Global Alliance for Genomics & Health

GA4GH Clin/Pheno: Phenopackets Cancer Task Team

20201112 - Meeting #5

Attendees: Peter Robinson, Alex Wagner, Moni Munoz-Torres, Jules Jacobsen, Thomas Campion, Ania Trelinska-Finger, Lincoln Stein, Ian Fore, May Terry, Andrea Sboner, Grant Wood

	Actions Arising	Deadline
1.	Discussion to continue, Peter and Jules will add an issue to Github	
2.		
3.		
4.		
5.		

Agenda

1) Welcome, Housekeeping
2) (cont. From last meeting) Review procedure/process and Sample issues

Meeting Notes:
PR: Reviewing the biosample element today. Phenopacket is talking about constitutional phenotypes. If we go to Biosample, might be referring to a piece of sample tissue, might have several biosamples, e.g. lymph nodes. Each of those would have a biosample. If we did seq on the tumor, would list their HTS files here.
Phenopacket is a top-level element. Phenopacket can have one or more biosample elements. Phenopacket building blocks - biosample is one of them. Have been talking to date of

phenopacket v2, question today is whether the elements we have in biosample v1 are adequate or need to be extended.
IF: Definition of a biosample and what it is. In this case, that first sentence, I read this as the immediate parent. In different sampling strategies, you could have had some hierarchy above this. Defining here the immediate parent from which the molecular extract was made for some type of molecular analysis. The general models that deal with this are along the lines of having a biomaterial that can beget a biomaterial and so on. You had to have this abstraction of a biomaterial. One fixed hierarchy works for a particular study, but doesn't work for others. In order to have generality need to have the hierarchy and retain the semantics of that hierarchy.
PR: Perhaps we were too simplistic. Makes less sense to do a deep analysis of what exists and discuss the better way of doing it.
IF: This is doing a specific listing of attributes. Depending on where you are in the hierarchy, the attributes vary. A biomaterial might be a tissue block from which you take a slide. The histological diagnosis might be done on a slide or collection of slides. In some abstractions the subject is a biomaterial. Where do you state that histological diagnosis. Stick it as close to the actual piece of material from which it came.
PR: Would it make sense for a recursive definition.
JJ: Have parent ID or child links. Say this is derived from sample whatever and give it a sample ID and build tree from that.
PR: Interested in phenotype in the context of a genotype. If we had an element that was derived from biosample, might need to change some of the wording.
LS: You've attached phenotypic information to the wrong object. Consider a single-cell seq where you start with pop of 1000 cells, and each becomes a biosample. Is the phenotype an attribute of each single cell, or the surgical specimen removed from the patient.
IF: The phenotype is too general. Phenotype of B cell is B cell. I've seen the attribute phenotype too variant for the disease. Phenotype is much too vague.
LS: Phenotypes appear at different levels, like tumor stage and grade. Attached to tumor specimen or block looked at. Stage does not belong at biosample level at all. Consider it either an attribute of patient or disease, or go further and say it's all the samples from the patient.
IF: If you want the relationship between stage and biosamples, say it's an attribute of the patient and patient came from that patient.
MB: Can make the stage related to biosample because it reflects the clinical stage the patient had at extraction. How to model this is a different story.
JJ: Biosample does specifically say phenotypic features of that biosample.
PR: The stage is not attached to biosample, except that you could infer by comparing time stamps.
MT: If you think about pathologic staging, it's tied to specific procedure. Biosample is in context. Goes back to the use case, can see both sides, but depends on how you will be using and interpreting that data. Difficult to clearly identify those use cases, going with own perceptions.
IF: Use cases could harm us. The difficulty is that we have so many use cases, and if we focus too narrowly we won't be useful.
https://data.humancellatlas.org/metadata/design-principles/structure#motivation

That basic model has general applicability to do this. Been looking at this for the human tumor analysis effort. This is an example we can get our heads around. No longer a hierarchy with multiplex, because multiple connections.
PR: I think this is an excellent idea. I love the top row of Alex to monocytes. If we connected the phenopacket to monocytes and jumped over the other stuff and went directly to VCF or BAM file. This packet would have a derivedfrom element that would point to this. This would then have another one. Phenopacket would connect to monocytes. Hoping other groups in GA4GH will standardize the protocols. Intention of having this in a phenopacket is to say which VCF file do we need so a computer program can pick up the path and do analysis.
LS: Are phenopackets supposed to support a complete description of an extended experimental protocol. Are you intended to reproduce that type fo detail of a series of experiments in phenopackets, or is phenopackets a component of a larger experimental modelling format.
PR: Think of phenopackets as a component. Need to know how to connect phenotypic descriptions to the correct file. The derivation of biosample is a borderline case where we could either say we'll keep the simplistic representation we have now but refer to another protocol with gory details. Maybe we should just try to adopt another one out there. Don't want to go very far down rabbit holes.
IF: I agree with the sentiment as a component. We need to be able to hang on to its presence in a bigger picture.
MB: Use case for Beacon, started to report this wrapper. One of the more typical use cases where you search for some phenotype and variants and get what the endpoint can deliver to you, but won't get a whole EHR. Want something logically sound for a large range of use cases, bottle neck is not so much in our modelling, but in the ability of data and resource providers to fil the parts of schema with data.
IF: Biosamples at EBI. Drilled down to a specific biosample from a colon cancer study, part of CRDC. This gets there because there's an exchange between NCBI and EBI. This is how it shows up on the site. It will tell us it's a human sample, it's blood, and that's the meaningful information. Has a field for age of individual at collection but in this case it's empty. The rest is the attributes that link to the ontology.
PR: Blood is not a phenotypic feature, might need to talk to them.
JJ: They don't have a schema, so pull whatever they have. They don't enforce anything on the people that submit. I agree it's meaningless in that you got blood from a human.
IF: The JSON gives us the same biosample, but have something much richer.
JJ: Purely because it's a key value thing they have in there.
IF: Going back to they don't enforce anything, because there's a diversity in the science people are submitting. Have to allow for that diversity and find some way for that to be communicated. We need to focus on a different minimum. Tell us what you measured in a machine-readable form, rather than discussion where people will be butting heads.
LS: The science is very diverse, the informatics skills are diverse, and if you start imposing schemas you get no data because they can't format. I'm uncomfortable with the idea of attached a phenotype directly to a material. It's a measurement you make, and phenotype is defined by the measurement you make on a biological material. Always associated with a biosample. Get a

lot more flexibility if you attached a phenotype indirectly via an assay. Then don't have to make a decision about whether the histology is with the block or tumor or slide. From that, could infer it's an attribute of the sample and the tumor in hierarchy.
PR: Would people would be able to instantiate this type of data structure. Are we going to get a lot less data.
LS: Could just stay at that lowest level. Have something very similar to what is already in biosample data model, but you've interposed the experimental measurement. Gives you ability later on to extend it and do multiple assays.
PR: Does reverse direction.
MB: I think it's a pretty standard data model to have an assay between the data model. You don't have to have it as an unbroken chain, link with ID.
PR: Is there something in GA4GH we could borrow and refer to.
MB: The old model from the GA4GH schema. Went from variant, to biosample, which is derived from individual. I have the general model alongside phenopackets in schemablocks. Doesn't solve clinical stage or attaching to biosample.
PR: We want to make this more sophisticated it sounds like. Need some sort of recursion. Either put in a parentof element, or a put a list with childrenof. Might be better to have the direction be childrenof. Should look and see if Beacon has a solution like this where we could put in an ID or something.
IF: Lincoln's point about not expecting scientists to data model. How to represent these complex structures. There are ways of dealing with it, example from GDC where if you go and look at the table the investigators submitted, that had everything. Were doing it anyway, only a few things to do to that table to work with what they were producing. When they submit to dbGap, tell us what columns you recorded, and gives us the definitions of them. Contrast that with EGA, who just take whatever.
MB: Connection to cancer disease vs histology. In principal, never get just a histology, but usually a description influenced by other measures. We love NCIt for these classes, but this demonstrates that there is a point in time. Not everything is nicely separated.
LS: Example from breast cancer. Take a chunk of frozen tissue and send for target seq of 6 genes, get a series of variants on those genes. Get ERPR+ after testing. Each of these combinations give you another data point. A pathologist will combine that all together to get a phenotype, an aggregate phenotype that attaches to the chunk of frozen tissue.
PR: Element called interpretation, and phenotype elements in phenopacket are intended to be observations with a minimum of inference. Interpretation wraps the phenopacket.
IF: What's happening here is a problem we tried to fix, the reproducibility problem. For the scenario we're describing, we can containerize, want the same thing to happen with the pathologist. No where near thinking in that way.
PR: This is my evidence and this is my conclusion, that's what we want to hear from the pathologist.
MT: If your use case is patient care, pathologist is not making the decisions, just one piece. If you're expected to include the list of actors as more than just the pathologist, makes it difficult to ask of the question of who implements.

LS: Hierarchical, pathologist is taking information and rendering a phenotype on the tumor. That information goes to the clinician who combines other information like patient family history, and then creates a phenotype.
IF: We have tumor review boards, at which the pathologist is at the table. The tumor board brings all the people together.
LS: Point I wanted to make was the hierarchy of biological objects. Phenotypes attached at each level, and different actors that make decision of what that phenotype is.
MT: Agree with that.
PR: Might not have come to conclusion. Suggest that Jules and I will put an issue up on Github, might have to continue this discussion.
LS: Get it more or less right we'll be in good shape.

20201029 - Meeting #4

Attendees: Lindsay Smith, Grant Wood, Alex Wagner, Moni Munoz-Torres, Lincoln Stein, Jules Jacobsen, Ian Fore, Peter Robinson

	Actions Arising	Deadline
1.	Revisit Biosample - plan next meeting around process and sample issues	
2.	Change cumulative dose to optional	
3.	Lindsay to send email checking that a meeting in two weeks (same time) works for everyone	
4.	Add Regimen ID - Peter and Jules to prepare proposal for regimen	
5.	Treatment - need to expand treatment types, potentially move chemo and hormone therapy into treatment	

Agenda

3) Welcome, Housekeeping
4) Walkthrough/Review of:
    a) Readthedocs

b) ICGC-ARGO Dictionary Mapping
c) EUCANCan Cancer Types

Meeting Notes:

PR: Phenopackets version 1.0 was approved about a year ago. Robust for rare disease but not ready for cancer or common disease. Goal for it to be a computational model for exchange and for bioinformatics. Talk about the new medical action element, way of representing treatment. Medical action is intended to be a general thing for clinical management. Defined in the schema as an optional 'oneof'. Medical action will become optional part of the phenopacket element. If you do use it, required to provide one of the listed elements.
In treatment element, there is an agent, could be an ontology term that represents a pharmaceutical.
LS: There is a stop reason ID, which is good. But in cancer treatment, there is an intent of therapy - curative, palliative, etc.
PR: Good point. Let's a present a few of these, and then discuss where to put the things like intent and stop reason. Agree that's a good idea.
Route of administration is by mouth, intravenously. We recommend using NCIt. Another new thing is dose interval. Have a unit, a value for that unit, a scheduled frequency. Things like twice daily, or three times a week, there are ontology terms for that. Interval that uses time stamp strings, not discussed here but for privacy reasons its possible to shift intervals by a non-disclosed amount and that can be indicated in the metadata.
MT: Is this treatment is context of admin or order?
PR: Use case for something like phenopackets is computation for research or potentially for differential diagnosis. More interested for if someone actually got the medication. Drug type is also the context of the drug administration. Currently have enumeration. If it's a prescription, can be sure they didn't take exactly as prescribed.
MT: Might be fine for oral drug, quantity is relative. If using for research, depends on the type of research. If absolute administration would need to know a lot more about the patient.
IF: In phenopacket, what are we attempting to capture, clinical practice or clinical study where this would occur.
PR: I think in cancer, dosages are given relative to body surface area. Phenopacket intended to be a general tool for any disease, rare for hypertensive medication relative to BSA. Interesting question, haven't thought enough about it.
JJ: These would be specific types of treatment?
LS: In a typical predicative biomarker study. The biostatistician would only look at the regimen applied and whether the patient completed it successfully. Typically would not dig into whether patient received all the doses at the correct intervals.
IF: Relative high level. But if relatively high, what's the depth to go to.
LS: Suggest a concept of a regimen that has an ID, so for a researcher you can pick out those that got a particular regimen.

IF: SHould have an ID, but probably a name as well. The detail level is going to be necessary, but maybe here it might not be. Leave people to decide depending on their purposes.
MT: In the case Lincoln mentioned, just looking at the regimen, which opens a big can of worms that remodels regimens. If it's more about admin, we already talked about the relative dose.
PR: I believe that NCIt had milligram per m sq for body.
MT: THey do.
PR: That would partially address some things.
IF: Looking at this, seems to me that this just becomes a unit. The ID points to the NCIt.
MT: If you do have it as relative dose, need more info about the patient to capture weight and body surface area.
IF: I don't think you would go there. Somewhere in the clinic, that calculation was done. If you want to cross check what they did, you want the actual amount and the body area. Otherwise, just say the dose at a particular rate per body area. Same story as how we follow up on microarray.
PR: Medicine is not a data driven science, not capturing phenotype reliably. We need to say that this unit can be relative and provide a table of examples pointing to the relevant NCIt terms. We don't have fields to calculate body surface area but easy to add.
JJ: Add to individual message at the moment. Scant right now, trying to reduce PHI in there.
LS: I think it's fine to capture BSA, just as it's fine for patients weight and mass. As someone using the data, would prefer recommendation dosage be given in mg for m sq. So I can reach into database and find patients with similar treatment regimens and not have to divide by BSA. A drug is usually dosed in a way relative to surface area, should do so relative to that dosage.
JJ: Good for downstream analysis, all had the same thing done to them rather than having to group. With regimen, I think that's a great one to shortcut or group things. What info do you need in there? Medical actions?
PR: Don't want a definition of regimen, just want to refer to regimens. See that there are regimens for certain diseases. Do they have NCIt terms?
MT: Crux of a lot of issues. Not one overarching standard for regimens. A lot of places define NCCN but that's not the only set of regimens. Could be anything in some respects. They all have similar characteristics. Some may have alternating regimens.
PR: We just want to give people the opportunity to refer to things. If we could refer to regimens in at least NCCN, that would be a great start.
MT: NCIt has a hierarchy of some regimens. Probably not up to date. But would give you a start.
IF: May be worth a look at this -
https://ctep.cancer.gov/protocolDevelopment/policies_nomenclature.htm

IF: NCIt, some of the drug codes feed what goes on in pharmacies.
PR: Use of phenopacket not meant to be a competition for FHIR, OMOP, i2b2. More a basis for computation, especially when connected with genomics data. Do not claim you can take a phenopacket and recreate a FHIR. Want a data structure to fit into some statistical program. If there are things important for admin, probably don't want to model in phenopacket.

PR: Should we have an element called regimen, and regimen object could have an ontology term that refers to regimen and maybe some other data that says is the regimen completed or not.
LS: Yes, I think that would be helpful.
PR: Boolean for completed or not.
ATF: Not just completed or not, also need unknown, or why not completed. I think for not detail questions it's enough to say yes or no.
PR: All of this is a list, would be possible to create a phenpacket with an enumeration for completed or some other things. Could also put all details into these other elements if desired.
JJ: What do you mean by other details?
PR: I'm hearing that yes or no is good for some cases, for others we want to know the details. In theory, possible to provide both in a phenopacket.
JJ: Could end up with repetitive stuff.
PR: Easy to put a regimen as one of these options, but perhaps it goes somewhere else. Make a proposal and touch base to finalize.
LS: The implementation point of view, may prefer being able to define or use previously defined regimens, and just relate whether patient received that or a different one. May ease data entry.
JJ: Could take medical actions and use a regimen instead of medical action. Actions then go inside the regimen. Could separate regimens for different conditions.
LS: You have a series of diagnoses and a series of actions, and you'd like to know which therapies are intended for which diagnosis?
JJ: No I don't think you can do that, would be a good thing to have. Goes back to intention.
LS: The patient can have multiple diagnoses and therapies and we link them together. Possible for therapy to be applied for multiple medical issues.
PR: Intended use case of phenopackets is a bit more general.
Procedure is intended to be something like an operation, or diagnostic procedure. Codes from many places for things like punch biopsy, stent implantation. In some cases these codes include bodysite, some they do not. Think this can be done for various operations. Element currently does not have a lot of other information. Initial use case really for biosample object for how a biosample was obtained. One question is do we need more information to refer to procedures in general.
IF: Overlap for whether these (region) are going on the procedure or the specimen. This area has had a lot of work done on it. Happened in the FuGE initiative that came out of MAGE. Also covered in the HCA project. Deal with both procedure and sample, and how you get from one to the other. Worth looking at. With sample collection, starting material is the person and procedure is the biopsy.
PR: Part of that could go into biosample element.
JJ: Already in there, procedure in the biosample. Maybe join the two together.
PR: More to discuss than we have time for. Would like to go through medical actions. Revisit biosample.

LS: Note that there is a difference between the specimen that was surgically removed and the sample that went for molecular analysis. A lot of tumor heterogeneity and it matters what part of the specimen is sent.
IF: The beginning of a labyrinth. What we need is the string that gets us through the labyrinth.

PR: There are three elements that are almost specific to cancer. Chemotherapy has all of the elements of the treatment we just looked at.
MT: When I think of cumulative dose I think of radiation therapy. Not aware of cumulative dose of systemic chemo.
LS: I tried to convince the Terry Fox cancer cohort to record cumulative chemo dose and there was no enthusiasm for doing that.
PR: I've seen this in publications with small cohorts. Likelihood of side effects depending on the agents. Maybe shouldn't be required.
LS: Recoding this was requested by data analysts, but clinical folks weren't going to provide it.
IF: Chemotherapy is a complex regimen, and have to talk about cumulative dose of each thing.
PR: Even though there might be an use case, we should remove?
IF: There will always be people who want that level of detail. Scope decision, for the purpose we're intended, it's probably not what we need.
PR: Could calculate the cumulative dose if phenopacket is complete. Think about removing that.
LS: Or, just make it optional.
JJ: For cumulative dose, just have a quantity. Is there anything else we might need for that?
PR: In theory could have just one element called treatment. Would it be appropriate to not have separate things for chemo and hormone therapy?
ATF: I think it's a better idea. For me, it's confusing, treatment is a parent term for chemo and hormone. Hormone therapy is not as important as immune therapy. I'm missing important kinds of medical treatment, and the list is not complete. In a registry, you have normally three types of treatment - medical drugs, radiation and surgery. They are not mutually exclusive.
IF: I think registry data models are potentially useful input to this activity. I'm somewhat familiar with the North American ones. How do things cross over in Europe?
ATF: Not very different but there are differences.
IF: As a source, these registry data models are highly relevant. Key resource to get info about cancer patients.
ATF: Similar models that present core data elements and every country specifies a little bit. Big hospitals have some registry and interface. Find out what is used in each country and take a look at them.
PR: Phenopackets is not intended to replace a data model that exists, but would be a wrapper to say you can export your data into a phenopacket. People would only need to write one algorithm for phenopackets, wouldn't need to write one for TCGA, ICGA, etc. Easily possible to take fields labelled chemotherapy and put them into treatment. We can maybe get rid of the hormone and chemo elements and put them under treatment, add optional cumulative dose and explain that this can be used for these three things.

JJ: Prefer to get rid of it, it's a subclass of treatment. Think we should just remove the confusing subtypes.
LS: Are specific subtypes for treatments, it won't be clear which are required as a set. Put common fields together in treatment, then have specialized subclasses for chemo, radiation, etc.
JJ: Probably need to expand treatment types.
ATF: Patients get more than one type and more than one drug.
PR: Phenpacket element will include a list of medical actions, for patients can have multiple elements. Cannot subclass in the schema. Can't make chemo a subclass. But from an object oriented perspective it looks wrong.
PR: Plan the next meeting around the topic of process and sample issue.

20200714 - Meeting #3

Attendees: Grant Wood, Lindsay Smith, Alex Wagner, Moni Munoz-Torres, Peter Robinson, Andrea Sboner, Ian Fore, Thomas Campion, Jules Jacobsen, May Terry

	Actions Arising	Deadline
1.	Please review and comment on the ICGC-ARGO early investigation/mapping document	
2.	Specific issues discussed today to be added to Github issue tracker - please review and comment when online	
3.	Please connect with Moni for materials about CCDH efforts	
4.	Lindsay & Moni to organize a CCDH Presentation	

Agenda

1) Welcome, Introductions
2) Walkthrough of ICGC-ARGO Dictionary
    a) Document - please review and comment!

Meeting Notes:
CCDH - Melissa leading an effort to harmonize 15 models. Would make sense to synergize efforts. First year dedicated to putting forward a prototype for representation of cancer data.
IF: CRDC is a GA4GH DP, we hope we can tie these efforts together. Also ties in to work in Discovery Search. Looks at how one makes a model available and query it.

MMT: Matt Brush is the lead designer on the CRDC prototype, find a way to have Matt share these efforts.

PR: Briefly, phenopackets started from more of a rare disease focus, and now being used in many projects. We have been working on cancer but are only really getting started on what the full scope should be. We'd love to have close integration and not reinvent any wheels.
IF: NCPI - interoperability effort within NIH. Starting with focusing on standards on AAI. Also has a FHIR group.
PR: We have possibly too many standards. Our goal is not to be separate but to integrate with other parts of GA4GH and beyond. BUilding block standards are starting to reach maturity. Phenopackets are likely to be approved as an ISO standard by 2021. Hope to approve v2 by the end of this year. Haven't gotten any negative feedback but more feedback from users would be helpful. There is a version 1.1 that covers some COVID and temporal elements. In the meantime, we've been talking about mCODE, ARGO, CCDH - see what elements are missing from phenopackets, and are the way we modelling things appropriate for oncology. Planning on starting to look at the ARGO data dictionary.
JJ: Had a look at the CCDH, and it's incredibly close to what we're trying to do. How does ARGO fit in with that? They've taken 15 models and mapped to BRIDG. CDMH in HL7.
MMT: CCDH has some members connected to HL7. Prototype model released at the end of June [Presentation slides here].

IF: A high-level opinion about where these efforts are - still early stages. One of the key challenges is a recognition of diversity. There are things that are genuinely different in the scope of cancer data. To relate to phenopackets, when talking about a phenopacket, you're defining literal fields for a particular set of attributes for a particular disease, but the diversity of disease in cancer is enormous. A question of how to represent what different disease experts define. Or the AACR staging criteria.
PR: One justification of phenopackets is it should provide a flexible standard that can be specified for particular uses. If in a consortium, you would consider defining certain fields as being required, but depends on the use case. A lot of models have slots for things, but for phenopackets some of these can be represented as things in a list. A question of how to structure that.

PR: Looking at the ICGC data dictionary. Few things that are relatively uncontroversial to map. ARGO has IDs, program ID is something that doesn't have any correspondence in phenopackets. Would need to consider adding that for research groups. Could do that by making a new issues on the Github. Wouldn't make it specific to ARGO but there are analogous needs.
JJ: Could put into a metadata external reference.
PR: Phenopackest are implemented in protobuf, which is an easy way to get into a JSON environment. No problem to add fields. The donor ID is pretty close to individual ID, just

representing hierarchical nature. Gender is the first place where you get to mappings that are not quite the same. Phenopackets regards sex a biological sex. Initial exploration of the closest thing. Phenopackets does not represent self-reported gender.
AS: Could be an useful addition, distinguish biological sex and gender, clinical considerations for gender.
PR: Will mark comments in yellow to the doc, and will add as issue.
IF: NCBI/Biosamples. It would be interesting to see how the mappings are being done, and if it's an useful mapping. Could end up with different phenopackets representations of the same thing.
PR: Only prepared the first half of ICGC for today. Please comment on this document.

PR: In ICGC, there is a specimen_tissue_source.In Phenopackets, Biosample tissue is an ontology class, but in ICGC there's a fixed list, one of the entries being OTHER, arguably a less flexible solution. Could go from ICGC to phenopackets by assigning an ontology term. But if we want to export to ICGC, and the specimen taken from a tissue not represented in the list, could only really enter 'other' which leads to info loss. Goal for GA4GH is to try to unify data, so the use case of going from phenopackest to ICGC is less important than going from various other things to phenopacket as a common medium for exchange and computation.
Specimen_type: not really an element in phenopackets that fits this. From ICGC, could be cell lines, xenograft tumor. Preferred way would be to use an ontology term for this. Does is sound like a good idea to include this field?
JJ: I think it's a good idea, might be the only way to map some of these things, otherwise it would be very ICGC-specific.
IF: Another aspect, it's illustrated by cell lines derived from tumor, the overall question is the depth. There exists different hierarchies of specimen. There's the thing taken from the subject, but then there may be extracts from that, and so on. The tumor itself might be a thing that needs to be represented as a specimen, but also the cell lines for that tumor. Are they capturing that in any way? To get the meaning of the specimen, it's important to understand the hierarchy.
PR: Situation where we might have one specimen that's a primary piece of tumor, and then they have cell line derived from that they do an analysis on. Want to connect those two things.
IF: Within the scope of what we want to track.
PR: From my reading, it's just a list. I do not believe there is a way to related two samples together. Is there a way to do that in CCDH we could adopt?
IF: One of things that effort has tried to do is look at that hierarchy.
JE: I agree, hierarchies are important. But might work best to make it flat, it's not perfect, but might be a way to include it.
PR: Can mCODE record relationships between samples?
MT: Not between samples.
PR: Understand that yes it would be great to have, but too difficult to enter.
IF: I don't think it's too difficult, because there are standards that do this already. The column that has the type is a perfect one. The column indicates the ID of the parent.

MT: Depends on who you're trying to satisfy. When I see that list of qualifying tumor I see a lot of pre-coordinated terms, which if I were to create some kind of model behind it I would have further broken it out and made it a qualifier. I would have concerns in mCODE where if there's a list. In terms of ability to create relationships, we are looking into a codeX use case. We would potentially revisit this in light of pathology.
PR: Would storing ID of the parent tumor work?
MT: It might, I think it's gonna depend on the use cases. Seems on the surface it can.
PR: We will put this up on Github, and send out a summary with some links so everyone can enter comments later.

PR: Sample_type is another thing phenopackets doesn't have in this form. Partially represented in hts seq file. Specimen type is tumor or tissue, and sample type is what did you do with, ie. get RNA or DNA.
IF: Would have a single field for both of these things, parent would by tumor and child would by type.
PR: Phenopackets intends to concentrate on phenotype, not sure if appropriate to put this sample type into phenopacket, or whether something else is more appropriate.
JE: Demand for this list, but not sure if optimal choice. The values are valid and important to know.
PR: The thing that worries me, is if I have one tumor, I might have duplicated rows if I do multiple analyses.
IF: If you look at ISA, you get exactly what you described. Have 'biomaterial' instead of sample_type.
JE: sample type is related to specimen type.
PR: Phenopackets tries to be normalized. We're going to start to put this up as issues into Github, and hope everyone can find time to comment on the issues. A lot of items in ICGC that are not currently there in phenopackets. Decisions we need to make: staging in phenopacket is in disease, can be a list of items. In cancer there are multiple specifications of that. Staging is something to do with the patient, not the actual specimen, but ICGC puts this into the specimen.
IF: There's clinical stage and pathological stage. Pathological is tied to the specimen.
PR: Distant metastases has something to do with the patient, not the tumor.
IF: That info may well come from pathology, that there is a metastasis
MT: Related to solid tumors, there are other cancer staging systems we might want to consider. With stage groups you probably want to call out what the system is, like TMN.
IF: If talking about phenopacket staging, one would not invent the phenopacket, would build off of what the AJCC manual is. Provide the attributes to record. That manual is put together by several committees & experts. Beyond our capability to define what they should be.
MT: In-line with what mCODE decided to do. TNM by itself will not determine stage group, have three prognostic factors in Breast Cancer.
PR: Phenopackets will not make recommendations, just provide a system of slots. One question is whether it's better to put staging into biosample or to keep it where it is as an attribute of the individual.

MT: If it's pathological, it makes sense to reference a specimen, but if clinical you won't have a specimen.
IF: Key thing to represent is the 'event'. There is a surgical event in which specimens would be collected. When asking the question of where it belongs, clinical staging would belong with the event that derived radiology data.
JJ: Valid to have a pathological stage on the biosample, and a disease stage on the disease.
IF: Patient at that point in time.

20200623 - Meeting #2

Attendees: Peter Robinson, Lindsay Smith, Alex Wagner, Lincoln Stein, Andrea Sboner, Adriana Malheiro, May Terry, David Hansen

	Actions Arising	Deadline
1.	Go over the resources (ie. mCODE, Recist, ARGO dictionary) with the goal of phenopackets being able to support any of these resources. Be able to discuss a final first proposal of the elements and, in some cases, the value sets in mCODE would be enumerations or ontology terms that you could specify	
2.	Continue with meetings every 2 weeks - Lindsay to circulate a poll	ASAP
3.		

Agenda

1) Welcome, Introductions
2) Driver Project Needs
    a) What do you hope to get out of this?
    b) How are current models not meeting your needs?
3) Driving use cases/requirements from the cancer community
4) Demo?

Meeting Notes:

PR: Summarize the last meeting - learned a bit more about mCODE, would like phenopackets to cover that. Got down some notes about things that are present in mCODE but not in phenopackets. Going to assume familiarity with phenopackets and protobuf.
There is a message for procedure, which has its first element as an ontology class - term ID and label. Example of this taken from NCIt. All elements in phenopackets are optional, but might

also use the code for punch biopsy. Finally, a time element, which is itself a series of options including a time stamp or an ontology code for age or age range. In the GA4GH context we are concerned about data privacy, so if we're transmitting phenopackets across firewall we might not use the exact time but some fuzzy ontology term like age 40-42 or something. Something called treatment intent in mCODE. Phenopackets are not just for cancer, so there is another treatment intent called preventative with a default vault of unknown. One proposal is to extend our procedure element with a fourth treatment called treatment intent.
LS: Intent is important, difficult to determine from a clinical record. Would need to include biopsy done for research purposes. Sometimes there's cosmetic treatment ie. breast reconstruction. There is pathology associated with that and would want to report it. In controlled trial, you might have a placebo, so there could be a placebo or treatment control (sham therapy - not common anymore).
PR: One issue for modelling is the enum in protobuf, it's not optional. So if you declare an enum it always has to have a value. Other things can be null. To have something optional we might need a boolean and then an enumeration. Perhaps next week, we can try to document this for next week. Would welcome collaboration on this. Should this be an enumeration or not.
LS: Example of encoding bodysite vs reporting more generic procedure plus a bodysite. Offering flexibility, but then difficult to find if encoded in different way. Is there a reason for this flexibility?
PR: I don't claim that any decision are good, but we thought that looking across the entire spectrum of medicine, it's difficult to make a standard where all of the elements are required. mCODE has both a semantic and syntactic standard.
MT: Bodysite has been constrained to SNOMED codes.
PR: Our vision is that if a consortium decides to use phenopackets for a certain project or to transmit data between clinic and a lab, then there additionally has to be agreement on what fields are required. This is not done in protobuf, everything is optional. Would be done by some software that ingests a phenopackets. We have some software for java, C++. Need software that will define required elements for a project. Will add a layer of complexity. Best compromise between making a universal standard for computing phenotypic data across any field but still have some utility.
AS: In this situation, is it possible to on the back end map the two instances. Something in the middle to understand they're talking about the same thing.
PR: Ontologies can be complicated, but not hard if using a terminology such as NCIt where the terms are logically defined. Typical ontology term in HPO, it would be defined logically with reference to anatomy ontology. NCIt offers a similar definition.
LS: In NCIt, presumably, two different biopsies share a common term of "biopsy". Does NCIt refer to the same ontology of bodysites. Language mapping, there are other ways of describing soft palate. This automated mapping of the procedure code + bodysite might required hand curation.
PR: Partial yes - colleagues in Monarch having funding in NCI to modernize the thesaurus, map internal concepts to Uberon. Result has been online for a year to two.
MT: Who are we anticipating would receive all of the different terms? Would that be an EHR, academic center?

PR: For practical use case, would be much preferable to have a consortium agree on codes to be used. In GA4GH, vision to use healthcare data for research. Maybe it will come from a hospital, could use SNOMED codes here. If I was doing analysis on the other side, would replace EHR codes with codes designed for more research-y things.
MT: mCODE goes through this with disease. Have certain SNOMED codes. We try to have a tent strategy, so if you're going to put in cancer condition code, won't stop you from having pre-coordinated codes as long as there's consistency. When you have something like punch biopsy, anticipation that community provider systems are expected to have intelligent terminology server, that may be possible, it's the binding and the data store that's the issue, not the logic needed to identify those terms.
PR: So in essence, the same format could be used either to take everything and do your best to map ontologies as described or it could be used as a consortium that would have agreed on the ini file that would specify what is required and what ontologies and terms are to be used.
MT: If you keep it loose, there's trade-offs, depends on who adopts. If this file is more restrictive on ontologies, that might make sense, it might be hard to get buy-in. In FHIR, if you store it, you should send it, if it's received, you should process it.
PR: One of the possible differences in scope and intention, phenopacket is meant to be a model for computation and analysis.
Not currently in Phenopackets, and a well-accepted standard in cancer (Karnovsky). Should this be put into a phenopacket?
LS: Standard in North America, not sure how well it's used in Europe or Asia.
MT: Would add that there are standards of practice, NCCN, that have certain guidelines of treatments options based on performance practices. If there are use cases where you might want to compare, makes sense to go ahead.
PR: Other option would be to have a class like message performancescale, make it oneof.
LS: I like this flexibility better.
PR: The one drawback is that we could wind up with hundreds of enumerations of various scales that might be good, or it might be regarded as bad. Hearing consensus that having one performance scale is not good. mCODE is ecog and lansky. Prototype three enumerations with a oneof element. This would make sense if we are wiling to extend list of enumerations and keep up to date. Decision has to how much the standard will grow with time.
AS: Are we including clinical trials? Recist criteria.
PR: in mCODE there is a qualitative element that is the subjective judgement of the physician about patient improving or getting worse. There is a field in phenopackets that can accept recist criteria, which are NCIt codes. To Do: find more info re: Recist.
MT: Used in certain points of treatment.
PR: Multiple ways of modelling that. Is this something that we should have over a course of time? There is a field we can add, that would be the Recist judgement at time stamp of the phenopacket.
LS: Is something that would be associated with time, both clinically or in a trial. Patient unlikely to know recist status but physician will.
AS: Won't find in EHR. Mostly for patient trials.

MT: I know EHRs that capture it but don't force to put it in. Could identify effectiveness of treatments. Not done for every patient. One of the reasons we struggle with mCODE, has status that doesn't match recist. General well-being status.
PR: For phenopackets - could be more general than mCODE - provides codes for mCODE status levels and people could add that to the same slot as recist criteria. Essentially have a list. Depending on context.
LS: ARGO data dictionary has a response to therapy field. Enumeration of codes taken from recist. There are reps from North America, Europe, Asia.
MT: Need a LOINC code for response to treatment. Submitted to LOINC. Hopefully can have something generated that is more international.
PR: What I would like to for this group is go over the resources with the goal of phenopackets being able to support any of these resources. Be able to discuss a final first proposal of the elements and, in some cases, the value sets in mCODE would be enumerations or ontology terms that you could specify. Will do this exercise, my default will be to work with Jules to make a document everyone can look over, please send me an email.
LS: Very grateful for you for taking on this task.

Meet every two weeks - maybe another three meetings like this. Have doc to present to larger GA4GH group. Put on readthedocs page.

20200527 - Kick-off Meeting

Attendees: Peter Robinson, Jules Jacobsen, Melissa Haendel, Lincoln Stein, Juergen Eils, Olivier Elemento , Andrea Sboner, May Terry, Lindsay Smith, Grant Wood

Conclusion:
We ended with a proposal to think about mirroring what mCODE is doing, such that it would be possible to get a cancer phenopacket from FHIR message that is coming from an mCODE implementation.

	Actions Arising	Deadline
1.	Schedule another meeting in ~ 2 weeks - see here for a Doodle poll. Please feel free to share the poll with any interested colleagues.	ASAP
2.	Jules and Peter to prepare a demo for the next meeting	
3.		

Global Alliance
for Genomics & Health

Meeting Notes:

Representation of Phenopackets for cancer can be improved.
Answer questions, introduce ins and outs of phenopacket format. Think about gathering requirements to create a format for the needs here.

Recap vision of phenopackets
PR: We have vcf files for variants but we don't have a format for transmitting phenotypes. Often get sent word or excel files. Many people publish phenotype/genotype reports, and this information is only available in journal articles. Discussing with journals, who are open to asking authors and possibly requiring authors to submit this information as a phenopacket. In genetic labs, often did not get phenotypic information, often just the disease. A form for allowing transmission of phenotype information would help with phenotype driven diagnostics. Also talking with groups in Canada about possibly transmitting phenopackets via a FHIR version from clinics to labs. Also a computational model. Might be interesting to have standard computational representation. Phenopackets would like to work with the community.
MH: Goal of phenopackets is to keep in mind that is the lightest weight, freely sharable nugget for sharing phenotype information. The first release was largely developed by Jules, and focused on that. In cancer data modelling, there are a lot of different models that don't align so well. What we're not trying to do is try to solve all of the problems, but what would be the most useful nuggets to share at a case level. Temporality representation is critical. Also want to make entities as openly useful as possible, don't want to constrain terminologies used.

LS: How do phenopackets and ontologies interact?
PR: Ontologies are basically a collection of terms or concepts connected by sub-class interactions, but doesn't tell you how to package up a collection of terms for transmission or computation. Phenopackets is a framework upon which to put ontology terms. Might be interesting for phenopackets to have recommendations on terms to use, so everyone uses the same collection.
LS: Have certain packets like evidence, which have an OntologyClass. Can you constrain a field in phenopackets to be a particular ontology or to be several different? How to do this? This becomes an issue in tumor staging, where there are multiple tumor staging systems, and would like to specify which system you're using and then constrain the values to be values that are valid within that system.
PR: In a simplified world, yes we could constrain. But people will not agree on the ontologies. Possibly EUCANcan might agree, so we could constrain EUCANcan to use phenopackets plus terms from an agreed upon set of ontologies.
LS: Would like to be able to say, you can put anything you like in that field, but you have to be able to point us to a ontology that describes what that term means.
PR: Phenopackets have an element called metadata, you can do a quality check of phenopackets to make sure the ontology terms match to the ontology mentioned in the

metadata. This doesn't happen directly from the protobuf quality control but we envision outside software doing more quality checks.

Introductions

MT: mCODE has a need for further research use cases, bring in a cancer interoperability perspective. Have synergy where possible.
LS: Head of AO at OICR. Lead the data coordinating centre for international cancer genomes consortium (ICGC). Part of EUCANcan project. Co-chair of data coordination and clinical management group for the Terry Fox Cancer Network.
JE: From EUCANcan, driver project of GA4GH. Also part of German ICGC activities. Would like to bring together international efforts, time to harmonize datasets. Fan of mCODE, used with many groups.
OE: Director of institute for precision medicine at Cornell. Implemented clinical genomics program at Cornell. Very interested in delivering precision medicine to patients here, merge genomic and clinical information.
AS: Director informatics and computational biology at Cornell. Helped infrastructure for clinical genomics.
SM: Medical director of clinical informatic at Cornell. Spend most time of EHR aspects of clinical decision support, terminologies, integration of genomic data. Personal interest in FHIR genomics. Cornell was part of the Sync4Genes pilot.
GW: Intermountain Healthcare, involved in family health history work.
JJ: Work at QMUL in London. Work with Peter and Melissa as part of the Monarch Initiative, Peter and I were the primary developers of phenopackets as it stands. Also the primary developer of Exomiser.

Discussion

PR: Several discussion points: is this useful for the cancer community? Is there a format similar enough in intent for phenpackets, should we use that? What do we need to do to satisfy the needs for the cancer community?

LS: There are different use cases: research cohorts - in my wheelhouse. There are numerous projects around the world collecting imaging and proteomic data, associating those with phenotypic information. Optimize treatment for patients. In this context, it's really useful to be able to combine smaller cohorts into larger ones. Combining phenotypic data is a wild west - there are a number of formats that can be used, primary one is mCODE - they've put tremendous effort into a high-quality and complete schema to describe phenotypic data for cancer. mCODE doesn't do everything that all of the projects I'm involved in need. Other use case is primary point of care, which mCODE focuses on. Missing things like point in time. Alternative to phenopackets would be mCODE. Within ICGC-ARGO, we have a well-established data dictionary and schema that works well in our hands but it's project specific. Resource to draw on but not a general alternative.

PR: What does mCODE think about this? Can we collaborate closely? Two end of spectrum: phenopackets extended to include everything that all groups need, other would be that GA4GH adopts mCODE as community standard.
MT: We are standard for trial use. With regards to collaborations, mCODE is part of a greater effort in a FHIR accelerator called CodeX. I would encourage that there are other people who run that, and to discuss with them. Could mean that phenopackets becomes a lot more influencing initiative, could have a greater role in the CodeX community. Can connect you with the right people. Greater committee that determines what the elements are.
PR: How to download, what is the native format?
MT: If you download the full spec as a zip, there are html instructions. Comformant with FHIR.
PR: This is essentially a FHIR profile
MT: Yes, JSON definitions are FHIR profiles
PR: The relationship now of phenopackets to FHIR, if you look are FHIR there are a lot of stuff in FHIR message you might not want to have to research. A pheno packet is substantially simpler than what you would see in a FHIR message. We have a contract with NLM to extract rare disease phenpackets from FHIR systems, being done at CHOP with GRIN. Prototypes are working. Wondering if a similar parallel relationship such that a cancer version of a phenopacket is a distillation of a FHIR message with the intent of something to be immediately used in analysis.
MT: Models and standards are fit for purpose. In the case of FHIR that's true, initial origins are provider oriented. There is an evolution in HL7 to evolve towards research, but you're right it's not lightweight.
PR: Computing power won't be the limiting factor. Also depends on where FHIR goes. A lot of existing data will be in OMOP or i2b2 and those people probably won't want to turn that into FHIR.
JE: Set up a platform for all uni hospitals, and changing our data within 4 weeks. All working with OMOP and i2b2. All are now using FHIR, converted the data. Not too complicated. Bioinformaticians who did it. Researchers working on FHIR.
PR: in the US, Melissa leading an effort to unite 50 hospitals, settled on OMOP. Distinction between medical informatics and not medical informatics.
JE: i2b2 is also working the FHIR in Germany.
PR: People from bioinformatics community have not seen FHIR, a lot of stuff there you would not want to use for research. We don't want to replace FHIR.

LS: For research projects, would have to heavily redact FHIR information, don't want to gather more PHI than we need. Only approved research fields are pulled out. Might as well put into more easily interpreted format.
PR: Phenopackets doesn't have the fields for PHI.
MT: FHIR doesn't need to take over research, definite benefits to OMOP and i2b2. There has been a reverse ecosystem of the formats to come in. If the intent is to make more use of provider-based data, then that is regulatory. FHIR is not going away anytime soon, need to better combine research to provider based data. Open discussion and collab.

OE: How do you envision phenopackets being used - files in the phenopackets format and exchange those files?
MT: Last week I did a comparison of the phenopackets FHIR IG, probably need to have a better feel.
PR: Core GA4GH use case of phenopackets would be when someone is studying a certain type of cancer and the spectra from other hospitals. Extracting cohorts from hospitals and putting them together. GA4GH wants to be closer to the genomics, and enable a day where you can look at genomic results from a hospital, and do this in multiple hospitals. Emerge an ecosystem of software to enable this.
MT: From an integration perspective, given the depth of research data needed for clinical genomics, goes back to notion of genomic archive communication server and translation layer that would take pieces of it - depends on direction you're going to. If it's outside-in, research to provider then it becomes a translation to take what is relevant from that study and have to do a mapping of those key elements to make correlations. I can see it being a mapping exercise, even within HL7.
PR: Might be interested for you to talk to OMOP group. Related to GA4GH projects to gather cancer knowledge bases.
MT: Back in November, gave a presentation on mCODE to OMOP ontology subgroup. Response is in line with what you're saying, the needs of research are things like deidentification. We want the evolving and greater participation. Trying to find that better alignment with research needs. Encourage folks here to be involved in that working group.

PR: Do you have competitors?
LS: Competitor in a collegial way, other standards to be considered?
MT: As far as it relates to FHIR as an implementation, no. There are other standards out there, like OMOP. Some more proprietary ones like FlatIron that make a lot of their business through sharing their own models.
PR: If you look at TCGA, ad hoc list of items that get collected, but there's no one standard schema collected. mCODE is moving into that space from a FHIR perspective.
MT: Could say that within FHIR, that's where we found the need. For CCDA, another HL7 product, are specific to registry reporting needs. Even the CDC is working with the oncology community because they recognize that CCDA has limitations. It's more about who's being involved.
PR: We should be thinking about mirroring what mCODE is doing, such that it would be possible to get a cancer phenopacket from FHIR messaging coming from an mCODE implementation.
MT: Outbound-in, worthy exercise might be structural and semantic mapping.
PR: Great to continue this discussion about possible elements from phenopackets that could be added into mCODE and vice versa.
OE: Idea of every hospital having a phenotype file makes a lot of sense. Concerns about mCODE, doesn't seem open source type of initiative.
MT: In terms of open source, it's public.

JJ: Are we going to mirror mCODE?
PR: Not sure complete consensus, but that's my opinion. We need something to explore. Why don't we touch base and prepare a demonstration.