Cancer Community Presentations

Presentation Expectations

With the mission of the Cancer Community in mind, all presentations provided at the GA4GH Cancer Community should aim to address the following points in 30 minutes or less:

- Background on your initiative/project
- Are there any GA4GH standards that you already use?
- Relevant use cases that may benefit from the implementation of GA4GH standards (see use case template under 'Resources')
- How can the community help you? (e.g. feedback on implementation, advice for other standards that may be relevant, help make a connection with one or more of the Work Streams)

Discussion following the presentation will focus on:

- Identifying and defining use cases that can feedback into the work of GA4GH standard development teams
- Identifying areas for collaboration (e.g. between two organizations in the meetings or between the presenting organization and a Work Stream)
- The suggestion of a GA4GH standard and/or implementation

Resources

Please see below for a list of relevant resources that you may find useful when creating your presentation:

- GA4GH toolkit
- Past presentations
- Slide template (optional)
- Use case template
- Follow Up Form (for completion after the presentation)

Presentation Schedule

If you are interested in presenting, please add your name to the sign up sheet. For more details on the specific day and time of the meeting, check the Cancer Community calendar invitations or email neerjah.skantharajah@ga4gh.org.

DD/MM/YYYY	**Project/Use Case**	**Presenter**
09/12/2024 21:00 UTC	WAYFIND-R Program Roche	Erika Schirghuber
11/11/2024 13:00 UTC	Zero Childhood Cancer	Marie Wong-Erasmus/team
14/10/2024 21:00 UTC	Canceled	
09/16/2024 13:00 UTC	Melbourne Connect session	
12/08/2024 21:00 UTC	CSIRO: LLM assisted querying for health APIs (Beacon)	Yatish Jain
08/07/2024 13:00 UTC	GA4GH Variation representation specification (VRS)	Wesley Goar
10/06/2024	Canceled	
05/2024	New Leadership Orientation	Cancer Leadership
04/2024	GA4GH Connect	Jordi Rambla
03/2024	Cancer Beacon	Jordi Rambla, Lauren Fromont
02/2024	CancerModels.org	Zinaida Perova
01/2024	CRDC + Velsera implementation of DRS	Surya Saha
12/2023	ONCOLINER	Rodrigo Martin
05/2023	Structural Variations Special Interest Group	Alex Wagner
04/2023	UNCAN	Eric Solary
02/2023	ICGC ARGO & Lancet Oncology Journal: Commission in Cancer Genomics and Precision Oncology	Rafaella Casolino, Amber Johns
01/2023	EOSC4Cancer Project (Squad #3)	Xenia Villalobos

11/2022	EOSC4 Cancer Project	Salvador Capella-Gutierrez
04/2022	CHARM	Benjamin Wilfond
03/2022	b1MG	Giovanni Tonon
02/2022	Pan Cancer Analysis of Whole Genomes (PCAWG)	Lincoln Stein
01/2022	EuCANCan	Jordi Rambla
12/13/2021	Genomics England	Alona Sosinsky
11/2021	Following up on BRCA Exchange Use Case	Melissa Cline
10/2021	VICC, ClinGen	Alex Wagner
6/2021	1- Cancer variant representation – gene fusions and categorical complexity 2 - Mining germline variant co-occurrences in order to move the needle on interpretation 3 - Enabling Passports to work with ERACommons ID and dbGaP	1 - Alex Wagner, Sharon Plan 2 - Melissa Cline 3 - Anne Deslattes Mays
1/2021	1 - Mining germline variant Co-occurrences 2 - Pediatric intracranial germ cell tumor 3 - Kids First Data Resource Center	Melissa Cline, Anne Deslattes Mays, Allison Heath

Past Presentations

GA4GH Plenary 2024 - Melbourne Cancer Data Commons: Insights gained and future directions		
Presenter: Bernie pope, Tanja Davidsen, Jordi Rambla, Dylan Spalding, Michael Lukowski, Marie Wong-Erasmus, Bob Grossman	**Slides**	**Recording**

Key takeaways

GA4GH Connect September 2024 - Cancer Community Agenda.docx

- Data commons promote open science, high-quality data curation, and streamlined access through cloud-based platforms, accelerating cancer research.
- The CRDC is developing a centralized submission portal & data discovery dashboard to improve data access and integration. Zero Childhood Cancer also expressed future plans to move towards a single multi-purpose user portal.
- Platforms like BRH powered by Gen3 support cross-repository data exploration through FAIR APIs, enabling interaction between various cancer data commons.
- Future data commons will incorporate AI models, aiming to enhance cancer research through tools like Retrieval Augmented Generation (RAG)

CSIRO: LLM assisted querying for health APIs (Beacon)

Presenter: Yatish Jain	Slides: N/A Youtube demo full Youtube demo short	**Recording**

- *AskBeacon* is an implementation built on Serverless Beacon (sBeacon) which uses live language models to make querying across the global Beacon network more accessible as it's able to process natural language
- The 15 minute AskBeacon demo can be accessed here. A 2 minute pitch version can be accessed here.
- Depending on the level of access a user has, they will have access to different levels of granularity (one of the query parameters in the Beacon schema). Not everyone can have record level access which is a user management safety precaution
    - Admin access is dependent on how your institution uses Beacon and who is designated part of the data access committee
- Q: how do end questions get translated into Beacon? The results will always be an "or"
    - LLMs are able to understand the Beacon schema so when an end question is posed, the LLM will suggest a 2 step approach
- Following the demo, Yatish went more in depth with a live example of a query, which can be viewed in the meeting recording starting at 23:45

Opportunities to collaborate

- The CSIRO team is open to collaboration and encourages feedback from the community on 1) the AskBeacon tool, 2) Severless Beacon, 3) what types of data are hosted in the sBeacon

- The AskBeacon system is flexible, being able to query genomic and metadata together. If the Community has specific questions or uses cases, they can reach out to Yatish and the CSIRO team to help customize the Beacon interface for their needs

GA4GH Variation representation specification (VRS)

Presenter: Wesley Goar	**Slides**	**Recording**

How the Genomics Knowledge Standards workstream (GKS) can help the Cancer Community

- The health data evaluation pipeline experiences multiple bottlenecks at different steps from data generation, preliminary reporting, annotation, human-interpreted resources, knowledgebase integration, and final reporting in clinical reports/studies
- GKS aims to alleviate this interpretation bottleneck & transmit info between computer systems with accuracy across a federated variant evidence ecosystem
- The use case categories GKS addresses include: Knowledge/evidence APIs, variant API/Repos, Knowledge classification, testing results interpretation, test result annotation
- Optimizing interoperability of clinical workflows is a priority of GKS
- For more information about initiatives that employ GKS standards, please refer to the slidedeck

VRS projects

- **ClinVar-GKS Project**: Objective is to enhance the utility and accessibility of ClinVar datasets through developing a data transformation pipeline to incorporate GKS standards that would make it more interoperable.
- **gnomAD-GKS project**: GKS is working to incorporate VRS (a way to create computable identifiers for specific variation) into gnomAD to avoid issues of ambiguity and encourage precision, while providing evidence lines. The entirety of gnomAD has been VRS-ified, so variants can be searched using specific computable identifiers
- **MaveDB - Multiplexed Assays of Variant Effect (MAVEs)** - refers to a high throughput scale assay that enables the creation of a vast library of genetic variants that have undergone an experimental condition, which produces a variant effect map.
- VRS enables the contextualizing of MaveDB libraries, by mapping the data in the platform to the human reference genome, attaching VRS IDs, which allows for the transfer of that data into different resources that also recognize VRS IDs. Mave also developed a minimum information standard so that data models can speak to each other
- For more examples of platforms that are employing VRS IDs, refer to the slidedeck

Obstacles to knowledge matching and curation

- Categorical Variants (CatVars) link assayed variants to genomic knowledge
- (CatVars) descriptors provide a means to differentiate variations that result in the same outcome under broader categories
- CatVars represent classes of assayed variation and are related to each other in hierarchical patterns
- **VICC MetaKB V1**: purpose is to harmonize information from different knowledge bases (OncoKB, PMKB, CIViC, etc) and put them under a common data model
- **VICC MetaKB V2:** some partners privatized their work and didn't participate in V2, it uses publicly accessible knowledge bases

Variation Categorizer (VarCat) -in progress

- Purpose of this project is variant interpretation and classification to enable efficient, reusable, and interoperable classification information, incorporating public evidence/data, data curation platforms, data normalization, variant interpretation/visualization, standardization, and industry/government/academic resources

GA4GH Connect 2024 - Ascona
Implementation of Beacon in cancer use cases

Presenter: Jordi Rambla	Slides	Recording

Meeting minutes:

- Minimal Dataset for Cancer from GDI can benefit from feedback from community to be expanded upon
- Abstract concepts common in cancer queries (examples below) need to be better defined

Healthcare cancer use cases - ELSI

- **EXAMPLE QUERY 1**: Uncommon motivation of EGFR gene in NSCLC patient → look for a cohort (variants, treatments, outcomes, survival)
    - The concept of an "Uncommon mutation" needs clarification, considering factors like the population, allele frequency thresholds, and if its somatic vs germline
    - How do we specifically define an **outcome** and **survival**?
- The GDI Minimal Dataset for Cancer does not address all of these concepts

- The community should determine the definition of these terms (e.g. length of disease is not the length of a hospital stay, it's when you have stopped receiving treatment)
- These types of clear definition should be integrated into models to improve the accuracy of queries
- **EXAMPLE QUERY 2**: Acquired resistance mutation occurring in Ros1-rearranged NSCLC metastasis during crizotinib therapy ➜ look for similar cases
    - What is "acquired resistance", "rearrangement", "during", "similar"?
    - Not everyone will have the same criteria, and this also may differ according to the disease type
    - MS: other groups such as ClinGen may have definitions we can use in this context

Discussion

- ZP: There's a lot of guessing because every single word brings uncertainty, however it seems that there is a certain weight to each word of the query. It touches on the comment by Tony in chat about doing it through an engineering perspective. If you have the option to ask this question to who wrote the query the person could redefine what they're looking for. Is there any way of limiting the choices? Aligned with Tony's comment on limiting scope.
    - JR: [...] once clinicians specify their criteria we can determine an answer, however when you open this up to a community that shares a common dataset we need a universal criteria, presenting the potential options to users might be too much and get unmanageable. At least we can say when you're using GA4GH standards, we can define the criteria
- MB: There are communities that do this. There are developments of abstract clinical ideas ontologies and data models.
    - JR: Yes, ontologies are a potential solution. Not saying that GA4GH should provide ontologies but the Communities of interest could work towards this (Cancer, infectious disease etc.)
- MB: Is there something flawed in our model because it doesn't serve the recurring need of a DP?
    - The only model that tries to tackle this is OMOP. defines events e.g. lines of treatment
- In this space we've started work to identify how lage data models can ease the process. Where the issue comes is how end users ask the query to beacon. This is where LLM has put in some work.
- TB: My initial comment was pointing out that Jordi is making a clear explanation that we can't achieve interoperability if you can't define what you're interoperating on. Definition is impossible if you're trying to create a complete dictionary. You need to add scope, and accept some fuzziness. Can't do it all. Data models can help here.

    - JR: the option of showing the barriers that work. When something depends on several components (e.g. survival)
- SC: In research mainly we speak in english. So it's important to consider different languages as well. [...] Scope to define similar cases. The scope comes from the community - so whatever is relevant for them.
    - JR: Assuming we can map everything to english, or map most of the things to ontologies that are easily translatable
- DB: The community's use to work in the situation where the client knows what he or she wants. In this case it's not the case that they don't know, they are not here. I'm with Tony that we should first accept that it's not going to be comprehensive at the beginning. We have to show them the possibilities first with not perfect. When you go to the genomes, if you have these questions 10 years ago, it would've been different. But the new genomes cannot answer these questions. Would try to find a balance between being comprehensive and realistic. Matching what the community wants, but also having the freedom to do what you want and check against something.
- **EXAMPLE QUERY 3**: Find optimal Tumor Mutational Burden threshold to tier patients with same tumors (harmoniza TMBs, compare biomarkers, treatments, survival)
    - These are really high level questions where you would deploy an analyst to see what they would do. You wouldn't use an API, you would go to the data. There are different pipelines and metadata involved in this. Noise rate in pipeline.
    - JR: The user understands those questions but not the implications of the answer. We need to know what to model
    - You are trying to automate the cohort creation question. This is an undetermined optimization problem. You have a natural language query that has some constraints that could be mapped onto some standard model. That's a challenge. Another challenge is once you've represented it, it's still an undetermined problem so you still need to figure out how to identify what best fit is.
    - It's a two step problem if you have an intermediate problem

Research Cancer Use Cases

- Molecular characterization of rare cancer types (obtain WES/WGS, reanalysis)
- Noncoding and regulatory regions: beyond exons to assess treatment outcome (find WGS, reanalysis)
- Characterizing cancer genomes for patterns of somatic retrrotransportation
    - These are generic and requires further description
- MB: Optimizing research through query.
- There is a distinction between healthcare and the research community. Within the research community there are bioinformaticians and doctors. So the range of scenarios is quite broad

- LLM can also be sued to prepare input data that the query is based on
- This isn't precise and there will be errors, but that's okay - it will just be something that possibly match
- Discovery queries are not the same as analysis select queries
- Discovery enriches for things that possible match our query
- Augusto: comment from DP AI/LM workshop - in a world where we have large language models, does that give us in GA4GH a different perspective of where we define the level of APIs. the api should just pass a pre-text query that the implementer resolves with one of these. Alternatively, the user/client can use free text and it understands how the pai should be called. How do we advance that conversation

Cancer Beacon

Presenter: Jordi Rambla	Slides: N/A	Recording

Beacon for Cancer

- Beacon is a simple server that allows for discovering data available using genomic variation information + annotations, as well as Phenoclinic information (mostly Phenopackets)
- Beacon V2 has a model that includes individuals (diseases, treatments, pedigrees etc), Biosamples (origin, tumor vs normal etc), Genomic variations (position, type, molecular attributes etc), alongside different levels of access to data
- Oncologists from EOSC4Cancer & TCGI clinics shared what features they look for in a Beacon specifically geared towards cancer. Two categories were mentioned: 1) healthcare queries (eg - how has a variant been classified, case-centric queries which are more sensitive) and 2) research focused queries
- A common issue that arises is mapping queries using natural language (which are difficult to define in a standardized manner) to Beacon
- There are 3 areas to align in terms of tailoring Beacon for cancer implementations: questions, data, tools > following this, group discussion started

Discussion/Q&A

ZP: what needs to happen to make the presence of a patient derived cancer model (PDCM) accessible through Beacon?

JR: To tailor Beacon for cancer, we have been replacing the base Beacon model with a cancer registry model (draft stages) and extending this model with images for cancer. The question is whether to add an extension data model to Beacon V2 or replacing most of the Beacon model with the schema used for PDCMs

SP: Realistically how do hospital systems with highly complex and firewalled systems participate in Beacons?

> **LF:** We're working right now with several hospitals in Catalunya on deploying beacons in an inter-hospital network, and unraveling new challenges for sure! But it is possible. Right now, we're talking to their IT departments in order to see how they can deploy a beacon within their system — once we've had enough use cases, we'll publish some guidelines
>
> **JR:** Beacon has different use cases. It can be installed inside the firewall, in which case there is no issue. Another use case is having a Beacon outside the firewall with filtered information. Another use case is having a Beacon both inside and outside the firewall, which can communicate (the external Beacon will receive answers from the internal inquiry).

JL: has Beacon implemented the approved GA4GH VRS model, the Variation Representation Specification?

> **LF:** some VRS examples
> **JR:** yes, in the Beacon response you have VRS, but annotation is not done yet.

SP: An important consideration left out of conversation has been pediatric cancer patients apart in clinical trials. Looking at how Beacon could be incorporated this type of work could be more effective than just working through hospitals (clinicians may not always be involved in the terms/structure of trials)

ZP: what is the limiting factor/what determines which data is available through Beacon?

> **JR:** what data is available depends on the institute/research team/individuals that are deploying a Beacon (what they have access to). There is no limit but the current Beacon model is focused on clinical genomic diagnosis. The Beacon model also makes suggestions of ontologies to use

CancerModels.org

Presenter: Zinaida Perova	Slides: N/A	**Recording**

CancerModels.Org: An open global cancer research platform for patient derived cancer models (PDCM)

- Discrepancies b/w preclinical data + clinical outcomes, low approval of cancer drugs, drug resistance etc are some bottlenecks to combating increasing rates of cancer

- Genomics-based precision oncology aren't always clinically actionable and have other limitations like not representing all aspects of tumor biology, among others
- Functional precision oncology in comparison uses PDCMs, a field of research that is rapidly increasing
- PDCM fall into 3 categories: 1) cells + 2D models, 2) Patient derived organoids + other 3D models, and 3) patient derived xenografts
- CancerModels.org tries to incorporate the FAIR principles (Finable, Accessible Interoperable, Reusable) by using a Minimal information standard for PDCM, alongside query filters, providing harmonized metadata with a model characterization score (based on amount of information available), integrated sources (like cancer annotation resources), working to expand partnerships to enrich metadata, populations and annotations
- CancerModels.org is a database of different types of cancer models that exist that spans bodily systems, ethnicities and available genomic data
- Zina shared pre recorded demos that demonstrate how to use the resource (view meeting recording)

post presentation discussion

- There is integration of CancerModels.org with NCI Patient-Derived Models Repository (PDMR)
- Funding of the platform comes from NCI, looking to diversify
- Annotations found in the tool come from various sources, they use processed data and link out ot the raw data for those interested
- Submitter indicates where the raw data came from, do not pull all models from providers (e.g. dbGaP)
- Potential GA4GH connections - BAM Files already have GA4GH DRS IDs
- Can deposit metadata from submission process, if it's available from other sources already they don't need providers to attain the data and can do it independently if possible

CRDC + Velsera implementation of DRS

Presenter: Surya Saha	Slides: N/A	Recording

Surya's Presentation: Multi-omics research on Cancer Genomics Cloud (CGC) driven by GA4GH DRS + National Cancer Institute (NCI) Cancer Data Aggregator (CDA)

- The NCI Cancer Research Data Commons (CRDC) comprises different data commons, including the Genomic Data Commons, Proteomic Data Commons, Imaging Data Commons plus others.

- The primary goal of the CDA team (NCI, ISB/GDIT, Velsera, Broad) is bringing all the data in CRDC into a single search interface
- Finding existing data isn't straightforward and being able to do a data-first search across various repos/file types etc would be more efficient (this is where GA4GH DRS comes in)
- The CDA now covers several data commons including: imaging, genomic data, proteomic data and cancer data services via a swagger API but a python interface also exists that is less technical (but still needs computational savvy users)
- CDA has open access metadata and users can search, but to access the files, a user requires credentials
- CDA meets GA4GH: once you access that primary dataframe, for each file in the collection, there is a GA4GH DRS URI which can be brought to cloud platforms or downloaded
- The CGC is powered by Velsera and part of the CRDC, it can be used to analyze NCI data
- Typical user flow: create a project, select data sets of interest, selects tools (like R), run analysis
- Within CGC, GA4GH DRS + NIH RAS Passports (one step removed from GA4GH) are being used + HL7 FHIR for clinical metadata
- Surya provides a use case using data from the genomics + proteomics data commons > selected breast cancer as the tumor of interest > specified file types (VCF + RAW) > used MaxQuant to correlate the 2 data types > cleaned data + multiple test corrections
- Point of the use case was a proof of concept to show you can use CDA to look at subjects across multiple data nodes and bring them together in a meaningful way

ONCOLINER

Presenter: Rodrigo Martin	Slides: N/A	**Recording**

Oncoliner Overview

- FAIR data principles guide the way we use genomic data in oncology > Find, Access, Interoperate, Reuse > the presentation focused on the last two principles in the context of somatic variant discovery
- Problem: different tools and analysis methods work differently, making interoperability challenging and subsequently making it hard to reuse data
- A study they conducted among 3 oncology centers found a significant lack of homogeneity along 2 metrics: performance heterogeneity score (measuring how much the performance (recall + precision) of centers differ, & gene discordance ratio (measuring how many genes are not detected by every center vs the total number of

genes detected by any center) >> this means the 3 centers would be unable to work on a unified cohort of samples
- Rodrigo explained how the issue of lacking interoperability could be resolved with the Oncoliner tool > a comprehensive platform with 3 models for assessment, improvement & harmonization of variant calling.
- Users of Oncoliner can interact with a GUI (graphical user interface) to access the output of any of the 3 models to address their research needs
- The 3 models work like a chain > assessment feeds into improvement which feeds into harmonization, resulting in a final HTML output report
- The result of the study showed a decrease in discordant variants/genes/driver genes, meaning the 3 oncology centers were calling more similar variants > improving harmonization + average performance
- Oncoliner also offers addition softwares like VariantExtractor (which is readily available python package)
- Question raised in the chat: does the VariantExtractor map to the VRS/ClinGen work on canonical naming of events? Rodrigo said yes and that they're working on moving towards VRS
- David Torrents mentioned that this is one of the first efforts to harmonize analysis across oncology centers > specifically in post processing of variant notation
- Question: to what extent can Oncoliner be applied to non whole genome datasets? you need a VCF you refer to as "truth" (this includes validated variance) and one referred to as "test" (this will generate your pipeline) so technically it can take any input in this format but Oncoliner wasn't tested on this type of data, could introduce some problems
- Oncoliner uses user defined benchmarks which makes it easier to play with for future uses

Structural Variation Special Interest Group		
Presenter: Alex Wagner	Slides	Recording

UNCAN		
Presenter: Eric Solary	Slides	Recording

EOSC4Cancer (Squad 3)		
Presenter: Romina Royo, Xenia Villalobos	Slides	Recording
The EOSC4Cancer project is a European-wide foundation to accelerate Data-Driven Cancer research. The main project objective is to prepare EOSC service with data tools and services for the cancer community. In the process, we would like to ensure that we are adhering to standards and recommendations set out by GA4GH. EOSC4Cancer has five use cases covering the patient's trajectory from cancer prevention, diagnosis, treatment, and medical management. The use case we feel is most relevant to the cancer community focuses on connecting omics data from multiple sources to a clinical decision support system for precision treatment of metastatic CRC. The biomedical questions they are trying to answer include: ● How can we optimize the use of molecular data in the context of precision cancer medicine? ● How do we better match profiles and investigate drug biomarkers? ● How do we efficiently allocate patients to clinical trials? ● How do we manage incidental findings?		

Cancer variant representation – gene fusions and categorical complexity (VICC, ClinGen)		
Presenter: Alex Wagner and Sharon Plon	Slides	Recording
Description: ● VICC was established to build expert curated interpretations and integrate those into reports, search consistently across knowledgebase (KBs) and use in aggregate, standardize cancer variant knowledge ● CIViC used by ClinGen to curate info about gene fusions and other variants causative / applicable to clinical interpretation in cancer ● Contents of KBs can vary quite a bit, eg. fusion notation and specificity. Variability of data depending on diagnostic assay used in clinical setting, ambiguity takes up lots of time when trying to interpret gene fusions ● VRS - semantically well-defined computational model could enable automation of the process. Quickly ID when observed variation actually matches literature curated in VRS supporting KBs ● **Goals:** ○ Define and disambiguate gene fusions		

- Recommendations to collect/curate salient elements
- Recommendations for fusion representation

- Problem scope: what defines an oncogenic gene fusion in the first place?
- Open community project, pulling together into an SOP, strong recommendations ready for initial draft, looking for feedback interest from the community for example ClinGen Cancer Variant Interpretation WG is interested in this issue.

Mining germline variant allele frequencies and co-occurrences in order to move the needle on interpretation

Presenter: Melissa Cline	Slides	Recording

Description:

- Largest population of variants (~40%) in ClinVar are of uncertain significance, existing data in different individual databases could address this problem of interpreting them but accessing data is problem
- Best traction not from co-occurrences but from allele frequencies, looked for co-occurrences of VUS with pathogenic variants, measured allele frequency as analytical control
- Allele frequency info is a great way to preserve patient privacy because it's aggregated - Federated analysis offers the potential to accelerate variant interpretation by analyzing sensitive patient-level records to generate informative variant-level summary evidence which is less sensitive and can be shared more broadly.
- Successful proof of concept analysis with BRCA Exchange and Biobank Japan (BBJ) data, shared with them containers for co-occurrence analysis of the BBJ cohort. Now are using data to interpret some VUS, classic case of data we wouldn't be able to access directly
- **Vision:** DPs or other interested initiatives (eg. NIs) who can query through Data Connect or Beacon, sending workflows via WES. We have the technology (analysis containers are ready, meets WDL standard), we just need data holders. Take the idea of containers to the next level and share results on an aggregated level. Could integrate containers with add'l GA4GH workflow standards (WES), GA4GH phenotype standards

Genomics England

Presenter: Alona Sosinsky	Slides	Recording

100,000 Genomes Project

AS: We were able to put this genomic data and match it to clinical data which is stored centrally in public health england databases in NHS digital and also in the national census* We have data on tumor staging, typography, histological subtype etc.
SP: Do they have tumor or normal genomes?
AS: for 17000 patients we have tumor and non-mal.
We partenred with academic research and so far it's 84 institutions which are registered with us and half a thousand of the research projects. We also partner with pharmaceuticals. We witnessed about half a thousand covid genomes.
ADM: For this group it would be useful to point out that one of those is the lifebit cloud OS platform. That is next flow work flows. What is the second research environment?
AS: The standard HBS.
ADM: On premise.
AS: Most of the data is based on HBS. Covid data and our pilot on the oxford nano board goes into lifebit.
So when data is being used in life bit. We should distinguish the compute and * platform. Are yo prvodign any of the compute resources?
AS: Depends on the project. Majority of the 100K data sits on our local storage.
ADM: So researchers including pharmaceutical companies do you have them go directly into on premise.
AS: Majority sits on premise. Only covid data and data from long read sequencing sits on Lifebit.
IF: this may be going back a bit, but we heard that all of the data was on a former military facility and the only way to compute it is to go there. Could you put this into contexT?
AS: Majority of the 100L data sits in HBS environment. We are developing new research environments on AWS using life bit, but not all data is there. If you want to access paediatric cancers you'd have to go into our legacy research environment. IT's on * Cloud. We are migrating to AWS so in the future everything will be moved to AWS but can't predict how long that would take. We finished the 100K Genome project. As an outcome, NHS established genomic medicine services. Clinically accredited WGS service for cancer and rare diseases. NHS commissioned pediatric, hematological, cancers and sarcomas.
SP: Genome data is done as a primary test in a few centers, but it's not the going test. In the UK, are these patients who would've had their tumor analysed in a RNA/DNA panel, or is this the only genetic testing that would have been done?
AS: For lung cancers we are still at the panel stage. For paediatric cancers, every patient has access to WGS. We've had quite a few cases where we've seen variant fusions in WGS which were not spotted by the RNA panels. Same for hematological cancers. Depends on the hospital and genomic labs. I would say that since pediatric has become eligible for WGS i can see a lot of enthusiasm among pediatric clinicians. So far 92% of patients sign consent to become a part of the national genomics research library.

New at Genomics England

AS: Cloud native trusted research environment on AWS using Lifebit's CloudOS platform is still in the process of adoption. We started a large pilot on screening newborns using WGS. Also started a diversity program so we can sequence genomes from different ethnic groups in the UK. Our genomic data is supplemented by clinical data. In the future we hope that this will show up as live data in our environment.
IF: Any GA4GH standards tha you've looked at which might be relevant to this new AWS environment? E.g. cloud standards to help enable compute on AWS. it would certainly help. The GA4GH standards themselves can help with this gap of knowledge. Was able to make use of the DRS standard and enable use of it from the cloudOS platform. WES and DRS could help with things like containerization and access to data on cloud storage.
ADM: And there are other approaches. Like instead of using WES*
AS: I think next flow is what is in store.
ADM: Are people using next flow also on prem?
AS: Yes, I think people still have a preference for on prem*but I think in Lifebit we are probably forced to use nextflow because it's the only thing that is enabled.
IF: DRS you're not suppose to get the data out. You would compute with it on place. Where there are use cases like this it helps drive what that solution needs to do.
IF: For data on that research platform,w hat would you need to do to be able to authorized to access it?
The institution needs to sign a * agreement with genomics england, submit proposal to GC* domain and once that's approved you can start submitting your workflow and run analysis.
IF: that's a fairly typical pattern.
IF: What you've described is similar to what you'd see in EGA and dbGap. DACReS which standardizes approaches across data access committees.
SP: Maybe we should consider changing this name.
ADM: One thing genomics england has is the consent. It's very transparent. Don't know of anyone with such far reaching consent.
SP: Nothing has this kind of scale.
AS: Cannot re-consent retrospective patients so have to deal with the consent they signed for research studies. It's usually difficult and limited. Given that this consensus is given so much flexibility more than 90% of our patients are happy to sign.
ADM: Don't think we think this way because in the US we don't have universal healthcare. Not sure Sharon if you're seeing a change in that. In consent going forward is there an effort to nationalize that?
AS: The AGHA has developed a proecm called control which is a dynamic consent program where patients can go back and review it, and also see if it's being used for commercial purpose or not. How that might go in the future in terms of national agreement is still ongoing.
SP: the closest is the children's oncology project Every Child. 90% of children are treated at the child oncoloyg centre. They are attempting to enroll every child at diagnosis, there's an epidemiology questionnaire.. In the pediatric cancer realm there is an attempt to do that. In the adult realm there's more pieces.

AS: On the question is consent, specifically pediatric consent, childrens grow and once they past 18 they have to be reconsented. We are facing that issue now with patients from the 100K genome project.
IF: The other new standard announced at the GA4GH SC was a pediatric consent standard. There's some mileage there for exchange of information.
SP: I don't know how much it applied in a cancer setting. Typically in a cancer setting you have alot of language about additional tumors, relapses etc. so I didn't know how much more of how...
AS: Standards I had in mind are a bit different. Standards for DNA Sample Collection, sequence, mapping/variant calling, and interpretation. This bit about collecting gremlin samples, I've gone through so many things over the few years. We put together sample collection guides and thought it was thorough but a year later we reviewed this germline that we collected* Now we try to develop pathways for collecting skin biopsies.
IF: Biospecimen Research Database: Database that NCI setup and continues to maintain specifically on this question. Nobody publishes a research paper specifically on methodology. This was intended to do exactly the thing you are describing. It looks at different analytical platforms of how handling and collection methods effect the outcomes.
AS: At the beginning of the program we tried to convince the pathologists to do WGS we need to collect fresh frozen samples. It was not easily transferable to the new platform. What is the DNA of sufficient quality? What is the genome of sufficient quality? Started using Illumina platform and faced some resistance from research community. What are we using as reference. What validation can we use across platforms. Finally, what are the standards for interpretation. We have seven genomic hubs and they all run their in house classifications. We did work 5 years ago on an overview of somatic changes in FF and FFPE samples. Quality of genomes from samples in FF vs FFPE. seen lots of false positives in FFPE.
LS: Re: quality metrics - for anyone interested, a few national initiatives (led by Singapore and AGHA) have been working QC metric definitions - https://docs.google.com/document/d/17bVufpacyoUM4UDKlwkr-0KOG-yZrcVSvM6yC9gbrk4/
AS: This is our current pipeline for Genomic medicine services. We adopted Dragen alignment. Until three weeks ago you had to have Dragen hardware to run the algorithms but at the end of November GSDK announced that you could run it* We are doing benchmarking and validation. What we do for clinical accreditation is a bit different. So there are a lot of comparisons with DNA RNA panels. Research users want to see comparisons with the ICGC pipelines. Different users are looking for different validation. Also looking to adopt this graph maker from Dragen. Saw big improvement in accuracy especially decreasing rate of false negatives.
MC: This is very interesting. What are you anticipating with the future human pangenome reference?
AS: My thoughts now are with this telomere to telomere reference that were generated from the long- read sequences. Would like to map the long reads to the telomere to telomere reference. But it still doesn't have much annotation so can't use it in clinical space.

IF: What's the right way for us as a group to continue to bring these kinds of questions to this forum within GA4GH. Which of these need to go off into one of the GA4HG Work Streams. Perhaps we can discuss this more in future calls. One of the questions would be whether we serve some of these things up in the group or if they're better picked up in a work stream.
LS: Part of how I see this group is as a clearing house. When we see things of interest to WS we feed them in. For example, Melissa' container work being moved to FASP. Part of this group should be directing the efforts. If there isn't a place yet then we can do work here or finding the appropriate spot. We need to do some exploring about how to organize this.
DT: Do you have a catalogue to search for data?
You can use the open CGA catalogue. We also provide a lot of tables which can be interconnected. Clinical data exists in the form of tables.
DT: Do you have in your plan ways of connecting to the outside (Europe, US, Canada) in the future?
AS: In the end, if we're all on the cloud.
IF: Want to access Genomics england data. Does access extend internationally?
AS: Yes. There are some clauses about IP in our agreement that not all American Universities are happy to sign so it also depends on Universities.
IF: What we can provide is the motivation and use case to specific working groups. If DURI is the right group to hand you off to, we want to stick with it. Taking your motivation Melissa partnered with Alona we can carry that over to DURI. To set the target for that group of solving the problem.
MC: I'M GAME

EUCANCan

Presenter: Jordi Rambla	Slides: N/A	**Recording**

EUCANCan
JR: Focus on the usage of GA4GH standards in EUCANcan. Create a network of federated data, both technically and legally. The idea is to demonstrate or show that by analyzing data in a uniform way and sharing it with legal aspects addressed, you can use it for solutions. Benefit end patients.
4 work packages. Clinical and genomic data collected and shared via nodes. Data should be shared and maintained in the long term and be able to be reanalyzed. Can generate a network of data to share in a uniform way and keeping it available for long term usage. Focused in making this happen.GA4GH standards in WP5 (standards for clinical data, phenopackets) and WP3 (infrastructure, cloud standards)
SP: Are these research genomics, or do they include clinical panel type projects?
DT: In principle we are considering whole genomes, but the concepts could be applied to exomes and gene panels.

SF: For WP2, sounds like an area that variant annotation work might be applicable?
DT: Idea was not to reach interpretation, but more identification. Project includes a report for the clinic, could find space for interpretation, not a priority.
SF: VA is having discussion about unambiguous representation of variants. Very detailed discussions, but don't often hear European voices on there. [Action: connect European projects with VA/VR work - variant standards are a shared topic of interest]

JR: WP3 - to define a circuit for long term raw data and metadata storage and flow, as well as access protocols. Metadata, phenotypic and raw or genomic data. With the metadata, we have no standards in GA4GH that applies to what we need, using INSDC model. Using Data Use Ontology (DUO) and OIDC (open id connect) for authentication of users. Passports doesn't make sense for us yet, but base for passports. Does not include phenoclinic data. Raw data is Crypt4GH format.
Looking at phenopackets for phenotypic data. Piloting with submission, encrypt and validate with a tool provided by Peter Robinson, made accessible for data findability.
Using passports for data access.
What happens when data couldn't be send to EGA and needs to stay locally? One use using Federated EGA, different countries aggregate data locally. Hospital doesn't have the capability to store data locally, so they send to their national centers. In other cases, we expect data goes out from the institution itself, "EGA Community Platform". Fostering a modular approach: manage the data, mechanism for allowing access, and then discovery and how to expose data. In some cases the institutions have something already, maybe a solution for browsing or storing data. Only thing we ask is to reuse what they have. If they have something but it's not easy to integrate, then we ask them to extend their solution, like putting a Beacon on top of the solution.
Let's assume you use cBioPortal, can analyze and visualize but cannot integrate. In this case we would say Beacon on top. You also don't have management of the files, so need GA4GH standards applied.

You have a patient visiting clinician and that clinician is sending information to the lab. Lab comes back with a report and maybe a vcf. Sometimes the repository is internal or external. Data is in this cycle. Then, researchers dig into the EHR and generate one database with the data they need for a study.
Problem of moving from distributed process to something much more organized. Have to get the data in using security, need to store is somewhere, need to analyze the data, and need to index that to make is discoverable. Can pick the pieces you need.

WP4 - cloud infrastructure
EUCANcan contributed requirements to Beacon v2. Strongly based on the scenarios shared here - eg. being a clinician needing to do analysis and share information to third parties.

We are EUCAN image using cancer images associated with phenotypic and genomic data. Seeing if we can reuse the model as a phenopacket. Should be able to get information from oncology domain and share in a portal. Want people to use the same model as much as possible, using the same ontology, dictionary terms, etc.

DT: Can you share the slides with me?
JR: Yes

SF: You mentioned nextflow - Jeremy has been doing work to making WES compatible with nextflow. Interested in your point about defining controls, sounds like something to feed to cohort representation. Also interested in the idea of the community platform, if someone was interested in finding about more about this, is there somewhere I can point them to?
JR: Not yet, more of a brand. Newbie friendly “how to” that shows how components can fit together. Working on a document about this.

IF: Is the community platform for filling needs shown on the first slide? Could you comment on how this works from a EUCANcan wide perspective, is this platform used in Canada too?
LC: The canadian component of EUCANcan is the Overture platform for data sharing and management, which is also used by ICGC. Happy to see it fit into this larger vision. Canadian researchers generally submit genomic data to EGA rather than dpGap in the US. Given the GDPR restrictions which are inhibiting access to the US, Canada will become more closely associated with European efforts. Discussion about creating a sequencing archive, need to discuss if time is right for Genome Canada to fund an EGA node.
IF: A lot of US data goes to EGA as well. What’s your sense of the the US content in EGA?
JR: 15-20% overall I think, including both UK and US.

SP: How much clinical data do you actually have? Treatment, outcomes?
LS: EUCANcan has been a great success at the technical level, but when it comes to our clinical partners committing to share real patient data, there is a lot of caution on the part of hospital admin in making data sharing happen, there is neither the incentives or the risk reduction needed to allow hospitals and healthcare providers to actually make it happen. We have servers running, but doing demos with synthetic data right now. One of the issues WP5 is addressed but will require more than just policy papers to break through this final barrier.
SP: In the US, the Childhood Cancer Data Initiative (CCDI) has patients already in a trial, funded by the government but designed to get all of the rich data in. Have the advantage that patients consented to be in a cooperative trial. One way to overcome this is to start with patients who were in a treatment trial who we already have genomes for.
IF: Jumping point for something to follow up on in a future call. CCDI is closely allied with the CRDC, we could probably bring in some people from that to talk about it.
LF: In EUCAn image, we have the hope that some issues with sharing reluctance will be resolved. EUCANimage has some use cases to use AI tools in order to predict outcomes

depending on diagnosis and treatment. Some clinical partners are willing to share as much data as they can. The number of variables that are available depend on the use cases, but there is quite extensive information about the different kinds of treatments. I think the clinical partners are still talking about the clinical parameters that are to be shared.

Here are the fields we are considering in EUCANCan that derive from ARGO: https://docs.google.com/spreadsheets/d/1Ph5tBD3tG0h1bOpv5sj0KLalx75_ZA-V/edit?usp=sharing&ouid=112786736772019968234&rtpof=true&sd=true

Pan Cancer Analysis of Whole Genomes (PCAWG)		
Presenter:	Slides	Recording
Description		

b1MG		
Presenter: Giovanni Tonon	Slides	**Recording**

An Overview of B1MG and Cancer WP9
GT: 1+MG Project, goal to collect 1 million genomes by the end of 2022. This is quite ambitious. Quite a few difficulties. Most countries are proceeding with panels and have decided not to include this kind of tools and platforms inside 1+MG. Panels are very heterogeneous so considered to be difficult to include them into comprehensive assessment. EU countries agreed to cooperate in linking genomic data across borders. 22 countries have now signed and 6 are observers. Working group 2 focuses on ethical, legal and social issues (Marco Morelli) difficult because of the heterogeneity across Europe. Working group 3 Standards and quality guidelines, probably the most relevant to this group.

Cancer Use Cases
GB: **Case 1: uncommon mutation in target cancer gene: rare somatic mutation in EGFR gene in NSCLC cancer patient**

- Questions that might be asked include:
    - Has the mutation previously been observed?
    - How many cases have been treated with the SOC...?
- Outputs will typically include the number of patients fulfilling the query, survival rates...
- Would be useful to have a quick list of mutation and approved drugs

Case 2: acquired resistance mutation occurring in tumor sample or liquid biopsy during a targeted therapy

- Patient with a peculiar mutation, not a single nucleotide. This patient has undergone SOC but doesn't respond. Want to see whether a new mutation arises.

Case 3: Consistently compare TMB across studies

- TMB is often calculated in the context of a specific study cohort. Each patient tumor differs in terms of mutation burden*
- Can use aggregated data from cBioportal to put together a cohort of patients.

Cancer cases you have a range of possible choices of what kind of aspect of a cancer you'd like to understand from single point mutation to timeline of events which happen during the lifetime of the patient. This is why cancer is also difficult to include across standards.
GT: One of the challenges that 1+MG is facing is modeled against the rare disease. There are some ocuntries in Europe which do WGS for diagnostic purposes.
SP: There's much more interest in RNA DNA panels and some interest in clinical genomes. Many fewer large scale projects doing genomes. Does the project include transcriptome data?
GT: At the present time, no.

Cancer POC
MM: Interplay between infrastructure, ELSI and need from use cases
POC is focused on use cases from rare disease. Have only two countries which exchange synthetic data. The standards in Rare Disease PoC was actually from GA4GH (Beacon, DUO, SAM/BAM/CRAM etc.). Here is a diagram of what a researcher might want to do. First, search for data e.g. what patients have a particualr genetic trait (done with Beacon). You might also want to query ..Then you have a couple of infrstrctures to store the data and manage identity of people who access the data. Then you have a part where you want to do quick analysis on the data you have carried out. Done through GPAP platform.
3 questions:

Re-analytics of raw data is not always a requirement. Need something that can do analysis on the fly. We broke down the diagram to adapt it for cancer. The genetic query is not doable with Beacon v1, but there is beacon v2 hat is about to come out and it may be able to address questions that come out. The search on clinical data must be changed because it cannot handle the complexity of cancer data. The analysis should be updated to fit cancer.

Minimal Data Set for Cancer
MR: Would like to propose a data model fitted for cancer genomics in Europe. Minimal for easiness of use but comprehensive for both clinicians and researchers. Capturing longitudinal and complex aspects of cancer. Considering data on treatments and their outcomes...
Took into consideration existing data models mCODE, OSIRIS, ICGC-ARGO etc.)
We compared the three of them and found out they share certain aspects.

The process towards a minimal set for cancer started with dialogue with clinicians and researchers. Afterwards we created a definition of a set of tables, sub tables and items. It underwent revisions by the WG9 NMG community. Minimal set consists of tables focused on both observations and longitudinal aspects. Consists of 13 tables where a definition is provided, examples, sources and whether it is mandatory, recommended or optional.

IF: The observation about not needing to access raw data is common. However, the calls are often against different reference genomes and from different calling pipelines. How does that affect whether access to raw data is needed? What might it say about harmonizing variant calling?
That question intended mainly for 1MG - but is obviously applicable more generally.
MM: Just to be clear: the access to raw data, if privacy allows, MUST be always an option.

MM: There are couple weaknesses with reanalysis of the data. Firstly, it is double the time, and it's a bit static in the sense that pipeline analysis often needs the usage of * tools. As a proposal could see requiring a set of data standards and then receiving the raw data and call from submitters, and give the user the opportunity to redo analysis if they want. But the query would be done on the call made by the submitter.
DT: One way of population the 1+MG project, Spain has parallel activities. The countries are starting open calls to generate data. We are in a project where we probably will start at the end of the year sequencing and analyzing genomes. Would be good to have a small table. Want to filter for the clinical, technical sample issues. Not sure whether there is time for this.
MM: Don't have the right answer. I've had experience trying to harmonize pipelines across Italian centres. Tools were the same but results were very different. At the level of raw VCF, as it comes out from the caller, if the caller is ** they start diverging a lot. You can ask people to document how they go about the process. If you want to reanalyze you should do it on the platform where the data is stored.
DT: We should have a very basic recommendation for those genomes that will end up in 1+MG project.
MM: There was a paper that suggested 5 scores on the data. If your scores are nto good, we won't accept your submission, but if they are good we don't ask you how you obtained your data.
DT: Quite an old paper and may need a bit of an update.
SF: If you're ding exomes vs gnomes, many of the drivers in cancers are at a minimal levels because of tumor heterogeneity. Are those tables available?
Caution that these are not validated by !+MG. It's an ongoing effort that is soon to be validated so you can use it with a grain of salt.
SF: Just had cancer disease in table 5. Which ontology are you using?
We stay away from ontology and only specify which fields should be filled. WG2 focuses on ontology.

IF: Going back to use of GA4HG standards. Ad something on their called ** storage. Seemed to be touching upon questions asked by the DRS standard. Oriented around cloud storage. Doesn't mean it can't be relevant to storage outside of the cloud. What is FEGA storage and is standardized access in the way DRS would it useful to this?
MM: The idea is that you have EGA as a database for sensitive data. FEGA = federated EGA. data generated cannot leave the country. The idea is to build a node of EGA in your country so the data created in your country does not leave. Then we federate these nodes in a type of EGA network. This is just for storing. Easier said than done.
IF: I think this is an important area to see how we might look at it differently. Ties in with the use case that I've brought to this group. The use case was about finding pancreatic cancer data set in EGA and dbGAP and wanting to federate analysis across those two. Dbgap is looking at drs storage. EGA data was generated in US. Highlights an area where these standards can be useful to us.
MM: The ELIXIR consortium provided an identitymanagement layer.
IF: Would having them begin to merge be a useful thing (dbgap and EGA)?
SCG: EGA is working on passports and visa as long as dbgap is doing the same.
IF

CHARM

Presenter: Benjamin Wilfond	Slides: N/A	Recording

Ben Wilfond: Co-PI of CHARM study. The CHARM study is investigating ways to increase access to genetic testing for those at risk of hereditary cancer in low income, low literacy and minority population. During the CHARM study we also looked into methods of authorship tracking.

Our study has a series of team leads. They lead different aspects of the study. The team leads would identify a paper and create a concept sheet. The concept sheet would be used to identify scope of paper, identify team members, audience, and timeline, identifying resources. Allows PI to prioritize analysis and balance work across team. Realized that there were issues we were facing. How do we promote equity and inclusion in publications? The ICJME guidelines retrospectively tell us who should be an author during submission, but doesn't tell us what the roles should be from the onset. These perspectives are based on discipline, institution, and career stage. Equity = who should be leadin the paper, inclusion = who should be co-authors. The process that CHARm used to determine authorship is the most important thing. Had a team meeting, breakout sessions to discuss authorship (perspectives, challenges,

thoughts on solutions), larger report out meeting, drafted guidelines based on outcomes. Shared guidelines with team, finalized and implemented it.

The guidelines had three components

1. Identify author roles and responsibilities
    a. Team leads (identified early on) ➔ lead authors ➔ primary writing group ➔ secondary writing group
    b. Being led by lead authors, pwg pretty involved, but secondary group is peripherally involved
    c. CHARM authorship values = equity, inclusion, efficiency
    d. Principles to promote...
        i. Equity: team leads should distribute lead authorship to team members are guiding key aspect of the study. Team leads should encourage junior team members to be first authors with mentorship from a senior author.
        ii. Inclusion: who else other than the lead authors should be included? Who has been involved with related aspects of the study and would be interested in manuscript development?
        iii. Efficiency: be realistic about the capacity of the team.
2. Develop values and approach
    a.
3. Implementation

SP: Is the data freely available? Same people get asked over and over. Do hte leaders reach out to people who haven't self-nominated.
BW: Yes, that's the intention with using smart sheet. Team leads tried to identify who was missing. We do not rely on self-nomination
AS: For Breast Cancer Association there are a lot more authors. Not sure we always reach ICJME guidlines. When people have to tick a box, there is a difference in people's interpretation of how they contributed. Practically speaking, when dealing with a lot fo authors have you come into arrangements for how you can track affiliations, acknowledgements? We've done that within ***** we keep a database in endnote of all those different things, circulate it and expect people to circulate it. More time spent on our papers and not who's going to write it. Top journals are actually ok for you to enter top 10 authors and leave the hundreds others.
IF: Went through this with the GA4GH papers. At some point, we were self-reporting our contribution through dropdowns.
AS: Think our role should be to get journals to follow the same process.
IF: How do we build trust? Talking about who's going to be responsible for a certain area of a paper.
BW: Anybody can meeting the journal guidelines simply by reading the paper in a thoughtful

way. It's more who do you ask to read it and send it out to. Within a consortium, who are authored papers vs. who are non-authored contributors? Can make that decision in a variety of ways. We do use smart sheets to track actual authors themselves. It never occurred to us that we would do our process based on institution because people contributed in different ways. Didn't make sense to break it up based on institution due to the differences in contribution levels regardless of institution. Having an institutional data doesn't seem a fair way...
AS: Missed out a few details. They would be core people who were involved in running the database.

EOSC4 Cancer		
Presenter: Salvador Capella-Gutierrez	Slides	Recording
Description		

1 - Cancer variant representation – gene fusions and categorical complexity **2 - Mining germline variant co-occurrences in order to move the needle on interpretation** **3 - Enabling Passports to work with ERACommons ID and dbGaP**		
Presenters: 1 - Alex Wagner, Sharon Plan 2 - Melissa Cline 3 - Anne Deslattes Mays	1 - Slides 2 - Slides 3 - N/A	**Recording**

VICC: Alex Wagner
Challenges in variant representation, impact on ability to rapidly classify variants, esp gene fusion events
Overview of efforts in ClinGen and VICC and others
Working on protocol for characterizing gene fusions
Background: VICC established to build expert curated interpretations and integrate into reports, search consistently across knowledgebase (KBs) and use in aggregate, standardize cancer variant knowledge
CIViC used by ClinGen to curate info about gene fusions and other variants causative / applicable to clinical interpretation in cancer

Eg. BCR-ABL, ALK Fusions, single genomic coordinates provided on CIViC, but otherwise all plain text
Contents of KBs can vary quite a bit, eg. fusion notation and specificity
Standards to computationally semantically describe don't yet exist, working toward that in VRS
Variability of data depending on assay used in clinical setting
Scientist has to take preliminary datasets, from different assays, annotated in human readable text, based on observations etc., how likely does variant apply to this patient/tumour
Ambiguity takes up lots of time when trying to interpret gene fusions
VRS - well defined semantically define computational model could enable automation of the process
Quickly ID when observed variation actually matches literature curated in VRS supporting KBs
Export into report to variant scientist, who validates it, stores it and reuses in downstream studies
Scale variant interpretation
Cross-consortium initiative, ClinGen, CGC, VICC, CAP/ACMG cytogenetics consortium
Goals:

- Define and disambiguate gene fusions
- Reccs to collect/curate salient elements
- Reccs for fusion representation

Problem scope: what defines a gene fusion in the first place?

- Deregulated or novel transcripted formed by the the interaction of functional element of swo or more genes
    - Those that lead to loss of product are not gene fusions, out of scope

Chimeric Transcripts and gene fusions that drive them, Marilyn Li HGVS
MVLD Structure, Gordana Raca and Angshumoy Roy
Curation, ...
Gene fusion curation workflow captures elements of the model
Open community project, pulling together into an SOP, strong recommendations ready for initial draft, looking for feedback interest from the community
Q&A (in chat)
Ian Fore to Everyone (10:26 AM)
Alex - are the knowledge bases you would like to use this representation of fusions respresented here? Is this group a way to get more Knowledge Bases involved.?
alfonso valencia to Everyone (10:30 AM)
Alex: Milana Morgenstern first in my lab and now in her own lab has been working in fusion genes product of the combination of a direct and reverse copy of a gene. I wonder If you have deal with this type of cases.
Alex Wagner to Everyone (10:35 AM)
Ian: The VICC has several KB partners (CIViC, OncoKB, JaxCKB, PMKB, MOAlmanac, CGI, others...) that we work with to develop these guidelines and tools. We have been discussing specifically with CIViC on how we will capture these recommendations in the curation interface.

The other half of this challenge is how to precisely characterize fusion assay data from clinical & research laboratories and the tools to match to these data. I think this group may have more representation / interest in the latter half of this equation.
Michael Baudis to Everyone (10:36 AM)
We run Progenetix (somatic/cancer CNVs) on a Beacon v2 stack, with e.g. relative variant frequencies per diagnostic code. So var frequencies per phenotype are +1 in Beacon v2.
Alex Wagner to Everyone (10:36 AM)
Alfonso: sounds like a very interesting case! Would be great to get input on challenges in representing or searching this.
alfonso valencia to Everyone (10:38 AM)
Alex: I will follow this examples/paper by email to see what do you think. Beside being biologically interesting, the issues of detecting and representing are also potentially interesting.

Mining germline variant co-occurrences: Melissa Cline
Slides
Largest pop of variants are of uncertain significance, data out there could address this problem to interpret them but accessing data is problem
Successful proof of concept analysis with Riken on Biobank Japan (BBJ) data, shared with them containers for analyzing BBJ cohort, now interpreting some VUS, classic case of data we wouldn't be able to access directly
Data on japanese population also helps address the fact that most research data is on caucasians, need more ethnic diversity
Best traction not from co-occurrences but from allele frequencies, looked for co-occurrences of VUS with pathogenic variants, measured allele freq as analytical control, that was where we interpreted most data
Even compared to what we'd get analysing same variants with data in gnomAD, able to get more information simply by having the 24K controls in the BBJ (vs. 10K east asian controls in gnomad, small handful of Japanese per se)
Allele frequencies is a nice and simple place to start in terms of data collection, also requires very little information
Useful to see phenotype to know if working with affected set, but "one person's disease group is another person's control"
Allele frequency info is also a great way to preserve patient privacy, aggregated, privacy nuances in aggregating data on variant level, come in when you have highly rare variant, unique to a single family or individual; for collecting variant evidence for allele freq's you can filter out that scenario
See a variant at some threshold frequency, filtering out highly rare observations
Would like to analyse highly rare, but starting with ID'ing common observations still gives us traction

How can we do this with a greater number of datasets? Can send container to another health initiative
This is a great use case for a Beacon or Data Connect API
Next step: working more directly with phenotypic data, case counts of variant by phenotype are basis of many diff forms of variant evidence, can address that with HL7 FHIR and Phenopackets to ask two questions:

1. What phenotypes do you see in your cohort for this variant
2. Given this phenotype and variant, how many patient observations do you have

Approaches for federated data query are Data Connect and WES starter kit implementation in conjunction with container at siloed dataset to specify parameters to a container that is located remotely
Someday: Pedigree analysis, co-segregation analysis, one of most powerful forms of variant evidence
Complex computation not viable with beacon query, pedigree subgroup of Clin-Pheno work stream developing standard to address this in future, moving in right direction but not quite there yet
Pedigree data is hands down identifiable, debatable whether one can look at an image of a pedigree, but great information, commercial testing labs are using this line of analysis more and more in conjunction with cascade testing
Q&A
What would be your vision out of the cancer community group? DPs who can query through Data Connect or Beacon, sending workflows via WES, which of these projects represented would implement this?
Should/can we engage the national initiatives (NI) in this?
RIKEN wanted data represented in research cohorts, won't be the only nation that has this wish
Are there national initiatives prepared to participate?
We have technology, we have data holders in the NIs, are they in a position to start being data providers
Take idea of container to next level, 150 different country genomic efforts (beyond just DNA, transcriptome, proteome etc) - how to share workflows that contain containers
Get everything into repositories, make sure tested, and then share those, share results on aggregated level
Commercial vendor as a compiler of workflows, use GA4GH standards to get it done
What is in the container - what do you ask of the dataset? Looks at vcf input, takes genotype data as input and asks two questions - where are there unknown variants that co-occur with known pathogenic variants, and what is frequency of each over the

Enabling Passports to work with ERACommons ID and dbGaP: Anne Deslattes Mays
Want to go faster! Answer specific driving cancer question, access data world wide
At last call lincoln suggested talking about making ERA commons a means to access data

Containers as smallest element, one function in larger workflow, all pieces containerized so can be used
More standards to help see what data is out there
Tens of cases for intracranial germ cell cancers in US, many more in Asia
Lifebit is compiler of workflows, compiled on nextflow, many different workflows just in time fashion to ask question
Share aggregated results on to know whether variant, allele or isoform level does what we think
GEL - biostatisticians, clinicians, together
US has no medically embedded data initiative per se
Ga4gh take ID of you and datasets you're authenticated for, pull direct from github

Q&A
Should that discussion be part of DURI? Need passport discussions to get out of "what ifs" and into real concrete use cases
Some possible now, GEN3 - 4 NIH cloud platforms, CRDC, CGDC - not actually using passport yet though
Though ERA commons is very NIH focused
Is there an opportunity in GEL to find scientific use case, send compute to GEL, federate something with them
NEED to go international, diversity doesn't exist in datasets in single countries
Technical people discussing technical things, but serious cancer questions we could answer now - what would it take to cross borders?
Figure out how it's done now and how we can improve on it
FASP - can I do a federated compute? Yes, but gone where there are GA4GH implementations, but not driven by scientific question, data in these two places I need to ask a real question of
Getting to those global examples is a key goal of this group
Research community has been developing tools, infrastructure, for many years; GA4GH comes from this community; now seeing in Europe at least that whole design and control is from health, far from our environment
Coming from one side developing nice things and standards, need also to connect initiatives that are racing to control / set rules for use of data - coming from health
Okay on genomic part, research controls that, but clinical data is key
Pharma companies want to come in and access the GEL data
N3C
H3Africa
Not a technical problem needed to be solved, we have the technical bits
ACTION: (now and forever!) document the use cases - this has been done through slides and elsewhere, how do we establish that as an accessible body of information?
Where are there gaps that GA4GH technical staff can help move this forward?

1 - Mining germline variant Co-occurrences
2 - Pediatric intracranial germ cell tumor
3 - Kids First Data Resource Center

Presenters: 1- Melissa Cline 2 - Anne Deslattes Mays 3 - Allison Heath	Slides: N/A	**Recording**

Mining Germline Variants Co-occurrences

Federated analysis for cancer variant interpretation

Presenter: Melissa Cline

- Algorithmic access to data we might not be able to analyse/share directly
- Proof of concept analysis with BioBank Japan
    - shared docker container for them to apply to protected patient level cohort, sharing variant level data, being analysed by ENIGMA for interpretation of co-occurring variants
    - Looks like it will impact several VUS, will be in ClinVar (and publication) relatively soon
    - Container: on Dockstore and integrated with WDL; per variant lists co-occurrences, freq. Of cases, and freq. Of control
        - not sufficient for re-ID w/o a high false positive rate; info BB Japan was comfortable sharing
        - Sufficient for qualitative assessment of data by ENIGMA

Pediatric intracranial germ cell tumor

Presenter: Anne Deslattes Mays

In the time of the genome project, more standardization than today

Data can't move, 90+ genome projects happening world wide

In 2019, write a layer on top of google cloud platform, needed to know how to spin up machines; but reproducible, access TCGA, spin up docker VM

In 2020, we do this differently, have a platform agnostic nextflow workflow, aggregate data can be stored in Zenodo; from this you can reproduce the entire analysis from that point on, large compute against 9K gtex files needs to be done in cloud

New Gtex rules limits ability to access , not accessioning new seq data in SRA; ability to search for data using bioproject and biosample

Teaching people with no prior command line experience how to do this kind of analysis, teaching to trainees in cancer centers; "dry bench skills for the researcher"

Gone are the days when we can toss solutions over a fence, not everyone can be a system administrator.

We are solving problems now in FAIR and rigorous manners

SRA explorer - put in general terms, eg., cancer, cell line, rnaseq, some are access controlled, some not; I write a nextflow, access data in reproducible manner

We shouldn't inhibit that

SDOs typically create standards, standards drive commerce, SDOs provide testing and certification

What is the test? WES is on terra? Is WES the standard? WES can write WDL, I write Nextflow

Kids First Data Resource Center

Presenter: Allison Heath

Kids First is at intersection of RD and pediatric cancer, more and more is in that RD use case

Data resource started 2-3 years into start of program

Child cancer and structural birth defects

Increasingly we know that across developmental landscape, utility in cross analysing/ understanding these different diseases

Portal includes significant amount of data, 20 studies released in 2020

Some overlap between pediatric cancer and structural birth defects

Over 20K more participants coming through

Intracranial germ cell cancer coming in (ADM's use case)

Down syndrome and pediatric cancer defects, lots of overlap

Way to have intersecting platforms and ecosystem, standards are key

For us the number of people looking t specific diseases is not enough; need people looking at other diseases; common disease, looking for interesting variant, more eyes looking at them

Genomic Workflows

- We do all WF in CWL, already a pain point of goal of having anyone bring their own WFs
- Working hard on the idea of functional equivalence and what that means
- Somatic WGS/WXS - tried to make it modular so if you want a piece in CWL only need to port the one you're interested in; other techniques like that
- RNA-seq, mainly WGS, focus on family based cohorts, cancer tumor and rna seq as well
- NIH Common data ecosystem
- Interoperability across NIH resources, RAS in collaboration with DURI work stream, australia biocommons, elixir
- Lots of data in dbgap and SRA researchers want access to, how do we have right standards for DRS, passports, etc so people can use the tool they want to rather than the only one data is available in

Kids variant workbench

- Alpha stage
- Data continuum challenge, wide spectrum from those biobanking to those addressing clinical outcomes
- Different tools / standards support along that continuum
- Runs on AWS, open source things;
- Focused on annotations, pulling all this together in way that is useful for people was most work, tech, big data, platforms can handle that; but what do people really want to do with this, so when they show up in makes sense
- ID'd our own use cases for variant QC, filter out lower to focus on higher quality

- First question is “do other datasets have phenotypes I’m interested in?” AND “how do I do this in bulk?
- Our goal is really to focus on high level, what do people want to do commonly, how do people bring own datasets, stay aligned even if view differently, so can exchange under the hood

Other modalities

- Genomics but also increasingly proteomics, esp in pediatric cancer
    - Rare tumors behave like other tumors but don’t have key variant
- Imagingg: files are there, but want a more integrative experience for folks wanting to do things like radio genomics

Clinical data flows

- Opportunity to use data to inform real time things like molecular tumor boards, how do we intersect these on behalf of kids with poor prognosis where time matters
- FHIR to connect to clinical systems, API to create apps and different tools, intersect research and clinical data flows lots of opps
- EMR, imaging, how do you get data out, de-ID, use in research
- NCPI FHIR working group
- Is this a path to interop with dbGaP?

Children’s Brain Tumor Network + PNOC (clincial trial arm)

- WG and RNAseq characterization
- Releasing PNOC integrating with landscape of CBTN data, what goes into clinical trial
- How do we do this and how do partner with the right standards
- Clustering uses CBTN as backdrop, see patients most similar, access in real time

- Zero Childhood Cancer (zerochildhoodcancer.org.au)
    - Scaling up to sequence every child with childhood cancer in Australia, WGS and RNA seq
- International Pediatric Cures Project (ipc-project.eu)
    - Federated learning, high level analytics