Global Alliance
for Genomics & Health

ARCHIVED

Latest Minute document 2021

GA4GH Discovery Work Stream

Minutes and Actions 2017-2018

These are the minutes for the GA4GH Discovery Work Stream. For further information, please visit the GA4GH Work Stream page (will link to the page on the website when it is ready) at http://ga4gh.org.

Table of Contents:

Meeting Protocols

Links to Sub-Group Minutes and Actions Documents

List of Meetings
- 2018-08-29: Bio-Link Model
- 2018-07-11: Sub-Group Status Updates
- 2018-06-12: Toronto Debrief / Global Genomics Network
  - Toronto Debrief
  - Verily's Data Explorer (part of Data Biosphere) - Melissa Chang
  - Pilot of Global Genomics Network
- 2018-05-08: Product Update / Toronto Meeting Plans
- 2018-04-11: RNASeq Search Use Case / Beacon v1.0 Prep
  - Product Progress - Marc & Harindra
  - RNASeq Search Use Case
  - Beacon API at 1.0 with GA4GH Approval
- 2018-03-13: Genomics England / Variant Transforms and BigQuery
  - Discovery Work Stream Update
  - Variant Transforms and BigQuery
  - Genome Interpretation Platform
- 2018-02-01: Case Log
- 2017-11-29: Initial Meeting
- 2018-MM-DD: Template

Global Alliance
for Genomics & Health

Meeting Protocols

- Please note that by participating in meetings, attendees agree to adhere to the GA4GH Standards of Professional Conduct.
- Meetings may be recorded for note-taking purposes. Recordings will be deleted within three months of the meeting taking place.
- Dates should be specified in the international format yyyy-mm-dd

Links to Sub-Group Minutes and Actions Documents

This Work Stream has meetings for the following sub groups.

1. Beacon API - 2017
2. Beacon API Technical - 2017
3. Discovery Search - 2018
4. Discovery Networks - 2018

List of Meetings

2018-08-29: Bio-Link Model

Chair: Harindra/Marc
Attendees “Name (Affiliation)”: Orion Buske, Deepak Unni, Rishi Nag, Heidi (called in)

	Actions Arising	Assigned To	Deadline
1	Re-schedule BioLink model, demos (Search/Beacon with Library Cards)		
2			

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda	3
2.	Previous Actions Review	7
3.	Sub-Group Review - Discovery Search (Harindra) - Beacon API (Marc Fiume) - For Plenary Beacon/DURI planning a demo of secured access beacon which respects GA4GH claims - that can be ascribed and put into a Library Card, allowing	20

	- Shift authorisation from white lists - Dataset might just be 1000 genomes, immaterial - Idea that this module can then be taken and placed in front of other GA4GH modules - Opens the way for the Beacon to provide more information if researcher can provide their claims - Data Discovery Explorer (Marc Fiume) - Discovery Networks (Miro Cupak)	
4.	Bio-Link Model schema - will be re-scheduled	Chris Mungall / 25
	A.O.B.	5

2018-07-11: Sub-Group Status Updates

Chair: Harindra Arachchi (Broad)
Attendees “Name (Affiliation)”: Jordi Rambla (Beacon & EGA), Deva, Olusola, Jessica Chong, Orion Buske (Gene42/PhenoTips), Rishi Nag (GA4GH), Steve from CNAG in BCN

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda - Introduction to Deva, Olusola. Interested in Beacon. Is there a reference implementation, or is more about making contents of the repo searchable? - JR: GA4GH is making the standards, and other projects are doing the implementation. Beacon API has a reference implementation with ELIXIR Beacon, currently at 0.3. In the ELIXIR Beacon GitHub repo. - Is it production ready? - JR: There is something to take VCF, load into DB. Depends on the	

	dB size, this may not scale to 100s of millions of variants. Could be extended and scaled further.	
2.	Previous Actions Review	
3.	Beacon API - Product Approval Progress Report - Re-identification paper plans for October	Jordi
4.	Data Discovery Initial Meeting Report	Marc
5.	Discovery Search Report - MME outline: 100k patients among 9 nodes - Sola: Search API + reference implementation in October? - HA: The plan, optimistic but achievable - Implementations	Harindra
6.	Discovery Networks Update	Miro
7.	Product approval process	Discussion led by RIshi
8	Discovery GitHub Approval Processes (initial chat, I think we need some help from the cloud work stream here, and possibly a GA4GH-wide process)	
	A.O.B.	

2018-06-12: Toronto Debrief / Global Genomics Network

Chair: Marc, Harindra
Attendees “Name (Affiliation)”: Harindra Arachchi (Broad/GA4GH), Jordi Rambla (EGA & Beacon/CRG), Gary Saunders (ELIXIR), David Lloyd (ELIXIR) Asha Rostamianfar (Google) Jessica Chong (Univ of Washington/MyGene2), Orion Buske (Gene42/PhenoTips), Juha Törnroos (ELIXIR-FI), Jörg Hakenberg (Illumina), Rudi Koopmann (biologis GIM), Ben Hutton (Sanger/DECIPHER)

(~13 participants)

	Actions Arising	Assigned To	Deadline
1	Data Biosphere and Cloud WS plans		
2			

	Agenda Item	Person/Time

1.	Welcome & Confirm Agenda	5
2.	Post-Toronto Product Status Review and discussion - Discovery Search - Discovery Networks - Beacon API v1.0 status	Marc, Harindra / 20 Orion Buske
	Presentation of Verily's Data Explorer (part of Data Biosphere)	Melissa Chang
3.	Data Discovery plans (See https://doodle.com/poll/zsw7tn2psiem2twn for meeting plan)	Marc / 15
4.	Pilot of a Global Genomics Network	Harindra / 15
	A.O.B.	5

Toronto Debrief

Discovery Search API Strong conversation led by Orion, Ben, Ryan and Harindra. API has been sent around, meeting with Melissa planned with Data Security this week.

Beacon Group had a meeting as well.

Conversations with Moran Cabili led to ideas for a **Data Explorer** - 'find me data that looks like this'. Not doing phenotype interpretation, depth search. Matching data/metadata.

Orion Buske: **on Clin/Pheno WS - they have three activities**

- Education and Publisher support to structured phenotype data
- Suggested best practices - which ontologies to use when
- **Phenopackets** - work to align concept and the schema by Jules of the Monarch Initiative to be aligned with FHIR standard
    - HA: How far along are they? Is it a tool or a specification?
    - OB: It has some history - sort of both and changing over time. The concept of a phenopacket is a small chunk of phenotype and metadata around a sample. Something to be attached to a publication. Extensions then to support some of the more common genomics data e.g. bundled in a VCF file, maybe with pedigree. This extension has tooling around that idea.
    - HA: Raw data formats? So tagging output from sensory machine to be represented across the wire.
    - OB: That would be a stretch goal

GA4GH Steering Committee meeting to vote in policies. Beacon is going through at 1.0 for next steering committee. Product Process is in place, going to be used for Search API.

- JR: We need implementations for the process?
- HA: Yes. I wanted security to be done earlier as well.
- JR: Are we planning to have any?
- HA: There are a few groups on MatchMaker side. Should be straightforward to adapt our existing tools. Can use limited 50 patient test case.

Verily's Data Explorer (part of Data Biosphere) - Melissa Chang

Suite of tools for Life Science Researchers, may undergo rebranding change
GitHub available
Currently one data set, needs to expands to multiple.
Need a schema - not sure if we will adopt or develop
DataBiosphere is essentially next generation of FireCloud, which has a Data Library to search across datasets, this will be similar
Identity Aware Proxy allows restriction of apps to Google Group members. Easy to set private data explorers.

MF: What are the first datasets to be hosted?

- Nurses Health study for 20 years
- Project Baseline - 10,000 people
- NIH AMP project to aggregate data from multiple Parkinsons disease studies

MC: Eventually more integration between cohorts, so accessible with notebooks, genomics workflows.
MF: Will this connect with Cloud Work Stream/WES?
MC: Not heard, will have to check with Moran on this.

MF: Moran mentioned to come up with facet IDs is a standardised set of facets. E.g. sex vs gender, so if multiple data providers standardise
MC: Do you have a concept of base standards?
MF: To be developed by the community, but we have enumerated facets is use between certain 6/7 repositories. There is a 40-50% overlap in gross level overlap that can be searched on. Some areas will be very different, e.g. autism is a different use case to cancer.
MC: Looking at HCA they have 100s/1000s. Brian O Connor suggested there may be a detailed schema for Commons data sets. We needs 10s rather than 1000s.
MF: We did a cursory look, sharing that would be great.

MF: **How should we approach this? In GA4GH, In Discovery Work Stream, or Clinotypic/Phenotypic Data Capture?**
HA: Has a lot of overlaps. Need to establish drivers. See the need, but not from drivers perhaps.
MF: We can see who is interested and see who puts their hand up
HA: With Phenopackets, ClinPheno speak a different language, but there is scope for overlap.
JR: I think ClinPheno group, I think maybe this should go in the query thing.

MF: Doodle Poll set up - Moran giving an in-depth Firecloud, folks from ELIXIR as well.

MC: This is currently the UI. Intended for researchers. Not considered exposing the APIs.

Rough outline of an API shown by Marc:

1. POST /query and GET /facet
2. Standardise on the ontology of facets

MC: Can't answer if this would fit into our roadmap regarding exposure, Moran can advise better. Looks fine technically.

Pilot of Global Genomics Network

OB: Meeting of stakeholders - GeL, AGHA, Discovery WS leads, Heidi Rehm. In what capacity can clinical reference data be shared between the big National Initiatives to improve the evaluation and curation of variants to help solve cases.

- What to focus on, how do we get there was the questions. Taking inspiration from MME - variant level needs to be shared. Challenge is the clinical data is senstive.
- Hypothetical Genomics England - AGHA sharing data example mooted, potentially with US, Canada projects well.
- Needed legal counsel on this. Augusto Rendon to discuss this with GeL first.
- Seemed people felt some NGO needed as a central umbrella over this that could sign legal agreements
- Sense was this was on the rare disease scope - variant case level clinical data at the individual level, together with the annotations for that variant and that individual.
- HA: Some background: This discussion started with discussions from Augusto. Last few months we have had some discussions on the best fit for the genomics or clinical data. For it to be useful needed to be clinical data. GeL data - 20% rare disease/80% cancer we think. Making this available to an AUS hospital was a suggested goal. GeL would have to sign an agreement with that hospital.
- OB: Also an issue was commercial labs relation here. Perhaps think of this as a restricted access case-level beacon. The query initially being centered around the variant of interest is a very Beacon like query.
- HA - this is currently worth investigating, but have to await the legal discussions GeL has to have

2018-05-08: Product Update / Toronto Meeting Plans

Chair:
Attendees "Name (Affiliation)":

	Actions Arising	Assigned To	Deadline
1			

2			

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda	3
2.	Previous Actions Review	2
3.	Product Status Review and discussion - Discovery Search - Discovery Networks - Beacon API v1.0 status	25
4.	Toronto Discovery Work Stream plans	20
	A.O.B.	5

2018-04-11: RNASeq Search Use Case / Beacon v1.0 Prep

Chair: Marc Fiume, Harindra Arachchi
Attendees “Name (Affiliation)”: Michael Baudis (UZH/SIB); Jordi Rambla (ELIXIR, EGA, CRG), Saif Ur-Rehman (EGA EMBL-EBI), David Lloyd (ELIXIR), Asha Rostamianfar (Google), Bashir Sadjad (Google); Sean Upchurch, Barbara Wold

	Actions Arising	Assigned To	Deadline
1	Send a more definitive list of human readable examples to you	Sean Upchurch	2018-04-30
2	Meet with Gary to discuss ELIXIR Beacon repo location, submission process	Rishi	2018-04-18

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda	3
2.	RNASeq Search Use Case (slides)	Sean Upchurch / 15
3.	Beacon API v1.0 for GA4GH Approval	Marc / 15

4	Product Progress discussion with Driver Projects	Marc, Harindra / 22
	A.O.B. - https://ga4gh-discovery.github.io - will be the main entry point from the GA4GH website. Harindra has set up search. Networks and Beacons to come into play	5

Product Progress - Marc & Harindra

Search - went through early draft, plan to have v1 in a couple of months and then look at implementations. Built off work from Orion Buske, Tony Brooks and some new items

Network - Service Registry API is drafted. Beacon Network expects to be transitioned to use new Service Registry by Toronto meeting

Submission Product possibly on hold as ClinVar are delayed. Some talk with GKS and Andy Yates. We may just have too much on our hands

- MF: Folks from Norway, Canada are also interested. Think we should shelve but look at this maybe after Beacon has come into play

RNASeq Search Use Case

Sean Upchurch, Barbara Wold (slides) - additional points and discussions below

Data types

- Library prep is mainly cDNA
- Sequencing
    - Image based RNA files looking at coming online for the end of the year with HCA
- quantification takes place in a couple of steps, dependencies on the exact programs used
- Transcript mapping is non-genomic sequences (splices, polyA tails) and multiple transcripts per gene

Data Duplication is an issue - different pipelines may run off the same data set, producing different results. Need help here.

HA: Do these pipelines run on the Cloud?
SU: Traditionally run locally but being set up the run on the Cloud
BW: HCA setup to run from the cloud

HA: What use type of the data?
SU: All RNA data queryable, but mainly cancer in Driver Projects.

Search Use Cases
(more details listed on the slides)

- Metadata based queries most common current use case
- Data heirarchy based search
- Profile matching - most requested future use case
- Slice on threshold expression values - another new item
- Retrieve raw data files - use of this should diminish

The reads have no introns, so when aligned to the genomic reference, are the tools adapted for this?

- **Yes, there are special tools like Star etc**

HA: Centralisation of data? Would **our query system be querying a network or a single access point**?

- BW: Current structure of participating driver/public projects is primarliy its own database. But they are all looking at internal federation, so a bunch of nodes. So there may be a super aggregator, but we can't rely on this existing.

BW: People are sensitive to **how things are processed allows how well things can be compared with each other**. The first impulse for someone doing comparisons of a new data set to their existing processed data, is to acquire all the new data and reprocess data so they can be compared more directly to that persons existing way of doing things. This will result a range of possibilities where searchers will find things ranging from structure duplicated samples up to in major overlapping sets of data having been reprocessed multiple times. End user may want to do a search either wanted to see duplicate samples that have been processed in the same way or wanting to eliminate.

- HA: A lot of what you said maps to the
- HA: We have a 'provenance' area documenting what methods and versions of API were used in generating the data. The decision is offshored to the querier.
- SU: How might this scale for several million single cells e.g. 80% HCA, 20% GDC, some from other bits, want to ensure provenance description
- HA: It should scale, it is provided by the data hosts and API itself doesn't mind the data.

HA: On the **raw data**: We are planning to return different types - phenopackets, handle to a cloud resource. I think we would look at this as a handle to a Cloud resource.

- MF: API being designed so users can do a deep dive into the raw data, but stewards can decide on the levels of aggregation

HA: Authentication and Authorization is another thing to consider. Have you decided on closed or open system?

- SU: HCA open, other drivers have closed elements

MB: **Beacon supports OAuth in the ELIXIR Beacon implementation**

- MF: Is there something that does both?
- MB: EGA has this, but I am not really following in depth. Dylan Spalding or Jordi will be able to provide more info. ELIXIR Finland building this

HA: Do you have software that networks these nodes together
SU: No, going to start on implementation items

MF: What is the **back end data store**?

- HCA using large commercial clouds
- GDC don't know
- I believe they use a database of some sort

HA: I'm thinking there will be generic software on a specific URL and then keep track of that using our Service Registry which will give you an expandable network. It won't handle authentication but it can be baked into the syste.

MF: I like what you've provided as a set of use cases. We want to use whether or not our Search API can deliver on those, and if it is pleasant or painful to do so, as a litmus test for if the API delivers or not. Please provide more if you can.

HA: It looks like the use cases map to what we have though we will need to add extra types. We would have to add a type for expression data. Sounds like you have multiple options here

- SU: FPKM, Counts, PPM
- HA: Those would be the extra filters then

MF: I am not sure how we translate between search API outputs

- HA: return type is not a solid type, but a framework that returns an ontology. Ontology could be Beacon, MM, RNASeq
- MF: Server side operation?
- HA: This is server side, but not every format needs to be supported by a server. No way to have a single return type for everybidy
- MF: In this case we need to have initial reference implementation of each of those to allow people to use this

MF: Hosts can decide to serve multiple formats. Some MM nodes are thinking about becoming Beacons

- SU: This is useful - we are looking at multiple formats of expression matrices and we may need to resolve this
- MF: Either **query parameter or 'here's what I can give you'**
- MF: Could be a case where Beacon part says 'yes' we have this variant and MM format is for the patients

- MF: I haven't seen anything for fields for a DOS pointer to link to the raw file? The soft landing for these queries is the raw file. We need to check with whoever is developing the data model that there is a way of accessing in DOS URL

HA: The Search API is succint but has areas for multiple supporting types. Start off with initial use cases, then expand to new types. Michael Baudis had good ideas in Orlando

- MB: We can discuss in a Search meeting

So next steps are

- Enumerate different search queries
- Sean: I will get a more definitive list of human readable examples to you

Beacon API at 1.0 with GA4GH Approval

MF: There is a formal process for GA4GH
RN: Final step is Steering Committee approval in Toronto, so working back needs to be approved to them before May 25.

Does specification built in to OAuth2?

- JR: Not part of the specification itself
- RN: The one review that happened so far was for htsget. **Data Security requested OAuth tokens as part of this** so might be needed.
- MB: Gary Saunders has taken this on, and strategy was to fix 0.4 and use this as a basis for this. Dylan Spalding has had a look at the security sheet.

Specification is located in ga4gh-beacon organisation but there has been discussion about moving to ga4gh-discovery.

- HA: A product from GA4GH is a standard, so we can keep all the Discovery products, but the implementations anywhere
- RN: Happy to move this across?
- MB: I think this has to go through ELIXIR Beacon leads. I think we thought GA4GH version that can reside anywhere and this could be whatever structure. The development version sits in ga4gh-beacon.
- RN: Will look at setting this up then

2018-03-13: Genomics England / Variant Transforms and BigQuery

Chair:Harindra/Marc
Attendance: Augusto, Asha, Jordi Rambla (DP EGA & DP Elixir Beacon)

Attendees "Name (Affiliation)": Asha Rostamianfar, Antonio Rueda, Augusto Rendon, Chrysanthi Ainali, David Lloyd, Deepak Unni, François Schiettecatte, Gustavo Glusman (ISB),

Global Alliance for Genomics & Health

Heidi Sofia, Jordi Rambla, Mark Fiume, Matthew Brush, Miro Cupak, Nara Lygia Sobreira, Orion Buske, Rudi Koopmann, Sabela, Harindra Arachchi

	Actions Arising	Assigned To	Deadline
1	https://ga4gh-discovery.github.io - help needed in development	All	
2	Sharing data meeting setup	Augusto, Nara, Harindra	

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda	1
2	Sub Group quick update	10
3.	Presentation - Variant Transforms and BigQuery (slides)	Asha / 20
4	Genome Interpretation Platform - Augusto Rendon (slides) Driver Project Presentation - Genomics England	Augusto / 25
	A.O.B.	4

Discovery Work Stream Update

- This is a monthly meeting
- Monthly leadership meeting established
- https://ga4gh-discovery.github.io has been setup, needs some development
    - Kanban board setup

SubGroup reports

Beacon

- Working 0.5 API, hopefully basis of API
- ELIXIR Beacon face to face next week, will fature a hackathon

Search Team

Product: Search-API

1. MME technical meetings are now merging with DWS Search (beta).
    1. They will still use some of the MME-general meetings to discuss updates on their v1.0 API.
2. We have started compiling and gathering drafts and potential solutions at
    1. https://docs.google.com/document/d/1jPPVhSvmzW5kK_rKTxkjPvxlcGicHhHPgO4cqqP3CBU/edit
3. DWS Search meetings had good participation
4. Presentation by Kyle Farh and Victoria Popic on Illumina Case Log

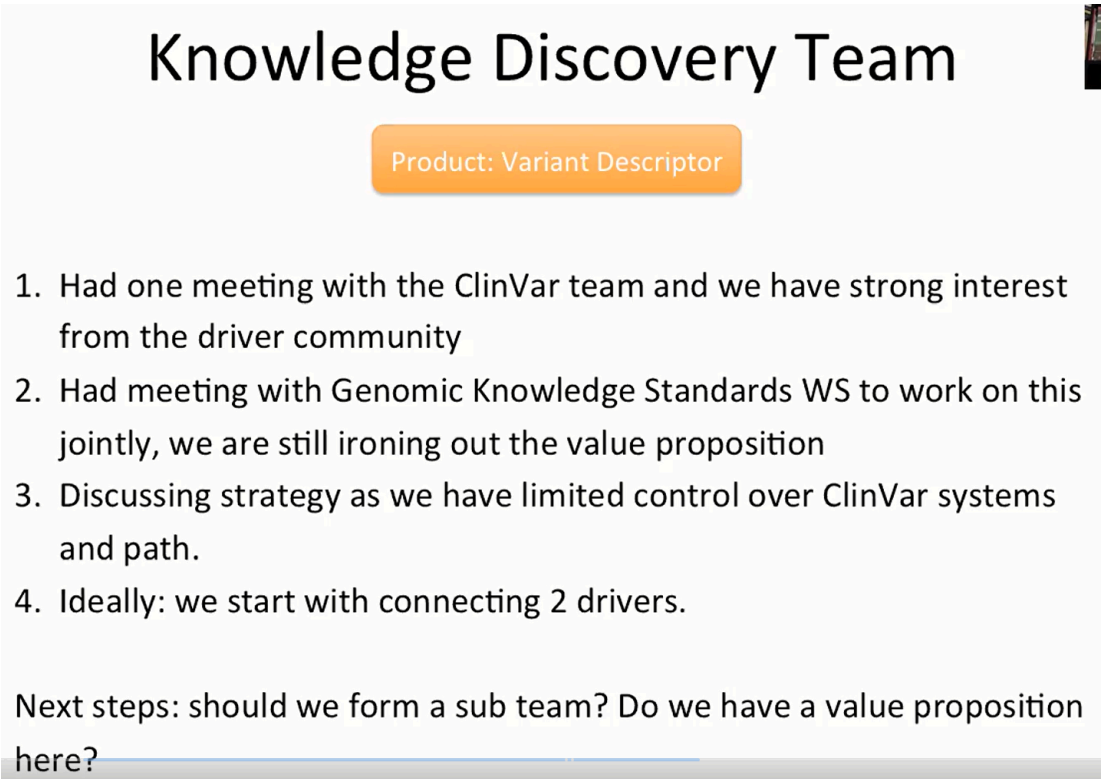
MME and DWS search better aligned now. Narrowing down drafts to one specification.

Networks Team

Product: Service Registry

1. Presentations on Genome Fingerprints and Beacon Network on the last conference call.
2. Scheduled a regular meeting every 4 weeks.
3. Started work on DOS Registry in collaboration with the Cloud Workstream.
4. There is a release (1yrs ago) of a service registry built by Marc Duby https://github.com/marcdubybroad/Ga4ghRegistry/

DOS registry work hopefully cover DOS and Beacon



Sub-Group meetings not started. It is an item on the roadmap, working with GKS to looking at a good value proposition.

Variant Transforms and BigQuery

(slides) (BigQuery site) Variant Transforms (link to GitHub)
Additional Notes:
VCFs passed into Variant Transforms to produce a BigQuery
BigQuery Table - rows are entries from VCF file, common entries are merged
Google hosts public data
Roadmap

- Robustness: rely on VCF files having valid and complete headers, not always true
- Roadmap available on GitHub as well

Augusto: Do you expect VCFs to be normalised (left-aligned, trimmed etc)? Ensures same variant addressed in a different way.

- No normalisation is done, but we want to look at that aspect.
- Phase not provided then we assign it to the same phase

Marc: How do table sizes scale?

- Looking into this at the moment, developing a tool to estimate size of BigQuery table

generated from a set of VCF files

Harindra: Have you discussed with the Cloud WS on Testbed platform?

- Not so much on that, more on the pipelines
- Marc: They are also working on Data Object Schema - representing objects on multiple clouds. Brian mentioned they could extend the URLs to items beyond files, e.g. could be BigQuery table or some search results

Harindra: How couple is the pipeline to VCF?

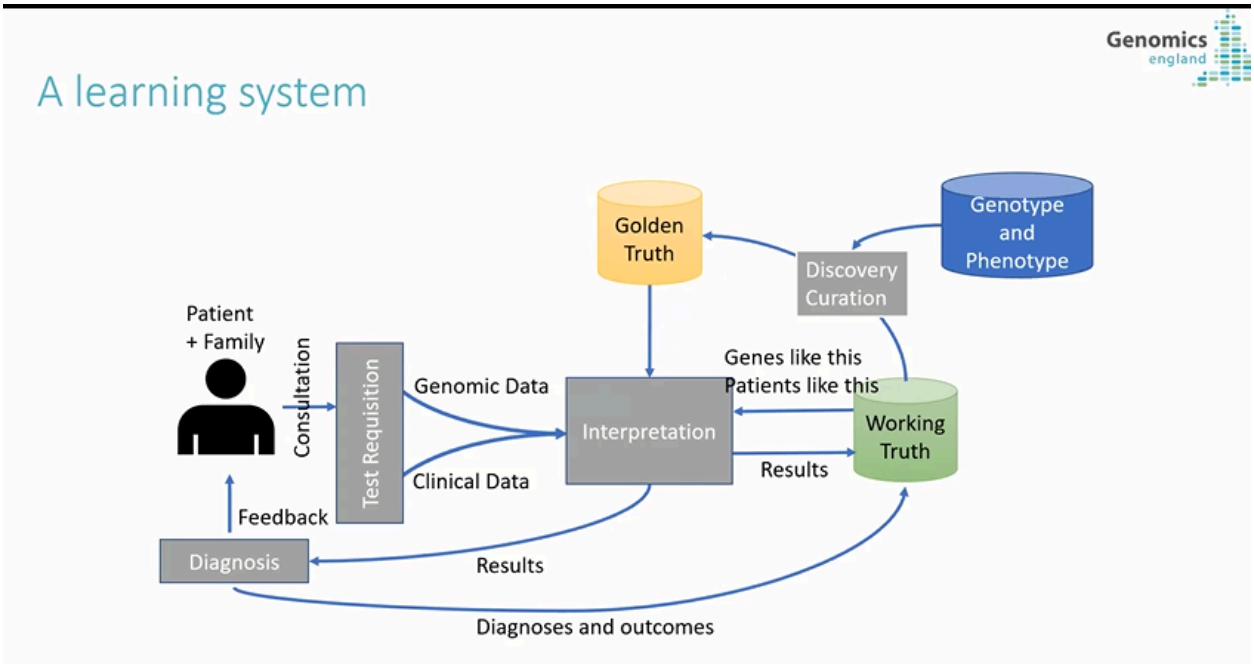
- Have an in-line intermediate class at the moment, going to move towards protos

Genome Interpretation Platform

Augusto Rendon, Genomics England (slides)

We want to discuss our internal developed items, and get feedback to see if should can be hardened etc.

Learning System



- Test Requistion for us this is whole genome sequencing and clinical data
- Interpretation uses literature (the ‘Golden Truth’)
- ‘Working Truth’ is genes and patient info. Events accumulate here. Curation happens after a certain number of events, then gets added to literature.

Interpretation Ecosystem

CVA: is the ‘Working Truth’
Outcomes Questionaire used to close the case
API and Visualisations available.

Harindra - **What part of the data are you willing to share**?
AR: We can write per institution data sharing agreements. Once there is a legal basis, an option that can be considered is per query costs.
Nara: Are you planning to share identifiable data?
AR: All is identifiable. Patient names, DOB, number will be removed. Only possible if there is a sharing agreement, this would not be for global usage.
Nara: Do you have consent to share indetifiable data?
AR: Clinical purpose, not for research
Nara: We find it tricky to define a line between clinical and research
AR: It depends on the institute we have the agreement with. If it is a service provider of clinical purposes. E.g, there is a network of labs that shares the information. We could do that if there is legal governance. For us the hospitals are fine with but across national boundaries is another step.

2018-02-01: Case Log

Chair: Harindra Arachchi
Attendees “Name (Affiliation)”: Ben Hutton, Saif, Heidi Rehm, Ilkka Lappalainen, Victoria Popic, David Lloyd, J Stone, Heidi Sofia, Orion Buske, Jordi Rambla, mhb, Kyle Fahr (Ilumina), Miro Cupak, Antonio Ruerdo, Milan, Michael Baudis, Susheel Varma

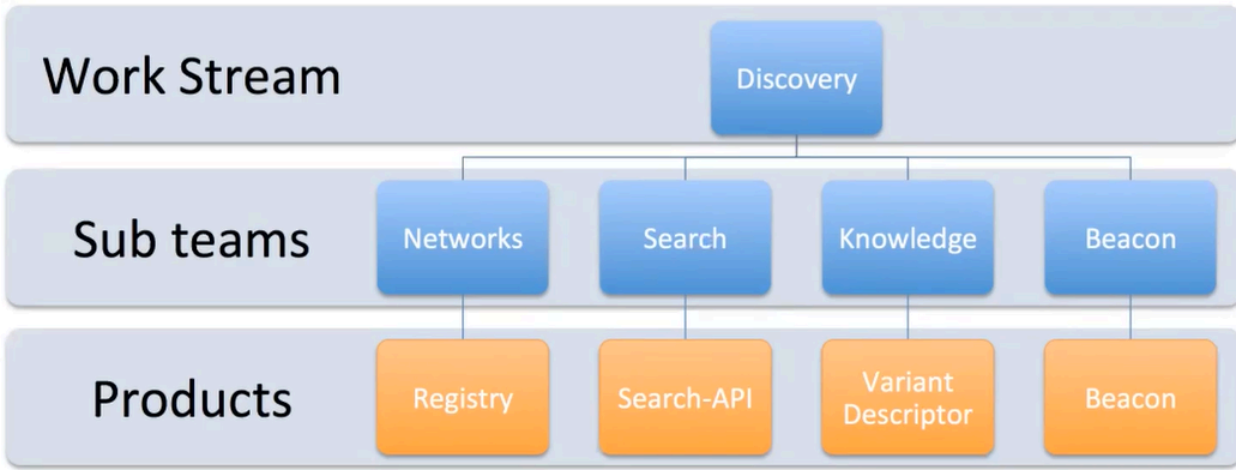
	Actions Arising	Assigned To	Deadline
1	Next meeting in one month with tech presentation from Asha	Rishi	2018-02-28
2			

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda	3
2	DWS state of the union including Discovery Sub Groups Update	10 (Harindra)
4.	Caselog Tools	30 (Kyle)
5.	Driver Project Feedback	12
	A.O.B.	5

Progress Update

Global Alliance
for Genomics & Health

Discovery Product Teams



Products from needs of the Driver Project community.
Currently one product per team - could be more

Search - building Search API - Beacon and Matchmaker led. Other teams are expressing an interest. Aim to have more sophisticated queries. Going to start a first draft.

Networks - the product is a service registry of services and data sources. These should have some sort of API. Working on a list of deliverables.

Beacon - legacy project, moving towards a 1.0 release. Beacon API 0.4 specification released.

Knowledge Discovery Team - Orlando showed need for ClinVar submission. Specification to describe a variant in a clinical application, using ClinVar as a driving use case. Joint with GKS. Need to clarify the value add.

Ben: Seems that Search might be moving quick. Want to chat about
HA: We haven't decided on path forward, but we do have strategies converging. Solutions are not final. We want to start with writing a draft down.
Ben: Jules spent a while moving forward.
HA: We can bring this up in the next search meeting
Rishi: There will be another meeting in 3 weeks or so. Jules was on the call yesterday. However there was a European interest groups in progress. Minutes will probably come through next week.

HA: We want this other groups to join in other discussions
Michael Baudis: Actively explore Beacon/MM have an input
Ben: We also want user stories which we haven't been capturing so far
Ilka: It is important to then have Plan A and then Plan B. If joint search not possible, where do we fall back to
Rishi: One thing is to look at a Solr reference implementation which would guide our thoughts on this. Within MME there would be variation on MM nodes being able to take this up.

CaseLog Presentation
(awaiting presentation from Kyle)

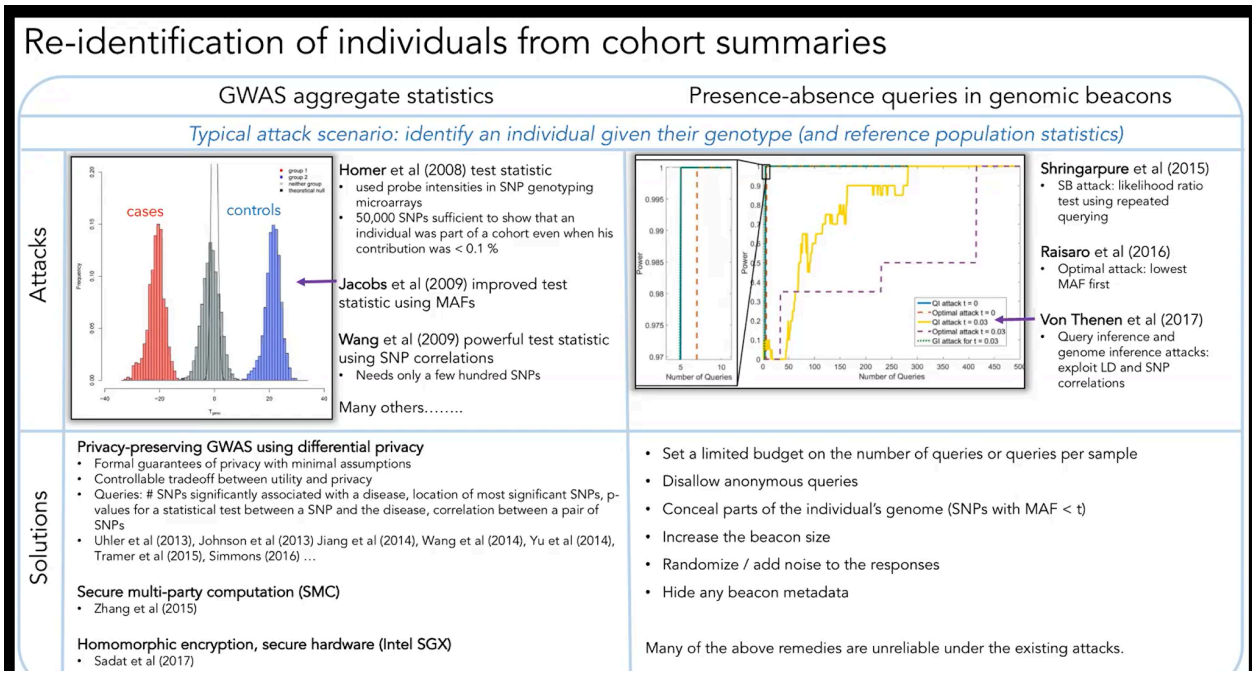
Harindra - do each of the federated nodes control how much data they share?

- Kyle: no restrictions, can work with people to add further

Harindra - Does the aggregation system on top understand underlying technologies to be different in each of the nodes?

- Kyle: An unsolved problem. Looking at machine learning approach to this. Open to ideas for collaboration on
- Milan (lead of engineering team on the project): yesterday there was a discussion on underlying technologies. Illumina has taken an agnostic approach. You should be able to access a web front end, data behind should be used as needed.
- Milan: Individual data elements can be controlled, controlling API access is easier.
- Milan: Aggregation of data across federated results is complicated

Re-identification slide from Victoria:



Michael Baudis: Just to remind everybody that the "Bustamante-Style" attack scenario was predicted at the first GA4GH plenary 2014, before launching beacons; and this was seen as part of the "social experiment" (willingness to implement nevertheless).
kfarh: Yes, especially when it comes to rare variants -- unfortunately rare variants are the most clinical relevant.

2017-11-29: Initial Meeting

Chair: Harindra Aranchi (Broad), Marc Fiume (DNAStack)
Attendees “Name (Affiliation)”: Heidi Sofia (NIH/TopMed), Jessica Chong (MyGene2 - a MME node), Junjun Zhang (OICR-ICGC Argo), Jenina Jeff (Illumina, representing Elliott, Jen Stone), Rishi Nag (GA4GH), Juha Toonroos (ELIXIR Beacon), Kyle Farh (Illumina), Malachi Griffith (VICC/WashU), Ben Hutton (DeCIPHER Project at WTSI), Orion Buske (Phenome Central/MME), Sabela de la Torre (ELIXIR Beacon), Jules Jacobsen (Monarch Initiative/Phenopackets), Antonio (Genomics England)

	Actions Arising	Assigned To	Deadline
1.	ClinVar - Ensure they will code an implementation - Working alongside VICC for germline variant interpretation - Identify diagnostic labs who regularly submit data - More information leads to a more intense submission process		
2	Product Drivers Matrix Confirmed/Updated by Driver Projects	Driver Project representatives	
3	Doodle for Search Team call	Rishi/Marc	
4	Doodle for Discovery WS call. Next one to feature caselog tool	Rishi	
5	Nominating governance and representative from product teams	Marc/All	
6	Set up product team mailing lists	Rishi	

	Agenda Item	Speaker	Time

1.	Welcome	Chair	5
2	Strategy Document / 2018 Road Map Review	Harindra/Marc	25
3	Merits of a Search Product Team	Marc	25
	AOB		5

Introduction (Marc/Harindra)

- Still trying to listen to the needs of the community
- Would like to present the organisational structure to deliver the needs of the Driver Projects. This is detailed in Strategy Document
- Mixture of existing projects and new items in Orlando

Introduction on the four distinct products

1. ClinVar submission - meeting on December 8th.
    a. We won't write the code, just the submission items.
    b. Need to get ClinVar buy in to develop the code
2. Beacon
3. Genotype-Phenotype Search
    a. MME has been functioning independantly so far. MME Nodes very closed but want to allow a richer set of queries
    b. Beacon very simple, want to go deeper into their query system
    c. Potential for alignment here
4. Internet of Genomics - a moonshot!
    a. In Orlando, many Driver Projects raised how difficult to share clinical data. How you describe a patient? Sharing data between hospitals? Take account of regulations?
    b. Can leaning on existing systems. Need to identify the good items and bring them together.

Kyle: what is distinction between enabling genotype-phenotype data vs full data sharing as identified on number 4?

- Harindra: Most institutions, countries won't allow patient data to leave. Solution might be sharing data without sharing data - abstracted in some way
- Marc: I view IoG as a coming together of multiple facets
    - Network Infrastructure to be made more general
    - The genotype-phenotype engine
    - Can be subsetted by the existing Beacon/MME type standards. Initial discovery AAI hoops

Q. Would the Beacon Network cease to exist?

Marc: Two options, we haven't decided between these. There is interest in both approaches

- Maintain the status quo, not changing expressivity of the language but hand-off

- Or allow MME/BEacon to converge

Harindra: Common ground between a Enhanced Beacon Query system and MME 2.0
Beacon if it stays as is, is easy. Making something 'new' would be easier to adopt
Harinda: We can't make MME 1.0 to look like Beacon, so enhanced query system in parallel

Harindra: The underlying goal is to define something, but there has to be a user on the other side. Driver Project will be the original builders.
Rishi: To clarify - a GA4GH standard will have to have implementations to be Approved

Harindra: We will have a meeting with ClinVar/ClinGen leads to discuss the first proposed product with them. **Who submits data to ClinVar?**

- Malachi: VICC have submitted data to ClinVar.
    - There are many groups who would like a more streamline process. One item to discuss is verification. I think ClinVar likes the slow and steady submission process, it allows them to understand and start to trust you, the submission has back and forth reviews, which leads to more trusted submissions. On the flipside you may be losing people due to the complicated/slow submission process
    - Other efforts exist like the VICC that have similarity to ClinVar. Representing a variety of resources that accept data, interpretations from others like CIVIC. Submission currently through a GUI. People want a sumbission API for this as well.
    - ClinVar a good starting point
- Marc: We want want two people for the standard. Are ClinVar and VICC close enough for a standard in this space?
    - Malachi: Yes. We could start by supporting germline cancer variant interpretations and then ClinVar variant interpretations that are mostly germline, and then extend to somatic particular items.
- Marc: We were caught off guard by the technical challenge facing people
- Harindra: ClinGen is supportive of this. The fine print is the coding of this. ClinVar will not allow outside parties to touch their code.
- Malachi: It would be interesting to identify diagnostic labs who regularly submit data, they would have a vested interested in this process.
- Ben: We do for **DeCipher**. Initial idea to push data through an FTP system. But an API would be much better
- **Illumina** did it in a bulk way, after cleaning up internal data.
- Jessica: People are putting in a minimal amount of information, particularly for supporting evidence. We just submit position and variant
    - Malachi: Maybe with a
    - Jessica: Adding provenance results in more questions, and longer road to getting the data
        - Doesn't sound like an API issue
        - Harindra: I think if we can automate the process, it is a first step to reducing error rates.

- Malachi: They are trying to understand how your idea of the data model fit with their data model
- Orion: There is a representation issue to be looked at
- Jessica: Maybe the questions should be specified in the API
- Malachi: e.g. What is the evidence that makes a variant pathogenic?
- Jessica: At the moment we can submit something and label it as pathogenic, and it gets accepted. If we say "...in four patients", we get a whole other round of questioning to deal with.
- Harindra: we have a Discovery + ClinVar + ClinGen meeting, and then go from there

Marc: We also need to review the Products and Teams section. This can be done in your own time if desired.

- Kyle: You should be able to query the variant with the same variant model which you submit it with.

Harindra: One product team per product, or one team working on multiple products? I guess only the first has multiple products.

- Marc: We found the Beacon was an incubation space. This could occur in this space
- Harindra: Do we want the Driver Projects
- Marc: Drivers may not have good visibility. We need to make sure the Kyle point above is addressed.

Consent Codes - we can remove this product, they are used so needs to be accounted for.
Marc: We have a draft Driver Project:Product Teams matrix at the bottom of this document - please update if required. This can be done in a two week time frame

Logistics Items

Governance: Each of those product teams should have Driver Project Representative
Harindra: High-level Discovery Work Stream to meet monthly. Alternate ones to have

- Caselog tool for the next item.
  - Yes, this can be made to happen.
- Each product team can have its own mailing list

Marc: Search Team discussions - we need to organise these. We would like people to enumerate the queries that people would like to be able to make. Had some good discussions with Antonio from Genomics England

Any TopMed requirements?

- Heidi: nothing specific at this point

MyGene2 requirements?

- Nothing missing

ICGC?

- Junjun: we are more observing at the moment

Phenopackets items?

- Nothing additional

Marc: Will MM v2 fit the phenotype-genotype model we are setting up?

- Orion: v2 has not evolved since a discussion from two years ago. Need to revive this item and see how it goes
- Marc: Broad, Google may be looking at this area
- Harindra: We need to gather our needs together and then meet with Beacon
- Orion: Would like to see how it fits with the ClinVar submission item. It would be good to maybe define something in ProtoBuf that could be shared. Variant representation
- Orion: The variant data itself may not be communicated,

2018-MM-DD: Template

Chair:
Attendees “Name (Affiliation)”:

	Actions Arising	Assigned To	Deadline
1			
2			

	Agenda Item	Person/Time
1.	Welcome & Confirm Agenda	
2.	Previous Actions Review	
3.		
	A.O.B.	