PRECISION Human Pain Data Subcommittee Monthly Meeting

Zoom Info:
https://pennmedicine.zoom.us/j/91953694656?pwd=W2bkTasFdEOOBBirxht4ExPkmpxdXg.1

Meeting ID: *919 5369 4656*
Passcode: *070816*

What: *Monthly PRECISION Human Pain Data Subcommittee*

Who: *Human Pain Data Subcommittee Members, DCIC PM, DCIC members, NIH*
- *Optional: open to the full PRECISION Networks*

When: *Friday, December 12 from 11:00-12:00 EST*

Meetings typically occur on the second Friday of each month from 11:00-12:00 EST

Key Documents/Links:
- *Data Subcommittee Meeting Recordings*
- *Dataset Curation and Publication Slides*
- *NIH PRECISION Human Pain Network Resource Page and Office Hours*
- *Linking the same subject across Multiple Datasets*
- ***Dataset Curation & Publication 3: Reality vs Theory***
- *NIH PRECISION Human Pain Network Resource Page*
- Big DRG Paper Preprint

Future Discussion Topics:
- UTD Spinal Cord Paper

2025-12-12

Attendees:

Agenda Overview

Action Items/Challenges

Action Items from 11/14/25 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
53	Circulate updated document listing paper titles that we're considering including in paper package	Rob	11/20/25	11/20/25
54	Draft abstracts and a sentence advocating for papers to be included in a particular journal	Rob, Ted, Wenqin, Will	11/26/25	
55	Reach final decision about external data sharing and communicate plan for sharing processed data with the DCIC	Rob, Ted, Wenqin, Will	11/25/25	
56	Send a link to the updated Dashboard	Joost/Sam	11/19/25	11/20/25

AI Descriptions

51) Discussed during the 11/14 meeting. Conversations will occur in parallel with paper package discussions.
52) Shams confirmed that the DCIC had the most recent Seurat file on 11/7.
53) Ted Price to add diabetic neuropathic pain and spinal cord information to the existing list.
54) *Science* requested initial collection information by 12/1. *Nature* hasn't provided a specific date but is likely to request the same type of initial information (list of paper titles, abstracts, and a sentence explaining rationale for inclusion in a journal).

- The updated version of the collection is here (note added 11/20/25)

55) Nothing is published until people specifically request publication. If it's published they are versioned if there are changes. Sample Dataset
56) *Completed as of 11/20 (see PRECISION Dashboard).*

2025-11-14

Attendees: DP, Julia Bachman, Rachel Weinberg, Wenqin Luo, Will Renthal, Ted Price, Rob Gereau, Diana Tavares, Kevin Boyer, Guoyan Zhao, Sam Kessler, Joost Wagenaar, Jyl Boline, Xianjun Dong, Barbara Gomez, Ish Sankaranarayanan , Mingyao Li, Peter Jin, Huma Naz, Camryn Wellman, Himanshu Chintalapudi, Shams Bhuiyan, Saad Nagi, Sijia Huang, Bryan Copits

Agenda Overview

Action Items/Challenges

Non-Neuronal Cells Preprint Findings (Guoyan and Kevin)

Paper Package Planning (Led by Rob)

PRECISION Dashboard Updates

General Discussion Topics:

- Data Submission/sharing Updates
- Symposium/Conference Planning

Reminders

Metadata standard modifications went into effective on 11/1/25

Suggestion 1

- *Add "donor status" as a required field for all subjects.*

Suggestion 2

- *Standardize missing value labels, use, and definitions. Use only "Not Applicable" and "Not Available" for missing values that are, at minimum, measured in a subset of patients. Variables that are not collected at all because they were never issued and/or they were not part of the study should be left blank.*

- *"Not Applicable" should be used only when the variable could not have been collected, e.g., level of education is not applicable to an infant; and "Not Available" should be used when the variable could have been collected but was not for some reason, e.g., level of education for an adult was not provided.*

- PRECISION Shared Publication Tracking Document
- Tools/Resources Submission Link and Registering RRIDs

Action Items from 10/10/25 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
49	Remove specific Fiber type "[β/δ]" from workflow.	Shams	10/21/25	10/21/25
50	Create a physiology flow chart	Saad	11/14/25	11/12/25
51	~~Consider best approach for submitting processed data to SPARC and update DCIC on final decision~~ • Combined with new AI # 55	U19 PIs	11/7/25	

52	Share processed data with DCIC	Will/Shams	10/22/25	

AI Updates
49) Completed
Description: Wenqin, Will, Shams decided that it might be best to make [Fiber Type] on nomenclature workflow more inclusive and avoid listing alpha/beta.
50) Completed
Description: Wenqin, Will, Shams, and Saad all thought it'd be helpful to create something similar to the schema/workflow document that Shams shared,
51) Added to 11/14 Meeting agenda. Shams confirmed that the DCIC had the most recent Seurat file on 11/7.
Description: Maryann previously noted that you can put all the processed data into one data set or keep it separate. The DCIC can create a data collection that describes that data set and gives you one DOI so you don't have to cite all the individual ones individually. The DCIC can provide more information about available mechanisms and has the ability to handle multiple data sets or one big data set.
52) Shams confirmed that the DCIC had the most recent Seurat file.
Description: During the October Subcommittee meeting, Will suggested that he thinks he currently has all of the processed data from all of the groups and he can share a Dropbox link with the DCIC.

11/14/25 Meeting Notes

Action Items/Challenges

No new challenges mentioned

Julia suggested that the NIH is catching up emails and that PRECISION members should reach out again if they have anything urgent

Some of today's discussion topics, including the paper package and the discussion about processed data are AIs from prior Data Subcommittee Meetings and the SC meeting.

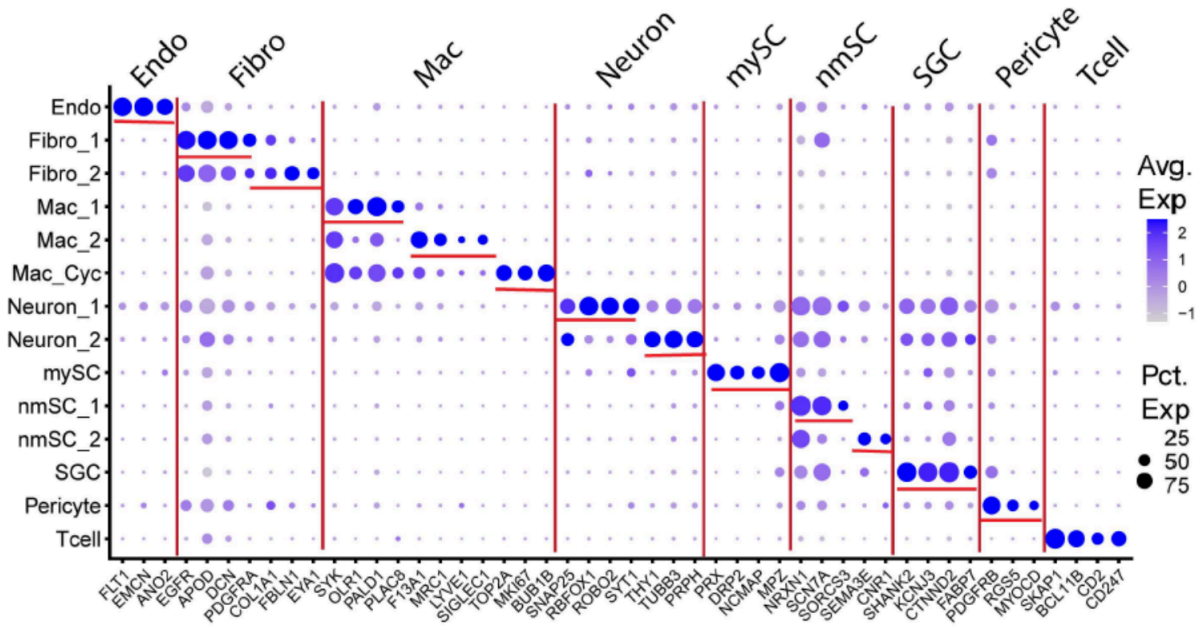
Big DRG paper is now in pre-print. The logic/flowcharts discussed during the last meeting are in paper and this folder.

Non-Neuronal Cells **Preprint Findings (Guoyan and Kevin)**

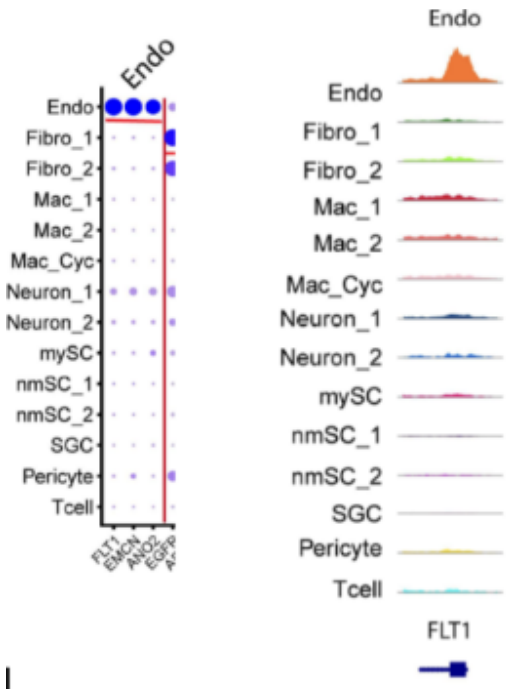
Kevin Boyer: I'll summarize findings from our multiomic data on the DRG which revealed some contribution of non-neuronal cell types to pain. Here's the composition of the non-neuronal cells. We identified the Satellite Glia Cells (SGCs), the Schwann Cells, different Immune cells (Macrophages and T Cells), Fibroblasts, Endothelial cells, and what we're now calling the Mural Cells. This staining image shows the neurons with the NEFH in this teal color. The GFAP is staining the SGCs around them, and then the collagen shows the connective tissue.

But when we look at the actual breakdown of the cells, we found only about 2% of them are neurons which is pretty consistent with what we found in our multiome data. So we took 7 donors and 11 total samples and performed this multiome and we also performed deep sequencing targeting 150,000 reads per cell on the RNA and 100,000 on the ATAC-seq data. Here's the different major cell types we identified. We identified the transcriptional regulatory networks in these different cell types. Then to tie it back to pain, we used GWAS studies to investigate the potential role of these non-neuronal cell types in pain.

This UMAP shows the major cell types that we identified.This is an older figure, and we are currently in the process of updating the names to match that of the big DRG paper. So in total, we have about 66,000 cells, and about 1,100 neurons (1.7%). The dot plot [*below*] just shows some of the markers that we used to identify these major cell types.



Kevin: Then we also were able to identify cell-type-specific peaks for all of the different clusters. [The image below] shows one example for the endothelial cells and FLT1.



I

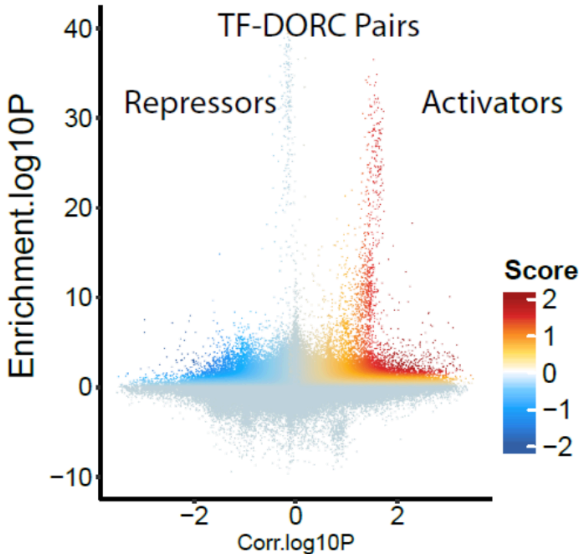
Kevin: Here are the neuronal populations. We only identified the two neuronal populations. Using just Pan-neuronal markers, we can see higher expression in this Neuron_2. And then we found that Neuron_1 is unique for this ROBO2 marker, whereas Neuron_2 is higher for the NEFH. And at the protein level we see that ROBO2 is expressed almost exclusively in the smaller diameter neurons, whereas NEFH is pretty much spread throughout. PRPH–which has a similar expression in Neuron_2–is also in the smaller diameter neurons. So Neuron_1 seems to be only those smaller diameter neurons, and Neuron_2 is kind of a combination of the two. This shows the SGCs and the non-myelinating Schwann Cells. We found that GFAP is not expressed at all at the transcript level, whereas at the protein level we can see it very clearly marks the SGCs around these neurons. And then SCN7A is marking the non-myelinating Schwann Cells which you can see form a ring around those SGCs. And then the NEFH is just showing that those black holes are the neurons here. This shows Somadense areas.

We also looked in the Axon Dense areas, and we can see high enrichment of MBP-which is for the myelinating Schwann Cells–and then much less of the non-myelinating Schwann cell marker for SCN7A.

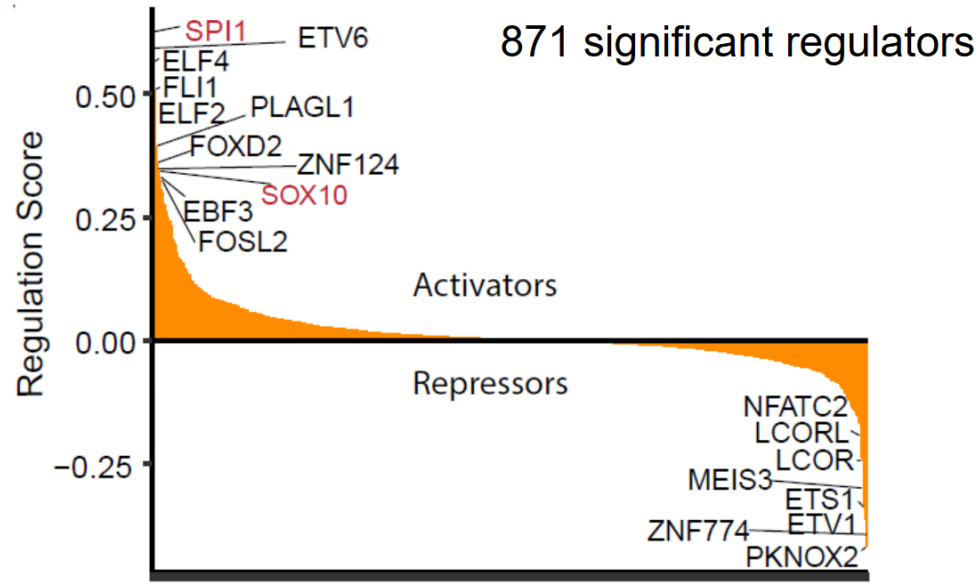
Kevin: So we then decided to look at the transcriptional regulatory networks governing these different cells using a tool called FigR which identifies the significant distal peak-to-gene expression interactions and then identifies these genes as Domains of Regulatory Chromatin (DORCS). So these are genes that have a high density of peak-to-gene interactions. This plot on the left shows the genes with the most interactions. To be considered a significant DORC gene the default is to have n ≥ 7 gene-to-peak interactions. The image on right shows the RNA expression in this Macrophage cluster is consistent with the DORC score assigned to the same gene…showing that these DORCs are consistent with our gene expression.

To look at the transcription factor regulating these genes we used the same tool that performs enrichment of transcription factor binding sites within these DORCs. The plot *[below]* shows TF-DORC Pairs. Each dot is a pair of a gene and a transcription factor. Those with a positive

score are putative activators and those with a negative score are putative repressors of those genes.

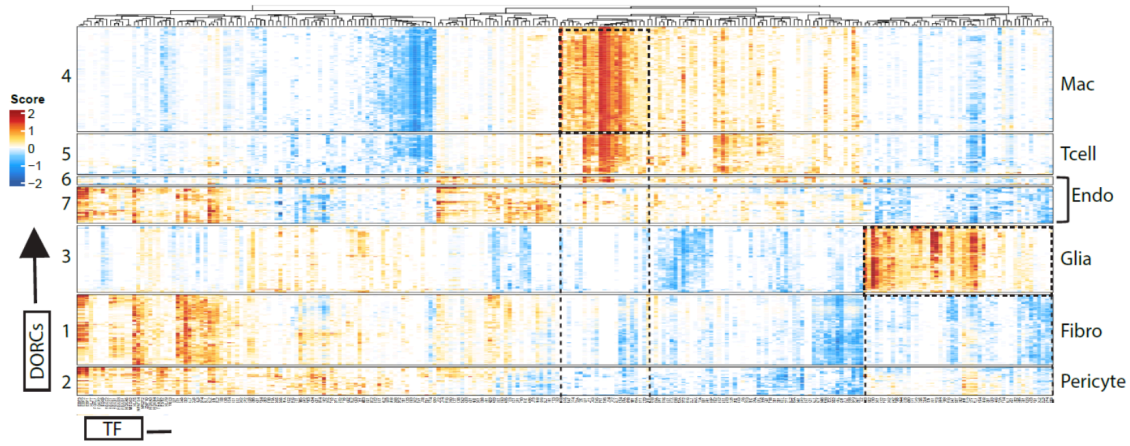


[The image below shows] 871 total transcription factors were identified as being significantly either activating or repressing these DORC genes.

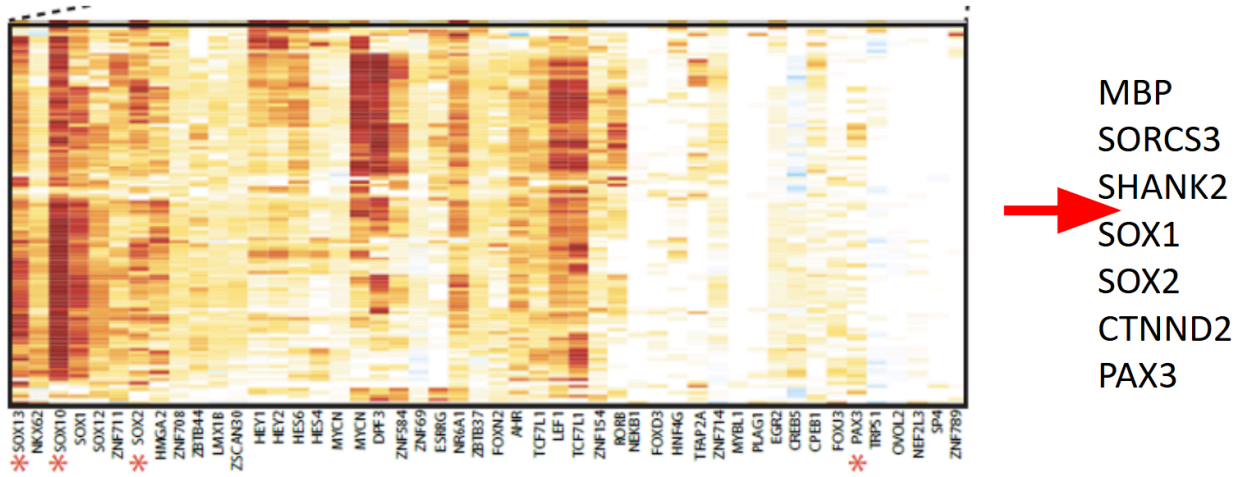


Kevin: So we decided to perform Kmeans clustering to see if we could tie any of these 871 transcription factors to specific cell types. We found these modules separated based on cell type. So we have Macrophage, T Cell, etc. The rows here are the DORCs and the columns are

the transcription factors.



Kevin: Zooming in on the glia cells *[image below*], you can see these are some of the genes that are present in this module which are all present in the different glia cells–whether they're myelinating Schwann cells, non-myelinating Schwann cells, or SGCs–and what we found highlighted in red here are transcription factors that are already known to regulate glia cells.



Kevin: We took all of these transcription factors and first plugged them into the string R Network. Most of them have some sort of known interaction with one another. We also plugged them into Pathway Enrichment analysis and found that the most enriched terms are glial-related, whether they be gliogenesis, glial cell differentiation, etc.

We zoomed in on SOX10 understanding that it's a master regulator of glia cells in the DRG. We see that at the transcript level, SOX10 is mostly enriched in those different glia cell clusters, and a little bit in our neurons. The plot on the right shows TF footprinting. It's basically scanning the different regions and all of our different cell types for enrichment of the SOX10 motif. And it shows the top 3 cell clusters where that motif is most enriched, and there are the 3 glia cell types in the satellite glia, and the 2 non-myelinating Schwann cell clusters.

To look at the expression at the protein level, we can see that the SOX10 not only is co-expressing with the satellite glia around the neurons, but also with the SCN7A–the

non-myelinating Schwann cells. The arrows point to where those genes are co-localizing. Again, this would be in the Somadense region on the top, and the Axon Dense regions on the bottom.

Kevin: To tie everything back to pain, we decided to look at GWAS studies. I should mention that with this dataset, we couldn't do any direct pain vs no pain analyses because many of the samples are pooled together with multiple donors. Some were with pain and some were without, so they couldn't be separated or be able to tell which cells are from the pain donors.

We decided to use this approach to study general pain association in the DRG. We took as many pain-associated SNPs and gene sets as we could find from different GWAS studies or other literature and compared them with our data set. In total we had about 2,024 unique genes from these publications

We first looked at the overlap with the markers from our different cell types. This is Hypergeometric testing between those pain genes from the previous slides (now listed on the right side) and seeing if there's any significant overlap with the cluster markers from our multiome data. Interestingly, we found that Neuron_1–which we presume are those sensory neurons or the smaller diameter neurons–show more overlap with the different pain genes. For Neuron_2 we found that the non-neuronal cell types seem to have the most significant overlap. Specifically, the macrophage 1 cluster–which will be renamed to the MAC_OLR1 to match the big DRG paper–is the only cell type that significantly overlapped with every data set of genes.

Next to go from the genes to the SNPs…or the regions, we used a tool to study gene regulatory networks. We know that transcription factors bind to these distant and proximal regions, but then you have these clusters of binding sites called CIs-Regulatory Modules that can be promoters, enhancers, silencers, insulators, etc. So, theoretically, if there's a SNP inside of these CIs-Regulatory Modules, it can affect the coding of this gene. So we developed this tool called VRMOD which uses computational tools to study transcription regulation.

This has been evaluated on over 300 vertebrate genomes currently to connect transcription factors with binding sites. So it uses only the genomic sequence as input and has been extensively computationally evaluated with many different genomic and epigenomic data (as you can see here in the list of 11 different databases on the right), and has also been experimentally validated in the chicken embryo.

To tie this back to our study, we took these CIs-Regulatory Modules that were predicted by the VRMOD tool and overlapped them with our pain SNPs from our GWAS studies. We then identified a list of CRMs that are associated with these SNP regions and took the snATAC-seq peaks that are significantly associated to genes–our multiome data–and overlapped them with the CIS-regulatory modules. We only found 8 unique genes that met all criteria–meaning that there was a CRM overlapping with a SNP, which overlapped a peak that was significantly associated with the gene. Then we looked further at the other information to see if there was any other evidence in these regions. These are the 8 genes I previously mentioned, including some known and some novel. I've highlighted SVEP1 and SLC4A7 which I'll expand upon.

Kevin: Here SLC4A7 was associated with a chronic pain SNP from one of our GWAS studies. This encodes a sodium bicarbonate cotransporter NBCn1, its dysregulation has been linked to several diseases. This plot is showing the peak gene links in our multiome data. If we use the human genome browser we can zoom in on this region and see that there are two SNPs in this small region, including one that directly overlaps with the CRM. The pink highlights our linked

peak that directly overlaps with that CIS-regulatory module. Then this peak also overlaps with ENCODE cCREs.

When we look at where SLC4A7 is expressed, we can see it's highly enriched in the macrophage clusters. When we zoom in just on that region of the peak that overlaps with the CIS-Regulatory module, we can see it's highly unique in all 3 of the macrophage clusters which may mean that this is a Macrophage-specific SNP regulating this gene. Looking at the same thing for the SVP1 gene we found two SNPs that directly overlapped the CIS-Regulatory module in this gene region, and when we zoom in we see many, many pain-related SNPs in this small region. There's two directly overlapping the CIS-regulatory module here, but many also overlapping are linked to peak as well as this ENCODE cCREs, suggesting this may be a very significant region for pain-related SNPs.

When we look at this gene, we can see it's very unique in the fibroblasts with some expression in the Neuron_2 cluster. When we zoom in on just the peak we see that the attack signal….the open chromatin is highly enriched in the fibroblast clusters with some in the neurons and myelinating Schwann cells well.

In summary, we identified 8 major cell types present in the DRG and found cell-type specific transcription regulatory networks. For example, SOX10 in the glia cells. Then we identified cell types which are associated with pain-related SNPs and genes, namely highlighting the potential role of non-neuronal cell types' contribution to pain sensation in the Macrophage and Fibroblasts.

Will: With the neuron clustering that you've assigned as 1 vs 2, when you anchor these to either the current reference atlas or the older ones; do the cell types line up for this kind of big vs small or are you defining this based on just a couple of genes?

Kevin: It's hard because we tried to do subpopulation analysis just on the neurons, and for 1, there's so few neurons. For 2 the data remained too messy even after we tried several rounds of cleaning to get this object. We also tried zooming in on the subpopulation level and tried to subcluster just the neurons to see if we can find all of those different subtypes

Will: Are you assigning the cell types based on RNA expression or based on attack?

Kevin Boyer: We tried to use both. For the neurons part of the problem is there's so few…There's basically no markers that are conserved across all of our samples. And the same thing with the attack….I've looked at several different markers and regions and there's nothing that's really unique to those regions.

Guoyan: So, Kevin didn't show one of the supplemental figures where we show the canonical sensory neuron markers. They are actually expressed in… enriched in the Neuron_1 population…and also the PRPH….So here we annotated based on that and also based on the staining. So the PRPH is a known small diameter neuron. And we saw the co-localization between the ROBO2 and the PRPH. So the PRPH at the transcript level is actually expressed in both populations. But at the protein level, we're actually only seeing it in the small diameter neuron. This is the rationale that we're naming those two populations

Kevin: Yeah and I think Neuron_2 more than Neuron_1 seems to be a combination of the many different neuron cell types, so large and small diameter which makes it harder since it also is the one that has no conserved markers. So it's harder to dig into that one given its heterogeneity.

Will: This is something that with small numbers that everybody struggles with. We also see different qualities of neurons coming through and sometimes they subset based on just the quality of the sequencing. So we get a Neuron_1 vs Neuron_2 cluster, for instance, and sometimes one is just a bad neuron cluster compared to a good neuron cluster. One of the ways that we can sometimes salvage some of that is just through identity of….you can just anchor to the reference atlases and see whether the probability scores match what your manual identities are. It's not perfect, and you can get artifacts, but you might be able to salvage some of your neurons that way too.

Guoyan: Yeah, it's an interesting observation because the Neuron_1 is actually the one that's much higher in count and features, but the marker gene expression shows Neuron_2 still has those markers for the large diameter neurons. So we think that might actually represent the biology…and there may still be some kind of mixture of small vs large diameter neurons. Maybe we really need more neurons to be able to look at this.

Xianjun: I'm not familiar with the FigR software, but how are the interactions between the Peak and the gene defined here? Is it based on correlation, or…?

Kevin: Yeah, significant correlation between gene expression and the peak.
Xianjun: When you look at one DORC, for example, in one cell type, do you see a case where the peaks might change for the same gene between different cell types?

Guoyan: That's a good question. I don't think we've dug into that yet.

Xianjun: It'll be interesting to see the repertory network change between different cell types. You also showed the SNP data co-localizing with the peak. I realize that your sample size may be too small right now, but for those SNP locations do you see a change in the binding affinity of the motif for any case you show?

Guoyan: Good point. We haven't done this for this project, but it's a good next step. For Tassia's paper we did transcription factor binding site analysis. We found a transcription factor on the BIN1 SNP, which identified four potential transcript factors, and one of the transcript factors actually was shown to regularly be in one expression in the periphery system. So this is definitely a great way to try to identify the transcript factors, but we haven't dug far into this data.

Xianjun: You define two subtypes of Macrophage, Mac_1 and Mac_2, are those subtypes linked to N1, N2 for Macrophage?

Kevin: Right now we're changing the names based on the big DRG paper, so they'll be MAC_OLR1 and MAC_MRC1. I know the MRC1 is the perivascular macrophage…like similar markers, like CD163, but I don't think we're making any sort of functional claims. We don't know what they are. We're just naming them based on their markers at this point.

Will: One thing that we've done back with the TG paper and a few other studies is we've tried to look at SNP associations with gene expression. And we've been surprised by how unstable some of those analyses are depending on what cut-offs you use. So I wonder if you guys have used any controlled—understanding that it's really tough to say what's controlled GWAS. But the question is how often would you see these variants by chance in a different GWAS in a totally different disease state, or how are you thinking about that? Because some of these things are

associated with multiple diseases. I don't have an answer to it. So I'm just asking if you guys figured it out.

Guoyan: No, we haven't. We had similar questions and have been thinking about it because those persons may not actually carry those mutations.

Will: Maybe a call to action for a big multiome paper down the road. Because I think you're right. We do need to aggregate our data to really get good enough coverage.

Paper Package Planning

Rob: We thought it'd be pretty impactful if we could put together a collection of papers to be sent to a journal family and have been collecting a list of titles of the papers that might be included. [*Displaying a list of the papers that are in consideration*] Titles are in bold and we include the principal authors and where they're from. There are currently 9 listed, but I don't think the SCN9A WashU paper (currently #9) paper is likely to be included though we might give it a try at *Nature Communications*.

Our next steps are to put abstracts together and come up with a one-liner advocating for why we think these should land in a particular journal.

Ted: The editors at *Science* and *STM* were both enthusiastic and asked for a list of the paper titles, the abstracts, and one sentence on why it should go to which journal within the AAS family.

Will: Clifford and I met with the editor at *Nature Neuroscience* and *Nature*. Usually those collections are led by *Nature* and not by *Nature Neuroscience*, but the latter will be doing their very first soon-to-be-released collection with the *Scorch* Consortium. They seemed very interested in leading it, or punting it to *Nature* if we felt we had enough papers to be led by that journal. It really depends on what impact journal we want to have lead the consortium. At least at *Nature*, a different editor would be owning the collection depending on the majority of papers going there.

Ted: Maria Amoroso from *Science* said they'd evaluate first, but it'd be good for us to come to a decision about which journal we think would be the best fit. They do like the idea of having a collection from PRECISION and liked the idea of us having a timeline and wanted us to send what we have to them by December 1st. After that they wanted to talk to us after they had seen this. So I think there would be lots of back-and-forth about how to actually go about it.

Will: I pushed back a little bit in terms of what a collection timeline was because if you look at BICAN and others, some of them all get co-posted at exactly the same time, and some are sent at slightly different times and they all get aggregated in the collection. I was trying to understand their reasoning around how they make those decisions and Shari Wiseman from *Nature NeuroScience* told me that they prefer everything to come out so they can have a press release around one publication, but suggested they're open-minded with the timeframe. Six months is more on the far end and they'd prefer to have everything within a few issues of each other.

Rob: We need to collect basically the same information for both. Would that be the same thing for *Nature* that we'd send them that same information and then have a discussion with them.

Will: Yeah. We just replace Science for Nature.

Action Item: *Rob to circulate list and group can start collating all these things into that document.*

PRECISION Dashboard Updates

Sam Kessler: We can skip this topic for today for the sake of time and because we discussed it during previous meetings. We were planning to discuss some new features we added but we'll send out a link.

AI: *Send out a link to the updated Dashboard.*

General Discussion about Data Submission/sharing Updates

Will: Since posting the Big DRG data people have been requesting the data sets. We haven't posted the actual counts matrices online yet because these often change from batch-to-batch as we add more data during revisions. What do people think about posting those data online now vs asking people to wait until it's submitted? I'm pretty flexible.

Rob: I personally feel like we should wait.
Will: It will change, so I worry about getting fractured datasets out, but it's probably not going to change that much. So I don't think there's a strong argument either way.

Ted: The requests are pouring in though.

Will: We should have a unified response to those questions. Either we provide the data or we inform the community that once we get revisions back and have a better sense of the final product we'll let external people know.

Xianjun: Is there a mechanism for sharing data between the U19 groups? It'd be helpful to view other data (including what Guoyan/Kevin shared) from different angles

Guoyan: I'd love to share, but don't know the policy. It's always determined between the teams.

Ted: We'd already decided that we'd share data across the different PRECISION Centers. I know we have an NTA thing that we set up that I think everybody has signed. I don't know about DUAs, but I don't know if we need them. We've been sending data back-and-forth with the other centers..

Julia sent a reminder that the NIH PRECISION Human Pain Network Resource Page has the linked Data Sharing and Publications Agreement.

Xianjun: Are those data in a repository somewhere? I know the PIs might already be corresponding via email.

Julia: Yeah, I think it'd be helpful if Joost was here. Because before the data is published through SPARC we should be able to give other people within the network access to it. So as long as the datasets are up in Pennsieve—at least the processed ones–there should be a mechanism. Joost was developing a mechanism to share across without publishing it. I don't know how that exactly is going to work if the primary sequences need to be shared through the

repository of dbGaP. So that's something we'll have to look into, but there should definitely be a mechanism within Pennsieve to do this.

Will: I agree with Ted, and individually everybody has been sharing on an ad hoc basis. For sharing externally, are we going to share now or wait until we hear back from the journal?

Ted: You're talking about, like, the massive object that has everything integrated, right?

Will: Correct. It's all online now. So if someone wants to go in they can access the data on a website GUI-based. So it's available for everybody publicly, but the raw counts matrix will change from this version to the next. I think the fewer forks of those datasets that exist, the better for the community especially if cell types change.

Ted: A lot of the datasets are already available on SPARC. I know ours and Penn data is, but I don't know about the other data sets. I get Rob's point, but we've been getting so many requests. It's unlikely to change in a major way and it's hard to tell people they can't view it until it's finalized.

Rob: Given the fact that there's a potential year-long timeline for the papers to actually be published the question is if there's any risk.

Ted: Yeah. I hadn't thought about that in a lot of detail, but it is going to be a longer timeline if we do the collection.

Jyl: You can publish it and have a disclaimer at the top of the data set description suggesting that things might change over time. Also, every time you add new data on SPARC, you get a new version, which is also a new DOI. So, if somebody gets an older version and they're referencing that, you can always compare and contrast that way.

Will: So maybe the way to do it is have the current processed version as a DOI. My primary concern was that we're going to change the version multiple times and people are going to be using the wrong versions that we won't have a way to track. So if there's a way on Pennsieve to do that, that'd solve that concern. It's really a matter of as a group whether we're concerned about sharing the data openly.

Rob: Yeah, I don't have a strong feeling about this data sharing thing, but I don't want to speak for other people. I haven't talked to Guoyan or Valeria about this yet.

Guoyan: Yeah, we received these types of requests and discussed this a little and prefer to wait for our paper to have some kind of decision before we share that.

Jyl: Here' s a link with Ted's data that has two versions. It shows the current version and has a DOI, and you can see other versions. So if people are citing the DOI, then they should be pointing to the right set of data.

Will: Let's follow up by email and make a final decision. I'll review what Jyl and we can coordinate amongst the senior leaders of the paper and make a decision.

***AI**: Make a final decision about data sharing.*

General Discussion about Symposium/Conference Planning

Sam: To allow more time for conversation, we'll skip this topic. We were going to talk a little more about the Code-a-thon that we touched on during the last SC meeting. Now that the NIH is back, we'll discuss this a little more with them and get back to you.

Reminders

Sam stressed the importance of updating the PRECISION Shared Publication Tracking document linked at the top of the agenda.

Sam also noted that metadata standard modifications that were suggested during the 10/15 SC were adopted. An SC follow-up email was sent on 10/22 requesting final feedback by 10/31 with a note that if there was no additional feedback that the metadata modifications would go into effect on November 1. The modification language is included above.

2025-10-10

Attendees: Saad Nagi, Bijesh George, Dmitry Usoskin, Sam Kessler, Sue Tappan, Jyl Boline, Himanshu Chintalapudi, Diana Tavares Ferreira, Ilias Ziogas, Camryn Wellman, Bryan Copits, Maryann Martone, Rob Gereau, Shams Bhuiyan, Mingyao Li, Xianjun Dong, Guoyan Zhao, Huma Naz, Wenqin Luo, Joost Wagenaar, Will Renthal, Sam Kraft, Fahim Imam, Ibrahim Saliu, Peter Jin, Ruifeng Hu

Agenda Overview

Action Items/Challenges

Big "DRG" Neuron Paper and Dashboard Discussion

Non-Neuronal cells Pre-print (Guoyan)

Updates/Reminders:

- Abstract Deadlines for symposium abstracts for USASP and IASP
    - Call for Symposia | USASP
    - IASP Call for Symposia-Pain Research Forum
- PRECISION Shared Publication Tracking Document
- Tools/Resources Submission Link and Registering RRIDs
- SC Meeting on 10/15
- Technology Meeting on 10/20

- The meeting will focus on recent spatial proteomics progress.

Action Items from 9/12/25 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
47	Share pre-print big DRG paper, Seurat object, and relevant datasets/dataset details with Maryann and Joost	Ted/Wenqin	9/26/25	9/22/25
48	Explore V1 dashboard and share feedback/suggestions	All Investigators	10/10/25	10/10/25

AI Background Notes:

47) These materials (and an improved understanding of the naming conventions that the U19s have agreed to use amongst each other) will be particularly helpful for K-CORE as it plans out next SCKAN steps.

10/10 Update: Completed

48) Please share your feedback/suggestions in this newly created Dashboard Feedback Document live document.

10/10 Update: Completed. Dashboard Feedback Document was shared on 9/18. No additional comments were added to the document, but the PRECISION DRG Atlas Resource was created.

10/10/25 Meeting Notes

Action Items/Challenges

Sam Kessler: I created a Dashboard Feedback Document after last month's Subcommittee Meeting including some general notes and feedback we've already received. While no additional comments were added to the document, we received an updated Seurat file on 10/7 and Shams kindly agreed to lead a discussion about the PRECISION DRG Atlas Resource that was linked in the preprint and provide a general overview of the data harmonization process he used for this big DRG paper.

Big "DRG" Neuron Paper and Dashboard Discussion

Shams: Will and I are going over some of the final details of the manuscript and we hope to post it on BioArchive shortly. It's been a pretty big and good effort by the entire group. It was an integration of data from 126 donors and it yielded an atlas of just under 54,000 neurons and we identified 22 cell types across these data sets, and saw a pretty nice consistency of these cell

types between the different sequencing methods. Then in the process, we've also been working on developing a good nomenclature schema that we hope will help alleviate some of the discrepancies of cell type names between studies, or at least give people like us a workflow to work with and better understand how we approached labeling these cell types.

This started off as a discussion between me, Ted, and Patrick, and then with the help of Dimitry we pinpointed some of the molecular features to go along with this nomenclature.The first question was…we do have this one cell type where we have a bit of an exception compared to the rest of the cell types, and it's this ATF3 cell type that also expresses markers like UCN, pink, and and certain injury-associated marker genes, and we decided to leave this as ATF3. For all other cell types, the nomenclature follows this standard that we're proposing,

[Fiber type][β/δ] – [Propr/LTMR/PEP/NP/Thermo].[Molecular Features]

where it starts off with fiber type, and then a β vs a δ on conduction velocity, and then a physiology, either proprioceptors, low threshold mechanoreceptors, PEP to indicate Peptidergic Polymodal. We're still working out what that NP should stand for…and then Thermosensory, so unimodal for at least temperatures, in the sense that it is temperature detecting for hot and cold. We might also consider adding in HTMR as another potential physiology to include into this nomenclature, just for situations where we have a peptidergic cell type that isn't polymodal.

AI: *Remove specific Fiber type "[*β/δ]" from workflow.

Shams: I can update that to reflect this workflow. Then after the physiology, we'd have molecular features. Right now we're hoping for two genes maximum that mark the cell type. And if possible, the aim would be to have a consistent molecular feature between the human and mouse, although we've learned that that's not always possible for a lot of these genes especially in cases where we haven't been seeing one-to-one orthologs or the lines we've used to study cell types in mice aren't from marker genes that cleanly map to the equivalent human cell type…then for the rest of the workflow for each one of these components, we've mentioned how we've guided our decisions and what we're labeling it.

Fiber type was based on the Myelination scores that Dimitry generated and then for physiology, we went with this workflow where there's a question, is there human physiology? And in the case where there is, we label for human physiology…..Wenqin, your Nature Neuroscience publication from last year, and I believe what you were calling hPEP NTRK3 was a unimodal HTMR that you didn't see temperature sensitivity. Is that true?

Wenqin: I think that's what Saad thought was most likely based on what he was recording

Saad: Yeah, so we typically find two types of HDMRs both in the A-beta range. Both are mechanoreceptors and there's one type that responds to cooling and we combined some targeted pharmacology during matroneography recordings to show that it relies on TRPM8 or its cooling response and there is this other type that seems to be a pure mechanoreceptor. So for the cooling response we now have a lot more data. There's another preprint out as well, showing that it's the PEP kit. But for this other type, excluding other possibilities, we think it might be the NTRK3, but we haven't directly tested that.

Shams: And so for that one it'd be an HTMR and not polymodal then, if I understand correctly, for that NTRK3 type.

Wenqin: But it's a big maybe. Because, for example, we haven't really tested whether the NT-3, for example, could sensitize this population, so for microneurography, if we want to have more direct evidence that's usually more tests of a bunch of the receptors expressed in that cell population. Like whether that can modulate the activity. So, I think what we need to decide here is really whether we want to just keep the PEP name, I think it's important to think about what we consider HTMLs. Because I think those are kinds of nociceptor things, but if we call them that, we may get criticized and be asked how exactly we know these are nociceptors.,

So that's the reason we go with the name PEP. I feel that name would be harder to dispute. For example, for this NTRK3 population, it'd be our best guess, like we think that it could be HTMR that Saad's group has recorded that does not have the cooling or heating temperature sensation because TRPV1- negative. So we know they are TRPV1 negative. They don't have this heating sensitivity, they are also TRPM8 negative, so they don't have cooling sensitivity, but the PPK has TRPM8, and that the other population has cooling, but not heating. So, that seems consistent with the molecular profiles, but to clearly align them will need more work, and I think as the whole field is pushing here.

And I think it'll also open us up to questions about why we call some PPEs HTMRs and not for others. So, it feels like a rabbit hole. So that's why I think for the physiological property thing, since we are more confident about the low threshold, the LTMR and Proprioceptor side–Proprioceptor is also a type of LTMR in a way—So I think we just use those names for the LTMR side, but for the HTMR side, it's not necessary to go into so much detail for now, because that's much more complicated.

Guoyan: Because of the absence, I'm still a little bit concerned because, for example, GFAP, we don't see the transcript expression, but when you do the protein staining, you actually see the strong expression in satellite glia cells. I'm concerned, if we make the conclusion based on the absence of a transcript that that'd cause problems down the road.

Wenqin: No, I think that's a really good point, but I think we kind of validate that with 10X Xenium and we see really a good correlation so far for our single-soma sequencing data for that TRPV1 TRPM8 expression or PLC 2 for those to predict their temperature sensitivity vs mechanical sensitivity. I think so far the correlation seems to fit based on the sequencing result that we saw, but I think that's a very valid concern, and that's why I think it needs spatial transcriptomics validation in addition to just sequencing results.

Will: Yeah, I guess to Guoyan's point ... .even with transcriptomics, sometimes with GFAP, you see the protein, but not the mRNA so there could be situations here where there's a very low expression of these channels that are functional, but we're just missing. The reason why we started putting this flowchart together is because there are a lot of more branches as we start thinking through the naming. So now, there's one that we hadn't considered here, which was the HTMR vs LTMR, which I guess you could branch out…..Like, we trust the microneurography for the LTMRs more because they're easier to record from, and in the HDMRs there's just more variability. So they could be polymodal and we're just not detecting it in that physiological state….Is that how to articulate that difference?

Wenqin: The simple reason is that from the physiologist perspective, the HTMR category is more complicated.

Will: One of the things that I find so striking—and Shams reminded me with David's most recent cell paper–is that some of these classical polymodal cells are actually not that polymodal with the things he tested. Some of them are really only mechanoresponsive, and some of them are really only heat-responsive. This is including a paper that showed that, basically, these SSTR2 cells that David recorded from by calcium imaging is purely C heat are actually stretch-responsive in that kind of ex vivo patch recording they did. So it may be that some of these things are actually more polymodal depending on the state than we think from that study.

Wenqin: Not just the state, but also the prep, including how you record and how you stimulate. Calcium imaging is different in a way from direct physiological recordings. So I think there are a number of considerations that I don't think this paper will be able to homologize all those and reflect in the name…and I don't think that it will be possible to do that in this paper for nomenclature.

Will: Completely agree. It'd just be helpful for us to have some sort of systematic approach articulating why we made certain decisions, even if it is something like, "it's just too complicated for this class of cells."

Maryann: We previously talked about computational ways of handling these names, and there are standards for how you do that and we'd be very happy to either present again in depth about how we would handle this situation, or even be able to help you come up with a naming convention. Because when you've added functional properties to anatomical structures or cell types when they turn out to be involved, or channels involved in multiple things that causes confusion down the line, but I'd strongly recommend that you put some sort of computational infrastructure behind this, because this is the type of problem that is always encountered in trying to deal with complex properties where things are always sort of changing and we don't have a complete understanding of how these things function and what they function in; that's always sort of changing.

Shams: I think we want to have a nomenclature converter. My goal today is to articulate the logic for the nomenclature that we're presenting into the paper. It's less about saying what it can be converted to in other studies, but it's mentioning that this isn't quite the nomenclature that other people have used, but it's a mushing of previous people's nomenclatures.

Rob: I think including this logic/actual flowchart is a good idea because every time we read one of these new papers the naming conventions are changing a little bit and this is going to continue to evolve, but if we can at least articulate that this is how we've done it here… we have data here that is a completely new compared to where we have been in terms of the scope of what we've been able to do computationally in terms of identifying cell types. And if you have this flow chart and this table that Wenqin was talking about, I think it'll be much easier for the field to see where things are at least at this point in time, and it'll continue to evolve. as we do get more detail on protein and on function.
Will: Saad, I'd love to hear your thoughts on this bifurcation here of the physiology differences between the LTMRs and the HTMR-polymodal cells or your confidence level in spotting differences between the LTMR recordings and the HTMR recordings.

Saad: Yeah, in physiology it's clear-cut. Our stand-alone criterion is a soft brush. LTMRs respond very well to a soft brush. HTMRs don't. We'll find a neat separation in mechanical thresholds in a number of studies. So, those that respond to soft brush low mechanical thresholds are typical LTMRs. And those that don't, we classify them as HTMRs. Now, whether

all HDMRs are nociceptors or a subset of HDMIs are nociceptors, I'm not sure how to tell them apart.

Will: But even taking nociception out of the equation, just whether it's polymodal versus HTMR, how confident are we on that decision? Because I feel like that's really the key thing about this branch here is whether we're going to say this is a polymodal cell or a mechanosensitive and not heat/cold responsive.

Saad: Within A fibers we only got to 50 degrees, and we have never seen a heat response, but we do clearly see the cool-positive ones at baseline without inducing any sensitization. And we think that's TRPM8-dependent. A challenge with HDMRs is that they're quite prone to sensitization too. And that always has to be very carefully considered, but from our data it seems mechanoreceptors, A-fiber mechanoreceptors, high threshold mechanoreceptors–which comprise about 10% of the A-fiber population—those that we call HDMRs do not respond to soft brush, have high mechanical thresholds, and encode forces in the noxious range. So those are the three criteria we use to classify them as a HDMR.

Will Renthal: Is there a standard protocol for how to determine whether a cell has been sensitized or is that one of the complexities Wenqin mentioned earlier?

Wenqin: So what I'd meant by complexity is this field has a long history and people sometimes used different experimental settings and stimuli. So, because the mechanical stimuli is the easiest one to deliver, that's the reason why this LTMR is more prominent in the field when compared with HTMR, but if you see a lot of studies a complication comes from when people call it HTMR, it's a response to stronger mechanical force. That doesn't mean they are not polymodal. They could very well respond to chemicals or to temperatures. We haven't studied it exhaustively so the ground of truth isn't there. Alternatively, for LTMR on proprioceptors I feel like we have a clearer idea about physiological clustering. That's the reason why it probably makes sense for us to put it in.

Saad: I agree. To add to that, the polymodality within LTMRs hasn't been studied well either. For instance, the cooling or heating response of CLTMRs or the cooling responses of slowly adapting A-Beta efforts haven't been well studied either. So even within the LTMR range, we've largely restricted ourselves in physiology to mechanical stimuli.

Will: Yeah, that is complicated. Because I guess they're all polymodal if you stimulate them the right way.

Wenqin: Yeah, possibly

Will: The naming convention we're using in the paper is largely, for most of the cell types, a one-to-one correspondence with the way they cluster, or with the way they integrate with mouse names.
Wenqin: I agree Basically, I feel this name is a mixture of, like, field convention and cross-species alignment, It's a mix of several different aspects to come down to this name and to the functional assignment,
Will: I completely agree, and I was thinking that if put into something like a flow chart, it would just allow other people to use the same way we approach it, or at least allow us to understand it.
Saad: Would it help if we create something similar for physiology in terms of how we classify units? Would it be worth adding that into the paper?

Will: I think it would be very valuable for me, but I'm not sure, I don't want to speak for everybody.
Shams: I think it would be valuable for me. I can't speak for everyone either.
Wenqin: It's something good to have, at least even for the supplementary data for, like, recording because, like, then they see how you classify those as LTMRs or A- -beta LTMRs, or SA1 versus SA2…
Saad: Exactly, and likewise for HTMRs as well.

- ***AI:*** *Saad to create a flow chart for physiology*

Web Resource Discussion

Shams: The goal here was to get a web resource that was public-facing out as soon as possible. We use the Shiny Cell 2 package to generate it. It works fairly similar to other resources that use Shiny Cell, including our own from last year and I know Patrick's group has a number of them as well. We are aiming to get people onto the resource where they can type in gene names and look at gene expression patterns. We also will include a nomenclature converter. That's under construction. The way I approached this was similar to what we had talked about in previous conversations. When Will and I sat down, it was hard to imagine what other people would use, but we could imagine what we would use and the key thing that we always look at is that if somebody is talking about a gene, then what does the gene expression pattern look like. [*Shams selected TRPM8 on the screen*] And as we've discussed there's this TRPM8 high population that's associated with the C fiber, and then there's this KIT population that also expresses a fair amount of TRPM8, albeit not as high. People can also explore this using the violin plot. Sometimes that's easier for looking at gene expression information and so, take a look at TRPM8 and here's this TRPM8 high population and this A-PEP Kit along with a few other C-fibers that have some background expression.

Shams: So compared to UMAPs it often helps to add some quantification to how we look at the data. Those are the two big features I think people use the web resource for. There are other features like if you're interested in the co-expression of two genes,if you want to generate a dot plot using these data or look at segmentation of the data across different metadata columns. These are all features included, but I think the gene expression pattern is really what people care about the most,

Will: If there's anything that you guys would like added for many things, it's relatively straightforward. Also if you see anything that's typed wrong, or for instance, some of these labels aren't user-friendly. We'll be editing those over the coming week or two, but certainly if you see anything that you'd like to have added, or you see wrong, please do let us know, because it hasn't really been released widely yet.

Joost: This is fantastic, Shams. It helped me tremendously to figure out what people are interested in seeing. And so now we have access to the latest data set, we're trying to see what kind of functionality would be nice to bring in the SPARC portal and mimic these things. And it makes total sense that you guys are focusing on getting something up now and then as part of the publication we can iterate on and add functionality. And I don't see any problems with us lagging behind you and seeing how we bring some of that functionality to the SPARC portal as part of published datasets.

Maryann: We wanted to make sure that datasets and any resources that are used in these papers are properly cited according to standards for data citation and resource citation. This is very important to HEAL and in general for assigning credit and making sure that the data sets that are referenced in papers are unambiguously found…so not just data is available here, but this exact data that we are talking about here is available here, and also for tracking them through the literature. For other consortia, we've typically just gone through the papers and provided the citations for you. The only thing in the manuscript that we received on the big DRG paper is one data set referenced by DOI, but I think there's a lot of datasets that go into this.

The one thing we want is to be very consistent even across resources. What often happens when multiple resources use the same data set is it's not apparent that this is the same data set. So, we just want to make sure that the referencing of these datasets is done appropriately so that it's very clear what it is you're talking about. But as I said, we'll just provide a list, and it's a simple matter of just adding the citations to the paper when you submit it or during review.

Shams: Sure, that sounds good. We still haven't posted it yet.

Maryann: Okay, just give us what you need, and that's what we'd like to do. We are also planning on taking whatever you decide about cell names and cell types and contributing that to community resources like the provisional cell ontology.

SPARC has also been working with HubMap on the cell types. They have something called Anatomical Structure Cell type and biomarker tables, and they've really been waiting for good data coming from the human on biomarkers for peripheral nervous system cell types.

So after the papers have undergone review and things, we do want to contribute those cell types to their ASCT plus B tables for the PNS, which we maintain. So I just wanted to let you know about those two opportunities, and I think it's a little premature right now, but we will revisit this.

Guoyan: Is this data submission to SPARC sufficient for publication because I realized that the raw data actually is not included in the submission.

Maryann: The raw data, I think, is in dbGaP, correct?

Guoyan: Well, it could be the… I thought later also great. It can also be submitted to GEO, right?

Maryann: Yeah, I think all of the raw data that's human data has to go into DBGAP. That's the policy and it's what NIH wanted but those will be linked to the SPARC data sets, so by virtue, we will just say in the citation that data has been deposited in SPARC, and all the identifiers in dbGaP will be linked through SPARC, so people will be able to find it.

Guoyan: Is submitting to SPARC sufficient for journal requirements?

Maryann: Yeah, if it's raw data. So the one place they do require that the raw sequencing data go to is dbGaP, but then they leave it up to you for the other repositories. So, it will conform to what's required. SPARC is cited all the time in journal papers.

Guoyan: So, we have to submit the raw data to fit the journal requirement in order to…

Maryann: Yeah, and also NIH had wanted all of that data in DbGAP as well. So I think they'd mentioned that they prefer it over GEO. So the raw data has to go and the derived data should go into SPARC. And then we can mint a collection DOI which would have the list of dbGaP references, and also the list of SPARC datasets. So we can help create that for you and the journal will accept it.

Guoyan: Well, yes, we have submitted two places. What's the purpose of having the processed data in SPARC? Because everything else also needs to be submitted to GEO or other data bases.

Maryann: Because the requirement of the PRECISION is that it goes into SPARC because SPARC is doing all the alignment with HEAL and the way we've been working with PRECISION datasets is using the ones on SPARC because managing them is a lot easier to access than going through dbGaP or GEO.

Joost: The HEAL program has requirements to put it in one of the HEAL repositories, which SPARC is one of them. and so that goes to the larger HEAL data ecosystem, because we integrate directly with the HEAL data ecosystem, so they're indexing all of the data sets. So a lot of the requirements come down from the NIH. This is why they set up these data coordination cores to support those efforts, and why this is set up the way it's set up.

Wenqin: So, if that's the case, is the action item, that every U19 center will submit your own raw sequencing data set to dbGaP, but then for this SPARC DOI thing where will we have one citation for this processed data, Shams and Will, will you just submit all of those processed data to SPARC so we have one common DOI from SPARC for the processed data? And each U19 center does their own dbGaP submission for the raw data?.

AI: *Determine approach for submitting processed data to SPARC and properly citing it.*

Maryann: Yeah, I think we need to leave that up to you, how it is that you are submitting this. We can handle it in SPARC… either you put all the process data into one data set or you can keep it separate, and we can create what's called a data collection that describes that data set and gives you one DOI so you don't have to cite all the individual ones individually. But it depends on how you're going to make the data available for the paper. It seems like that the raw data is all going to be submitted separately, because I think that's how you've been doing it, but the main thing that we care about is the data that underlies the claims in those papers.

Joost: Outside of the human data sets that need to go to dbGaP, the question that I think dictates whether you do one dataset or multiple datasets is how you want to cite this in papers. So a single data set that gets published gets a single DOI, so that will be the thing that you'd reference in a paper. If you'd always want to reference one data set, then it makes sense to submit it as a single data set.

So a dataset is typically a publishable unit that you'd like to be able to reference in a paper. The way that we have been doing this in SPARC is you put the data in there, there's metadata that comes with that, and then if there's an associated dbGaP link, then we put that as part of that dataset publication. So if you go to the DOI, it'd actually say, "this is the data set and this is the dbGaP reference for the volume and data." So the curation team can help with specific questions on how to organize that. But we handle both things. We have data sets on here that are spanning multiple patients. We have data sets that have a single data set per patient. We have data sets that are coming out, single data sets that come out of collaborative efforts, and

we have data sets where each group submits their own thing, tracks their own thing, and then create a collection afterwards, where we basically combine different published data sets into a single resource. So there's a lot of different options.

Maryann: Exactly, you just have to tell us how you'd like to do this, and we can help you if you'd like to discuss the different mechanisms ...We can handle one big data set or multiple data sets, but need you to let us know how you want it. We can make sure that if it's one big data set, that there are appropriate links to all of the dbGaP identifiers that say, here's where the raw data has been deposited. So we just need to know what you want to do.

Will: I think we have all the data from all the groups, the processed data, and probably can just share a Dropbox link with the Data Coordinating Center. If there needs to be a lot more metadata, we might need to get help from individual centers

> ***AI:*** *Shams/Will to send processed data to DCIC*

Maryann: That'd be very helpful.

Non-Neuronal cells Pre-print (Guoyan)

- This topic will be shifted to the November meeting.

2025-09-12

Attendees: DP, Rachel Werner, Julia Bachman, Joost Wagenaar, Maryann Martone, Jyl Boline, Sam Kessler, Camryn Wellman, Elle Mehinovic, Huma Naz, Kevin Boyer, Ish, Peter Jin, Diana Tavares Ferreira, Sijia Huang, Ted Price, Xianjun Dong, Ilias Ziogas, Himanshu Chintalapudi, Guoyan Zhao, Wenqin Luo, Sue Tappan, Khadijah Mazhar, Bijesh, Ruifeng Hu

Agenda Overview

Action Items/Challenges

Linking Subjects Across Datasets Subjects across datasets

SCKAN Presentation PRECISION: SKCAN & Cell types V 2

PRECISION Dashboard

Updates/Reminders:

- PRECISION Shared Publication Tracking Document
- Tools/Resources Submission Link and Registering RRIDs
- SC Meeting scheduled from 2:00-3:00 on Wednesday, October 1

Action Items from 8/8/25 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
45	Email Subcommittee a summary of the Initial Version of Dashboard	Joost	8/8/25	8/8/25
46	Share Seurat Object with Joost	Shams/Will	8/8/25	8/8/25

45) Completed
46) Completed

9/12/25 Meeting Notes

Action Items/Challenges

- No outstanding AIs or pressing challenges
- Future agenda ideas discussed include having a dbGaP curator join and talking about the harmonization process for the big DRG paper.

Linking Subjects Across Datasets

Maryann: This presentation was created to remind people about the importance of consistent subject and sample identifiers. We've got a distributed workflow in that datasets—even those with the same subjects and the same samples–may be loaded at different times and go to different places. That means consistency across the data sets is paramount if we want to be able to reassemble the data sets.

We're currently specifically talking about how people are now uploading their data to dbGaP. There's some DCIC support coming on this, but also the curators do a deep check to ensure that data sets can be put back together and we often find that the naming is inconsistent across these data sets.

So in order for it to work (*screensharing the example on this slide*), if you have a sub-2 in Dataset 1–assuming this is all coming from the same subjects–then sub-2 needs to be identically named across those different contexts.They can't be different identifiers, they can't be different forms of the same identifier because then computationally putting the data set back together again becomes very difficult.

So, this is just a SPARC example where we have two datasets on the same subjects. They were done at different times and we note for the subject file that all of the metadata is exactly the same. The subject IDs are the same. Because there can be ambiguity there is also a place in the SPARC dataset description where you basically give the provenance so you link those two data sets.

So when you go into dbGaP, we're also going to be looking for the accession number. And it's always a good idea to provide some information to a user as that some subjects are here and some subjects are there. There may be multiple legitimate reasons for why this is the case or why you may not have collected the same type of data on all of them, but it's always helpful to orient the user as to why there may be discrepancies.

We're always looking for computational tools and just want to ensure that provenance is absolutely clear by using the same identifier and that we're taking advantage of ways to provide the necessary data set accession numbers along with information about why are these two files may differ. All of that goes a long way towards making these datasets usable. So, then the curators are here to work on this. One of the most important things in curation is to make sure that those subject identifiers map.

SCKAN Presentation

Maryann: This is a brief introduction on the work that we're doing to develop some of the knowledge-based aspects of what PRECISION is producing for exposure through the SPARC Portal. I think I'd given a presentation several months ago talking about creating a computable representation of cell types that we're being produced. So, the idea is that you have cell types that you are identifying from transcriptomic data, patch-seq, whatever types of phenotypes that you are using and then we want to make sure that these cell types are both consistent across different products that are being developed.

We're not trying to impose uniformity, but we want to ensure that when cell types are identified based on a set of phenotypes that we make those phenotypes computable by mapping these different features to ontologies. And by doing that, we can create a structure that's called a knowledge graph, a Knowledge Base–they're related to each other–which means that we can display and manipulate this information in different ways. So, I just wanted to give a brief update on the types of tools that we have available, and some of the demonstrations that we are preparing.

A knowledge graph is a semantic network that represents a network of real-world entities. There are huge efforts across different institutes and funding agencies for groups to start expressing the information that they collect in terms of these knowledge graphs. So this is really part of a larger effort to start really ramping up on production of knowledge. They are also very flexible and powerful structures that let you do a lot of things.

If you have your knowledge expressed this way, one of the very simplest things is that we can create different facets across the data. So if you want to explore the data according to all these different facets that we have in SPARC–like sex, age, and experimental approach–those are all parts of the knowledge graph. And for them to work effectively, you have to have something in common between the two that says–just like in the case of subjects–these two things are the same thing.

So that's why we work with ontologies and controlled vocabularies. So, as part of our work with PRECISION, we want to be able to extend the SPARC knowledge graph to contain information on cell types and key molecules. This is not a data structure. It's a knowledge structure, so it doesn't necessarily take the entire expression pattern across thousands of genes and encode it.

The other type of knowledge that we have in the SPARC knowledge graph is connectivity. This is one of the major products that SPARC developed where the connections that were made between the CNS and the PNS were encoded in this formal ontological structure so that you can trace connectivity back-and-forth between the central nervous system computationally.

So this is an example of SPARC's flat map. All of these lines were generated computationally from a large knowledge base that's called SKCAN. The computer read this and said, "here's the origin of this particular pathway, here are the nerves that it runs through, and here are its targets." So, this is all computable connectivity knowledge, and we had proposed that we'd start to extend that into the rest of the peripheral nervous system because this is primarily an autonomic nervous system through projects such as PRECISION.

So we've taken some of those steps. Again, these are just proto-steps. We need to reach out to you as the experts, but I wanted to let you know that this work is starting to progress. The way that we did our connectivity knowledge graph was based on a model of the neuron itself. So these connections are not just that Region A projects to Region B, it is a population of neurons that sit in region A, send axons to region B and terminate there.

So there is a model of the neuron in there. We use something called the neuron phenotype ontology for this because this allows us to associate phenotypes–not just with an entire connection–but also with parts of connections so if there are subtypes of cells, if there are molecular profiles of cells, or axonal or terminal specializations, we can record that information at that granular level and still have it roll up into the type of regional connectivity that we're normally associating with connectivity databases. So it's a very, very powerful model. And it's the neuron phenotype ontology that we want to use to be able to express the phenotypes of the cell types that you're identifying for the DRG.

So, one of the things that we're doing in collaboration with HubMap is starting to put in the connections of the somatosensory system as opposed to just the autonomic nervous system. And we've been working on a lot of that cutaneous skin connections–just based on general knowledge about how these things work–and these have been added to the SKCAN database.

So in addition to being able to visualize that connectivity through a SPARC Maps, you can also interrogate the database directly and generate graphs so you can see different ways that these different structures hook up. So we've started to do that for these general populations, and we're starting to dive a little bit more deeply into DRG populations, specifically involved in pain sensation and this is something we'd really like your help with.

I know when we've talked about this before, we've also mentioned that you're dealing with other types of connectivity in addition to just pure anatomical connectivity. We want to then extend the knowledge graph, not just with these connections, but with more detailed phenotypic information about the cell types and molecules and we've been working with the 2024 Science Advances because we wanted to show exactly what we meant about making these phenotypes computable.

So we took this very nice table that was produced where they go across species and talk about marker genes and physiological properties and fiber type and everything else and linked that to formal ontologies. It looks like this right now because we're extracting information, but this is basically taking that information and turning it into something that a computer can process. We're using NPO for that. We don't yet have these imported into NPO, but the idea is that you translate the terminology that a group uses, you express it in terms of these ontologies, and then you can generate class names of any type that you want because it's just generated based on a rule. So if you want species first, or morphology, electrophysiology, molecular, anything in there, you basically can generate and say this is the rule by which I want to generate names.

I was involved in a committee on nomenclature for this, and there was a lot of disagreement about…what should go first and "are these the properties." This system is agnostic to how you want to name it, but allows you to generate these names automatically. And behind the scenes, these are all mapped to common identifiers, so–just like with those subject IDs–regardless of what you may call that concept across different representations, it all comes to the same ID and allows us to gather these things together. So we've started this process of translating the table just to show you how it works and we've also started to play with what we might be able to put up. Again, this is just a very quick demonstration, but we'll want your opinion on this.

Satrajit Ghosh is one of our close collaborators on several projects who developed a whole technology platform to allow you to take these knowledge graphs and essentially turn them into a nice card-like interface. So Jyl is working with him on that.

So you can basically take that spreadsheet that Ilias put in the knowledge graph–that has all of the basic pieces and notes and it automatically took that information that we're putting into that spreadsheet and turned it into a nice interface which you can see here has the name of the neuron, and then all of the properties that we have encoded.

If we can get all of this information into this knowledge graph structure, it'd allow us to create these very flexible interfaces where you'd be able to click on this and say, "I'd like to see all cells in mouse," or "I'd like to see all cells with this particular marker gene or physiology," and you can reorganize the knowledge graph around these different facets just like you can through the faceted interface of SPARC.

In order for this to work and be able to easily link across different tools we really do need to, ideally, standardize the cell nomenclature so that the cells that are present in the PRECISION dashboard. For example, when they pull these out, if we have a card or something for them we understand what the relationship is between the two. And for that, we really will need the help of this group.

So these were the recommendations that I'd suggested the last time–having consistent name and machine-processable cell type identifications really allow for very powerful search visualization capabilities. I think you can see that with those cell cards, that was taking existing code, and allowing the AI to program it. Again, these are flexible tools that take what it is that you believe and they just express it in a different form, but one that's much more flexible than just a table. And we'd really like to work with PRECISION on any connectivity data….if that's generated we'd want to import it into SkCAN. And that will also help us ensure that we have pain pathways because that is a means for linking across these different knowledge graphs

DP: Are the models and the net connectivity data that were developed based on publications of data from humans or from multiple species?

Maryann: The Connectivity Knowledge Base has any of the data that we have tagged according to which species we observe it in. So, some of it is human. The peripheral nervous system data that we put in is largely human. Again, it's at a textbook level for the most part, because we don't have really detailed experimental tracing in humans, but for SPARC we worked with experts-and we base the animal work on this sort of detailed experimental context. But you can filter that graph to find if data comes from humans, mice, etc…because a lot of detailed connectivity is inferred from other animals, in the model that we use we say that this has been observed in these species because we're finding all kinds of species differences between different nerve pathways. The work we did to get the peripheral pathways was done in collaboration with HubMap, and they are human-focused.

We're going to continue–unless someone has a different idea–working with the cell populations from that table. So, if there's work that is going on in the big DRG paper about cell names we'd love to be involved in that. It'd really be nice if the work that goes in there is represented in the knowledge graph.

Ted: For the preprint, we have a deadline of next Tuesday and a group of us have been working on this all the time. Wenqin is here, so she's one of the ones working all the time. We came up with a naming scheme for the preprint that didn't involve using the system and we'll miss the deadline if we change it at this point.

Maryann: No, that's fine. We can accommodate whatever naming system you have. But it'd be great if we could have a copy of the paper.

Ted: Absolutely. We should be sending a draft and a Seurat object, and other things Tuesday or Wednesday…there'll be a mix of the paper and a bunch of different data sets that went into putting it all together, and we can get that over to y'all.

- *AI: Ted or Wenqin to share a copy of paper, Seurat object, and data set information with the DCIC.*

PRECISION Dashboard

Joost: So in anticipation of the paper we've been working building a dashboard to display the data from this atlas. The data that is in this dashboard, is an early version of this Atlas that was shared with us. So what we're looking at here is the Atlas that is on the SPARC portal. So if you go to staging.sparc.science/apps/precision-dashboard you'll find this dashboard.

So the dashboard is a kind of universal framework that we can have widgets on. We can create or resize these widgets. We have multiple widgets that we can add as part of this initial effort. We created three specific ones—UMAP Viewer, the Data Explorer, and the Proportion Plot–and then those widgets interact with data on our platform.

The data in our platform, in this case, is one of those Seurat files. The Seurat file is a fairly large file that is uploaded to our platform that contains all of the data that is necessary to generate these plots. Because this file is really big, we can't easily make a dashboard that downloads that entire file and starts rendering it. So what happens on our platform first is that we convert this internally to a file format called Parquet that is optimized for cloud streaming and dashboard functionality. Parquet is a very common format that is currently being used for a lot of data visualization and data integration efforts, but this is something that users won't see

This dashboard actually runs an in-memory database called DuckDB, that can query that data over the interconnect connection, and then display that data. And so what we have here is two visualization plots that we mimicked from the PainSeq dataset. This is the UMAP viewer. It plots all of the points in the dataset, and we've tested that up to like 300,000 points and it seems to be working just fine. You can select different ways to color code the individual cells by study, but then also by atlas annotation. And even though it looks pretty small on this canvas now, you can resize this to make this a lot bigger because we set it up in a way that we can query it, we can use that same data then to create these these proportion plots where we can put something on the x-axis and something on the y-axis and we can plop these things alongside one another and allow people to explore this data in a more meaningful way.

Something that I think is pretty pretty powerful–and we hope to do more– is that you can actually provide a way for people to directly use SQL to interact with this data. So you should be able to run multiple types of queries and export that to a CSV file for people to do further analysis.….We're still on V1 and there's a number of things that we already flagged where we see possible improvements and we'd appreciate your feedback to ensure that we make this a useful tool for the community as we're rolling out this paper and this atlas through the PRECISION consortium. I'll be looking forward to getting the final data set to get that up there.

AI: Investigators to explore dashboard and provide feedback

Ted: We should have the data set ready to go pretty quick, again, and we intend to get it to y'all as soon as it's ready.… Although, it actually is going to be multiple data sets that led to the naming, and I guess we'll have to figure out how to kind of separate them out, but the names will be consistent, so it should be easy. I'd imagine what you could do is say, "this is the subtype, and here's one kind of data, here's another kind of data, and here's a third kind of data all demonstrating the transcriptomic subtypes." So there are some differences within them, but blending them together would probably be a little bit problematic.
Joost: Yeah, because, I assume we're not doing any recomputation of UMAP coordinates or anything like that. We take them from the Seurat file.

Ted: Yeah, it is remarkable how well the different types of data have aligned. And that analysis will all be in the paper….So give us, like, 2 weeks and it'll be ready to go.

Joost: That sounds good to me. I hope that once this is out that we can work together as a team to make this the most impactful resource—which is not just making the data publicly available—but also making it available in a way that actually reaches a lot of people. So please send any thoughts/comments our way and we'll iterate over the next year or as long as it takes to increase the impact of these types of visualizations.

Wenqin: We're currently under a great time pressure to get that pre-print out, but after getting the pre-print out, between the pre-print and the submission, the manuscript, there's probably going to take some time of refinement. I think this could be a good opportunity to include this dashboard in the formal submission or before the paper comes out.

DP: I totally agree. There's no pressure to include it at this point from a preprint side, but definitely before submitting the final version to the publisher, including this could be really, really helpful for the entire community.

Xianjun: Also, is it possible to search by the individual gene or visualize individual genes on the UMAP?

Joost: So that'd be one of the things that'd be great to get put on an email.

Xianjun: Yeah, it's like a feature plot from the Seurat… It'd be nice to allow the UMAP panel full screen. Right now, it's a little bit small.

Joost: That's one of the things we're working on and it should be very easy to change. We're working on a few things now, but the idea will be that users can have multiple dashboards with different configurations of widgets that you can save or select from a library.

Ted: The two things that people use our Sensoryomics site for by far the most are to do search for their favorite gene and to do co-expression of their favorite gene with marker genes. I think you should put that in there and have it be something that's quite prominent right at the start. Our users usually come over for only about two minutes–so I assume they're either writing or reading something and then they're exploring, "was that really expressed in this population?"

Joost: This is exactly the type of feedback that'll be super helpful.

Xianjun: I like your SQL query part as a computational person, but I don't know how useful that is for biologists. They like to modify the MySQL query. I think it'd be clear to define your main user.

Joost: Ultimately, I think there will be various dashboards that we do. And one might be where I can see the SQL and–it might not necessarily be for this specific dataset–but it's great for doing types of cohort creation. If in the end we use this same kind of technology to have patients or subjects or something like that being able to do that in a read-only way…in a sandboxed way. I agree that we need to figure out what the user group that is interested in this specific dataset, and tailor the specific dashboard to that user. And it's most likely not the computational user that really wants to have access to the raw data and to AI pipelines.

Maryann: At the very least, you need to make provenance explicit. I see a lot of these tools that have names but you don't know what those names mean or what the data sets are. The whole concept of these broader knowledge graphs is you can do anything you want, but that you just need the metadata and IDs and to make the provenance clear so people know where it came from.

Joost: Right. So as long as the curation process results in a final Seurat file that has the right terminology–which I think should be the final output for this paper–then the dashboard will use the right names…..right now, the data that is part of this isn't published on SPARC yet and hasn't gone through the standard curation processes on SPARC.

We generated this dashboard based on a file that was provided by us. What I fully expect is that as this data is finalized and being made public, this should go through a dataset publication mechanism. And so it becomes available as a dataset on SPARC, and then this will be a dashboard that sits alongside that dataset to have a better view into it. The reason we have this markdown file widget is so we can put that type of information in there, but right now it has boilerplate text.

Updates/Reminders:

Sam: I'd asked Ted to provide a brief update on Technology meeting that we've been working on scheduling.
Ted: I can just say that was gonna be about progress on spatial proteomics which has been substantial, so it should be a fun meeting.

- *AI: Sam to provide updated meeting time options for proposed Technology meeting.*

2025-08-08

Attendees: Sam Kessler, Maryann Martone, Jyl Boline, Xianjun Dong, Barbara Gomez, Bijesh George, Bryan Copits, Camryn Wellman, DP, Diana Tavares Ferreira, Will Renthal, Huma Naz, Guoyan Zhao, Himanshu Chintalapudi, Ilias Ziogas, Ishwarya, Julia Bachman, Kevin Boyer, Peter Jin, Mingyao Li, Rachel Weinberg, Rob Gereau, Ruifeng Hu, Selwyn Jayakar, Shams, Sijia Huang, Sue Tappan, Joost Wagenaar,

Agenda Overview

Action Items/Challenges

Metadata Standards

- Dataset Curation & Publication 3: Reality vs Theory

Web Portal Demo (Ruifeng Hu)

- BrainDataPortal_RuifengHu

Updates/Reminders:

- PRECISION Shared Publication Tracking Document.
- Tools/Resources Submission Link and Registering RRIDs
- SC Meeting scheduled from 2:00-3:00 on Wednesday, October 1

8/8/25 Meeting Notes

- Diana is the new Data Subcommittee Co-Chair

Metadata Standards

Maryann: [Sharing Slides] We talked about metadata a couple of months ago but wanted to discuss challenges that groups are encountering with metadata standards now that the curation team is receiving more datasets and has more use cases for different data types.

We've updated the slidedeck to remind everybody that we're working on allowing people to put their data into dbGAP and SPARC. Ultimately, all data is going into SPARC, being uploaded to Pennsieve with the goal that we can view and search that both through SPARC and HEAL.

I think we mentioned that the first datasets are now available through HEAL, which has these specific requirements for various CDEs. And ideally, we want each project (even if you're not using the CDEs), to adhere to a standard data dictionary to make it easier to query variable-level data through the HEAL platform.

The metadata standard for PRECISION Version 2 was approved in January. All of the documentation for preparing data for PRECISION and all the specific PRECISION requirements are available on the PRECISION Documentation Page on the SPARC Portal. As you start to collect your data you need to download both the PRECISION template for data collection and also the PRECISION Data Dictionary which describes what all of these variables mean and how you're supposed to use them.

A lot of the 99 fields that are required come because of the requirements of the HEAL project. The PRECISION data dictionary that has been prepared has different sheets for those requirements that come from HEAL for human subject data…So you can see here, there's 99 fields that are required, 4 fields that are recommended, and 2 fields that are "if relevant." And the curators and our validation instruments are going to check against those validated fields and if they're not present, then errors get kicked out.

Maryann: So we're trying to make this easier and more relevant for PRECISION. So, first of all, those who are collecting HEAL CDEs are doing it through REDCap. And whereas before people were transferring the data from the REDCap data file/data dictionary over to the PRECISION templates, we don't think that that's necessary. All HEAL needs is access to the REDCap file. So you can just upload that file with it and we don't have to do that detailed transfer over.

There are still some limited number of fields that have to go to the PRECISION template, particularly subject ID so that we can link across everything effectively….This REDCaP data file/data dictionary can be (but doesn't have to be) a data set on its own that can be used to provide metadata, for example, for data sets that you publish sequentially.

But basically, you don't have to copy over all those fields into the template. It's an error-prone process. What you do have to do is give us the REDCap file and the associated data dictionary with it and the curators will go through detailed instructions on how to handle subject IDs. And part of the reason is that the REDCap doesn't always provide us with the subject ID that you assign in an easy form. It tends to assign its own record ID to it. So some work has to be done there, but it's minimal.

Maryann: Data coming from surgical biopsies or cadavers does not have these same requirements. So we've been hearing from the curators that there's been very little problem with people filling out the section on surgical biopsies or cadavers because this only has 11 mandatory fields. In general, if any of these fields are included in the HEAL CDEs we try to align them because that makes it easier to do searches across all data.

However, if CDEs for these subjects HAVE been or WILL be collected, they should be submitted-either with the dataset or in the future. The REDCap export is also preferred for that. We confirmed with HEAL that you can publish your data in slices. We do have ways inside of SPARC of linking those together, but if you're going to do it that way, those HEAL CDEs do have to be present at some point. So it's absolutely critical that they be submitted so that all of the data in these slices makes sense and can be tied back to those CDEs.

Bryan: We've realized that not every surgical patient will complete the survey for the required HEAL CDEs. Does that mean that we cannot use samples from those Patient donors?

Maryann: Absolutely not…. If you don't have the HEAL CDes for some of them, I think that just gets explained. The idea is that if you can do it, you should. But we understand that it might not be possible to get that and we understand that you don't always have control over how these arrive.

Maryann: We have missing value designations. We won't cover them in detail now because we're still testing them, but the idea is we'll have a set of missing value designations, which is requested but not answered……So we're going to continue to put out detailed instructions on how you handle those things, but as long as you record the reason that is sufficient. If you don't even have the reason because a lot of this started before HEAL, I think that's understood as well.

Then if the CDEs are not relevant to that context (like the cadaver case), then you don't have to just put "not applicable" for every answer….You can just leave all the HEAL CDEs blank and provide the critical subject metadata that you do have. Because you always have some.

Joost: It'd be fantastic if we can work with REDCap files and include those into our structure. I do have a question about separating the metadata from the original data. It seems that you mentioned that you can do that as a separate dataset?....I'd think that we probably want to keep the metadata with the raw data, because otherwise we would have a completely separate DOI and author list description, image…

Maryann: That always happens if you publish slices of data. But ideally we'd want all of the metadata to be available when you publish the first data set….but we decided to check with HEAL and they said it's perfectly fine as long as it's linked to every data that requires it.

So that's why we looked really deeply into subject identifiers. Because that'd only work if the subject identifiers were identical across every data set that was published as part of that data set. So if you do that, you need to check your internal processes about subject IDs and you need to contact the curation team as soon as possible to ensure that in every case the metadata can be associated with the data files that are coming out.

We had a long discussion about this on Friday, and our first instinct was "no, you can't separate the metadata from this," but if this is the way that the investigators were planning to do it, there is a way to do it. It just has to be done very deliberately to make sure the dataset can be reassembled.

Maryann: The thing that the curation will really be paying attention to is the subject identifier. I think these slides show how you can do it, but that might be something that we can talk about more and think about building some automated pipelines.

Joost: Yeah, I was thinking about how we keep these things together because it's not the same as our current process where we have one thing linking to one other thing. We now have one data set with multiple samples that need to link to multiple samples in another dataset.

Maryann: Yeah, I agree. We should probably go through that. Yeah, so this actually would have to link to two data sets. You're absolutely correct. And I don't think there's anything in our data model that doesn't allow that to happen, but if this is the way that data have to be published,

then we'll figure out a way to handle it so that it's HEAL compliant…..So the goal of this now that we have real-life use cases is to modify and simplify our process and also focus on the things that we really need to focus on and not spending time cutting and pasting when that's really not necessary if you're using REDCap.

Web Portal Demo (Ruifeng Hu)

Xianjun: Ruifeng from my team led the development and will be giving a demo/presentation on behalf of us. This portal isn't developed by the U19 teams and wasn't funded for that purpose, but we developed this for a different project and thought it'd be nice to share with this community given that the portal can be easily applied to different data sets and projects.

Ruifeng: I will quickly go through a high-level introduction about this portal, and then demo the website. Here is the landing page including the navigation bar for going to the function pages. The main part listed is the brain region which is clickable and the number of datasets/assay types for the brain sample. At the bottom, there is a quick access button to a tutorial for uploading a new dataset.

For the Tech Stack, I'm using React for the front end, FastAPI for the back end, and SQLite3 for the database. So, in current design, there are three main data tables at the back end–study, dataset, and sample. Here is the relationship. In one study, there may be several different data sets, and in each dataset there are several samples.

Ruifeng: In the current version, we support importing data from Seurat and CSV files. So, here are the required slots and columns. In red, I highlighted the required data slots for different data types.

Here's the landing page and the data set list, and the supported views for different kinds of data visualization purposes

Here's the gene view that has a feature plot and a violin plot for genes, and it shows clusters based on select features if no gene is selected.

Here's a visual transcriptomic data view that shows the slices colored by different metadata. Also, the QTL view can help to visualize the SNP and associated genes.

The landing page is also easy to use for other projects because the backend was designed for generalized purposes.

Ruifeng: Then we can go to the dataset list page (Recording—scroll to around 31:30). I currently have 7 available datasets. I can give you a demo there. This is a single nuclei RNA-seq from our internal data.

So, first it's showing a U-MAP. The metadata is sometimes very big and takes a few seconds to load. After it's done loading, you can then the clusters based on different plotting, and you can select genes and it will show the MajorCellGene plot. Multiple genes can be selected.

At the bottom, there is a violin plot, but the group is based on what you selected here… Another function (Recording-scroll to around 33:20) shows a feature plot for genes and a violin plot for the selected genes in different groups. I also preloaded a cluster view….Once we selected a cluster, it will highlight it in the UMAP, and at the bottom, it will show the dot plot for

the micro genes, as well as the cell count in different condition groups, as well as the differential expression within the selected cell type.

….This UMAP is interactive, you can manually show or hide the selected clusters. So, these options, such as the default samples or default gene or features are in a config file when you are preparing the dataset.

You can also use the QTL view (Recording–scroll to around 36:00). So at the bottom is a selected gene and each row is another SNP and if you hover you can see some statistical information.

The Brain Portal is currently still under development, but I prepared this tutorial for how to set up this app on your local server on your own local computer ... .Once you set up the app, you can do the data preparation. I provided a step-by-step guide to prepare a dataset.

Joost: Could you put a link to GitHub in the chat? Does your web app or your API directly read from the Seurat RDS files, or is there pre-processing going on that you store the data in your database?

Ruifeng: We provided some scripts and have made it standardized so it's applicable to most cases. And you can use our provided scripts to process your thread RDS data or CSV file. Then to upload the processed data set–it's usually a folder–you can just upload that folder to the backend.
Joost: If I give you a pointer to a file, for example, on SPARC, would that directly allow you to visualize that data? Or does there need to be some process to take the data out of the RDS file into some database before viewing it?

Ruifeng: Currently you need to preprocess your dataset. It cannot automatically process yet.

Joost: So you store the actual values that you're displaying in your dashboard in some sort of SQL database somewhere?

Ruifeng: Actually, you can install this app. It's not a web server though it acts like one. You can install this app on your own server on your own computer. You don't need to upload your data to our server.

Xianjun: That's a good point you made to read the point URL from SPARC directly and visualize locally. That's definitely a direction we're working towards. But to your question, we will not save any data to the database. So basically, we just read the RDS files from URL or from uploads, and then convert to local format. That's it. There's no database for that raw data at all. Because that would be huge.

Maryann: It'd also be good in your metadata if you allowed for the DOIs of data sets to be directly published, instead of just the publication DOI. In BICAN we saw a lot of researchers put their data in a lot of places but then it gets out of sync…To Joost's point, I think it'd be helpful if you can read directly from a data repository so that provenance is there.

Xianjun: We developed this Portal largely to share new internal data amongst team members so that the team can explore the story behind the data, but I see your point… We don't really have a web URL yet–we go through the Yale VPN–but we will make it public in about a month.

Discussion about Next Steps for Data Visualization

Rob: Can our U24 colleagues comment on what the plan is for how we're going to do visualization?

Joost: We've been going back-and-forth with a number of you about how the U24 is going to set that up. And as we've demonstrated before, we have some similar functionality around taking datasets and creating these UMAPS from it. We also have a larger effort to generate dashboards within the SPARC Portal and integrating these UMAP viewers inside the dashboard.

I'll be sharing an email with all of you later today that outlines our plans. The primary thing that we need at this point is clarity on what files will be uploaded, and in what format—like, what are the expectations around the files that are uploaded that we need to parse into the viewer. The previous presentation was a great introduction to that because it outlines the requirements of this viewer, and I see that there is an RDS file with certain fields populated, such as the UMAP coordinates, etc. And so we will need to have similar requirements for the files that we can display in our dashboard.

It'd be great if we had an example of dataset PRECISION Workspace on Pennsieve that we can use as the example dataset that we build the initial visualization around. As far as I know, the last time that we talked, we discussed the Pain-seq dashboard and the plan was to mimic some of those visualizations and that requires that there is an RDS Seurat file that we can process. And then we can surface that to the viewer with some selection where you can highlight the different cell types in a similar way. And so that's kind of what we're doing now.

I think our current blocker is identifying as a group the files that we're going to display on the dashboard. And our thinking was that if there's this big DRG paper, then there should be a shared dataset somewhere where we have files that we can display

Rob: So that could be a good starting point for the transcriptomic data. But we're gonna have multimodal data, and single-cell, single nucleus, spatial, and proteomic data too.

Joost: So calling that out is helpful for us to know, and we also realize that we need to do things in an iterative way. So let's start with one of those modalities, and then assume that we layer things over time. So my suggestion was that we start out in a similar way to what was just presented ... like having a dashboard where we can pull in the data, show UMAPs based on cell classification or other UMAPs that are already defined in the RDS files, and build an interactive visualization around that.

I think that that is the best approach and the reason we, from a foundational point of view, started out with developing a dashboard framework that we expand on with a notion of widgets that we can put different tools on top of….whether they're kind of quantitative metrics, where there's a bar graph or visual metrics, such as UMAPS or imaging or things like that.

Rob: Yeah, and people need to be able to search for their genes, their cell types, all these things.

Joost: Exactly. So I have to put this in writing and will send an email out today and request getting access to the data sets that this group wants to display there. So, the set of RDS files that we want to include in this initial visualization.

Sam: During the last Steering Committee meeting, we discussed how this big DRG paper was going to be that initial use case. I think initially we were aiming to get that paper done by the end of the month, but I didn't know if that timeframe changed.

Will: I think things are moving along. I think Shams has shared a preliminary object with many of the authors. I don't know that that's the final one and maybe Shams can comment on whether the metadata needs to be pruned a little, but I think that was already shared widely across the groups I don't know if that's reached the U24. But that's essentially what the structure is going to look like with some modifications to the metadata once it gets finalized.

Joost: It'd be super helpful if someone could point me to where that is. Because part of this is being able to play around with that data and to see what works and what doesn't work. We're definitely ready for it.

- *AI: Shams to share preliminary object with Joost*

Rob: Yeah, I feel some urgency in getting something public-facing soon. Obviously. Preprint, getting this paper published is going to take who knows how long, but when you get the pre-printed, we need to have something public-facing with the special emphasis on being able to see progress and kind of accountability.

DP: I want to double down on that. So, anything on a preprint server which is public-facing, would be the way to go. We really need those.

Joost: In order to do anything publicly, we'll need to be able to publish and make the data publicly available.

DP: Even for a preprint server?

Joost: If we want to make a dashboard publicly available, then the data needs to be publicly available. We cannot have data be private while the dashboard is public because the dashboard would actually be publicly available at that point. We could focus on creating this dashboard for the consortium initially, and then link through that, or outline that in the preprint. And then we can publish that data, including the dashboard so that it provides access to the public data once the paper is published, or once the publishing is in process or accepted.

DP: Yeah, that'd be totally fine.

Maryann: Yeah. And we do that in another consortium. We say that it's being developed, but it is not available. They generally know the data sets, analyses, and other things are going to undergo change during peer review and a lot of people don't want to finalize them until it's done.

Rob Gereau: Shams or Will maybe you could comment on the notion that we might use data from the existing harmonized Atlas. Is the data structure going to be similar enough that we could have a way to have something that's public-facing and say this is built on an existing data set, and the data that's coming from this preprint and these publications will be incorporated?

I think we just want to have an update suggesting this is what we're going to do and these are the incoming data sets that we're seeing down the. We need the people who are trying to advocate on our behalf to be able to see what's coming because it could take a long time to get the big DRG paper actually published.

Will: Yeah, I feel like the data structure is near….We've been working hard internally across all the centers to discuss how to do annotations and finalize the way to actually generate the matrices. And I think that work has largely been done and agreed upon now. So that counts table can be shared publicly. The raw data underneath it gets a little bit more complicated. But that's also pretty easy to solve. We all have that data. We just need to sort out how to share that if that's what's important. But we could get a public-facing counts table display of all the process data within hours. The hard part is really getting all the final figures put together and writing a paper in a way so that it's polished and high quality.

Joost: That's great. It takes time to build things. You cannot turn an entire web app, like what we've seen before. Like, that's not something that you build in a day. It takes a while to build. So the sooner we have access to the data, the sooner we can continue to iterate on how our developers can play around with that. And then, like we agreed before, it's an iterative process.I do think that we can probably put something out publicly early on…like we have the whole framework and we're doing dashboards for SPARC and for RE-JOIN as well. I just want to make sure that this is something that's meaningful.

Will: I completely agree. I was mainly suggesting that if we wanted to put up a preliminary version of this using the same interfaces we've used in the past that's actually pretty straightforward and would just take hours, but I agree about getting a really polished final product that we're happy with.

Joost: Yeah, and also thinking about, like, how do people actually want to use that? How do you make it fast? But we're ready for the data and we've been working on the overall infrastructure and how we do that on our side. So I'm looking forward to getting the RDS file and seeing if we can seamlessly replace the files that we currently use, which are the ones from the PainSeq RDS files and go from there.

*The 7/8/25 Data Subcommittee Meeting was canceled. AI updates as of 7/11/25 are listed below:

~~2025-07-08~~

Action Items from 6/13/25 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
34	Explore if there are any remaining Strides Accounts for PRECISION	Sam	TBD	
43	Have Internal DCIC Meeting to discuss Single Cell Visualization Tool	DCIC PIs and PMs	6/23/25	6/17/25
44	Provide investigators with an update about next steps for Single Cell Visualization Tool	Sam	6/30/25	6/30/25

7/11/25 AI Updates:

- 34) DCIC and NIH briefly talked about STRIDES account and next steps during their 6/24 U24/DCIC meeting. The AI from this meeting was that Joost would send Julia an email about the STRIDES account and that she'd help him find the appropriate HEAL contact to answer STRIDES questions.

- 43) Completed: The DCIC discussed both long and short-term goals. They were particularly interested in ensuring that the V1 of the visualization tool will be ready to coincide with the publication of the big DRG Paper.

- 44) Completed: Sam sent the Data Subcommittee Follow-Up Notes and Action Items from 6/13 out on 6/18. The note mentioned that there are multiple internal conversations occurring regarding the Single Cell Visualization Tool and mentioned that we're considering some ad hoc visualization conversations and invited investigators to sign up to join conversations by adding names to the recently created Visualization Feedback Group File. Sam met with Xianjun and Ruifeng on 7/8/25.

2025-06-13

Attendees: Jyl Boline, Diana Tavares Ferreira, Ayesha, Barbara Gomez, Bijesh George, Bryan Copits, DP, Will, Dustin, Guoyan, Hanying, Himanshu, Huasheng, Huma Naz, Ilias, Julia Bachman, Jyl, Kevin Boyer, Maryann Martone, Peter Jin, Rachel Weinberg, Selwyn Jayakar, Shams, Sue, Xianjun

Agenda Overview

- Action Items/Challenges
- Center Updates:
    - WUSTL General Update (Peter Jin)
    - UTD General Update (Diana)
    - Harvard General Update (Xianjun)
    - UPENN Update on unsupervised approaches to detect neurons (Hanying)
- Single Cell Visualization Discussion
- Updates

Meeting Notes (6/13/25)

Center Updates

WUSTL General Update (Peter Jin)

Peter Jin: The main update is we submitted the RFC1, short-tandem repeat paper. In this paper we can replicate first the RFC1 recessive repeat expansion in the Idiopathic peripheral neuropathy patients, as well as a new finding of association between the monoallelic RFC1 short-tandem repeat with risk for developing IPN.

We received pretty positive feedback and we are working on a revision and hoping to resend it back to *Annals of Neurology* in about a month. My group is currently working on the gene burden analysis using the STAAR pipeline, as we proposed earlier.

One big step is that we really want to integrate the genetic finding from this burden analysis with some single-cell ATAC-seq and RNA-seq to see how those new risk genes affect the gene expression over time…see what cell type is enriched. And hopefully we can have some collaboration with you all if you are interested.

Guoyan: Kevin will discuss how he's been working really hard working on the data submission with Anka. We had our first draft for the first DRG single-cell multisome paper, so we're hoping to submit by the end of July and then we're working on finalizing the second comparing arthritis versus controls and now trying to work on the manuscript.

Kevin: Anka from the curation team t wanted to emphasize the standards that the U19s have put forward for data submission have to be pretty much followed to a T in order for them to accept our data submission. This includes the condition that the column names have to match required information that the U19 previously agreed to acquire. She emphasized that they're not trying to be difficult, but wanted to stress that they're just following the standards that the consortium has put forward. Luckily, the SPARC curation team has been very helpful with that process.

Another update is that I'm finishing up the initial QC for Three Pain Conditions Paper (Back pain, fibromyalgia, and rheumatoid arthritis). We have 27 total samples between the controls and the pain conditions and we'll hopefully get started on downstream analysis and are working on a manuscript for that.

Sam: The metadata standards are updated. The newest development is the controlled vocabulary for non-answer values. The Curation Team talked with several people to reach an preliminary agreement on this element, but they're also getting feedback on if there's difficulty gaining adherence to the controlled vocabulary. Another thing that I recently spoke to the Curation Team about was Subject IDs and how they're sometimes used differently within the U19. So we need to just ensure that they're coordinated in advance, and there's a way it's summarized so the curation is aware of that.

Anka is happy to talk about this in greater length, but in the meantime, she provided a helpful slidedeck she previously created for linking the same subjects across multiple datasets for pigs. I think that was kind of the framework they were going to use to create a new resource, so it's just something to keep in mind as working within your group.

JyI: This situation arises when you guys are using samples or collecting data from the same subjects. Essentially, we need to be able to track the subjects across those different experiments. So, anytime that you modify the subject IDs, we lose that ability for machines to read that. Anka and Marlena caught that when they were working with Kevin's group and were able to ask whether this is actually coming from the same subject. It's basically you maintaining the same subject IDs across experiments, and then we have ways to link different data sets together so that people know that they're coming from the same larger project or experiment. The curators are also happy to help and realize that a conversation is sometimes more helpful.

UTD General Update (Diana)

Diana: I'll provide today's Data Core Update**.** We've generated some data for C2 DRG that includes single nucleus sequencing, bulk RNA sequencing, and Visium. This data has been analyzed and most of the data seems to be uploaded to SPARC and are working on the metadata in curation.

The *Single-Cell Characterization of Human C2 Dorsal Root Ganglion Recovered from C1-2 Arthrodesis Surgery: Implications for Neck Pain* paper is also currently in review. It's posted on *BioArchive* and we're trying to upload the dataset to SPARC to coincide with it being published.

We've also generated a lot of data from DRG with the bulk proteomics. The initial dataset with 8 DRGs has been recently published in *PAIN*. And I think we have an additional 42 samples that are being analyzed. I don't think we have a timeline for publication.

Then for the single nucleus RNA-seq, I think for UTD, we have around 85 DRG samples. The data upload to SPARC, metadata, and data analysis are in progress. I think we have some plans from our center for individual papers, particularly looking at the diabetic neuropathy samples, but I think some of this data will also go into the big DRG paper.

We also have a separate dataset on DRG. I believe the Spatial (Visium)] Transcriptomics is the thoracic vertebrectomies in which we have both Visium and single nuclei data, and that data analysis is in progress. We also recently published spatial data related to the Nageotte nodules and Visium sequencing and uploaded that to SPARC. I think the plan for thoracic is to publish by the end of the year and work on the metadata and data upload to SPARC as we go.

Then the bulk RNA sequencing data LBP (low back pain) tissue, which includes disc, nerve, and other parts, is being analyzed and the protocols, metadata, and uploading are in progress. We're hoping to at least have initiated the publication process by fall.

And then for the spinal cord we have single nuclei RNA-seq data from 19 spinal and dorsal inventory horns. The data analysis is in progress and I think the plan is to initiate the publication process by the fall.

More recently we started generating Xenium datasets. I don't currently have the numbers here because there are a lot of data sets that have been generated for both the spinal cord and the DRG and the data analysis is in progress and I don't think we have any plans yet for publication.

Diana: For challenges and successes, I think the biggest challenge has been the Xenium segmentation. I've included this under the successes too because we have been talking with other centers a lot and making progress. It sounds like 10X is incorporating a better segmentation algorithm into their pipelines that actually works for DRG neurons. So we sent

them some of our samples and it does seem to work relatively well. So they will release this in the next couple of months.

Another challenge is integrating the single nuclei data with spatial data and finding the best approach for that integration. Another big challenge is the volume of data that we have to upload to SPARC and how to fill out the metadata. And then Ayesha and everyone here at UTD has been doing a really good job of keeping up with that, but it's been challenging to internally create streamlined processes.

Then our current area of focus has been exploring the best way to meaningfully analyze our data and integrate all our data sets.

Harvard General Update

Xianjun: This slide lists the number of samples we collected so far for both Project 1 (post-mortem brain) and Project 2 (Surgical Tissues). I'll talk about the samples we sequenced. So, the overview on progress here is we continuously get a new sample, and then reiterate into our pipeline for sample QC and concordance checkup and refresh our clustering result.
So this is still ongoing. So far, the analysis, based on these 42 samples of surgical samples, include the cell type annotation, differential expression between different segments per cell types, and also, we compare different pain levels (high pain, moderate pain, or low pain) to healthy nerves.

We also identified the ligand receptor interaction between the DRG neuron to the non-neuron cells. I'll quickly go through some slides we have in terms of cell annotation based on these 42 samples that we get and then we compare the self-proportion of each type of segment between different segments of different types of tissue, like neuroma or proximal to neural nerve.

And also the marker genome across different clusters, and the DE between different conditions in each cell type. An interesting finding is fibroblasts have a lot of up-regulated or down-regulated genes in the neuroma versus the nerve. We also look at the interaction between ligand receptors between different pain levels….

For the ongoing work: we are converting to the Version 2 metadata and then there's a lot of downstream analysis that can be done, including integrating and comparing with mouse data and LR interactions, but we are still internally discussing new stuff within this data.

For example, other components from the data centers, like spatial data brain donor data also need to be integrated here….And many other analyses for the RNA-seq data, not just expression, but also could be splicing, etc–those things can also be included in the analysis plan. Then the proteomics data isn't delivered yet, but it's going to be received soon. So we'll integrate protomics data and RNA-seq data. And then there's also whole genome sequencing data also being processed. So, I think our team needs more internal integration here before I present better to the other centers here.

UPENN Update on unsupervised approaches to detect neurons (Hanying)

Hanying: I'm from Wenqin Luo's Team and from the Mingyao Li Lab,and I want to give an update on our current effort to explore the unsupervised approach to detect the human sensory ganglion neurons. So our team uses 10X Xenium. This is the general workflow. And from Xenium we can get the transcript image, a nuclear staining image, and also the H&E image.

So here you can see it's a transcript image of two genes, the UCLH1, which is a neuronal marker in purple, and FABP7 is a glia cell marker in green. So this is a spatial distribution for those two genes and if we zoom into a single neuron, we can see with the transcripts and the H&E image together, we can manually annotate this new realm, boundaries.

Xenium also has automated software to detect the cell boundaries, but it often fails for our large neurons. We're not sure why, but it often segmented a single neuron into multiple smaller fragments. So, we can't use this for our downstream analysis. And although we can do manual annotation, it's really time-consuming.

Hanying: So we tested several other popular methods, including Stardist, Cellulose, and Yolo V8, but I don't I want to cover them too extensively today because they're all supervised learning-based methods…we need to provide enough manual annotated neuron boundaries as training data to those models for them to generalize and predict accurate segmentations on those new unseen data and the manual annotation is very time-consuming and labor-intensive.

So we want to see if there are any unsupervised approaches to detect these neurons and started by building a DBSCAN pipeline. DBSCAN is a clustering algorithm that can partition the data points into clusters based on their distance from other points. In the first plot, you can see that each point represents one usage, one molecule. And, I used DBSCAN to first cluster those molecules based on their coordinates and then I can calculate the convex hull of each identified cluster and treat this as the cell boundary–as this red line shows here in the left lower corner.

For DBSCAN there are two key parameters that can be adjusted. First, there's Epsilon which defines the maximum distance between two points for them to be considered as neighbors. Second, there's MinPts, or the minimum number of points that are required to form a cluster.

So this is a simple example from a 500x500 micron region from a DRG sample. You can see the performance is highly dependent on the Epsilon and MinPts we used compared to the ground truth manual annotation.

Our group is currently preparing a manuscript that evaluates the performance on our DBSCAN-based pipeline across different combinations of those two parameters. We manually annotated 4 DRG and 1 TG sample, summarized our results, and gave some suggestions on how to select the two parameters to achieve a higher accuracy for new data.

So another supervised method I recently started exploring is based on the HE image only. We know that human sensory neurons are super large compared to other types of cells– ranging from about 20-100 micron in diameter.

We found a recently published foundation model called UNI. It is designed to provide a high-quality visual representation of those Python images. It typically takes 16x16 pixel image patches, and then it can produce…like the image feature embeddings containing both the global and the local image information. So, we rescale our HE image to 0.5 micron x pixel. And then I

used a software overlap called HistoSweep to extract those high-quality 16x16 super pixel regions from the HE image. Then we performed the PCE clustering using the embeddings the UNI based clustering model generated. And here are the results based on the UNI embeddings.

So if we zoom in we can see this is a 500x500 micron region from DRG. So, those are the manual annotations of the neuron boundaries. This is cluster 17 from the previous pages. So it looks like Cluster 17 captures the location of those neurons in some sense.

My next step is to further explore how to decide the boundary based on these cluster results. I also want to see if changing the resolution can help improve the results (ie doing like .25 x pixel which would be 4x4 micron per super pixel).

Will: Like Guoyan mentioned, I think one of the things we all are struggling with is segmentation. That, depending on the cell and where and how you cut each cell, you're going to see the satellite glia on top or below, and the exact boundary is always a little bit challenging. And this is probably something that could be trained with a lot of iteration, but I wonder as a group if we want to just throw out cells that are blended or if there is a certain amount of contamination we're comfortable with. Because I feel like this will introduce a little bit of variability in our segmentation approaches based on where you draw that dotted line, and how the cut of each cell in a given section occurs.

Hanying: So in that manuscript, we also have the other supervised methods I mentioned, the StarDist, Cellulose, and Yolo V8. I think it evaluates the image…. So because for my DBSCAN, I can only use UCLH1 Transcript location for segmentation, but for the other supervised method I can generate a transcript image that includes both the UCLH1 and FABP7. And I think the results are much better than using HE-only images.

Single Cell Visualization Discussion

Will: Joost and have talked about this offline too. One of the things that we're putting together now is an integrated data set across all four centers of the single-cell RNA sequencing data and for the DRG. It'd be great to think of a way to very quickly and cohesively present this data on a website for people to be able to browse and interpret the data. So I think some of those things many of us have separately done in our own groups others have done really well, like the Allen Institute or CZI are actually better than some of the stuff we're all using. And I think it'd be great to think of ways to leverage some of your expertise to take the best from all the places and put it together to share our data as.

Joost: I agree 100%. I think that as we are doing our milestones for our side of our U24. This is the thing that I think we should focus on. I previously showed a proof of concept (see Painseq Demo Sample) in taking your PainSeq dataset embedding that in our platform so that we have a UMAP viewer that renders all of the points and that is interactive.

The important step now is to find out what the next steps are with the understanding that we can't build an equivalent of the Broad Institute or CZI as our first pass. I'd love to get a very concrete deliverable that we can work on, and set a roadmap together that probably includes things like, can we sketch out exactly what the first viewer is that we want to view? Can we define exactly what the dataset is and how we format the data? Like do we create the UMAP coordinates prior to putting that on the platform or after on the platform? And what are the interactions that as a user we'd have?

Then I think on our side we can make some sort of estimate that'd take and build that as part of our milestones for our award. Now is the time to figure out what a first success means, like document that in a statement of work, and then we can work on that together. So I think we should get a smaller group to get some agreement on what that first viewer would look like or the first component of the dashboard would look like. Then we can implement that and iterate. Then ultimately, like over the next year my expectation is that there are different types of viewers that we want to see and different types of use cases that we'd want to render. The first thing we did was look at Painseq, but I also want to better understand other case studies too.

Will: I think most of us have a similar minimum feature set in mind that we all use and agree with. We may have to make more artistic choices about color schemes and dimensions somewhat artistic, but the biggest research item that I would pose would be which of the available viewers would be the most straightforward to implement. Like, would it be the CZI approach or do you have your own? Because I feel like the scalability of the current method isn't really very good.

Joost: Yeah. So we built our own version for the UMAP viewer (see Painseq Demo Sample). So we're able to render your 300,000 cells in a UMAP viewer on the web in a dashboard interactively. So that was my proof of concept, but it's really a proof of concept and I'd love to work on with you now is how do we now turn this into a dashboard that is either useful if we make it publicly available…or is this something that maybe, as a first step, we make available only to this consortium…which is actually something we should make a decision about. So there's a couple very practical things that we should jot down in a document so that everybody has the same expectations.

Shams: I think 95% of what people will want to do, including me, is going to be looking at gene expression profiles across these cell types. So the fact that you're able to get hundreds of thousands of cells into this visualization is great. But I think, if we're gonna talk about next steps, and maybe this is for the smaller working group that you'd suggested–something like Azimuth, where people could take their new data and upload it and just anchor it onto an Atlas I think would be a real help for the field.

Joost: Yeah, you bring up two important components that are actually important. One is do we build something for the rest of the community vs do we build something for this consortium. The second thing is do we build something that visualizes the data in here or do we build a service-slash-resource where people can look at their data in context of what is developed within this consortium.

Will: I think for us just getting it rolling it out internally over the coming months would be a huge first start. I think a lot of us roll out our own work, just to kind of beta test it internally. My suggestion would be to try to implement something for us internally to start working with our own data as a consortium. Then as we start to notice the bugs, we can adapt it and then post it live to the community shortly thereafter. I also think particularly with what Shams was mentioning, this "anchoring method" so that people can annotate their own data sets is a really valuable service. It does require a little bit of compute on your end, so I think that's one question I would just ask, but I think that's not a huge burden, but we'd have to clarify what the budget would be on that.

Joost: Yeah, we have time to set our milestones with the NIH. I'm 100% working towards that. I think that having some compute—like that -- is exactly why we have these negotiations. If that's

the most impactful thing that we can do, then we should do that..I'll work on a small document that we've got fulfillment of work for the first for the first, probably internal, but recognizing that the moment that data is public, the visualization should become public as well on the SPARC Portal, and so we'll develop it so it can do both. I'd love to get feedback on that, and we can rapidly iterate. And then maybe at the end, we'll do a section of nice-to-haves, talk to the NH about in terms of what kind of longer-term milestones and roadmaps that we can work towards over maybe the next year.

Xianjun: My group have been exploring this for a couple years already on different projects, not on this U19. So we look at Parkinson with millions of cells, I think anything working on RShiny won't work because it's too slow.

Joost: Yeah.

Xianjun: So we have to use Python, and we have optimized many strategies to make this fast and the load times quick. I think the protocol is useful for a biologist but not for our computational people. So we have to really talk to them about their use case. Like Shams mentioned a good example for gene profile. So, we actually look at our browser or portal from different viewpoints.

Number one: gene view–people search a favorite gene, then show the gene profile expression…between age, genders, etc. between different cell types, so the gene center view is really the number one used need.

Number two: cell type viewpoint. People look at favorite cell types, and then check every gene, like marker gene, top marker gene for that cell type or networks for that cell type. And then there's a layer view if you have spatial data. There's certain layers, like layer 1 white matter, for example. And then there's, like Snip View, if you do EQTL, you want to search a snip and link any gene linked to that SNP. There's also a couple more views. So, I think we design our protocol in this viewpoint so users can enter the subpanel based on the viewpoint and link to the same dataset. So I'd like to share our prototype to the group maybe in a couple weeks. But our goal is that anyone can upload their standardized format, like thread objects or H5AD format and then this browser will generate the page for the dataset immediately. So that's our goal.

Joost: That's awesome. I think it'd definitely be worth demoing and seeing that as this consortium.

Sam: If you're interested in joining these discussions….I don't know if this will be a working group, or we'll figure out how to approach this in the next step, but it would be helpful to know who's interested. We'll discuss this during some data Subcommittee meetings too, but if people have input or are particularly interested it'd be helpful to know. I also think we should Zoom out a little and go back to some questions like what is our goal and who is our audience and we also need to figure out which features we want to prioritize.

~~2025-05-09~~

**Given some changes in availability, we decided to cancel the 5/9/25 Data Subcommittee. Sam will provide some key AI updates during the second half of the week of 5/5.*

Main AI Updates:

AI40: HEAL Codes for Missing Values
May UPDATE: Jyl suggested new standards to HEAL and received some feedback from HEAL Data Stewards. The Curation Team recently finalized incorporating this feedback and will be reaching out to investigators to get their feedback.

AI41: Common Reference Genome
May UPDATE: Peter Jin is still working on facilitating a consensus for a common reference genome and wants to more thoroughly review the literature, including this GENCODE article.

AI42: The big DRG Paper Follow-Up
May UPDATE: PIs confirmed that they are making good progress on the paper and that they are targeting a May 11 deadline to clean/post data.

Action Items from 4/11/25 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
34	Explore if there are any remaining Strides Accounts for PRECISION	Sam	TBD	
39	Share Suzetrazine commentary link with DP	Will Renthal	4/16/25	4/17/25
40	Send HEAL codes for missing values to Data Subcommittee members	Jyl	4/21/25	4/21/25
41	Facilitate reaching a consensus for common reference genome version	Peter Jin	4/30/25	4/30/25
42	Send Follow-up Email to U19 PIs and PMs about next steps for big DRG paper	Sam	5/2/25	5/2/25

AI Descriptions (from 4/11 Meeting)
34) This item is a carryover that we will revisit at a later date
39) Will mentioned that the commentary about the role of human pain research in the development of Suzetrazine and its FDA approval was accepted at Neuron and should come out next month. DP requested the link.
41) Peter will leverage the relevant email thread that includes representatives from each of the four U19s
42) Complete

2025-04-11

Attendees: Maryann Martone, Jyl Boline, Xianjun, Aldrin Yim, Bijesh George, Bryan Copits, DP, Diana Tavares Ferreira, Guoyan Zhao, Guoyan Zhao, Hanying Yan, Hao Wu, HImanshu Chintalapudi, Huma Naz, Ilias, Iswarya, Kevin Boyer, Khadijah Mazhar, Marlena Pela, Megan Uhelski, Mingyao Li, Peter Jin, Rachel Weinberg, Shams, Sue, Joost, Wenqin Luo, Will Renthal

Agenda Overview

- **Action Items/Challenges**
  - Primary AI Updates
  - 26: Met with Allen Group and learned more about some of their lessons learned and gained improved understanding of their tools and how Allen Institute tackles data harmonization/ingestion
  - 33: Reshared Banner and Consortium Data Sharing and Publications Agreement
  - 35/36: Regarding Big DRG Paper, PIs suggested we block off a little time today to discuss DRG paper
  - 37: Shared Data Reuse Resource
- **SPARC Data Repository Approval for Journals**
  - Journal Policies vs HEAL Policies
- **Data Check Up**
  - Dataset Curation & Publication 2 Slides
  - Required Metadata, Complete Metadata, Missing Values, Data Dictionaries
- **Repository Compliance Email**
- **Genome Reference Discussion**
- **Big DRG Paper Discussion**
- PRECISION Data Subcommittee Roadmap

Action Items from 3/14 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
26	Meet with Allen Group	Maryann/Joost/ Sam/Jyl	3/21/25	3/21/25
27	Update PRECISION Cell Explorer Roadmap and Next Steps	DCIC	~~3/28/25~~	
33	Reshare Author Banner and Consortium Data Sharing and Publications Agreement	Sam	3/18/25	3/18/25
34	Explore if there are any remaining Strides Accounts for PRECISION	Sam	~~3/21/25~~	
35	Plan Internal U19 meetings about Dataset Use	Ted, Will, Wenqin, Rob	3/24/25	

36	Plan Network-wide follow-up DRG Big Paper Meeting	Ted, Will, Wenqin, Rob	3/31/25	
37	Share Data Reuse Resource	Sam	3/18/25	3/18/25

AI Updates:
26) Complete
27) In Progress.
34) Sam to revisit
35 & 36) In Progress. Time allotted during 4/11 Data Subcommittee Meeting for PI update/discussion
37) Complete

AI Descriptions from 3/14 Meeting
35) Each U19 will consider which DRG controlled datasets that they can contribute to paper flagship the Big DRG Paper, plan initial analytical steps, and finalize the list of representatives that should join the group meeting. They should also discuss what this whole package thing might look like.
36) After identifying the datasets, there can be meetings with key people from each center to start bringing the data together

Meeting Notes (4/11/25)

Action Items/Challenge Discussion

- DCIC and Allen Group had a good discussion about best practices. DCIC gained an improved understanding of their tools and how they tackle data harmonization and ingestion. We are having some internal conversations about next steps and working through a variety of dependencies, including STRIDES accounts.

- Sam shared some follow-up items including the Data Reuse Resource, the Banner, and Consortium Data Sharing and Publications Agreement and had a good conversation with Nele from RE-JOIN about the manuscript submission process.

- There are a few follow-up items for the DRG Paper and the U19 PIs thought it'd be helpful to discuss this during today's meeting.

SPARC Data Repository Approval for Journals & Data Check Up (Dataset Curation & Publication 2: Reality vs Theory)

Maryann: SPARC is the approved repository for PRECISION. Our job is to curate the data to the standards that have been agreed to by HEAL, SPARC, and PRECISION themselves and then to publish your datasets and make them available through the HEAL data platform and, eventually, the HEAL Knowledge Graph. Since SPARC isn't an approved repository for sensitive data, we have agreed that the transcriptomic data and the raw or minimally processed data is going to go into dbGaP. That's already part of our workflow, but we do take some of that data into SPARC and we make this available through the HEAL platform.

Maryann: Some journals have requirements about where datasets can be deposited. Several no longer say it must go in one of these repositories, but there are a few exceptions. *Nature Journal* said that linked genotype and phenotype data for human subjects should be submitted to a public repository with appropriate access controls and dbGaP and the EMBL resources are listed for that. In general, for other types of data, authors are free to deposit their data in any repository so long as it meets the criteria. SPARC is not listed as a recommended repository for these journals. Again, they've backed off from those lists, even though some like *Nature*

Scientific Data still maintain them, but we will go through the process of having those listed so that people don't run into trouble. If you do experience any issue, please contact me (mmartone@ucsd.edu).

Maryann: The metadata standard for PRECISION was approved in January. Documentation for this metadata standard is available on the NIH PRECISION Human Pain Network Resource Page. The page has all kinds of information about the process. For example, if you are going to be submitting data to dbGaP, you do need to consult with the curation team because we do need to make sure that identifiers for subjects and samples are synced across the two platforms. If you go to the section of the documentation on the PRECISION Metadata Standard, there's the PRECISION Metadata V2, which is essentially the data dictionary that describes what the variables are, what they mean, and how they should be represented in any type of collection instrument. It also includes the permissible values for filling it out. The PRECISION dataset Template V2 is a metadata template which is designed to upload into SPARC. The collection instrument and the upload instrument may sometimes be a little bit different. It is apparent that people are not looking at the data dictionary when they're collecting their data and that's very important because we agreed to it.

Maryann: Here's a snapshot of the data dictionary. These red boxes tell you whether the variable is required or not required. If you look at the metadata standard that was agreed to, there are 99 fields that are required, 4 that are recommended, and 2 that are required if relevant. There is a split whether you're collecting data from a living donor or whether it is a postmortem subject. For postmortem subjects, there are 10 mandatory fields that have to be put in. Required means that a data set will not be accepted if you do not fill out this particular field in the way that it is supposed to be filled out…..There are some little wrinkles in data collection versus what you submit to SPARC. HEAL requires some CDEs to be collected—zip code and other identifiable information that is not supposed to be shared with SPARC. SPARC is not approved for sharing sensitive data.

Maryann: It's the responsibility of the investigator to ensure that PHI is not shared with SPARC, although the curators do look for this…..Required metadata is not optional. For HEAL in particular, we do need to make sure that you stick to the metadata standard. That means not only collecting these variables, but collecting them exactly in the form that is present on the metadata standard. "I didn't collect it," isn't a valid excuse. However, there are times when values are legitimately missing. HEAL Data Stewards didn't have a standard around this but they did have a recommended set of values that we could use. We can supply these codes to you.

Maryann: Here are some HEAL and PRECISION metadata best practices. Something that the group can consider is a metadata 3.0 is whether all these 99 fields actually do need to be collected as structured metadata or perhaps some of them can be automatically extracted, like the imaging metadata….In RE-JOIN, we identified a whole set of critical pieces of information that needed to be included in the protocol, but they didn't need to be collected as structured metadata fields. The way that we made that decision was we asked, is this essential for search or comparison? (Ie. does it have to be structured or can it be provided in a protocol?)

Maryann: After collecting the required metadata, creating similar but non-standard headers is not allowed. Using abbreviations not specified in the standards, entering not permitted values and then assuming semantic equivalence between related terms is also not allowed…..As part of the HEAL project, you are required to submit a data dictionary for your variable level metadata. This is to help them understand how to query it. We have worked with the HEAL Data

Stewards and we will be allowed to submit a data dictionary on behalf of PRECISION, but that only works if you follow it.

Maryann: It takes time to go through the data submission process. While the curation team tries to turn things around as rapidly as possible, they are supporting multiple projects and deal with multiple submissions simultaneously. We recommend that as you start to prepare your manuscripts, think about the data sets that you are going to be describing or using and work with the curation This slide provides an example of some of the things that curation needs to do.

Rob: I think we should discuss this with the real world realization of what's happening in terms of what we're able to collect, like the surgical samples and things like that again it's variable and challenging. 99 items is a mind-boggling administrative burden. I understand the more the better, but I don't want to be perfect to be the enemy of possible.

Maryann: A lot of them are required by HEAL, but I don't think all of them are. The 99 might include these different instruments that have scores. I think it is very good now to do a check. RE-JOIN looked into what they could extract from a file header, for example, for an image. There's also a lot of fixation protocols that you'd like to know, but you're not necessarily going to search on it and that can be in a protocol because you've already provided it in the protocol in the first place. So for RE-JOIN, we check the protocols for the required variables to make sure they're there, but you don't need to enter them again. There's many reasons why a variable is missing and that's okay. What we don't want is that I never bothered to collect it in the first place because I didn't know I was supposed to.

Rob: The challenge being if you're dealing with surgical samples. And it doesn't render tissue or a sample unusable, and it shouldn't mean we can't publish it.

Maryann: Absolutely. So in that case, you fill in the missing value code for "could not be collected," "could not be collected," "could not be collected,"

Bryan: When we put together the requirements for the HEAL CDES, we understand that they're required and we attempt to ask every patient that. But we've realized that only about 50% fill all those in. So they're required in the essence of us administering the questionnaires, but it's not clear what happens when we don't get those answers from the patients.

Maryann: If the patient doesn't fill it out, there's nothing you can really do about it. So that's considered a missing value. And then it just has a code. So you have to fill it out. It's just you fill it out with, "I don't have it."**...**If it's a required variable, we are expecting an answer, just an NA is uninformative. And so there are sets of codes that HEAL pointed us to that we could use that have different reasons why something was missing. I think we should revisit standards now that people have tried it and see how we can optimize it. We'll get the codes to people, and if there are other people in your lab who are actually working with this, we're happy to meet with them and go over best practices. We do really try to minimize the burden on manual entry and if there are certain things that people can automatically extract from files, we can do that. So it's really just a matter of, I think, revisiting the standard now that people have tried to use it.

Repository Compliance Email

Joost: I received an email as a developer of repositories stating the NIH now reaches out to every repository saying that we need to comply with Executive Order 14168 to defend women

from gender ideology, and in order to do that we need to post the following information on the banner on the landing page that says that this repository “is under review for potential modification in compliance with administrative directives.” I've had a couple of discussions with NIH Program Officers. This seems to be related to the term “gender” in data sets, and so for other efforts we are actively working with investigators to remove the term gender from all of their public data sets. I thought this might be related to the HEAL CDEs, and I wanted to give everybody a heads up that this is something that is currently actively being pursued by the NIH.

Maryann: We've done a review….the one so far that we saw….SPARC uses sex because we've mostly dealt with animals, but there are things like “sex at birth” and other things that we might need guidance on. I think we'll do a review of that.

Joost: I don't think that right now there's something directly that we have to do. There's one data set that we are going to update the description of. I haven't looked much into PRECISION HEAL CDES, but if there is more guidance, then I will let everybody know. And I'm sure that our NH representatives here will help us navigate this.

DP: Yeah, I'll try to collect some more information. Unfortunately, our HEAL CDE office was heavily impacted with the RIF (reduction in force). We're trying to collect information from other programs, not just HEAL, other CDE programs.

Genome Reference Discussion

Peter: This topic is related to an email thread initially started by Mingyao at Penn about the reference genome everyone uses. GRCh38 is the one most people use, but some sites, including WASHU, use a different version. Moving forward, the first question is if there's a common interest to use a sort of reconciled version. The second question is if there are any collaborative projects that each U19 is interested in pursuing jointly because that makes the uniform use of the reference genome reasonable.

Wenqin: We were exploring data integration for Project 1 and Project 2 data integration and wanted to use the same standard. And as we were talking, we realized that if we want to integrate data sets across different centers for the big DRG paper, we wanted to avoid reanalyzing and remapping this data again and again every time to do that.

Mingyao Li: If we want to harmonize the data across the different centers we should have a consensus in terms of the genomic reference.

Will: I think that during the generation of this first consortium-wide paper that we're currently planning, we'll probably need to remap our data since we haven't been doing this in a consistent way. I don't know how we want to approach that…if it’s most recent or whoever is contributing the most data gets to pick. We actually did a lot of this for Harmonized Atlas, like comparing different referencing like NCBI versus ensemble. We didn't see a ton of difference like on the data output. So it's really just annoying to have to do it. I don't think it's going to change.

Shams: It would be great if we were all consistent across the board. I'm not attached to any specific reference that we end up using.

Will: Maybe we could get a list of the genomes that are currently in use and then pick one. Unless someone has a favorite version letter that they prefer.

Hao: The main reason I raised this point earlier is because GENCODE very recently released their V47 build of the genome reference I've linked a recent paper link main advantage of the latest build of the V47 is that they dramatically expanded the number of link RNAs that are potentially experimentally validated. If you look at their curves…over the years, the number of linked RNAs is only progressively increasing, but there is a huge jump for both mouse and humans in the latest release, December 2024. So the discussion from my perspective is really just do we want to use the latest V47.

Guoyan: I remember seeing a paper that compared the different annotations. And I think the conclusion is even though you have a great expansion of this link RNAs, the performance was actually not as good as the previous version. I can try to dig out that paper, but it's not like the more the better.

Big DRG Paper Discussion

Will: There's an email chain going on trying to gather the data that people want to include from the different sites. Three of four sites had shared data [*WUSTL confirmed they were finalizing the data and would submit soon].* I think the initial draft idea would be to nominate 1 or 2 people from each center to try to have a small group meeting to discuss the way to aggregate all the data and how to distribute all of the analyses. I think there's a lot of different ways to do it. And I don't know which is going to be the best one. So I think it'll be a pretty active process.

Hao Wu: Yeah, I think 10X pre-build is great. I think if everybody agrees to just use that, I think within the consortium, we can all have one common reference. The raw data will be uploaded somewhere later. People can choose to use different references to map anyway. I agree to be cautious and not to use the latest non-verified version.

2025-03-14

Attendees: Ebenezer Simpson, Kevin Boyer, Huma Naz, Jyl Boline, Megan Uhelski, Marlena, Aldrin Yim, Ted Price, Bijesh George, George Murray, DP, Bryan Copits, Illias Ziogas, Sam Kessler, Joost Wagenaar, Guoyan Zhao, Mingyao Li, Ayehsa Ahmad, Maryann Martone, Khadijah Mazhar, Hanying Yan, Peter Jin, Diana Tavares Ferreira, Will Renthal, Himanshu Chintalapudi, Nikhil Nageshawar Inturi, Julia Bachman, Rachel Weinberg, Xianjun Dong, Selwyn Jayakar

Agenda Overview

- Action Items/Challenges
- Pennsieve Analytics
- Painseq Collaboration
- Leveraging PRECISION Tracking Document
- Updates
    - Papers

Action Items from 2/14 Meeting

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
26	Meet with Allen Group	Maryann/Joost	~~3/7/25~~ 3/21/25	
27	Update PRECISION Cell Explorer Roadmap and Next Steps	Maryann/Joost	~~3/11/25~~ 3/26/25	
28	Send targeted microscope/imaging follow-up email	Sam/Sue	2/21/25	2/26/25
29	Create shared Banner feedback Document	Sam	2/28/25	2/26/25
30	Create New Zoom for 2/27 SC Meeting	Sam	2/17/25	2/17/25
31	Ask Rob if he's willing to Chair SC Meeting	Bryan	2/19/25	2/19/25

AI Updates
26: Meeting with Allen Institute scheduled for 3/21
27: Additional updates will be made to the roadmap after meeting with Allen Institute.
28: Complete. Shared Imaging Metadata Presentation, relevant Data Subcommittee meeting note section and the recording link (relevant conversation occurs from 19:50-35:40)
29: Complete. Created Banner and Metric Tracking Shared Feedback Document
30: Complete
31: Compete (SC Notes)

Meeting Notes (3/14/25)

Citation/Acknowledgement Conversation

The meeting opened with a conversation regarding when it's appropriate to cite or acknowledge PRECISION in a paper. For example:

- Guoyan is resubmitting a paper that previously didn't use any data from the U19, however, the updated paper will incorporate some more recent tools and methods
- Harvard is working on a clinical paper that was put together by their neuroma group that wasn't working directly on the aims of the grant. However, Will thinks the PRECISION group benefited from conversations.

DP: The general NIH Policy is that if the product is an outcome from your approved project in any form, then you must acknowledge it. If a product is partially developed through your grant, you can acknowledge it.
Will: I think the challenge with those definitions. There are situations when an investigator's work is funded by the U19 team, but they discover something while working on another project.

So, there are thematic issues or situations when some people partially funded by the U19 are working with reagents and supplies that weren't directly funded by the grant.
DP: Yeah, that's pretty borderline. It's difficult to judge because each case is unique. I'd ask you to make your better judgment there how aligned it is to the project and to network. It might not be to your project, but it might help the network or in future the data or the method or the approach or analysis or anything would be really impactful or can be recruited, aligned to the network then you could. With this in mind, it might be best if you can help distinguish these types of contributions in your annual report. You can create a separate list for list for those which were directly aligned to your project
Will: In our case, where there's a thematic similarity I think most of the PIs in our center have been taking a relatively inclusive perspective because there's a lot of shared knowledge that goes across the network that's hard to really quantify.
Ted Price: RE-JOIN is using their banner a lot and sending out emails for people to opt in or out.
Sam: We do have a publication banner. Names were previously compiled and then the banner was created based on the RE-JOIN sample that Maryann previously provided.

- **AI:** Reshare Banner and Consortium Data Sharing and Publications Agreement Final

Pennsieve Analytics (including Painseq example)

Joost: I'll provide a little overview of our current thinking and how we might leverage our platform's ability to run workflows and conduct data visualization. I'm hoping over the next year that we work with all of the groups to figure out how we can take the data that is being uploaded as part of PRECISION and consider how we fit this into the dashboard and SPARC Portal.

Joost discussed how we used a dataset that came from the Harvard group on Painseq in order to explain the difference between visualization on your local computer versus in a cloud environment. Pennsieve is a cloud-based environment that helps to optimize data for visualization in a web browser at scale. Pennsieve relies upon Github and AWS integrations and Pennsieve users can link to their GitHub account and get Github repository change notifications. Joost also explained our new analysis tab and the work we're currently doing with the Penn Immune Health, including creating pipelines for site data that now allow Pennsieve to automatically generate QC reports that get emailed to clinicians.

Joost: My hope is that we're starting with this now in a few small examples, like proof of concepts. I see a really huge opportunity here for the PRECISION team to figure out if we want to generate integrated dashboards or we want to have more standardized pipelines across all of the sites.

Will: A take-home message I got from our recent conversation was you need the data generators to think determine the way we're going to process and integrate our data, and then provide you with either a workflow to take raw data into that new model or an integrated matrix that you can help visualize for all of us. And I think that requires a little bit of coordination on our end with the data generators and analyzers.

Joost: You are the domain experts in these areas that rely very much on you to tell us how to optimize the pipelines to leverage data, and we can help you with putting it at greater scale.

Will: Many of us have been talking about trying to contribute some single cell RNA-seq data sets to what Ted is kind of called the Big DRG Paper. This could be a great opportunity for our

group to work through some of these slight differences in how we're approaching very similar problems in terms of a pipeline. Then after we reconcile that separately. After that we can come to you because I envision putting the object that we agreed to use on a scalable platform. Then users, including people from our group, can then upload new data sets and have that either anchored or integrated with the new object using this established pipeline.

Joost: I see some nodding. I think we can start thinking pretty boldly in terms of what we can do. I think one of the big success stories would be if we take some pipeline that runs in one of your labs and make that available as an app or as a pipeline that other people can run or we look at the differences between pipelines and through a mechanism like this can benchmark data sets against both pipelines. It will be very interesting to see the differences between results if you run two different pipelines. There's a lot of things possible once we have a small layer of standardization in how we run those pipelines.

Peter Jin: I like the idea, but I do think we should start with single cell as that's the data set everyone is generating.

Will: Great point. We haven't really articulated it in detail, but the hope for this big DRG paper will be a joint collaborative project across the different centers. Basically the deliverable would be a matrix that we all could agree to use, but also the workflows of generating that matrix…..Pennsieve requires using our own AWS accounts, and it's a lot more expensive for us to run it through your server than it is to run it internally through our internal pipeline internally. I do understand there are certain cases where it's great to have scalability.

Joost: You're 100% right. In the long-term, one of the reasons why we have this idea of a compute node and a compute resource is that we do have collaborators that have an HPC here. and so we are very much thinking over the next year that your local HPC might be a compute resource as well, where now you can select if you want to run it locally or remotely which could bridge those situations. This isn't going to be a replacement for your current workflows, but there might be some kind of pilot projects where it makes sense to try this across the different groups, and we can use this as a use case where we inevitably run into some walls which then help us improve what we're doing here.

- **AI:** Explore if we have HEAL allocation of Strides.

Big DRG Paper (+Leveraging PRECISION Tracking Document)

Prior to this meeting Ted, Will, Rob, and Wenqin discussed the big DRG paper and agreed that each of the sites were going to try to identify data sets that have single nucleus RNA-seq data sets that have been generated through the U19. They also discussed the need for each center to look at their datasets in order to ensure that they didn't cannibalize individual projects that are led by trainees

Ted:... We want to create a very robust, large, publicly available data set for so-called control DRGs. So it would be mostly individuals who didn't have a neuropathic pain state or things like that because that would be going into individual projects within the groups because I think all of us have different kinds of pain that we have been focusing on.

- **AI:** identify data sets and we want to bring to together

Ted: People often don't know whether they can reuse the datasets to do analysis. The issue is you can't publish the same analysis twice ... .But you can certainly– within reason– use data for different purposes and different kinds of analysis and publish papers using the data.

Maryann: Here's a resource that helps clarify when you can publish a data set and then publish multiple papers on it…..BICAN uses paper packages. They have a lot of individual papers and then often a joint analysis flagship paper. That would be something that everyone participated in, but all the individual papers were given an opportunity to be published as part of that paper package. These were deals that were negotiated with the publishers ahead of time as to if they were interested in receiving this package. So that's another way of balancing this need for these joint efforts in a consortium, but not co-opting the individual studies that are going on.

Ted: So the big DRG paper will be the flagship paper and we have a number of other papers that would have subgroup analysis and maybe additional data sets related to that. And we were shooting to submit these papers in the fall. I think there are a couple of journals that have expressed interest in receiving papers from the PRECISION Pain Network to individual PIs. I had a good conversation with one of the *Science Translational Medicine* editors and think this is something that we could pitch to them ... .Also, just because the flagship paper would be the big Human DRG Paper doesn't mean that the other ones have to be related to that. So in my conversation with the STM editor, they had a deep interest in Human-based pain mechanism studies. So if it was a neuroma paper from WUSTL or Harvard that went along with these other ones, that's in my view, perfectly good and probably even better for us to have a chance at getting a whole package here.

2025-02-14

Attendees: Sue Tappan, DP Mohapatra, Julia Bachman, Khadijah Mazhar, Bijesh George, Shams Bhuiyan, Peter Jin, Hanying Yan, Rachel Weinberg, Jyl Boline, Kevin Boyer, Maryann Martone, Xianjun Dong, Sam Kessler, Guoyan Zhao, Himanshu Chintalapudi, Ish Sankaranarayanan,, Diana Tavares Ferreira, Huma Naz, George Murray, Illias Ziogas, Mingyao Li, Will Renthal, Joost Wagenaar, Bryan Copits, Selwyn Jayakar, Ayesha Ahmad, Aldrin Yim, Ted Price

Agenda Overview

- Action Items/Challenges
- Reminders:
    - PRECISION_Data_Dictionary_v2.0 - Google Sheets
    - Provide Updates
        - Tools/Resources Submission Link
        - Bioinformaticians
        - PRECISION Shared Publication, Protocol, Presentation, Poster Tracking Document
- PRECISION Cell Explorer (sample: Allen Cell Type Knowledge Explore Sample)
- Imaging Metadata Update
- PRECISION Banner ("Experiments in Progress")
    - PRECISION Metric Tracker

- Updates
    - Papers
        - White Paper/s
        - Big Human DRG Paper
        - Validation of Human Tissues in Pain Research
    - Meeting Planning

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
18	Add Sensoryomics Tool to SPARC	Ayesha/UTD	1/21/25	1/24/25
21	Share Sample Key Metric Slidedeck Presentation	Ted	1/10/25	1/10/25
22	Send final feedback request for key metrics	Sam	1/20/25	1/17/25
23	Plan out next steps to facilitate future gathering metrics	Sam/Ayesha/Ted	1/27/25	1/27/25
24	Update/Review HEAL Data Dictionary (ie. pooled samples for multiple donors & anatomical locations)	Jyl/Ilias	1/15/25	1/15/25
25	Send Slack message highlighting recent metadata, HEAL Data Dictionary, and HEAL meeting updates in the #precision-metadata channel.	Jyl/Ilias	1/21/25	1/17/25

18) Complete
21) Complete
22) Complete
23) Complete
24) Complete
25) Complete

Notes (2/14/25)

PRECISION Cell Explorer (sample: Allen Cell Type Knowledge Explore Sample)

Maryann: At a recent PRECISION Cell Annotation Task Force meeting, the DCIC got feedback that it'd be helpful for Task Force members to better understand the DCIC's bigger picture view. This included learning more about what the DCIC is envisioning will be displayed (and the tools that we'll use) on the Portal. I met with Joost and the team and we started sketching out what we can make available on the Portal.
Joost: Over the past year or so we've had separate conversations about how we develop dashboards for the SPARC Portal. In the SPARC Portal we have this notion of SPARC application that lives here, and we're going to use the same type of interface where users can

interact with data on the Portal for PRECISION. We discussed more widgets that can show different types of information–ie. summary information about cells or visual information like different pie charts, plots, distribution curves, or the anatomical mapping. In the next several months, we're planning to build a separate application that will give us a way to interact and view different cell type information that is captured within PRECISION.
Joost: We now need to really figure out what V1 could look like. So part of that is tapping into the knowledge base that is being created. We've been looking at the Allen Cell Atlas to better understand how we can build our tool out and tie it to data that is uploaded/captured/presented that might not necessarily be published. So we'd love feedback over the next few months.
Will Renthal: Allen is a great model for the single cell interfaces and I don't think we need to reinvent the wheel. CCI is another good one. Ideas can also be floated in SLACK to get quick feedback.
Maryann: We wanted to give an informational update at this meeting because it's very difficult for the community to weigh in without having something to throw darts at. So the idea is to sketch out a proposal and then have you guys comment on what you'd like to see. We're setting up a meeting with the Allen Group regarding whether we can just use their tools and the CCI explorer as well. Once we have that sketch–which we hope to have at the next Data Subcommittee meeting–then it's something people can give us feedback on. But I like the idea of throwing out initial ideas if you have them.
Will Renthal: I think the use case has been relatively well defined single cell interfaces. For the ATAC data and spatial data, there are explorers that have published ... .I'm not sure if there's a standard in the field yet. Maybe we should open that up for others who have thoughts. But that might take a little more innovation on our end.
Joost: The key here is to do this iteratively. We aren't the Allen Institute or Jen Zuckerberg and need to determine how we can tap into resources with the funding that we have available.
Will Renthal: I think of V1 as being RNA only. I think that's a straightforward thing that we can at least get a deliverable up and show that we're generating data and share it with the community ... .Is there a way to include analytics on site, in terms of metrics to include for NIH and what people are visiting.
Maryann: Yes. We track that. SPARC by the numbers are calculated and distributed monthly.
Sam: I'd created a really rough roadmap of next steps PRECISION Cell Explorer (Background/Roadmap). The Cell Annotation Task Force will still exist and focus on this type of tool, and we will continue to provide meeting updates during Subcommittee meetings.

Imaging Metadata Update (2/14/25 Notes Continued)

Sue: Thank you for uploading the microscopy image modalities that you are capturing as part of your research. I reviewed some of your metadata and wanted to share a walkthrough of each of the sample image sets that were provided. The first set of data I looked at were .LIF files which are captured from the Leica microscope family. .LIF files are container files, so they can contain any number of images. To convert files to be open, one of the first steps that you have to do is disentangle them from that .LIF container in and of itself. So once you take your images out of the .LIF and save them independently from each other, then they can be converted.

Sue: You can use FIJI to see all the image metadata. And when I extract a file from the .LIF, and take a look at it in MicroFile+, I can also see that all the metadata is present except for the channel description. So, if I come to MicroFile+, it has a number of options here to select what type of operation you want to perform on your image data. It is a pretty straightforward tool. It's multi-instance. So if you have a number of different types of images that you need to convert,

you can contain them all within each folder for image type and then set up a conversion for the conversion for types of files and then that way you can process through them rather quickly….

Sue: So the .LIF files really aren't a problem. They have lots of great metadata in them. All that is remaining is the ability to add the target label. You may be able to configure that within the acquisition software as well–in which case then it contains all the data. The only problem with the .LIF file is that it is a proprietary file and therefore not everybody can open it natively. FIJI is a great tool set though.

Sue: If we look at Keyence This is a slide scanner that you're using to collect images of tissue sections on slides. As a TIFF image, it's missing metadata, but with MicroFile+ you can do things a little bit better. This is the image that was generated and supplied by the Gereau Lab. So you can see that it is an RGB image….So if we look at more information, you can see that there's no scaling and we don't actually know what each channel refers to. That's what I mean by target label. So, I could just convert that to a TIFF and add the required metadata or, using the different operation in MicroFile+, you can retain the individual channels so that you see each channel individually. I noticed that Keyence was a series of individual channel TIFs.

Sue: [see around 24:30 into video] And so the operation to generate that is actually pretty straightforward, because you would select this option here, “combine images into a multi-channel image or stack.” Here, we can generate both versions of the files at once and if I drag in these, it'll ask if I want to create a stack or a multi-channel image. I want a multi-channel image, so I'm going to say load. When you do, this is what pops up. So you'd put in the modality. Here, it'll be a wide field, you can choose what the pseudo color will be. And then here is where you add the information about what the label actually entails.

Sue: The last major component that I want to highlight is the ability to create a preset. So in situations where you’ve got different sets of microscopy modalities but you're collecting a lot of data for a given experiment that just needs to be processed, you can create a preset that allows you to say what each of the target labels are (the microscope, the modality, all of that), and then you can select that as the dropdown to automatically apply metadata at the time of conversion. That allows you to set up a device. So in this case, it could be your Retiga Electro and then configure the objective that you use at the time of snapping the picture as well as what each of the channels mean. And then all you need to do is bring them in, select the preset and hit convert.

Sue: Switching gears to another WASHU team. There's a Nikon set of files from Cavali and Zhao. These are ND2 files. These have all the metadata except for what the modality is–I'm assuming it's a wide field–and the target label. So you can put the target label. I inferred the target label based on the order of the labels in your file name, but this writes it explicitly into the metadata and allows everybody to know exactly what you're doing

Sue: The last group from WASHU, had montage data. So, if we open that in FIJI, it looks like this. It's a really nice image. The stitching is done very well. Even the flat field correction looks great. I'm sure you're using this to give an indication of what the ganglia look like at the time that it was nerve sectioned. Scaling, modality (although it obviously seems this is a bright field), and what the stain is are missing. Adding this information is really quite straightforward. This is an exported TIF image, it looks like. And so if it was from a slide scanner, either Keyence or an SVS file like Aperio. You can use that native file to load into MicroFile+ and it may already have the scaling information.

Sue Tappan: UPENN provided a nice metadata spreadsheet. And it looks like you guys have Aperio files. If so, you can drop them into MicroFile+, add anything that may be missing. Again, the most common thing that's missing is the target label and once you write that into your file, it will be preserved through any other information source. So any other application that you might be utilizing that data for. I'm happy to go over any questions or answer things in greater detail

Guoyan: So some team members from my Lab who are actually using the microscope aren't attending so I was curious whether you'd share this with instructions. For example, labeling the metadata that was missing from our project.

Sue: Yeah, that's why this is generally an overview. I would like to meet with anybody on anybody's team that wants to know how to effectively apply this as part of their pipeline. You also don't have to use MicroFile+. You're welcome to use whatever software you want. MicroFile+ does enable you to add the metadata that has been considered essential for the SPARC metadata standard.

Sue: What you should take away from this is that there's two steps at which you can ensure that your imaging experiments have the necessary metadata to allow someone else to understand your experiment. One is at the time that you're actually doing the imaging. And then the second one is what we're talking about here. So if you are able to configure your microscope and you get in that habit, I would consider that a best practice because then you just have less to do later on.

Sam: I recognize that not everyone who uploaded the microscopy files or works with them is not attending this meeting, but now that I have the people and the names of the people who uploaded to the different folders, I can create an email including Sue, timestamp the appropriate parts of this discussion, share the presentation, and suggest that if people want to meet, we can follow up.

PRECISION Banner (2/14/25 Notes Continued)

Sam: The PRECISION "Experiments in Progress" Banner was put together by each U19's primary PI and the DCIC PIs. During last month's Data Subcommittee meeting, we reviewed metric-related tools/templates that we could leverage. The group expressed a strong desire to show the ongoing consortium work occurring prior to publication. We created this PRECISION Metric Tracker based on an agreement on some simple metrics that are fairly easy to collect and track. The timeline of displaying the banner was accelerated given a paper deadline, but we still welcome feedback and would like to revisit the idea of more advanced metrics.
DP Mohapatra: The sample number looks very high and the way it's denoted can be interpreted in a number of ways so it'd be helpful to add some clarification. I see that it's a variety of tissues that are analyzed and some subjects have provided more than one type of sample.
Sue: We might want to ask investigators whether it should be subjects, tissues, samples, and methods. Like do you want to swap the order of tissues and samples to make that more explicit.
Maryann: DP, if I'm understanding, we just need to make it clearer what that number represents.
DP: Exactly. And I like the idea of bringing the tissue type first and then following it by samples. But within the sample, as well as the previous icon that is subjects, my suggestion would be to write participants versus subjects because now it's a more widely acceptable practice to denote human participants in human studies versus human subjects…because subjects can have some alternate meaning.
Maryann: Yeah, that's fine.

DP: There are no clear hard and fast guidelines there, but based on what's being now adapted more and more is human participants versus human subjects but, again, this is just a thought.
Maryann: Labels are easy to change, so we can change them to whatever we'd like. And if there's something that needs more explanation, I think there was a little hoverover or a little question mark to lead you to more information about what that is.
DP: Yes, exactly. Human participants, study participants, and tissue donors could be the labels. I know that tissue donors are the maximum numbers here and they necessarily did not consent to participate, but I'll defer to Ted.
Ted: For us, I would prefer a donor to a participant since they're deceased in almost all of our stuff. I think probably for the number of tissues we shouldn't include blood because this is not really about anything related to blood. It's about the nervous system. I know why we're using the blood samples but I think I would specify that the samples are nervous system samples and not use the blood….For the subjects, which I guess we can make donors…It could be donors and participants
Sam: Yeah, so I will immediately look at that number and then I can remove the inclusion of blood from blood cells.

Updates (2/14/25 Notes Continued)

White Paper/s
Ted: We wrote the White Paper. We decided to make it short and to the point and pivot from the Suzetragine approval a little over a week ago. Clifford sent what I think was a very nice to-the--point document to an editor at *Science* and we had no response. So Clifford and I have been emailing a little bit and we're going to try to submit it formally to *Science*. I'm not quite sure how we actually would do that, but I'll try to figure it out today. And then I was thinking if *Science* is a no-go, we should go to Nature Review's, drug discovery. I think we'd have a pretty good chance of getting it in there and they publish a lot of stuff like this.What do y'all think about that?
DP: Is it possible for you to send me, Jules, and Rachel a draft of that? We'll definitely keep it completely within us.
Ted: Yeah. Sam, can you put the Google Doc link in the chat?
Ted Price: And DP, I'll send you and Jules what Clifford said. Editors of science.
DP: Yeah, just send us a draft. I mean we cannot suggest edits or modifications or anything. It's absolutely just to look at it.
Sam: Related to papers, I think there was a Steering Committee meeting, so I just wanted to make sure we're all on the same page and can differentiate papers. I think there was a decision made to make two types of white papers. So on the agenda I listed out three papers (White Paper/s, Big Human DRG Paper, and Validation of Human Tissues in Pain Research). I think we decided there's two kinds of white papers–a shorter one which was just submitted and discussed and another longer one that I think we plan to discuss in the future. I wanted to clarify next steps, because I believe there's an upcoming Steering Committee meeting that the NIH added to the calendar for "a working meeting for a network-wide manuscript." So I was curious what the next steps might be and hoped to talk a little more about the goals of that meeting.
Ted: I'm going to be on an airplane during what we have scheduled as our next steering committee meeting. Should we keep the meeting and somebody else run it or should we reschedule?.....I'm not hearing anything from anybody so let's keep the meeting. So I'd suggest that Will, Rob, or Wenqin should lead the meeting. So why don't we include the PIs on setting the agenda. We're all discussing it internally over here so other people can speak for me quite easily…. The shorter one is done, right? And we hopefully will find a home for it soon. But we have the idea to do this longer form of that that we should figure out where that should go and who's going to lead it and how we're going to do it.

Meeting Planning
Sam: The final topic that I wanted to discuss was related to meeting planning. Given timing, I won't open it, but we created a spreadsheet similar to the one that we use for Human tissue Subcommittee meetings with dates. Joost is not currently here, but I had created a spreadsheet similar to the one that we use for the Human Tissue Subcommittee with just dates. And right now it's just dates. The idea that Joost had mentioned that wanted to pitch and get immediate feedback is that he suggests the possibility of maybe having some kind of alternating meetings where some are DCIC led and others are kind of more U19 run. He wanted to kind of also do some more active projects. So if that's an idea that people might be interested in, we can continue to kind of flesh out specific timeframes and meeting dates when people want to present, but I did want to get some immediate feedback if people liked the idea of alternating between the DCIC and U19s. The hope is U19s will have more opportunities to provide Data updates/challenges and we can prioritize the areas U19s are most interested in speaking about ... .Does anyone have any immediate feedback?....
Sue: I wanted to give a plug to have people sign up for the SPARC newsletter.
If you haven't and you're interested. We'd really appreciate it. We can let you know about the upcoming webinar series. Members of Dr. Price's team will actually be kicking off the series with the first webinar. And so if there are other folks on this U19 that would like to present their Research publicly, please do reach out.

2025-01-10

Attendees: Peter Jin, Ted Price Guoyan Zhao, Huasheng Yu, Aldrin Yim, Elle Mehinovic, Sam Kessler, Jy Boline, George Murray, Wei Feng, Kevin Boyer, Himanshu Chinatalpudi, Megan Uhuelski, Ayesha Ahmad, Xianjun Dong, Shams Bhuiyan, Selwyn Jayakar, Ish Sankaranarayanan, Rachel Weinberg, Bryan Copits, Huma Naz, DP Mohapatra, Julia Bachman, Wendy Dong, Diana Tavares Ferreira, Wenquin Luo, Ilias Ziogas, Ibrahim Saliu, Harsha Matwani, Sue Tappan, Bijesh George

Agenda Overview

- Action Items/Challenges
- Reminders:
    - Continue to think about Human-Specific Markers/QC white paper ideas
    - Remember to submit Tools/Resources Submission Form
        - Currently working on Sensoryomics
        - EnsembleTFpredictor
        - Robust principal component analysis (rPCA) for high-dimensional data outlier removal
- PRECISION Preliminary Numbers/Metrics
    - Samples
        - https://hubmapconsortium.org/
        - https://data.smaht.org/
        - https://portal.brain-map.org/atlases-and-data/rnaseq
        - https://portal.gdc.cancer.gov
            - https://github.com/NCI-GDC/portal-ui
- Updates
    - Metadata (PRECISION_Data _Dictionary_v2.0 - Google Sheets)
    - Papers

        - Validation of Human Tissues in Pain Research White Paper
        - Consortium White Paper
        - “The Big Human DRG Paper”

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
15	Create Human-Specific Markers/QC Shared Feedback Document	Sam K	12/18/24	12/18/24
16	Plan next steps for Task Force	Maryann	12/23/24	1/6/25
17	Remind U19s to submit their microscopy metadata samples	Sam K	12/19/24	12/19/24
18	Work with K-Core to Get Sensoryomics Tool on SPARC	Ayesha/UTD	1/10/24	
19	Share metadatav2 update	K-Core/Illias	12/24/24	1/10/25
20	Send SPARC Phase Two Emerging Scientists Call for Judges Reminder	Sam K	12/19/24	12/19/24

15) Complete
16) Doodle shared with Task Force Members aiming for January meeting.
17) Complete. Following samples received:
    - Penn:
        - UPENN Microscope Metadata Samples
    - WUSTL:
        - Project 2 - Cavalli, Zhao
        - Project 3 - Gereau, Copits
    - UTD:
        - UTD Microscope Metadata Samples
18) In Progress. Expecting to finalize next week.
19) Metadatav2
20) Completed

1/10/25 Meeting Notes

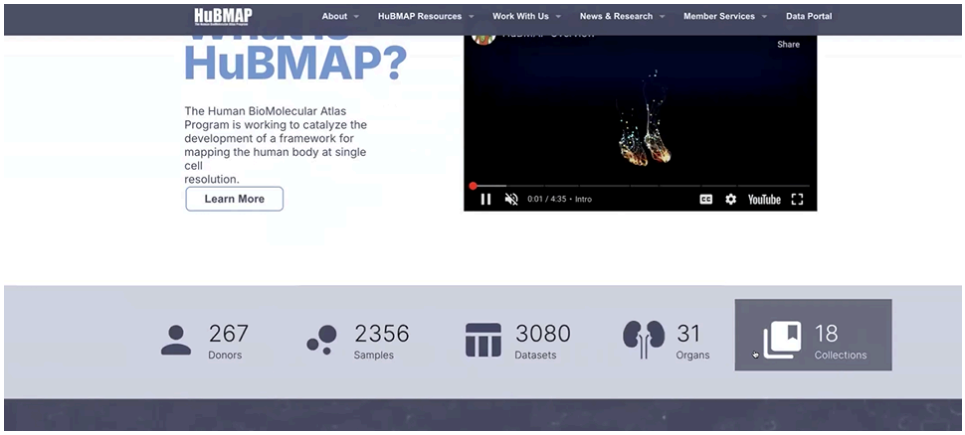
Action Items/Challenges/Reminders (1/10/25)

- Please continue to update Human-Specific Markers/QC
- Maryann and I have been in communication regarding the Cell Annotation Task Force and are aiming to have a January meeting to plan out next steps and set expectations. IWe want to also make sure it's integrated into other efforts occurring in this space.
- Received Microscopy metadata samples from multiple groups. Sue is currently analyzing
- Remember to submit Tools/Resources Submission Form.
    - Currently working on Sensoryomics
    - EnsembleTFpredictor

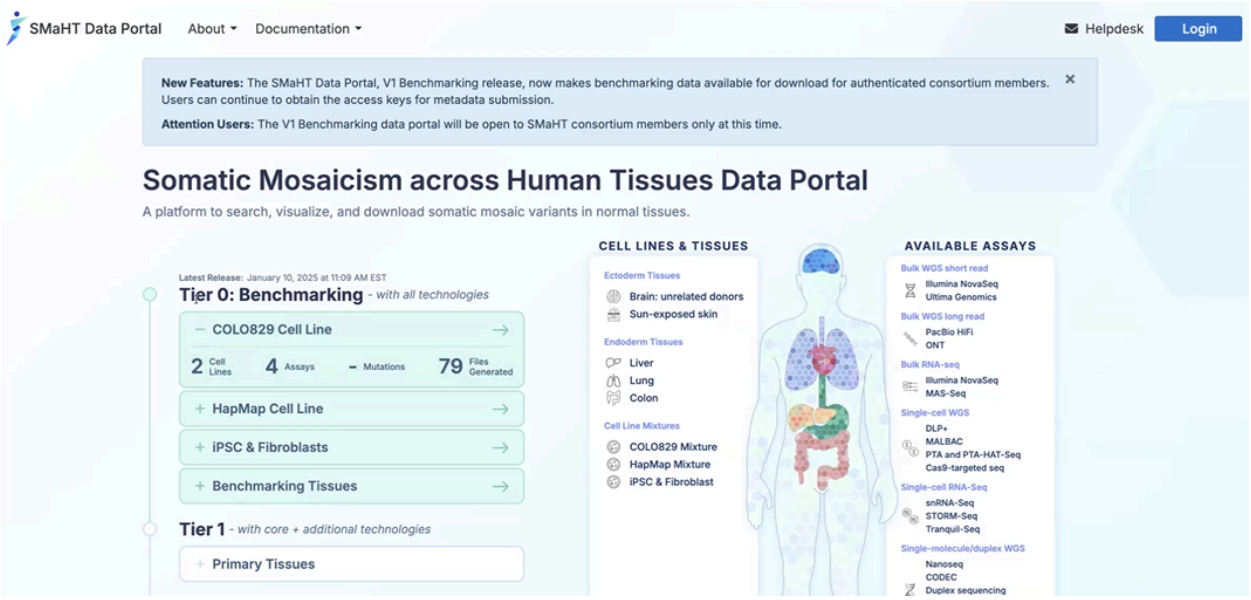
    - Robust principal component analysis (rPCA) for high-dimensional data outlier removal
- K-Core will provide an update on Metadata Dictionary. Metadatav2 has been finalized, and we'll provide updates after next week's meeting with HEAL.

PRECISION Preliminary Numbers/Metrics

Peter Jin: Today's meeting will focus on preliminary numbers and metrics that U19s are collecting/tracking. Given the complexity of the data that we generate, we're also interested in determining how best to share the metric information with the community. HuBMAP is a super simple example where metrics are shown in the middle of the web page, and they highlight the number of donor samples and some high level metrics.



Peter: SMAHT (Somatic and Mosaicism Across Human Tissues) Data Portal, is a new common fund program initiated in 2022 with a goal of generating mixed data and lots of technology across 15-20 tissues for around 150- 200 donors. SMAHT clusters the donor samples into tiers and displays the tissue based on the human anatomy. PRECISION probably doesn't want to use the same anatomy, but it's a type of format we can consider.



Additional templates suggested are:

- BRAIN Initiative platform (https://nemoarchive.org)
- GDC Data Portal (https://github.com/NCI-GDC/portal-ui).
- Allen Brain Institute which focuses on single cell data

Xianjun: All these options are very nice, but we should consider our goal. Eventually we want to have a portal that people can explore at the gene level. For example, you search for your gene of interest. You can see the UMAP or any special level change across different stages of disease, different brain regions, or different cell types. This kind of exploration will be part of this portal too, right? We don't want to invent two portals, but we want to do everything in one place.

Guoyan: This is probably hard when you're having bulk data versus single cell data. Those may be totally different interfaces to explore those things.

Ted: I really liked what Peter showed (SMAHT) with the body and how many datasets are available, but just want to ensure that we're still planning on using that kind of format through the SPARC portal. For example, you would go into a site, click PRECISION PAIN, and you would go into a body map that would say, "X number of datasets from all these different things using the body map that SPARC already has," and then people could dive in from there.

A conversation about searchability occurred, but it was decided that the more pressing need for PRECISION investigators was to define and post numbers to let the Pain Community know what we're doing. The DCIC will revisit this conversation.

Will Renthal: We're missing numbers in terms of donor numbers and modalities….We could do quarterly or twice a year, where each center sends you, "we've done this many donors of this tissue type," in something like five-column Excel sheet where we could post big numbers at the bottom to give the community a sense of how much data we're generating

Ted: I completely agree because it can take a long time to get these papers published. We have some slides including some numbers that I imagine are similar to other groups. It's important to try to get this up on the website sooner rather than later and get some of the datasets to be available in at least some type of way so that people in the community can start to use them prior to publication and so people in the Pain Community know what we're doing.

Jyl: We can get statistics up pretty fast. It's a matter of getting some standards across the group for what you guys want to share.

Peter Jin: I agree. We just need to make a consensus about what are the 5-6 or 10 metrics that we're going to put up there.

Will Renthal: The key metrics that come to mind are site, tissue type, modality/method, and donor numbers.

Decision: Start with a simple big number template similar to HubMap and then work towards something more sophisticated like SMAHT.

Ted: Here's the goals/metrics we have to date (see slide below), which I think we could probably simplify. I think everybody can put something similar into an Excel sheet. So we have donor and surgical tissues–I think it's probably a good idea to separate them overall–but for the front page, we could probably just put the overall number of donors or surgical samples across all the

centers. Then we're doing Visium and Xenium–we probably combine things in a single cell, single nucleus, or spatial–and then they can break out into other platforms. Then I know some other places have trigeminals and nerves. So we'd have DRGs, trigeminals, nerves. We don't have trigeminals or nerves, but we also have spinal cord.

DRGs

Tissue type	Sequenced to date	Total project goals	Minimum Metrics/Notes
Single nucleus sequencing (donor)	71	168	Most single nuc-seq data collected is from 10X kits, though some initial data is from Parse kits
Single nucleus sequencing (surgical)	30	68	10X FLEX Kits: 10% neuronal nuclei, 5,000 reads per nuclei, 500 genes per nuclei
Visium/Xenium spatial sequencing (donor)	16	16	Visium: 100 neurons per frame, 25,000 reads per spot, 2,000 reads per spot
Visium/Xenium spatial sequencing (surgical)	18	36	Xenium: 100 neurons per section, 100 median transcripts per cell, 480 gene panel
Proteomics (donor)	50	60	Completed samples were processed for bulk proteomics. Analysis metrics based on membrane protein detection

UT DALLAS — Center for Advanced Pain Studies, THE UNIVERSITY OF TEXAS AT DALLAS | SCHOOL OF BEHAVIORAL AND BRAIN SCIENCES

Ted: [*Below]* We have a fairly good number of samples, and then we also have some ATAC-Seq stuff that we've been doing that's not listed here on the DRGs, but we had that in our projects, so we need to add that.

Spinal Cord

Technique	Sequenced to date	Total project goals	Minimum Metrics/Notes
Single nucleus sequencing (donor dorsal or ventral horn)	19 DH, 19 VH	19 DH, 19 VH	Same as DRGs
Visium/Xenium spatial sequencing (donor dorsal horn)	8	8+ as needed to complete project objectives	Similar to DRGs, will report based on ongoing analysis results.
Spatial ATAC-sequencing (donor dorsal horn)	20	20	300M reads per sample, TSSe over 3, FRiP over 0.1
Proteomics	-	60	
Bulk sequencing (other lumbar tissue)	25	84	50M reads with 75 bp read length

UT DALLAS — Center for Advanced Pain Studies, THE UNIVERSITY OF TEXAS AT DALLAS | SCHOOL OF BEHAVIORAL AND BRAIN SCIENCES

Ted: So I was thinking that we could get a list of all the tissues–which I think we can do very quickly if everybody just puts together a similar slide deck or an Excel sheet. And then basically all the data core leads can just kind of check off on everything, and I feel like we could get that together pretty quickly.

Will and DP suggested that U19s are going to need to have these numbers in place by March mid-year review check-in anyway so we could plan around that.

AI: *Ted to share slides.*

AI: *Ayesha, Ted, and Sam to work to reach consensus on metrics and subsequently help facilitate the compilation of metrics so that the timeline aligns with March mid-year review.*

Metadata Update

Ilias: So we have a clean version of the spreadsheet we shared with you that now incorporates the adjustments that were made based on your feedback. And we have color coded the requirements of each field so you can easily find required ones. And we also included the measurement unit when this was explicitly mentioned in the description.

Jyl: Most of our conversation with the HEAL group won't impact you, but we'll be discussing permissible values and things like that. We agree that there should be multiple options when you're not able to answer the question, including, for example, N/A or you were unable to get to a question. Hopefully, we'll have a good answer for that after our HEAL meeting next week.

Guoyan: We have some samples that were pooled from multiple donors. Do you have some information about how to handle that?...It would've been from the experimental sample prep level…It was initially for trying to do some benchmarking, because the sample didn't have sufficient tissues, so we liked to put multiple tissues together and use a single pool to test the technology.

AI: *Jyl and Ilias will go back and check how to handle samples pooled from multiple donors. They will also revisit anatomical locations to ensure that they're clear and people are using them the way they are supposed to be used.*

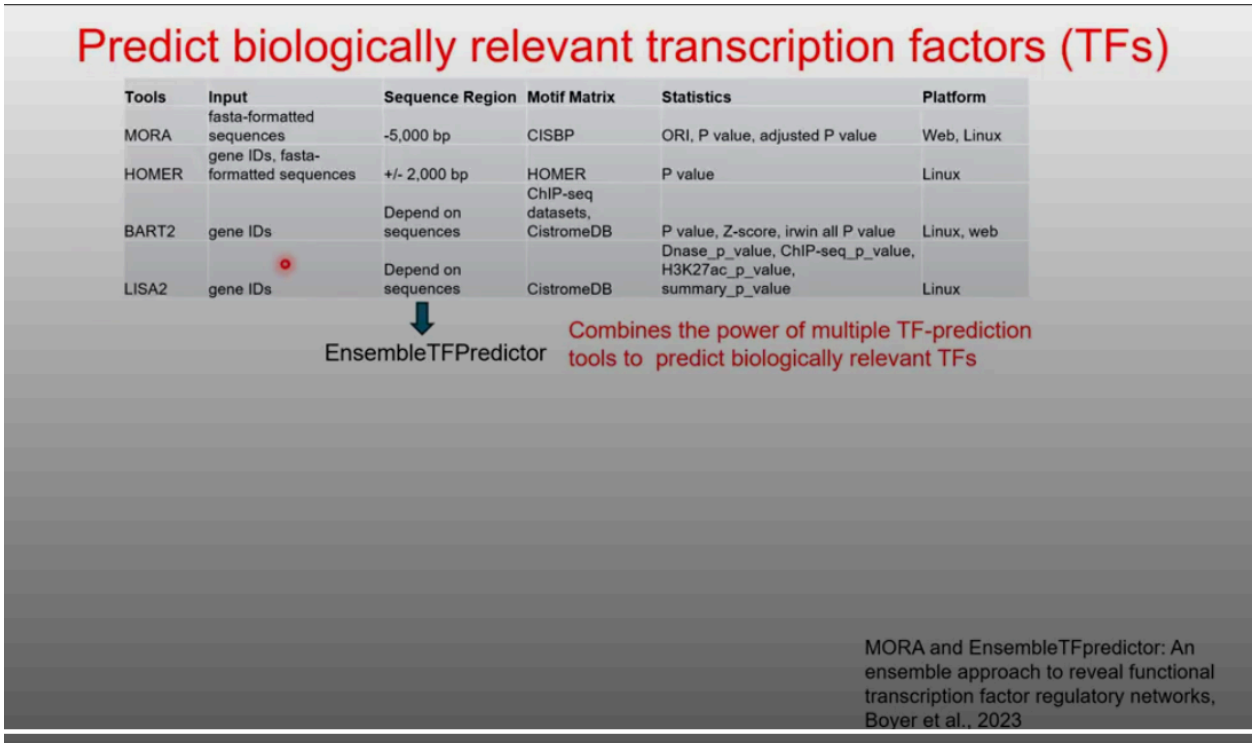
Sam: So after the meeting with the HEAL Stewards, someone from K-Core will follow up. We'll continue to leverage the #precision-metadata Slack channel for major updates and/or pending questions. At one point, I think we were considering creating a Slack channel for metadata-version2, but I don't think there are any immediate plans to create this anymore. Understanding that not everyone has access to Slack, I'll leverage K-Core's text and send it out to different listservs to ensure everyone is aware of this.

Paper/Manuscript Update

Ted: We've made good progress and talked about the manuscripts mentioned during the Steering Committee. We're going to plan to have a meeting sometime soon to get our ideas a little bit more firmly together on the larger white paper. And then I think everybody gave us the folks that they thought should be involved in the so-called big DRG paper…While we haven't agreed on what exactly that should be, I do think having a group to start discussing it will get us going in the right direction.

Tool/Resources Updates and Examples

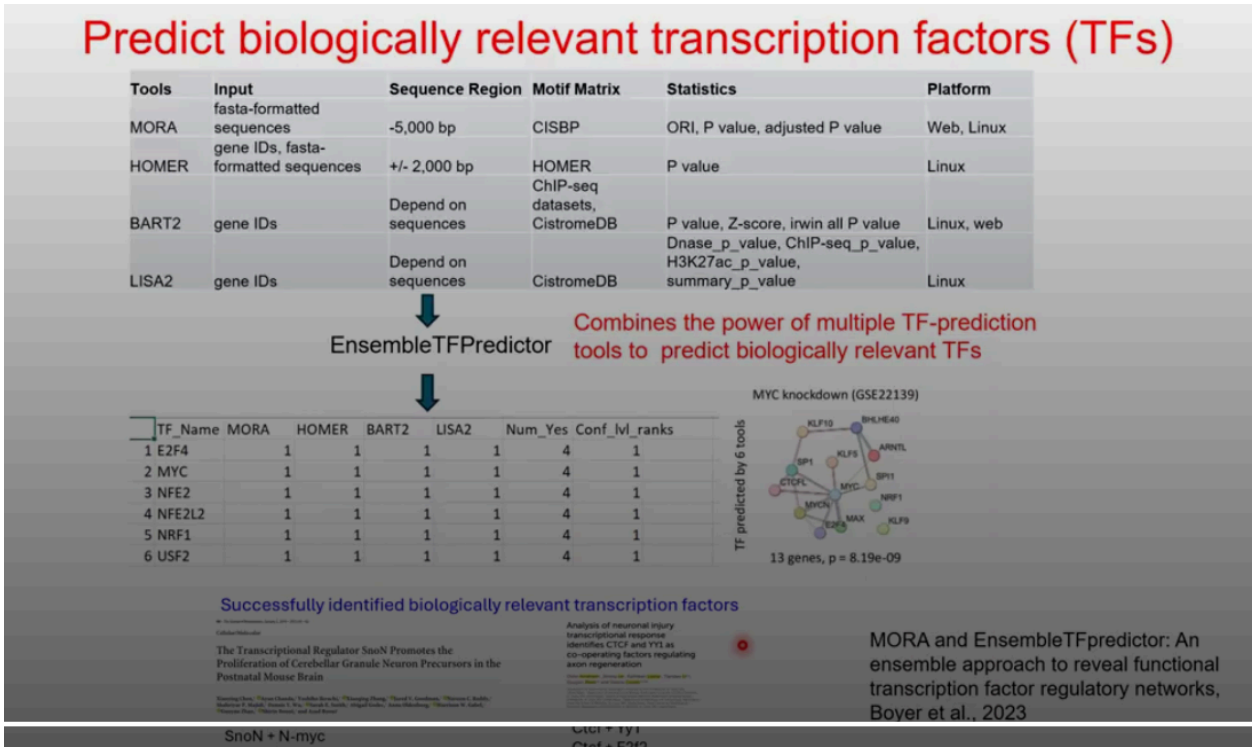
Guoyan: This *presentation* touches on some tools we've been working to predict biological relevant transcription factors. We developed a tool called MORA (motive overrepresentation index analysis), so if you find a bunch of DEGs or find that the single cell clusters have unique molecular markers, you want to find what the transcription factors are that might be regulating this set of genes. So we developed MORA to look for transcription factors binding motifs that are enriched in the set of genes you're interested in comparing to the background genes. And we compared our tool with six other tools, including in motif enrichment analysis. So, our tool performs better than Homer as well as two other tools that I didn't show here. Then the other category is BART2 and LISA2 which uses metagenomic data and explores which TF might be responsible, and the MORA is comparable to the other tools.



Guoyan: And we developed a second tool which combines the power of a different TFprediction tool to predict biological relevant TFs. The basic idea is that if we have TF being predicted by different tools which use different inputs, different count metrics, or different statistics, and they are predicting the same–like the transform factor being biologically relevant—it's more likely to be truly biologically relevant…..The EnsembleTFPredictor will combine different outputs and then rank them based on how many tools predict the TF. And we have used multiple different data sets to evaluate how well the tools are performing. One is using known targets for known TF. Basically, you perturb the TF by overexpression or knockout and we get the DEGs and we make the predictions.

Guoyan: So we know what TFs should be there. And also, we use the CHIP-Seq data from encode because we know those are the CHIP-Seq peaks for the transform factor being like IP. And we evaluated whether we are making the right prediction and we have been showing the top ranked TF, and they are all known targets. And also TF don't cooperate with other transform

factors together to regulate target gene expression. So we have been showing that if you look at the top TF, they actually interact from this network and interact with each other to regulate the target gene. So our tool has successfully applied in two different biological conditions. We made the prediction and with my collaborators were able to validate all the predictions correctly. I also recently had another collaboration. We sent a set of DEGs and we did the data analysis blindly and then found out the knockout TF being on the top list. So, it really works.



Guoyan: The second tool is called [robust principal component analysis for outlier detection](). This tool…because when you have outliers that use conventional principle analysis, they will scale how the cells were grouping together. Whereas in the robust principle analysis you actually can detect what the outliers are, and we have tested on simulated data, the actual data, and then it showed this PCA great….However, they actually can have like 100% sensitivity and 100% specificity in terms of identifying outliers in RN-seq, but it also can be applied to any kind of genomic data like CHIP-seq. And so far it has been like one of the fastest growing cited papers. It seems to be pretty useful to the genomic field. If anyone interested in using i'm happy to share more information

2024-12-13
Agenda Overview

Attendees: Aldrin Yim, Ted Price, Jyl Boline, Asta Arendt-Tranhold, Ayesha Ahmad, Bijesh George, Bryan Copits, Diana Tavares Ferreira, DP Mohapatra, George Murray, Hanying Yan, Hemant Mydugolam, Ilias Ziogas, Ish, Jeffrey Milbrand, Julia Bachman, Kevin Boyer, Mark Chavez, Nikhil, Peter Jin, Rachel Weinberg, Rob Gereau, Selwyn Jayakar, Shams Bhuiyan, Sue Tappan, Vivien Li, Wendy DOng, Ying Li, ZT

Agenda Overview

- Action Items/Challenges
- Reminders:
    - Review Data Sharing and Publications Agreement so we can finalize at SC Meeting
    - Continue to think about Human-Specific Markers/QC white paper ideas
        - SK to create a shared document
    - Remember to submit Tools/Resources Submission Form
- Sensoryomics/Visualization Environment
- Metadata Feedback Follow-up
- Updates

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
11	Schedule SPARC/PRECISION/HEAL Meeting/s to discuss alignment	K-Core/HEAL	12/3/24	12/3/24
12	Put out Requests for Cell Annotation Task Force	Sam K	12/5/24	12/5/24
13	Get Feedback on Data Dictionary	Ilias/Sam/Jyl	12/3/24	12/3/24
14	Get Feedback on Anatomical Terms from Human Tissue Subcommittee members	Ilias/Sam/Jyl	12/3/24	12/3/24

11) Program-level meeting scheduled to occur on January 13th.

- Alignment of HEAL SLMD and VLMD with the SPARC metadata model
- Expected timeline for PRECISION Data Subcommittee
- Alignment of metadata curation workflows between SPARC and Platform
- Clear guidance for investigators that prevents duplicative or conflicting processes dbGap, SPARC, and HEAL alignment
- Definition, adoption, and curation of pre-clinical CDEs between SPARC, Platform, and HEAL Semantic Search
- Governance of license-protected CDEs
- Use of SPARC publication data citation templates

12) Cell Annotation Task Force
13) Data Dictionary Feedback. We can take some time at the conclusion of today's meeting to discuss pending questions.

- Key Data Dictionary Notes - Google Docs
- Microscope Metadata Samples: UTD Submission

14) Completed PRECISION_Data _Dictionary_v2.0
(PRECISION-HEAL Data Dictionary v2.0 - Google Sheets)

12/13 Meeting Notes

Action Items/Challenges

- January 13th PRECISION/HEAL/SPARC program-level meeting scheduled. *See item 11 above.*
- Requests put out for those interested in joining Cell Annotation Task Force. Maryann and K-CORE will help facilitate that. We want to ensure that PRECISION efforts aren't siloed from other efforts. Maryann is very plugged in with relevant communities/committees. Currently determining next steps.
- There's been a lot of progress on the Data Dictionary and we've received a lot of feedback, particularly more recently from the Human Tissue Subcommittee. We'll use some of today's meeting to go over additional questions. We've also received feedback on the anatomical terms questions. More recently, we requested microscope metadata samples.
- We will discuss Data Sharing and Publications Agreement as more of a formality during SC as it's been distributed widely and multiple requests for feedback were put out.
- We discussed the human specific markers and QC white paper ideas, so I put that on here as a reminder.
    - AI: Create human specific markers and QC white paper ideas
- Reminders to:
    - continue submitting Tools to Tools/Resources Submission Form
        - Update: We're in the process of getting Sensoryomics on SPARC
    - share new contacts/team members with Sam Kessler so he can update lists and add people to Slack

Sensoryomics/Visualization Environment

Ted Price Sensoryomics Presentation

- Sensoryomics.com website was started several years ago. Initial effort was to ensure data we generated (at the time bulk sequencing of experiments on human DRG), could be widely available. During recent meetings, it was eye-opening to see how a large proportion of meeting attendees were using the website when searching for genes.

- When UTD previously sent sequencing data files to people, those investigators didn't understand what to do even if UTD sent them in Excel sheets. UTD decided it'd be helpful to make searchable databases. UTD now uses Shiny for most of the data sets that it makes, but didn't when Sensoryomics.com was created. Ted decided to buy the domain name as part of continued effort to ensure that the data that UTD generates is as widely available as it can for people in the field–including academics, industry people, and people who don't have any specialized knowledge of working with sequencing data—to be able to access and utilize the data.

- We have our spatial data, our long read sequencing data, our brand new proteomics data, and our Interactome tool…The most accessed data sets that UTD has are the spatial data that was published about two and a half years ago, and the bulk sequencing from a fairly large number of thoracic DRGs from thoracic vertebrectomy patients, which is all in collaboration with Pat Doherty's group at MD Anderson.

- Slide shows that you can toggle in between things to look at the data in different ways. You can visualize data, create a gene list and look at the spatial transcriptomic data. You can also switch over to look at those in the neuropathic pain bulk data, download, make your own plots, or download the data sets or the plots you make.

- The two most widely accessed datasets here are our new bulk proteomics data and Asta's long read sequencing work.This one is represented in a slightly different way because a lot of this is about splicing. We integrated this with the USCS genome browser page.

- We developed this tool before those other tools were available. We kind of couched it more as a pain-specific tool. We should have made it a lot broader and made the tool a little bit more widely available, but we have kind of integrated this with some other tools and you can now easily take data sets that we have generated for human DRG and preload them and then take some other sequencing data set that you have to look at interactions between cell types. We have everything from our data sets at this point preloaded so that people can do that and then you can look at intersections, you can filter and rank, and you can also visualize so the tool will generate Sankey plots using the code from SankeyMATIC. We've tried to make it easy for people to use our DRG data sets so that they could easily look at what might pain centric interactions.

- And this slide just shows how much it gets used. So, the spatial transcriptomic and neuropathic pain stuff gets used the most. The proteomics just went up. We have a relatively smaller number of people that are using the human DRG and interactome, but the people that are using it give us really positive feedback. People also probably bookmark our site when they hear of some gene that they don't know about and want to know if it's expressed and or increased in pain DRGs, because they don't typically spend a lot of time on the site but you can see we have a pretty good number of users that come over and use the transcriptomics and neuropathic pain site.

- This is really a team effort here.they're supported by our center which is great to have the local support to make sure this is sustainable over the long term.The reason we have stayed committed to this is because we feel like it's really important that there's a kind of sustainable easy to find resource for all the data sets that we generate and we like getting feedback from people about how they use the data. We also maintain a pretty wild set of box folders that we often share with people because people do also give us feedback that they have a lot of trouble getting in the dbGaP to access our data for people that we know are people that we know qualified researchers that work in the area, we often share the data sets through these box folders that we have. We don't share the raw sequencing data though.

Sam: We asked Ted to present today because we wanted to consider ways to approach our visualization environment moving forward. We wanted to have a general conversation and also provide an opportunity to provide Ted and UTD with feedback.

Rob Gereau: We talked about this at the outset of the U19 effort. Ted's Lab has created an incredible resource for the field. Everybody knows about it and uses it, and I send people there all the time. One way that we can make this effort really impactful is to have a really user friendly visualization and search tool.

Maryann: It's a very nice tool. We'd want to ensure that if data sets are in SPARC that, ideally, they could be directly imported over and visualized, and if they can't we make it so that the references and the identifiers make it clear what the provenance is of data, and that it could possibly be part of the workflow. If we consider using this as a primary tool for representing data sets that are coming in from SPARC, then we want to make sure you're uploading data in dbGaP, you're uploading data in SPARC, and then it gets visualized over here. But I don't know the architecture. Sam, we just need to have a sit down with the technical group to consider how it could be integrated into this or if it makes more sense to use this as a visualization tool that's listed amongst many.

Ted Price: There's currently no consortium data that is part of our Sensoryomics site, including the Harmonized Atlas. We didn't incorporate that in any way into our own site because it's not really ours. So the only data that we have on the Sensoryomics site is the data that's UTD-generated. And none of the Sensoryomics activity is funded by our U19 so the people that are doing the work on this are funded by our center and we're just creating tools to allow people to access the data….I had a lot of discussions, I think before the U19s, with some of the folks from the GTEX Consortium and our original idea was, "hey, we'll just give you all the DRG data that we have and you guys can incorporate it into there." And then when people do searches for a gene on GTEX, it'll just pop up in there and then there it turned out there were all kinds of reasons that that couldn't necessarily be done. One of the things that I learned from that is the importance of having a way to ensure that these resources have longevity and some of the challenges that come along with that.

Sam: Certain things are also unclear to me and I'm trying to figure out what's distinct versus integrated. I think even having a general conversation about the audience, like whether we should assume they have domain expertise. So I think just getting the right people in the room to discuss what questions need to be addressed.

Ted Price: If you look at the stuff that I just presented, the long read sequencing, the proteomics data is on SPARC. We didn't yet build an epigenetic landscape site and don't know that we necessarily will ... .the spatial transcriptomics and the neuropathic pain stuff isn't on SPARC because that all predated the U19. As far as I'm concerned, the SPARC part of this is great because it means that we no longer have to maintain all these box folders everywhere so people can download and get access to the data. For all of these new data sets, we just point them to SPARC. For instance, for the long read sequencing people have asked me how they get the data, and I can just share a link to the SPARC site, and nobody has ever asked me another question. People would get frustrated in the past when they were going through the rigmarole of getting access to dbGaP and give up or come back to us and ask if we could just share the data. And for years, we'd say we could share the processed data and Deanna would share the box files and I'm sure she's sick of it.

Peter Jin: The resource that Dr. Price's group has generated is amazing. There are a couple other external groups outside this U19 that have generated very useful tools. Should we consider using them for reference or for annotation or do we know any other useful tools that we can use for our research?

Ted: Peter, are you asking specifically for naming cell types?

Peter Jin: Not only that. Cell type would be one. The other would be splicing gene expressions in probably not just human DRG but the other relevant tissues.

Ted Price: Yeah, it'd be great to integrate similar data sets in such a way so that you can. A benefit of the GTEX data is that a single search gives you an amazing viewpoint of gene expression across a number of tissues that includes splicing data. So most of the tissues that we work with aren't part of that effort. So, you can get a little bit of tibial nerve data from there, but it also doesn't have anything like the clinical characteristics that we have here, and I don't think it has single cell data for tibial nerves. It'd be great if we had long read sequencing and you could look in the DRG. You could also look at the splicing differences in the spinal cord and then the peripheral nerve, but I also don't know that other groups are creating long read sequencing data sets. So, that might be a little bit limited, but I viewed single-nuc for ERG, spinal cord peripheral nerve neuroma, etc, as being something that SPARC would eventually do to integrate similar types of data sets. And we had that presentation on the anatomical maps and I can imagine how that would all get integrated into something like that.

Metadata Feedback Follow-up (12/13)

Jyl: From a high-level standpoint, we're closing in on Version2 which we mentioned at a recent Human Tissue Meeting. Illias is working on a cross group data dictionary, so it might be confusing if you look at the Data Dictionary. We want to be in a position where we can go back and look at the different values and understand if they're required either by your group or mostly HEAL. SPARC has a few minimal requirements. We're also still in the phase of deciding if some of the things that were in Version1 will still be required for Version2 or be recommendations. Thank you for sharing your samples because that's helped Ilias identify where there are some disconnects between Version1 and what you're actually collecting. We're also expanding it a little bit for the imaging side. As a reminder, we're going to have required–those would be the ones that you guys all agree need to be collected by everybody–and then recommended.

Illias: Thank you, everyone, for sharing your feedback on the questions and updates we had on the metadata. I will try to ingest this and we'll provide you with the final version of the Version2 and see from there if you agree with that version, and then we can solidify that. There's no pending questions we have, and we're hoping to wrap it up before Christmas and then we can go about presenting it and having it approved.

Hanying Yan: We have the live human subject and I wondered if there's any additional metadata for live human subjects besides what's identified in the data dictionary that Ilias has created. It's just a little bit different from the HEAL CDEs and what you guys submitted before.

Maryann: It depends on what this consortium does with the tissue in terms of techniques. So we expect the HEAL required metadata, and then if you all have agreed to certain required fields from PRECISION, which generally are at the technique level, then you would be required to fill that out for those techniques.

Jyl: We requested that everybody share samples of their image files with us. Sam recently sent out that email/Slack messages so that we can evaluate to see if your collection devices already collect the information that we need.

Sam: UTD has already uploaded some samples within the last few days. I've added notes in the chat again for the other 3 U19 centers and Sue Tappan is also on this call and can address specific questions as she'll be the one evaluating them. (Instructions and relevant links are included within the Key Data Dictionary Notes**)**

Rob Gereau: The list of metadata was extremely daunting to think about the number of image files that we have and the diversity of microscopes. How much are we going to have to go in and edit for this metadata document?

Sue: So there is an application that SPARC has created called Microfile Plus that allows the image files to be read in the metadata that we identified in that spreadsheet are all the fields that relate to this device, software, and it provides a mechanism to add metadata. Microfile Plus converts your file to an open format so other people can utilize the image data for their own purposes. So we support an OME TIF version of your image data, the raw data as you acquired it would also be included. And that processing is just to add metadata. It changes no pixels. You can change the compression level if you want, but that's not a requirement of the prospect procedure. Working with other groups in the past, we've requested that everybody uploads the images they typically work with. Then I've gone in to look in the application, see what's present and what's configurable automatically from your acquisition software, and then go forward from there. Ideally, the required metadata is something that's kicked out by your acquisition device rather than something that has to be added by a human.

Sam: To bring some conversations together, I think we're doing this to determine what's required versus kind of good to have. We really want to figure out what's really automated versus manual. I think people often will want to submit a bunch of metadata, but find that more difficult to do in practice.

Maryann: When we went through this process with ReJoin microscopic imaging metadata. There's 27 required fields in microscopy imaging metadata, but 22 of them were automatically extracted. So if it's automatically extracted from a file header, you can have as much metadata as you want…But if you have to do it manually, then you have to consider whether you are willing to go through and put structured metadata in. If it's critical, then it's got to go there. But if it's something that could be in a protocol, for example, we allow things to go into protocols. because that's often a good place to discuss details of tissue processing rather than having fields for 100 different steps right in your protocol. So we really want to get to something that you can use because you're going to be the ones who have to fill this out. That's why we want to make sure you understand what required means because if it's got a technical requirement there, you'll have to pass that.

Rob Gereau: I would like for us to be very careful about what we say is required then.

Julia: I know we've been focusing on the metadata that is common among the techniques that are across different centers, but there's still a couple methods that maybe only one center is using. So it's not something we probably approach in terms of required versus recommended metadata. So, how should investigators go about those types of metadata fields?

Maryann: We know that there are some techniques…for example, there's the physiology. There are some standards that are out there on that. We've got multiple groups working with Patch-Seq and we thought we would try to consolidate across them. We haven't found a community standard that everyone has reached an agreement on what fields to acquire. If you're using any technique that isn't common across the consortia, please consult with us and we'll see if there's a community standard that you might want to adhere to, but if there isn't one we would be happy to work with you on determining the metadata that's required. The way that we think about this, at least from an information technology point of view, is what information is going to be required to search for a dataset that is being used for this technique.

Sam Kessler: My understanding is also the curation team has been quite busy with this. For example, I know they've been working with Bryan Copits on the Patch-Seq.

Hanying: I have a question for the metadata dictionary, especially for the bioinformatics analysis tools. We may have new software coming out that we may use in the future. So what should we do about the required fields if those new tools might sometimes skip some of the analysis steps? Should we just say "not applicable" in fields for new tools?

Maryann: We've been working with PRECISION investigators on what to do with values that aren't there. When there's a lot of technological flux and these things may not be required in the future, then you can make them recommended and fill them out but then there's no requirement that you put it there. But I like the NA solution right now. And this is a suggestion we're going to make to NIH that they come up with some standards around this.

2024-11-08

Attendees: Ted Price, Anthony Juehne, Bijesh, Diana Taveres Ferreira, Kevin Boyer, Tassia Mangetti Goncalves, Hanying Yan, Joost Wagenaar, Julia Bachman, Elle Mehinovic, Aldrin Yim, Maryann Martone, DP Mohapatra, Ilias Ziogas, Anka Pilko, Julia Bachman, Guoyan Zhao, Elle Mehinovic, Xianjun Dong, Selwyn Jayakar, Himanshu Chintalapudi, Ibrahim Saliu, George Murray, Guoyan Zhao, Bijesh

Agenda Overview

- Action Items/Challenges
- Discussion on Human-Specific Markers/QC
- Publication Process Overview
- Metadata Update, Metadata 2.0 discussion PRECISION
- SFN
    - Cell Type Nomenclature Meeting
        - Maryann's presentation to the SC meeting about cell types
          Metadata discussion PRECISION: Cell annotation standards

- Updates/Reminders
    - NIH PRECISION Human Pain Network Resource Page
    - Tools/Resources Submission Form

Notes:

- Action Items/follow up:
    - Good progress on metadata analysis moving forward based on metadata standard V1 and experimental metadata, as well as incorporating HEAL & SPARC and now dbGaP requirements into it. We'll continue reaching out with questions
    - Data Sharing and Publications agreement approved at the SC
    - Sam sent instructions for submitting PRECISION Tools/Resources and will continue to share that with this group. We have a few that are getting close to the publication on the Portal
- Human-specific Markers QC:
    - DP: Start to think about tissue requirements and build confidence that samples are from the same species tissues, etc. More towards functional or biochemical data. A lot has been happening outside of this consortium, but all of it will be taken from AI in the future for integration, we need to have a clear, identifiable source. So future experiments are accurate and build confidence. Within this subcommittee, what sort of data should be stored for experiments. What should be shown to identify an experiment. What are the unique human DRG neurons in electrophysiology? Food for thought to start talking about this in this network.
    - Ted: talked about this in the Tissue meeting, some ideas related to cell size. Some genes specifically found only in human. What form could this take, potentially a paper
    - DP: come up with some ideas, how to test and what would be needed to show in terms of data and results for a white paper
- Publication Process Overview
    - Curation for high-standard space like SPARC and to coordinate with HEAL, etc. the process takes some time, so work on it early.
    - This slide shows different steps and estimated timelines, but it varies based on complexity of the dataset and the what's required for dataset to be in compliance.
    - Share your dataset with the Curation team on Pennsieve. After curation performs the final check and proofs dataset for publication, the investigator has to sign off. Most datasets are published on Fridays.
    - SPARC has an embargo function, it'll resolve to a page that has metadata so you have a landing page that describes the data and how you get it. You can't access this data without investigator permission.
    - Cite the DOI of your dataset in your paper (not the URL).
    - To summarize, it's always good to consult the curation team before submitting for the first time or a new type of data, and inform them about special timelines/needs/potential complications. Curators are happy to preview your data set before you formally submit it.

        - Remindeer: Once you upload it to Pensieve, be sure to share it with the curation team or they won't see it.
- Towards Metadata V2.0
    - In the meantime, work with V1, the more the better
    - PRECISION metadata standard-fields identified as required are not all being submitted
    - Basic demographic variables even post-mortem could match HEAL Core CDEs
    - We'll continue back and forth with questions
    - Anatomical:
        - Work has to be done on anatomical terms to align them to community standards. The current list has a lot of ambiguous names and we're aligning those to the community standard that is used across a lot of projects. Also try to capture spatial specificity
    - Ted: what is AI readiness?
        - Maryann: Several groups are still trying to define this, but general agreement is well-documented standards, rich metadata, provenance is clearly recorded and machine readable, permissions and what the data can be used for is specified. Bridge2AI is one of the groups we https://commonfund.nih.gov/bridge2ai
    - Anthony: Points of follow up for HEAL
        - --Alignment of SPARC metadata v2 standard with HEAL SLMD and VLMD (https://github.com/HEAL/heal-metadata-schemas/tree/main/variable-level-metadata-schema)
        - --Alignment of HEAL CDE implementation and VLMD curation across SPARC and HEAL Platform MDS
        - --Use of publication data sharing citation template across HEAL
    - Working on recommended modifications as well as Data Dictionary
    - Will need to work with PRECISION investigators and get feedback on many questions
    - Next month with an update, potentially proposed updated version later
    - Xianjun: V1 to V2, when will we create a consensus V2, they made a lot of changes and added several fields
        - What we consider as a consortia to be required vs. recommended.
            - Manual entry generally tends to have low adherence, but if it can be automated then it's easy. If automatic we can get as much as we want.
            - Some metadata may be very experimental, but they should be in the protocol and not in a metadata form. As long as we have it, we won't require it in a form.
            - A lot of flexibility for getting this information. We want to minimize the number of manual required metadata that would mean it would get kicked back to you
        - However, always feel free to include extra metadata-it's possible we may want to see if others want to match what you're contributing or look at metadata elements

        - Xianjun: they determined what was required internally
            - Researchers know best what to absolutely require
- Cell annotation standards: follow up from SfN
    - Attended 10/6 Toward a unified nomenclature for somatosensory neurons. Diverse experimentalists-no informatics people. Disagreement about important aspects of a nomenclature and what to capture. Is it based on transcriptomics, physiology, connections, morphology?
    - These arguments will persist and a single nomenclature isn't on the horizon for the needs of this group. We can proceed in the practical approach described in September presentation. We can support it with modern technologies that tag things with phenotypes. if you record your phenotypes in a machine processable way then you have a lot of flexibility in how you report neuron names. Machine processable nomenclature using the table in https://www.science.org/doi/10.1126/sciadv.adj9173, https://painseq.shinyapps.io/harmonized_painseq_v1/ formalization using same methods used in BICAN, BICCN, Cell ontology, Provisional Cell Ontology.
        - Turn them into computable phenotypes-cell names are constructed computationally
    - How would this group name their neurons? Can turn it into something machine computable and work to have it show up in the DRG/TG reference atlas.
    - Xianjun: when we annotate, we use expression markers & we're giving at least a marker gene from biologists, and then we look at those genes expression profile in the human app, and then we manually annotate each cluster to a cell type.
        - How do we convert to this method you have described here?
        - Take marker genes (don't take full expression profile), formally express the phenotypes associated with it. Any other
        - Jyl in chat: I think slide 9 has some good examples
        - Maryann: has taxonomy standards they use, we could build on that so that cell types can be more easily compared using the features extracted from these datasets
    - Guoyan: where will this information be recorded?
        - Maryann: The actual name of the neuron type is just an ID. We'll build the ontology and knowledge base of the cell types reported by the PRECISION consortia, and will give you the information you need to reference them.
        - It can have cell types that aren't neurons
    - DP: feels like this consortium should be playing a major role in this field
    - Maryann: have Sam put out a call to help work with them
    - Xianjun: would like to have some examples
    - Maryann: BICAN had a proposal for how to deal with data driven cell types. We would like to take this use case and come up with a solution for it. We know the issue of cell types is difficult, and data driven ones are even more difficult (based on a profile). Others are grappling with similar issues

2024-09-13

Attendees: Joost Wagenaar, Maryann Martone, Aldrin Yim, Asta Arendt-Tranholm, Ayesha Ahmad, Diana Tavares Ferreira, DP Mohapatra, Elle Mehinovic, Guoyan Zhao, Hanying Yan, Julia Bachman, Jyl Boline, Ibrahim Saliu, Kevin Boyer, Khadijah Mazhar, Mingyao Li, Nikhil Inturi, Peter Yin, Rachel Weinberg, Shamsuddin Bhuiyan, Suzanne Tamar Szak, Xianjun Dong

- Action Items/Blockers
- Finalizing the Data Sharing and Publications Agreement
    - Re-Join Banner Sample
- Cell Type Naming and Annotation Standards (Led by Maryann from K-Core)
- General Discussion/Updates

Action Item Summary

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
7	Send Instructions to Prepare for 9/13 Metadata Discussion	Sam K	8/21/24	9/13/24
8	Send Instructions to Prepare for 9/13 Cell Type Discussion	Sam K	8/27/24	9/13/24
9	Do Final Review of Data Sharing and Publications Agreement Draft so we can finalize at Data Subcommittee Meeting	Full Committee	9/3/24	9/9/24
10	Send Email Instructions for PRECISION Tool Catalog Discussion	Sam K	8/27/24	9/9/24

7) Revisiting the metadata dictionary and requesting metadata samples.
8) DCIC decided to use the 9/13 meeting to provide a broader overview of what we're looking for and how we can leverage expertise. The two main resources we'll use to frame this conversation will be:

- ***Harmonized cross-species cell atlases of trigeminal and dorsal root ganglia - PMC (nih.gov)***
- ***Cross-species transcriptomic atlas of dorsal root ganglia reveals species-specific programs for sensory function | Nature Communications***

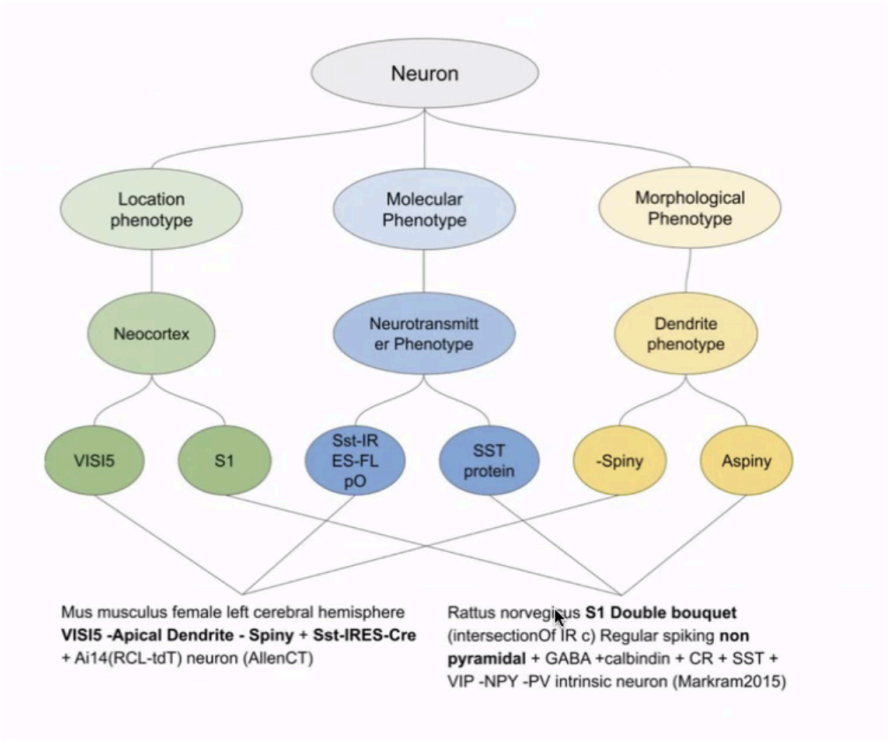
9) Data Sharing and Publications Agreement Draft

- Document updated and link shared in agenda email sent on 9/9

10) SPARC website updated so investigators can submit PRECISION tools and resources. A submission form link was shared via the general Slack Channel. To see examples of other tools/resources that were previously provided, click here

09/13/24 Notes:

- Metadata overview: Metadata standard for PRECISION is evolving. In the meantime, just capture what you can from V1 or other metadata you think is important. Sam is collecting metadata examples samples. K-Core will be following up about what people are doing and work towards ensuring PRECISION investigators meet HEAL and SPARC requirements. In the meantime, people should collect metadata even if what we have may not be thought of as a "final" version. Use Data dictionaries, variables + description of what they are.
    - Julia: clarify that we're working on a V2 ov metadata. In the meantime use V1.
    - Jyl: some people are collecting data types that aren't captured in V1, but they should still collect metadata they consider important, match standard to extent possible.
- Cell Type Naming and Annotation Standards-Maryann
    - Essentially when people work with cell types, how do they name them, and can we formalize some of it so we can use information systems more easily (See slides 2 & 3)
    - Sitting behind the SPARC Portal is a knowledge graph (KG) that supports search and query (SPARC annotates key entities with controlled vocabularies, which are nodes in the KG and objects of the search/query). Ontologies put controlled vocabularies into computable and inferrable format. SPARC wants to extend some of these key entities based on key terms for PRECISION.
    - SCKAN-knowledge base that specifically holds connections (cell bodies in anatomical entity a, axon terminals in anatomical entity b, axon travels through anatomical entity c (sometimes entities c, d, e, etc.). Having this information in this form allows dynamic visualization and extensions
    - Extending the knowledge graph to include PRECISION cell types and molecules. A lot of structured work has already been done by this group. (See slide 4)
    - Overview of what an ontology is. Ontology is a formalization of human knowledge in a form that can be reasoned over by a computer. Add hierarchies/chains that can be used to define an entity. (See slide 5)
    - There can be many many ways for how to classify a cell. Basically, we still don't know about what constitutes a cell type. We have some agreement at high level, but there's less consensus as we get to subcategories.
    - Proposed cell types, we call these evidence based types in the wider community. These are called provisional cell types. We can have as many of these as we want. The important thing is they are mapped to a formal ontology. (See slide 6)
        - So even if we may not agree on cell types, we at least can talk about cell types in a common language which makes it easier to travel across the different nomenclatures and find things that are in common. Example:



    - We don't make claims about the underlying science, we just say someone asserted this by this technique. People in the scientific community can review and evaluate the information. Maryann tends to think of this as a way to find the diverse information that's available and potentially pull the bigger picture together as well.
- BICCN example, extracted marker genes in the primary motor cortex and modeled it using the Provisional Cell Ontology (PCL). They've built a Cell type knowledge explorer at the Allen institute. Where they link information to the primary data. Ontologies don't capture everything in a profile, but they capture what is asserted to be the marker genes. (See slide 7 and 8)
- Using this framework allows powerful search and visualization, but we also don't need to have consensus, these methods can still capture the formalizations of these phenotypes.
- Recommendation: if there are consensus types, capture them into the taxonomy and issue these IDs. If there isn't agreement, each lab should register their cell types as they produce them and work with our knowledge engineering team to express them into these formal ontologies. Consistent nomenclature generates labels and synonyms, so you can use whatever names you want.
- If there are points of comparison between these connections we'll be able to see that.
- The Cell atlas that goes along with the paper seems like a good start to our tool and resource catalog that will be available on the SPARC PRECISION page. We're also collecting other tools you use
- Shams: It would have been a good idea to put together a graph of the ontology for their publication. Maryann-we can still do that
    - Shams: what about species differences? SMR in mouse doesn't have an ortholog in humans.

        - Maryann: this is a common issue. These are species differences in connections too. Hierarchies generally have a set of phenotypes that at a high level associate with a specific cell type. We can in an ontology say that something lacks something (e.g. a gene). As opposed to not knowing if something is there or not. That use case could still be formalized and still call out that something that was looked for wasn't found in human. Species are also tied to NCBI taxonomy ID. So you can see where on the tree that particular thing disappears. We can put you in touch with Tom about that.
- Shams: connecting granularity of single cell data to single nuclei data. With single nuclei data it's just called PEP, but single cell data there's granularity there.
    - Maryann: we can talk to Tom about that, one of the reasons he insists on having the technique is there molecular phenotype, but it might differ at the technique level. I think we can accommodate that, but we need to touch base with Tom.
- Guoyan: how about the sequencing devs will that be captured?
    - Maryann: that's in the data. Ontology won't be able to capture 4000 genes. We try to capture what the scientists are asserting, we don't assert that what they say is correct. We say this has been observed in human, but you can't assume it's in everything else because we haven't looked for it. But this is why we capture it. These are experimental cell types. They have to be attached to datasets. Not a perfect capture of an analog world
- Guoyan: Brain region, cell type in cortical vs. subcortical
    - Maryann: pretty good about capturing neuroanatomy. Because layer in cortex is important, e.g. L5 V1, hierarchy all the way to visual cortex is in the ontology. People should annotate to whatever level of granularity you have. We have two top notch ontologists on our team.
- Joost: how do we go from vision to practical? Work on formalizing more of these as part of the SPARC Curation process.

2024-08-09

Attendees: Joost, Maryann Martone, Hanying Yan, Aldrin Yim, Jyl, Ibrahim Saliu, Rachel Weinberg, Khadijah Mazhar, Kevin Boyer, Elle, Asta Arendt-Tranholm, Xianjun Dong

Agenda Overview

- Action Items/Blockers
- Metadata and Controlled Vocabulary Metadata discussion PRECISION
- Representing Data on SPARC Portal
- General Discussion/Updates

Action Item Summary

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
1	Facilitate Elle and K-Core meeting to help with uploading data	Sam	6/20/24	6/20/24
4	Explore rescheduling Data Subcommittee Meetings	Sam	6/26/24	6/26/24
5	Add metadata progress and next steps note to general Slack channel	Peter	6/17/24	6/15/24
6	Determine metadata action items to take place in interim of group metadata meeting	Sam/Peter	6/26/24	7/15/24
7	Schedule metadata meeting	Sam	6/26/24	

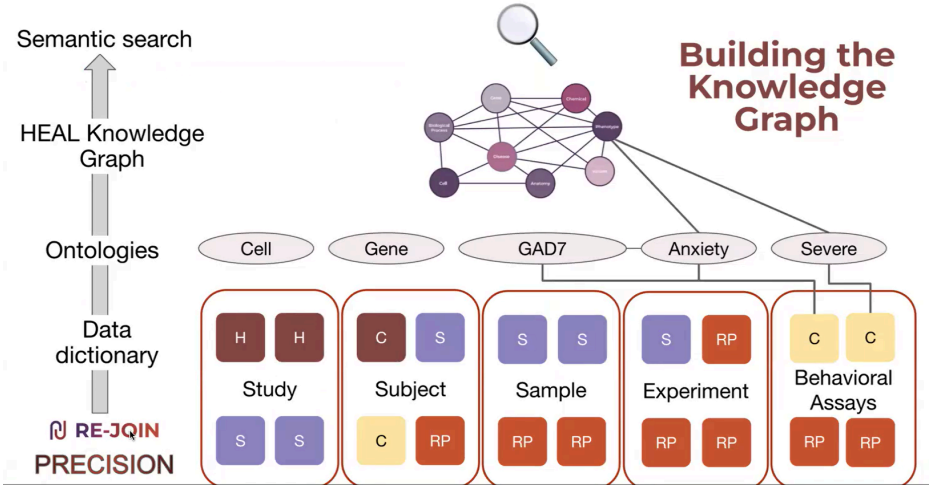
1) Completed
4) Given expected light attendance, we decided to not have a July meeting. We are likely sticking with recurring current timeframe, but are considering changing this pattern every 3 or 4 meetings meetings to incorporate Map-Core and/or plan adhoc meetings
6) Prioritized Finalizing K-Core Questions
7) New Doodle will be shared next week along with today's presentation. We'll just go with the best option.

8/9/24 Notes

- **Maryann Martone: [Sharing PRECISION HEAL Wrapping Up Metadata V1):** Presentation is an overview of the current state of metadata in PRECISION. HEAL has a set of very prescribed data elements that they want collected in a certain way, and those have to be satisfied. SPARC, the data platform you're using, has various standards. PRECISION has its own requirements and these may not entirely be encompassed by what HEAL requires or what SPARC requires, so we need to understand what those are. And those also can be divided into cross-lab activities because there's a lot of data that you are collecting in common, but I've also seen from the data survey that there may be data specific to an individual lab. The DCIC's job isn't just to facilitate the collection of this cross-lab and w/in a lab data, but to ensure that that we can present it to HEAL consortium and the community at large in a way that is FAIR. There's a certain amount of standardization that has to happen, so PRECISION data can be integrated with all the other data that is being collected by HEAL and biomedicine in general.
- **Maryann:** Various HEAL requirements have been put in place. Study level metadata describes the study as a whole, and looking at HEAL's scorecard, it looks like everyone has completed it.

Today, I'll be focusing on the Variable Level Metadata. This is in the realm of the common data elements (individual items of data that you collect and how you organize them and structure them).

- **Maryann:** HEAL has the NIH CDEs for human data that are prescribed and must be collected in a certain way. PRECISION has done a very good job in doing that PRECISION-wide, which is what I call a data dictionary that covers a lot of what it is that you are that you are planning on covering. It's in a structured Google spreadsheet with dropdown menus. The PRECISION-wide data dictionary is going to have to be submitted to HEAL so that our data can be properly queried. So you'll be working with the DCIC, and the DCIC will serve as an interface between the Variable-Level metadata and HEAL to ensure that HEAL has everything they need to appropriately structure and query PRECISION HEAL data.
- **Maryann:** [Slide below] shows how information flows from the individual working groups that are underway in PRECISION. We'll be producing a data dictionary that reflects the variables that you are collecting and how they'll be structured. These will undergo further refinement inside of the DCIC and HEAL so that they can fit into the HEAL Knowledge Graph, which is basically a large graph structure that is able to go across the individual studies, across the variables, that are semantically related to each other.



- **Maryann:** Core CDEs: minimal and defined set of patient reported outcome screening tools for each pain domain that all HEAL pain clinical trials are required to collect. These are made available on a box and the HEAL website. As I mentioned we've reviewed the PRECISION Metadata V1.0. Curation has gone back-and-forth with some individuals based on some feedback to the questions that we had, but in a deeper review we can clearly see that the document is clearly informed by HEAL requirements, but it doesn't appear to be exactly compliant with everything in the Core CDE requirements. There's detailed consensus metadata that's provided for bioinformatics, pipelines, and genomics data, but according to the data survey that you filled out, there's other data types that will be collected (physiology, spatial transcriptomics/proteomics, and imaging). And I wanted to go through some strategies for how we might be able to bring some coherence to that, especially given that Map-Core is working on the spatial aspects. It's very important that we agree on how location is going to be annotated.
- **Maryann:** In terms of the coding of CDEs, HEAL recommends coding your datasets with the HEAL CDEs before data collection rather than re-coding later. We need to ensure that the variable names, permissible variables, etc, correspond with the CDEs in the datasets, and even the wording of the questions in the CRFs are the same as in your own CRFs.
- **Maryann:** When looking at the data elements that you provided in the metadata standard, I saw that, for example, you have the same household income categories that HEAL does, but what HEAL is expecting is a permissible value. We'll review and figure out where these sorts of things

exist and ensure that we conform to the extent possible. There's always some re-coding that needs to be done based on data already collected, but if we can align the collection instruments to what HEAL requires, it'll make things a lot easier down the line.

- **Maryann:** We also need to consider other data types. Especially metadata to be able to understand them across the labs. There's imaging, spatial transcriptomics/proteomics, and physiology, and we want to focus on places where more than one group is collecting the same type of data as potential points of correspondence that we would like to explore. 2 groups are collecting Patch Clamp and REJOIN is also now collecting patch clamp and asked us what it is that we need to do in terms of metadata standard. We think there's an opportunity to deal with this data type in a more coordinated way.
- **Maryann:** Some imaging strategies include:
    - Ensure microscope is configured to provide richer imaging metadata
        - For imaging metadata SPARC already has 21 optical microscopy parameters that it expects. But most of these are extracted automatically from imaging using the SPARC image pipelines. So, provided that you've configured your microscopes appropriately so that all the data that is captured by whatever system you are using is accurate, we can extract those automatically. If that can't happen, then we ask that people fill in the required metadata that you need.There's certain critical metadata that you need to interpret an image, so if that happens, we will ask you to fill that in manually. But hopefully, most of the file formats that you're using won't be required so this kind of comes automatically when you upload to SPARC.
    - For Physiology Patch Clamp, we did see that one lab (I think the Gereau Lab) is planning on converting their data into Neurodata without Borders (NWB), which is a fairly well recognized community standard. There might be a joint activity that can go on with REJOIN..
    - Set up vocabulary in Interlex by extracting terms from existing anatomical vocabularies and augmenting as necessary with PRECISION specific terms
        - We want to make sure that spatial location is consistently annotated and described across the different projects so that at the end data can be put back together again inside some of the visualization that Map-Core is doing. So, usually my group handles the vocabularies and Map-Core handles the spatial referencing, including fiducial marks and other sorts of things that people may need to define in order to do this localization. So our recommendation is to take what you've already put in, map these to community standards for anatomical terms–that's the UBERON Ontology...We do have a system whereby, if there are custom terms, or terms that are not in the community ontology, we can add them into a custom vocabulary that manages it as a computable artifact. I recommend that we set some of those up for PRECISION, start to understand what sorts of tags we're going to need for annotating data.
- **Maryann:** Interlex UBERON/FMA identifiers are very important to the overall interoperability across HEAL. The Knowledge Graph is built on these identifiers. So we will be taking the metadata standard and also indicating where some of these identifiers need to be supplied. A lot of them are automatic, meaning we can do the mapping of anatomical terms to UBERON, but there are some that you will have to provide yourself, and so we will indicate where those are. We do have unique identifiers for things like antibodies and Gene IDs.
- **Maryann:** For Subject and Sample, we do note that you are good about putting in these identifiers, and, of course, if the same subject or same sample is being used across centers, they need to have the same non-identifiable ID. SPARC actually tracks these, so that if you do use the same subject across multiple studies or the same sample across multiple studies, that is something explicitly tracked by SPARC in its schema so that we know that subject A over here and subject

A over here are, in fact, the same individual. It's always best practice in laboratories to be working with unique identities for your subjects and samples.

- **Maryann:** For next Steps:
    - Wrap up PRECISION metadata V1.0, including the HEAL Core CDEs, the metadata standard that has been developed and proposed already, and I would also just add imaging metadata to that because that is something that already comes pre-baked into the SPARC pipeline.
    - We are going to host the official version of the PRECISION Metadata standard on SPARC. We have a dedicated page just for PRECISION. We will be responsible for taking the metadata standard that was created and put that in the form of a data dictionary. And that's what we're going to communicate with HEAL about to make sure that they know every PRECISION dataset is going to organize their data according to this way. We'll help with semantic mapping and other things
    - **T**he DCIC will submit variable level metadata to HEAL on behalf of PRECISION once we have everything done and once we start submitting data to PRECISION, and that will be done in conjunction with data upload. When we upload the data, it'll go over to SPARC.
    - Then for imaging, we're just going to use the SPARC standard and pipelines unless anybody has an objection because it's already there and there isn't one in HEAL.
- **Maryann:** In the future, we can go into the next version of the Metadata Standard (Metadata standard 1+). I think these are the spatial and semantic standards that need to be agreed to, and how we want to proceed. We will consult with Map-Core and with the different groups. And we'll talk about that in a future meeting. I'm hoping, again, we will contact the people who are working on Patch Clamp and that they will agree that we work together with ReJoin because I think that just will be a lot simpler, and it will also again give us a degree of interoperability across HEAL.
- **Maryann:** Next month, I'd like to start talking about the standard for cell type annotation. It's a big topic of discussion. We have various tools and approaches. There's been various standards that have been proposed, but I think we need to have a dedicated meeting on that and have a path forward. So, I'm hoping that by our next data meeting, the DCIC can continue to go through some communication about this last point (for imaging metadata, we will use the SPARC standard and pipelines), so that we can present a candidate to you at that meeting.
- **Peter Jin:** Regarding the genomics data, I think it'd been suggested that dbGaP was the place to go, but I want to know if we've reached a consensus about where to deposit the genomics data.
- **Joost:** Yes, we are expecting to move forward so that all of the high throughput sequencing files (like the Bam files and the Fastq files), and we're going to work with the teams to put that in dbGaP, and then have that be linked to the SPARC Platform. The DCIC worked with your groups to be listed as a data submitter for dbGaP, so we actually have access to your studies in DbGaP now. We will first manually work with you to submit those data to DbGaP, work with the DbGaP team in combination with the HEAL team to figure out if there is a mutually agreed upon data dictionary that works for the DbGaP team, that works for the HEAL data ecosystem team, and that works for the SPARC team. And that's probably a pretty simple data dictionary.
- This year we're trying to leverage that initial manual step to determine where we can introduce some automation, and explore if we can automatically generate the dbGaP data dictionary files from a data submission to SPARC or something like that. So, the primary goal is to have the very large raw sequencing files to be in DbGaP,. Have all of the derived data and the data that we might use for visualizations and further analysis have that go to the SPARC Portal.
- **Maryann:** A discussion on cell annotation needs to happen. But I think that's a separate topic that I want to prepare for next time. In terms of how the data is presented on the SPARC portal…A visualization environment for the types of data that you are providing probably built from community tools, as we say…But then we need to conform to the standard. The good news is that multiple groups around the world do seem to be talking about the schemas for annotating cells,

and I think if we can conform to that there will be a lot of tools. The question is to what degree we have consensus on cell types. And we do have strategies for dealing with that. So having a dedicated discussion about where we are on that is critical. I would recommend next time we dedicate this meeting to that topic.

- **Joost:** That's great. And if some group wants to present their visualization at some point to show what they are doing internally, that might be a good jumping board for further discussion.
- **Maryann:** Next time we talk I would like to just wrap up the metadata standard and ensure everybody understands this. But one of our other action items is—and Joost we can handle that in the context of this– is there is a resource catalog that we are going to deploy for PRECISION tools. So, perhaps getting the visualization tools and other tools that people use, we can pull those in because those would be entries into the catalog. I don't think that this is one discussion, as I say, with all of these there has to be a kickoff so everybody understands, A) what the problem is, meaning where we are right now, and then we can define what steps we're going to take for Version 1. So, I think if we start to do that at this meeting, it's going to take a while, but we can start the conversation next time. Eventually we may want to invite in other groups to present because there are other options that are developing in BICAN and HCA, but I think that it's going to take a couple of iterations for us to settle on the minimal viable product.
- **Xianjun Dong:** For your presentation note about the identifier to be added….For the subject sample ID, you were suggesting that the best practice for the laboratory is the unique ID within the lab. I'm wondering for future reference how SPARC is going to track the same subject that we used across labs.
- **Maryann:** We're preparing a presentation for ReJOIN that shows the SPARC Schema. So SPARC does want to know if a same subject has been used across multiple studies. So, in the schema we basically allow people to say–and we can actually detect ourselves—if the names are unique within a lab, whether that same subject has been used from the same group across different experiments. So there's a specific schema item to say that this subject was also used in this; this sample was also used in this. The problem that we always have is that, if you were to do a query right now on all of the subject IDs provided by the investigators across all our studies, Subject 1 would be in 100 studies… because everybody names them Subject1, Subject2, Subject3.
- SPARC does concatenate experimental ids, and subject ids, and sample ids together to try to make things a little bit unique because it's the same with samples (Sample1, Sample2, Sample3). So, the best practice in a lab is to say this sample, in my lab (regardless of what study it's been used in) always has the same identifier. And it's something that's fairly unique. So a lot of people say, "well, I'm going to take an experiment number or date and put that together with it, but you don't want different groups to have Subject 1s, that, if that's its identifier, and we don't know it's the same subject…You might say the metadata will tell you, but the chances of 2 mice having the same age and sex, you might be high, right? So the best practice in a lab is to always have some convention whereby subjects and samples have unique IDs that can transfer across Studies. We will be able to detect it because we ingest all of your sample IDs, all your subjects, and if we see two that are the same, we will ask if these 2 are, in fact, the same. But that becomes very difficult if your subjects aren't unique.
- **Xianjun Dong:** I think concatenating Subject ID and Sample Id into unique identifiers within each lab makes sense and we'll make sure that the IDs within the lab are unique per subject. But my question is more across labs. Let's say Subject1 in my Lab may be Subject2 in your Lab, and without checking the genotype, how do you know these 2 are the same person?
- **Joost**: I think that there are different ways that you can do this. One of the things that we are doing right now–and we can figure out whether we want to leverage that for PRECISION as well—is having a tool where basically you ask the person who knows who the patient is to provide some identifiable information (maybe a combination of last name last, 4 digits of the social security number) that then generates a unique hash, or like a unique Id. And if we do that or we are doing that for a different project, that means that no matter who sees that patient again (it

could be a completely different institution/different study), that as long as they use the same mechanism to generate a GUID, you would end up with the same ID for that patient because it will still have the same last name and the same last 4 digits of the security number

- **Xianjun Dong:** That's exactly my question. Do you have any suggestions for the rule to make a GUID for each of them?
- **Joost:** So, we're doing this from a different project on our side, but we can do that here as well. The tricky thing is different teams and institutions typically have different access to types of PHI data that we can include in this. So, maybe one of the things that we can do here with PRECISION is come up with a list or combination of things that we do have access to. For example, do we have access to a last name and place of birth, and then we come up with a standardized way of creating that hash.
- **Maryann:** Yeah, the other way is that whoever initiates it creates the hash, and then that hash is just used by others so they don't have to recreate it…because they might be just getting a sample. I don't know to what extent the same subjects are going to be doing it.
- **Joost:** Yeah, it just depends on how generic you want it to be. If you have good collaboration,then all you need is a unique Id and pass it along. But you might enroll a patient in some study without even knowing that that patient was part of another study, but you still automatically link that…You need to have some kind of unique ID that is consistently generated based on real information. I'm happy to put something in a document to figure out the flow for this consortium in terms of what might make sense and see if there's agreement across people to do it. If that is too complex, the easiest thing might be to just go with what Maryanna mentioned, like have documentation that outlines best practices for naming your subjects internally and be consistent. And then leave it up to each individual group to do that for their patient population
- **Maryann:** Julia, are you going to talk about NIH GUIDs?
- **Julia:** I think the latter approach to what you said is probably more ideal because for a lot of our biomarker type studies, and in the neurodegenerative space, NINDS has their GUIA generator system, but lot of this is working with the same patient in longitudinal studies and the biomarker samples are getting distributed for other studies which is really applicable to PRECISION. We have a much smaller patient population, and a lot of it is more about getting donor tissue. So, I think it's more that the subject ID is not so much going to be an issue as linking all the different samples to a subject and the samples, even within a lab that are used for different technologies. So I do think there still needs to be a concerted effort in terms of the best approach to naming strategies, but I don't think we're gonna have so many issues with patient following.
- **Maryann:** So, if we could collect the use cases for where things transfer from one place to the next (how that happens), every lab is going to be responsible for making sure that every sample can be traced to a particular subject and that that is consistent..Let's take this simplest strategy we can given the used cases that we have.
- **Joost:** Especially because this will be part of the submission to dbGaP, we should consider whether we want to have a HEAL subject ID that isn't lab specific so that we don't end up with a lot of Subject1s, but that we could have a unique Id for a subject across all of HEAL.
- **Maryann:** Yeah. So, I went over that part rather quickly and I think I was confusing because ReJOIN had a strategy already. But you have a lot of things in there that can form the basis of a strategy already in the metadata standard. So, I think we should just gather the use case and put that proposal together for how we're going to handle it.
- **Peter Jin:** In the past, we've had analytical path point discussions across the core to reach a common ground for, at least right now, 10X Single cell and Atac analysis. So, in the future are we going to reconcile, for example, Patch Seq or Spatial Genomics analysis.
- **Maryann:** That's my question to you. We might have a little working group on that. At minimum, I want the metadata standardized to the extent possible across studies, but you can go deeper than that…and it's how deep the consortia wants to go. Standardizing pipelines and analysis tools and other things is always a little bit more difficult. I'm obviously for as much as we can reasonably

do, but that's up to the groups. And that's why I suggested we have just a tiny task force that considers how deep they want to go. At the very least, we don't want everything named differently or incomplete metadata. We want to make sure we standardize that. But I really leave that up to the consortia unless you've had some agreement already with NIH as to what they would like to do with that type of data.

- **Joost:** I do think that we want to be practical about what we aim for because there'll be a lot of walls if we say our aim is to standardize everything.. So, I think that having a working group or a team that says like, "we are interested in doing that and let's see what we can do…If we can do that this year like that would be fantastic. I don't think that we want to go the route of forcing the entire consortium to update their experimental protocols and pipelines to be a standard. But as Maryann mentioned, we are looking to, on the SPARC Portal, to have a resources and tools page where we can list different tools and pipelines that are being used. So maybe the beginning of that is if there are public pipelines that are well described, let's let's turn that into like a findable page on the SPARC Portal, so we can use then use that as a way to say, "let's see what happens if we apply that to somebody else's data."..... Longer term, we are excited about doing that, as well as from a technical side of things. We are getting closer and closer to allow people to do that as part of the platform. Looking further ahead, we had originally proposed that in the latter couple of years, like Year 3 we'll start having yearly hackathons, where we actually bring some of the people that run those and build those pipelines together for a while to see what is possible there, but I think it's a little bit too early right now.
- **Julia Bachman:** We sort of had some of these discussions in the Data Subcommittee meeting in the past, and the consensus that I've received is everyone is doing something slightly different or doing something similar, but with slightly different technologies. So I think, coming to a consensus on what pipeline to use is going to be a bit of a challenge. But maybe that 1st step is running your data with other peoples' pipelines across the consortium and see what is maintained and what is not, and that might help the group eventually come to best practices and stuff. But, as Joost said, that comes as a later stage activity ... .One thing the NIH would like to see is some sort of cross validation across centers of the data that's coming out. So, that might be a good way to do that (JW/MM agreed)
- **Maryann:** Physiology pipelines and interfaces are a little harder to reconfigure since there's not as many standards that govern everything in physiology compared to genomics. That's why I said that even NWB is a much heavier lift than a lot of the other types of standards, but if we could even agree on that so the data are in a format where they can be reused by others that would be great. There's a lot of conversion tools and things, but we don't yet know–unless somebody actually converts their own data to the conversion tools, capture everything that's in that data–there's a lot of tuning that would have to go on to make this work. So we usually leave that to the scientists ourselves. We focus on metadata, the spatial standards and the right data formats that improve interoperability.

2024-06-14

Attendees: Wenqin Luo, Peter Jin, Elle Mehinovic, Diana Tavares Ferreira, Xianjun Dong, Joost Wagenaar, Guoyan Zhao, Julia Bachman, Wendy Dong, Julia Bachman, Megan Uhelski, Jyl Boline, Bryan Copits, Hanying Yan, Mingyao Li, DP Mohapatra, Ayesha Ahmad, Suzanne Tamara, Urzula Franco-Enzastia, Asta Arendt-Tranholm, Peter Jin, DP, Ibrahim Saliu, Iris Lopez, Kevin Boyer, Mingyao Li, Qingru Xu

Agenda Overview

- Action Items/Blockers

- U19 Updates (5-10 minute overview from each group)
    - Peter Jin (WUSTL)
    - Xianjun Dong (Harvard)
    - Diana Tavares Ferreira (UTD)
    - Wenqin Luo (UPenn)
    - Suggested Topics:
        - Data production to date
        - Data coordination and analysis efforts
        - Areas for Synergy/Collaboration
        - Spatial genomics, technologies, and tools
        - Cell Types
- Updates:
    - Metadata (Peter)
    - PRECISION Resource Page
    - Pennsieve PRECISION Workspace
    - Data Sharing and Publications Guidelines

Notes 6/14/24:

Action Items/Blockers

Sam: Much of today's conversation will be devoted to center overviews. A big aspect of this project is spurring collaboration. For action items, we shared what we initially called the Publications Policy and the DUA policy and combined them into the **Data Sharing and Publications Guidelines**.

There are several meetings coming up for MAP-Core regarding the surveys each U19 had representatives fill out. We created a metadata Slack channel today that includes the people that each center requested should be included. We'll make that channel public so everyone can see it, but it will be devoted specifically to ongoing metadata conversations.

The DCIC wants to leverage several good conversations that were had before the DCIC was onboarded. We're going to continue trying to use Google Drive and Slack more moving forward. I recently created a resource page that will serve as more of a direct access point for key Google Drive links and materials that were previously shared. It also reminds people of office hours. Access is limited to people who were invited to Google Drive, so you'll need to ensure you're signed into the appropriate email address. Please let me know if you have issues signing in or if you'd like me to make any changes. In Slack, I've also pinned a link so it's easier for you to see.

U19 Updates (5-10 minute overview from each group)

- Suggested Topics:
    - Data production to date
    - Data coordination and analysis efforts

- Areas for Synergy/Collaboration
- Spatial genomics, technologies, and tools
- Cell Types

Peter Jin (WUSTL Center Update)
I will quickly summarize the data production for Project 1. We have generated about 270-ish whole genome sequencing data sets from a sample. We are planning on submitting a manuscript for single cell spatial analysis, Wendy Dong, who is a trainee with Jeff and my lab knows more detail, but I know we have finished single nuclei analysis for around 13 human samples and we applied multiple technologies, for example, PIP-Seq and 10X and compared their yield resolution and the cost in results.

Wendy: We currently have 21 libraries made and have done Xenium, and are looking into Visium. We've tried a lot of different 10X technologies like GenX and FLEX.

Peter: I'll need to survey my team and get back to you about cell types. I also wanted to mention that for the data we are ready to upload and start the test run. So, I think Elle is here and she'll be in charge of that when the U24 team is ready to go.

Guoyan: For Project 2 we have 11 samples where we performed the deep sequencing. We're currently performing QC and trying to analyze this single nuclei multiome data with RNA-seq and ATAC-seq data. And then we have ongoing sequencing for control and for patients with arthritis. The current plan is to have 3 controls and 3 arthritis with single nuclei multiome data. Then we identified 24 DRGs, 4 controls, 6 with fibromyalgia, 7 with rheumatoid arthritis, and 5 with back pain. So, we're planning to do sequencing in the next year or two.

Peter: I'll need to survey my team and get back to you about cell types. I also wanted to mention that for the data we are ready to upload and start the test run. So, I think Elle is here and she'll be in charge of that when the U24 team is ready to go.

Jyl: For cell types, we'd be interested in learning how you are identifying them and naming them. In terms of your request for uploading, you can reach out to me and I can provide the contact information for the curators and we'll set something up.

Xianjun Dong (Harvard Update; Presentation Link)

We have been collecting data for both Project 1 and Project 2. Project 1 is mostly focused on Post-Mortem Brains. We have optimized the protocol for RNA isolation and sequencing and the data has been generated and I'm expecting to be able to review it soon. Project 1 includes:

DRGs: 53 donors
TGs: 189 donors

For Project 2, for surgical tissues, we focused on human neuromas, and there were 102 neuromas from 55 donors (avg 1.9 neuromas per patient), and there were 116 non-neuroma

from healthy control nerve tissues. We have spent time optimizing the protocol to single cell versus single nuclei because neuromas are large and there are enormous cells there. We tried to enrich a homogenous cell population. The protocol needs to be optimized, and we spent a lot of time trying different enrichment methods and isolating the cells-neuromas vs proximal.

For the Data Core, we set up the workflow and SOP of how to work with different sub teams. There is a workflow we set up within the center. We also shared the scRNAseq Nextflow Pipeline with the Data Subcommittee. This Github link for Pain Initiative pipeline is publicly available. We also created a standard of operation of procedure for how the Data Core should work with other teams when they receive an email from the sequencing core, and the next steps regarding who does the downloading and the QC time point. We're still optimizing this version, but I'm happy to share this with the team if that's helpful.

We also developed the metadata schema for Project 1 and Project 2. I believe you've received the link in Peter's email. This is the metadata that we actually used for our project, but you have a link and you can make a clone for yourself and then modify it. There are many updates from our site compared to the Version1 schema that we discussed. We reached a consensus in the Data Subcommittee when I was co-Chairing there last Year.

So, some of the fields like the red ones, or the orange ones, we might like to include in Version 2. But we still use this for our internal purpose practice. A take home message from there is that we want to divide the metadata into subjects that we had already into Version 1 and Sample Table for another table, and then the assay table—the assay table is basically experiment IDs for each type of assay. Whatever assay they do, you should have your own table, because each QC is different for each type of assay for the next version of the metadata

For our site, in terms of assay optimization, I'll talk briefly about single cell or single nuclei. We showed they are largely consistent but some cell types are quite different. We also tried different locations of the tissues, neuroma vs proximal neuroma. Some cells have also shown differences. Again, this cell type naming is based on marker genes, but the community hasn't standardized cell type naming yet and we're happy to join the other cores with that effort. And we have already analyzed 26 scRNA data from Project 2 and the Batch 2 results just came out recently and there are some annotations and marker genes there that look pretty good. In summary, we are ramping up collecting in Y2 and Y3 and Projects 1 and 2 are currently generating data.

Diana Tavares Ferreira (UTD) (Presentation Link)

Diana: For the data collected so far, for the Human DRG, we have done single-nuclei data on 48 organ donors so far. We have done 8 thoracic vertebrectomy, 10 C1/c2 fusion—so those are the surgical samples. We have been mostly using the 10X FLEX kit, but 6 of the 48 organ donor samples are from the Parse V2 kit. We're also trying to compare those 2 technologies and I don't think we have a conclusion on this yet, but one of our goals is to compare these two technologies.

For most of our samples, we got about 8-10%, and we're still waiting for some of the data to be sequenced so that the protocol is completed, but so far we have a reads per nuclei in all of the samples. We are expecting to get more for the latest batch as we are using the Nova6X platform. As a quality control, I did want to highlight that most of our nuclei have less than 10% mitochondria.

For spatial, we have been using the Visium Technology. For some of the samples, we have done Visium V1 technology which is a polya-based technology. More recently, for the latest organ donors, in the thoracic and C1/C2 fusion, we have been using the newer Visium (Visium v2, which is probe-based). But this is still the 55 micro resolution. Again, here our goal is to look at barcodes that overlap single neurons. So, I think for the surgical, because it's___ and thoracic, (22:38), so the DRGs are smaller. So in those samples we get less single neuron resolution, but in most cases for the organ donors we can get more than 500 neurons with single barcodes.

In terms of other technologies, we have also been using ATAC-seq [9 organ donors (w/ bulk ATAC-seq) 13 organ donors (w/ spatial ATAC-seq)]. And then long read sequencing in 3 organ donors. And then for Proteomics, we're still trying to determine the best technology for samples. We tried Somalogic in 11 organ donors and 13 thoracic vertebrectomy samples from surgical patients. More recently, we tried the phospho-proteomics from Kinexus KAM-2000 array from cultures. We just got that data, so I'm not sure which is best

For Human Spinal Cord, we have single-nuclei using the 10X Flex Kit in 38 samples from 19 donors. So we have separated the dorsal and ventral horn. We just recently got the data from the latest batch, so we're still in the process of analyzing, but overall from the 1st samples we were getting about 9% of the cells as neurons and we are getting 10,000 reads per nuclei. And then we still have analysis ongoing for the ATAC-seq for 20 organ donors for the spinal cord.

A brief overview of the analysis. For the single nuclei, we will need to compare with the Harvard pipeline, but we mostly use Cellranger + Seurat v5. For ambient RNA removal, we use Cellbendar, but in some cases we have used SoupX.

Also, for the DRG, we have been using DoubletFinder to find Doublets for the spinal cord. Actually, I don't think we have done that, I think we have been doing manual methods for identifying doublets. And then we use Harmony for Batch Correction, and Seurat for Graph-based clustering. Then for post processing for cell interactions, we mostly use Cellchat.

Then in the Visium analysis, we have a built-in pipeline that we adapted to automatically detect neurons in H Stained images and then identify the barcodes that are associated with those pixel coordinates. Then we follow it up with a QC step to remove low quality barcodes or barcodes that don't express neuronal markers, and then adapt it to a neuronal single cell like Seurat. We are also trying to develop this other pipeline to identify distances between cell types, mostly in relation to the neuronal cell types. We are also working on integrating single nuclei in spatial data.

For some of our samples, we do have single nuclei and Visium spatial data for the same donors. We're hoping to be able to integrate that data in. We are now testing several packages to see which one worked best. For ATAC-seq, the DRG data was posted earlier this year on BIOArchive by Urzula Franco-Enzastiga, a postdoc in Ted's Lab. So you can see details on the analysis there, you can see the preprint. And I think she's initiating the data submission to SPARC.

Then Asta led the work on long read sequencing. This paper is now published in *PAIN* and I think this was the first data set being submitted to SPARC.

Then for other questions you asked about, in terms of synergy/collaboration, Ted mentioned that the Center PIs have been discussing this and want to discuss it more at the next SC meeting. Then in terms of spatial genomic tools, our center is very interested in Xenium moving forward and thinks it has a lot of really good capabilities.

In terms of cell types, the Ginty lab is leading DRG, and we also have the Harmonized Atlas for DRG cell types that's led by the Harvard group. Knowing Ted, I think our only criteria is that it should be based on humans and not mice. UTD is currently working on spinal cord cell types.

DP: Is the Ginty Lab doing nomenclature on human tissue?

Diana: I don't think so. I think they released a genetic toolkit where you can study any particular cell type, but I think it's based on mice. Our center still thinks it should be based on humans.

Wenqin: I think it's going to be an ongoing discussion because for the field, the reality is most of the knowledge previously was generated using mice as the model system. So, I think the cell nomenclature will take time to continue to evolve and that will take some time.

Diana: I agree, and there are some differences. But I hope that with the U19s and everyone working on this data, that we can have a reference that eventually is based on humans.

Guoyan: What's your target sequencing depth? In our data, we observe that you have to have pretty high sequencing depths to be able to detect SST….Because I know in mice, you see SST neurons versus in this Atlas you compare and do this integrated analysis and you find SST neurons in humans. I'm just wondering what's the detection recount in those data.

Diana: With the 10X Flex Kit, because it's probe-based, I think 10X recommends 10,000 reads per nuclei, and we're already going above 20,000. And I think with the latest batch, we are going to have more…I just don't have it yet. But we detect most of our neural markers. So, we are able to compare…most compared with the existing reference set, like the harmonized Atlas for example. We can probably do some comparisons with the newer data set once we have the full data set ready. I do remember you mentioning cell depth, but it's kind of a balance between the neurons—like sequencing more cells to be able to get more in.

Guoyan: I agree. I just want to have an idea because I haven't worked with mouse data. So, I'm wondering in the identification of the human SST positive neuron, if is it based purely on the

integration (the co-clustering of the mouse data) versus if you want to see the SST expression in the population. Because in our pilot data, we sequence to 150k per cell. And now we only see—I can't remember the exact range—but it's something like less than 100 reads per cell in those SST neurons. So, I was wondering if I have like 20K per cell, how many reads do you actually detect?

Diana: That's something we can look at. Because I can't remember if we detect SST. I just know the major cell types, but I can bet back to you.

Wenqin Luo (Presentation Link)

Because this is Year 1 for us, we're still going through a lot of regulatory and administrative processes. I'll focus mainly on Project 1 because that's what I'm involved with. For both Project 1 and Project 2, when we get the Human TG samples, we cut sections using adjacent sections. For the soma part of Project 1, we do this via Laser Capture Microdissection and Smart-Seq deep sequencing and all the spatial transcriptomics, like 10X Xenium and 10X Visium. So those are all about 20 micron section thickness. But in the same samples, we will leave 6 sections for Project 2 so they can do disassociation to get the single nuclei for epigenetic and the nuclei analysis.

For Project 1 have developed a technique using laser microdissection to cut out those neuron somas and then generate deep analysis sequencing libraries. The strength of this technology is that we can reliably isolate the soma of individual human TG neurons. The deep sequencing also allows us to detect much more unique genes. We are working on a manuscript focusing on human TG neurons that was recently accepted by *Nature Neuroscience*. Hopefully it will come out, but any of you are interested in that data we are happy to share among the U19 Centers.

By using this technology, on average, we get close to 10,000 unique genes per neuron soma. So, that's our sequencing depth from the human TG neurons.

We have done pilot experiments with 3 pilot TGs. Using laser dissection, we dissected more than 200 neurons, and we sent out about 60 neurons for the sequencing library. And out of those, we have an average of about 5,000 unique genes per neuron soma. So, this slide is less than what we get compared with human DRG, but I still think it's a very good number.

For human DRGs, in total we sequenced more than 1,000 human genomes and we got really good results. There is ongoing work being done on the migraine donors. So, the control is probably ready to be sequenced, but the migraine one we are still doing dissection.

We tried the 10X Xenium by using Human DRG samples and the human TG samples. Again, we have done 4 human DRG sections. The 10X is a very nice result from the human DRG neurons. Basically we have a cryosection with 2 donors and 2 sections from each donors. We used 100 probes, which is a customized probe. Those are the top mark genes we got based on our single-soma deep sequencing. And then when we did the segmentation, we discovered a problem because the 10X company, the automatic platform doesn't really work well for their segmentation. So for the publication, we used other criteria to manually segment some of the

neurons and did data analysis because we were in a hurry for this publication. So, we are working this summer to develop an automatic platform to outline this area of interest and then convert back for this analysis

So, what we did for the data analysis is we actually used the 10X Xenium to do the cell clustering independently. We didn't map this to our previous work….Without getting into the strengths and weaknesses between reviewing cell types using the LCM single-soma Dataset and 10X Xenium Dataset, I think independently the numbers are comparable between the 2 approaches and the grand truth about the relative percentage or portion of each population probably lies in between the two of them.

Then with the 10X Xenium, we can visualize the spatial distribution of all the 16 different populations of human DRG neurons now in a given section for the first time. But now with 16 different populations and similar colors it's hard to see the pattern, so we also did this one-by-one. So, for given cell type and given neurons we have 500 diameter circles around that neuron center, and we can get the density of that population within that circle and then compare it to the theoretical if it's an even distribution. So, we can calculate that P-Value. So as you can see actually for most of this they do have like about 40% of those neurons show significant clustering together in the DRG sections. We think that might reflect chronogenesis of a given population of neurons. And then another way to visualize this is to use a heat map of the grid. So we are developing automatic platforms and algorithms for this data and we are happy to share this presentation if other groups are interested.

Updates:

Metadata

Peter: So, I just wanted to remind you that we put the original metadata standard used by the PRECISION Network, as well as the reconciled donor and surgical reconciled metadata forms in the Slack Channel as well as an email. Please review them as well as the DCIC questions because this will facilitate the finalization of the metadata. Bryan Copits should be able to add more.

Bryan: I'm happy to answer questions about this. It looks like the ones that were created in June, Xianjun created internally–where you guys have separated the surgical and organ donor cases. If there's questions from the DCIC or Data Subcommittee, I'm happy to smooth over sticking points.

Jyl: From our side we just want to ensure that everything that has been decided as a whole, should be collected by everybody in a standardized way. Also, when you guys are providing your metadata, we want to map it to our standards. We're also interested in interoperability and cross analysis because some of that will hinge on what each of you are doing in common in terms of protocols and then capturing metadata in a consistent way. So, we want to explore where we can harmonize across the groups as much as possible. For metadata, we want to make sure that we understand that when people say that they're collecting something, that we have the same meaning. So, we'll be creating a data dictionary based on those, and then, of

course, people are allowed to share as much as they want on top of that. The other thing about metadata is that HEAL has several required metadata elements, and those are called Common Data Elements (CDEs). So we can help you guys enforce collecting that for your clinical studies as well. So, that's why I'm bringing up some of these things about metadata, and then if there are protocols, trying to get some collaboration there. So, that's also why I brought up the ideas of building on top of what you guys are all doing individually. So, as a group, that's why I brought up the idea about having Map-Core attend.

You had some great spatial resolution experiments there, and I don't know if the group has talked about a somewhat standardized framework for spatial coordination, but those are some things that can help with tighter collaboration and we can help facilitate that here.

Sam: Jules in the chat mentioned that the Human Tissue group was recently asking people to present and the group has recently requested that people present their work, turn their work into SOPS, and then start adding to protocol.io. So, it'd be helpful if people started sharing the names of likely protocols. In terms of future meetings, we want to include Map-Core. Again, Map-Core is based in Auckland and getting them involved is difficult given that they're in a different time zone. That being said, we want to incorporate them, potentially even next month given that one of the U19 teams might have an overlapping conference. They are also currently having some individualized meetings with U19 teams. We may also revisit rescheduling additional Data Subcommittee meetings.

Action Item Summary

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
1)	Share first draft of publications policy	Sam	5/10/24	5/10/24
2)	Share first draft of DUA policy	Sam	5/10/24	5/10/24
3)	Share Next Steps for Visualization Surveys	MAP-Core/ Sam	~~5/17/24~~ 6/5/24	6/5/24
~~4)~~	~~Follow-up with NIH about connecting with NIH Genomic Program Administrator~~	~~Sam~~	~~5/17/24~~	
5)	Connect with dbGaP team to determine best of streamlining data submissions and linking different platforms	Joost	5/31/24	In Progress

6)	Share surgical and donor metatables	Sam	5/13/24	5/13/24
7)	Dat-Core and K-Core to meet to discuss metadata next steps	Sam/K-Core	5/24/24	5/24/24
8)	Determine which members of U19 team need to be at upcoming metadata conversations	U19 Data Leads	5/20/24	5/20/24

1 & 2) Decided to combine
3) Map-Core emailed respondents and is scheduling meetings
4 & 5) GPA Ran Zhang suggested that Dr. Anne Sturcke from NCBI would be the most helpful person to meet with. Dr. Sturcke leads the team responsible for dbGaP submission and data processing and will be best able to connect the team with dbGaP data curators and facilitate further discussion.
6) Shared Xianjin's files
7/8) New PRECISION metadata Slack channel created.

2024-05-10

Attendees: Diana Tavares Ferreira, Joost, Ted Price, Julia Bachman, Peter Jin, Jyl Boline, Guoyan Zhao, DP, Wenqin Luo, Elle Mehinovic, Guoyan Zhao, Ayesha Ahmad, Mingyao Li, Xianjun Dong, Elle Mehinovic, Rachel Weisberg, Hanying Yan, Ish, Marlena Pela Asta Arendt-Tranholm, Bernard de Bono, Tassia Mangetti Goncalves, Julie Choi, Qingru Xu, Ibrahim Saliu, Sarah Rosen

Agenda Overview

- Action Items/Blockers
- Updates
    - Anatomical Survey Mapping (**Sam)**
    - DbGaP (**Joost)**
    - Consortium-Based DUA (**Sam)**
- Output of the PRECISION Data Requirement Summary (**Joost)**
- PRECISION Metadata Discussion (**Peter**)
    - PRECISION_Metadata_template_v1.xlsx
        - Copy of PRECISION metadata feedback from K-Core
- BICAN Data Levels

High Priority Blockers

Milestone /Action Item ID	Issue/ Item	Assigned To	Expected Completion Date	Actual Completion Date

Notes:

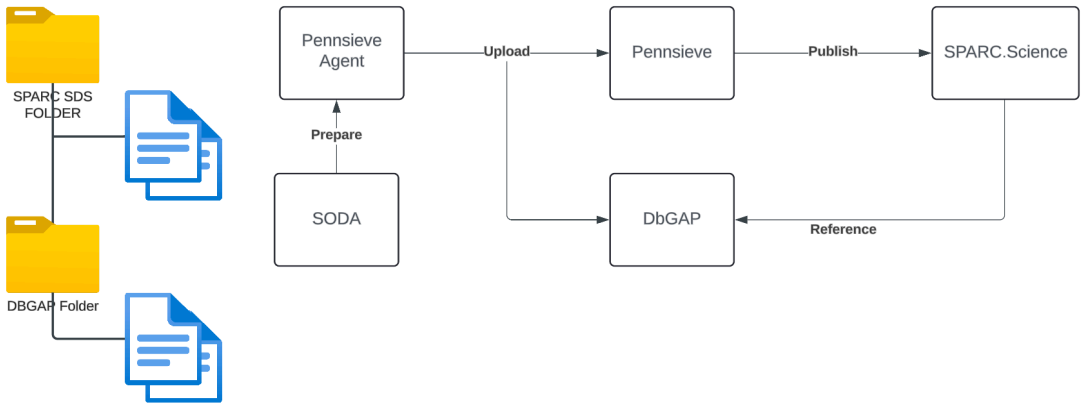
Action Item Summary

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
	Share Metadata/omics questions and comments form	Sam	4/9/24	4/8/24
	Share example of BICAN data levels	Sam	4/9/24	4/8/24
	Schedule Data Subcommittee Roadmap Discussion with Peter	Sam	4/12/24	4/9/24
	Distribute data visualization surveys	Sam	4/9/24	4/9/24
	Share consortium-based DUA example from HubMap, BICAN, and Human Cell Atlas	Sam	~~4/12/24~~ 5/7/24	5/7/24
	Reach out to NIH regarding next steps for dbGaP after reviewing submission guidelines	Joost	4/12/24	OBE

Notes from 5/10/24 Meeting

- Action Items from April Meeting have been completed
- Best Practice is to include affiliation in Zoom name
- Centers should inform subcommittee about new members joining
- Anatomical Survey Mapping Update
    - All U19s had at least one rep submit their survey responses. Sam and/or MAP-Core will be in touch about scheduling meetings with MAP-Core. Currently deciding between group discussion and/or individualized U19 Center discussions.
- dbGaP:
    - Joost met with Diana Ferreira for walkthrough

    - Primary friction is getting sign-off from the dbGaP team around the data dictionary. Need to integrate with DCIC Curation processes
    - Data upload is not as challenging as Joost initially thought
    - The reason we selected dbGaP is because we know that some centers may have more protected data given the type of patient data and because we are working with NIH on long-term stability. One of the reasons we selected dbGaP for the raw data is because it is managed by the NIH itself.
    - Current thinking is that we'll use DbGaP for the very large raw sequence data. We'll probably focus on FASTQ, BAM files to support long-term sustainability.
    - Most of the rest the data would go through the processes we're setting up with the SPARC curation. Very rough draft of current thinking is included below.



    - Questions for NIH
        - Who is the PRECISION Genomic Program Administrator (GPA)? This person should be at the NIH IC and designated to support Genomic Data Sharing. We need to talk to the GPA as this person sets up dbGaP to know what data should be expected. This then gets verified by the NIH program officer (PO).
        - Would it be possible to have people upload to Pennsieve and then move data within AWS to NCBI funded storage?
- GPA should be able to help us determine the expected metadata and files so we can create a plan to determine the best next steps for how we share data with dbGaP. We can then work with the dbGaP team to figure out the best way to streamline data submissions so that we can link to different platforms.
- Several Committee members indicated they thought this approach made sense.
- Ted Price supported getting data on dbGaP, but mentioned that an issue that he's run into is that collaborators, particularly pharmaceutical companies, really struggle to get data off of DbGaP. However, they don't really want the raw data. They want the processed data. So as long as the processed data is easily available on SPARC it sounds reasonable.
- Julie mentioned part of the reason we're pushing dbGaP over GEO is that GEO isn't a HEAL-compliant repository. If data will live on SPARC, that might be a workaround right now, but we're not currently proposing people put their data in GEO.
- If people have questions, they should feel free to reach out to Joost or the U24 team

- Consortium-Based DUA
    - Sam shared slides covering PRECISION DUA and Publication Policy drafts. He'll share the first draft documents in near future.
    - Ted mentioned some of these topics will be discussed during Monday's SC meeting
- Output of PRECISION Data Reqs:
    - Encourage people to read through the document and reach out to Joost with questions.
    - Documented file-types and expected data delivery
    - Clear that there is a lot of emphasis on ensuring that data is actionable after the data is on the platform. Investigators wanted to focus on visualization tools so people can interact with the data once it is submitted. We talked about CELLxGENE and other resources.
    - Need for getting list of 'approved' data formats
- PRECISION Metadata Discussion (**Peter**)

PRECISION_Metadata_template_v1.xlsx

Copy of PRECISION metadata feedback from K-Core

    - Peter & Sam spoke prior to the meeting and felt like a more targeted group discussion about PRECISION metadata and where people should focus, what people need clarification about, and next steps would be helpful.
    - Peter mentioned data leads at each U19 would require their U19's team input to complete the forms and gave background on how metadata sheet was created
    - The metadata template has been touched by multiple people and displayed/talked about during multiple group meetings. Peter mentioned that data leads may currently have some limitations on the data that they can currently access and that they may have limited information about subjects and tissues. This will require a center effort to populate the required columns.
    - Peter is also trying to reach out to Brian Copits and Guoyan to get sample subject tissue information.
    - Peter mentioned that we need to determine how we work with individual-level data and if we want to have one single file combining all the subjects across centers. We should start designating people in each center in charge of collecting/importing information for future usage
    - Joost-Getting this document to the standard that works for dbGaP and also for SPARC makes a lot of sense.
    - Diana mentioned that Harvard had some more updated documents (Metatable_U19_Project2(surgical) - Google Sheets and Metatable_U19_Project1 (donor))
    - Wenquin suggested Human Tissue Core are the people primarily working with metadata and we should really work to figure out overlap between centers.
    - Jyl mentioned that everyone will be submitting data to SPARC in the end, and SPARC has standards that they require, so it'd be good for this Consortium to agree to some required parts while it was understood that certain aspects can be at the discretion of the group. It is also crucial to work from a singular ground truth, a data dictionary essentially so that we all know we're entering the same thing and that it's human and machine-readable.

    - Joost will connect with curation team to get to a final version as soon as possible, and then Curation will reach out to the teams to set up a single person who speaks for their team to verify/validate that this applies to them
    - Jyl mentioned that SPARC standards for metadata are pretty basic. It's very likely that groups already have what's required. Then the consortium can determine what additional metadata to require. On top of the bare minimum, we're trying to determine what additional data we'll require from the groups to increase the interoperability amongst the groups. That's what the DCIC is here to help facilitate, and then further down the road we can work on visualizations. HEAL has requirements for the human subjects for the CDEs.
    - Julia: There is metadata, some is required, some is recommended/nice to have, and some are optional. So when you review the requirements for the human subjects, there was actually very little in terms of what every group knew they could actually get. So the minimal data standard for the metadata isn't too cumbersome. Beyond that, it's really about balancing what you want for your U19's studies and maybe cross-center stuff, but the main point of this consortium is to generate data sets that the entire community can use. Having a metadata standard that's at minimum able to be reused and reanalyzed more broadly needs to be kept in mind. It's important to remember SPARC standards and file type standards and that's why one combined sheet might be best approach
    - Xianjun displayed Metatable_U19_Project2(surgical) - Google Sheets.

2024-04-05

Attendees: Joost Wagenaar, Peter Jin, Diana Tavares Ferreira, Xianjun Dong, Ted Price, Mingyao Li, Wenqin Luo, Maryann Martone, Julia Bachman, Anthony Juehne, Jyl Boline, Hanyin Yan, Ayesha Ahmad, iPhone, Ish, Anthony Cicalo, Suzanne Tamara, Zitian Tang, Elle Mehinovic, Khadiah Mazhar, Ibrahim, Saliu

Agenda Overview

- Admin
    - Data Subcommittee Name change on Slack ("data_core" to "data_subcommittee")
    - Google Drive (folders transferred from Huddle to Google Drive)
    - Peter Jin nominated new Subcommittee Co-Chair
- U24-U19 Requirements Gathering Update
- Data Sharing Discussion
- Action Items/Blockers

High Priority Blockers

Milestone /Action Item ID	Issue/ Item	Assigned To	Expected Completion Date	Actual Completion Date

Notes:

Action Item Summary

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
	Share Metadata/omics questions and comments form	Sam	4/9/24	4/8/24
	Share example of BICAN data levels	Sam	4/9/24	4/8/24
	Schedule Data Subcommittee Roadmap Discussion with Peter	Sam	4/12/24	4/9/24
	Distribute data visualization surveys	Sam	4/9/24	4/9/24
	Review Harmonized Atlas Paper	Joost	4/15/24	OBE
	Share consortium-based DUA example from HubMap, BICCN, and Human Cell Atlas	Sam	~~4/12/24~~ 5/7/24	5/7/24
	Reach out to NIH regarding next steps for dbGaP after reviewing submission guidelines	Joost	4/12/24	OBE

Notes:

- *JW: [showing slides on data collected]* I wanted to use today as a working session to discuss the right data to share, think through data formats, and consider how we can harmonize data across the different centers so we can build functionality on top of it in terms of queries and search. We sent surveys and had 1:1 meetings with each center and worked on collating it and making it actionable. We also want to talk more about data visualization and how we integrate with other platforms. Surveys will also be sent out soon related to anatomical mapping and metadata.
- *JW:* Some groups sometimes capture that original data not within NWB file format (Neurodata Without Borders) but indicated that they are planning to share that part of the conversion into the file format. This is something that we'd like to promote. So, if we could agree on transforming all the electrophysiology NWB data prior to uploading it to our platform, that'd be great. If there are cases where that is not best from your perspective, we should gain a sense of what sort of other file formats we should expect in those cases.
- *XD:* With the single-cell nucleus sequencing listed [*on this slide*], it is hard to tell whether it's RNA or DNA right now.
- *JW:* Yes, so this is my best interpretation of this right now, but if I should avoid collation or split multi-omics out, then we can do that. What's important for us to know is what sort of file formats these come in.
- *XD:* The fastq. file is very nice and a must have, but do you also want the process file (like the single-cell H5 file for single-cell RNC which is way smaller)? I see you don't list anything for spatial files (I know it's an ongoing discussion), but we often use 10X Visium and they often have both image and the raw fastq file together. And for the long-read sequencing you put .BAM, which is a post alignment file, not the real raw format. So, if you prefer fastq., we should also put it for long-reads as well.
- *JW:* Yeah, so this is coming from what you originally provided to us, but if as a workgroup, we review and decide on standardizing on fastq. or something else, then this is the best place to do it.
- *XD:* Yeah, that will depend on who the user of the data is and whether they want to work with raw data and do the analysis themselves or want pre-processed data. The .BAM file is great for some purposes too. It's much smaller than fastq. or raw data. A general question is what kind of data we want files in and who our target audience is for using this stuff?
- *PL:* Do you want to have a suggested naming pattern for each type of data? So, for example, "whole genome."
- *JW:* Yeah, this is something that I was going to bring up, because if we know what types of data and types of file formats we have, how does that then fit into a standard in terms of uploading files and organizing the files? One thing that I think HubMap does really well is that for each file format, they have a specification for how everything is stored within a dataset. I was wondering if that's something we should adopt here. Because these files [*that the U19s are working with*] are quite different and not as simple as a single file, So, I know that HubMAP has pretty good guidelines for all these different data formats or data modalities in what they expect to be uploaded and where that is within

the dataset. Obviously, we can deviate from that if people think it's too cumbersome or not necessary. That is one of the things that I think we should discuss here.

- *MM:* The BICCN Consortia has developed this concept of data levels and when and where they should be shared or not shared because they have a requirement that they need to release their data on regular cadences. They go through raw/unprocessed data to level 1, to level 2, which has annotations and more types of data with it. I'm happy to share information with this group. There are a core of things that people from HubMap, BICCN, and Human Cell Atlas are starting to agree on and it would be great for HEAL to try to adhere to that.
- *XD:* I agree with the proposal to define the data in the different levels/permissions. I think other cohorts like dbGaP use similar methods. Like a fastq. needs a higher level of data use agreement while other post post-processed data are less restricted.
- *JW:* Yeah, based on the survey responses, fastq. are mentioned as the output of the project. And this goes towards a question for the NIH…We can make a decision that we just want to do .Bam files or some post process data, but I do think there are probably some requirements around having some of that original data available.
- *DT:* For the PacBio long-read sequencing, the output is actually .Bam files and TBI files. So, if you want TBI. So you'd have to convert .bam to fastq. So, I think that's why we shared the .BAM. And in the single nuc sequencing, I think we will also be doing the 10x for the technology.
- *JW:* I know there was some other data that was presented as the output that was more survey or patient-oriented text-based, but I intentionally left that off here because the majority of the effort will be providing the support for these major categories of large data in terms of primary vs derived data.
- *PL:* For PacBio long-read, generally the files produce an H5 .Bam file and fails, but sometimes the failed reads are useful. But they are both very large files. Just something to consider.
- *XD:* I agree with Peter. We often see the H5 format instead of .Bam one.
- *JW:* So, I did a deeper dive in the CELLxGENE and they do require their data to uploaded as an H5ad file and I know that some of you have uploaded some test datasets in this format which seems to be some sort of combination of the raw data, plus some of the embeddings and associated metadata in a single file. If we're interested in building a visualization tool like this or we're thinking about integrating into the CELLxGENE platform, we'll need to think about whether we're willing to adopt this format as some sort of standard that we're requiring for data that is uploaded by HEAL PRECISION.
- *JW:* Is that something that seems doable? We only get real benefit out of this if we make this data actionable. I would love to move towards some type of visualization and tools around it, but if we really need to figure out the standard file that we require to be uploaded or what are the processes to convert into that file format so we can build tools on top of that. Looking at CELLxGENE, there are different ways we can approach this. Their software is open source so we could make our own CELLxGENE tool running within the SPARC portal or within HEAL, or we could reach out and see if we can integrate that in some way into CELLxGENE once this data gets published. I'm not sure

if that's something they'd be interested in as this would be slightly different from the data they support, but I'm happy to email them. So, which groups have worked with this H5ad format and where is the opportunity to standardize around that if that is the outcome of this conversation?

- *DT:* I'm less familiar with it.
- *AC:* It's fairly simple to take an R-object and transfer it. We've written some scripts for it that we can share.
- *XD:* We really like CELLxGENE and we tried to install this in our local server and test our datasets for other projects and it worked pretty well. We'd also like to invite the U24 to join one of their meetings so you can see the work and collaborate. We not only have a single cell, but also ATAC Seq and spatial data. Right now Spatial Transcriptomics data is a really hot topic in the field, and I don't see many available tools. It would be helpful to hear thoughts on that. In terms of whether we should use H5ad, it's a post-process format not a raw format. It's already annotated cell types. If you go to the CELLxGENE browser, you can pick the cell types and subcluster them. So, it's a post analysis result. So it's not the same as Fastq. Fastq. is a raw file and it's a long process to get it from pipeline to h5ad.
- *JW:* If you have a fastq. and you turn that into an h5ad file, how standardized is that? Is that something that we could hypothetically take a pipeline for and run that after files are uploaded automatically over old files or is there a lot of customization there that is specific to your parameters?
- *XD:* During the process there are many steps that require experts. For example, annotating cell types requires a lot of work and sitting down with biologists and having them look through markers. Also, each lab has different parameters for the protocols. The QC standard might be different. Also, the single-cell/single-nuclei might be different. So, among different U19s we might have different opinions on parameters to use for different pipelines.
- *WL:* Yes, I agree. I also wanted to mention that I looked back and saw that our Electrophysiology is not in NWB format. If it's a cell type assignment, that is a very tricky process that takes a lot of thought from data scientists and biologists. So that would make standardization quite difficult.
- JW: So, if these are different and we rely on investigators to do this before submitting it to the U24, is there any concern that merging the data into a single visualization would be difficult? Or would we still be able to work towards something where, at the end of the pipeline, we could look at the data in aggregate? Or can we only look at data in the context of the team that collected that data?
- *ML:* I think cell type assignment is a very challenging task, it really depends on what level each team wants to annotate. For visualization, it will probably be best to do it by data generated by each team separately as it's going to be very difficult to combine all of them in one visualization.
- *PL:* We have to spend a lot of time trying to reconcile single nuc, ATAC seq data. We agreed that we would use the Nextflow language as a starting point. But I don't think we want to go back and discuss that. Anthony, can you share a link?

- *AC:* Once you have a processed file, you could probably automate that to where it becomes an h5ad file and you can visualize that. I think having that, even before annotations, you can click the marker genes and see where they highlight on UMAP.
- XD: Great point. Peter to your point, I think single-cell nuc seq pipeline is more ready to share than single cell patch-seq. We are working on a definition and discussing those kinds of components right now but we'd like to share as soon as possible with the group.
- *JW:* Yeah, that would be great. And when we're a little further along, I'd like to work with someone from your team so I can see where there is an opportunity to run something that is uploaded. Like we have the opportunity to run workflows over data in the cloud. So especially around standardization, for something like an h5ad, I'd be very interested in exploring whether that's something we can automate and standardize on our side. But it seems like we're a little early to dive into that.
- *PL:* Is there a way to know which dataset has been uploaded to ACC? That would be helpful for us to know how to do this general practice and know what data is available. Should we set up a temporary deadline for uploading some of the data sets?
- *JW:* Yes, we have some test data that has been uploaded that hasn't been shared yet. But we need to explore whether there needs to be some consortium-based DUA in place for data sharing. Anyone that uploaded data to our platform can share data with anybody else, but we don't yet have a system that automatically shares it with other teams. We are currently talking to the NIH about this. We had a similar process for SPARC, and had each team sign a DUA to be part of the consortium and allow more sharing.
- *JW:* It'd be helpful to have an example of all the expected data types on the platforms so that we can set a ground-level roadmap. This helps us prioritize what we're working on. And one of the things that Maryanna's team is working on is the organizational structure The sooner we know the data that we expect, the easier it will be to know how this fits in with the SPARC data standard and folder structure standards. I'm happy to put a date on there. Would next month be too soon to have a representative dataset example from each of the teams available too soon?
- *TP:* The Harmonized Atlas paper from Will's group is probably a good starting point and there's data from each of the labs that was generated from the U19s. And it's all single nucleus DRG and they have different mixes of methods, but I think Sham's ran it through the same pipeline. We also have other DRG datasets that we'd be happy to upload that are from the technique that I think we landed on which we'll be using most of the time, which is the 10x. As far as DRG, I think Will's paper is a good starting point.
- *JW:* Yeah, that sounds great to me. And for the other data types that are shared in this chart, including some of the electrophysiology, it would be great to have examples of that as well. Because we're focused on the single cell right now, but expecting some other types of data to come in will be helpful. So, maybe one of the things we can do is to send an email request to update the data survey that they filled out and then aim to have each team upload that data as a platform as a test data. Simultaneously, we could work on the data use agreement so that we can figure out how to harmonize that and where overlaps are as a team.
- *JW:* Another thing that came up in DAT-CORE/U19 conversations is supporting the uploading/submitting data to DbGaP as part of the end location for the raw data. We

really need to figure out which part of the DbGap submission process is feasible for us to help with, but I want to work with the Penn team to figure out what this process entails and then where we can tie this into an automated process. We can also consider if there are components where we can facilitate some of the standardization or the creation of the files that then can be more easily updated by individual investigators. So, there are still a number of things that we need to figure out, and separately we need to work with the NIH to ensure it's captured in our milestones. So, it's probably unrealistic to hand this off too soon to us, but if you have thoughts or experience with DbGaP, please reach out to me cause it'd be great to have a discussion to incorporate some of that effort.

- *TP:* Yeah, I think that would be great. Many of us set up accounts with DbGaP, but haven't started to upload the data. But if we could figure out a way to facilitate that, it'd be very helpful. Because it's been a laborious process in the past.
- *JB:* Yeah, on the NIH side we're working with Dr. Ran Zang to determine the best process. DbGap suggested that they keep all four centers as registered studies in DbGap, and then if we get to the point where the U24 is able to submit on the centers' behalf, the U24 would be added as a submitter for each of those four centers. Another thing that they strongly recommended was changing the study names of each of the 4 studies to start with a prefix of "HEAL PRECISION" to make it easier for linking in the future. So, unless there is an objection to that, then we can go ahead and do that. Also, the earlier we can figure this out, the better.
- *JW:* So, I think there is agreement that we should leverage dbGap and that we should use U24. So, Julia, I'll follow up with you separately to figure out our next steps for that. So, I think over the next we can figure out what we can expect in terms of file formats and then create some sorts of proposals around creating certain standards such as the neurodata without borders.
- *MM:* I just wanted to mention that curation had viewed the metadata/omics that we had gotten from you and had some questions/comments. So, I put the link in, but can have Sam distribute it. We can also determine if that needs to be discussed and put that on a future agenda.

High Priority Blockers

Milestone /Action Item ID	Issue/ Item	Assigned To	Expected Completion Date	Actual Completion Date

Notes:

Action Item Summary

ID	Action Item	Assigned To	Assigned Completion	Actual Completion

			Date	**Date**
1	Continue discussions amongst groups in Slack in between monthly meetings	Xianjun	1/26/24	Complete
2	Follow up with each group on establishing DUAs	Sam	2/29/24	NA
4	Discuss improvements to the metadata sheet with the HTPP core	Xianjun/ DCIC	~~Feb Meeting~~ ~~March Meeting~~ Will revisit after SOPs finalized	

2024-01-24

Time: 4pm
Attendees: Julia, DP, Diana Tavares, Xianjun, Guoyan, Asta, Ayesha, Sam Kessler, Peter, Bijesh, Rachel, Huma Naz, Wei, Khadijah, Anthony, KBoyer, Gabi, Urzula, Hanying, Mingyao, Suzanne, Megan, Elle, Jenna, Nikhil, Dana, Hao

Agenda Overview

- High Priority Blockers
- Action Items
- Admin:
    - Intro new PM Sam Kessler
    - Leadership of Data Subcommittee group will transition to co-Chair with the DCIC and a U19 leader
- Feedback and comments on the metadata template v1 from the practice use — 20min
- Interaction with U24 — 10min
    - Follow up on the PRECISION DCIC Survey 1: Data Types
    - DUA progress
- Review metadata document examples from other consortia previously posted on Huddle and Slack: BICCN (Guyana) and KPMP (Peter) — 20min
- Other items?

High Priority Blockers

Milestone /Action Item ID	**Issue/ Item**	**Assigned To**	**Expected Completion Date**	**Actual Completion Date**

Notes:

Action Item Summary

ID	Action Item	Assigned To	Assigned Completion Date	Actual Completion Date
1	Continue discussions amongst groups in Slack in between monthly meetings	Xianjun	1/26/24	
2	Follow up with each group on establishing DUAs	Sam	2/29/24	
3	Review metadata document examples from other consortia previously posted on Huddle and Slack: BICCN/KPMP	Guoyan	1/24/24	1/24/24
4	Discuss improvements to the metadata sheet with the HTPP core	Xianjun/ DCIC	Feb Meeting	

- [1]: Need to start discussions in slack closer to meeting end
- [2]: JBB: will be passing this to Sam. Will be discussed at the 1:1 Dat/ U19 meetings
- [3]: Review metadata document examples from other consortia previously posted on Huddle and Slack: BICCN (Guyana) and KPMP (Peter) – Guoyan screenshare on the BICCN metadata Submission. Reviews the workflow. They like the stepwise organization. Metadata table - sample inventories
    - Data collection inventory (similar to the one PRECISION has already)
    - Sample inventory (similar to PRECISION but less info)
        - Comments from wetlab technicians are that our metadata is too much and simplifying it would be best
        - Sample metadata table not template; different from subjects
        - Julia via chat: I think Diana based the our version 1 template on one of the BICCN metadata tables that I had previously uploaded to Huddle
    - Goal is to get an idea of what other consortiums are doing
    - Next step would be to compared to the SPARC existing metadata
    - Julia: NIH steered you towards splitting the sample and subject
    - Anthony: alignment of variable level CDEs meta data. DCIC working behind the scenes with HEAL Platform team and data stewards to
    - Xianjun: version one does split the metadata down by protocol and different assays and then also split between donors and surgical patients
    - JBB: next steps will be to compare the
    - Julia: Maybe we need to take a step back and revisit the goals of this group. Look at minimum required vs need to have; look at Anthony's slides from a year ago about thinking about things
    - Diana: are we going to have common metadata for all data types or metadata?

        - Julia: both
    - Julia: this is why I wanted to steer the groups towards the rna seq metadata as that is common across most u19s. I think there is only one U19 doing patch so they can do whatever they want though I encourage them to look at the standards out there
    - Peter: screenshare.. KPMP metadata is split by technology. Very minimum header/columns for users to populate. PRECISION metadata covers all of it. It will depend on how we want to display our metadata.
    - AI: share sparc metadata templates
    - Julia: Just to keep in mind that these other networks are using the same protocol but we do not
    - Diana: the sample and subject metadata may have more overlap and differ more in the analytical metadata
    - Peter: maybe we can defer to the sites that are the only ones doing that data type
    - Asta
    - JBB shared:
        - SPARC Dataset Submission Walkthrough
        - SPARC Dataset Structure (organization)
            - More detailed file organization info
            - Also see links within the above articles
    - PRECISION Metadata v1
        - HTPP: first for tabs
        - DAT: first for

Notes

- Feedback and comments on the metadata template v1 from the practice use — 20min
    - V1.5: Metatable_U19_Project2(surgical)
    - Feedback from HTPP (Bryan)
    - Assays
        - Include the red field as recommended from tissue core (Dr. Dong's team)
        - Red columns are required; black columns are recommended.
        - Blue is the data analysis QC
    - What do other people think?
        - Enter some user control (not free text)
        - I.e. Sample type (is free text and not useful) ensure the column name means the same thing
        - Julia: step one was deciding on the metadata elements - step 2 might be creating a data dictionary so that you don't have so much free text or string categories
            - Work with U19 on sanity check
    - Review of dataset
    - AI: share omics dataset existing on sparc