall de inte finner det lämpligt att använda koder ur NPU-databasen. Dessa koder administreras lokalt och ingår inte i NPU-terminologin.

Inera har i ett utvecklingsprojekt tagit fram GLOO 4.0 (GLOO - GetLaboratoryOrderOutcome) som ska fungera även för resultat på analyser avseende infektionssjukdomar (mikrobiologi). Tjänsten GetLaboratoryOrderOutcome returnerar patientens kemilaboratoriesvar och/eller mikrobiologisvar.

Varje svarspost innehåller ett laboratoriesvar, vilket består av en eller flera analyser, där varje analys har sitt resultat. Ansvaret för informationen delas mellan den som beställt laboratorieundersökningen och den på laboratoriet som står för resultaten. Signering förekommer för hela laboratoriesvaret samt per analysresultat. Ansvarig laboratorieläkare är normalt den som signerar analysen/analyserna. Läkaren på beställande enhet signerar också informationen.

SNOMED: Systematized Nomenclature of Medicine. Det är lätt att blanda ihop olika generationer av dessa system. Oftast avses idag “SNOMED CT” när någon säger “Snomed”, men det finns fler generationer varav en hette just SNOMED: *SNOP (1965), SNOMED (1974), SNOMED II (1979), SNOMED International 3.0 (1993), SNOMED RT (2000), SNOMED CT (2002)*. Sedan 2017 har licenser och underhåll av föregångarna till SNOMED CT upphört, men vissa av dem används dock fortfarande i Sverige inom t.ex. delar av patologi. Övergång till SNOMED CT pågår.

SNOMED CT: SNOMED Clinical Terms är ett stort internationellt begrepps- och terminologisystem, exempelvis inom områdena anatomi, sjukdomar, fynd, procedurer och mikroorganismer med över 350.000 aktiva begrepp. Socialstyrelsen arbetar med översättning, underhåll och utgivning. Sverige är anslutet som nation och alla invånare och organisationer i Sverige får använda SNOMED CT via en gratis licens från Socialstyrelsen. Bläddring/sökning kan göras via SNOMED International SNOMED CT Browser, välj “Swedish Edition”. Det pågår arbeten och studier med kopplingar mellan SNOMED CT och andra system, t.ex. HPO och ICD-10.

NCI Thesaurus (NCIt): Används som en bred referensterminologi i många system från NCI och andra aktörer. Innehållet kan undersökas via ncitbrowser. Dessutom finns där NCI Metathesaurus (NCIm) som kopplar innehåll i NCIt till många andra system.

6.3. Datafångst och omtolkningar (dokumentationsmodeller, meddelandeformat etc.)

För att hantera och lagra patientspecifik data i vård-IT-system (t.ex. journalsystem) används datastrukturer av olika slag. De kan vara konstruerade av leverantören själv (proprietära) eller bygga på öppna specifikationer/standarder som kan användas av flera olika leverantörer.

För bilddiagnostik har länge standarder som DICOM använts för att undvika att de lagrade bilderna (och delar av tillhörande metadata) blir bundna till ett specifikt system och därmed inte kan återanvändas i andra leverantörers system (t.ex. när en vårdgivare vill byta system, eller när data skall skickas till en annan vårdgivare). DICOM innehåller både sätt att standardiserat lagra och överföra information.

För strukturerad text- och sifferbaserad dokumentation som mätvärden, strukturerade journalanteckningar, laboratoriesvar m.m. har standardisering av lagring respektive överföring varit mer uppdelad. Ofta har datafångst och lagring gjorts med proprietära/leverantörsspecifika modeller inuti system men överföring/systemintegration har baserats på **överföring**s-standarder som HL7 v2.x, HL7 CDA, Ineras nationella tjänsteplattform/tjänstekontrakt m.m. Tittar man framåt så är den mer REST-API-baserade standarden HL7 FHIR intressant som överföringsstandard vid nyutveckling av integrationer som behövs om man t.ex. inte kan påverka så mycket i källsystemens datafångst och lagring.

Tittar man framåt för standardiserad **datafångst och lagring** av strukturerad text- och sifferbaserad dokumentation inuti system börjar openEHR användas i allt fler system i Sverige (Cambio, TietoEVRY med flera) samt hos enskilda vårdgivare (t.ex. Karolinska Sjukhuset) och internationellt (t.ex. Norge, Finland, Storbritannien, Tyskland). Svenskt standardiseringsarbete med exempelvis urval och tillämpningsanvisningar för SNOMED CT och andra terminologier/ontologier är andra exempel som förbättrar jämförbarhet för datafångst.

En viktig strävan för datakvalitet och jämförbarhet är att titta på och försöka påverka hur primärdata av betydelse för GMS (t.ex. genotyp-, fenotyp- och metadata) fångas och lagras i källsystem. Om man inte har tillräcklig jämförbarhet, struktur och kvalitet från början så är det svårt eller omöjligt att åtgärda bristerna eller olikheterna i informationen senare i kedjan (se exempel i presentation, bild 13 och 18). Indata till nya poster i svenska variantdatabaser eller motsvarande är t.ex. en bit ner i kedjan och beroende av kvaliteten på primärdatakällorna och eventuella omtolkningar längs vägen. I de fall viktiga delar av källsystemen är tillräckligt flexibla och konfigurerbara så kan vi dock, om vi vill, i Sverige (över tid) till stor del samordna den primära datafångsten (och minska omtolkningsbehov) även om systemen internt inte bygger på specifikationer som openEHR (se presentationsbild 62-63 där möjligheter till samordning i nuvarande och kommande journalsystemsprodukter i Sverige beskrivs).

Det vore lämpligt att titta på de stöd som finns i HL7 FHIR och openEHR (kombinerade med SNOMED CT och andra terminologisystem/ontologier) vid all nyutveckling i GMS som hanterar patientspecifik data, i några följande rapportavsnitt beskriver vi bland annat exempel relaterade till sällsynta diagnoser och cancer. Hur dessa modeller och format relaterar till variantdatabaser beror lite på vad man vill göra och hur; det bör diskuteras vidare inom GMS. Vill man t.ex. lagra det mesta av data/metadata om patienten, provet och provresultatet inuti själva variantdatabasen så kan dessa standarder förhoppningsvis (om sändande system stöder dem) användas för att hämta ut data från källsystem. Dessutom kan de användas för att definiera eller inspirera delar av lagringsstrukturer i själva variantdatabasen och relaterade verktyg. Vill man å andra sidan lagra relativt lite data/metadata i själva variantdatabasen så kan man välja att främst länka variantdatabasposterna till källsystemens motsvarighet och vid behov via API hämta data från källan eller samköra frågor (se även rapportavsnittet om HiGHmed). Man bör även fundera på i vilka riktningar kopplingar görs mellan identifierad patientspecifik data och avidentifierad generell data.

Både HL7 FHIR och openEHR är breda standarder som syftar till att täcka mycket inom journalsystem, laboratoriesystem m.m. så därför är de grundläggande bitarna i standarderna (“resources” m.m. i FHIR respektive “Reference model” m.m. i openEHR) mycket generella för vård i allmänhet. För att komma ner på detaljnivå om t.ex. cancer så använder FHIR “extensions” och “profiles”, openEHR använder “archetypes” och “templates”, där man utöver att detaljera struktur även kan peka ut innehåll från terminologisystem och ontologier (t.ex. SNOMED CT och HPO). Både HL7 FHIR och openEHR beskriver även sätt att skapa och integrera beslutstöd. När genetisk information avsedd för både forskning och behandling behöver integreras med journalsystem och tillhörande beslutsstöd kommer dessa standarder sannolikt vara aktuella (se t.ex. review-artikeln Electronic health records for the diagnosis of

rare diseases som diskuterar olika sätt att göra detta och refererar till bland annat openEHR, FHIR, SNOMED CT och HPO).

6.4. Nationella tjänsteplattformen och Ineras ”tjänstekontrakt”

Svenska befintliga specifikationer och en tjänst som används inom hälso- och sjukvård för överföring till bland annat tjänster som Nationell Patientöversikt och Journalen via 1177. De flesta vårdgivare är redan inkopplade. Se mer information om Nationella tjänsteplattformen och leveranser för laboratoriemedicinska svar.

6.5. openEHR

Allmän information om patientjournals-standarden openEHR nås på svenskt forum för openEHR. Svensk förvaltning etableras i SFMI:s regi, en av SLS medlemsföreningar. Journalsystemsleverantören Cambio vars journalsystem Cosmic snart används i alla svenska regioner utom Skåne, Stockholm och Västra Götalandsregionen använder openEHR i delar av sitt system och håller på att växla över större delen av journalsystemet till openEHR. Karolinska Universitetssjukhuset i Region Stockholm skapar journalfunktioner och datalager baserade på openEHR och det är sannolikt att hela Region Stockholms framtida journalsystem kommer vara openEHR-baserat. Även för system som inte är openEHR-baserade, t.ex. Cerner Millenium som håller på att införas i Region Skåne och Västra Götalandsregionen, så kan man samarbeta om att ta fram openEHR-inspirerade konfigurationsunderlag[6] till formulär etc. i systemen på ett sätt som gör det lättare att återanvända samma mallstrukturer mellan vårdgivare och att dela information i openEHR-format eller med gemensamma mappningar mellan openEHR och FHIR.

Patologi-rapporteringsstrukturer, som är en sannolik källa för GMS-relevant ”metadata” inom cancerområdet (solida tumörer), håller på att tas fram i openEHR-format i ett nationellt projekt där bland annat Regionalt Cancercentrum Väst (RCC Väst) är en huvudpartner. RCC Väst förvaltar INCA-plattformen, en teknisk plattform för kvalitetsregister.

Ur GMS perspektiv innebär detta en chans att sådan metadata som traditionellt sett kommer ur patientjournal, laboratorieinformations- samt remiss- och svarssystem kan komma att fås på ett redan standardiserat sätt från källsystemen och att det för denna kategori av information vore klokt att använda kompatibla strukturer (t.ex. openEHR:s förenklade JSON-format) för lagring av genomikdata.

Genomik allmänt

- Forskningsartikeln OpenEHR modeling for genomics in clinical practice, innehåller en introduktion och även resonemang om kopplingar till arbetsflöden kring varianter och VCF.
- Exempel på hierarkisk struktur av “template” innehållande “archetypes” i en “Molecular Pathology Report” i engelskt/internationellt respektive tyskt samarbete, som innehåller beskrivning av genetiska varianter m.m. Man kan även se strukturerna i relaterat engelskt och tyskt exempel i formulär-liknande vy. Användbarhet och mer dynamisk inmatning går att optimera betydligt bättre än exemplet med andra openEHR-baserade verktyg, så “formuläret” ska inte ses som ett realistiskt gränssnitt, dessutom vill man sannolikt integrera datakällor så att allt som går förifylls automatiskt.

[6] Configuration of input forms in EHR systems using spreadsheets, openEHR archetypes and templates: https://ebooks.iospress.nl/publication/52410

Sällsynta diagnoser och Cancer, exempel på openEHR-användning

- Genomics England Phenotypic Datasets datasets for Rare Diseases and Cancer. (Se även avsnittet om Genomics England i denna rapport).
- Forskningsartikeln Development of a pilot rare disease registry: a focus group study of initial steps towards the establishment of a rare disease ecosystem in Slovenia nämner att de baserar sig på bland annat openEHR...
- ...och i en relaterad "rare disease"-diskussion i openEHR-forumet tipsas det om brittiska exempel på "template" baserad på EPIRARE (fenotypdata)
- HiGHmed:s onkologiprojekt delar öppet sina "archetypes" och "templates" och är öppna för samarbeten.
- Norska journalsystemsleverantören DIPS använder openEHR för att fånga data som sedan överförs till Kreftregistret (cancerregistret).

Svenska exempel (inom sällsynta diagnoser & Cancer)

- Karolinska Institutet och Region Stockholm jobbar en del med openEHR. Under 2021 genomfördes exempelvis openEHR-modellering kring bröstcancer.
- Linköpings Universitet och Region Östergötland har arbetat med openEHR- och Snomed CT-baserade patologi-svarsmallar för bröstcancer

6.6. HL7 FHIR

Allmän information om integrationsstandarden nås från HL7 FHIR:s hemsida. Svensk koordination sker via föreningen HL7 Sweden och exempelvis INERA arbetar med FHIR för överföringsändamål. Alla större journalsystem i Sverige har stöd för eller kommer att stödja FHIR för att exportera/importera vissa datamängder. Vi har i denna rapport inte hunnit granska vilka datamängder/kategorier av intresse för GMS som stöds i dagsläget.

Genomik allmänt

- Genomics Implementation Guidance från HL7 FHIR ger en bra översikt och beskriver en "Molecular Sequence Resource" samt hur "profiles" kan användas inom genetik, t.ex. Observation-genetics, DiagnosticReport-genetics, ServiceRequest-genetics HLA-genotyping-results.
- Genetic Reporting Implementation Guide innehåller bland annat ett avsnitt om variant reporting.

Sällsynta diagnoser och cancer, exempel på FHIR-användning

- I ett av HL7:s europeiska nyhetsbrev (sid 24-26) beskrivs samarbete om sällsynta diagnoser via RD-Code.
- Initiativet Minimal Common Oncology Data Elements (mCODE) samarbetar med bland annat FHIR och försöker beskriva viktiga onkologi-relaterade dataelement och samband i journalsystem och exempel på användningsfall.

Svenska exempel (inom sällsynta diagnoser och cancer)

- Socialstyrelsens rapport "Att kunna följa patientens väg genom vården" nämner bland annat FHIR Care Plans som en möjlig väg till standardisering.
- Umeå Universitet: Kandidatuppsats i datavetenskap om semiautomatisk mappning av patologidata till HL7 FHIR.

6.7. OHDSI

Samarbetet Observational Health Data Sciences and Informatics (förkortat OHDSI, som ska uttalas "Odyssey") har genom sin önskan att forska på journalinnehåll m.m. liknande mål som de beskrivna engelska och tyska projekten, men siktar mer på global storskalig analys via multicenterstudier m.m. (vilket även GA4GH delvis gör). Inom OHDSI mappar/översätter man olika källsystems innehåll till den gemensamma förenklade forskningsinriktade modellen "OMOP". Man håller nu på att växla sätt att i den beskriva genomik (inklusive varianter) från ett sätt (med överskådligt diagram) till ett nytt samt har en arbetsgrupp för fenotypning och en för onkologi.

6.8. ISO, CEN, SIS – officiella standardiseringsorgan

International Organization for Standardization (ISO) och motsvarande europeiska (CEN) och svenska (SIS) standardiseringsorgan håller på med flera standardiseringsarbeten relaterade till genomik, bioinformatik m.m. ISO samarbetar även med GA4GH och andra initiativ/ organisationer inom området. SIS tekniska kommitté för genomik och precisionsmedicin (SIS TK/620) där GMS medverkar, håller på att inventera och bekanta sig med befintliga och kommande ISO-arbeten inom området. När den inventeringen är klar blir det lättare att se vilka standarder som skulle kunna vara relevanta för arbete med variantdatabaser, relaterade system och integrationer.

7. Förkortningar

Nedan listas några av de förkortningar som används i denna rapport.

Förkortning	Beskrivning
AACR	American Association for Cancer Research
ACMG	American College of Medical Genetics and Genomics
API	Application Programming Interface
ASCO	American Society of Clinical Oncology
ASH	American Society of Hematology
BAM	Binary Alignment Map
BIC	Breast Cancer Information Core
CCDS	Consensus CDS protein set
CIViC	Clinical Interpretation of Variants in Cancer
ClinGen	Clinical Genome Resource
CRAM	The Genomics Compression Standard
dbNSFP	Database for Non-Synonymous SNPs and Functional Prediction
dbSNP	The Single Nucleotide Polymorphism Database of Nucleotide Sequence Variation
dbVar	Database of human genomic structural Variation
DECIPHER	DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources
DICOM	Digital Imaging and Communications in Medicine
DO	Disease Ontology
DRS	Data Repository Service
EBI	European Bioinformatics Institute
EFO	Experimental Factor Ontology
ENIGMA	Evidence-based Network for the Interpretation of Germline Mutant Allele
Entrez	The Entrez Molecular Sequence Database System (at NCBI), providing access to a number of databases within NCBI.
Entrez Gene	The Gene database (at NCBI), accessible via NCBI's Entrez database system.
ESMO	European Society for Medical Oncology
ESP	Exome Sequencing Project
ExAC	The Exome Aggregation Consortium

FAB	French-American-British
FDA	The U.S. Food and Drug Administration
FTP	File Transfer Protocol
GA4GH	Global Alliance for Genomics & Health
GDC	Genomics Data Commons
GENIE	Genomics, Evidence, Neoplasia, Information, Exchange
GRCh37/38	Genome Reference Consortium Human Build 37/38
gnomAD	The Genome Aggregation Database
HGMD	Human Gene Mutation Database
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HL7	Health Level Seven
HL7 FHIR	Health Level Seven standard Fast Healthcare Interoperability Resources
HPO	Human Phenotype Ontology
HUGO	Human Genome Organisation
ICD-O	International Classification of Diseases for Oncology
ICGC	International Cancer Genome Consortium.
ID	Identifikation
IGV	Integrative Genomics Viewer
ISCN	International System for Human Cytogenetic Nomenclature
InSiGHT	International Society for Gastrointestinal Hereditary Tumours
ISO	International Organization for Standardization
LOVD	Leiden Open Variation Database
LRG	Locus Reference Genomic
MeSH	Medical Subject Headings
MSK	Memorial Sloan Kettering Cancer Center
NCBI	National Center for Biotechnology Information
NCCN	National Comprehensive Cancer Network
NCI	National Cancer Institute
NCIt, NCI Thesaurus	National Cancer Institute Thesaurus
NHS	National Health Service
NIH	National Institute of Health
OMIM	Online Mendelian Inheritance in Man
OncoKB	Oncology Knowledge Base
OncoTree	Oncology Tree
REAL	Revised European American Lymphoma
RefSeq	Reference Sequencing database
SAM	Sequence Alignment Map
SEPIO	Scientific Evidence and Provenance Information Ontology
SFMI	Svensk Förening för Medicinsk Informatik
SIS	Svenska institutet för standarder
SLS	Svenska Läkaresällskapet
SNOMED	Systematized Nomenclature of Medicine
SNOMED CT	SNOMED Clinical Terms
SO	Sequence Ontology
SOP	Standard Operating Procedure
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
TES	Task Execution Services
TSV	Tab Separated Values
UMIN	University Hospital Medical Information Network
UMLS	Unified Medical Language System
UniProt	Universal Protein Resource
UTA	Universal Transcript Archive
VCF	Variant Call Format
VRS	Variation Representation Specification

WES	Workflow Execution Service
WHO	World Health Organization

Bilaga A. Information i internationella databaser och projekt

ClinVar

A ClinVar record contains the following elements, as per defined on the ClinVar website, https://www.ncbi.nlm.nih.gov/clinvar/intro/:

ClinVar Accession and version

1. Submission accession number/version number separated by a decimal (SCV000000000.0) assigned to each submitted record.
2. Reference accession number/version separated by a decimal (RCV000000000.0) assigned to sets of submitted records about the same variation/condition pair.
3. Variation accession number/version separated by a decimal (VCV000000000.0) assigned to sets of submitted records about the same variation.

Identifiers for each variant allele or allele set

1. HGVS expressions
2. Published allele names
3. Database identifiers

Attributes of each phenotype

1. Name
2. Descriptions
3. Defining features
4. Database identifiers

Description of the genotype/phenotype relationship

Review status of the asserted relationship

1. Submitter of the assertion
2. Clinical significance
3. Summary of the evidence for clinical significance
    1. Number of observations of genotype/allele in those with the phenotype
    2. Number of observations of genotype/allele in those without the phenotype
    3. Family studies
    4. Description of the population sampled
    5. In vitro studies
    6. In silico studies
    7. Animal models
4. Mode of inheritance
5. Study design
6. Citations, including URLs

Submission information

1. Submitter description
2. Dates submitted and updated
3. Data added by NCBI computation

DECIPHER

Struktur DECIPHER records

Patient records och gene records följer liknande struktur med avseende på metadata.

Overview tab

- DECIPHER ID
- Kromosomalt kön och/eller fenotypiskt kön
- Patientens ålder vid senaste bedömning
- Open access data (ja/nej)

Genotype tab (se även rubrik Sequence Variant Data)
Översikt av hos patienten identifierade genvarianter;

- Genomkoordinater, varianttyp, gennamn, variantstorlek, annotering, ärftlighetsmönster/genotyp, patogenicitet och i förekommande fall bidragandegrad.
- Länkar (Ensembl, USCS)

Phenotypes tab

- Associerad fenotyp hos patient och i förekommande fall föräldrar

Assessments tab

- Kliniska bedömningar (datum, status, varianter, fenotyper, konklusioner)

Karyotype tab

- Karyotypdata

Citations tab

- Referenser

Contact tab

- Information för kontakt med ägare av respektive patient record

Sequence variant data

Browser tab
Genombrowser där tracks representerande olika typer av data kan väljas för att visualisera bland annat:

- Loss intolerance, haplo-insuffiency m.m.
- Sekvensvarianter identifierade i patienten
- Konservering över arter
- Sekvensvarianter och strukturella varianter identifierade i andra DECIPHER-patienter
- Varianter associerade med kända sjukdomar enligt GeneReviews och OMIM
- Predikterad konsekvens, allelfrekvenser och coveragedata utifrån gnomAD
- Sekvensvariantdata från HGMD, ClinVar, LSDB, dbSNP med flera källor

Gene tab

Clinical;

- Gen/sjukdom-association:
  - OMIM, Morbid, The Developmental Disorders Genotype-Phenotype Database (DDG2P), ClinGen
- Predictive scores:
  - pLI, LOEUF och sHet (LOF-tolerans); (pHI (%HI): Probability of being a haplo-insufficient gene

- Sökdatabaser:
    - PubMed, Genetic Testing Registry (GTR), Genomics England PanelApp, Locus-specific databases (LSDB), DECIPHER entries
- Kvantitativ data
    - Ackumulerad kvantitativ data för varianter som hos DECIPHER-patienter bedömts som sannolikt patogena eller patogena. Resultaten för en specifik genvariant jämförs mot kvantitativ data från alla DECIPHER-patienter med potentiellt patogena sekvensvarianter.

Protein/Genomic;

- Proteinidentifierare (UniProt, InterPro, PFam, 3D Structures (PDB))
- Genomidentifierare (Ensembl, UCSC, RefSeq, NCBI, HGNC)
- Sökdatabaser (Expression Atlas, GeneCards, The Human Protein Atlas, DECIPHER entries)
- Protein: interaktiv proteinbrowser (se nästa stycke)

Protein tab

- Protein-viewer; visar DECIPHER sekvensvarianter mappade till protein, vid sidan av (om tillgängligt):
    - Proteinstrukturdata (PFam, Uniprot, PDB/Uniprot)
    - Proteinförändrande och loss-of-function-varianter från ClinVar och gnomAD
    - Missense contraint-data

Annotation tab

- Konsekvensprediktion (Ensembl Variant Effect Prediction (VEP) samt SIFT, PolyPhen, CADD, REVEL och SpliceAI)
- Allelfrekvenser (gnomAD)
- Transkript och proteinförändring (RefSeq, Ensembl/GENCODE)

Matching patients tab

- DECIPHER-patienter med varianter som är funktionellt identiska med aktuell variant, eller matchande sekvens-/copynumber-varianter i samma gen

Matching CNV syndromes tab

- Matchande microdeletion/microduplication-syndrom

Pathogenicity evidence tab

- Evidens för patogenicitet så som registrerats av patientdataägaren

Database structure

A HGMD entry consists of:

1. HGMD accession number
2. Reported disease/phenotype (as reported in literature)
3. Variant class
4. Gene symbol (HUGO)
5. Codon change
6. Amino acid change
7. Codon number
8. Reference to first literature report of a mutation
9. Literature citation (text + PMID)
10. Citation type
11. Support
12. Notes
13. Coding strand genomic sequence
14. Genomic coordinate (specified build)
15. HGVS nomenclature
16. HGMD variant class
17. Annotation data. Some selection:
    - PolyPhen2
    - SIFT
    - CADD
    - gnomAD
    - 1000 Genomes

Genes:

1. cDNA sequence (RefSeq)
2. RefSeqGene
3. Reference to first literature report of a mutation
4. Associated disease state as specified in report
5. Gene name
6. Gene symbol (HUGO)
7. Chromosomal location
8. Gene Ontology

gnomAD

Variants

Divided Exomes/Genomes:

1. Filter flag
2. Allele count
3. Allele number
4. Allele frequency
5. Number of homozygotes

Annotations:

1. Number of transcripts
2. Annotation per transcript
    - Ensembl iD
    - HGVSp
    - Polyphen
    - SIFT

Population frequencies

1. Number of alleles in different populations in the dataset

Age distribution

1. Split per Exome/Genome and Heterozygos/Homozygos

Quality metrics per individual (split Exome/Genome)

1. Genotype quality
2. Depth
3. Allele balance for heterozygotes

Quality metrics per variant (split Exome/Genome)

1. BaseQRankSum
2. ClippingRankSum
3. DP
4. FS
5. InbreedingCoeff
6. MQ
7. MQRankSum
8. Pab_max
9. QD
10. ReadPosRankSum
11. RF
12. SiteQuality
13. SOR
14. VQSLOD

ClinGen

Resources that provide core Allele record keeping - focused on the registration and canonicalization to support Allele indexing.

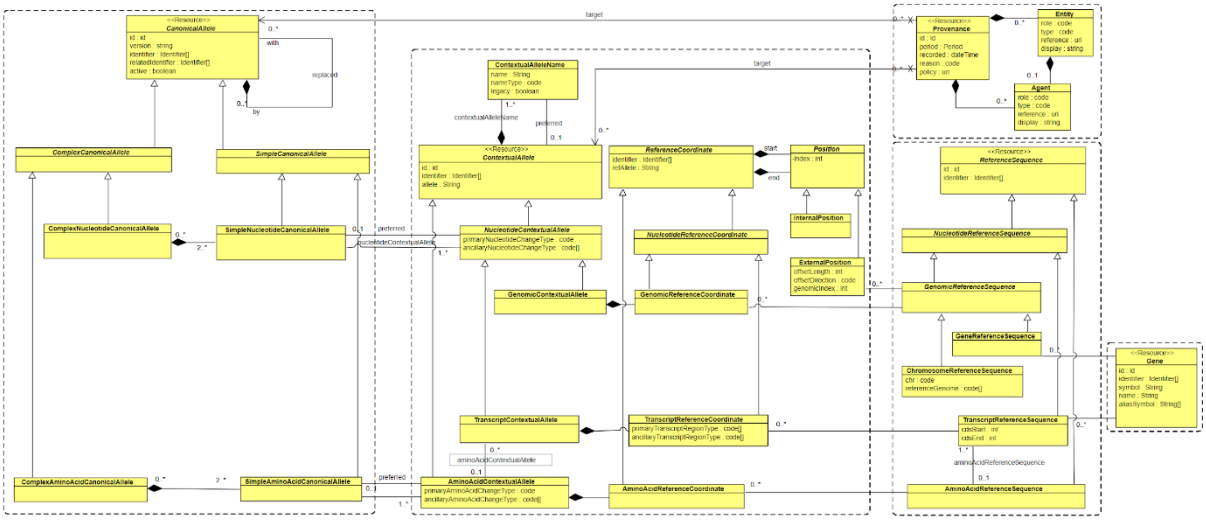
Allele

Resources that provide core Allele record keeping - focused on the registration and canonicalization to support Allele indexing.

Name	Aliases	Description
CanonicalAllele	allele identity, measure set, normalized variant identity	One of a set of coexisting sequence variants.
ContextualAllele	contextual allele, allele, variant, sequence variant, snv, amino acid variant, dna change, aa change	A representation of one of the multiple variant sequences at a contiguous region in a particular ReferenceSequence.
Gene		A genomic region related to a collection of transcript ReferenceSequences, given a name by one or more naming agencies.
Provenance		Provenance describes the context in which a resource was created.
ReferenceSequence	reference sequence, refseq, sequence, accession	A versioned sequence of nucleotide bases or amino acids.

http://dataexchange.clinicalgenome.org/allele/resource/

Full conceptual model:



dbVar

Struktur dbVar
Varje studie/variant tilldelas ett ID enligt följande:

- (n|e|d)std för studier inskickade till NCBI, EBI respektive DDBJ
- (n|e|d)sv: strukturvariant-ID
- (n|e|d)ssv: supporting variant-ID

Study browser
Browserdata kan sorteras/filtreras enligt:

- Publikation och publikationsdatum
- Tilläggsidentifierare enligt ovan
- Studietyp (exempelvis case-control, somatic, tumor vs. matched normal)
- Metod (exempelvis Sequencing, SNP-array)
- Organism
- Antal variantregioner eller antal variant calls

dbVar Study Page
Allmän information om studien [(n|e|d)std000]:

- Organism
- Studietyp
- Dataleverantör
- Beskrivning
- Publikationslänkar
- Datum för senaste uppdatering

Detailed information/Variant Summary tab:

- Nedladdningsbar data
- Variant Summary tab
    - Sequence ID (NC.00000.0)
    - Antal variant calls och variantregioner för varje kromosom
    - Placement type (angiven assembly och/eller remappade assemblies)
    - Länkar till grafisk visualisering av varianterna i NCBI's Sequence Viewer

Samplesets Tab:

- Detaljer inkluderande namn, beskrivning, materialstorlek och i förekommande fall relevanta fenotyper

Experimental details tab:

- Varje unik kombination av metod och analys anges som ”Experiment”.
- Detaljer kring metoder och analyser som användes i studien (ex SNP array, Affymetrix Cytoscan HD) samt antal variant calls per experiment

Validations tab:

- I förekommande fall beskrivning av valideringsexperiment som tilldelats egna unika ID:n.

dbVar Variant Page
Allmän information om variantregionen:

- Organism
- Studie [n|e|dsv]
- Varianttyp
- Metod
- Antal variant calls som stödjer varianten

- Regionstorlek
- Valideringsinformation
- Kliniska påståenden/patogenicitet
- Datum och publikationslänkar
- Kromosomideogram

Genome View tab:
- Visar den aktuella varianten i NCBI:s Sequence Viewer i kontexten kända gener och övrig variantdata i samma studie

Variant Region Details and Evidence Tab
- Visar koordinaterna för variantregionen på den assembly som dataägaren angivit samt assemblies som den senare mappats till, HGVS, sekvens-ID (NC.000000.0).
- Score-kolumnen anger kvalitén på en eventuell remappning
- Detaljer över supporting variant calls som definierar regionen, inklusive genomposition, kopplad fenotyp och ClinVar-ID.

Validation information tab:
- I förekommande fall valideringsresultat, med detaljer om metoder och analyser samt utfall

Clinical assertions tab:
- Information om klinisk relevans – koppling variant-fenotyp om sådan fastslagits;
    - Variant call ID samt ClinVar ID (inkl länk)
    - Typ av event (insertion, deletion etc.)
    - Allele origin (parental)
    - Fenotyp
    - Författarens bedömning av patogeniciteten

Genotype information tab:
- I förekommande fall genotyp/allelfrekvensdata

Data download:
- FTP site, data organiserad per assembly och per studie

dbVar Entrez sök:
- Accession – interna och externa identifierare
- Ålder samt genus/kön på studiesubjekt
- Accession/namn på assembly
- Accession – identifierare/namn på kopplat BioProject
- Kromosom samt Start/end på kromosom för varianten
- Klinisk tolkning samt Fenotyp
- Global VAF
- Namn/alias för gen
- MeSH termer kopplade till publikation samt NLM MeSH Browser-unikt ID
- Metodtyp
- OMIM
- Objekttyp i dbVar (study, variant)
- Allelursprung (germline, somatic)
- Range för överlapp med patogen variant
- Studiepopulation
- Publikationsnamn, författare och/eller publikationsdatum samt PubMed-ID
- Prov-ID för studiesubjekt eller referensprov
- Antal prover (N) för studien

- Studie-ID, alias, visningsnamn
- Studietyp tilldelad av NCBI
- Taxonomi-ID
- Unikt nummer som tilldelats studien eller varianten i Entrez
- Variantstorlek samt typ av variantregion eller call
- Zygositet

TCGA

Se GDC data portal.

GDC data portal

Listed below are data submittable to the GDC. Note that not all programs or projects will have data available for all types.

Entity Category	Entity Name	File Format	File Metadata Template
Administrative	Case	--	TSV, JSON
Biospecimen	Sample	--	TSV, JSON
	Portion	--	TSV, JSON
	Analyte	--	TSV, JSON
	Aliquot	--	TSV, JSON
	Read Group	--	TSV, JSON
	Slide	--	TSV, JSON
Clinical	Demographic	--	TSV, JSON
	Diagnosis	--	TSV, JSON
	Exposure	--	TSV, JSON
	Family History	--	TSV, JSON
	Follow Up	--	TSV, JSON
	Molecular Test	--	TSV, JSON
	Treatment	--	TSV, JSON
Data File	Analysis Metadata	SRA XML, MAGE-TAB (SDRF, IDF)	TSV, JSON
	Biospecimen Supplement	BCR XML, GDC-approved spreadsheet	TSV, JSON
	Clinical Supplement	BCR XML, GDC-approved spreadsheet	TSV, JSON
	Experiment Metadata	SRA XML	TSV, JSON
	Pathology Report	PDF	TSV, JSON
	Run Metadata	SRA XML	TSV, JSON
	Slide Image	JPEG, SVS, TIFF	TSV, JSON
	Submitted Unaligned Reads (Illumina Platform)	FASTQ, BAM	TSV, JSON
	Submitted Aligned Reads (Illumina Platform)	BAM	TSV, JSON
	Submitted Genomic Profile	MAF, TSV, VCF, XML	TSV, JSON
	Raw Methylation Array	IDAT	TSV, JSON

Listed in the table below is the data generated from the various in-house pipelines run on all data that is submitted to the GDC.

Entity Category	Entity Name	Access (Open, Controlled)	File Format	File Metadata Template
Analysis	Read Group QC	--	--	TSV, JSON
	Alignment + Co-cleaning	--	--	TSV, JSON
	Alignment	--	--	TSV, JSON
	Genomic Profile Harmonization	--	--	TSV, JSON
	RNA Expression	--	--	TSV, JSON

	miRNA Expression	--	--	TSV, JSON
	Germline Mutation Calling	--	--	TSV, JSON
	Somatic Mutation Calling	--	--	TSV, JSON
	Structural Variation Calling	--	--	TSV, JSON
Data File	Aggregated Somatic Mutation	Controlled	MAF	TSV, JSON
	Aligned Reads	Controlled	BAM	TSV, JSON
	Gene Expression	Open	TSV	TSV, JSON
	Masked Somatic Mutation	Open	MAF	TSV, JSON
	miRNA Expression	Open	TSV	TSV, JSON
	Structural Variation	Controlled	TSV	TSV, JSON

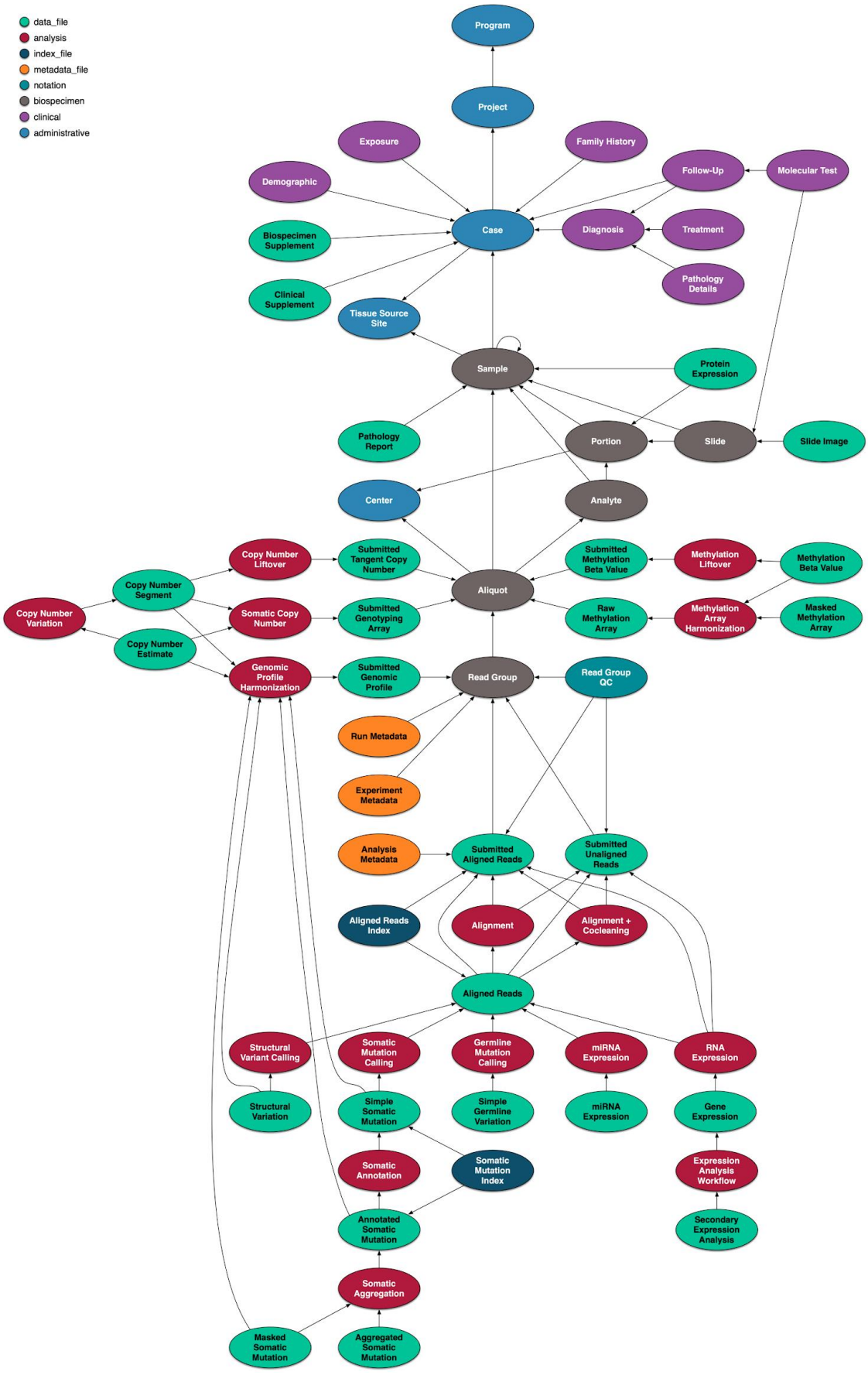


Figure. Graph representation of the GDC data model. https://gdc.cancer.gov/developers/gdc-data-model

Evidence Items Gene Variant Statement Evidence Level (validated association, clinical evidence, case study, pre-clinical evidence, inferential association) Type (predictive, diagnostic, prognostic, predisposing, functional) Direction (supports, does not support) Clinical Significance Origin (somatic, rare germline, common germline, unknown, N/A) Disease Associated Phenotype Drug Drug Interaction Type Citation PubMed ID Clinical Trial Evidence Rating **Assertions** Assertion Statement Assertion Description Gene Variant Origin Disease Type Direction Clinical Significance Drug Associated Phenotype Clinical Trial AMP Category ACMG Codes NCCN Guideline FDA Approval & Test	**Variants** Name Summary Alias HGVS Expression ClinVar ID Sources (PubMed IDs) CIViC Actionability Score Variant Type(s) Primary Coordinates: - Reference Build - Chromosome - Start, Stop - Reference & Variant Bases - Representative Transcript Secondary Coordinates: - Same as Primary MyVariant Info **Genes** Name Summary Sources (PubMed IDs) MyGene Info **Variant Groups** Name Summary Sources (PubMed IDs) Variants **Sources** Citation Title Authors Abstract Publication Date Source ID Journal PMC ID Status

Cancer Hotspots

The following data, regarding identified hotspots, is visible via Cancer Hotspot's website.

1. Gene
2. Residue
3. Type (single residue, in-frame indel)
4. Variants
5. Q-value
6. Samples
7. Tumor Type Composition

The resulting data of the second analysis, contains the following data set.

SNV-hotspots	**INDEL-hotspots**
1. Hugo_Symbol 2. Amino_Acid_Position 3. log10_pvalue 4. Mutation_Count 5. Reference_Amino_Acid 6. Total_Mutations_in_Gene 7. Median_Allele_Freq_Rank 8. Allele_Freq_Rank 9. Variant_Amino_Acid 10. Codon_Change 11. Genomic_Position 12. Detailed_Cancer_Types 13. Organ_Types 14. Tri-nucleotides 15. Mutability 16. mu_protein 17. Total_Samples 18. Analysis_Type 19. qvalue 20. tm 21. qvalue_pancan 22. Is_repeat 23. seq 24. length 25. align100 26. pad12entropy 27. pad24entropy 28. pad36entropy 29. TP 30. reason 31. n_MSK 32. n_Retro 33. judgement 34. inNBT 35. inOncokb 36. ref 37. qvaluect 38. ct 39. Samples	1. Hugo_Symbol 2. Amino_Acid_Position 3. log10_pvalue 4. Mutation_Count 5. Reference_Amino_Acid 6. Total_Mutations_in_Gene 7. Median_Allele_Freq_Rank 8. Allele_Freq_Rank 9. SNP_ID 10. Variant_Amino_Acid 11. Codon_Change 12. Genomic_Position 13. Detailed_Cancer_Types 14. Organ_Types 15. Tri-nucleotides 16. Mutability 17. mu_protein 18. ccf 19. Total_Samples 20. indel_size 21. qvalue 22. tm 23. Is_repeat 24. seq 25. length 26. align100 27. pad12entropy 28. pad24entropy 29. pad36entropy 30. TP 31. reason 32. n_MSK 33. n_Retro 34. judgement 35. inNBT 36. inOncokb 37. Samples

OncoKB

Gene Gene summary Gene background Gene alias, isoform and RefSeq number Classifying a gene as oncogene or tumor suppressor **Alteration** Name[7] Oncogenic effect of an alteration[8] Biological effect of an alteration[9] **Tumor type** Tumor type	**Tumor type-specific clinical implications** Clinical summary Diagnostic implications: - Level of evidence[10] - Description of evidence - Additional information (optional) Prognostic implications: - Level of evidence[11] - Description of evidence - Additional information (optional) Implications for sensitivity or resistance to therapy: - Therapy - Level of evidence[12] - Description of evidence - Additional information (optional)

[7] Naming conventions are described in the OncoKB curation SOP. They concern alterations such as missense mutations, truncating mutations, fusions, copy number aberrations, in-frame deletions or insertions, oncogenic mutations, tumor suppressors, oncogenes, hard-coded alteration names and hotspot mutations.
[8] Oncogenic, likely oncogenic, likely neutral, inconclusive
[9] Gain-of-function, likely gain-of-function, loss-of-function, likely loss-of-function, switch-of-function, likely switch-of function, neutral, likely neutral, inconclusive
[10] Dx1 = FDA and/or professional guideline-recognized biomarker required for diagnosis in this indication.
Dx2 = FDA and/or professional guideline-recognized biomarker that supports diagnosis in this indication.
Dx3 = Biomarker that may assist disease diagnosis in this indication based on clinical evidence.)
[11] Px1 = FDA and/or professional guideline-recognized biomarker prognostic in this indication based on well-powered studies.
Px2 = FDA and/or professional guideline-recognized biomarker prognostic in this indication based on a single or multiple small studies.
Px3 = Biomarker is prognostic in this indication based on clinical evidence in well-powered studies.)
[12] Level 1 = FDA-recognized biomarker predictive of response to an FDA-approved drug in this indication.
Level 2 = Standard care (NCCN or other expert panels) biomarker predictive of response to an FDA-approved drug in this indication.
Level 3A = Compelling clinical evidence supports the biomarker as being predictive of response to a drug in this indication, but neither biomarker nor drug are standard care.
Level 3B = Standard care or investigational biomarker predictive of response to an FDA-approved or investigational drug in another indication.
Level 4 = Compelling biological evidence supports the biomarker as being predictive of response to a drug.
Level R1 = Standard care biomarker predictive of resistance to an FDA-approved drug in this indication.
Level R2 = Compelling clinical evidence supports the biomarker as being predictive of resistance to a drug.

Omic data types	**Mutation (continued)**
Mutations	Start Pos
Clinical data	End Pos
DNA copy-number alterations	RefVar
mRNA expression	HGVSg
DNA methylation	HGVSc
Protein och phosphoprotein levels	Mutation states
	Validation status
Study	Mutation type[15]
Name of study	Variant type (ex. SNP, DEL)
Description of study	Center
Number of patients	Allele frequency in the tumour sample
Number of samples	Variant reads
	Ref reads
Gene	Variant reads (normal)
Gene name	Ref reads (normal)
RefSeq ID	Copy number status[16]
Ensembl ID	mRNA Expr
CCDS ID	- mRNA z-score
UniProt ID	- Percentile
	Cohort (mutation frequency in cohort)
Mutation	Number of COSMIC occurrences
Sample ID	Exon
Cancer type	gnomAD population allele frequencies
Gene	- Population (European Non-Finnish, South Asian, East Asian, Ashkenazi Jewish, European Finnish, Latino, African, Other)
Gene panel	- Allele count
Protein change	- Allele number
Annotation OncoKB, CIViC, My Cancer Genome and Cancer Hotspots	- Number of homozygotes
Functional impact:	- Allele frequency
- Source[13]	ClinVar ID
- Impact[14]	dbSNP ID
- Score	
Chromosome	

[13] MutationAssessor, SIFT, PolyPhen-2
[14] MutationAssessor: high, medium, low, neutral.
SIFT: deleterious, deleterious_low-confidence, tolerated_low_confidence, tolerated.
PolyPhen-2: probably_damaging, possibly_damaging, benign.
[15] Ex. missense, nonsense, splice site, frameshift insertion or deletion, in-frame insertion or deletion, nonstop, nonstart, fusion.
[16] Ex. low-level gain, high-level amplification, diploid/normal, shallow deletion.

Copy number alteration Gene Gene panel Copy number alteration[17] Annotation OncoKB, CIViC, My Cancer Genome and Cancer Hotspots Cytoband mRNA Expr - mRNA z-score - Percentile Cohort (alteration frequency in cohort) **Patient** Patient ID Cancer type Sex Age Race category Metastatic site Oncotree code Overall survival (months) Smoking history	**Samples** Mutation count Cancer type Cancer type detailed Oncotree code Primary depth Primary mitoses Primary tumor site Sample type Immunohistochemistry Gene panel

[17] Ex. homozygously deleted, heterozygously deleted, diploid, gained, amplified

Clinical trials
Clinical trial title
Clinical trial identifier (NCT ID)
Description
Related conditions
Recruiting status[18]
Phase[19]
URL to the clinical trial
Trial eligibility
- Summary of cancer disease and associated biomarker criteria
- Disease states
- Disease criteria
- Biomarker criteria

Treatment context
- Therapies
- Therapeutic context

Document
Title
- Brief title
- Official title

Clinical trial IDs
- Org study ID
- Secondary ID
- NCT ID

Conditions
Interventions
- Drug
- Synonyms
- Arms

Purpose of study
Trial arms
- Name
- Type
- Description
- Intervention

Eligibility criteria
- Inclusion criteria
- Exclusion criteria
- Maximum eligible age
- Minimum eligible age
- Eligible gender
- Healthy volunteers

Document (continued):
Primary outcome measures
- Measure
- Time frame
- Safety issue
- Description

Secondary outcome measures
- Same as Primary

Details
- Phase
- Primary purpose
- Overall status
- Lead sponsor

Diseases
Disease name
Associated genetic biomarkers
Overview
- NCI definition

Biomarker-directed therapies
- Summary
- Biomarker criteria
- Clinical setting(s)
- Note

Clinical trials
- Summary significant genes
- Gene name
- Summary

Disease details
- Synonyms
- Parent(s)
- Children

References

[18] Recruiting, Active not recruiting, Completed, Not yet recruiting, Terminated, Unknown status, Withdrawn, Suspended, Enrolling by invitation, Closed.
[19] Early phase, Phase 1, Phase 2, Phase 3, Phase 3, Phase 4, N/A.

Biomarkers Biomarker name Associated diseases Associated genetic biomarkers Associated pathways Biomarker overview - Variant type - Gene - Summary - Location - Pathway - Affected exon number - Protein - Protein domain - SIFT prediction - Synonyms Biomarker-directed therapies - Summary - Biomarker criteria - Predicted response - Clinical setting(s) - Note Clinical trials - Summary of related clinical trials Significance of the biomarker in diseases - Disease name - Summary References	**Drugs** Drug name Associated genetic biomarkers Associated diseases Drug overview - Generic name - Trade name(s) - NCI definition Biomarker-directed therapies - Summary - Biomarker criteria - Predicted response - Clinical setting(s) - Note Clinical trials - Summary of related clinical trials Drug details - Synonyms - Drug categories - Drug target(s) - NCIT ID - SNOMED ID References **Pathways** Pathway name Upstream pathways Therapies Genes Pathway overview - Summary - Upstream pathways - Drug categories targeting the pathway Biomarker-directed therapies - Summary Clinical trials - Summary of related clinical trial References

Cytogenetic Characteristics Morphology Topography Karyotype Case no (local database number) Reference **Gene Fusions** Morphology Topography Abnormality Genes Immunophenotype (B-Lineage, T-Lineage) Reference **Clinical Association** Morphology Topography Abnormality Genes Immunophenotype (B-Lineage, T-Lineage) Reference **Structural Chromosomal Abnormalities** Aberration Type (balanced, unbalanced) Morphology Topography Genes Band Abnormality Cases **Numerical Chromosomal Abnormalities** Aberration Type (monosomy, trisomy) Morphology Topography Abnormality Cases	**Chromosome Abberation Case Info** ***(information of cases where an aberration was found)*** Reference No Case No (local database number) Karyotype Patient Characteristics: - Sex - Age - Ethnicity - Country - Series - Hereditary Disorder Present Tumor: - Topography - Immunophenotype - Morphology - Tissue Previous Tumor: - Topography - Morphology - Treatment **References** Ref No Title Authors Journal name Volume Year **Other** Investigation no (local database number)

Variant Names Gene Symbol HGVS Nucleotide HGVS RNA HGVS Protein Genome (GRCh38) Genome (GRCh37) Transcript Identifier Abbreviated AA Change BIC designation ClinGen Allele Registry GA4GH VRS Identifier **Clinical significance (ENIGMA)** Clinical Significance Date Last Evaluated Comment on Clinical Significance Clinical Significance – Citations Assertion Method Collection Method Allele Origin **Clinical significance (ClinVar)** Clinical Significance Submitter SCV Accession Date last updated (ClinVar) **Clinical significance (LOVD)** Submitter Clinical Classification Variant Data Type (Origin) Variant Frequency Individuals Variant ID (Database ID) Variant Haplotype Created Date Edited Date Variant Remarks **Clinical significance (BIC)** Patient Nationality Ethnicity Family members carrying this variant Literature Reference Allele Origin (Germline/Somatic)	**Multifactorial Likelihood Analysis (ExUV)** Posterior Probability of pathogenicity (ExUV) Prior probability of pathogenicity (ExUV) Missense analysis probability of pathogenicity (ExUV) Co-occurrence likelihood (ExUV) Segregation Likelihood Ratio (ExUV) Summary Family History Likelihood Ratio (ExUV) Pathology Likelihood Ratio (ExUV) Case-Control Likelihood Ratio (ExUV) **Allele frequencies (gnomAD)** Allele Frequency (gnomAD non-cancer cohort) African/African American (AFR) Latino (AMR) Ashkenazi Jewish (ASJ) East Asian (EAS) Finnish (FIN) Non-Finnish European (NFE) South Asian (SAS) Other (OTH) **Allele frequencies (ExAC)** Allele Frequency (ExAC minus TCGA) African/African American (AFR) Admixed American/Latino (AMR) East Asian (EAS) Finnish (FIN) Non-Finnish European (NFE) South Asian (SAS) Other (OTH) **Allele frequencies (1000 Genomes)** Allele Frequency AFR Allele Frequency AMR Allele Frequency EAS Allele Frequency EUR Allele Frequency SAS Allele Frequency **Allele frequencies (Exome Sequencing Project)** Allele Frequency (ESP) EA Allele Frequency (ESP) AA Allele Frequency (ESP) **Other – fields from Findlay et al** If applicable, the database contains information on functional assay results from Findlay et al and In Silico prior probabilities of pathogenicity.

Genes General Information	**Variants**
Gene symbol	Individual ID
Gene name	Chromosome
Chromosome	Allele
Chromosomal band	Affects function (as reported)
Imprinted	Affects function (by curator)
Genomic reference	Classification method
Transcript reference	Clinical classification
Exon/intron information	DNA change (s)
Associated with diseases	Published as
Citation reference (s)	ISCN
Refseq URL	DB-ID
Curators	Variant remarks
Total number of public variants reported	Reference
Unique public DNA variants reported	ClinVar ID
Individuals with public variants	dbSNP ID
Hidden variants	Origin
Notes	Segregation
Date created	FrequencyRe-site
Date last updated	VIP
Version	Methylation
	Average frequency (large NGS studies)
Transcripts	Owner
Transcript name	Gene
Gene name	Transcript ID
Chromosome	Exon
Transcript – NCBI ID	DNA change (cDNA)
Transcript – Ensembl ID	RNA change
Protein – NCBI ID	Protein
Protein – Ensembl ID	
Protein – Uniprot ID	**Individuals**
Exon/Intron information	ID_report
Remarks	Reference
	Remarks
	Gender
	Consanguinity
	Country
	Population
	Age at death
	VIP
	Data_av
	Treatment
	Panel size
	Diseases
	Owner name

Phenotype Individual ID Associated disease Phenotype details Diagnosis/Initial Diagnosis/Definite Inheritance Age/Examination Age/Diagnosis Age/Onset Phenotype/Onset Protein Owner name **Diseases** Official abbreviation Name OMIM ID HPO Project Inheritance Individuals reported having this disease Phenotype entries for this disease Associated with genes Associated tissues Disease features Remarks	**Screenings** Screening ID Template Technique Tissue Remarks Genes screened Variants found Owner name

Research Environment content (de-identified) ● Genome sequence data ● Variant call files ● Phenotype/clinical data ● Outputs from the Genomics England interpretation pipeline, such as tiering results ● Hospital Episode Statistics ● Diagnostic Imaging Dataset ● Patient Reported Outcome Measures ● Mental Health Services Data Set	**Clinical data** ● DNA source – germline/tumour ● Standard Operating Procedure version(s) used for collection, extraction, QC and logistics. ● Tumour type (if cancer– primary/recurrent/metastatic) ● Topography and Morphology (if cancer) ● Tumour cellularity percentage (if cancer) ● DNA concentration (nanogrammes/ ● microliter) ● DNA volume (microliters) ● DNA QC metrics (to be finalised with sequencer) ● Genomics England identifier for data subject (site ID, local patient ID (NHS Number), sample ID) ● Gender and ethnicity ● Date of birth ● Pedigree structure and affectation status (rare disease) ● Version of Consent and Patient Information Sheet used ● Syndrome or disease name ● Named site contact and email (to verify receipt of samples by collection hubs, for queries etc.)