**PsychENCODE Consortium Workshop**

February 9, 2016

6001 Executive Blvd, 1st Floor Conference Rooms C, D, & E

Bethesda, MD

8:30 AM – 5:00 PM EST

**9:00 AM – 11:25 AM: Open Session Presentations**

**9:00 – 9:10: Welcome and Introductions** – *Geetha Senthil, Ph.D. and Thomas Lehner, Ph.D., M.P.H.*

**9:10 – 9:20: Transcriptional and Epigenetic Signatures of Human Brain Development and Autism & Functional Genomics of Human Brain Development** – *Nenad Sestan, Yale University*

**9:20 – 9:30: Cis-Regulatory Epigenome Mappings in Schizophrenia** *– Schahram Akbarian and Pamela Sklar, Icahn School of Medicine at Mount Sinai*

**9:30 – 9:40: Genetic variants affect brain gene expression and risks of psychiatric disorders** – *Chunyu Liu, University of Illinois at Chicago,* *and Kevin White, University of Chicago*

**9:40 – 9:50: The USC PsychENCODE Consortium** – *James Knowles and Peggy Farnham, University of Southern California*

**9:50 – 10:00: Gene regulatory elements and transcriptome in iPSCs and embryonic human cortex** – *Flora Vaccarino, Yale University*

**10:00 – 10:10: Epigenetic and Transcriptional Dysregulation in Autism Spectrum Disorder** – *Daniel Geschwind, University of California, Los Angeles*

**10:10 – 10:25: Break**

**10:25 – 10:35: Long non-coding RNAs in gene regulatory networks underlying Autism** – *Dalila Pinto, Icahn School of Medicine at Mount Sinai*

**10:35 – 10:45: Decoding schizophrenia-From GWAS to functional regulatory variants** – *Gregory Crawford, Duke University*

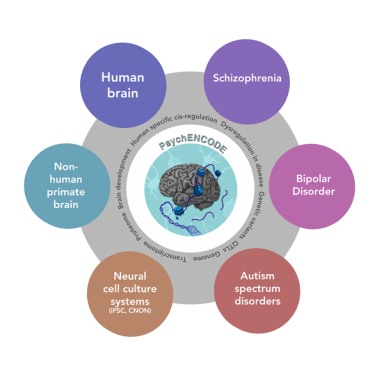
**10:45 – 10:55: RNA Sequencing of the Limbic System in Bipolar Disorder** – *Peter Zandi,* *Johns Hopkins University and Thomas Hyde, Lieber Institute for Brain Development*

**10:55 – 11:05: Establishing comprehensive and quantitative maps of DNA methylation in the developing brain** – *Andrew Jaffe, Lieber Institute for Brain Development*

**11:05 – 11:15: GABA Epigenomes in Autism** – *Stella Dracheva, Icahn School of Medicine at Mount Sinai*

**11:15 – 11:25: Synapse and Website Demo** *– Mette Peters, Sage Bionetworks*

**11:25: Adjourn**



**PsychENCODE Consortium Workshop**

February 9, 2016

6001 Executive Blvd, 1st Floor Conference Rooms C, D, & E

Bethesda, MD

8:30 AM – 5:00 PM EST

**11:25 AM – 5:00 PM**: **Closed Session (Consortium members and NIMH Program staff only)**

**11:25 – 12:05: Lunch**

**12:15 – 1:20: Consortium-wide reference tissue project progress updates** – *Nenad Sestan and Schahram Akbarian*

* **Nenad**: Schahram and I will start the first part of the afternoon, which will be on Consortium-wide reference tissue project progress updates. We have two more sessions after that one on data coordination and analysis core activities. I didn’t realize that there were many questions for Mette about Synapse. Pamela, Mark, Zhiping, and Kevin will talk in that session and that is 45 minutes. And, then we have a longer session which is a Consortium-wide analysis, joint and collaborative effort. Schahram and I will give you an update on the Consortium reference tissue project. We have two goals with this; the first one was to construct a reference brain map. The idea was to use a fetal and adult specimen. For the first goal, was to create a reference assay platinum brain and transcriptome was genome and so we will do that on any sorted or any NeuN cells and tissues plus some pilot brains. The second was to facilitate a cross-comparison analysis and to do that we were also distributing for this case only adult specimens and 9 samples. We have allocated multiple adult brains. These are all complete brains that are in my freezer. We will be going down the list and will be shipping tissue to Schahram, and we have allocated 4 mid-fetal specimens that are also in our freezer. Fetal would again be NeuN – and as well as tissue samples from the same pulverized samples and fetal would only be tissue specimens. So, that is the current status. If you need tissues, let me know besides this.
  + **Dan**: Can I ask a quick question? Clinically unremarkable? Is there a cause of death?
  + **Nenad**: Usually accidents or something like that.
  + **Dan**: I see. So, they have no chronic diseases?
  + **Nenad**: Yes, that’s true. So, we do check these brains just to let you know that we have some sense that they aren’t like gliosis or something bad.
  + **Andrew**: I think we are doing DNA-Seq on them, so we’ll be able to call that.
* **Nenad**: Moving forward, what we are doing now at least for adults, Schahram’s lab has been isolating nuclei. He has received specific protocols for each of them. Pilot samples means that they are not from the same brain, so you can mess up and stuff like that. And, for those who will remember for fetal we are receiving tissue, not nuclei, and for the adult you will also receive tissue that I will provide. There is one thing that we will unfortunately, before the supplement; we were unfortunately not able to collect adult fibroblasts. We have the brains and we are not going to postpone this. For the fetal ones, we have slightly more but not intact and unfortunately, we only have 4 and collection has really dried out because of political situations we are under.
* **Schahram**: This slide summarizes what we had discussed by e-mail, the options in some ways in terms of sorting and, more important than sorting after the NeuN + and -, is the issue of how to store the nuclei because with the nuclei are, when you do FAC sorting each nucleus comes in a little volume of FAC sorting solution. So, you have to ask, in some ways, the choice of depositing the solution or pelleting the nuclei first and then sending it out to the investigators. I think the safest one in terms of yield; maximizing the number of nuclei recovered after FAC sorting is associated with option number 2, which is our classical protocol. In essence, you’ll notice here we’re adding CaCl2 and Mg(Ace)2 provide the right ionic concentration medium to maximize pelleting yield and then we ship out the nuclei. So, I think that most of you want the nuclei. This summarizes the sorting that already happened for these various reference brains. As we speak, we have started to ship out things. Most of you should have received the FedEx or UPS from Royce Park who is organizing all this.
  + **Nenad**: The sorting is not trivial.
  + **Schahram**: It is very time consuming. To be honest, I was regretting that I signed up for this, but we are up and running with this. I am not sure if we’ll get done within a year everything that we proposed.
  + **Nenad**: One other important thing, don’t ask me, ask Sirisha. If you ask me, there is a good chance nothing will get done. So, if anyone needs tissue, send me an e-mail and I will cc Sirisha. She’s our project coordinator for PsychENCODE in the lab.
  + **Geetha**: So out of eight brains, we have one done for sorting? And, option 1 is pulverized tissue, right? Option 2 is sorted nuclei?
  + **Schahram**: Yes.
  + **Geetha**: How many groups are getting pulverized tissue and sorted nuclei. How are we coordinating on that?
  + **Nenad**: Every group will receive Option 1, which means pulverized tissue, except for fetal ones which will just be pulverized. There will be no sorting and that Sirisha will send to you directly. The reason we did not send yet is that I want to be sure everything is in place because once you get that tissue there is no going back and getting more tissue. For adult, we can always get more tissue.
  + **Kevin**: Did you already send out an initial set to all the groups? I am a little bit unclear.
  + **Nenad**: No.
  + **Schahram**: There is one important thing. The really limiting step I’ve noticed is the communication between all of us. From our division, Royce Park sends out a communication to your team on what address or where to ship, give me a date…
  + **Kevin**: I guess what I am getting at is, can we have a tight well-documented process, maybe a little flow chart that everybody has and everyone sees that also has maybe some goals in terms of deadlines, when to expect what?
  + **Schahram**: What is even way more important is that you give us a name of a reliable worker in your lab and make sure you always cc-ed on our e-mails. That is what will work the best.
  + **Kevin**: Maybe we need a list of that. It would be nice to have a public webpage or something that shows us what the process is and has everyone listed with their contact information. Maybe that would help us coordinate and speed things up so that we’re not playing this game of e-mail tag with a dozen different groups.
  + **Nenad**: I think that’s an excellent idea. I don’t know how to do that. Should we a wiki?
  + **Mette**: Yes, I already started this in Synapse. So, I communicated with Royce and Sirisha.
  + **Nenad**: What I want to say is this: I think the most important thing is to keep people in all of this besides Schahram and I and Royce and Sirisha in the lab, two key people. Because Sirisha has access to all specimens in our lab. She is the one to send samples to Royce and for tissue she will be the one sending the pulverized to you.
  + **Geetha**: Since we have Synapse liaisons could we be in touch with liaisons for this project as well? It seems like you are getting good responses from them, Mette.
  + **Mette**: Right if it is the same people that have been uploading and documenting data for the data deposits then we have a list of those people in Synapse.
  + **Kevin**: Synapse has data wranglers and what we are talking about here is more sample wranglers. We need just a point sample wrangler.
  + **Nenad**: Right, so that would be Royce in Schahram’s lab and Sirisha in our lab.
  + **Peggy**: Can we just add the e-mails to this table you have? If everyone has their contact e-mail for the person in their lab, we could all see this table that you have then our contact people can contact others if they have questions. They could be discussing amongst themselves by just adding another column to this table and sending it to everyone would be very useful.
  + **Geetha**: I agree with Kevin with having a flowchart with the goals and the timeline would be very helpful. And also, we may want to discuss if sorting is a very limiting step here, should all the groups go ahead with using pulverized tissue as a first step, test their protocols and how they can sort, and then later on how to apply to sorted nuclei as they are shipped, or do you want to do this in a batch mode where groups can consolidate all 8 brains and do it at once? What would be the right thing?
  + **Nenad**: The only reason we were hesitant to do that earlier was because when you sort, you want to be sure you don’t waste tissue. But, considering where we are in the project, I think that’s probably the best idea.
  + **Schahram**: I think that is the best idea. The tissue is going to be sorted anyways.
  + **Geetha**: We would like to get comments on that.
  + **Flora**: Can I also raise another issue? For us, the fact you can’t obtain the dural fibroblasts from the adults, if it’s essential, the main goal of our supplement or proposal would be to compare iPSC-derived organoids with the brain itself. If we don’t have fibroblasts from which we can make iPSCs then if there is someone else who has fetal tissue, I don’t think we technically need to do it. We could concentrate on the fetal and get more data on the fetal.
  + **Nenad**: I have been collecting fetal tissue; it’s just that I have not received anything in the last several months.
  + **Flora**: But for the fetal tissue, do you have the fibroblasts or not?
  + **Nenad**: No, for the old one I have that and I can share with you, but in the last several months I cannot get even anything that is remotely usable when it comes to fetal tissue. And, the reason for that is, as you know, not to go into details, every place I’ve been collecting it’s just been hard to collect them and the way they collect the tissue it’s just really not usable for what we would like to do.
  + **Flora**: So, for the fetal tissue, you don’t have any tissue to share right now?
  + **Nenad**: No, fetal tissue we have. These are 4 brains that I have. Unfortunately, this is from my already existing collection, and I was very optimistic when we got the supplement.
  + **Flora**: So, you don’t have the fibroblasts, not even for the fetal? Neither for the fetal, not for the adult specimens.
  + **Nenad**: For fetal one, no I don’t have them because the one we had we already gave you a piece of the tissue. The 4 fetal we have now are intact brain, but we did not think about collecting fibroblasts is what I am trying to say.
  + **Geetha**: Should we, as a group, discuss and vote on the options we have in front of us? That is, go ahead with pulverized tissue first, all the groups for 8 different projects and process the sorted nuclei in a batched mode later on once Schahram has all the samples ready? Would you all agree with that? Can I see a show of hands, please?
    - **Peggy**: I don’t think everyone has to do it the same. We can use frozen tissue now because we still have to do the analysis to see which of the options is better for nuclei. But, maybe other people are ready to go with nuclei.
    - **Geetha**: I am thinking the same thing, Peggy. Go ahead with the pulverized tissue and wait at a later date for the sorted nuclei.
    - **Schahram**: Right, the big question of the supplement was reference tissue, analyzing the very same subject in brain tissue across labs and assays and so I would say we should go ahead with the homogenate and while we keep shipping and sorting later. I am just strongly agreeing with what Geetha suggested. The rationale was to do the homogenate and get the reference project moving.
    - **Nenad**: Yes and what we will do is I will have Sirisha send you all at once, all specimens.
    - **Geetha**: Any comments?
    - **Schahram**: Geetha, I think you wanted to write down that the group wanted to go ahead with the reference tissue project or not because with the show of hands I thought everyone were afraid of fessing up. Are we doing this or not doing this?
    - **Jim**: I don’t know if I’d do Nome-Seq on the bulk tissue this month and then when I do my nuclei six months from now whether those data will be comparable. They might be, but we just don’t have any data on them. If we could even do one first and then do a second bulk later or something like that and then figure it out.
    - **Stella**: Is it possible to not pulverize the tissue because, from our experience, we get much better results with ChIP-Seq than we do with pieces of tissue.
    - **Nenad**: The answer is yes. That is why we dissect the whole brain. We have not yet dissected fetal just because once it is dissected there is no going back. I am happy to send a chunk of tissue if people want to do the pulverization themselves. But, then again we are introducing variations and the goal was not to introduce variations. I think we should stick with sending the pulverized tissue just so that it is processed the same way. If I send someone a chunk of tissue, you could have less or more brain matter that would lead to introducing bias into the data. So, the idea of this reference is again the goals, what were the goals…?
    - **Stella**: You can still sort the nuclei and it doesn’t matter the ratio of the glia to the neurons.
    - **Nenad**: I understand, but for example, we sort the nuclei in the lab, so why do I need Schahram? The reason is you either standardize or don’t standardize it. But, if you don’t standardize it there’s absolutely no reason to do the second thing. That’s the goal. And, so I think we need to stick with what are the two goals of why we are doing this. One is to create a multi-modal reference on the same samples, so if I look at Greg’s ATAC-Seq and I look at some of my histone marks I can be sure I am not looking at a rotten apple and good apple. We can get the same in the platinum brain. I think it’s important to have the same specimens and now for cross-comparison it’s even more important to have that and how we process it. I do agree with you that either you send it all now and have someone sequence all the tissue and then sequence six months later or a few months later, nuclei does this, and that could introduce bias. So, again, what are our goals? We have two goals and I think we have to stick with the original plan just to make sure we are not introducing additional bias. So, I am for sending pulverized tissue rather chunks of tissue.
    - **Jim**: I think doing it right versus fast, unless Program feels differently, then that’s their call I suppose.
    - **Geetha**: Fast is not necessarily the right thing, I agree, but what how long would it take to the do the right thing. What time frame do we have in mind because this project started in the summer of last year?
    - **Nenad**: It was September. I am always for doing it careful and taking time and just make sure that we do a good job. I mean this is a really Consortium-wide effort, so I completely agree with Jim.
    - **Geetha**: We did have a conversation before we put together the supplement this particular common experiment and you all agreed that there were many layers of opinions and you gravitated towards one consensus. That is the plan in the supplement. And, that is for ATAC-Seq, we said we should not do it on nuclei, but we should do it on homogenate tissue. And, for RNA-Seq let’s try both and for whole tissue that is what Chunyu said. We do have an idea of who does what and now we know a lot of information we haven’t had. It’s just coming up with a layout of the plan and setting some goals so we don’t stray away.
    - **Schahram**: So, my suggestion, listening to everything and integrating everything maybe here is room for a hybrid pathway. Hybrid would mean we start with two fetal brains and two adult brains studied as homogenate, effective immediately. And, at the same time, we keep sorting from additional adults brains, both NeuN + and – and then those for which we sort the neurons and wait until everything is sorted for the homogenate NeuN + and – and we process all together for all of these brains. But, we get going with the first part of the study for both the fetal and adult brains, studied only as homogenate and only as homogenate, and everyone generates data and nice reference brains until we have everything sorted.
    - **Nenad**: Jim has introduced a little bit of thinking about this. What are we going to do with this data is also another thing I think we need to discuss. The way is see this is two ways. One, we can say, we can do all of this on fetal and adult specimens or should we use this as some kind of common analysis that we can use as a nucleus for some of the joint, meta-analysis to have a really big Consortium paper? To me, that, in my opinion, is something we should discuss this afternoon and decide how fast we should process with this reference. We have specimen that only caveat is that we unfortunately can’t collect fibroblasts and unfortunately, the chance of collecting good, intact fetal brains at the moment is just basically minimal. So, should we rush with this or should we really make sure we do everything proper? And, I really do think that by next fall we should have all the samples and have all analyzed. To me, that is good timing to think about if we want to have a Consortium paper. And this could be something in addition to our individual studies and joint analysis that we could use in that study and really boost what we have created, or we should think about this should be a separate paper. In that case, I do think we should rush because the more you wait, new ideas are coming and everyone is working on the same thing. So, I am more for scenario number 2. Use this as a nucleus to have a really nice joint Consortium paper where we have some Consortium-generated data that will help in the interpretation of joint analysis and meta-analysis. I would then want to be a little more careful. Those are the two scenarios I see and determine how fast we should process all of this. Again, the brains are here and we even can sort them in two places, but I am against this because otherwise we are introducing bias. And, I think this is something that whatever we do in the future better. So, I think we should the do the best job and go with the plan that we had and use this as a joint paper and analysis.
    - **Dan**: Have you guys looked at the variability between your sites in terms of the sorting results?
    - **Nenad**: I have no idea. And, I would be surprised if there are none, and even if there are none, it would be hard to…
    - **Flora**: Maybe we could explore it with the tissue that you will send to do that. That would actually be something nice to do if you, Schahram, provide a uniform sorting and then with part of the pulverized tissue we all do some of our own sorting. Then you could compare uniform sorting vs. site specific sorting if that’s an important problem people want to resolve.
    - **Schahram**: I agree that would be more to speed up the whole process. That was the sorting capacity and so the idea is that some brains get sorted and only sorted at Yale and some only get sorted at Mount Sinai.
      * **Nenad**: I am not for it, just to let you know.
* **Schahram**: I wouldn’t see a major problem with that as long as not one brain is sorted at two different sites.
* **Jim**: I think you’re going down a wormhole there because then you need to test it on every method. It’s going to be as long as the sorting part is going to be.
* **Schahram**: I just wanted to summarize what Nenad said. So, even if we want have it done in an optimal, there’s no way we have to wait until everything is sorted or processed properly, then out of that we may have a good high quality reference paper. The other alternative is to do the fast-tracking like starting on this homogenate…
* **Nenad**: What I meant on the two scenarios is what do we do with this data? There are two goals, one is reference map and the other one is a cross-comparison. And, I think that the reference map would benefit a future joint analysis and, hopefully, a Consortium paper. We can start by saying these groups have their individual techniques and datasets from diseases and development and this and that. And, we wanted to standardize this to create reference map and we did this so that it’s a fresh new data. Hopefully, we’ll all have published something on our own by that point in time. For that, I don’t think you need to rush but we should be really careful to make sure everything that’s done is as standardized as possible. Scenario 1 was like we should really be published as fast as possible, irrelevant and uncoupled for many other plans for joint analysis. In that case, we really do need to hurry up. I am more for the second scenario because I think as a Consortium we have to have the best product as possible. We all have established labs, we are all publishing, but I think why do we even have a Consortium if we don’t do a good job? I thought that this could be used as a nucleus for joint analysis and comparison, and that is anyway, the goal too.
* **Jim**: I kind of like Schahram’s hybrid approach though. I don’t know if two is the right number for each of the tissues, but even one so we all make one set of our data type because that is what is going to facilitate the downstream analysis of putting them together and comparisons. And, all that analytic work won’t start until there is any data. So to even see that process a bit would be very valuable.
* **Nenad**: But, we can do that. The first brain we are almost done.
* **Schahram**: Everything is almost done in terms of sorting. One is in progress, one is on hold, but we are very few weeks away.
* **Nenad**: And, Sirisha can send you pulverized tissue from the same specimens for the same pulverized tissue sent to him to sort nuclei. We can do that tomorrow. I think we should just do it one by one. Is that a good plan? So, let’s finish one brain, sort the nuclei and then send the tissue and then we’ll move to another one. I would like to keep fetal as long as possible for two reasons, let me repeat: they are small and for example, we shipped Schahram tissue the New Year and ended up finished. We don’t know what happened; we thought we had enough dry ice. So, if something catastrophic happens to the adult specimens we still have more tissue from that brain, but for fetal it is really getting harder and harder.
* **Schahram**: There are a lot of logistical issues that sometimes can happen.
* **Nenad**: The hybrid approach the way I see now from what Jim said is to basically do one by one brain. And, I think we should start with adults that way if there are some screw-ups or something that we have enough tissue because we have left and right hemisphere and then we can accomplish something and really be sure it’s done properly. As soon as that sorting is done we will send you the same tissue from the same brains and specimens. If someone wants more tissue from the same brain, we have more tissue.
* **Geetha**: So, is the plan that the groups would sequence one by one, not in a batch?
* **Nenad**: That’s a good question because the ideal situation would be to sequence everything, but also many of you have not worked with the nuclei so maybe it’s good to do it at least one brain because if you make a mistake, you make a mistake on 40 samples. So, I do think that this is the right approach. Let’s do one brain and as soon as we are done then everyone gets the data. We’ll start processing the other one before that, but that’s why I call it hybrid. I am not sure it’s an ideal scenario because it’s not something that we can replicate easily. So let’s finish the brain we are now doing and everybody will get their nuclei and as soon as you want, we will send you tissue. We can send it tomorrow from the first brain and we will continue with the next brain, the nuclei, and as soon as you do the first brain, please report to us so that we know how good the data is and if you need more tissue and stuff like that. Then we all have two adults and if everybody is happy with the data and that the nuclei we send are correct, we can do fetal as well. Also, one thing, we do have scrap brains that are perfectly fine for fetal and adult. So, if you want to practice we can send you one to test, just to let you know.
* **Geetha**: So the plan is the one brain and the nuclei for groups to test and see how the data is and test their protocols and the next batch would be all brains at once or one-by-one?
* **Nenad**: No, one-by-one.
  + - * + **Peggy**: Can you send us your scrap tissue to us as soon as possible?
        + Nenad: Yes, we can absolutely do that. So, for scrap tissue, I will send you both fetal and adult. These are actually not scrap brains, they are perfect but I’ve decided not to reserve them. So, you can play with these brains as much as you want.
* **Nenad**: So, just to repeat, we will do this hybrid, one by one brain. The first brain will be sorted, sent you within 2 weeks. We will send you within two weeks pulverized, adult tissue from the same brain, and I will also send you a pilot fetal and adult from a different brain so this way you can work around. (**CONSENSUS**)
* **Geetha**: Does it take a month to sort one brain? Why are we talking about eight months then to complete this project?
* **Schahram**: I would say realistically it takes more than one month to sort a single brain. It should take less than two months to sort a single brain.
* **Geetha**: There are two issues*,* one is logistics and another one is the question of joint Consortium. How does this serve as the nucleus for the Consortium-wide analysis and my question goes to Pamela and Mark because you are the DAC and data coordination. In terms of data analysis, how is this going to impact the plans we have for the Consortium-wide analysis and conducting sequencing at different time points?
  + **Mark**: Maybe the best thing would be to have the discussion of the analysis in relation to this. We have the analysis discussion at the end. We did want really want to use this as a reference center point of the analysis which Nenad is eluding to that it could be the centerpiece of a major paper and so forth.
  + **Pamela**: There are two different things between it taking a while for Schahram to do all the sorting and send the cells to each of the individual labs. That doesn’t necessarily apply to each individual lab that they do their experiment only as they get them. So, whether you store something in the freezer one month vs. five, there will be differences, but I am less concerned about that than the batch effects that come from making library preps one month, next month, those are the bigger batch effects usually.
  + **Geetha**: This is what I am concerned about more.
  + **Daniel**: To deal with that, is to either have some randomization protocol across things we want to compare or to take a small subset of the samples and just keep letting them with every batch that we can kind of use as an anchor for harmonization. Both of those are plausible at 5% of the samples and the other might not be feasible, the randomization part. There’s an experimental issue that it takes several months for each brain between fetal and adult. You don’t want to put these in the freezer for two years or 10 months. It’s better to get the data out. If Schahram took some and just ran this after sorting the first brain we could use this as an anchor. So, you run that with the fetal and run that with the other adult and use that as your normalization.
  + **Pamela**: I would still try to wrap it in a smaller amount of batches. I think that the randomization will be very difficult. We’ve made ourselves randomize mostly RNA – Seq and ChIP – Seq and its tortuous and very time consuming.
  + **Geetha**: So, the batch effects is a concern if you sequence at different time points with different library preparations one at a time over 8 or 9 months is actually concerning. And with the randomization being tortuous, what is the compromise here?
  + **Pamela**: It’s not like we are doing a huge number of samples in anything. It’s really a very small number, so there will always going to be variability. It is about partly assessing what goes on. I do like Dan’s idea of having one set of tissues that are done that first time and a little bit of that is done every time. We’ll do our best.
  + **Zhiping**: What about the reference sample you are going to ship now? Why not use that?
  + Dan: Yes, that is what I was saying.
  + **Nenad**: The problem with this is for Schahram to sort again the same brains 7 or 8 times would be very difficult.
  + **Daniel**: No, just save or freeze it.
  + **Nenad**: What I am saying is FAC sorting is limited to how many nuclei you can push through the sorter. That is the limiting factor because you cannot push it too fast and they are too fragile. You can’t run it 24 hours a day. So, if we are now deciding we should repeat the reference brain each time…
  + **Dan**: Yeah, just do the whole tissue. The idea is to do one nucleus and one whole tissue. Make it simpler, but the idea to account for batch effects, flow cell variability.
  + **Nenad**: I actually do like this idea then. But, how about we take not PFC but something that as long as we have a large amount of tissue. I think we should stick with the cortex just because we are all cortical-centric people here. We have additional sample that is only pulverized and it’s like a dorso-lateral posterior parietal cortex because I don’t know anybody who is studying that, and I can try and dissect a huge piece and this way you can run this and we have this hybrid approach. You get nuclei, tissue, and this super-duper platinum reference sample. Would that be okay? So then we adopt a one-by-one hybrid approach, first adult, you sequence them as you get them and in addition you are getting one super platinum reference sample from DLPPC. Is that okay? Does anyone have any concerns?
    - **Stella**: Do you sequence it every time for each sample?
    - **Dan**: So, I think that’s an issue. With techniques like Hi-C with an entire flow cell for one sample, that’s not something we can do lightly. We’ll figure something out, but I think we should get some input from the statisticians and people who do analysis to ensure that we do something that is useful for it, but the key is to have a reference to compare across. It doesn’t mean that you necessarily have to run if you’re doing RNA-Seq 50 million map-able reads as long as you have something you can use as an anchor.
    - **Pamela**: Another thing is that part of this exercise; there are things that need to be calibrated across labs more than other things. Like, Hi-C may not need to be calibrated. It’s a much newer and difficult technique, but the RNA-Seq and ChIP-Seq, we wanted to have because remember we wanted to be able to leverage into ENCODE and Roadmap. And, so, just being able to do it in such a way that it allows us to leverage into that…
    - **Sherman**: In the early days of ENCODE, you took the biggest differences between different laboratories doing the same analysis. If you took the data from several groups and did a principle component, the first component was always the laboratory, so I think comparing the same assay and same material in two different laboratories is an important point.
    - **Flora**: And, I think that following on that, what do we do if we do all of this digging and we do find such a difference...In resolving those differences, with this binary comparison first, so we actually move ahead and we do publish a paper that makes sense, right? Otherwise, we publish a paper and we know there is a big difference and there’s nothing constructive in it.
    - **Mark**: Later on in the discussion we have the coordinating with GTEx and ENCODE as well and I was wondering if Dan’s suggestion of using a reference over and over again as a calibration, if maybe we could just use a cell line. Maybe it’s not just sample, but it could be a cell line like the SK-NH1 or something that everyone can use that is sort of easy to get your hands on that we could standardize everything according to that. I don’t know if that’s easier for you guys.
    - **Andrew**: What about ERCC spiking and other sorts of synthetic spike-ins that we could insert into all of the different libraries?
    - **Peggy**: Whatever we do decide a flowchart explaining what we are all supposed to do would be essential.
    - **Schahram**: The last thing we want to decide is do we want to have a platinum reference brain every time we assay or a reference cell line that should be assayed every time? We have to do the sorting and doing both would be very costly.
    - **Nenad**: I am not for cell line. Just because we are a brain consortium…
    - **Pamela**: But, it’s a perfectly reasonable thing to use as a positive control calibration resource. You are trying to use something that is high quality and there’s nothing wrong with that. I would just stress that we should come to a decision here even if it takes a little bit longer because this is a critical current problem. The other things that we are going to discuss are much easier to continue over time.
    - **Geetha**: I agree. Over the course of these 3 or 4 hours, we can be flexible on how we distribute our time for different topics.
    - **Greg**: One potential problem in doing a cell line is that in order to do it right you have to grow up one big batch and then freeze it and then put it in aliquots, which is fine and can be done. But the same thing was going to be done with this extra calibration tissue where it’s basically pulverized the exact same way and aliquoted and so in some ways it’s going to be a lot easier.
    - **Nenad**: I completely agree and second everything you said. Again, we can grow that and I am happy to contribute.
    - **Peggy**: I think the RNA samples will be different from the genomic samples. We’re not going to do Nome-Seq on something more than once. It’s just too expensive and the same for WGBS, the same for all of us. So, I think the RNA guys may have a different flow chart than the chromatin, 3-dimensional, and DNA methylation people.
    - **Andrew**: With the spike-in that would theoretically be biology independent, so whatever technology that you are making samples should contain things like a beacon. So, it depends on how you are partition the variability and look at by looking across labs.
    - **Geetha**: Can we do eight brain samples or after finishing the first one can we do the next seven, pulverized tissue first and then do sorted nuclei all at once 4 or 5 months from now? Schahram can consolidate the nuclei using the same calibration tissue, does that make sense?
    - **Nenad**: You are saying to finish the first brain, which sorting is close to being done, then send the pulverized tissue for that brain and then for all other 7 pulverized tissue, and then do sequencing as a calibration? That is fine with me. What Schahram and I will do is send you a flowchart next week. **(ACTION)**
      * **Summary**: The plan is to finish sorting and pulverized tissue for the first brain, and then send other 7 and calibration samples. Is everybody for that?
    - **Stella**: If you do the 7 samples continuously, why do you need calibration samples if you do everything together? We don’t need the calibration sample if we do the 7 in one batch.
    - **Geetha**: This is to compare the first brain and also nuclei that you all will be doing. This will be the anchor to compare. The people who are doing more expensive analysis don’t have to everything at once.
    - **Schahram**: I think we accomplished something and we can move on to the next item.
    - **Geetha**: Can you please summarize, Nenad and Schahram?
    - **Schahram**: The plan is to finish the sorting of the first brain and ship it out. And that comes from my lab shipped out to all the individual labs. We will send out homogenate tissue from this sorted brain and homogenate tissue from the 7 other brains to everyone so everyone has the homogenate that the need, plus the reference brain sample that is important for being the anchor to be repeatedly assayed with every sequencing and library preparation you will do because the bottom line is that for the remaining brains for which we have to sort nuclei we will do every month up to one brain at best and send it out and in a couple of months everybody should have a completed set of sorted nuclei and then they’ll run the assays again on the reference brain. And the flowchart we will send out next week. **(ACTION)**
    - **Mette**: We will put it up in Synapse and we will also add this with the data of things that have been shipped and what remains.
    - **Geetha**: For the flow charts, are we going to do for each assay?
    - **Nenad**: One simple flowchart.
    - **Mette**: How much of the reference tissue will be shipped to each group?
    - **Nenad**: Schahram and I will decide. I just don’t want to guesstimate now. We will probably send another e-mail.
    - **Schahram**: And, this is where we need help from you.
    - **Nenad**: For fetal I will send what I think is appropriate because it’s impossible to calibrate because it’s a completely different measurement compared to the adult.
    - **Geetha**: So, should we say by the next time we meet in October, I propose that we meet every six months that we can get through this by our next meeting.
    - **Nenad**: We will do our best.

1. Milestone updates from each group
2. Data coordination with GTEx and ENCODE

**1:20 – 2:30:** **Joint Consortium-wide analyses** – **Moderators**: *Pamela Sklar, Mark Gerstein* an*d Nenad Sestan*

1. Bionimbus data transfer – *Kevin White*

* **Kevin**: I don’t have an overview slide of the Bionimbus platform, so I will just describe it in words. So, the concept here that as datasets are being uploaded into Synapse, coordinating with Mette and her team on creating a smooth workflow to move the data into the Bionimbus protected data cloud system, which is currently housed on a whole bunch of servers in a data center. We, among other things, are handling all of TCGA, all of NCI’s data, and we’ve also had an instance of this set-up for PsychENCODE and we’ve done ENCODE, and so forth in the past. We also have access to some supercomputing resources. The whole idea here was we would use the Bionimbus platform to sort of at least start copy of the raw reads data as well as do a uniform processing with the help of Mark and Zhiping and all of you.
  + We’re really starting out with uniform quantification of transcripts so that all the datasets are normalized and harmonized using the same pipelines and parameter settings. The purpose behind this is to provide raw input for studies by all of us. We are all going to do our own project specific analysis but then there’s also the power of aggregating the data. These aggregating datasets can also be used as a resource for the community looking at splices and other sort of discoveries, especially because I think we all ended up with a Ribo-Zero protocol and potentially other types of transcripts. And then we hope what we’ll do is create a standardized product that can be shared with the broader community through the Synapse platform which we moved the results back onto or through Bionimbus.
  + We’re starting with the RNA-Seq that Mette put together. It looks like we’ll need to start thinking sooner rather than later, putting in the ChIP-Seq pipeline as well. The good news there is that we’ve been running the ChIP-Seq pipelines for quite some time due to the ENCODE project, but this is just a less granular view of the various datasets. We’ve basically pulled in all the datasets that currently are available or were available in the recent past from the various groups. So, we haven’t started the actual running of pipelines yet.
  + Mark was a little afraid of what he calls the pipeline overload. So, he presented the ENCODE pipeline. One of the issues with the ENCODE data in respect to our data, is that our collective data here is that ENCODE primarily focused on Poly-A RNA. And when you start getting into the nitty gritty of running these pipelines, the settings, parameters, and even the algorithms used by the ENCODE project, we may not as a Consortium end up want to use. So, we have three pipelines we’ve got in this system as this point. One we call the Chicago flavor. This is also being used in part on the Epigenome Atlas data. The next pipeline is the one that you just saw from Mark’s slide. This uses STARR and I think Cufflinks. The next one we call the Trendy flavor one. This one uses Sailfish and kallisto, which are relatively newcomers on the scene and a lot of folks are using this. So my real point in putting these up here is that these are the three different possibilities. We can run them all. If members of the Consortium have their favorite pipelines they run, we can run those as well. We create virtual images, we can doctor-ize pipelines and put them into the system or better yet, you doctor-ize them and we run them. And then within some kind of reason can compute, computing takes dollars because it uses electricity and equipment and cooling, but within some reasonable parameters we should be able to run multiple pipelines of this data and create datasets that we can share, use, take, and figure out ways to normalize and compare across datasets. That’s where we stand right now. We’re ready to start doing the processing.
  + The last slide is again on analysis. In terms of ENCODE, are there pipelines we should be starting up under PsychENCODE? Have people found that with their datasets that are parameterization or particular algorithms they would prefer to run? We should probably start having those discussions. At least two of our groups are doing ATAC-Seq and there’s a lot of activity around how to analyze the data right now. So, at some point, we may want to be doing that in a unified way as well.
  + **Mark**: Do we want to have some discussion about the unified processing? And also, we could talk about the idea that we’d put all this data or results onto Synapse and stuff like that.
    - **Kevin**: I didn’t show a slide, but people on Mette’s team and my team and Mark and Zhiping’s teams have actually put quite a bit of work already into figuring out the workflow on getting data into the Bionimbus system and, once things are determined, put back onto the Synapse system so that they are actually in some sort of format we are all comfortable with using and there’s a central way to have to rest of the Consortium look at the data.
    - **Pamela**: We should not launch into a long discussion exactly what the pipeline issues are until we then talk about what our mission critical mode of analysis is and then fit pipelines and priorities around that.
    - **Mark**: I agree with Pamela. We have break and then we have the joint analysis and we have some presentations set-up. We should probably go over all of that first…
    - **Peggy**: Do we want to talk about a DNA methylation pipeline? We have two different groups doing DNA methylation data and we should probably think about doing some of this the same way, right?
    - **Pamela**: I think that as we talk about how we are approaching different things that would prioritize which pipelines get written first or which pipelines get written by certain sub-groups.
    - **Dan**: I think this is great. Some of these methods, even though we’re doing RNA-Seq all the time, we’ve never even tried the newest fashion. I guess my main sense is that when you compare HCG exon gene union or cufflinks you notice different methods and generally you get a number of genes that are different at it and there’s a high correlation between the results. I am not worried that that is going to screw things up. What I am more worried about are what variables we stick in the model whether you are going to use something like tiers, and I think Andrew could talk more about that.
    - **Pamela**: I certainly agree with that. I think we’ve spent more time way more time on the model. We’ve tried various ones out to find that no package is perfect.
    - **Dan**: So, I think that is going to be at some level a couple of gold standards that kind of bracket the different approaches rather than going wild because I think that the bigger problems are going to be the biological covariates and the technical covariates.
    - **Pamela**: And, I will say later that what is most special about us is the ChIP-Seq s tuff not the RNA-seq. I am afraid we will launch into an enormous maelstrom of RNA-Seq analysis that will somehow delay us getting where we really need to be. Not that I do not think that it’s important.
    - **Mark**: I agree with Pamela very strongly in the sense that we don’t want to reinvent the wheel, we don’t want to repeat the endless discussion of with GTEx, ENCODE on these subjects.
    - **Zhiping**: It’s kind of important no matter which pipeline we stage. It matters that have we same data in the same pipeline. Otherwise, we will see a lot of architectural defects.
    - **Pamela**: In my presentation, there are some other datasets that we want to decide when and that we have all the appropriate datasets that we might want processed at the point in which we send it through.
    - **Mette**: Perhaps we can have a discussion on how to incorporate data as it is coming into the pipeline so that we can go beyond the heavily controlled data that we currently have.

1. Update on data submissions and downloads – *Mette Peters*

* **Mette**: As part of the discussion for the data analysis and data coordination core activities I created a spreadsheet listing all the data that has been deposited to date so that can frame the discussion of what data analysis can be done at this point based on what data is available. In your package there should be a document that looks like this, 2-pages, that has been divided into assays on data that is available, RNA-Seq, ChIP-Seq, ATAC-Seq and ERRBS. This was from the pilot, animal data. So this is what is there now.
* **Pamela**: I have been annotating all morning with what additional things are there. So, I will show this in a different way later.

1. Analytic platforms for RNA-Seq and ChIP-Seq – *Mark Gerstein and Zhiping Weng*

* **Mark**: Zhiping and myself made a quick refresher slide pack of three slides for the RNA and ChIP-Seq on where we all agreed. We will go through those really quick and then I would suggest Kevin’s slide pack and then we can talk about 2 and 3 together. This is just to refresh people’s memory. The goal is to have uniform processing of all the data and to do that we needed a standard pipeline for RNA-Seq, which is one of the major data types and also ChIP-Seq which is assays.

**2:30 – 2:45**: **Break**

**2:45 – 4:45**:

1. Joint analyses and analytic foci – **Moderators**: *Mark Gerstein, Pamela Sklar, Nenad Sestan, Daniel Geschwind, Flora Vaccarino*

* **Pamela**: The question now is how to operationalize all of this, and I suggested something 6 – 8 or 9 months ago, and so I’ve framed what I’m talking about today using some of the same terms. But, after all the presentations, I think I would not necessarily use long-term vs. short-term as descriptors. What I think about now as we are starting to collect signification amounts of data but in specific areas is what would our capstone kind of paper be? Even the short-term goals that would lead to papers I had outlined 6 or 9 months ago are a huge piece of work. None of them were simple. So, I kind of flipped around where instead of focusing on largely on the RNA-Seq based analyses, which in fact have been done in a variety of formats since the early 2000s in the brain, what is unique is to the project, not that the RNA-Seq shouldn’t be done and shouldn’t go along, but what is really unique are our characterization of the epigenetic landscape using that in its integrated form with RNA sequencing and whatever other data is appropriately ready and then connecting that landscape with diseases and doing the same thing across development. Now, I think that’s actually more than one paper, but the question is heading towards this as a capstone a reasonable thing to do? What also occurred to me is that the kinds of smaller individual studies that each lab wants to do on their own data would fit very well nicely over the course of the next year to fifteen months while we are finishing and heading towards this larger paper and all of those preliminary analyses would be part of PsychENCODE. Not that they would share full authorship, but they would be things that have emerged from the U01s from this grant. So, would in anyone’s slide in the future be attributed of PsychENCODE projects? It has a number of advantages. Since this wasn’t a consortium like ENCODE that was established together, we have a variety of data types, some of which talk to each better. It allows the individual groups to simply proceed and learn about their data, which is going to be a value to the pipeline to the internal group and generate ideas and analyses. And, if we were to agree to do something very simple as one of our authors, it could always say “PsychENCODE Consortium”, not carrying any people with that, but then it would always be clear. Then we would have a way to move forward.
  + What I did while we were talking, is I put together where things are now and what we might think about. I posed a lot questions we could think about. I put together where I think ChIP-Seq and where I think it will be in six months and it’s not necessarily in the most straightforward way to interpret it, so I’ll just describe it. I put in acetylation and me3 marks across a couple of regions in SZ, CNON, and ASD patients and some across development and maybe 50 in SZ and 50 in controls that by the next data freeze in August will be what is that. That is a non-trivial amount of data that we can think about.
  + Then there are a bunch of other data types we will want to think about and prioritize where they go. So, I think we could think for a capstone by the end of 2017. Not everyone’s data would make it in necessarily in this publication necessarily, but we could start pulling off aspects of analyses that individuals would be responsible for. They’ll like publish their own individual papers, but then things would get applied to the full integrated dataset, which doesn’t exist yet.
  + The short term goals were largely around three things. One was RNA sequencing, eQTLs, differential things that are already set to go, in some ways, into the pipeline. And a sub-flavor of those is in the differences between diseases both in the basic analyses and the more high tech co-expression analyses. These may or may not be useful without the epigenetic marks because they’ve all be done and published. And so, the other area is that Mark is particularly interested in and also forms what we could envision one of these sub-analyses which are comparing the brain to other organs with other data that are out there. Obviously, at each point here is a parallel brain analysis that could go with RNA-Seq data and whether that can survive independent of pulling in the data from the ChIP-Seq marks and other epigenomic is not clear to me. All of these things can be done by our individual labs and we’ll generate solid important data, but maybe not the highest profile, highest impact. I am thinking there are several kinds of papers that we want to have.
  + From what people have said today and what might be done for analysis over the summer would be another 400 or so controls and then around 500 SZ cases, and 300 BP cases and 60 AT patients. I don’t yet have a strategy for iPSCs and individual cell work and the developmental data. I pulled in GTEx because we’d want to use all high quality RNA-Seq data available from the brain.
  + Before we had discussion today, I was thinking the individual groups would take part in the shorter term goals or maybe there wouldn’t be enough RNA data to make it worthwhile, but it does look like there is a substantial amount of RNA data. So, over the course of the day, I’ve been thinking a new parallel and important pathway of big overall projects that we can contribute to, and then there’s the endless number of the individual projects that have been outlined all day long where two groups doing AT and SZ would want to analyze at the co-expression levels together or Jim and Andrew will come together because they want to make a methylation map and a pipeline around that. I just think we need to articulate where those sub-areas area. I also think that it would be nice that if someone has a bigger dataset and doing a larger analysis and someone who has a littler dataset, please accept them into the group because we don’t have to always be angling. There are unbalanced datasets where one or the other is more of a prime interest. What I would open it up to is whether this resonates with other people and if it does we can spend some time operationalizing that before we leave today and then move that forward.
    - **Zhiping**: Maybe a short-term goal would be an RNA-Seq centric paper and then in parallel a ChIP-Seq paper?
    - **Pamela**: I think the RNA-Seq paper will be basically just as hard. I think ultimately, they are going to run parallel. That’s my guess realistically.
    - **Thomas**: Can we think of a strategy where we could try to do something similar to what ENCODE is doing. They had a smaller phase and then a larger phase and they got their papers in at the same time to Nature Neuroscience.
    - **Pamela**: I thought I was outlining something quite similar to that. And that would be, for example, if you can imagine in the next year, Jim Knowles will have RNA-Seq and ChIP-Seq in some interesting aspect of his CNON cells. Flora will have something that characterizes her organoids. All of these will carry the name of PsychENCODE. And this will in some way be competitive because she wants to be out there having been published in these organoids on ChIP-Seq before we all get around to getting the big Consortium paper out. We talked about because we’ve done the epi-map whether we are able to refine even though we have small sample of ChIP-Seq data right now, will that help is refining our understanding of SZ. For example, I was talking to Dan today about how we should be looking at the co-expression networks and a couple of things in AT and SZ. That would be a series of papers that would emerge in that same way pilot phase. I think because this project isn’t organized in the same top down way ENCODE started, I don’t think there’s a way to just call it “pilot phase” because we are doing the same thing, but bigger. I thought we’ve all outlined mission/disease critical questions in each of our grants and those would form the equivalent of the pilot phase.
    - **Thomas**: I was wondering if there was a strategy to approach one of these journals and ask if they’d be interested in doing a focus edition. Nature Neuroscience or one of those journals could a focus edition on PsychENCODE. Or is it just too difficult to coordinate the various projects to converge at approximately the same time?
    - **Pamela**: Since I didn’t see anything today that looks exactly like people are at the end-game for all variety of papers, I think it’s hard to answer that question right now.
    - **Mark**: Coordinating on many papers is also very complex, and I would argue if our goal was to cover Nature or Science, I think that the idea of having one really good paper that would make it onto the cover of Science or Nature, I don’t think that’s going to happen without a tremendous amount of other work where people do the types of analyses they want for their projects where people analyze their specific datasets and then coalesce. I think it’s still a worthwhile aim. But, what I was trying to get at with ENCODE is that those “big papers” didn’t happen first. They’re the roof on the house as opposed to the floor under your feet.
    - **Pamela**: The response we got from Magdalena when we submitted the marker paper, I think as a response we are going to be fighting up against that all time no matter what we do. Why are you really different and better and what are we going to learn about this project above and beyond ENCODE and Roadmap? We have to do something that’s really meaningful to get to that. I think we can and that we’ll have a dataset that we’ll be able to, but the way you could sort of sell something as that’s never been done before is harder with this.
    - **Greg**: I remember the ENCODE capstone papers. The original goal was to coalesce all the datasets and most of the paper was stuff that couldn’t have been figured out by looking at one dataset by themselves, and I think that’s something we could think about here. Everyone kind of goes about and does their own analyses and then the capstone paper would be a meta-analysis of everything that couldn’t have been done with individual datasets. It seemed like that is how it worked.
    - **Mark**: This may be a little too prescript, but you can imagine the scenario where we everyone try to finish these smaller analyses or the types of individual group analyses and after that take all that sort of published stuff and coalesce it into something bigger. I don’t know if that makes sense.
    - **Kevin**: These sort of capstone ENCODE papers, their main utility and value is having organized the data from the Consortium in to the public domain so that other people can do interesting analyses with it. We should take a cue from them as well. It’s not like we won’t be doing our own interesting analyses. We will. But, part of the big consortia paper is just simply saying, here is all the data, it’s all organized, here’s our meta-analysis and little add-ons, but most people don’t care able the analyses in those little ENCODE papers anyway.
    - **Mark**: If you look at the citations of the main consortia paper, it’s that they want people to use the data. They track people using the data and they want people to use the data. Part of organizing the data is just putting it into a unified corpus so it’s accessible. And that’s a tremendous amount of work. Dumping data is not trivial. It’s a tremendous amount of analytic work just to homogenize it.
    - **Pamela**: I would say that we have one huge advantage over ENCODE though because we have a disease framework. There are going to be chances to make some fundamental observations if we get lucky. We may or may not have enough power, but we have some chances to do stuff in the disease framework.
    - **Kevin**: I think particularly, the set of diseases we are working with here, there’s a common biology. This is a hypothesis we are testing here. I think the capstone paper could be a lot more interesting, but I am pointing to the utility of ENCODE is where the baseline of this project should be and then anything we add to that makes this Consortium look pretty awesome.
    - **Pamela**: It looks to me like we’ll have more RNA Seq data than GTEx for the brain.
    - **Flora**: So, there are two basic ideas. For common analysis, what is really the meta-analysis taking advantage of the sheer data, which we will generate to do something with high impact? Another one is a small focus project that utilizes the reciprocal discoveries that we will make, so why don’t we integrate RNA and Chip-Seq across two stages of development or ATAC-Seq or maybe, in our case, we were thinking to use in vivo models to test or validate data in the brain. That could be something we are working on now. Take advantage of what each of us has to offer to do something, but and that can be done with whatever schedule we want to follow. Maybe that could be going in parallel perhaps?
    - **Pamela**: The individual differences are one of the long-term goals. I listed a map of the histone marks, map of methylation, chromatin accessibility, region cell type specific, eQTLs, and an integrative map. And, then those same things in a disease context. These are not the only things that could happen. We could spend weeks saying we’re actually going to do everything, the question is question would those things be novel enough to be high profile?
    - **Dan**: Our value added is our biological and with us having to do with being neurobiologists, coupled with a lot of other genomicists and moving this into really to show how the brain data informs our understanding of disease. I’d look at this in a really practical way if we could cut out using the charts you’ve put out what kind of data we will have and when, one could easily come up with a series of 3-4 papers that are linked to each other using the same datasets, some that have to do with showing the kind of resource aspect of it, combined. But some of those have moved down the corridor into various diseases both across and into individual ones. I just see it as a practical issue.
    - **Pamela**: That is what I tried to start doing. I think RNA-Seq is one and the two ChIP-Seq marks. I think they’ll be a lot of data from those, not all of it necessarily. And, I don’t think we need to wait for all the data for every cell type or everything that may be practical.
    - **Dan**: If you could imagine one of all of them together and one that shows it’s utility by the other data that is around, non-neural data, non-developmental data and show how it might improve, giving examples of how to improve annotations of psychiatric genetic elements.
    - Pamela: Have I missed something in my going through today that we think is going to be done in that same time frame or that would be extremely helpful to put together with this? We don’t need to necessarily answer that today, but the framework I’ve outlined, let us think about all of the people who are doing the analysis who are going after those datasets, finding them, getting them, and getting them in one place and processed.
    - **Nenad**: I do agree. I am not aware of comprehensive analysis of known brain tissue during development. And, I have collected several specimens of other organs that I don’t plans to use it. And, I think that is one interesting resource. So, maybe those are things to help. Maybe we don’t need to do a lot of analysis.
    - **Dan**: One of the things could be just to imagine a series of 3 – 4 papers that are linked but involve different types of expertise and analysis. If you could think of one network type paper carefully trying to integrate ChIP-Seq data with expression and methylation, chromatin structure and all of that and then make working groups that are working on setting up plans. Then what happens if you have 3-4 papers going out, it helps push things along and gets people together, and I think it also gives people their kind of mission and the post-docs know they’ll have their thing. You all with ENCODE must have huge experience with this.
    - **Pamela**: Although, more and more for the network analyses, you want the results of both ChIP-Seq and the RNA-Seq at this point. So, getting those really well analyzed so those different types, because that’s what needs to get fed in at this point more than what’s been done in the last three years.
    - **Kevin**: Based on work in my own laboratory where we don’t study the diseases that this Consortium studies, we study other diseases like cancer. I think one of the things we could do as a Consortium is sooner than later not have some paper ready, but some kind of a global analysis that help to nominate candidate genes, candidate regions and so forth, for biological exploration in validation by the laboratories in this Consortium that do actually do that kind of work. If we can do it early enough, it may be possible that some of these papers that come out involve the kind of level of validation we saw in the Broad C-4.
    - **Pamela**: I would imagine that would be what happens in the individual groups this year. We are certainly intending to do that.
    - **Kevin**: I’m making the point that there may be candidates that emerge from a sort of joint integration.
    - **Pamela**: I understand and that’s the third bullet point on the slide.
    - **Dan**: If we got the top 5 things that are really solid right now, we could start modeling them…
    - **Pamela**: Obviously, I can’t wait to look up the top 5 things for SZ in the datasets that…
    - **Kevin**: I’m just saying that is something that maybe the DAC could have as a high priority that Zhiping, Pam and other people who have biological models that want to be following up on this.
    - **Pamela**: Psychiatric diseases are just really different from all the other diseases. The idea of refining a signal has proven you gotten lucky because there is a structure abnormality there and there is the one super strong signal where you can sub-divide it into additional analyses and still have power to find something. In SZ and BPD and in AT it’s probably a little different because it has a slightly different architecture but this has proven to be really, really hard. The linking between eQTLs and signals, there will be many chip peaks and there will be many eQTLs that nominate many genes. The question is at what point do you want to send people down the goose chase for candidate genes? I think in psychiatric diseases, there is such a terrible history of premature closure on candidate genes that have little to nothing to do that, as a Consortium, I think we want to be careful and rigorous about that.
    - **Kevin**: So, all the DAC can do is say, here’s what I’ve found, and no one wants to go on goose chases. I realize it is extremely tough and you need to modify what you accept as validation because anything you can hold on to and doesn’t slip away is a big success. It’s not like cancer where we can transform a cell and then make a tumor in a mouse and come up with a therapy to cure the tumor and get it into people. It’s a completely different ballgame, but even if we can identify some enhancers that are expressed in the right kind of places and then validate those and the right kind of models, maybe? I don’t know what the right assays are. Those are for the wet labs in the Consortium to decide. I was just expressing a hope and dream because I think this is the core principle upon which the Consortium was founded on. It was to try to advance the ball forward in terms of understanding the brain and how a change in population variation relates to differences among individuals in regard to these diseases. Anything that we could do to feed that is all I am trying to argue for.
    - **Nenad**: I completely agree with you. And, I just have another question, who do we want to appeal to as a Consortium? Do we want to appeal to ENCODE people, and the answer is “no”. That is not my goal and I don’t think that should be our goal. Our goal is to appeal to neurobiologists. If we are appealing to neuroscientists, they think completely different. We really need to think about this, and I completely agree with Kevin. I understand it’s not easy because we will be judged by them and criticized. At least we can learn about what those elements do in development.
    - **Mark**: If the goal is to appeal to the neuroscience community, we want to make the resource useful to them. The current website is definitely in the genomics mindset, and I think that is important to have. My sense is the neurobiology community will want a different interface. I am not exactly sure what that interface is and I am not so sure how easy it is to build that, but something where like a gene highlighter where they can…
    - **Zhiping**: Maybe like the TCGA portal which is very user friendly. Having the processing power would really add to our appeal.
    - **Mette**: We’d really need the data before we could implement that, but maybe we could consider a wall of targets or whatever we want to call it where people can provide this sort of information of genes if you want to go in that direction.
    - **Pamela**: Certainly, a map early of 108 loci would be useful.
    - **Jim**: I would make it less about whom we want to appeal to and make it more about how can we make progress on these diseases? And, I think the best thing we have is the GWAS and that is only going to get better. So, how do we integrate most quickly if it’s the 108 or whatever into that framework and start to describe it? I think the common thinking at the moment is that it’s not protein coding, it’s regulatory and it’s common variation and those variations probably lie in regulatory elements, so let’s find the relative regulatory elements and find which genes are going up and which genes are going down based as a result of this genetic risk and then move from genome space into transcript space and then figure out what are the key regulatory elements. And, what is the right model for that? Can they go on our cell lines or do you have to really organize them to see the effect, or is there some other assay we don’t even have in this project because we don’t even know what these changes are doing and when in development. So, I would keep more on that discovery path more.
    - **Sherman**: What we are doing is making bricks, not building a house.
    - **Thomas**: Are we discussing strategies for cross-Consortium analyses or which papers to push first?
    - **Pamela**: The summary is that we think the data is most far along in RNA-Seq, and then one promoter and one enhancer mark and that those will be moved more aggressively over the course of the Spring and hopefully people will start depositing the rest of the data that will enhance that for a full turn of a crank over the summer that would lead to findings. That in the other set of things will be the individual projects which I can’t list for you right now, but perhaps the thing to do is then canvas each group to have just a rough title or bullet point or what are the analyses people are pursuing and generate a list for you and for us so that we can make some cross-Consortium groupings that are sort of sub-groupings. The first two I outlined, everyone participates in because everyone is generating some form of those datasets. That’s my idea.
    - **Geetha**: For the map of the genetic landscape, is the DAC who is going to generate that map or if all groups are contributing data, who would be generating the map?
    - **Pamela**: I think there are two parts to it. There’s the part where you have to assemble all the data and get it together so it can be analyzed. Then there’s an analytic group, which most of the DAC members will be part of, to figure out what you are actually going to do with that data.
    - **Jim**: What if we broke into disease specific groups? Of course everyone has their controls so that would become the common map. The question is how much is the disease map like the common map? But, we’d all be contributing to the common map with our controls, and then break into disease-specific groups to try and answer the question of what is causing these diseases?
    - **Pamela**: I think that’s going to happen naturally.
    - **Geetha**: There are the long-term goals and short-term goals that Pamela and Mark brought up in September. Since we are meeting in person, this might be the best venue to discuss which sub-groups could be formed and who would be involved in which kind of analysis. Which sub-groups would be interested so that we can establish a framework right here.
    - **Pamela**: My guess is that all groups are going to be interested in both of those capstone projects, but there’s at least 6 months of DAC work unless you to me differently.
    - **Mark**: I agree with what Pamela said. The way toward the capstone is to get through a lot of the individual group papers. I sense that would be a very good thing because each of those analyses people are looking at in great detail in the datasets they are working with. The thought process is to leverage that experience productively towards building the roof on the house.
    - **Pamela**: We could go around and state what we think our individual projects…I’m sure you are all like me and are thinking about what the papers are going to be, but it’s about what is the next step.
      * **Jim**: We have a draft manuscript that’s almost ready to go on the differential gene expression, then we have all of the microRNA data, and we’ll decide where that is going to be just from the first half than the whole thing. Then there will probably be a second long RNA thing. Then there’s going to be correlations to genome structure. Then we move into the ones that were funded by this mechanism like the 27AC and how it correlates with gene expression and structural variation for outliers and then we also have the NOMe-Seq data. I think we’ll have the whole ChIP-Seq dataset by early Fall. The NOMe-Seq data is going to be a little further along. It’s the intent to try and get the data generated by early year 3 because we want to write it up and try for a renewal.
        + **Peggy**: I would think that some of the things we want to do is to specifically focus in on the regulatory elements using k27, nucleosome depleted regions, and CTCF and ask how do these vary across controls, from schizophrenics and how do they vary from person-to-person using the same tissue? We’ll probably focus a lot on the regulatory variation elements. That would be the first thing probably, and then try to lay down the chromosomal domains and see how that fits in with some of these GWAS areas. I think that will be a little longer, but that’s something else we want to do.
        + **Pamela**: If you were to say which one would lead to a paper, what would you be able to draft by summer?
        + **Peggy**: I think we’re going to have quite a bit of the H3K27 data and probably quite a few of the NOMe-Seq datasets. I think that trying to look at an analysis of what is the variation and position of these regulatory elements and if are there nucleotide elements that are causing them to go across the population, I think that could be the first thing that we do.
* **Dalila**: We have been working on the paper for that describes the methodology being developed for Iso-Seq cap and it is almost ready for submission. And, this is a little bit more than just profiling the brain. The other project not funded under the R21 but somewhat related is that we have been analyzing the CMC SZ data, which is about 600 cases and controls, and we have basically been leveraging the Illumina Seq data for differential expression lncRNA paper which I plan to submit in the next 2-3 months. We have not moved the AT dataset because I would prefer to establish some paper on that since it’s only 40 samples per brain region, so I think the bottleneck there is that we need to try to get more samples, and right now in my lab we only have 10 samples, 10 cases and controls for each one of the brain regions.
* **Flora**: It would be great if we could just do what I said in the beginning to compare iPSC derived organoid models and see if it does model outside the fetal brain development. To what extent we could go beyond the early fetal stage by all the things we are doing in the lab, like transplanting the organoids and getting to go to a later stage and seeing to what stage we can finally get to model brain development. And the differences are going to be interesting as the similarities because we want to know what we are missing. What kind of gene regulatory elements and transcripts, it could be non-coding RNA and all of that in our proposal and ChIP-Seq data cauterized these organoids vs. normal brain. We are hoping in the second stage to use all of the data that are generated in this project to model variants that people have discovered or will discover, but without this preliminary validation, so to speak, it is going to be very hard to do.
* **Pamela**: Do you have enough data to do something with that?
* **Flora**: We have data generated using these models in psychiatric patients, we have families with ASD that we have from another project, but we could also use genomic variants you will generate here. I know of course there is a lot of stuff that has to do with harmonizing ChIP-Seq and RNA-Seq data and of course they cross-validate each other. But, again, this is a very specific model so it would be mutually interesting to see what extent what can we use them for? For example, you discovered some potentially interesting things in post-mortem human brain, you are not going to know exactly to what extent these are doing something at the RNA-Seq level and maybe these are expressed very early in development and it’s going to be hard to do in embryonic brain. So, having something that actually develops in vitro because time zero is very, very, early and we never going to get human brains from that stage.
* **Peggy**: So, it sounds like that map, even though it’s not been refined, it will be, at the end of your project, will be very useful for everyone and would generate you as a collaborator on a ton of independent projects.
* **Chunyu**: The first thing we want to do is the co-expression network. We are quite interested in building BP and SZ cases and controls. And for that we will look at long non-coding and coding RNAs. We already have some data and those lncRNAs and copy number regions; this can be even broader using RNA-Seq data. The second thing is a technical issue I tried to address. The technical variations and population variation is an issue. I am trying to gather enough data with technical replicates for the ATAC-Seq. So we’ll create many alleles if you have any those data, either you can work with us or send it to us, I would like to work with the technical variation and population variation.
  + **Kevin**: The next big publication that is going to come out of my lab is a large number of fusion genes and their functional characterization in cancers. We should probably, on a wiki page on Synapse, put together a spreadsheet of what people are planning. That’s worked out fairly well with the ENCODE Consortium. What inevitably happens is that we all promise to do many more papers than we actually get around to doing, but at least the ideas get down on paper and some of them do get processed all the way through, so I think that would be useful.
* **Dan**: Our project pre-dated a lot of this. So, we have two papers that under review now. One on Hi-C in the fetal brain and how you can use that to annotate in the specific regulation and the psychiatric loci. The second is the replication of the AT microarray data in RNA-Seq, microRNAs, and using the normal RNA-Seq protocols for long non-coding RNAs. Then there’s the cross disorder stuff we have the microarray data, but not much RNA-Seq data yet. We’ve used the publically available microarray data to look at that issue but would be worthwhile to do that much more. What will happen in the next year, is that we’ll get hopefully get 5 but a maximum of 8-10 cortical regions from a good number of AT and controls and we’ll have that RNA-Seq finished to understand the patterning and the extent of the changes across the cortex. Those are some of the major things going on right now. Some of the PEC stuff that is funded will be doing much more Hi-C. We find that if we look at fetal brains and compare it to the published HI-C data, half of the loci really that we are identifying the fetal brain can be identified in the fetal cell types but only half of them can’t be. So, it shows us that there’s really a utility to having the tissue specific chromatin data which isn’t really a surprise. I think that’s what going on here is going to be really exciting in tissue and cell-type specificity.
* **Schahram**: We have a methods paper under revision that has a quality control pipeline. For Aim 1, we want to at some point start, by the end of the year, writing some of this data which goes after subject specific signals and maybe some region specific signals. I think that is what we are trying to accomplish by end of this year. And then, if the next wave of PsychENCODE application gets funded, we will look to do similar data things in caudate neurons and then export it to the CMC SZ data collection to look at the three-dimensional genome to examine in SZ and controls tissues. And one last thing, because I’ve been working with Zhiping, the last 5 – 8 years on ChIP-Seq datasets that we established at UMass, we are still having a trickle of papers from these 50 or 60 ChIP-Seq datasets, which we did in the first generation of NeuN sorting.
* **Mark**: Most of what I would say would fall under Flora or Nenad’s project. The other thing that we are working on is two focused projects like the slides on data analysis looking at transposable elements.
* **Pamela**: There were three things that I highlighted that are kind of what I would like to get through this year. One that is most directly PsychENCODE focused, which is to use our Epi-map data to look at inter-individual differences. That will be plus or minus using the epi-QTLs in co-expression format depending on how things go. The other things that are reasonably far long; we are doing a fair amount of conditional eQTL analysis as part of looking at heritability. We’ve looked a lot at how many eQTLs per gene. The outline for that as a paper also involves using the ChIP-Seq data to refine that. The third thing is what we submitted for the other part, which is to do a variety of cell specific work like sorted nuclei and ATAC-Seq but also to do a fair amount of Drop-Seq in freshly dead human people, so that we can try to make various transcription maps that I think would be great for all of us to leverage.
* **Nenad**: I am hoping to finish three key studies. Human developmental ChIP-Seq would be really relevant here, macaque developmental RNA-Seq and mid-fetal RNA-Seq.
* **Greg**: The initial study would probably be to look at what Chunyu, Kevin, and Schahram said specifically looking at regulatory variation from case and controls. In addition, at least for a sub-set of the GWAS loci, we’ll have some of the STARR-Seq data to quantify how strong of a sub-set of these regulatory elements are. And longer term, I think we are interested in looking at 2 different regions of the brain but also the Pop STARR assays that Tim is working on to look at regulatory variations in these high throughput assays that handle which variants are contributing to altered enhancer function.
* **Stella**: We just published a paper DNA methylation in GABA and Glu neurons. We are doing ChIP-Seq… And, if we have enough money left, we will do co-methylation in these samples from different subjects. For future, I would like to collaborate on developing some strategies for regulatory elements for a map.
* **Andrew**: We’re going to write the paper on the WGBS data and then integrating some of the changes we see across development to expression changes. I also envision there being some cell type specific vs. homogenate comparisons. Other papers that might be relevant to PsychENCODE are methods for RNA-Seq data analysis, other ways of looking at transcription besides gene count and transcript count, and also some adjustment approaches for using new measures on RNA quality that might dramatically change results through re-analysis of a lot of public data. Those are forthcoming and probably in the next few months we can circulate a draft to people who are interested and then the methylation will probably be done by the end of the summer.
* **Peter**: We’ll be working on the project that we talked about, finishing the RNA sequencing papers we’ll be working on towards related to the aims. Of course, refining the GWAS loci from BPD and looking at their transcriptional effects. And then, a second paper focused on differential expression on BPD vs. controls both at the gene and the network level. Amending one thing that would be helpful in terms of thinking about this sort of cross-disorder or cross-Consortium analyses that we can do see in addition to the different papers people are working on because to finish out the matrix on the data we are collecting in terms of the genomic features, whether it’s RNA-Seq or whatever methylation mark vs. sample type based on disorder, what brain region, and whether its cell sorted or not I think that’s going to dictate what we really capture and combine. And, one note of caution that I’m concerned about is the little whole we have in that matrix. For example, if I think of BPD, you look at it and say, oh wow, there are 400 samples, but half of them are in dorso-lateral, some are hippocampus, so I think seeing that matrix will really help guide the kind of analyses we do.
* **Pamela**: We did that exercise at the very beginning at all of this, and it’s really interesting how it’s changed a bit for a lot of people and now we are getting to where we can re-do it in a way…we we’re all just projecting when we started. We had no idea exactly how long it was really going to take to do each of these things. So, I think your point is excellent and we should re-do that. I think now we can do it more extensively and fill that in and have a really good idea of what there is and the timeframe.
* **Peter**: I would have two versions of that – what we have now and then another projection.
* **Pamela**: I do have a feeling that this is a much more comfortable group now than the second in-person meeting. Everyone sounds much more with the sharing aspect, which is in human nature of what you would expect to happen. It was much more difficult assembling this in the beginning, and now it actually sounds like it won’t be so hard. I think we will have some areas where lots of people will want to do very similar things. We may have to figure out a good strategy for that.
  1. Timeline for short-term vs. long-term goals – *Mark Gerstein, Nenad Sestan, Daniel Geschwind*
  2. Individual vs. joint publications – *Nenad Sestan*
     1. ENCODE presentation – *Mark Gerstein*
* **Mark**: To set the stage here, I thought I would just say a few words about actually happened with the ENCODE publications. And this is just my view, and I’ll give you a very data oriented discussion.
  + One thing that is worth mentioning that is you have to see these ENCODE papers as kind of a result of a very long process. The first ENCODE grants were funded in 2003. The pilot of that was just that one Nature paper, which was 2007, and the last thing is 2014, which is 11 years later. That’s quite a long period after the initial period of funding and there’s quite a lot funding from that.
  + I wanted to give you my own take on how to think about. I am excited by these publication roll-outs. You can think of these roll-outs as huge, gargantuan terabytes of information being organized and if you think about it, you have the raw data and above you have all the processed summaries of the data, the raw peaks and what-not. And then above that you have the results from the particular table of the code and more specialized papers and supplements and at the very top you have the Consortium papers. And, the Consortium papers you can see it at the top of this huge edifice and to get to the roots you have to build the walls first and build the foundation. What is important to realize is when the reader comes to read about ENCODE, they usually start the top level. So, the average reader is going to read the Consortium publication first or even the abstract of the Consortium publication. And, they’re going to get a very top-down view and they’re going to walk down through this very high level thing, then into the specialized paper, eventually into the data and then eventually to the reads. That’s the way the reader perceives it, but the way the creator or author felt is the complete opposite. They started with all way at the bottom, then the foundation or the specialized publications before working their way up to the top. So, I think we’re going to contrast this top-down vs. bottom up perspective.
  + The other thing to emphasize, the reality is that there have been almost 1,000 ENCODE papers. It’s amazing. NHGRI tabulates it. They have an official tabulation called ENCODE papers which are done by ENCODE authors and they have non-ENCODE papers that use ENCODE data. There’s about 1,200 papers that are not written by non-ENCODE authors but use ENCODE data. You can see that actually the tremendous increase in non-ENCODE papers and the large number of papers being written by the ENCODE Consortium, not just the Consortium papers. It’s not just all these big Consortium papers, which is important to note.
  + You can make a network of all the authorship interactions, co-authoring of papers over time. What we have found when we looked at this network, is you can see clearly that it is important that the people who link the Consortium with the people outside of the Consortium publish inside the consortia and outside the Consortium in terms of connecting the Consortium to other people and getting the data being used.
  + If you look at the Consortia as a network of people that come together and publish separate, the original consortium rollouts you can see how they separate and then came together. If you look at the connectivity between each of the authors and whether they are ENCODE members or not, you can see the special role of these broker individuals who are publishing extensively with people inside the Consortium as well as with people outside the Consortium. That’s just the history of the ENCODE papers and a little bit of the analysis.
  + There a few lessons that can be learned from this. The first is that this stuff takes a long time. Second of all, this concept of top-down vs. bottom-up. To some degree, people get very fixated on the Consortium paper and that’s literally sitting on top of hundreds of smaller analyses. And to build that roof, you have to build the walls. With those sentiments, I thought I would say a few general things for what we could think about for PsychENCODE. General things we might want to think about, this is from the Google Doc Pamela and I worked on – we want to process the data uniformly and we kind of want to characterize healthy brains, build large sets of eQTLs over brain development, and we also want to do integrative analysis across psychiatric diseases. Another thing we can think about is imagine creating a uniform data resource and thinking about small sub-projects that either individual groups are doing or a particular student or post-doc is doing that sort of goes over a number of these datasets, then looking at them together, and gradually looking at these things as larger themes.
    1. Joint publication authorship rights/guidelines and data sharing
* **Geetha**: Should we discuss, in terms of the framework, if people are thinking about publishing, would they include PsychENCODE projects? Would you all be happy to share manuscripts before they are submitted? What does GTEx and ENCODE follow? Basically, the groups have smaller papers and then this larger capstone paper. If we are thinking along similar terms, because we are multiple projects and each project has its own interests, then are joint interest people are thinking of, and we had publications that came out that we did not share in beginning...should we talk about that? GTEx does discuss when they have smaller papers.
* **Pamela**: There are two different kinds of manuscripts, so there’s sort of a secondary analysis of the whole dataset or pieces of the whole dataset where something like the DAC would make the whole dataset and people would use it for something. Those papers get circulated and people can use their own data for things they want to do without circulating.
* **Peggy**: So for ENCODE, there’s a Google Doc where you have to put the title of the paper, the authors, and then the stage it’s in, like “abstract is ready”, “almost submitted”, “under review”. What I was thinking for the smaller papers, before they were actually submitted, if people could present their papers on one of our calls. And that way everyone would know what is going out there because what you actually say in the paper or in the text takes a lot longer than when you have the nice figure set all ready to go.
* **Mette**: I just wanted to see how the acknowledgement statement is and authorship. Is just the Consortium named and not just individual members or…
* **Pamela**: Mostly, they are named as individual members even if the author is just a banner. Sometimes it’s just a Consortium banner, but all authors are authors. So, SZ Working Group of the PGC is the banner author, but listed somewhere in the back are all of the authors. In that case, everybody participates. So, that would be what I assume that’s what we would want for the main papers and that that would be true for someone who didn’t make it in because we’ll all in this process of evolving the Consortium. So, the question is for the individual papers, sometimes there are subsets where the consortium gets named and people are carried as independent authors and sometime not. We have groups, for example the BP group, that are obsessed with being the authors on any sub analysis that happens, so we have all these complex agreements around making sure that anybody else who does secondary analysis does that. Here we’re actually a relatively small group and we could probably agree that for the individual papers that regular scientific standards would work. So, for example, if it was a sub-analysis of something that came out of the DAC that had a centralized dataset that would probably require everyone as authors because a whole bunch of people contributed to making the data and a whole bunch of people worked on putting the dataset together. For the foreseeable future, there would be sub-analyses that might be hybrid so let’s say Jim and I want to do an analysis. If we choose to use our own pipelines, then it would probably be just our own groups. If we decide to take the data from the uniform, then we would pull the people from the DAC. I would suggest that for all of the papers that we have PsychENCODE Consortium just like it’s a person when it’s; let’s say just mine and Schahram’s Epi-Map data that doesn’t involve anyone else, that we put the authors and the PsychENCODE Consortium.
* **Mark**: I think having the Consortium name as an author is a great way to get the name out and stuff like that. But, the thing that I strongly want to point out and that we should do is have two names. I think we should have one name, the PsychENCODE Consortium, when we actually need all the people of the Consortium and it’s not an advertising name and it actually names people. The other name would be the PsychENCODE advertisement that’s just the same name with no person. It’s so important because it’s so confusing when you’re looking at a paper and trying to figure out does that consortium name representing 200 or 0? You have to sort of decide. 1,000 Genomes is an incredibly good consortium unfortunately didn’t do this. For instance, the authorship of the main 1,000 Genomes is the consortium. That is like 1,000 people. They have a paper in NAR with 5 people of those 1,000 of the 1,000 Genomes Consortium. Those thousand people are not the original authors of the paper, but they never really differentiate this, so it’s sort of confusing. If you go into the PubMed and search me, you will find a lot papers that I didn’t know existed until they came out. I was even contacted by a reporter to comment on my paper and I didn’t even know what paper they were talking about.
* **Pamela**: Are you listed as an author in the end?
* **Mark**: Within PubMed there’s a concept of authors and also the concept of collaborators. If you go over a certain number of authors, they have to become collaborators and it’s deeply confusing. The bottom line is that I have experienced this, but I just say that we should have two different names. We should have something like, the “PsychENCODE Group” or the “PsychENCODE Members” or something, but two different names where we actually know what we’re saying when we put this out.
* **Flora**: How long we keep having a code or the PsychENCODE people for?
* **Pamela**: Usually, after a couple years there’s some re-discussion of everything because things change and maybe the group becomes three times as large. We could argue about all this, but I feel so good that we’ve even agreed to the first part of it.
* **Thomas**: There may be a PsychENCODE 2 and the Consortium might continue.
* **Mark**: I think authorship means somethings so you should be precise with authorship. So, yes, you can have these bizarre pathologies that represent 50 people now and then 5 years later there’s another 20 people who are added. Are they now retroactively authors of papers that were previously published?
* **Pamela**: What PGC has done is there’s a freeze. There’s the PGC 1 free. You use the PGC 1 data, you get those authors. Not all of them are on the PGC 2. I also think it is fair for the funders to request an acknowledgement of where the funding is coming from in a certain way and in that sense the pseudo-authorship doesn’t matter so much how long that gets. You’re still going to have a first and last author.
* **Mette**: I assume that for the respective collaborative analysis and individual analysis that will need to be an acknowledgement statement on those publications.
* **Nenad**: I think we have a specific acknowledgement that does not just include the number. I think we should have a standard sentence for individual projects. The question is should we require that from people who let’s say I publish something from my lab and someone uses that dataset, should they also acknowledge the P50 in the sentence?
* **Pamela**: I think they have too. I think we’ve already written that language.
* **Mette**: We have an acknowledgement right now that lists all the grant numbers and PIs as a group.
* **Nenad**: The question is, should we have one specific for PsychENCODE Consortium or something because the grant number search will not get hits automatically. And, I think we should put that on our portal, so that people know…
* **Pamela**: But, they have to agree to that when they download the data.
* **Nenad**: But, add PsychENCODE in that sentence.
* **Mette**: I’ll take a look and if it doesn’t have it, I’ll add it.
* **Daniel**: I have a question in the very sense of where we had started this. Like this one paper that we have under review right now was done before all of these discussions. And, we haven’t included the PsychENCODE Consortium, but moving forward, I’m wondering if this is a separate issue or a personal issue.
* **Pamela**: I have a similar one because with PNC…
* **Nenad**: You should not put the entire Consortium, but having a sentence that has the grant number…
* **Pamela**: For the PGC, they’re trying using ORCID IDs because you can just generate, if you fill out ORCID with your place and appropriate specifiers, and then it’s easy to just generate and update it.
* **Geetha**: Should we talk about the short-term goals? We are talking about when the groups could start joint analysis or keeping in the mind the long-term goals, Consortium-wide analysis, what steps are we going to take on how to have some goals, which we do. As a group what could be accomplish? I heard all the individual projects’ ideas. Jim suggested that for cases and controls maybe we could form disease-specific working groups and for the epigenetic landscape all of the groups could be involved. Is this something we should start right now or six months later when Mark, Zhiping, and Kevin process all of the data? What is the framework? I think in the next six months we will have substantial data, a lot more data than we have today. What are we going to talk about six months from now? Would it the same things or will we be steps ahead?
  + How are we organizing ourselves? We do have case controls as something we are thinking about, comparing how the regulatory elements differ? Greg I think mentioned what projects couldn’t accomplish with their own datasets. Could this be accomplished using datasets from the Consortium and how do we organize ourselves together rather than individual groups?
* **Pamela**: I thought we had the organization around that. The DAC would get their hands really wet with the data and that they analyses would be part of the larger epigenomic RNA-Seq.
* **Kevin**: I agree with Pam. I think this past discussion is gone a long ways to clarifying what we are doing at the individual level and also in terms of a collective. My point of view is that we’re studying RNA, because that’s the most data we have, but then we’d move into other types of data like the ChIP-Seq data via the centralized processed set of data and that will be made available to the group and the people in the group like Greg, Chunyu, etc. that are doing certain kinds of analysis like the eQTL kind of analyses will then churn on that and will have weekly or bi-weekly phone discussions with everyone interested in participating in the conversation and the analysis and once the data area ready will then have a snowball effect where more and more analyses get done, more and more presentation get done on the phone calls, and that will be the forum of how the fine details get put together for the paper and how the Consortium actually is functionally working together. I think right now what we can say and Pam has already articled is that we got a nice broad outline and now we have to get certain tasks done, which fall largely with the DAC when it comes to the collective effort. Once that is done, if not before, we should start having some regular phone calls to talk about the progress on the various analyses and who is doing what and so forth. I am not personally thinking right now that we need to figure that out in the next hour.
* **Geetha**: That’s what I was thinking too.
* **Kevin**: I think that the individual groups are going to develop their own analysis pipelines and get to their point where they’re submitting or published their own papers on their own datasets and, in the meantime, we are doing this collective uniform processing and so forth. And, it’s going to become fairly evident who’s in a position to do which analysis and we’ll work all that out at a future date.
* **Geetha**: I know most of these common analyses are dependent upon DAC processing and the groups right now are thinking on their own because, this is my project and I’m analyzing this data and we’re interested in pursuing this particular question. And, then we have the DAC processing and thinking of how we can best leverage all of the RNA-Seq datasets being generated. And how do we have groups continue this conversation on a regular basis?
* **Mark**: I think some of the key messages from today are that everyone would agree that we want to continue to do the uniform processing and the focus is that we’ll eventually have this nice corpus of data that would be used for some capstone paper. I would also say that there’s a sentiment that we don’t want the capstone paper to be cross-purposes with individual efforts. So, there’s this thought process that we want people to move quickly to get their own individual efforts done. And, I think the way to approach the cross-work is for people to do their individual efforts fairly quickly, not 5 years and the sort of open to discuss in a community forum. I think people should be encouraged to do their own thing but they should discuss it and we should know what is going on with the idea that we’ll all build on that in the future.
* **Pamela**: I think the other thing that happens in this kind of larger project it’s often easier to participate in it once your own group has published on of something. It feels as though you’ve established yourself in that area. It’s certainly true for me and all the things I’ve done. This is a different thing to do in this gene expression. I don’t feel like every piece of the project.
* **Jim**: I think the single most useful product we could do first is make a good eQTL map for brain. I don’t think that’s too much. GTEx is not really stepping up to the plate and doing it well. They’re doing Poly-A, they’re not doing fetal, which is clearly what the field needs. A really good powered eQTL map of the brain. And, yes you have the CommonMind samples, which do a pretty good job, but I mean adding in all the other samples would be even better. Apropos that if we could make this the first thing, then we can also, right now, set a deadline of when the genotypes and RNA-Seq have to be into the database to be part of that first analysis. If that’s this summer, in three months, or next fall, and then if you’re in by then, you’re in it. The one after that would be k27AC, right?
* **Pamela**: I think we have enough methylation as well. And, I think that they should proceed in parallel because in fact the ChIP-Seq marks are really new. I would not be prioritized to do the ChIP-Seq data.
* **Mark**: I would say I think that’s a good idea, Jim. I’d like to agree with the both of you. One thing about making a larger eQTL map could be seen as a common resource project that doesn’t threaten anyone. That might be good thing. They’re lots of eQTL maps and it’s just sort of putting all the data together and doing something that we could do collectively rather than individually. I’m not sure what I said is completely rue though.
* **Kevin**: The comment that I was going to say earlier is that just pragmatically the point about first having the individual labs get their work done and then having the group paper come out makes a lot of sense because it is going to take some time to do these Consortia papers. There’s a lot of complexity, a lot of heterogeneity.

**4:45 – 5:00: Wrap-up discussion**

**5:00:** **Adjourn**