HCA Metadata Call

Every 4 mondays, 8am PST 11am EST 4pm GMT 5pm CET 1am JST BlueJeans

---

This meeting is recorded. You can find the recording for each meeting in this google drive folder Call Videos

If you are attending this call and would like to sign up for the metadata-community google group please fill out this form https://forms.gle/mwcKuD5WDR17aea69

November 16th 2020

Agenda theme this week: ASCT+B

Announcements

Topic ideas

Ray - If going forward you have ideas for these meetings please, email the wrangler email address wrangler-team@data.humancellatlas.org

Agenda Items

Katy Börner - "Anatomical Structures, Cell Types, and Biomarkers (ASCT+B) Tables: Design and Usage"

PDF of slides

Some meaningful links related to the discussion today.

https://hubmapconsortium.github.io/ccf/

https://hubmapconsortium.github.io/ccf-asct-reporter/

KB - This is a team effort including 300+ people in the wider HubMAP initiative.

Information comes from many labs, many individuals etc. So data needs to be harmonised. That is being lead by Rahul Satija.

CCF consists of ontologies and reference object libraries. It needs to be able to work with multiple tissues and we need to be able to match the samples to the coordinate framework.

CCF portal contains 3D Object Library, RUI, EUI, all linking to the CCF Ontology in building. Visible Human MOOC is a nice training to use with your families etc.

CCF Requirements: We must capture anatomical structures, cell types and biomarkers and multiple levels of resolution. It should be semantically explicit (ontologies like Uberon etc)

The tables are quite extensive and there needs to be lots of expertise as well. There are people working on trying to compile these tables. We are now focused on 10 tables, 1 per organ. AS we go updating the numbers in these tables go up and down. In order to review the tables we have reporter interface. In red anatomic structures, in blue cell types and in green biomarkers. By checking on each of them you can find the linkage.

For the lung for instance it might be important to capture smoking status. I want your help on trying to understand what needs to be captured for all organs. And it would also be nice to try and understand what metadata is specific for a particular organ. You can go into these tables and provide input.

For vasculature based CCF macro and micro scale detail of vasculature is important. The tables for vasculature, lymphatic system is a bit particular, more branched.

We now have a working group lead by me and Jim Gee. Ellen M Quardokus serves as Knowledge Manager. Next Meeting on December 3rd.

ASCT+B Table Usage: They guide the ontologies. To develop the CCF all 3 types of ontologies are needed. Cell types also need to link to the tables. For 3D Object Library, we offered the labs generated the samples to register 48 blocks of samples.

hubmapconsortium/Github/CCF-3d-reference-library has available free 3D Reference Libraries

We have 15 different organs at the moment in HubMAP.

Sarah Teichman’s group for instance used 5 extraction sites.

CF Registration User Interface (RUI) allows people to register samples, and assist them to extraction sites. It is quite like a game. You can add new anatomical sides. You can even delete them.

HuBMAP Upload Portal: When extraction sites upload their data

You can zoom in and see all the tissues that were taken on a particular organ.

All tissue mapping centres are asked to collaborate with us on ASCT+B tables, to register all tissue samples using CCG RUI, review after submission in the CCF EUI. We ask all centres to identify FTU (Functional Tissue units). We are very interested in vasculature-based CCF, cell segmentation.

I hope you saw HubMAP MOOC.

Hacking the kidney Hackathon - We will all benefit from universal tissue segmentation.

Q1: When you mentioned the vasculature system, is that because it is a more complex tissue to map to an average?
A: We have quite a lot of experience with visualization etc. We are at the very beginning of the vasculature. We are open to collaborations.

Q: When you are determining the biomarkers, I guess you go to experts on the field, you are asking them what biomarkers that define a landmark etc. Do they all agree? How do you reconcile?
A: There is little agreement. There is not even a definition of what a cell type is or even what a cell state is. For biomarkers even more. There are new technologies being developed. These will change over time as we learn about robustness. What we can do on the technology site, we can provide technology so users see the gaps. The reporter has all the organs. You can see what biomarkers are being used for all mucosal tissues in all organs etc. It can help see what biomarkers are most scalable. I hope this will help find the gaps and see if we can bring the experts to define how to fill the gaps. People like David Osumi have gone through ontologies and bring to the tables. Efforts like the Kidney are working closely with curators.

Q: Is metadata captured for biomarker entries, such as the assay or experiment used, as well as to what it might be in reference (i.e. gene A is a biomarker of alveolar macrophages when looking at a population of lung macrophages)
A: We are trying to break it. We want to try to come to structures that serve as rosetta structures, so that experts can work on that and the tables serve as translators for the different levels of expertise. I would like to come to the expertise of your team as well to what types of metadata fields come be important for every organ. In HubMAP we are adding lots of different technologies. Most of the flows have links to Protocols.io as well. Ultimately to be able to interpret, you have to take the technology into account as well. It helps to have at least part of technology in the tables. Ultimately you want to know what technology you used.
Q: When you are using at Gene ontology and all the annotations that have been made to that. If you are trying to define a biomarker for a type of cell through scRNAseq, you are defining it by comparison to the rest of the clusters.
A: The experts started to work on this in March and managed to get a start with 10 tables. In the future the next step probably is where did they use as a protocol, and be able to connect

and link. We are working with ontology experts but also with UI experts to add all the complexity that all the bench and clinic experts find without the complexity.

Q: I can see a need for the software allowing a deeper dive on biomarkers. For example, while CD31 is a good generic biomarker of endothelial cells, other biomarkers are needed to subdivide endothelial cells into lymphatic endothelial cells, venule endothelial cells, capillaries, etc. There should be a way to further subset cell types based on a richer biomarker list.
A: We would like to have biomarkers that are more and more specific. That's also coming to the point of what biomarkers can be used on the different organs. Cause we don't currently have that.

Marion - Depending on where the information comes from and if they were interested in something in particular for their study we might or might not get the smoking status. If you go to a published dataset, you might not find it. We don't have a good understanding of this either. I don't know if there are any other organ experts that want to talk about what they believe is essential for their organs.

General: Sex, age, ethnic origin, height, weight, girth, BMI Pregnant, menstrual cycle Organ Specific: Heart (8/24/2020) - hypertension, diabetes, cancer, pulmonary disease, liver disease, echocardiography (LVEF %) Lung – smoking Skin – sun exposure

Certain types of diets also change things.

I would love to connect to anything that you see in trying to agree on all organ fields and organ specific fields.

for reference: https://www.gutcellatlas.helmsleytrust.org/

Q: You express interest in environmental factor, but where would you draw the line to establish the average, the healthy state vs disease state?
A: We need to capture metadata that are derived from for instance pregnancy. It is still normal but it changes a lot of things. I think you all already agreed on sex, age, etc. I think we should add pregnancy and even menstrual cycle.

Marion - i think we under-collect 'environmental' type metadata when we know how much impact it hs over life. 'we' being biology researchers in general.

Metadata per Organ
General:
Sex, age, ethnic origin, height, weight, girth, BMI
Pregnant, menstrual cycle

Organ Specific:

Heart (8/24/2020) - hypertension, diabetes, cancer, pulmonary disease, liver disease, echocardiography (LVEF %)
Lung – smoking
Skin – sun exposure

August 24th 2020

Agenda theme this week: The Human Heart Atlas

Announcements

Topic ideas

Agenda Items

Henrike Maatz - The Human Heart Atlas

The Adult Human Heart Atlas - Overview of metadata that is collected

It is a complex organ. We knew that the different regions were made up of different cell types but we did not know the extent and the physiological differences of the cell types.

The heart team emcompasses people from many institutions.

We got samples from CBTM and also from Mazankowski Heart Institute.

Sex, age, ethnic origin, primary diagnosis, cause of death, heigh, weight, girth, BMI, hypertension, diabetes, cancer, pulmonary disease, liver disease, echocardiography (LVEF %) are some of the main metadata fields we collect.

We have a broad range of samples from different ages but it would be needed to have more individuals from each age range.

On the ethnic origin we have a very homogeneous range of data. We would need to expand.

We used hematoxylin and eosin staining of hearts. All our donors had medication to keep alife. A heart might look different depending on age as well.

Different datasets we checked them for quality and duplets.

Cardiomyocytes are too big to detect unless we do single nuclei.

Having different data sources we get complimentary views of the heart.

You can see cardiomyocyte types when you enlarge. Some types are more common in left or right ventricle or atrium.

We managed to identify a wrongfully labelled sample thanks to the cardiomyocyte subtypes that are specific to a particular site of the heart. The gene expression profiles of cardiomyocytes are very useful metadata.

HM - Single nuclei has the advantage that you can freeze the tissue.
KR - I see your primary diagnosis. This might change overtime. Do you usually investigate or do follow ups?
HM - It's one point in time that we record.
HE - How do you obtain the consent? Cause of death would be metadata that you would include?
HM - This came from the heart partner in Canada.
CES - In North America next of kin can give consent. At the sectioning stage that was when the labelling error happened. We see a signature of the second heart field being very enriched and that makes sense from a developmental perspective.
HE - On annotations, what happens if something has been annotated as being obtained from a particular source and it's not true. Who is responsible for making it right?
LC - I would say that if we are made aware that after analysis something does not seem right. At EBI people added metadata to BioSamples, we can tag it with curated by flag. I would be intrigued Henrike to know how you handled the error sample.
HM - We removed the library from the dataset. We did not feel comfortable with just switching the label of the sample.
CES - This is a broader question. The bigger numbers will show us the outliers. I believe we need a large number of samples.
Michela - We need a bigger amount of donors and more organ representation. But also we need lots of metadata. We need to bring back the classic histopathology and put the two things together.
LC - Are there samples were you can do that combination?
Michela - We have some samples but we have to compare more. Some patients are older so you might have expected normal features for someone that age.
LC - Heart like brain are organs that are more difficult to get. Hearts tend to go to transplant before they go to science.
CES - We can do biopsy and follow up (clinical). Doing a second biopsy is usually not possible. Some of the hearts we get were perfectly fine but they did not comply either distance, recipient
HE - You said some hearts not the whole heart could be transplanted, and you got the other part of the heart. Was there any valvular part left?
Michela - Maybe a future step would be to make the most and characterize the small relevant areas.
MH - Have you come across a case where you had two nuclei that were so similar that you think they came from the same cardiomyocyte?
HM - No.
HE - In the nuclei sequencing did you do hashing on them? Or you did one region at a time?

HM - We did one region at a time. Isolating cardiomyocytes is already quite challenging as they are very sensitive to enzyme digestion. Samples tended to come in the middle of the night.
CES - Pathologists can possibly comment on this. Apparently cardiomyocytes contain lots of interstitial tissue around so to separate them is quite challenging.
HM - For single cell sequencing is not possible at the moment to separate individual cells easily.

June 29th, 2020

Agenda theme this week: Developmental stage metadata

Announcements

Agenda Items

Steven Lisgo - HDBR - Review of staging and terminology associated with the stages for prenatal samples

PDF of slides
https://drive.google.com/file/d/1oxoxZ2Td8ExIIb4meCc0duhwPpuQS7wK/view?usp=sharing

Using a classification system like the CS makes sense. Counting the number of somites is very useful but we can't do that for all the samples. We also look at some morphological landmarks and how they should look like at a particular CS stage.

For instance, in the image in the screen we counted the somites, in this case, and this places the sample in CS10. If we look at a particular landmark that also corroborates CS10.

After 8 weeks most things we look at have developed already. There is the fetus growing. Using external landmarks before 8th week is very useful.

CS12 lower limbs start to appear. In the CS13 upper limbs start to form. CS14 we can also start to see developing hands. CS16 you can start to see development of the eyes, and CS17 the fingers start to develop. CS19 toes start to form. CS20 and CS21 eyes. CS22 there start to be eye lids.

This is all for the early age samples.

For post-8 week samples, we use certain measures: foot length (toe to heel). We also use hip to heel. Based on the measurements we define the post-conceptional age.

Laura - These seem like very useful guidelines. Have you published them?

Steve - The reference is on the slides. And you can also find them in the website.

We record the time of the last menstrual period. But we don't find that extremely useful. We find the estimation better when the ultrasound takes place. But we take the measurement over the samples so that we can have more consistency.

Question - Do you take images?

A - Whenever we have samples, we take pictures and describe why we staged a particular sample. If they have been sectioned, and histology is done, etc we would expand the information and record it in our database. You need to notice that all organs develop differently at their own pace.

Heather Etchevers (France) - Happy to hear that you keep doing the same protocol. We are also still doing the same. We wanted to have evidence of what was done. Link the evidence with all the information and have a database, also with images. How have you evolved to do this?

Steve - We have images in a server. Last 10 years we have been using a database that was built specifically for HDBR. It has information about the samples but also about the projects that use the HDBR samples and the requirements that they have. The database would tell whoever is collecting the sample what projects are using the samples etc. That's essential. We have lots of different research projects that use our samples. But fortunately we had something written specifically for us.

Laura - Any more questions? Or about staging prenatal samples?

HE - Is there anyone here from Sweden or any other group?

Paolo Giaconini (France) - I am from the same consortium as Heather and we are doing exactly the same thing. We are struggling with the space for images. We are facing a bigger problem with 3D images. But we are pretty much aligned in the staging.

Laura - Maybe Enrique could present the findings from the survey and then we can talk about how to capture this.

Review of development stage metadata survey - Enrique and Gabby from HCA metadata team
Google slide deck

https://docs.google.com/presentation/d/1VxBMKEvbl7uFG8R1_nbPC1-snu-mtt2q5RzN6-HiZlg/edit?usp=sharing

We have a couple of projects with developmental samples to give us an idea. But we don't have much to consolidate our knowledge.

We realised we needed more information so that we could have more confidence on the terms we use.

In the survey we asked 3 main questions.

We had 28 respondents to the survey from very diverse consortia. Most participants are from European countries or consortia. We might be able to reduce bias in terminology if we reach out to other countries.

Top insights from the survey were that Post-Conception weeks and Carnegie stages were the most remarkable, and were rated as essential or most important. CRL was also mentioned as important.

On the measurement side Number of somites, digit formation and others were also mentioned.

It seems like the participants also welcomed terms that they could transform to their preferred term.

We also noticed that participants who choose other extra terms were working on multiple projects and organs.

Most of the participants were not familiar with ontologies or controlled vocabulary but one participant was familiar with ontologies/Uberon.

We have some recommendations that we have come up with following the results of the survey.

- Currently the HCA metadata schema allows to enter the PCW in the gestational age field. We should make it clearer that we refer to
- Maybe enter as an optional field - how PCW was calculated
-

Steve - Are you planning on using PCW also for embryonic stage?

Enrique - Do you mean that CS has more granularity in these early stages? We do capture the different CS. But sometimes we did not have the right metadata. Would the PCW be helpful if there was no CS?

Steve - It would be helpful but not as helpful.

HE - You would not be able to compare to other tissues. Before 5 weeks we would use somite number for more granularity.

Laura - It's worth mentioning that we are not proposing to remove CS. We have had feedback from some contributors that they gave us the CS because we requested it, by someone looking at the sample and estimating it. Analysts might want a value so that there is a unifying value for all stages. So that there is a single consistent measurement for all embryo stages.

PG - So you're proposing that you would have the PCW and in brackets the CS?

Laura - For some people we would have all the information.

HE - So you will have the PCW as a requirement and then say that is highly encouraged to send more detail.

Marion - Sometimes we get more granular information in days. CS usually has a defined day range. But do you need to establish the CS by looking at the sample?

SL - I would be very careful about how people estimated this. And I don't know how accurate you can be. Usually [if you specified by PCW] these correspond to multiple CS. Usually you would [want to] compare samples among each other. I would recommend anyone working with samples less than 8 weeks to use CS.

HE - I would also advise people to check the information and tables in the HDBR website to define the morphological criteria to help determine the CS.

Marion - Is the length a proxy for the CS or not necessarily? Or all those other things you mentioned before.

SL - Usually the samples would have been processed in a sample bank right?

PG - Here we have the samples in Lille and then we distribute them in France. But at the moment there are ethical rules that delay access. At the moment we are trying to centralise the embryo collection and processing in Lille and some other places. At the moment, without specific authorization we can't send samples to other countries. There needs to be a ministerial authorization. If it's biological material there needs to be one.

HE - For libraries we don't need the authorization.
PG - Same for us. We are trying to change this [the need for a ministerial authorization for specific projects collaborating with consortium partners]..

HE - Histological slides are also under this regulation.

LC - I am happy to bring this issue to Tracey, on the HCA side to take it to leadership. I guess different countries have different regulations.

PG - It would be great if you could help with the conversation.

MS - When we get metadata from the researchers, sometimes they give us the metadata that they think is important or they use in their analysis. We don't necessarily know all the information that the tissue bank has. What should we ask them to make sure they give us all? Cause they don't give us everything. Sometimes we have a day age of an embryo or CRL. If we know that it should have been staged we can go back and ask, if that's what the community needs.

HE - The samples sometimes are destroyed after staging them. That's why it is so important to take pictures. So asking them what made them establish it was normal or a specific stage. We can stick with the weeks. And then adding the CS if we are sure of that. If we have the photos a trained eye would know. I am really keen on the photograph as evidence.

LC - We are coming to the end. We should gather feedback from this call and the survey and come back with a proposal. We could do another review to see that we don't miss anything. Another thing would be to have a work session with institutes doing regular sample collection and see how we could capture everything on EBI Biosamples and Bioimaging archives. Consents would probably be something to evaluate.

We will come with a proposal.

Gary Bader - Are you guys familiar with the evidence code framework? Making clear that PCW is evidence rather than something calculated.

LC - Do you have an example of ontology?

GB - I can send them to you later, Laura. Gene Ontology is the biggest user. I know that pregnancy weeks [weeks of amenorrhea] are different from embryological weeks by about two weeks. Do we know when you are getting PCW that everyone is defining it the same way? Hopefully these metadata terms have good definitions that make it clear.

HE - I think everyone understands PCW the same way because of the inclusion of "post-conceptional". Maybe not with other terms like "gestational"..

GB - If this was communicated to the community they might be encouraged to collect the information. It will ensure better data for the analysts. Just as a comment. Also it may be possible to establish **new staging benchmarks** with future scRNAseq data.

June 1st, 2020

Agenda theme this week: Analysis of Data

Announcements

Agenda Items

“Putative cell type discovery from single-cell gene expression data" - Zhichao Miao (EMBL-EBI,Sanger)

https://www.nature.com/articles/s41592-020-0825-9#Sec8

Nowadays cell types are often defined by expression of a number of feature genes.

Before it used to be definitions based on morphology and phenotype.

scRNA-seq is a high throughput technique.

The Problem : Cell Type Annotation. Many times we get overclustering or sub clusteringThe a. Only those with enough knowledge will be able to determine the cell types.

Single-Cell Clustering Assessment Framework (SCCAF)

SCCAF algorithm to identify cell types

Historical background: cell types.
Rober Hooke (1665)

Cell type classification today: feature genes
Markers, surface markers.

related terms, synonyms:
cell state, putative cell type, biologically meaningful cell group

2D cell expression matrix.

Unsupervised clustering. Check marker genes.

The algorithm has two parts, sel-projection approach and machine learning.

After the self-projection approach we put the model to test.

Some cells were confused by the machine learning model. Confusion tends to happen between clusters of the same cell type, but not between different cell types.

If one particular cluster is overclustered in two, that's when the model will encounter issues.

RNA expression profile to define each group of cells

A self-projection approach
* training
* test
* feed back test results

Detection of over-clustering

An automatic approach

We can use key feature genes in the machine learning model. These genes would define certain cell types.

With real data, we achieved same clustering achieved in the original paper. Over 4 rounds we did optimisation of the model

Machine learning and self-projection

Example: mouse retina dataset (Shekhar et al 2016)

Four rounds of optimization. After those rounds we could clearly define different cell types.

If you have overclustering, machine learning can identify it. But under clustering can be more complex.

There are some datasets that are being used as reference datasets, and using them in machine learning. Adding other datasets later can be sometimes challenging.

If we compare our method with reference-dataset method we achieve results more similar to the published results.

Applications:

Define cell states in continuous
Validation in another dataset

Define subpopulations
Zoom in the data (and define cells in the centre of a clustering)

Unannotated human brain data:
There are lots of different cell types. There is not good annotation. We identified good gene markers.

Benchmarking:

SCCAF restores 'ground truths'

L1, L2 seems best machine learning method to use

'ground truths' achieves better results than under-clustering or over-clustering

Define cell states in continuous process

Disease states can be detected by SCCAF

SCCA can define cell sub-populations

Reference based method vs. SCCAF: unannotated human brain data

confusion matrix

Laura - How did you obtain the metadata about the cell type annotation?

Chichau - Sometimes we would get it from GEO or AE. But sometimes we need to go to papers. Sometimes in the supplemental information. I spent quite a lot of time getting the cell type annotation.

Laura - If we were collecting the cell type annotation, do you have requirements that would make it easier to re-use? Is there something we can do that would make it easier?

Chichau - For cell type annotation is quite complicated. Sometimes we don't know how the algorithm can work on the annotations. Normalisation of data, batch effect, there are some fields that would be nice to have.

Marion - The cell types that you found when you tried to benchmark your method. How did you find the consistency in the metadata about the cell type annotation? Did you find lots of inconsistencies?

Chichau - I found the same annotation with typos. You need an assessment to decide if the annotation is correct. In terms of solving it, we need a cell type ontology. Especially with more or more data coming.

Marion - We have a few working groups trying to improve the ontologies.

Marion - With your algorithm would you be able to discover a brand new cell type? Is that an advantage?

Chichau - Yes. We can in a reference dataset define clearly defined cell type. I would recommend using both in parallel. The algorithm would find meaningful cell types

May 4th, 2020

Agenda theme this week: Summary of last 6 months

Announcements

Agenda Items

Metadata Community calls: 6 month review - Marion Shadbolt & Zina Perova (EMBL-EBI)

https://docs.google.com/presentation/d/1fSRwbWfZX0BRAmUopSp3vzCbB9qxBF8yeWM452SURYY/edit#slide=id.p1

The slides were put together with Marion and Laura.

I want to thank all the presenters we have had so far.

I will give a summary of what we have learned over the course of the last 6 months.

These calls emerged from a need to have a standard that was highly structured, agile, flexible, versioned and self-describing.

This diagram represents the 3 pillars. We have worked on this. The idea of the community calls were to be able to present and understand how to evolve the standard.

Emerging challenges:

- Which metadata fields are valuable?
- How do we capture spatial information?

- How do we harmonize cross consortium?

Valuable fields - Fields we have (location of sampling, smoking status, embryonic/fetal stage, sequencing assay type. See slides for more detail on needed information, existing attributes and new additions.

Valuable fields - fields we are missing (batch information, cell-type, cell state, sequencing (genome assembly, sequencing depth, etc not present in the metadata model at the moment…). See details in slides

Capturing Spatial information: How to capture spatial information has also been up for discussion on these calls. It would be ideal to be able to work with the community on establishing the linking in between anatomies (example lung)

HubMAP has the CCF as a key focus.

In terms of **cross consortia harmonization**, this is a slide that shows that the information that we collect is the same but we collect it differently. For the harmonization it would be critical that all of us record the information in a more comparable way.

If we made cross-consortia harmonization it would be easier for the researchers to use and compare the data. And we could minimise the effort.

For the cross consortia harmonization there is a regular call. See slides for call details and email list.

Ontology Improvements:

We have come across multiple cases during the calls and our work with contributors, when someone wanted to use an ontology term they would try and find it and either find that it was not accurate anymore or there wasn't one.

In most cases, when we talk with contributors, they would tell us about a lack of a term and we would talk to the ontologies. However, it is also possible for any researcher to go into the Github ontologies repo and add a ticket to add a new term. They could also undergo training for this.

We can organise ontology training workshops, so that the community can learn how to add new terms.

We are going to transition into having videos of these calls on YouTube.

Ideas For Action:

- Reach out to more computational scientists and analysts - Volunteer to present or spread the word

- Gather requirements for metadata fields from community - Tell us your ideas on how to collect these!
- Organize Ontology workshops (GMT time and PDT time) - Help organize and/or participate
- Provide contact details for other initiatives and Working Groups - Sign up!

Katy Bourner - Can you see to a game to see how to validate ontology terms? Have you seen interfaces in an easy way ontology terms? We need to have experts to provide understanding of organs, but if we could also have game style learning ontology interfaces. At the Allen institute they have interfaces for neurons. I haven't seen any for ontologies and metadata.

Zina - I haven't seen any but I think it is a great idea. There is a protein localization game. It was a massive effort, but it paid off. It is a great idea but it would require a lot of effort to make something like this. Can be helpful and engaging. If it's fun apart from helpful, it might make people more engaged.

https://www.ccpgames.com/news/2018/ai-and-eve-online-community-improve-cell-and-protein-mapping-in-the-human

KB - If someone knows a game like this, could you, please, add the link to the chat in the call? Is there any game I could play with to build on the CCF? It would require biomedical researchers, not only developers.

Laura - You mention that someone sent you a publication about ontology mapping.

KB - This was not done by experts.

https://www.sciencedirect.com/science/article/pii/S1532046416000277?via%3Dihub

KB - Games are good

Laura - Not known ontologies are perfect fit for the atlas effort. That's why we developed an ontology application. Being able to use the right ontology at each time would be best. We have at EBI different applications.

Laura - It was interesting from the presentation what might be interesting down the line. People analysing data it would be nice to have your understanding on what you might need that is missing. If people can volunteer about their metadata challenges, how they want the information available, in what format, that would be great

Zina - Anyone in the call from the analyst perspective?

Tim - Asking what is the use case would be the way to go. They need QC metrics for instance. If the use case is the analysts doing the analysis, it would probably be more convenient to pull and have the file and metadata together. Just because it's files

Laura - Cell type annotation, and some of the information might be good to have it searchable but the cell types might be needed in the metadata rather than inserted in the file. If we can recruit analysts now we can ask them what formats they need.

Tim - Your team had the flexibility for the scientists to also have their annotations. Probably there will be cell type annotations by consortias as well.

Kathy - I am tempted to have a cell type annotation based on ontologies and maybe later the scientists and the analysts doing a deeper cell type analysis and establishing states.

Laura - For standardization we probably need to interact with teams working on annotation, either manual or automatic.

Tim - I'm hoping great alignment between what you're doing and the cell annotation efforts, where John Marioni is also involved. Multiple cell states could be the same cell type.

Laura - You hit some of these challenges with the annotation curation you did right?

Zina - This curation is challenging. One of the problems, the nature of it, there is no strict definition and it's going to change with time. That's one of the challenges of the curation.

Kathy - I found these meetings really helpful. Researchers coming to talk about their data was really good.

Zina - next call I hope we will have the perspective of the data analyst. If anyone wants to volunteer would be great.

From talking to people who do analysis one question I have is whether they would say the more metadata we have the more we can learn about batch effects. But I would like to learn more about what that really means.

Tim - Batch effect correction is a tricky one. Scientists might have certain scope. They usually represent some technical and biological area of work. The batch effects need to be understood from the start. As you want to compare it to other data, you might be adding more technical batch effect. What events do we support? The events would require batch effect correction. We should do these things when they are required only.

Laura - We are going to supply as much as possible raw data in case a particular use case can't use our processed data. What purposes we are trying to serve will tell what is that we need.

Laura - Any other thoughts?

April 6th, 2020

Agenda theme this week: Imaging Data

Announcements

Agenda Items

Metadata acquisition in smFISH experiments - Simone Codeluppi (Karolinska Institutet)

Video of call:

Link to slides: Metadata_presentation_HCA_2020506

Metadata Acquisition in Cyclic smFISH Experiments

SC - We are still in very primitive stage, collecting as much information as possible. In the lab where I work we use 3 different smFISH

Serial approach

Barcoded approach if we are looking into a large amount of genes. For compensating we would combine the barcoded and serial steps.

Second part are the analysis steps:

Data mining - Collecting the RNA molecule structure

Clustering, atlasing, spatial relationships

We are collecting metadata on all these steps:

- Experimental Metadata
- Sample related metadata
- Data acquisition m
- Data mining m
- Data atlasing m
- Storage m

Experimental Metadata:

- Motivation
- People responsible
- Institutions
- Information regarding experimental design

Sample related metadata:

- Age
- Sex
- Genotype
- Disease phenotype (if any)

Tissue related info

- Anatomical informations
- Time of collection
- Storage
- Shipping condition
- Travelling time

All this information can help understand if a problem is sample or procedure related

Data Acquisition Metadata:

Pre-Imaging:

- Sample cutting
- Protocol sample preparation
- (txt file or forked version of protocol.io)

Imaging:

- Logs acquired from automated fluidic system (matching protocols.io)
- Logs from the microscope during acquisition (size of pixel of the camera, light conditions etc)
- Image specific information

Data Mining Metadata:

In our processing pipeline we organise the data by field of view. Some of the analysis parameters that you use are changing depending on the region that you're analysing. For us it makes it really easy to change the analysis. We cna just edit and relaunch the whole pipeline.

Data Analysis Metadata: Code plus parameters

- Cell type clustering methods and parameters
- Spatial clustering methods and parameters

Storage Metadata (Backup/Cold storage):

We usually safe raw data in a separate cluster

We have all in yaml file.

We are happy to discuss the best way to approach this.

Different people are running the experiments and different people writing the code. I'm just an example.

Discussion points:

- Sort the metadata in searchable vs non-searchable and where to store them (some information can be saved with images and some other information can be indexed)
- Standard for file format for the different type of metadata (ex. yaml/json/csv we are currently using .yaml
- How to integrate analysis code or processing

MS - Any questions?
Deborah Hoshizaki (NIH/NIDDK) - How do you decide of what method to use for clustering?

SC - Depending on what number of genes you need to assess you might go for barcode or serial approach. And then you might need a different clustering method.

LC - How often do you think that people will want to re-run the pipelines? Vs re-run the raw data?
SC - People are not going to re-run everything. If you can run multiple approaches to have the count, probably there would be no need to go back and run the raw data.
LC - I liked your searchable vs non-searchable, when people try to compare different tissues, what you think are going to be the searchable information vs what would be nice to have?
SC - I don't have an answer yet. There are things like info on the pixels (to asses quality of images) might be important.
LC - What are the experiments to do to see what information is important
SC - SpaceTx is going to help understanding many aspects of the imaging technologies and what information is important for each of the imaging technologies.
ZP - You mentioned that most of the time you try different methods and then use the one that works best with each type of data. Do you envision a future world without raw data?
SC - In the data mining, in order to get read of the images, we will need some QC standards run at the beginning. When you're working with human samples, there is a big variability in samples. The analysis steps are very robust on some cases. But we are not removing them because there is not reliable QC. You still have to look at the tissue. It would be nice to measure the structure of the tissues. QC is something big that needs to be fixed.
ZP - I guess that having so many imaging technologies in the imaging field gets thinks more complicated.
SC - Illumina standardised processes. There are lots of technologies, but they all depend on the output. SpaceTx was trying to understand what you were having as outputs from the experiments.
MS - You mentioned that having a tissue QC would be ideal. But you would still collect some fields to understand that the tissue was fine

SC - Sometimes knowing how the tissue was fixed, the background (you can estimate how good the tissue is sometimes). You need to run some analysis on some sections, if possible, and identify RNA morphology etc to score the sample.
MS - Are there standards. You mentioned shipping times etc to give you an idea.
SC - With human samples they are so precious, if we get something we are not confident about, we can track it. Is the technology that is not working, or is something in the sample? Is not like you would discard it but it might help you understand better.
MS - Code as metadata. I never thought about that before. There is no standard pipelines right?
SC - It's interesting to know how your RNA segmented, what criteria (parameters) was used to run a particular analysis. The code, the environment would be important if you wanted to re-run. The criteria or overview of the processing pipeline would be enough to understand what was done (on the data mining side). We are doing data mining you're using lots of data. So you depend on the infrastructure. And it might take a while to get results.
LC - Another question about imaging data. Moving data vs metadata. Data good enough that you could move it without having to move the whole data stack.
SC - Segmented data. Your reference for the segmentation will be there but not the entire raw data. With the improvements you might not even need the nuclei anymore.

SC - Before trashing the data, people can come back and re-process the data to bring better results. But there has to be boundaries. Probably once you feel like the QC are fine etc you move on.
RS - I have a question regarding the probe design. I see that you mentioned genes. Isoforms
SC - With our technologies we need many probes so usually we don't distinguish isoforms. So usually to try the probes matching the isoform most common on that tissue. We look at the fasta sequences to see that they are the isoforms. There are different ways of generating probes so it's useful to record what probes were used.
LC - Even if raw data is not available anymore, we should know the probes that were used.
SC - Yes.
LC - How good are people are naming the genes
SC - We try to use the standard nomenclature. The probe usually would be location and the isoform is matching to.
MS - Any more questions?

March 9th, 2020

Agenda theme this week: The Common Coordinate Framework

Announcements

Agenda Items

Challenges toward building a common coordinate framework for the human respiratory system - Tommaso Biancalani

Publication - “Toward a Common Coordinate Framework for the Human Body”
https://www.sciencedirect.com/science/article/pii/S0092867419312759

https://drive.google.com/drive/u/0/folders/1LlyKwSboIc0yYINIG5XNN-N1n7T0jB0_

At the core of the mouse atlas there is the CCF.

For me the CCF is the most essential part of an atlas. There are big differences between geological and biological atlases. No single ground truth when representing a biological organ.

What Allen has done with the brain is use several independent brains and average them.

Adult mouse brain is an ideal organ for atlas construction due to the big amount of available data. There is the CCF, the ontology information, the cell positions information, the connectivity information, and the sc/sn RNA-seq data.

We plan on using the process that we establish in the brain in other tissues like lungs.

That’s where we encounter most of the problems

We take and transfer the image to the Allen CCF and then use all the annotations into the CCF into the image. Prior can be anatomical information but also gene expression and cell map.

If we sum up the cell clusters in a particular ROI and the prior information contained as annotations (gene expression map, anatomical region map, cell map) we can predict cell type densities mapped in the ROI.

How do we build the CCF?

- There is not much data to play with.
- The anatomy of a tumour is not regular. CCF for different organs might be challenging due to the different nature of each of them.
- The lung seems like a good case for us to work with. Our plan would be to come up with a good pipeline. That can be used in the whole organism.

Data Collection: Extract lungs and inflate them and take CT scans. I have now 5 different lung samples. After ultra CT scan we have lobe separation and slicing (that would provide mapping) that result in block sections where histology is performed.

Computational Strategy:

- Identify anatomical landmarks for image registration
- Iterative construction of an anatomical template
- Overlay a reference coordinate system

For lung we have been using two different strategies, airway branch points and Silhouette and shape (shape and size)

On the first strategy the airways branching points serve as anatomical landmarks.

Off the shelf methods do not yield satisfactory results, with lots of false positives.

We are trying to establish a computational method to determine the branching points.

Machine learning would be great but would require pre-annotated data, which we don't have.

Human airways have a fractal-like structure that we can use (broccoli shape structure) for machine learning.

Deep machine learning identifies airways branch points in lungs without pre-annotated lung images.

We generated a set of annotated airway-looking fractals and we trained 3D-UNET on these synthetic datasets.

Then we transferred our 5 ultra-CT scans of single lungs and detected branch points in human airways.

Trachea and bronchi are very accurately recognised by deep training methods. Lower in the lungs it gets more complex and less accurate.

Poisson distance maps are the second strategy. They are very useful when you have irregular shape tissues.

Before we did a PDM of the whole lung. Then we establish a landmark. And referred the PDM distances to that landmark.

This is ongoing work by 3 PI Aviv Regev, Rahul Satija and John Marioni.

Q&A

Q - Have you used the broccoli approach to vascular areas?
A - No, but I plan to do that. I should probably first check what others have done.

Q - Have you checked if in some ontologies there are different relationships between lung areas, or vascular areas. Have you compared airways with ontology representation?
A - I have not. I guess we could. But we are still very much focused on the fragmentation process.

Q - You said that as you go down accuracy goes down. Do you know by what magnitude?
A - I don't have a figure right now. But it's a very good question. We're currently exploring this. It seems like 1-4 we're doing good inferring the airways. 4-7 not so good

Kathy Reinold (Broad) - I'll reach out. I work on data modeling at the Broad.

Laura - On the ontology side, Katy it would be nice to see what you are using and compare with our ontology application.

Q - Do you plan to collect sequencing data apart from CT scans?
A - They have already produced bulk RNA and some sc RNA-seq data, but I haven't seen the data yet.

Q - HCA does not have a consensus on collection. Any particular piece of information that you find would be essential?
A - I would advocate for collecting as much as possible information prior. I have not yet mapped the lung. Unless we have data from all sorts of sources we can't.

Q - If we are trying to extend to other organs like the liver, uterus etc where the landmarks might not be as clear, how would you approach the CCF on those?
A - Aside from mouse brain, using and creating CCF focused on landmarks seems like the best approach. I am starting to try and adapt the ones for the lung to the digestive system for instance. I would like one method that works for all tissues.

TB - We have the CCF and then we have scRNA data. On the latter we lose track of where the data comes from. We know broadly the tissue/area the samples came from and then we have the plots. The goal of my research is to be able to take all cell types and say where, what region they come from.

Katy Borner's paper about CCFs and vasculature: https://arxiv.org/abs/1911.02995
KB - Any comments welcome

February 10th, 2020

Agenda theme this week:Heart

Announcements

Agenda Items

“How to improve the importance of metadata from the clinic to the lab” - Michela Noseda and Sara Samari (Imperial)

Slides: https://drive.google.com/drive/u/0/folders/1FfvThrekc4g4ewIXw-RLtb9ylPH26v77

Michela - It’s a large project involving many groups, including Sarah Teichman. It is a very exciting project, but very complicated. We have many collaborators.

We have three different tables, with Clinical Data, Experimental Data and Histology Data. These are the 3 types of metadata we are particularly interested on.

For clinical data, we consider the metadata in the slide the minimum essential.

There is nothing out of the ordinary on the experimental data.

Histology gets more complicated.

If we have two donors and we are focusing on some parameters like age etc we don’t have problems merging datasets.

But if we go to diagnosis, it won’t be uniform for every donor. We would need to do a pre-filter for the nomenclature to uniform it.

If we look at the Infection at admission, or a simple one like the Alcohol use. In UK is recorded as units. In the US as number of beers. Sometimes it’s difficult to make qualitative descriptions quantitative.

Look at the definition of a smoker. These fields have a big relevance if you’re studying cardiovascular disease.

We would like to be able to

Kathy Reinold - Smoking duration, in our data model we found it useful to use start date and end date.

Michela - That sounds good

There is information on allergies that might be irrelevant to us but might be very useful for others.

Across countries the blood tests might be presented with different units. If we want to take the raw data from the clinician we might need to convert in order to compare.

In these north American patients when we asked for glucose, in the North American sample we were given 3 reads. In the UK we were given a range. Unexpected information that if you plug into an algorithm you might get.

The more we can think about this early on the better it will be down the line.

Units, different ways of describing diagnosis, consumption of alcohol, smoking etc were the main things I wanted to raise attention to.

We are trying to acquire as much histopathological information for all the samples. My idea is that in v1 of the Human Cell Atlas, we have histo staining of every tissue. I believe this is key stage to move towards molecular resolution. 1 single sample per tissue is not the best scenario, but it could be a start point.

We have histology for each region of the heart that we have analysed. There are lots of images coming. And they require lots of space. We store the raw images and then users can do what they want.

We need to try and give minimum information about the area where the sample came from. This is quite vague at the time.

More things to worry about, simple images and how to link to the anatomy.

Once pathologists go through the images and describe the samples.

Is one section from one patient enough?

In a clinical setting, if a clinician was assessing a patient they would do more stainings. If we were thorough we would ask different stainings. Can we do the interpretation of these images?

Does the DCP want to interfere and elaborate on this?

Marion - We are still trying to understand what the community wants from the imaging data.

Zina - At the moment we don't have much information. I have a question. You say that you take a section, how is the metadata stored at the moment?

Michela - We do the analysis as we want to understand the tissue we are working with. Each section will be analysed by about 2 different pathologists. Can we integrate the findings that each of us groups do?

Zina - Is it now very abstracted?

Michela - How do we bring it together to make the HCA?

Angela Pisco - A question for the DCP. What about create an index, searchable so that researchers know what people are working on. Is it out of scope?

Marion - We are trying to figure out at the moment. If that's useful for the community we could maybe take the proposal to our oversight committee. We can take that as a feedback.

Malte - There are imaging repositories so we probably would like to be in touch with EBI.

Marion - Does anyone have any other questions regarding storing images and metadata?

Michela - People are doing RnaScope to do validation. I am not sure if that should be included or not. In terms of images, that could be useful. I don't have a clear opinion on whether this should be added. It could be useful for certain genes.

Zina - We are working with researchers on spatial transcriptomics. It's still early stages but we are working with them. And also with people doing RNAscope.

Michela - We don't have any more slides. We just wanted to give you examples of metadata that we are collecting and that would be nice to share with the community.

Zina - These 6 areas you are collecting, is this something you do on all samples?

Michela - if we have the whole heart then we are focusing on the 6 regions. When we move on, we will start sequencing more regions and get finer as we know more about the tissue. If we don't have access to all 6 regions, that's fine but we try our best to get 6 or more.

Marc Halushka - Are you adding the date of the study? I am curious about that.

Michela - It's an information I can't give you just yet. In my opinion is important to record that. But I don't know now. In samples from North America it would be good to know where the 3 readings are coming from in time. It would be good to go back and check.

Marc - Did you say the samples were coming from donors?

Michela - At the moment they are coming from donor hearts?

Marc - Did you collect heart weight, atherosclerosis extent etc?

Michela - If we get the whole heart here we would weigh it. But sometimes the valves were taken. We should organise to wait at least when we can.

Sara - We don't weigh the whole heart. Different groups take different samples, and each weights their samples, so we could add up.

Marion - Any other questions? Going to the 6 regions. The regions are still quite vague, I was wondering within your different sampling procedures, how do you make sure samples are comparable. We work with the CBTM. They struggle to be specific about where the samples come from, as organs can have different sizes…

Michela - We have been talking to Krishna and we have gone through a couple of rounds of discussion. Now she has a drawing where she can point out where the sample came from. If we have graphics. We take a number of pictures of the heart when we receive it, and where we sample it. If everyone can record it. For CBTM they take samples in theatre is more complicated as there needs to be asking for permission, and taking photographic recording might be more challenging.

Michela - You can use points of reference but if they are in a rush they might not be able to measure. But at least they can reference visually.

Marion - Maybe we need to think about having a map and sharing it.

Zina - Is there a heart reference atlas for anatomy? I'm coming from working with mice brain.

Michela - We would probably need to talk to pathologists and ask if there is one. Pathologists would be the ones knowing this. A surgeon will only look at one aspect. The pathologists need to look at the whole heart. The idea is to get pathologists more and more involved.

Marc - I am a cardiovascular pathologist. As for atlas, we don't have a defined atlas as for the brain. There aren't as many anatomic landmarks.

Marion - From clinicians sometimes you get different readings. You're still trying to figure out what is the best approach. But do you have a feel for what would be the approach? Choose the latest value?

Michela - For blood tests probably I would go for the latest read. We haven't done the mean for duration of hospitalization, but I think it's been short. But it will depend on the field you're investigating.

Marc - When you have unhealthy patients as donors, I don't think you said you were collecting medication. Some medication might affect expression.

Michela - I didn't point out but we try to collect antibiotics and other drugs. People who collect the organ don't always have the information about the history. They need to be proactive going to gather the information. But we agree that medication and history are important to record.

January 13th, 2020

Agenda theme this week: Kidney

Announcements

Agenda Items

Video of call - https://drive.google.com/open?id=1fSFdpFkkqIM9mBw6cax82Ef_8mEChbVP

Update on the Kidney Cell Atlas (20 minutes) - Anna Greka (Broad Institute), Olga Troyanskaya (Princeton)

Slides (shared with second presentation) - https://drive.google.com/open?id=1kuLmInF9KXMs7U_9Lsp7O3Bqo2CrDf1I6WSejbyDs0A

Anna Greka - I don't think Olga is going to join us today. I will keep some of my comments to a minimum so that Becky and Evren can expand on the metadata. They are a little bit ahead of us in creating the kidney atlas. There are lots of points of interaction. I will allow them to focus on some of those points.

I will share a couple of slides only.

On the last Scientific meeting I presented this slide. There were 62 individuals at the time. This number has probably increased.

Seed Networks - In the UK Menna (Clatworthy) and Sarah Teichmann would be focusing on very specific areas of the kidney and scATAC Seq. In Boston area we are focusing on spatial transcriptomics (slide-seq, HCR/FISH)

We all work with pieces of kidney that come from patients that had kidney tumor but were normal.

The most pristine kidney sample would come from living donor that comes from Princeton. It comes from a living donor. After donating and before going to the recipient Michigan allows to take a sample for research.

In the UK there are ways to work with kidneys that were meant to be donated but they didn't in the end. They are going to be available to Sanger. They are not in the best condition, hence not being donated in the end but they will serve their purpose.

Boston group, Bay Area group, Princeton group, UK groups.

As part of the broader kidney community, Humfries' group for instance, there has been some benchmarking with organoids.

I now will pass the mic to Becky and Evren.

Marion - Does anyone have any questions for Anna now?

Angela (BioHub) - In terms of integration of data from all modalities, do you have a plan on how are you going to approach it? What role is the metadata going to play? How are you going to address batch effect?

Anna - Mattias, Evren or Becky should comment.

Mattias - What variables are present in the different assays and we have been thinking about the metadata. Comes in small flavours comparing some of the assays. We are quite advanced in conversations to align with not only HCA but HubMap and KPMP with regards to kidney.

Development and implementation of a metadata structure in the Kidney Precision Medicine Project (KPMP): clinical, histological, and experimental data capture process (20 minutes) - Becky Steck (Michigan Medicine), Evren Azeloglu (Icahn School of Medicine at Mount Sinai)

Slides - https://drive.google.com/open?id=1kuLmInF9KXMs7U_9Lsp7O3Bqo2CrDf1I6WSejbyDs0A

Becky Steck - Me and Evren are going to be talking and taking any questions.

Evren - We've been thinking about integration of metadata.

Becky - Kidney Precision Medicine Project is looking mainly into kidney and disease. With data coming up we are hoping to find new cells, get closer to an organised kidney tissue atlas.

We have recruitment sites and tissue interrogation sites. The first are the ones collecting the samples. The patients are followed for years. The second is responsible of data analysis, and integration.

Samples are sent to Central Biorepository. From there they are sent to tissue interrogation sites to do imaging or whatever is required. Data integration centre, data is stored, validated, standardised. When is there we can share it with the kidney atlas.

We have been happily influenced by the HCA project in how we thought about the infrastructure.

Before the HCA was a little bit ahead of KPMP.

We have a very small group in comparison to the HCA team.

We think of our clinical data, molecular data

Clinical Data: We collect a lot of phenotypical data about the samples. We have clinicians understanding what is important clinically. We also have a team of ontologists. We have case report forms. The ontologists would go over the forms and reference it to existing ontologies. Sometimes we would have asked to add one because it would be collected this way for kidney. Sometimes we have developed our own ontologies (OPMI and KTAO).

There is lots of metadata that goes along on this type of data.

Pathology Data: We want to co-relate pathological features. So we need more controlled vocabulary for what the pathologists see when they look into the slides. They are working with ontologies (HPO for descriptors, UBERON for anatomy, and CL for cell types). Spatial details. We need to reconcile tissue structural data. We have a pathology working group working on a common coordination framework. That is not as important in kidney.

We will not have resolution of common coordinate framework that other tissues might have.

Because we have lots of spatial information that we are gathering through transcriptomics.

I hope the KPMP will also help us determine healthy kidney.

Kidney disease is a disease of aging. If we look at a kidney of someone older, we might hvae tissue with features of a normal aging kidney, not necessarily kidney disease.

Single cell power means that we might also be able to differentiate healing kidney.

In the application in the example, the user looks for a particular cell to see where is located. It would highlight where it is that cell, and then show you the single cell data.

Molecular Data: Laser micro-disection, single cell, single nucleus, etc. On the right subway map there is a representation of the type of data and where it is. Harmonisation we want to see what happens when you compare metabolomics data and proteomics data etc.

Our approach to determining metadata standards:

1. DVC establishes Pyramid of data elements
2. DVC Compares the suggested data elements from the TISes against the pyramid, as well as against some other standard bodies.
3. DVC Has meetings with TISes (on a technology-themed basis) to discuss open questions (and ideally finalize the metadata)
4. DVC + TISes finalize finalize metadata standards by technology

We have been having discussions on a google group. We meet with the transcriptomics teams, and would ask what is the information that would be important to collect on tissue processing. The domain experts came up with extra information that they needed. The spreadsheet was much bigger and had information on whether a field had to be required or not.

We de-modularised it. We did this deliberately for the sake of re-using data later. All these modules come with SOPs. These are generated by the experts but there are levels of approval. The map and protocol of a module would be versioned.

We are in the process of moving to protocols.io, same as the HCA.

We have the same type of slide for all our type of data. We know that there is proteomics coming. We went through the same process, see what the grouping was. Metabolomics, the same idea, but there is just one group. Imaging data, this is probably the less mature. The imaging team has worked really hard independently to come up with this slide.

Evren - One of the things we thought that was very helpful to diversify the metadata component. We did discussions with metabolomics group, with one single group. We could have meetings with proteomics groups first.

Becky - We are open to questions

Kathy Reinold - I work at Broad. I have been working with ontologies etc for a long time, and I am very impressed about your use of ontologies and your methods. We have run into similar issues across modalities. I would like to get in touch with you off-line to discuss further.

Becky - I meet with part of the EBI metadata team in November, and we are definitely happy to start working together.

Matthias Kretzier - Spending some time making clear how helpful this is for the whole HCA community would be worth it.

Anna Greka - We are 80+ groups in the kidney seed networks. Many samples will be coming from Michigan so it should be easier to be compliant with collecting information. We have KPMP, HubMap, HCA. The idea is that when we have in March in DC the HCA meeting, we can spend some time on metadata and ontologies. Because if we can harmonise those between initiatives that would be extremely helpful. In the meantime Becky can connect with people at EBI, Broad…

Marion - I have a question. Sometimes in HCA, we have this disconnect between what scientists are willing to give us and what the users want to use.

Becky - We have the advantage that we work with the people who collect the data. And we can somehow dictate what metadata is collected. We tried to develop personas. I know HCA did that as well. We do interviews and mock-ups. All our personas are clinicians or users. And they are very close to us. Keep doing interviews etc. If they are involved, they give tonyideas, they feel part of it. I don't know if I answered the question.

Tony B - I want to agree with what Becky said. Having a transparent process to develop the software and engaging people. Getting people engaged early into the development is a great idea.

Logistical question - How do you publish your standards? At EBI we have thought about this. There should be natural alignment. The team here would be willing to get their hands into whatever you have to see how the alignment goes.

Becky - It is not published yet. We have been working on Google docs for sharing and collaborating. The users have given us excel spreadsheet and we have ingested it. It's what is going to be useful for our users so we want to support it. So we are going to create excel spreadsheets with the standards. For now, that's how it's going to be. All our code is open in GitHub. But we haven't used it that much.

Question - How soon can they at least have a look at what you have, rather than waiting for it to be published? Within the DVC I think we should not have to take spreadsheet and rather have drop-down menus.

Tony B - Spreadsheets are easier but you probably want a more structured way. We could share some examples. We would have json schemas and generate web forms or even csv from that json schema. It would be interesting to see alignment between the projects. It would be nice to have a look at what you've got. We would be delighted.

Becky - I want to tidy up and then could happen anytime. We are going to have automatic validation. We started with a UI. When we went to 80-90 fields, they requested to add the spreadsheet where they were already collecting their data.

Tony B - We had the same scenario

Kathy Reinold - At Broad, we're starting to provide drop down menus within Excel. Still a GUI that collects the info would be better!

December 16th, 2019

Agenda theme this week: Lung

Video https://drive.google.com/open?id=1C9GS2V24vvDHr8BefJre1YHwBMQV6K3o

Announcements

Agenda Items

Overview of the HCA-Lung initiative - Martijn Nawijn

Slides - https://drive.google.com/open?id=1l4dpYCv92dML0ZFQVEqByeq00xm0npPZ

I'll be presenting a general overview of the lung initiative and what we are doing not only in Europe.

I will explain what we are doing to try and integrate data from different sources. Hence the importance of the metadata.

Laura, if you could bring up the presentation, I have about 15 slides.

LC - If you could share the slides maybe it would be better so you have control of the slides.

I think this is a decent overview of the projects.

I am from Groningen, in the Netherlands.

For the Human Lung Cell Atlas, we started with an epithelial cell study. Since we have grouped with some groups. We have coordinated quite good. There is LungMAP that has been going for a long time. In the US there is HubMAP. And there is now LungMAP 2. Pascal Barbry in France got a CZI pilot award.

We have been quite ambitious looking at the white paper. Looking at what we promised in that paper we are not quite there yet, but we are getting closer.

Lung is an interesting organ to sample. You can collect samples from living donors.

There is extended annotation towards location of sample within the organ. There is a lot of groups that can contribute. We thought that we would need contribution from anyone who could contribute samples.

It is a community effort.

Papers: First in Nature in 2014, about 200 cells. The number of cells and papers has increased over the last few years. The number of donors is really important for an atlas effort.

For the human lung cell atlas we need to integrate all datasets so that they are in a database and you can search and compare them.

To compare your data to published data. Also to understand where your data fits.

There is lots of mouse studies as well in LungMAP. There was quite a lot of paediatric studies there as well.

Stockholm University and Karolinska are working on lung development.

MRC consortium is also working in lung development

HubMAP (CCF and Lung Tissue Atlas) is working on adult lung, and so are CZI Seed Networks (US mainly but also in EU): Run by Sasha Misharin

Recently awarded EU 15 partner consortium that I am running. We are using techniques used in the other initiatives.

There is multiple consortia and data out there. We will need to put this together.

In January we will start with the EU effort. Since I am the coordinator I feel confident I can share.

It will take healthy live donors and samples from deceased.

It is going to be 5 living donors, but the number of cells and different tissues will be high.

When we get limited amount of tissue we will focus on one single spatial technique.

For spatial, we use two FISH based methodologies, one with 70 probes and one with 30-50 probes. We will do this in the big blocks of samples.

These are being used in the lung development program.

What kind of metadata you need for the imaging data is not clear just yet, and it’s being debated.

snRNA-Seq is also going to be done.

Will all this sequencing data and imaging data.

We want to submit to DCP. GDPR we need to align to. We need to prepare to integrate this data.

There is datasets that are already there. On top of that there are datasets coming, both from sequencing and imaging perspective. If we want to integrate them, we will need to get the right data from all the groups in order to be able to do that. There is no current consensus in the group as to what is the right metadata. The sooner we reach consensus the better.

We need input from the entire community. I appreciate the opportunity to come and discuss this early on.

Absolutely required metadata:

- Location of origin (with respect to CCF): Might be easier in this EU consortium. But it might be more difficult if smaller groups also produce data.
- Basic demographic data, clinical/smoking data, geographical and ethical (Under GDPR we can't collect data on a nice to have. We need to collect data only because it's useful)
- chemistry
-
- DKH - If you don't use the same wording you might be limiting interoperability.
- LC - We are using ontologies to try and allow interoperability
- DKH - In kidney at least, there needs to go through tuning. Someone with biology of kidney to look into the ontologies in detail.
- LC - It would be great to get feedback from yourself and other experts
- Michela (Imperial College) - Do you have experts who can review the ontologies.
- LC - We don't have a consistent help. Maybe I could reach the community to get experts who could revise and comment the ontologies.
- MCN - In the lung we are trying to organise meetings. Any other organ will need to organise such meetings.

CCF:

There was a paper last week from Aviv and Rahul.
They use CT to measure and standardise bronco tree, and how it compares in between different individuals. The HCA lung community will generate lots of hrCT data.

The paper says that we need to be able to build a reference so that we can refer to that reference. For now I would advice we need to be very specific as to where the sample was coming from as lung is a big organ.

Pascal Barbry went to great detail to describe were the samples were originated from. There is a limit to the resolution on what you can record when sampling.

We will have a meeting in Israel in March. You can let me know if you want to register.

Any questions? If not, it makes sense to transfer to Lisa now, so that she can talk about the metadata.

Kathy Reinold - I would encourage people to use the cell ontology. Although it's not perfect, it is a single place where we can strive to make it better. Broad Institute is working hard to limit the ontologies we reference -- and sometimes recommend a limited vocabulary but based on a "preferred" ontology

Lung Data structure, analysis, pipelines, and metadata details - Lisa Sikkema

Slides https://drive.google.com/open?id=1yEvCujuBUnWoLMU2NyR7NgLqmo-GNmBp

Metadata and the Lung Cell Atlas:

I'm a PhD in Fabian Theis' lab and I will be working on taking all the metadata that has been produced in this initiative and try and integrate it. Before I was in Dana Pe'er's lab in New York.

I will now talk about how do we use the metadata from a computational side.

One of the main goals is to try and define and understand what normal tissue looks like. This includes knowing the variation. We also want to define and understand abnormal and its variation.

I have been busy mostly trying to get all the data together and process it in the same way. Different labs have been using different pipelines. There are organisational issues to share data.

There are different methods that are being used for data integration, like batch correction methods.

To give an overview of when we use metadata:

- To find datasets
- To assess the quality (that's not always extracted from matrices)
- Identification and correction of batch effects (determine if there is batch effect present, for which you need to have the biological knowledge, and determine the source
- Data analysis and interpretation (eg disease status)

LC - Is there any particular field that can be more useful identifying batch effects?
Any information can be useful depending on what you need, and the question you are asking. Platform, depth of sequencing, way samples are processed. It's difficult to assess which ones would help most.

Importance of identifying batch effect:

For lung data, due to the amount of data I think it's going to be easy to spot batch effects.

Cell State Annotation:

- Provided by lab
- Mapping to a consensus cell reference

LC - At the moment we are just collecting raw data. Is there any consensus on type of files that people are expecting to use or are using?

There is a few different ones, csv, loom, mtx. Those are the basic ones. Then there is the scanpy platform, that comes with a file that has annotations. Hd5id file? Works with scanpy

LC - We should look into that so that we can give people advice so that we get to shapes that can be interoperable.

Michaela - trying to collect much more data from Heart donors - blood counts etc can be difficult

KR - I would suggest that the cell or the sample can be healthy or diseased, but not the patient. The patient can have a collection of diagnoses that may not be relevant to the cell.

KR - use DUO to track Data Usage rights

November 18, 2019

Agenda theme this week: What are the valuable fields to collect during the tissue acquisition process?

Intro and Announcements

Video https://drive.google.com/open?id=1jXEkJRE-My6yMrFnCUDMZrfdBusqJRlj

Why are we all here? - Laura (5 min)

Slides - https://docs.google.com/presentation/d/1Xufl0hNSW0Xxdx-kibs8pWjp1dyzHwGtA4dhgxRKMfQ/edit#slide=id.g75662fe026_0_22

We are at a transition stage, with more and more data coming in.

The purpose is facilitating the goals on the slide. We want to understand what is needed, what the people collecting the samples can provide us and if the analysts on the user side consider the information useful.

Does anyone have any questions about the purpose of the meetings?

All - No

Agenda Items

CONTRIBUTORS PERSPECTIVE

Mapping fields between MBC Autopsy Metadata and HCA metadata - a model approach to integrating metadata capture. Mark Halushka and Zina Perova (20 min) Slide Deck

- We will present results of metadata mapping of metastatic breast cancer (MBC) Autopsy worksheet used for tissue processing at the Johns Hopkins Hospital to the HCA Metadata Standard.
- We will discuss what constitutes a “Good match”, “Possible match/related fields” and “No match” in a process of manual assessment.
- We will discuss with the community and propose a plan for collecting the valuable fields for the tissue acquisition process.

We wanted to assess what fields was the DCP collecting and what fields might be used to map to the human cell atlas.

The assessment was done manually. IT was divided into good match (green), Possible match/related field (blue) or no match (the later would not be visible in the data portal but still usable in free text field).

See slides for details on exact matching

https://docs.google.com/presentation/d/1zC8jpb3BcMjckDwnpRuebGhTUF3KOAcZ4C3iYd4RPNo/edit#slide=id.g6b4537700d_0_275

Summary:

40 fields with a good match, 32 fields with possible match/related fields, and 40 non-matched fields.

HCA Metadata standard - Motivation:

Data consumers: can search and interpret data efficiently and in a standardized way

Tool developers: rely on metadata to develop and implement analysis and visualization tools

Data contributors: describe their project in a useful way

HCA benefits from being able to access to FAIR data

Kathy Reinold - Did you manage to do this on any automatic way or did you do all the assessment manually?

ZP - It was done manually.

Kathy Reinold - Is the purpose to extend the HCA metadata standard to other consortia?

LC - There have been new awards related to tissue specific approaches. So the idea would be to start doing this rather than going retrospectively doing it later. We have Steven Lisgo, for instance capturing information.

Kathy Reinold - IN the US the NIH is doing metadata harmonization effort. Are you connected?

LC - If you know of initiatives you know and you could point us to that would be great. Apart from this call we also meet on a cross-consortia call to try to align. This is focused on the HCA community although is not exclusive. Matt if you don't mind, and Steven is on the call it would be great if he can share his experience now.

Steven Lisgo - The metadata we collect is standard to what we collect. All the information (like medical history) we are recording we are storing in our database. If we are taking a fetusx§, we collect the mother's details. All tissue that comes from that fetus is assigned individual/specific id. Not all the samples we collect we use for the HCA. Maybe an offline conversation would be useful.

LC - We might have similar discussions with CBTM for samples that might not be used in the HCA but we could take them at EBI.

AI - Laura, Steven and Krishna to have this conversation offline.

CONSUMERS PERSPECTIVE

Single Cell Expression Atlas: Programmatically Mapping and Consuming HCA DCP Metadata. Matthew Green (10 min) PDF

- Introduction to our use case
- Collaborative work on metadata standards https://arxiv.org/pdf/1910.14623.pdf
- Brief look at our current solution
    - To programmatic consumption of data
    - To metadata mapping

I am going to talk from the perspective of the Single Cell Expression Atlas that consumes single cell data from HCA but also from ArrayExpress, etc.

We try to automate the process but we also have a group of curators that validate. They also do differential analysis and annotate ontologies for instance.

We use tools that are in nextflow and we go one step further than the HCA to use scanpy and do representations and we also use anatograms of where the data is located.

We produce this sort of layout. Summary of type of data. For each project we have plots and marker gene tables as well .

Our user case in the HCA is we want to take the raw data to run pipelines and build this.

There has been a large effort for harmonisation of metadata. We wanted to produce a paper as we have done quite a lot of work on the metadata.

If you want specifics, we have a list of specific metadata attributes that we use. Some of them are absolutely required for analysis and you can't run analysis without them. Some other could improve a lot analysis.

Please, get in touch if you have any questions.

I have been working on an automated pipeline that takes the data from the HCA into our system.

I have been trying to combine.

We came up with a very configurable method. Maps entities, maps attributes. Schema changes and sometimes we might have more complex situations. It needs to be configurable. Next steps moving forward, there are quite a few APIs in the DCP. In the future

we are planning on using the new query service. Cause at the time being is difficult to see what data is in the DCP.

We currently use a lot the DSS API. They are going to change so we will need to change there.

Any questions?

LC - Given the previous conversations about tissues do you have a feel as to single cell expression atlas needing more information apart from what type of sample you are using?

Matt - We are multiorgan system. If there are new fields in the HCA we would take it.

LC - Are there any more people who are using the HCA data or trying to build on it that have an opinion on what might be needed from
University of Chicago -

LC - We are aware that some people are collecting information on tissue collection quality etc. We might have taken supplementary files. But it might not be useful from a consumer perspective. If Matt or someone from consumer side is on the call, if you had histology what would you want to do?

Matt - We are going to start working with imaging and we want image overview. If you have samples and images on those samples, and you don't have sequencing, we would be willing to link to bioimaging archives etc

At lungmap we are archiving the regions were samples have been collected from. Being able to go back all the way to the histology of where the sample was collected would be useful, especially in disease as well.

LC - Do you have a feel as if current ontologies allows you to determine location?

We have ontologies to describe branching etc. It's just a very big organ.

LC - If you had histology that could be shared with us so that we can understand how to link it would be great

We are meeting this week. The samples are available, as they can be shared. They are frozen samples. Fixed. Investigators can go back and analyse the samples.

LC - Anyone has anything else they want to talk about?

Request for the HCA-Lung groups starting tissue collection on January 1st: Could we have a metadata ingest form from the group as to align our internal forms to match this as good as possible? (Martijn Nawijn)

We are running an EU consortium on the EU. We are going to adapt metadata to fit DCP or other resources. We have ideas on what we need to capture. But if there are fields that you think that would be useful now is the time to share.

LC - We are going to update the portal to show the metadata standard so that people can understand.

AI - Zina and Matt would like to understand what other groups are working on this.

Currently there is no mailing list but in the future we will work on the best approach and people who are interested can subscribe to it.

There will be a call in a month's time.

Thanks everyone

A.O.B

Future topics - Laura Clarke