**Summary of the HuBMAP Planning Meeting**

June 15, 2016

**Executive Summary**

Eighteen experts in the fields of molecular mapping and single cell analysis met on June 15, 2016 to discuss challenges and opportunities for building maps of cells in human tissues to understand organization and cell-cell interactions. The discussion throughout the day focused on two themes: 1) the process for building an atlas of maps of cell types that compose human tissues and 2) technology development that would greatly enhance this process. An iterative approach of surveying using FISH imaging and deeper dives using RNA-seq that would then feedback to the surveying was identified as a well-developed process for building an atlas. There was significant interest in high-throughput imaging mass spectroscopy and in situ sequencing, with the feeling that further development of these technologies would benefit the program. A number of practical challenges were identified including acquisition of high quality tissue specimens, annotation of datasets, validation of data and analytical techniques, and collection of metadata. Beyond a focus on primary human cells and the recognition that complementary work on model systems would be very beneficial, the participants suggested that expert knowledge about individual tissues was needed to optimize data collection and guide analysis.

**Process for Planning the Meeting**

In early 2016 the HuBMAP Working Group was asked to contribute names of leaders in the field of single cell analysis of tissues keeping sight of the different areas of expertise that were needed – i.e. technology development, bioinformatics, experience with similar programs, scaling to higher throughput and higher content, etc. Participants were identified initially by multiple nominations from the working group. Thirty-eight experts were invited to the in-person meeting on June 15 in March – April 2016, twenty of whom accepted. The experts were asked to prepare brief presentations related to and be prepared to discuss a number of questions identified by the working group (attached at the end of this summary). Two invited experts were unable to attend because of last minute events, and three experts were unable to attend in person, but were able to contribute online via Webex.

The in-person meeting ran from 8:30am to 6pm on June 15, 2016 and the agenda (attached to this summary) was organized into 4 sessions of 105 minutes, giving each expert the chance to give a brief 12 minute statement in response to the questions posed in advance by the working group and then to have a round table discussion of the questions for half an hour. The session titles were:

1. Approaches to Studying Cell Organization and Interactions in Tissues and Distributed Systems
2. High-Content, High-Throughput Tissue Analysis
3. Integrating Data, Visualization and Analysis
4. Emerging Technologies

In addition to the invited experts, many of the working group members participated in the meeting either in person or via Webex.

**Summary of Key Points from the HuBMAP Meeting**

1. ***Approaches to Studying Cell Organization and Interactions in Tissues and Distributed Systems***

***Panelists:***

* Loyal Goff, Ph.D. - Johns Hopkins School of Medicine
* Eng Lo, Ph.D – Harvard University
* Ellen Rothenberg, Ph.D. – California Institute of Technology
* Steven Potter, Ph.D. - Cincinnati Children's Hospital Medical Center
* Mark Vidal, Ph.D. – Harvard Medical School (remote)

***Presentations:***

Loyal Goff described the use of RNA-Seq to segregate and identify paralogous neurons in mice, observing that lncRNA has the highest predictive power for cell type and that there is potentially a transcriptional signature of neuronal activity within a given population. Eng Lo described a range of studies of the mouse vasculome and cell-cell interactions within the neurovascular unit; he highlighted how the effects of aging, blood pressure and diabetes could be distinguished via RNA-Seq and the interactions of astrocytes in response to different challenges as well as the importance of distinguishing biological variation from technical noise. Ellen Rothenberg highlighted the need to understand relevant and reversible variance and in particular the response trajectory of cells, arguing that this would require validation in models systems that could be reproducibly manipulated. Steven Potter described the strengths and weaknesses of RNA-Seq in the GUDMAP and LUNGMAP initiatives, including that it gives a very deep unbiased look at cells without requiring labeling, but that dissociation alters gene expression and that pulsatile and low expression level RNA cannot be easily identified. Steve also identified the need for multiple maps covering aging, different organisms and healthy to diseased. Marc Vidal described a framework for a multi-scale atlas of functional maps, and the opportunities and challenges of building a molecular interactome model for a dynamic map of functional relationships. He noted that dynamical aspects should not prevent us defining reference maps.

***Discussion:***

The participants discussed a number of challenges and opportunities to studying cell type, organization and interactions in tissue. Overall, the group felt that the recent development and validation of a number of experimental technologies and data warehousing systems now gives us the tools to tackle building cell maps, but that the biggest challenges are in the data processing and analysis pipeline and how to intelligently sample cells and tissues.

**Human Cell Sources.** The challenge of sourcing high quality, different human tissues was discussed with surgical discards, biopsies, autopsies and organ donations identified as potential sources. The group identified some tissues such as kidney that are hard to obtain as high quality samples and the limits to using biobanks because of lack of compatibility between single cell analysis techniques and many tissue preservation techniques, but also pointed out that large samples are not needed in many cases. Furthermore, the smaller the cell and the more tightly embedded (e.g. bladder), then the harder to dissociate for ‘omics analysis – there are signatures of freezing and dissociation in RNAseq. FISH does not require dissociation and you can get subcellular spatial information and morphology. The group felt that going from in vivo to ex vivo and primary cell cultures decreased the ability to measure cells in their natural environment but allowed temporal / functional / perturbation measurements. Organoids were felt not to represent good model systems.

**Cell Characterization.** The participants discussed a number of ways to characterize cells and highlighted the challenge of capturing temporal dynamics and low copy numbers with low efficiency, point measurements. lncRNA was discussed as a case where they could be a useful biomarker, but their low abundance makes reliable measurements very challenging. The group discussed how to measure and understand cell state, and that a framework is needed along with a lot of data to map out whether states are continuous or discrete and their stability, and how underlying regulatory elements vary.

**Cell Organization.** The group considered the need for a reference map, and whether the vasculature is a useful system for identify and describing location. There is a need for better defined ontologies for describing expression profiles and clusters and the participants felt that different reference maps would be needed for different purposes.

1. ***High-Content, High-Throughput Tissue Analysis***

***Panelists:***

* Jim Eberwine, Ph.D. – University of Pennsylvania
* Garry Nolan, Ph.D – Stanford University
* Aviv Regev, Ph.D. – Broad Institute of MIT and Harvard
* Lani Wu, Ph.D. – University of California San Francisco

***Presentations:***

Jim Eberwine described several approaches for quantifying biomolecular, functional and morphological features of neuronal cells to capture temporal dynamics and the need for a robust ontology to integrate and process data from different approaches. He also described an emerging technology from his lab for performing transcriptomics of fixed human tissue (TISA). Garry Nolan described two approaches for highly content, high throughput data collection: 1) an imaging mass spec approach utilizing rare isotope labelling (MIBI-TOF) and 2) an automated barcoding system for 22 color imaging that can be retrofitted to most inverted microscopes (CODEX). He illustrated the throughout by showing a 1cm2 slice of human tonsil (>200,000 cells) that was imaged in 30 hours and emphasized the need to engage AI experts for automated annotation and inference. Aviv Regev described an atlas as identifying: 1) type, 2) state, 3) transitions, 4) lineage, 5) location and 6) interactions. She highlighted that drop-Seq and InDrop have pushed single cell throughput to tens of thousands of cells and thousands of genes analyzed per day, completely mapping the cell types of the retina, showing conservation of cell cycle expression patterns across mouse and human gliomas, observing that precocious dendritic cells (1%) coordinate response to LPS challenge, and that cell-cell associations in the gut can be studied. She noted that samples, technology and computational resources as well as community engagement are required for a human cell atlas, and that it is necessary to study engineered system to explain how tissues work. Lani Wu presented three cases to illustrate that the depth of molecular profiling, how much tissue to sample and how to infer dynamics from the biomarkers tracked could be optimized by picking non-redundant features.

***Discussion:***

The group discussed a number of the challenges and opportunities to scaling the depth and throughput of different technologies to be general purpose solutions for analyzing human tissues. One boundary given to the participants is that a concept could not focus on a specific disease, tissue or cell type; the group felt that wound repair was a general mechanism to consider and the immune system and epithelial stroma are present in many organ systems and is the origin of many diseases. Further, they felt that cross-species comparison would be enlightening, particularly with mice, though the focus should be primary human cells. There was support for any proposed program to demonstrate a cutting edge pipeline for “normal” tissue analysis that could be replicated at the IC-level for disease-specific analysis and that a program should include both technology development and a data resource. Ideally the program would be hypothesis-generating and lead to some sort of predictive parameters for understanding disease emergence and treatment response at the cellular level.

**Sampling Strategies.** The participants identified three approaches to sampling: 1) survey mode trying to cover large numbers of cells at low depth, 2) enrichment mode looking at a specific cell populations at higher depth, 3) adaptive approaches that can stream data and drill down selectively. Choosing a mix of reproducible and well-studied tissues (e.g. gut, retina) with ones less studied (e.g. stroma). The challenge of mapping plastic tissues that undergo regular remodeling or are dominated by motile cells were also considered. There were different viewpoints on how much data would be enough, though there was support for an iterative approach alternating between survey mode and enrichment mode.

**Validation.** The group identified the need to compare whether cell types identified by one method correspond with those from another and that taxonomies are consistent. Teasing out technological noise from biological variation was deemed difficult but an important challenge, particularly given the large variation that can occur during states and transitions. Tracking temporal dynamics or having a well-defined perturbation may help identify what measured variability is real and meaningful.

**Human Tissue Mapping.** In addition to short-range cell-cell communications, some of the participants pointed out that long-range communication is measureable and the boundaries of a cell as an information processing package are fuzzy. The group also considered whether there is a significant need to build maps of gender differences, developmental process and aging and whether a composite reference map, similar to the human genome, is a desirable outcome for the program. With the limited time and resources of a program, the feeling that a broad survey combined with deeper dives into a limited number of tissues would provide a realistic solution.

**Data Analysis Pipeline.** There was a discussion around how a cloud-based neural network may be trained to identify sub-cellular and cell organization features in an automated fashion. This was perceived to be a potential solution to how to annotate imaging datasets to understand cellular organization – running challenges or competitions or encouraging other approaches may be necessary to engage the computer science community to tackle some of the big data challenges this program would generate.

1. ***Integrating Data, Visualization and Analysis***

***Panelists:***

* Helen Blau, Ph.D. – Stanford University (remote)
* Wataru Fujibuchi, Ph.D. – Kyoto University
* Katy Börner, Ph.D. – Indiana University
* Cole Trapnell, Ph.D. – University of Washington
* Orit Rozenblatt-Rosen, Ph.D. – Broad Institute

***Presentations:***

Helen Blau described the advantages of a single cell analysis approach to identify drug targets to enhance aged muscle stem cell function using ATAC-seq, RNA-seq, microscopy and CyTOF and how to integrate these datasets to build a comprehensive model. Waturu Fujibuchi described the five dimensions (differentiation, classification, markers, conversion, organism) used in the SHOGoiN database and the university coding system used to disambiguate cell type. He estimated that to perform RNA-seq on all 3 x 1010 cells in a mouse would cost $900 billion and take 800 years but the drop in cost and increase in performance is easily bringing into the same ballpark as where the human genome project started from and that there is a lot of international interest in pursuing such a project. Katy Borner described how complex datasets can be visualized for interactive interpretation and analysis, highlighting that a book of maps is needed to understand the variety of patterns and trends in complex datasets. Cole Trapnell argued that single cell indexing (CSI) could provide a route to multi-omic single-cell assays, that the size (10,000’s of cells) and sparsity (97% empty) of datasets are difficult to interpret and that variation confounds unsupervised cell classification. Furthermore he provided examples of state transitions and argued that in vivo lineage tracing could resolve the question of recruitment versus reservoir of tissue resident macrophages. Orit Rosen described how analysis of RNA-Seq data from thousands of cells can be used to classify cells, their molecular status (cell cycle, drug resistance, activation), and infer cell-cell interactions to study regulatory circuits.

***Discussion:***

**Meta-Data.** The participants identified the need to collect a range of variables with the data, including experimental conditions, origin and processing of cells, perturbations, medical background of donor, data filtering and cleaning algorithms applied etc.

**Data Complexity.** The group did not offer any insights into the minimum set of information needed to create a cell profile, however they chose to highlight the complexity of potential data that could be collected, from molecular information to function, morphological and physical and that are math / stats tools can’t cope with the volume or variety of data. The balance between DNA sequencing of the donor, transcriptional profiling of many individual cells, and morphological, physical and functional measures of a select number of cells will be dependent on questions if interest – e.g. mosaicism vs. tumor vs. T-cell activation vs. development vs. neural circuits.

**Automated Analysis.** The participants felt that moving to unsupervised classification of data from single cell analysis is a significant challenge and going to be an issue for several years. They also emphasized the need to analyze differences between algorithms, data pipelines, experimental procedures in different labs and the influence of individual researchers.

1. ***Emerging Technologies***

***Panelists:***

* Long Cai, Ph.D. – California Institute of Technology
* Je H. Lee, Ph.D. - Cold Spring Harbor Laboratory
* Arjun Raj, Ph.D. – University of Pennsylvania
* Gajus Worthington – Fluidigm Corp.
* Xiaowei Zhuang, Ph.D. – Harvard University

***Presentations:***

Long Cai presented details on how sequential barcoded FISH (seqFISH) can be used for in situ profiling of single cells in tissue and is compatible with CLARITY and lightsheet imaging to study thick (1 mm) sections. Jay Lee described the sensitivity, reproducibility, scalability and limitations (100 mRNA / cell, tissue specific, sequence biased) of fluorescent in situ sequencing (FISSEQ). He described how these limitations may be addressed by heuristic in situ targeted oligopaint sequencing (HISTO-seq) that uses sequencing by ligation on RNA combined with temporal coding. Arjun Raj argued that high dimensional data will always allow us to cluster however the dimensions being measured may not matter and may obscure important hidden variables (e.g. cell volume). Gajus Worthington highlighted immunotherapy as a revolutionary treatment approach that will require single cell analysis of the engineered cell population for safety and efficacy and that laser ablation coupled to CyTOF can provide high throughput, 32-color analysis of tissue. Xiaowei Zhuang described multiplexed, error-robust fluorescence in situ hybridization (MERFISH) an approach for directly imaging the transcriptome via sequential imaging over 10-16 rounds. This approach can be extended to imaging chromosome organization in the nucleus and translation of individual mRNA molecules.

***Discussion:***

**Emerging Technologies.** The participants identified a number of technologies which are emerging at the single cell level including metabolomics, proteomics, sub-cellular spatial resolution of biomolecules, digital counting of RNA and detection of SNPs. The group also discussed how the definitions of cell “type”, “subtype”, “state”, “normal” tissue and biological “noise” continue to evolve as our understanding changes. In general terms “type” can be related to lineage and “state” to signaling (intrinsic and extrinsic).

**Human vs. Models.** Humans were perceived as the most compelling organism to map and that by mapping another model organism first would not necessarily be of any benefit. However the group noted that there are advantages including many sources of variability can be controlled, there are better quality antibodies available, tissue is more readily available, and ex vivo temporal and perturbation studies are easier to perform. There was also interest in comparing cell types and organized structures across organisms.