# Common Coordinate Framework Meeting

December 11-12, 2017

Bethesda, MD

*February 12, 2018*

This meeting summary was prepared by Kelly Beazley, Rose Li and Associates, Inc., under contract to the National Institutes of Health Office of Strategic Coordination (Contract No. HHSN2632016000241). The views expressed in this document reflect both individual and collective opinions of the meeting participants and not necessarily those of the National Institutes of Health. Review of earlier versions of this meeting summary by the following individuals is gratefully acknowledged: Nancy Tuvesson (Rose Li and Associates).

# 1. Executive Summary

The creation of an effective atlas of human tissues at high resolution requires the development of a common reference map. Coordinate systems enable the use of common landmarks for the integration of reference maps at differing scales in one common framework. A common coordinate framework (CCF) for the human body must uniquely and reproducibly define any location in the human body. The CCF is a projected coordinate system in space that is defined relative to one or more origins. The challenges to developing a human CCF are how to define a robust set of origin points that are practical over different anatomical scales and across the natural variance of human bodies, and how to handle the complexity of the projections from these origin points and the relationships among them.

On December 11-12, 2017, the National Institutes of Health (NIH) hosted the Common Coordinate Framework Meeting in Washington, DC, to foster discussion among and solicit guidance from experts on strategies for developing and implementing a successful CCF for the human body (see Appendix A for agenda). Organized by NIH, the Broad Institute, the Sanger Institute, and the Chan Zuckerberg Initiative, this meeting engaged anatomists, pathologists, clinicians, organ experts, technology experts, machine-learning experts, software engineers, and visualization experts from across the world in discussions about strategies for building and piloting the CCF (see Appendix B for participant list).

The discussion focused on (1) *existing knowledge and tools* that could be leveraged in the development of the CCF, (2) *key features*, and associated challenges, to create a durable CCF that can tolerate human variability and function across lifespan and disease, and (3) *potential pilots* to build and test the CCF. Meeting participants proposed pilot projects for both data collection and development of the CCF and supported an iterative approach to building the CCF—first as a spatial framework and later using probabilistic modeling. Important next steps were identified, including optimization of tissue collection, expansion of ontologies, and establishment of quality control parameters.

# 2. Background: Existing Knowledge and Tools

To map the human body at cellular resolution, human variability must be captured on a common coordinate framework (CCF). Considerations for development of the CCF are how to effectively distinguish between anatomical and cellular functions and how to determine what variations are meaningful. In September 2017, a small group met to identify perceived roadblocks and to frame discussion points prior to the CCF meeting. This group emphasized the importance of standardizing collection and documentation of primary samples, and highlighted contrasting experimental approaches to spatial mapping (gridding/barcoding versus direct measurement in situ) and the role of computational inference between them. The group identified a tension between building a CCF from existing knowledge and learning a CCF as data are generated; ontologies as a potential bridge between existing knowledge and new data; and the brain community as a valuable knowledge base for developing a successful CCF.

## 2.1. Human Anatomy and Atlases

There are many approaches to studying anatomy, including regional, systemic, clinical, developmental, surface, radiological, and microanatomical. Systemic anatomy is most relevant to developing a spatial CCF for the human body. Different levels of organization exist within the human body, from the cellular level, through tissues and organs, up to the whole body. Anatomical position refers to body position in a very precise manner that is adopted globally for anatomicomedical descriptions. This is a standard in anatomy as well as medicine and pathology, and this terminology is ingrained in the medical community.

A reference atlas is a series of images, ranging from low to high magnification, from individual specimens that show the physical details of cells, tissues, and organs, as well as their orientation within the human body. Idealized versions of human anatomy are used to create these atlases, and interpreting individual variation, even at the gross scale, is a challenge. At the anatomical scale, some variations are considered "normal" (e.g., different appendix orientations). At the tissue level, histologic similarities exist between distinct and functionally different tissues. For example, vagina and esophagus are cytologically the same epithelia, so discrimination between histologic sections of the two tissues cannot occur without the capture of esophageal glands in the section.

For the brain, there are two types of reference atlases: histological atlases (one brain, spatially mapped at cellular resolution) and probabilistic atlases[1] (functional maps drawn from many specimens at the resolutions obtainable with MRI). The first true cellular atlas of adult human brain[2] was obtained via a time-intensive process using a single brain that was slabbed, sectioned, stained, and then microscopically-mapped and annotated with digital cartography at a final resolution of 1-2 microns. However, this atlas remains a series of annotated plates, rather than a true 3D framework. Further, it illustrates an important lesson for spatially mapping other organs at cellular resolution: more efficient tools are needed to reduce the significant investment of time required to map a single complete specimen.

## 2.2. Lessons Learned from the Brain

Layers of complexity exist within the brain: it consists of hundreds of regions and nuclei with distinct functions, there is topography within regions, cell-type specificity is reflected in the cytoarchitecture, and both regional and cell type–specific anatomical and functional connectivity exist. Spatial organization is highly relevant in the brain, and robust spatial CCFs have been developed for both *Drosophila*[3] and mouse[4] brains, providing useful knowledge bases for developing a CCF for the human brain.

The Allen Mouse Brain Atlas (AMBA)[5] has been successfully used for large-scale data mapping, quantification, presentation, and analysis. The AMBA CCF was developed and refined through

[1] See MNI Big Brain: http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach, https://bigbrain.loris.ca/main.php; Human Connectome Project: www.humanconnectomeproject.org/; Allen Brain Atlas: www.brain-map.org/.
[2] See BrainSpan: www.brainspan.org/.
[3] See Virtual Fly Brain: https://v2a.virtualflybrain.org.
[4] See Allen Mouse Brain Atlas: http://mouse.brain-map.org/.
[5] Lein E, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature,* 2007;445:168-76.

iterations: the first version was a spatial framework built from the 3D reconstruction of a single brain, and successive versions refined this framework using hundreds of additional brains and probabilistic modeling across more than 12 parameters. Two steps were performed during each iteration: each specimen was deformably registered to the template and averaged together, and then the average deformation of all specimens was used to deform the average image. The shaped, normalized average was used as the anatomical template in the next iteration, and this was repeated until the magnitude of the average deformation reached a given threshold at increasingly smaller resolution. The AMBA uses an automated informatics data processing pipeline to deliver connectivity and projection data, single-cell characterization, and *in vivo* calcium imaging to the 3D reference model. Current strategies for mapping in the mouse brain follow a pragmatic approach: use the tools available (e.g., ontology-based structure mapping or collections of annotated 2D plates) and map at multiple levels.

Virtual Fly Brain[6] is an interactive tool that integrates neuroanatomical and expression data from the *Drosophila* brain onto a 3D viewer, and annotates these data using ontology. Like the AMBA, the *Drosophila* brain CCF is viewed as expression patterns painted on a standard brain. Both use affine registration, are mirrored across the midline, and offer a standard delineation of structures with a parts tree to complement. They also share a similar data pipeline (image registrations and some degree of automated mapping to regions). Reference boundaries in both systems are defined using multiple imaging modalities and are redefined as new data identifies cells that cross boundaries. Although both *Drosophila* and mouse systems use affine registration, *Drosophila* has no universal coordinates because whole brains can be imaged in one field. In contrast, the mouse brain uses a landmark-based coordinate system for whole brain and cortex. Current efforts at human brain atlases also employ a landmark-based approach to define the coordinate system. However, unlike both *Drosophila* and mouse atlases, the human atlas also uses landmarks for neuronal registration, complicating visualization of the registration. Painted volumes on standard brain are powerful, as seen in *Drosophila*, but very small differences in registration can produce confounding results.

Mapping of the *Drosophila* brain epitomizes the power of robust ontologies. The *Drosophila* anatomy ontology (DAO)[7] has more than 10,000 terms, 5,000 of which are devoted to brain. This richness of ontology, combined with the Virtual Fly Brain clustering tool NBLAST[8], allows for query by classification of parts, types, regions, location, innervation, fasciculation, synaptic connections, lineage, and function. In addition, it has given a standard or typical image for each neuron class. Moreover, machine learning is being used to great effect in cell ontologies to define novel cell types.

The human brain poses two additional problems: variability and size. Whereas the mouse brain atlas was built around an isogenic mouse line, the human brain is more variable across individuals. In terms of size, while the human cortex is only about twice the thickness in

[6] See Virtual Fly Brain: www.virtualflybrain.org/site/vfb_site/features.htm#tech.
[7] Costa M, et al. The Drosophila anatomy ontology. *Journal of Biomedical Semantics,* 2013;4:32.
[8] Costa M, et al. NBLAST: Rapid, Sensitive Comparison of Neuronal Structure and Construction of Neuron Family Databases. *Neuron,* 2016;91:293-311.

humans compared to mouse, and cell size is not significantly larger, the human brain is three orders of magnitude larger in terms of total volume. Despite the larger size and variability, some extensible principles from the mouse atlas can be drawn on for the development of a CCF for the human brain:

- start as a spatial framework and move to a probabilistic coordinate system as more samples are processed
- develop a standardized hierarchical structural ontology
- apply iterative annotations based on multiple modalities
- build in computational tools for both mapping data to and extracting data from the CCF

Single-cell resolution also presents new challenges, despite a rich history of mapping the brain. Meeting participants suggested a general strategy for constructing a foundational cell-type atlas for the entire human brain: unbiased transcriptome-based cell-type classification in different brain regions; spatial mapping of in-tissue sections (i.e., a "census"); analysis of features of transcriptomic types using multimodal analysis and post-hoc gene expression analysis; and then mapping to a 3D CCF.

### 2.3. Computational Approaches and Tools to Build the CCF

Defining characteristics of individual cells are encoded in their molecular signature (e.g., cell type, state, transitions, lineage), which includes positional information and interactions with other cells. Therefore, we can in theory reconstruct a positional map from sequencing data. The challenge is creating an idealized average with real-life specimens that have remarkable diversity. Mapping cells onto a CCF is a way to aggregate and organize this complex, variable data to facilitate quantitative analysis as a function of spatial position.

Meeting participants discussed two approaches to developing a CCF: spatial coordinate frameworks and generative probabilistic models. With the former, the average of a subset of specimens is generated as a reference, axes are defined on this reference in 3D physical space, and then these axes become the coordinate system of the CCF. With the latter, both the average and the variance of the specimens are used to represent the complete probability distribution of any given cell's location, and the axes of the coordinate system are defined in a potentially highly dimensional space based on latent variables. The spatial framework is more intuitive because the axes can be drawn and viewed in observable space, but the generative model is more powerful. A CCF based on the generative probabilistic model captures more information on variability that may not be readily visible and allows information to be intuited back to the reference (i.e., you can learn from the model).

Spatial coordinate models are intuitive, but they are based only on what is already known; in contrast, probabilistic models are based on what is known and what can be inferred. Interpretability of the axes in the spatial framework is straightforward, but complexity and variability are difficult to capture with a simple average. An example of the power of general probabilistic modeling comes from work by the Allen Institute for Cell Science,[9] in which images

[9] See https://arxiv.org/pdf/1705.00092.pdf.

of cells labeled with different combinations of nuclear and membrane stains were used to predict unobserved combinations of nuclear and membrane stains. Despite the enormous variability between sample cells, this approach yielded a model that could classify the localization of proteins based on cell and nuclear shape and could predict the localization of unobserved structures de novo. However, this model was generated from clonal cell lines, which reduces inherent variability.

The generative probabilistic model requires a comprehensive statistical model and posterior predictive checks to define the tolerable level of uncertainty, which many existing tools in machine language can accomplish. Two statistical models were suggested: bootstrapping and model-based posterior estimates. Although the generative probabilistic approach is powerful, regardless of how the latent system is constructed, it cannot learn relationships that are not built in a priori. Participants agreed that the CCF should be an iterative model: first building a spatial coordinate framework based on what is currently known and then increasingly moving toward modeling approaches once more data are in place. In the initial iterations, the model can be tested by comparing predictions to actual data. This testing will show where additional sampling is needed, and ontology (see Section 3.3.3) can be used as a tool to help anchor new data to prior existing knowledge.

For the CCF to be adopted and embraced, computational tools need to be developed to allow researchers to map their data onto the CCF. These tools should incorporate bi-directional flow of data: users should be able to input either (1) transcriptomic data to see what types of cells in the atlas their data maps to or (2) spatial histological data to see what cells are most likely being labeled. Desirable tools should be able to recapitulate known structure-function relationships and to predict new ones (e.g., the NBLAST tool in Virtual Fly Brain, which has the capability to score morphological similarity of neurons, both known and those predicted via unbiased clustering from single-cell transcriptomics).

# 3. Key Considerations for Building a CCF

A comprehensive reference atlas displaying the types and properties of all human cells will provide the basis for understanding and monitoring health and disease. The CCF is essential for mapping the anatomical location of every cell from each organ or system in a comparable manner. Such a mapping is difficult for several reasons, both intrinsic and extrinsic, and a robust CCF must be able to adapt to new knowledge, uses, and technological developments. The development of a durable CCF should consider the following important parameters: tissue/specimen collection and processing, proper identification of the anatomical origin of tissue/specimen, data acquisition and analysis, integration of scales, a common language, and quality control.

## 3.1. Tissue/Specimen Collection and Processing

Construction of the CCF begins with the collection and processing of target human tissues for analysis. Meeting participants discussed the factors important to identification of the ideal approaches to do so, including determination of optimal sources for human tissues, opportunities to leverage postmortem tissues, and refinements to tissue processing.

***3.1.1. Determination of optimal sources for human tissues*.** Human tissues for research purposes can be obtained through three main methods: biopsy, surgical resection, and autopsy. Autopsy offers distinct advantages in terms of the quantity of tissue that can be obtained and the range of tissue types available. However, a key question is whether tissue collected via autopsy is comparable in quality to live tissue collected via biopsy or resection. A participant suggested a ***pilot study*** to determine the relative equivalency of samples obtained by different methods, and the flexibility of a potential CCF to re-create tissue organization of varying complexity.

***3.1.2. Opportunities to leverage postmortem tissues.*** The use of postmortem tissues obtained from rapid autopsy programs offers several potential advantages to CCF investigators. First, these programs have existing tissue repositories and established pipelines for processing additional postmortem samples. Second, autopsy programs offer the opportunity for hypersampling—the collection of substantial quantities of tissue from across the entire organ—for normal tissues. This both increases the probability of capturing all key features of an organ and allows for distribution of the same patient sample to multiple CCF investigators for standardization and optimization of downstream protocols. Third, this method allows for greater capture of human variability. For example, patients are obese and thin, pediatric and adult, and have differing comorbidities. Meeting participants highlighted examples of sources for postmortem tissues for adult[10] and embryonic[11] tissues.

***3.1.3. Refinements to tissue processing*.** To improve spatial mapping at cellular resolution, distortion of the tissues should be minimized during sectioning and segmentation should ideally be automated to minimize human bias. Participants discussed three techniques to refine spatial mapping of cells within tissue sections, each with increasing stringency and degrees of difficulty to implement. The first technique, and the easiest to rapidly implement across all CCF investigators, is the tape transfer method for sectioning frozen cryotome sections.[12] Compared to the traditional method of transfer, which uses a brush, this method grossly preserves geometry and the relative positions of disconnected or soft pieces of tissue.

The second, more stringent technique uses differential geometry-based methods for atlas mapping guided by the alternate section's histochemistry to remove distortion error from the average spatial model.[13] The third technique, considered to be the gold standard, is the "automated anatomist" —a refinement of atlas mapping based on direct image-based segmentation of histochemical stained sections using machine learning as an alternative to manual segmentation. Although calculations suggest that this approach has the future potential

[10] See https://www.mskcc.org/sites/default/files/node/146506/document/appendix-d-program-brochure_new.pdf.
[11] See http://www.hdbr.org.
[12] Pinskiy V, et al. High-throughput method of whole-brain sectioning, using the tape-transfer technique. *PLoS ONE,* 2015;10(7):e0102363.
[13] Majka P, et al. Towards a comprehensive atlas of cortical connections in a primate brain: Mapping tracer injection studies of the common marmoset into a reference digital template. *J Comp Neurol,* 2016;524(11):2161-81.

to process an entire human brain in 2-4 weeks, it may not be fully developed within the 5-year timeframe of this project.

### 3.2. Proper Identification of the Original Location of Tissue/Specimen

Participants highlighted several key considerations pertaining to the ability to accurately identify the source of tissue specimens: what is the right metadata to gather, what is the right language to use at the point of collection, and what tools can be developed to capture the metadata.

***3.2.1. Metadata to gather*.** Participants identified three sources of information as minimally desirable to document the original location and orientation of surgical- or autopsy-derived tissue specimens: text labels, photographic images of the sampling sites, and some type of diagrammatic annotation. Potential pitfalls, and opportunities for growth, include the availability of comparable equipment at different collection sites. They discussed the consequences of not having metadata from all three sources and whether the location can be further resolved after collection with expression analysis and spatial transcriptomics. They agreed that the level of metadata detail necessary to capture proper location of a specimen needs surgical field testing, and that metadata collection will likely be an iterative process.

***3.2.2. Language at the point of collection*.** Participants noted that surgeons are unlikely to adopt the language of the CCF, which may necessitate the reconciliation of anatomical vocabulary with biological vocabulary within the CCF. The CCF may need to develop an ontology that deals with frequently used location terms, recognizing that precise location and orientation relative to something else is difficult. Although ontology is a necessary function of the CCF, perhaps as important as the mapping itself (see Section 3.3.3), ontology mapping will likely be applied at the point of data wrangling rather than data capture and the CCF inputs will need to adapt to the anatomicomedical vocabulary.

***3.2.3. Tools to capture metadata*.** The ideal tool will allow capture of as much information as possible about the location and orientation of a specimen within its original context in the body at the time of collection. However, the time required to process specimens must be considered: if there is an expectation to collect a large amount of data, people likely will not do it. Tools can improve minimum standards by reducing the time to input metadata; however, the optimal tool will not be over-structured. Allowing users free text ability to add comments may help to address inter-individual variability of specimens. An optimized tool will find the right balance between structured information and free text, and between minimal and complete data. Participants suggested REDCap[14]—a pre-designed web application with pulldown fields for consistency and free text fields that also allows for pictures to be taken and assigned to specific cases—as an existing model to leverage in creation of a custom metadata collection tool.

For diagrammatic annotation of specimen origin, members of one working group suggested a novel strategy that they termed the “anatogram” —loosely defined as a simple stylized representation of each system that will allow whoever is excising the samples to quickly note

[14] See https://www.project-redcap.org/.

where they think the tissue is obtained from. A first instance of an anatogram may be a simple paper diagram or drawing with gridded sections where an "X" can be marked at the location from which the specimen was removed. Later iterations could be an interactive online diagram in application form on an accessible tablet, perhaps tied into the overarching metadata collection tool. Key attributes of the anatogram, regardless of form, would be the ability of the surgeons and pathologists collecting specimens to include free text for nonstandard notations, and the ability for the tool to gracefully degrade (i.e., allowing scaled options for input data).

Participants noted the absence of surgeons at the meeting, who will be important contributors to the determination of best practices for ensuring that the location of specimens collected either pre- or postmortem is properly captured. Surgeons can better inform understanding of the tissue extraction process and the real constraints to human tissue collection, which often does not occur in optimized or ideal conditions. Surgeons can also provide valuable input on the development of common language and tools, such as the anatogram, to capture information on the location and orientation of tissue specimens.

### 3.3. Developing a Common Language

For the CCF to be adopted, a common language must be developed that is understandable for a diversity of end users and that is interactive with spatial data networks so that users can retrieve data rapidly. A common language must also be intuitive for tissue contributors, such as surgeons and pathologists, so that they can quickly and easily input their annotations during specimen collection. Different types of language should be considered during the development and implementation of the CCF, notably vocabularies, labels, and ontologies.

***3.3.1. Vocabularies*.** Historically, atlases of the human body have been described using anatomical terms. As we move into a single-cell world, we need to define the basis for an integrated common language that resolves molecular, cellular, tissue, and anatomical terms. Because surgeons will collect the specimens, the CCF should incorporate anatomical language into the data collection strategy. Further, because surgeons are unlikely to adopt a common language, anatomical language must be translated post-collection. Cross-links to Wikipedia or a similar site could be embedded to correlate vocabularies between capture, input, and output.

***3.3.2. Labels*.** Two meanings of the term "label" should be considered (and perhaps more importantly distinguished) when communicating both internally among CCF investigators and externally to end users. One form of label is a subjective assignment made by people, such as the text labels gathered as metadata, which the CCF should attempt to standardize. This may be done, for example, using structured portions of the metadata capture tool (see Section 3.2.3) or in the user input fields for the CCF interface. Alternatively, label can refer to the procedure of detecting cells or molecules using stains or tags. This type of unbiased label does not involve people.

***3.3.3. Ontologies*.** Ontology allows the integration of vocabularies and labels across species, across data types (i.e., physiological trace versus gene expression profile), across development, and when scales change. Participants generally agreed that mismatched ontology is acceptable because users can include multiple annotations of the same areas. Participants also agreed

that, although embryonic and fetal ontologies exist, they will require more work to achieve the richness that is necessary to use ontology mapping to resolve adult and embryonic anatomy. Participants discussed the benefits of using ontology across species in the absence of data-driven knowledge to create initial versions of the CCF (i.e., using mouse ontology to help define landmarks for the human brain), and agreed that pilot studies are needed to clarify this issue.

Ontology mapping in the CCF is as important as the coordinate system itself. Users with different experimental needs will likely also have different working languages. If there is no way to effectively translate between different languages, interoperability error may occur: terms from different languages that refer to the same structure or cell may not be recognized, or the same term from two different languages that refers to different structures may not be discriminated. Participants agreed that it would be beneficial to obfuscate ontology terms from the users, and instead use machine language to perform ontology mapping behind the scenes unless there is conflict. However, this goal may not be achievable in the first version of the CCF, and initial users may have to be educated on ontology services.[15]

### 3.4. Integration of Scale

The human body can be clearly viewed at three levels: (1) the gross (or macroscopic) scale, which is the anatomical view of the organs and body systems linked by the vasculature; (2) the fine scale, which encompasses the microscopic view of the human body at the histologic and cellular level; and (3) the mesoscale,[16] a transitional point between the microscopic level, where individual variation is prominent, and the macroscopic level, where a more stable species stereotypy is observed.

***3.4.1. Mapping different scales.*** The gross anatomical scale would be comparable to satellite view in Google Maps, whereas the fine histologic scale would be comparable to street view. The anatomical scale is largely stereotypical, and although some specific inter-individual variability exists, spatial mapping may be intuitive and sufficient for resolution of the CCF at this scale. In contrast, organization of the CCF at the fine scale may be best achieved using the generative probabilistic approach (see Section 2.3). Because the mesoscale is where things tend to break down, participants discussed in detail whether to define scales prior to data acquisition or to allow a data-driven decision of what scales are important. Although they did not reach consensus, they generally agreed that multiple scales will be likely for all systems, and that resolution of these different scales will be key.

***3.4.2. Integrating multiple scales.*** Although there was considerable discussion about efficient integration at different scales, there was general agreement that no one strategy of mapping will work for all systems and scales. Instead, the CCF could be viewed as a sequence of triage

[15] See, for example, Uberon (www.uberon.github.io/), HuDSeN (www.hudsen.eu/), Ontology Lookup Service (https://www.ebi.ac.uk/ols/index), emouseatlas (www.emouseatlas.org/emage/home.php/), MARender (https://github.com/ma-tech/MARender), Human Cell Atlas Ontology (http://github.com/HumanCellAtlas/ontology), The Monarch Initiative (https://monarchinitiative.org/).
[16] Mitra PP, et. al. The circuit architecture of whole brains at the mesoscopic scale. *Neuron*, 2014;83(6):1273-83.

events with relatively coarse positional information on larger structure and higher-resolution for local architecture. Based on previous work in the mouse brain (see Section 2.2) and suggested target tissues for pilot studies (see Section 4.1), organs seem to be a natural level at which to begin developing and piloting the CCF. Thus, in an iterative approach there may be at first development of multiple individual CCF systems for specific organs that will be linked with a hierarchical overarching description or physical map. Google Virtual Human,[17] a collection of rotatable 3D selectable images of organs, may serve as one template for such an overarching physical map.

As an alternative approach, meeting participants discussed using blood vessels to provide a common reference for different organs and major body systems at the gross scale. Although inter-individual variability of the vasculature exists, major vascularization of the organs is both stereotypical and recognizable at the histologic level. Use of blood vessels as a common reference to link different organs in the CCF offers an additional advantage because surgical resection is performed according to blood vessel location. Therefore, information on the positional information of specimens relative to blood vessels is likely to be easily obtained at the point of specimen collection.

### 3.5. Quality Control

Although not specifically discussed, quality control (QC) was identified as an important aspect of the CCF worthy of consideration. Participants agreed on the need to develop an appropriate set of molecular markers of tissue quality, as well as distinct languages for QC at the collection and analysis phases so that the quality of samples and data can be distinguished. The MDP briefly highlighted one study showing that preservation of RNA integrity in several tissues was preserved at different postmortem intervals (PMIs) was an indication of quality,[18] and 70 percent of tumor samples collected produced successful xenografts.

## 4. Next Steps

To build a CCF, meeting participants agreed that it is important to define specific pilot studies for the CCF in the context of an atlas of “normal” human tissues but recognized that it is equally important to implement a generalized solution to “future-proof” the CCF for future applications. Multiple spatial frameworks may be necessary initially to describe the global and local architecture for the specific targets in pilot studies and to match the resolution at which data is captured, but an over-arching hierarchical structure should be defined at the outset to link different CCFs developed as part of the pilot studies (e.g., Google Virtual Human or mapping to the vasculature).

At the level of tissue collection, meeting participants agreed that to maximize the re-use of samples, their origin must be clearly defined, and proper documentation is key (see Section 3.2.1). Metadata will improve in an iterative fashion once pilot studies commence. Meeting

[17] See https://www.biodigital.com/.

[18] Fan J, et. al. Quantification of nucleic acid quality in postmortem tissues from a cancer research autopsy program. *Oncotarget*, 2016;7(41):66906-21.

participants identified several important initial steps for the construction of the CCF, beginning with the tissue collection/processing pipeline:

1. development of a standardized metadata collection tool (e.g., REDCap-based application),
2. consultation with surgeons and pathologists for the creation of standardized ontology based on anatomical language at the point of collection,
3. creation of the anatogram for diagrammatic annotation of specimen location and orientation at collection, and
4. introduction of the tape transfer method in cryosectioning of tissue specimens.

Participants also noted that, for subsequent iterations of the CCF, embryonic and fetal ontologies must be expanded, structural ontology must be resolved with cell-type ontology, and existing ontology mapping tools must be incorporated into the CCF user interface. With the tissue collection pipeline and ontology mapping in place, target tissues and pilot studies must be selected and carried out.

## 4.1. Select Target Tissues for CCF Pilots

Participants agreed on the organ as a natural level of abstraction for the initial CCF and a good target for pilot studies. They suggested several potential tissues/organs:

- ***The eye*.** This organ presents a unique opportunity because it contains all known tissue types, histologically and morphologically, in one relatively small physical space.
- ***The brain*.** This highly structured organ has a wealth of curated knowledge, and ontology can be leveraged to inform initial iteration(s) of a human CCF using the mouse CCF.
- ***The lung*.** The lung represents an intermediate organ in terms of organization: less ordered than the brain, but more ordered than the liver. Of interest, the same cell types within the lung give rise to different cell populations based on their location, providing a unique use case for the CCF.
- ***Prostate cancer*.** The only non-organ-based tissue suggested, prostate cancer benefits from imaging modalities (e.g., fusion MRI) that are not currently available for other tissues/organs. It is also highly relevant from a funding standpoint due to disease prevalence.

## 4.2. Identify Potential Pilot Studies

Meeting organizers provided criteria for consideration of pilot studies: these studies should have a testable milestone within a 1- to 2-year timeframe; may consider approaches to tissue collection, assays, analysis, and/or mapping and visualization; and should clearly state either a scientific or technical goal. Pilot studies may help to address missing labels during collection, how to best QC data, how to make data interoperable, and what system and length of scale should be used for the CCF. Four specific pilot studies were suggested:

***4.2.1. Comparison of tissue sources*.** To determine the optimum source of human tissues, and to empirically determine the resolution of the CCF for tissues of varying inherent complexity,

meeting participants suggested a cross-comparison of two tissues of differing complexity that are obtained from either resection or autopsy. Specimens would be obtained for three to four autopsies, and from resections, to compare complex 3D architecture recreated by the CCF. Liver and skeletal muscle were offered as two tissues that are routinely accessed in many patients at all institutions, and a potential framework would be hypersampling combined with multiplex immunohistochemistry to integrate data with single cell DNA and single-nucleotide variant sequencing for phylogenies.

***4.2.2. Field test for anatogram*.** The purpose of this pilot study would be to determine the minimal granularity of metadata sufficient to accurately identify the specimen's origin. The comparison of diagrammatic and/or photographic documentation (i.e., initial metadata input) about the source anatomical location of the specimen will be compared to the location inferred by the CCF following molecular analyses of the samples. This could be combined with a field test at the point of specimen collection for the digital anatogram pre-loaded on a tablet.

***4.2.3. Molecular characterization of Idiopathic Fibrosis (IPF)*.** IPF has a high public burden but unknown etiology. It is not clear whether IPF is a disease or a syndrome, or whether it is local or systemic. This is in part because of the patchy nature of presentation, and in part due to ontology, because diseases tend to change names as they progress and/or when they appear in other organs. A benefit of studying IPF is heterogeneity within a single sample, which allows for an internal control for the tissue collection and processing pipeline. The goal would be to determine the resolution of the CCF to distinguish between "normal" and diseased tissue.

***4.2.4. Variation in heart thickness as a function of gene expression*.** For this pilot, structural data on heart thickness obtained by structural MRI (or similar methodology) would be obtained from a relatively large group of people (100-200), and gene expression profiles would be captured from a subset of this group. The -omics data from the subset group will be used to predict heart thickness in individuals from the larger group and will be compared to their MRI readings. A potential limitation of this study is the rarity of ventricular biopsies.

# Appendix A: Agenda

**Common Coordinate Framework Meeting**
**December 11–12, 2017**

**Day 1**

| | |
|---|---|
| **8:00 AM – 8:45 AM** | **Registration** |
| **8:45 AM – 9:00 AM** | **Welcome and Opening Remarks**<br>Richard Conroy (NIH, Office of Strategic Coordination)<br>Aviv Regev (Broad Institute)<br>Sarah Teichmann (Sanger Institute)<br>Jonah Cool (Chan Zuckerberg Initiative) |

**Session 1: Framing the Framework**

| | |
|---|---|
| **9:00 AM** | **The Anatomy of the Human Body: what makes us so similar and yet so different**<br>Rosalyn Jurjus (George Washington University) |
| **9:20 AM** | **Facts and Myths of Using Postmortem Tissue for Research**<br>Christine Iacobuzio-Donahue (MSKCC) |
| **9:50 AM** | **HDBR: a fetal tissue resource enabling human developmental research**<br>Susan Lindsay (Newcastle University) |
| **10:10 AM** | **Round-the-Table Discussion – Scope and starting points for a human body Common Coordinate Framework** |
| **10:30 AM** | **Break** |

**Session 2: Using Our Brain**

| | |
|---|---|
| **11:00 AM** | **Brain Common Coordinate Frameworks: extensible principles and new challenges in the single cell analysis era**<br>Ed Lein (Allen Institute) |
| **11:20 AM** | **Ontologies, Atlases and Co-Ordinate Systems, From Flies and Mice to Humans**<br>David Osumi-Sutherland (EBI) |
| **11:40 AM** | **TBD** |
| **12:00 PM** | **Round-the-Table Discussion – Developing a common language for all people** |

| | |
|---|---|
| **12:30 PM** | **Lunch – Meals and light refreshments are at the expense of attendees.** (Attendees will be responsible for meals and/or light refreshments on their own, at their own cost. The government and/or government contractors are not involved in facilitating the provision of food and/or light refreshments). |
| **1:30 – 3:00 PM** | **Working Group Discussions**<br>Group 1: Integrating spatial and omics information over multi-scales – how to streamline and optimize the workflow<br>Lead Discussants: Anne Plant (NIST), Alex Shalek (MIT)<br><br>Group 2: Building the language of a CCF – collecting the right metadata and integrating ontologies<br>Lead Discussants: Laura Clarke (EBI), Maryann Martone (UCSD)<br><br>Group 3: Building the computational infrastructure for storing, visualizing, and searching a human body atlas<br>Lead Discussants: Alex Wiltschko (Google), Robert Murphy (CMU) |
| **3:00 PM – 3:15 PM** | **Break** |
| **3:15 PM – 4:45 PM** | **Sharing Ideas & Group Discussion**<br>1. Working Group 1<br>2. Working Group 2<br>3. Working Group 3 |
| **4:45 PM – 5:00PM** | **Wrap-Up - Day 1**<br>Richard Conroy (NIH, Office of Strategic Coordination)<br>Aviv Regev (Broad Institute)<br>Sarah Teichmann (Sanger Institute)<br>Jonah Cool (Chan Zuckerberg Initiative) |
| **5:00 PM – 6:00PM** | **Demo Session** |

## Day 2

| | |
|---|---|
| **8:45 AM – 9:00 AM** | **Framing of the Day**<br>Richard Conroy (NIH, Office of Strategic Coordination)<br>Aviv Regev (Broad Institute)<br>Sarah Teichmann (Sanger Institute)<br>Jonah Cool (Chan Zuckerberg Initiative) |

**Session 3: Building on Knowledge, Models and Statistics**

| | |
|---|---|
| **9:00 AM** | **Sampling Cells by Organ, by Location and by Individual**<br>John Marioni (EBI) |

| | |
|---|---|
| **9:20 AM** | **Capturing Variation in the Cell Atlas Across Healthy Human Populations**<br>Barbara Engelhardt (Princeton) |
| **9:50 AM** | **TBD**<br>Partha Mitra (CSHL) |
| **10:10 AM** | **Round-the-Table Discussion – Assumptions, inferences and registration – intelligent sampling, using sparse data and comparing individuals** |
| **10:30 AM** | **Break** |
| **11:00 – 12:30 PM** | **Working Group Discussions –** Building a CCF that is robust and provides insights…<br>Working Group 4 …. Across the lifespan<br>Lead Discussants: Marius Linguraru (Children's), Kristin Ardlie (Broad)<br><br>Working Group 5 …. Across inter-individual variation<br>Lead Discussants: Jason Swedlow (Dundee), Alexis Battle (JHU)<br><br>Working Group 6 …. Across the health-disease continuum<br>Lead Discussants: James Gee (UPenn), Zorina Galis (NHLBI) |
| **12:30 PM** | **Lunch – Meals and light refreshments are at the expense of attendees.** (Attendees will be responsible for meals and/or light refreshments on their own, at their own cost. The government and/or government contractors are not involved in facilitating the provision of food and/or light refreshments). |
| **1:30 PM – 3:00 PM** | **Sharing Ideas & Group Discussion**<br>1. Working Group 4<br>2. Working Group 5<br>3. Working Group 6 |
| **3:00 – 3:30 PM** | **Wrap-Up and Adjourn Meeting**<br>Richard Conroy (NIH, Office of Strategic Coordination)<br>Aviv Regev (Broad Institute)<br>Sarah Teichmann (Sanger Institute)<br>Jonah Cool (Chan Zuckerberg Initiative) |

# Appendix A: Participants List

Tarmo Äijö, Flatiron Institute
James Anderson, Office of the Director, NIH
Kristin Ardlie, Broad Institute
Chris Armit, GigaScience, BGI-Hong Kong
David Balasundaram, Center for Scientific Review, NIH
Alexis Battle, The Johns Hopkins University
Andrea Beckel-Mitchener, National Institute of Mental Health, NIH
Richard Bonnea, Flatiron Institute
Kristen Browne, National Library of Medicine, NIH
Albert Burger, Heriot-Watt University
Robert Carter, National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH
Laura Clarke, European Bioinformatics Institute
Richard Conroy, Office of Strategic Coordination, NIH
Jonah Cool, Chan Zuckerberg Initiative
Barbara Engelhardt, Princeton University
Christopher Ferguson, Booz Allen Hamilton
Jeremy Freeman, Chan Zuckerberg Initiative
Zorina Galis, National Heart, Lung, and Blood Institute, NIH
Deep Ganguli, Chan Zuckerberg Initiative
James Gee, University of Pennsylvania
Ronald Germain, National Institute of Allergy and Infectious Diseases, NIH
Sharmistha Ghosh-Jangijian, National Cancer Institute, NIH
Daniel Gilchrist, National Human Genome Research Institute, NIH
Katrina Gold, Wellcome Trust
Barbara Goldstein, National Institute of Standards and Technology
Sean Hanlon, National Cancer Institute, NIH
Martin Hemberg, Wellcome Trust Sanger Institute
Michelle Holko, Booz Allen Hamilton
Shannon Hughes, National Cancer Institute, NIH
Christine Iacobuzio-Donahue, Memorial Sloan Kettering Cancer Center
Etai Jacob, DC Fatherhood Initiative, HHS
Stephen Jett, National Cancer Institute, NIH
George Johnson, Duke Center for In Vivo Microscopy
Abdo Jurjus, The George Washington University
Rosalyn Jurjus, The George Washington University
Jacob Kagan, National Cancer Institute, NIH
Jane Lee, Broad Institute
Ed Lein, Allen Institute for Brain Science
Sara Lin, National Heart, Lung, and Blood Institute, NIH
Susan Lindsay, Newcastle University
Marius Linguraru, Children's National Health System
Mary Maleckar, Allen Institute for Cell Science

Christine Maric-Bilkan, National Heart, Lung, and Blood Institute, NIH
John Marioni, European Molecular Biology Laboratory
Maryann Martone, University of California, San Diego
Andrea Mitchener, National Institute of Mental Health, NIH
Partha Mitra, Cold Spring Harbor Laboratory
Evan Molinelli, BioDigital
David Morse, University of Cambridge
Robert Murphy, Carnegie Mellon University
David Osumi-Sutherland, European Bioinformatics Institute
Dana Peer, Sloan Kettering Institute
Ajay Pillai, National Human Genome Research Institute, NIH
Anne Plant, National Institute of Standards and Technology
Aviv Regev, Broad Institute
Sonia Rosenfield, National Cancer Institute, NIH
Ananda Roy, National Institutes of Health
Joshua Sanes, Harvard University
Sirarat Sarntivijai, European Bioinformatics Institute
Kamran Sayrafian, National Institute of Standards of Technology
Frank Sculli, BioDigital
Alex Shalek, Massachusetts Institute of Technology
Bishen Singh, National Institute of Mental Health, NIH
Jessica Smith, Office of Strategic Coordination, NIH
Pothur Srinivas, National Heart, Lung, and Blood Institute, NIH
Ram Sriram, National Institute of Standards and Technology
Sudhir Srivastava, National Cancer Institute, NIH
Mike Stubbington, Wellcome Trust Sanger Institute
Jason Swedlow, University of Dundee
Norbert Tavares, Office of the Director, NIH
Sarah Teichmann, Wellcome Trust Sanger Institute
Reiko Toyama, National Institute of Child Health and Human Development, NIH
Caitlin Trasande, Memorial Sloan Kettering
Jose Velazquez, National Institute on Aging, NIH
Jeremy Wertheimer, Google
Elizabeth Wilder, Office of the Director, NIH
Alex Wiltschko, Google
Dan Xi, National Cancer Institute, NIH